Applications and Applied Mathematics: An International Journal (AAM)

Volume 16 | Issue 1 Article 4

6-2021

Nonparametric Estimation of the Conditional Distribution Function For Surrogate Data by the Regression Model

Imane Metmous
Djillali Liabes University

Mohammed K. Attouch
Djillali Liabes University

Boubaker Mechab
Djillali Liabes University

Torkia Merouan
Djillali Liabes University

Available at
http://pvamu.edu/aam
Appl. Appl. Math.
ISSN: 1932-9466
Vol. 16, Issue 1 (June 2021), pp. 55–74

Applications and Applied Mathematics:
An International Journal
(AAM)

Nonparametric Estimation of the Conditional Distribution Function For Surrogate Data by the Regression Model

[1]**Imane Metmous,** [2]**Mohammed Kadi Attouch,** [3*]**Boubaker Mechab and** [4]**Torkia Merouan**

Department of Probability and Statistics
Djillali Liabes University
Laboratory of Statistics and Stochastic Processes
Sidi Bel Abbes 22000, Algeria
[1]metmous_imane@yahoo.com; [2]attou_kadi@yahoo.fr;
[3*]mechaboub@yahoo.fr; [4]merouan-to@hotmail.com

*Corresponding Author

Received: July 10, 2020; Accepted: April 20, 2021

Abstract

The main objective of this paper is to estimate the conditional cumulative distribution using the nonparametric kernel method for a surrogated scalar response variable given a functional random one. We introduce the new kernel type estimator for the conditional cumulative distribution function (*cond-cdf*) of this kind of data. Afterward, we estimate the quantile by inverting this estimated *cond-cdf* and state the asymptotic properties. The uniform almost complete convergence (with rate) of the kernel estimate of this model and the quantile estimator is established. Finally, a simulation study completed to show how our methodology can be adopted.

Keywords: Functional Data Analysis (FDA); Conditional distribution function; Nonparametric kernel estimation; Surrogate data; Conditional quantile

MSC 2010 No.: 62G05, 62E20; 62G20

1. Introduction

Conditional distribution function (CFD) estimation is an essential field in nonparametric statistical analysis; this technique helps us understand the relationship between a response variable and covariates set.

One of the branches of modern statistics is Functional Data Analysis (FDA). This has become possible thanks to the computing techniques' progress, both in terms of memory and storage capacities, which allows us to consider increasingly voluminous data, regarded as an observation of curve or surface. The reader can consult the books of Ramsay and Silverman (1997), Ramsay and Silverman (2002), Bosq (2000) and Ferraty and Vieu (2006), which offer a good introduction both for the theoretical or applied aspect with various applications, including economics, sociology, and biology. It should be noted that extensions of probability theory to random variables taking values in normed spaces (e.g., Banach and Hilbert spaces), including extensions of some classical theorems, are handy tools in the literature.

Note first that the study of the conditional distribution function of real data was obtained in the early 1960s by Roussas (1968) who studied the kernel estimator's asymptotic properties conditional distribution function where it showed convergence in probability. In the case of functional data, many researchers have been interested in the study of this function. For example, we cite, Ferraty et al. (2006) who estimate the conditional distribution characteristics in nonparametric functional models. In the same framework, Ferraty et al. (2005) use the conditional distribution function to obtain a nonparametric estimator of the conditional quantile when the data is weakly dependent.

It should be noted that most of the results involved in the nonparametric literature (and not only on the conditional distribution) only deal with completely observed samples. While in many practical works, including, for example, sample survey, reliability, or pharmaceutical tracing where data is often observed incompletely, and parts of the responses are missing randomly (MAR).

The most popular method to involve missing data is the imputation method that fills or retrieves the missing data in the response variable Y. In this context, we can cite various works that used this technique. We can cite Yates (1933) for the linear regression model. The kernel estimation of the mean functions is considered in Cheng (1994), the nearest neighbor imputation for the data survey is addressed in Chen and Shao (2000), the robust regression model with missing data is considered in Pérez-González et al. (2009), the asymptotic properties of the regression operator estimator when the regressor is functional and completely observed, and that missing data at random in the scalar response variable are investigated in Ferraty et al. (2013), in the case of dependent data, the reader may refer to Ling et al. (2015). In this work, we investigate the unavailability of response data because sometimes it is default or very expensive to measure some response observations; the main idea is to recover (or fill) this missing data by a surrogate validation data set. In this context, we cite Duncan and Hill (1985), Wittes et al. (1989), Carroll and Wand (1991) and Pepe (1992). The principle of this method is to incorporate both surrogate data and the corresponding observations of the covariate X.

This paper aims to study the conditional models (conditional distribution function and the conditional quantile) for missing response by the kernel method. We explore in this work the aspect of missing data in the response variable. First, we consider the estimator of the conditional distribution for complete data, then by using the validation data set (see, Ibrahim et al. (2020) and Wang (2006)), we build our new estimator with surrogate data and we obtain some asymptotic results for the conditional distribution and the quantiles. In the end, we realized a simulation study to improve the efficacity of our estimator.

The rest of the paper is organized as follows. We present our model in Section 2. The required notations and assumptions are introduced in Section 3. The main results of strong uniform consistency (with rate) and the quantile estimation as a direct consequence of our asymptotic result obtained from CFD estimation are formulated in Section 4. For the numerical results, a simulation study that shows the performance of the proposed estimator is presented in Section 5.

2. Model and Estimator of the Conditional Distribution Function

2.1. Estimation of the *cond-cdf* with complete data

Let $(X_i, Y_i)_{i=1,\ldots,N}$ be a random variables independent and identically distributed as (X, Y), where $X \in \mathcal{F}$, Y take values in $\mathbb{R}$ and $(\mathcal{F}, d)$ is a semi metric space with a metric $d(.,.)$. The conditional cumulative distribution function of Y given $X = x$, denoted by $F^x(.)$ is defined

$$F^x(.) = \mathbb{P}(Y \leq y | X = x), \quad \forall y \in \mathbb{R},$$

and by the regression model, we have

$$\mathbb{E}\left[H\left(\frac{y - Y_i}{h_H}\right) \Big| X_i = x \right] \xrightarrow{h_H \to 0} F^x(y),$$

where $H(.)$ is a cumulative distribution function and h_H is a sequence of positive real numbers tending to 0 when n go to infinity.

The estimator of conditional distribution function by the kernel method defined by

$$\widehat{F}_C^x(y) = \frac{\sum_{i=1}^N H\left(\frac{y-Y_i}{h_H}\right) K\left(\frac{d(x,X_i)}{h_K}\right)}{\sum_{i=1}^N K\left(\frac{d(x,X_i)}{h_K}\right)}, \quad \forall y \in \mathbb{R},\ \forall x \in \mathcal{F}, \tag{1}$$

where $K(.)$ is a kernel function and h_K is a bandwidth sequence tend toward 0.

2.2. Estimation of the *cond-cdf* with surrogate data

We have the sample set of the size N and the validation set of size n, where the observations are independent and identically distributed. Here, Y is not accessible (available), so we replaced it by a surrogate variable $\widetilde{Y}$.

Let V the index set of the sampled validation set and $\bar{V} = \{1, ..., N\} \setminus V$. Note that, for the surrogate data we have

$$\mathbb{E}\left[H\left(\frac{y-Y_j}{h_H}\right)\Big|X_j, \widetilde{Y}_j\right] \xrightarrow{h_H \to 0} F^x(y),$$

and

$$\mathbb{E}\left[\mathbb{E}\left[H\left(\frac{y-Y_j}{h_H}\right)\Big|X_j, \widetilde{Y}_j\right]\Big|X_j = x\right] = \mathbb{E}\left[H\left(\frac{y-Y_j}{h_H}\right)\Big|X_j = x\right], \tag{2}$$

then, the distribution function can be estimated by

$$\widehat{F}^x(y) = \frac{\sum_{i\in V} H\left(\frac{y-Y_i}{h_H}\right) K\left(\frac{d(x,X_i)}{h_K}\right) + \sum_{j\in\bar{V}} u(X_j, \widetilde{Y}_j) K\left(\frac{d(x,X_j)}{h_K}\right)}{\sum_{i=1}^{N} K\left(\frac{d(x,X_i)}{h_K}\right)},$$

where

$$u(X_j, \widetilde{Y}_j) = \mathbb{E}\left[H\left(\frac{y-Y_j}{h_H}\right)\Big|X_j, \widetilde{Y}_j\right],$$

and the function $u(.,.)$ is unknown.

So, we estimate this function by validation data set:

$$\widehat{u}(X_j, \widetilde{Y}_j) = \frac{\sum_{i\in V} H\left(\frac{y-Y_i}{h_H}\right) W\left(\frac{d(X_j,X_i)}{a_n}, \frac{\widetilde{Y}_j - \widetilde{Y}_i}{a_n}\right)}{\sum_{i\in V} W\left(\frac{d(X_j,X_i)}{a_n}, \frac{\widetilde{Y}_j - \widetilde{Y}_i}{a_n}\right)},$$

and $W(.,.)$ is the two-dimensional kernel function in $\mathcal{F} \times \mathbb{R}$ and a_n is a sequence of real number which tend to zero when n tend to infinity.

3. Assumptions

Let $S_{\mathcal{F}}$ be some subset of $\mathcal{F}$ such that $S_{\mathcal{F}} \subset \bigcup_{k=1}^{d_n} B(x_k, r_n)$, where $x_k \in \mathcal{F}$, and (d_n) is a sequence of integers which satisfies the assumption (A5).

Let us introduce $B(x, h_K)$ a ball of the center x and radius h_K defined as $B(x, h_K) = \{x_1 \in \mathcal{F} : d(x_1, x) \leq h_K\}$. Furthermore, we have x a fixed point in $\mathcal{F}$, and $S_{\mathbb{R}}$ a fixed compact subset of $\mathbb{R}$.

Our assumptions are gathered below for easy references.

(A1) $\forall h_K > 0$, $\mathbb{P}(X \in B(x, h_K)) =: \phi(h_K) > 0$.

(A2) The operators $F^x(.)$ and $u(.,.)$ are Lipschitzian, such that, $\forall (y_1, y_2) \in S^2_{\mathbb{R}}$, $\forall (x_1, x_2) \in S^2_{\mathcal{F}}$ and $C, A_1, A_2 > 0$,

(a) $|F^{x_1}(y_1) - F^{x_2}(y_2)| \leq C\left(d(x_1, x_2)^{A_1} + |y_1 - y_2|^{A_2}\right)$.

(b) $|u(x_1, y_1) - u(x_2, y_2)| \leq C\left(d(x_1, x_2)^{A_1} + |y_1 - y_2|^{A_2}\right)$.

(A3) The distribution function $H(.)$ satisfy

$$\begin{cases} \forall\,(y_1,y_2)\in\mathbb{R}^2, |H\,(y_1)-H\,(y_2)|\leq C\,|y_1-y_2|\,, \\ \int |t|^{A_2}\,H'\,(t)\,\mathrm{d}t<\infty. \end{cases}$$

(A4) The bandwidths h_K and a_n satisfy

$$\lim_{N\to\infty} h_K=\lim_{n\to\infty} a_n=0 \quad \text{and} \quad \lim_{N\to\infty} N\phi(h_K)=+\infty,$$

and

$$\lim_{N\to\infty}\frac{\log N}{N\phi(h_K)}=0 \quad \text{and} \quad \lim_{n\to\infty}\frac{\log n}{n\phi(a_n)}=0.$$

(A5) For some $\beta>0$,

$$\lim_{N\to\infty} h_H=0 \quad \text{with} \lim_{N\to\infty} N^{\beta}h_H=\infty,$$

and for $r_n=O\left(\dfrac{\log N}{N}\right)$ the sequence d_n satisfy

$$\frac{\log^2 N}{N\phi\,(a_n)}\leq d_n\leq\frac{N\phi\,(a_n)}{\log N} \quad \text{and} \quad \sum_{n=1}^{\infty} n^{\beta}\exp\left\{(1-\eta)\log d_n\right\}<\infty \text{ where } \beta>0 \text{ and } \eta>1. \tag{3}$$

(A6) The kernel $K(.)$ is a continuous function from $\mathbb{R}$ into $\mathbb{R}^+$ such that $\int K=1$, and there exist some positive constants C and C' such that

$$C\mathbf{1}_{(0,1)}\leq K\leq C'\mathbf{1}_{(0,1)}, \tag{4}$$

where $\mathbf{1}_A$ denotes the indicator function on the set A.

We assume the two-dimensional kernel $W(x,y)=W_1(x)W_2(y)$ is a continuous function with a compact support satisfies (4), however, there exist positive finite real constants C_3 and C_4, such that

$$C_3\phi(a_n)\leq\mathbb{E}\left[W\left(\frac{d(X_j,X_i)}{a_n},\frac{\widetilde{Y}_j-\widetilde{Y}_i}{a_n}\right)\right]\leq C_4\phi(a_n).$$

Remark 3.1.

The concentration assumption $(A1)$ depend to the distribution of X and has an important role, which is linked with the semi-metric $d(.,.)$. Note that the correct choice for $d(.,.)$ is through the corresponding function $\phi(.)$ a key to the curse of dimensionality. The assumption $(A2)$ is linked with the nonparametric structure of the model and it's used it for determine the bias term. The assumptions $(A3)-(A6)$ are a technical condition similar to the hypothesis in Ferraty et al. (2006) for obtain our results.

4. Results

4.1. Uniform almost complete consistency

The uniform almost complete $(O_{a.co.})$ convergence of $\widehat{F}^x(.)$ is given by the following Theorem and Lemmas.

Theorem 4.1.

Under assumptions $(A1)-(A6)$, we obtain

$$\sup_{x\in S_{\mathcal{F}}}\sup_{y\in S_{\mathbb{R}}}|\widehat{F}^x(y)-F^x(y)| = O(h_K^{A_1}+h_H^{A_2}+a_n^{A_1})+O_{\text{a.co.}}\left(\sqrt{\frac{\log d_n}{n\phi(a_n)}}\right)+O_{\text{a.co.}}\left(\sqrt{\frac{\log N}{N\phi(h_K)}}\right).$$

Proof:

Let $\widehat{F}_N^x(y)$ and $\widehat{F}_D^x(y)$, defined by

$$\widehat{F}_N^x(y) = \frac{1}{N}\sum_{i\in V}\frac{H\left(\frac{y-Y_i}{h_H}\right)K\left(\frac{d(x,X_i)}{h_K}\right)}{\mathbb{E}\left[K\left(\frac{d(x,X_i)}{h_K}\right)\right]}+\frac{1}{N}\sum_{j\in\bar{V}}\frac{\widehat{u}(X_j,\widetilde{Y}_j)K\left(\frac{d(x,X_j)}{h_K}\right)}{\mathbb{E}\left[K\left(\frac{d(x,X_j)}{h_K}\right)\right]},$$

and

$$\widehat{F}_D^x = \frac{1}{N}\sum_{i=1}^{N}\frac{K\left(\frac{d(x,X_i)}{h_K}\right)}{\mathbb{E}\left[K\left(\frac{d(x,X_i)}{h_K}\right)\right]}.$$

The proof is based on the following decomposition and the Lemmas 4.2, 4.3 and 4.4 given below:

$$\widehat{F}^x(y)-F^x(y) = \frac{1}{\widehat{F}_D^x}\left\{\left(\widehat{F}_N^x(y)-\mathbb{E}[\widehat{F}_N^x(y)]\right)-\left(F^x(y)-\mathbb{E}[\widehat{F}_N^x(y)]\right)\right\} - \frac{F^x(y)}{\widehat{F}_D^x}\left\{\widehat{F}_D^x-1\right\}. \tag{5}$$

■

Auxiliary results

We put the quantities, for $x\in\mathcal{F}$, $(y,\widetilde{y})\in\mathbb{R}^2$ and $i,j=1,...,N$:

$$K_i := K\left(\frac{d(x,X_i)}{h_K}\right),\ H_i(y) := H\left(\frac{y-Y_i}{h_H}\right), W_{ij} := W\left(\frac{d(X_j,X_i)}{a_n},\frac{\widetilde{Y}_j-\widetilde{Y}_i}{a_n}\right).$$

We note for $j\in\bar{V}$:

$$\widehat{u}(X_j,\widetilde{Y}_j) = \frac{\sum_{i\in V}H_i(y)W_{ij}}{\sum_{i\in V}W_{ij}} := \frac{\widehat{u}_N^x(y)}{\widehat{u}_D^x}.$$

We need the following lemma to establish the uniform almost complete convergence.

Lemma 4.1.

Under assumptions $(A1)-(A6)$, we get

- $F_1=\sup\limits_{x\in S_{\mathcal{F}}}|\widehat{u}_D^x-1]|=O_{\text{a.co.}}\left(\sqrt{\frac{\log d_n}{n\phi(a_n)}}\right)$ and $\sum\limits_{n=1}^{\infty}\mathbb{P}\left(|\widehat{u}_D^x|\leq 1/2\right)<\infty$.
- $F_2=\sup\limits_{x\in S_{\mathcal{F}}}\sup\limits_{y\in S_{\mathbb{R}}}|\widehat{u}_N^x(y)-\mathbb{E}[\widehat{u}_N^x(y)]|=O_{\text{a.co.}}\left(\sqrt{\frac{\log d_n}{n\phi(a_n)}}\right)$.
- $F_3=\sup\limits_{x\in S_{\mathcal{F}}}\sup\limits_{y\in S_{\mathbb{R}}}|u(x_j,\widetilde{y}_j)-\mathbb{E}[\widehat{u}_N^x(y)]|=O(a_n^{A_1})+O(h_H^{A_2})$.

Proof:

(1) As F_1 is a particular case of F_2 (by taking $H(.)\equiv 1$), then the proof will be omitted. Now, we have

$$\mathbb{P}\left(|\widehat{u}_D^x|\leq 1/2\right)\leq\mathbb{P}\left(|\widehat{u}_D^x-1|>1/2\right),$$

thus, by applying the result above, we get $\sum\limits_{i=1}^{\infty}\mathbb{P}\left(|\widehat{u}_D^x|\leq 1/2\right)<\infty$.

(2) We conceive the following decomposition, where for all $x\in S_{\mathcal{F}}$, we set $k\left(x\right)=\underset{k\in\{1,\ldots,d_n\}}{\operatorname{argmin}}|x-x_k|$ and we use the compactness of $S_{\mathbb{R}}$, where, we can write $S_{\mathbb{R}}\subset\bigcup_{j=1}^{q_n}S_j, S_j=(l_j-l_n,l_j+l_n)$ and take $y_t=\underset{l\in\{l_1,\ldots,l_{q_n}\}}{\operatorname{argmin}}|y-l|$, to obtain

$$\begin{aligned}\underbrace{\sup_{x\in S_{\mathcal{F}}}\sup_{y\in S_{\mathbb{R}}}|\widehat{u}_N^x(y)-\mathbb{E}[\widehat{u}_N^x\left(y\right)]|}_{F_2}\leq&\underbrace{\sup_{x\in S_{\mathcal{F}}}\sup_{y\in S_{\mathbb{R}}}|\widehat{u}_N^x\left(y\right)-\widehat{u}_N^{x_{k(x)}}\left(y\right)|}_{P_1}\\&+\underbrace{\sup_{x\in S_{\mathcal{F}}}\sup_{y\in S_{\mathbb{R}}}|\widehat{u}_N^{x_{k(x)}}\left(y\right)-\widehat{u}_N^{x_{k(x)}}\left(y_t\right)|}_{P_2}\\&+\underbrace{\sup_{x\in S_{\mathcal{F}}}\sup_{y\in S_{\mathbb{R}}}|\widehat{u}_N^{x_{k(x)}}\left(y_t\right)-\mathbb{E}[\widehat{u}_N^{x_{k(x)}}\left(y_t\right)]|}_{P_3}\\&+\underbrace{\sup_{x\in S_{\mathcal{F}}}\sup_{y\in S_{\mathbb{R}}}|\mathbb{E}[\widehat{u}_N^{x_{k(x)}}\left(y_t\right)]-\mathbb{E}[\widehat{u}_N^{x_{k(x)}}\left(y\right)]|}_{P_4}\\&+\underbrace{\sup_{x\in S_{\mathcal{F}}}\sup_{y\in S_{\mathbb{R}}}|\mathbb{E}[\widehat{u}_N^{x_{k(x)}}\left(y\right)]-\mathbb{E}[\widehat{u}_N^{x}\left(y\right)]|}_{P_5}.\end{aligned}$$

- For P_1 and P_5, we have from $(A3)$ and the boundness of $W(.,.)$ we can write

$$\begin{aligned}P_1&\leq\frac{C}{\phi(a_n)}\sup_{x\in S_{\mathcal{F}}}\sup_{y\in S_{\mathbb{R}}}\frac{1}{n}\sum_{i\in V}|W(x,\tilde{y})-W(x_{k(x)},\tilde{y})|\\&\leq\frac{Cd_nq_n}{a_n\phi(a_n)},\end{aligned}$$

and analogously, for P_2 we obtain

$$P_2 \leq \frac{Cd_nq_n}{a_n\phi(a_n)}\mathbf{1}_{B(x,a_n)\cup B(x_{k(x)},a_n)},$$

by applying Bernstein's inequality, with

$$Z_i = \frac{\epsilon}{a_n\phi(a_n)}\mathbf{1}_{B(x,a_n)\cup B(x_{k(x)},a_n)},$$

which gives, for n tending to infinity,

$$P_1 = O\left(\sqrt{\frac{\log d_n}{n\phi\left(a_n\right)}}\right) \quad \text{and} \quad P_2 = O\left(\sqrt{\frac{\log d_n}{n\phi\left(a_n\right)}}\right).$$

Moreover, using the fact that $P_5 \leq P_1$ and $P_4 \leq P_2$ to get, for n tending to infinity,

$$P_5 = O\left(\sqrt{\frac{\log d_n}{n\phi\left(a_n\right)}}\right) \quad \text{and} \quad P_4 = O\left(\sqrt{\frac{\log d_n}{n\phi\left(a_n\right)}}\right).$$

- Now concerning P_3. For all $\eta > 0$, we have

$$\mathbb{P}\left(P_3 > \eta\sqrt{\frac{\log d_n}{n\phi(a_n)}}\right) \leq q_nd_n \max_{x_k\in\{1,\ldots,d_n\}}\max_{y_t\in\{1,\ldots,t_{q_n}\}} \mathbb{P}\left(|\widehat{u}_N^{x_{k(x)}}(y_t) - \mathbb{E}[\widehat{u}_N^{x_{k(x)}}(y_t)]| > \eta\sqrt{\frac{\log d_n}{n\phi(a_n)}}\right),$$

we can use the Bernstein's exponential inequality to Γ_i, where

$$\Gamma_i = \frac{1}{n\phi(a_n)}\left\{W_{i,j}(x_{k(x)},y_t)H_i(y_t) - \mathbb{E}\left[W_{i,j}(x_{k(x)},y_t)H_i(y_t)\right]\right\}, \quad \text{for } j \in \bar{V},$$

and we have $|\Gamma_i| \leq C_4/\phi(a_n),\ \mathbb{E}|\Gamma_i|^2 \leq C/\phi(a_n)$.

However, take $C\eta^2 = 2\beta$ and $q_n = O(l_n^{-1})$, we get

$$q_nd_n\mathbb{P}\left(\sup_{x\in S_{\mathcal{F}}}\sup_{y\in S_{\mathbb{R}}}|\widehat{u}_N^{x_{k(x)}}(y_t) - \mathbb{E}[\widehat{u}_N^{x_{k(x)}}(y_t)]| > \eta\sqrt{\frac{\log d_n}{n\phi(a_n)}}\right) \leq q_nd_n2\exp\{-C\eta^2\ln d_n\},$$

then, by $(A6)$ we get

$$P_3 = O_{\text{a.co.}}\left(\sqrt{\frac{\log d_n}{n\phi(a_n)}}\right). \tag{6}$$

(3) We have for $j \in \bar{V}$:

$$\begin{aligned} F_3 &:= \mathbb{E}[\widehat{u}_N^x(y)] - u(x,\widetilde{y}) \\ &= \mathbb{E}\left[W_{ij}\left(\mathbb{E}\left(H_1(y)\,|X,\widetilde{Y}\right) - u(x,\widetilde{y})\right)\right], \end{aligned}$$

and we have $\mathbb{E}(H_1(y)|X,\widetilde{Y}) = u(X,\widetilde{Y})$, then, from $(A2)$, we get

$$|u(X,\widetilde{Y}) - u(x,\widetilde{y})| \leq C(a_n^{A_1} + h_H^{A_2}).$$

Finally, from (F_1), (F_2) and (F_3), we finished the proof of Lemma 4.1. ∎

Lemma 4.2.

Under the assumptions $(A1)-(A6)$, we obtain

$$\sup_{x\in S_{\mathcal{F}}}\sup_{y\in S_{\mathbb{R}}}|F^x(y)-\mathbb{E}[\widehat{F}_N^x(y)]|=O(h_K^{A_1})+O(h_H^{A_2})+O(a_n^{A_1})+O_{\text{a.co.}}\left(\sqrt{\frac{\log d_n}{n\phi(a_n)}}\right).$$

Proof:

We have $|V|=n, |\bar{V}|=N-n$,

$$\begin{aligned}F^x(y)-\mathbb{E}[\widehat{F}_N^x(y)] &= F^x(y)-\mathbb{E}\left[n\frac{H_1(y)K_1}{\mathbb{E}[K_1]}+(N-n)\frac{\widehat{u}(X_j,\widetilde{Y}_j)K_1}{\mathbb{E}[K_1]}\right]\\ &= F^x(y)-n\mathbb{E}\left[\frac{H_1(y)K_1}{E[K_1]}\right]-(N-n)\mathbb{E}\left[\frac{\widehat{u}(X_j,\widetilde{Y}_j)K_j}{\mathbb{E}[K_1]}\right]:=T_1+T_2.\end{aligned}$$

- Concerning the term T_1:

$$\begin{aligned}F^x(y)-\mathbb{E}\left[\frac{H_1(y)K_1}{\mathbb{E}[K_1]}\right] &= F^x(y)-\mathbb{E}\left[\mathbb{E}\left[\frac{H_1(y)K_1}{\mathbb{E}[K_1]}\Big|X_1\right]\right]\\ &= F^x(y)-\mathbb{E}[H_1(y)|X_1].\end{aligned}$$

  We know that

$$\mathbb{E}[H_1(y)|X_1]=\int_{\mathbb{R}}H'(t)F^{X_1}(y-h_Ht)\mathrm{d}t,$$

  and

$$|\mathbb{E}[H_1(y)|X_1]-F^x(y)|\leq\int_{\mathbb{R}}H'(t)|F^{X_1}(y-h_Ht)-F^x(y)|\mathrm{d}t.$$

  So, from $(A2)$, we get

$$|\mathbb{E}[H_1(y)|X_1]-F^x(y)|\leq C\int_{\mathbb{R}}H'(t)\left(h_K^{A_1}+|t|_2^Ah_H^{A_2}\right)\mathrm{d}t,$$

  then, $T_1=O(h_K^{A_1})+O(h_H^{A_2})$.
- Concerning the term T_2:

$$\begin{aligned}F^x(y)-\mathbb{E}\left[\frac{\widehat{u}(X_j,\widetilde{Y}_j)K_1}{\mathbb{E}[K_1]}\right] &= \mathbb{E}\left(u(X_j,\widetilde{Y}_j)-\widehat{u}(X_j,\widetilde{Y}_j)\frac{K_1}{\mathbb{E}[K_1]}\right)\\ &\quad+\mathbb{E}\left(F^x(y)-H_1(y)\frac{K_1}{\mathbb{E}[K_1]}\right)\\ &\quad+\mathbb{E}\left(\left(H_1(y)-u(X_j,\widetilde{Y}_j)\right)\frac{K_1}{\mathbb{E}[K_1]}\right).\end{aligned}$$

  Thus,

(a) Firstly, we have

$$\sup_{x\in S_{\mathcal{F}}}\sup_{y\in S_{\mathbb{R}}}\left|\mathbb{E}\left(u(X_j,\widetilde{Y}_j)-\widehat{u}(X_j,\widetilde{Y}_j)\frac{K_1}{\mathbb{E}\left[K_1\right]}\right)\right|=O\left(\sup_{x\in S_{\mathcal{F}}}\sup_{y\in S_{\mathbb{R}}}\left|u(x,y)-\widehat{u}(x,y)\right|\right),$$

by the following decomposition for $j\in\bar{V}$:

$$\begin{aligned}\widehat{u}(X_j,\widetilde{Y}_j)-u(X_j,\widetilde{Y}_j)&=-\frac{u}{\widehat{u}_D^x}\left(\widehat{u}_D^x-1\right)+\frac{1}{\widehat{u}_D^x}\left\{\widehat{u}_N^x(y)-\mathbb{E}[\widehat{u}_N^x(y)]-(u-\mathbb{E}[\widehat{u}_N^x(y)])\right\}\\&:=-\frac{u}{\widehat{u}_D^x}T_{2,1}+\frac{1}{\widehat{u}_D^x}\left(T_{2,2}-T_{2,3}\right),\end{aligned}$$

then, from (Lemma 4.1), we get

$$T_{2,1}=T_{2,2}=O_{\text{a.co.}}\left(\sqrt{\frac{\log d_n}{n\phi(a_n)}}\right)\text{ and }T_{2,3}=O(a_n^{A_1})+O(h_H^{A_2}).$$

(b) Secondly, we have

$$\left|\mathbb{E}\left(F^x(y)-H_1(y)\frac{K_1}{\mathbb{E}\left[K_1\right]}\right)\right|=|F^x(y)-\mathbb{E}[H_1(y)|X_1]|,$$

and $\mathbb{E}[H_1(y)|X_1]=\int_{\mathbb{R}}H'(t)F^X(y-h_Ht)\mathrm{d}t$, so, from the hypothesis $(A2)$, we get

$$\left|\mathbb{E}\left(F^x(y)-H_1(y)\frac{K_1}{\mathbb{E}\left[K_1\right]}\right)\right|=O(h_K^{A_1})+O(h_H^{A_2}).\tag{7}$$

(c) Thirdly, its clear that after (b), we get

$$\left|\mathbb{E}\left(\left(H_1(y)-u(X_j,\widetilde{Y}_j)\right)\frac{K_1}{\mathbb{E}\left[K_1\right]}\right)\right|=0.\tag{8}$$

Finally, from T_1 and T_2 the proof of Lemma 4.2 is achieved. ■

Lemma 4.3.

Under the assumptions $(A1)$ and $(A3)-(A6)$, we obtain

$$\sup_{x\in S_{\mathcal{F}}}\sup_{y\in S_{\mathbb{R}}}|\widehat{F}_N^x(y)-\mathbb{E}[\widehat{F}_N^x(y)]|=O_{\text{a.co.}}\left(\sqrt{\frac{\log N}{N\phi(h_K)}}\right).$$

Proof:

We keep the same notation used previously, in the definitions of $k(x)$ and y_t. The proof is based on

the following decomposition:

$$\begin{aligned}\sup_{x\in S_{\mathcal{F}}}\sup_{y\in S_{\mathbb{R}}}|\widehat{F}_N^x(y)-\mathbb{E}[\widehat{F}_N^x(y)]| &\leq \sup_{x\in S_{\mathcal{F}}}\sup_{y\in S_{\mathbb{R}}}|\widehat{F}_N^x(y)-\widehat{F}_N^{x_{k(x)}}(y)|+\sup_{x\in S_{\mathcal{F}}}\sup_{y\in S_{\mathbb{R}}}|\widehat{F}_N^{x_{k(x)}}(y)-\widehat{F}_N^{x_{k(x)}}(y_t)|\\ &+\sup_{x\in S_{\mathcal{F}}}\sup_{y\in S_{\mathbb{R}}}|\widehat{F}_N^{x_{k(x)}}(y_t)-\mathbb{E}[\widehat{F}_N^{x_{k(x)}}(y_t)]|\\ &+\sup_{x\in S_{\mathcal{F}}}\sup_{y\in S_{\mathbb{R}}}|\mathbb{E}[\widehat{F}_N^{x_{k(x)}}(y_t)]-\mathbb{E}[\widehat{F}_N^{x_{k(x)}}(y)]|\\ &+\sup_{x\in S_{\mathcal{F}}}\sup_{y\in S_{\mathbb{R}}}|\mathbb{E}[\widehat{F}_N^{x_{k(x)}}(y)]-\widehat{F}_N^{x}(y)|\\ &=: E_1+E_2+E_3+E_4+E_5. \end{aligned}\tag{9}$$

- Concerning E_1 and E_5, by following the same lines as for studying the terms P_1 and P_5, we obtain:

$$E_1=O_{\text{a.co.}}\left(\sqrt{\frac{\log N}{N\phi(h_K)}}\right)\quad\text{and}\quad E_5=\left(\sqrt{\frac{\log N}{N\phi(h_K)}}\right).$$

- Concerning the term E_2, by using the Lipschitz's condition on the kernel $H(.)$, we can write

$$|\widehat{F}_N^{x_{k(x)}}(y)-\widehat{F}_N^{x_{k(x)}}(y_t)|\leq Ch_H^{-1}\underbrace{|y-y_t|}_{l_n}\Big(\underbrace{\frac{1}{N\mathbb{E}\left[K_1\right]}\sum_{i\in V}K_i}_{\widehat{F}_D^x}+\underbrace{\sum_{j\in\bar{V}}\frac{K_j}{\mathbb{E}\left[K_1\right]}}_{\widehat{F}_D^x}\Big),$$

under $(A6),(A4),(A5)$ and from the almost comply consistency of $\widehat{F}_D$ (Lemma 4.4), and take $l_n=N^{-\beta}$, we get

$$E_2=O_{\text{a.co.}}\left(\sqrt{\frac{\log N}{N\phi(h_K)}}\right)\quad\text{and}\quad E_4=O\left(\sqrt{\frac{\log N}{N\phi(h_K)}}\right).\tag{10}$$

- For E_3, we have

$$\begin{aligned}E_3&=\sup_{x\in S_{\mathcal{F}}}\sup_{y\in S_{\mathbb{R}}}|\widehat{F}_N^x(z_y)-\mathbb{E}[\widehat{F}_N^x(z_y)]|\\ &\leq\sup_{x\in S_{\mathcal{F}}}\sup_{y\in S_{\mathbb{R}}}\Big|\frac{1}{N}\left(\sum_{i\in V}\frac{H_i(y_t)K_i(x_{k(x)})}{\mathbb{E}\left[K_1\right]}-\mathbb{E}\left(\sum_{i\in V}\frac{H_i(y_t)K_i(x_{k(x)})}{\mathbb{E}\left[K_1\right]}\right)\right)\Big|\\ &\quad+\sup_{x\in S_{\mathcal{F}}}\sup_{y\in S_{\mathbb{R}}}\Big|\sum_{j\in\bar{V}}\frac{K_j(x_{k(x)})}{\mathbb{E}\left[K_1\right]}-\mathbb{E}\left(\sum_{j\in\bar{V}}\frac{K_j(x_{k(x)})}{\mathbb{E}\left[K_1\right]}\right)\Big|\\ &=: E_{2,1}+E_{2,2},\end{aligned}\tag{11}$$

then, for $E_{2,1}$:

$$\mathbb{P}\left(E_{2,1}>\kappa\sqrt{\frac{\log N}{N\phi(h_K)}}\right)\leq q_nd_n\max_{x\in S_{\mathcal{F}}}\max_{y_t\in S_{\mathbb{R}}}\mathbb{P}\left(\Big|\frac{1}{N}\sum_{i\in V}\left(\Lambda_i\right)\Big|>\kappa\sqrt{\frac{\log N}{N\phi(h_K)}}\right),$$

with

$$\Lambda_i=\frac{H_i(y_t)K_i(x_k)}{\mathbb{E}\left[K_1\right]}-\mathbb{E}\left(\frac{H_i(y_t)K_i(x_{k(x)})}{\mathbb{E}\left[K_1\right]}\right).$$

So, by the Bernstein's exponential inequality for Λ_i, where, $|\Lambda_i| \leq C/\phi(h_K)$ and $\mathbb{E}|\Lambda_i|^2 \leq C'/\phi(h_K)$, as usually, we take $q_n = O(l_n^{-1})$, $C\kappa^2 = 2\beta + 1$, such that

$$q_n \max_{y_t \in S_{\mathbb{R}}} \mathbb{P}\left(\left|\frac{1}{N}\sum_{i\in V}\Lambda_i\right| > \kappa\sqrt{\frac{\log N}{N\phi(h_K)}}\right) \leq q_n 2\exp\{-C\kappa^2 \log N\}$$
$$\leq CN^{\beta}N^{-2\beta-1},$$

so,

$$\mathbb{P}\left(E_{2,1} > \kappa\sqrt{\frac{\log N}{N\phi(h_K)}}\right) \leq CN^{-\beta-1},$$

now, by take $H(y_t) = 1$ for $E_{2,1}$, we obtain $E_{2,2}$ in very easy manner.
So,

$$E_3 = O_{\text{a.co.}}\left(\sqrt{\frac{\log N}{N\phi(h_K)}}\right). \tag{12}$$

Finally, the Lemma 4.3 is achieved. ■

Lemma 4.4.

Under the assumptions $(A1)$ and $(A3) - (A6)$, we obtain

$$\sup_{x\in S_{\mathcal{F}}}\left|\widehat{F}_D^x - 1\right| = O_{\text{a.co.}}\left(\sqrt{\frac{\log N}{N\phi(h_K)}}\right),$$

and

$$\sum_{i\in\mathbb{N}}\mathbb{P}(\widehat{F}_D^x < 1/2) < \infty.$$

Proof:

We have

$$\widehat{F}_D^x - 1 = \frac{1}{N}\sum_{i=1}^{N}\frac{K_i}{\mathbb{E}K_1} - \frac{1}{N}\mathbb{E}\left(\sum_{i=1}^{N}\frac{K_i}{\mathbb{E}K_1}\right)$$
$$= \frac{1}{N}\sum_{i=1}^{N}\frac{K_i}{\mathbb{E}K_1} - \frac{\mathbb{E}K_i}{\mathbb{E}K_1}$$
$$= \frac{1}{N}\sum_{i=1}^{N}\Delta_i,$$

where $\Delta_i = \dfrac{K_i}{\mathbb{E}K_1} - \mathbb{E}\dfrac{K_i}{\mathbb{E}K_1}$. Under (A6), for $m = 1, 2$, we have

$$0 < \frac{C'}{\phi(h_K)} < \mathbb{E}(K_1^m) < \frac{C}{\phi(h_K)},$$

then

$$|\Delta_i| < \frac{C}{\phi(h_K)} = \theta_1,$$

and

$$\mathbb{E}\Delta_i^2 < \frac{C'}{\phi(h_K)} = \theta_2.$$

We apply the Bernstein-type exponential inequality, for all $\varepsilon \in]0, \frac{\theta_1}{\theta_2}[$, we get

$$\begin{aligned}\mathbb{P}\left(\left|\widehat{F}_D^x - 1\right| > \varepsilon\sqrt{\frac{\log N}{N\phi(h_K)}}\right) &\leq 2\exp\left(\frac{-\varepsilon^2 \log N}{4\phi(h_K)\theta_2}\right) \\ &= 2N^{-\varepsilon^2/4\phi(h_K)\theta_2} \\ &= 2N^{-C\varepsilon^2}.\end{aligned} \tag{13}$$

It follows that for ε^2 large enough

$$\sum_{i=1}^{\infty} \mathbb{P}\left(\left|\widehat{F}_D^x - 1\right| > \varepsilon\sqrt{\frac{\log N}{N\phi(h_K)}}\right) < +\infty.$$

For the second part, we have

$$\begin{aligned}\mathbb{P}\{|\widehat{F}_D^x| \leq 1/2\} &\leq \mathbb{P}\{|\widehat{F}_D^x - 1| > 1/2\} \\ &\leq \mathbb{P}\{|\widehat{F}_D^x - \mathbb{E}\widehat{F}_D^x| > 1/2\}.\end{aligned}$$

We deduce that

$$\sum_{i\in\mathbb{N}} \mathbb{P}\left(\widehat{F}_D^x < 1/2\right) < \infty.$$

■

4.2. The consistency of the conditional quantile estimator

In this section we study the asymptotic behavior of the conditional quantile. Obviously, we will estimate it by mean of the conditional distribution estimator. We introduce $\hat{q}_\gamma$, the estimator of q_γ defined as

$$\hat{F}^x\left(\hat{q}_\gamma\right) = \gamma,$$

where $\gamma \in]0, 1[$.

To achieve our result, we need the following hypotheses.

(A7) $H(.)$ is strictly increasing *cond-cdf*

(A8) The distribution $F^x(.)$ is strictly increasing, continuous and differentiable in neighborhood of q_γ.

Note that (A8) control the flatness of the conditional c.d.f. around the quantile to be estimated.

Corollary 4.1.

Under assumptions of the Theorem 4.1 and $(A8)$, we obtain

$$|\widehat{q}_\gamma - q_\gamma| = O(h_K^{A_1} + h_H^{A_2} + a_n^{A_1}) + O_{\text{a.co.}}\left(\left(\frac{\log N}{N\phi(h_K)}\right)^{1/2}\right) + O_{\text{a.co.}}\left(\left(\frac{\log d_n}{n\phi(a_n)}\right)^{1/2}\right).$$

Proof:

We present briefly the proof, where Taylor expansion of $F^x(.)$ drive to the existence of some q^* between $\hat{q}_\gamma$ and q_γ and under the condition $(A8)$ we get:

$$\hat{F}^x\left(\hat{q}_\gamma\right) - \hat{F}^x\left(q_\gamma\right) = \left(\hat{q}_\gamma - q_\gamma\right)\hat{F}^{x^{(1)}}\left(q_\gamma^*\right),$$

$$|\hat{q}_\gamma - q_\gamma| = \frac{1}{\hat{F}^{x^{(1)}}\left(q_\gamma^*\right)}\left[\left|\hat{F}^x\left(\hat{q}_\gamma\right) - \hat{F}^x\left(q_\gamma\right)\right|\right].$$

If we could confirm that

$$\exists \delta > 0, \sum_{n=1}^{\infty} \mathbb{P}\left(\hat{F}^{x^{(1)}}\left(q^*\right) < \delta\right) < \infty,$$

we obtain

$$\begin{aligned}
\mathbb{P}\left(|\hat{q}_\gamma - q_\gamma| > \epsilon\right) &\leq \mathbb{P}\left(\left|\hat{F}^x\left(\hat{q}_\gamma\right) - \hat{F}^x\left(q_\gamma\right)\right| > \delta\left(\epsilon\right)\right) \\
&= \mathbb{P}\left(\left|F^x\left(q_\gamma\right) - \hat{F}^x\left(q_\gamma\right)\right| > \delta\left(\epsilon\right)\right) \\
&\leq \mathbb{P}\left(\sup_{x\in S_{\mathcal{F}}}\sup_{y\in S_{\mathbb{R}}}\left|\hat{F}^x\left(y\right) - F^x\left(y\right)\right| > \delta\left(\epsilon\right)\right).
\end{aligned}$$

Under assumption (A8), and by comparing the rates of convergence given in Theorem 4.1, we have

$$\sum_n \mathbb{P}\left(|\hat{q}_\gamma - q_\gamma| > \epsilon\right) \leq \sum_n \mathbb{P}\left(\sup_{x\in S_{\mathcal{F}}}\sup_{y\in S_{\mathbb{R}}}\left|\hat{F}^x\left(y\right) - F^x\left(y\right)\right| > \delta\left(\epsilon\right)\right) < \infty. \qquad \blacksquare$$

5. Simulation

In this section, we evaluate the behavior of the proposed estimator by conducting a number of simulation studies. Let $\widehat{F}_V^x(y)$ be the standard Nadaraya-Watson estimator with the true observations in the validation data set. That is,

$$\widehat{F}_V^x(y) = \frac{\sum_{i\in V} H\left(\frac{y-Y_i}{h_H}\right) K\left(\frac{d(x,X_i)}{h_K}\right)}{\sum_{i\in V} K\left(\frac{d(x,X_i)}{h_K}\right)}.$$

A simulation was conducted to compare the proposed estimators $\widehat{F}_R^x(y)$ with $\widehat{F}_V^x(y)$ and $\widehat{F}_C^x(y)$, where $\widehat{F}_C^x(y)$ is defined above in Equation (1). It should be pointed out that $\widehat{F}_C^x(y)$ can serve as

a gold standard in the simulation study, even though it is practically unachievable because of the measurement errors.

We generated the response variables Y such that

$$Y_i = m(X_i) + \varepsilon_i \quad \text{for } i = 1, ..., 250,$$

where the functional regressors X_i are defined (see Figure 1), for any $t \in [0, \frac{\pi}{2}]$, by:

$$X_i(t) = 3W_i \sin(2\pi t) + A_i t \quad \text{with } W_i \sim \mathcal{N}(1, 0.5) \text{ and } A_i \sim \mathcal{N}(0, 1),$$

the error ε has the standard normal distribution and it is independent of X, and $m(X_i)$ is given by

$$m(X_i) = \frac{5}{1 + \int_0^{\frac{\pi}{2}} X_i(t)\mathrm{d}t}.$$

A sample of smooth curves $X_i(t)$ are plotted in Figure 1.

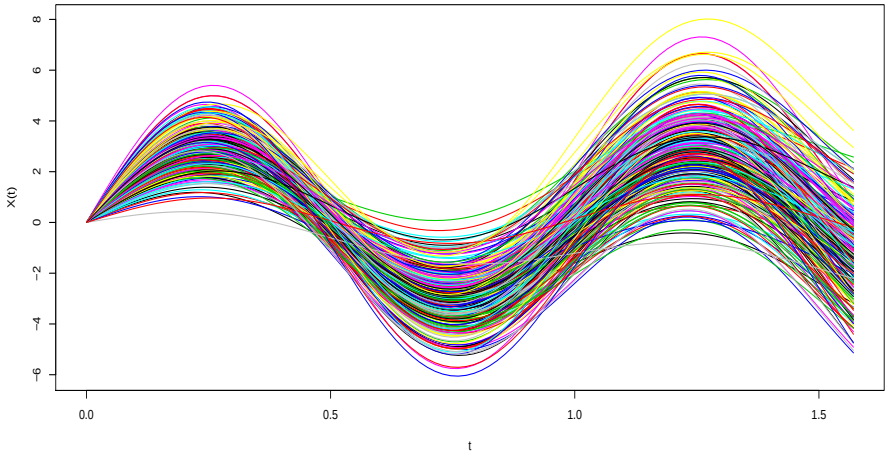

Figure 1. Curves (N=250)

Now, let $S_0 = \{1, \ldots, 200\}$ and $S_1 = \{201, \ldots, 250\}$ be two subsets of indices. Then, we choose $\mathcal{L} = \{(X_i, Y_i)\}_{i \in S_0}$ as the learning sample and $\mathcal{T} = \{(X_i, Y_i)\}_{i \in S_1}$ as the test sample. We have from Ibrahim et al. (2020) that the surrogate variable $\widetilde{Y}_i$ of Y_i, for all $i \in S_0$, was generated from

$$\widetilde{Y}_i = \rho Z_i + e_i,$$

where Z_i is the standard score of Y_i and $e_i \sim \mathcal{N}(0, \sqrt{1 - \rho^2})$. In such a way that the correlation coefficient between Y_i and $\widetilde{Y}_i$ is approximately equal to ρ which would not be controllable in practice. In the sequel of this simulation study, we take $\rho = 0.35$ or $\rho = 0.75$.

From the learning sample containing $N = 200$ functional data, we randomly choose a set V of n validation data $\{(X_i, Y_i)\}_{i \in V}$ which allows to build the functional kernel estimator $\hat{F}_V^x(y)$ of $m(x)$. The estimator $\hat{F}_R^x(y)$ is then constructed by using the surrogate data $\left\{\left(X_i, \widetilde{Y}_i\right)\right\}_{i \in \bar{V}}$ with

the help of the validation data, where $\bar{V} = \{1, \ldots, N\} \backslash V$. It should be pointed out that for $N = n$ (complete observations), we have

$$\hat{F}_V^x(y) = \hat{F}_R^x(y) = \hat{F}_C^x(y).$$

The bandwidths h_H and h_K are selected by a cross-validation method. Because of the smoothness of the curves, we have built the predictors through the semi-metric based on the first derivatives (see Benhenni et al. (2007)). For the bandwidths a_n, we used the same principal steps in Ibrahim et al. (2020), the kernels $K(.)$ and $W(.,.)$ are chosen to be the quadratic and the integrate quadratic kernels, these latter are Epanechnikov kernels.

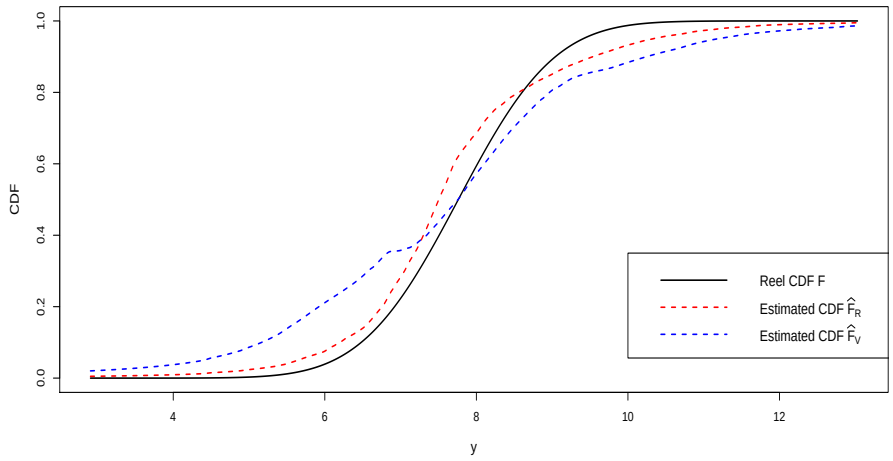

Figure 2. CDF comparison

Figure 2 represents the curves of the CDF with $F^x(y) = \int_0^y \frac{1}{2\pi} \exp \frac{-(z-m(x))^2}{2} \mathrm{d}z$, where, it is clear that our $\hat{F}_R^x(y)$ is closer to the real curve which represents the complete sample and consequently, $\hat{F}_R(y)$ performs better than $\hat{F}_V^x(y)$.

Hereafter, we will apply our result on the median and obtained results are given in Figure 3.

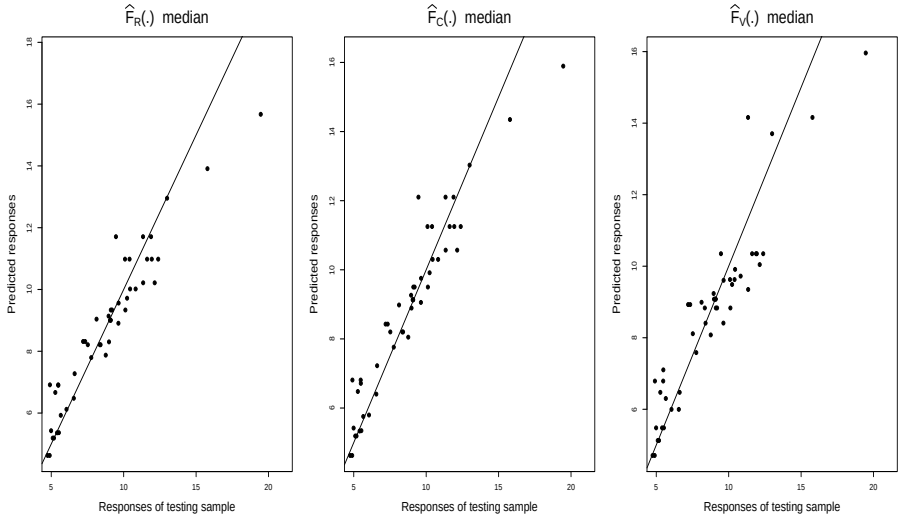

Figure 3. Comparative prediction between the median for each: $\hat{F}_R^x(y)$, $\hat{F}_C^x(y)$ and $\hat{F}_V^x(y)$

Table 1. MSE result

	$n/N \rightarrow$ ρ $\downarrow$	0.125	0.25
$\hat{F}_V^x(y)$	0.35	0.6543	0.7127
	0.75	0.6729	0.7149
$\hat{F}_R^x(y)$	0.35	0.5503	0.5922
	0.75	0.5692	0.6018
$\hat{F}_C^x(y)$	–	0.5248	0.5248

It can be noticed from Figure 3 that the estimator $\hat{F}_R^x(y)$ is better than the estimator $\hat{F}_V^x(y)$. Also, it appears clearly that in this case the performance of both estimates is closely linked to the correlation coefficient and the ration n/N since the values of MSE-error increase substantially with respect to those parameters (see Table 1). In this table, we summarize the MSE-error for two values of n/N and ρ, this error increases with respect to those parameters. It is noted that the results are sufficiently good for all sample size, and further results are given for large sample sizes in Figure 4.

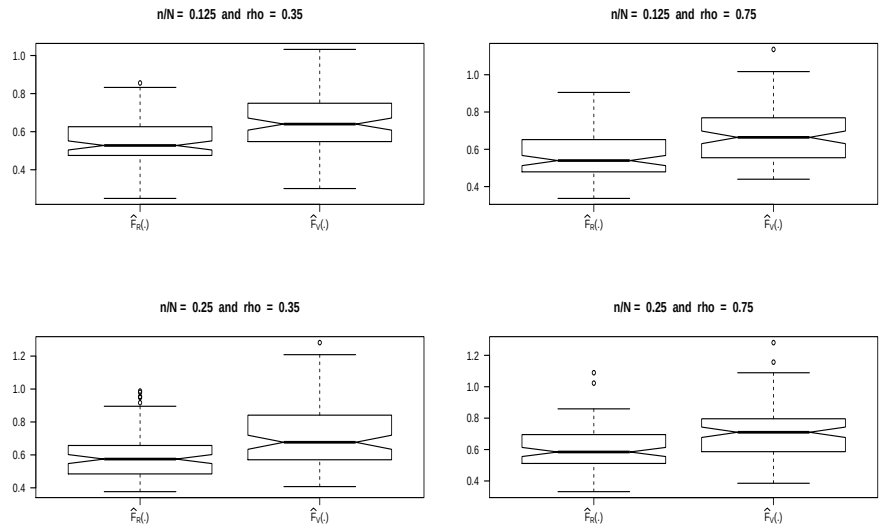

Figure 4. A boxplots of the MSE of $\hat{F}_R^x(y)$ and $\hat{F}_V^x(y)$

Figure 4 displays the boxplot of MSE. It can be seen from this figure that our estimator $\hat{F}_R^x(y)$ remains more stable than $\hat{F}_V^x(y)$, and we can conclude to good asymptotic performance of $\hat{F}_R^x(y)$.

6. Conclusion

This paper presents the conditional distribution function's estimator using the kernel method for a surrogated scalar response variable given a functional random one. This estimator is built from the validation data. We obtained the uniform, almost complete convergence of this model using kernel estimate and the conditional quantile estimator under some classical assumptions. To improve the performance of our proposed estimator and the theoretical results, we realized a simulation study. Other research issues are possible, such as extensions to local linear method estimation and the semiparametric linear regression model which can also be studied using this kind of data. Finally, the k nearest neighbor method can be adapted to treat the outliers in the data set as proposed in the literature by Attouch et al.

Acknowledgment:

The authors greatly thank the Editor in chief and the reviewers for the careful reading, constructive comments and relevant remarks which permit us to improve the paper.

REFERENCES

Attouch, M., Laksaci, A. and Rafaa, F. (2017). Local linear estimate of the regression operator by the kNN method, Comptes Rendus Mathematique, Vol. 355, No. 7, pp. 824–829.

Barrientos-Marin, J., Ferraty, F. and Vieu, P. (2010). Locally modelled regression and functional data, Journal of Nonparametric Statistics, Vol. 22, No. 5, pp. 617–632.

Benhenni, K., Ferraty, F., Rachdi, M. and Vieu, P. (2007). Local smoothing regression with functional data, Computational Statistics, Vol. 22, No. 3, pp. 353–369.

Bosq, D. (2000). *Linear Processes in Function Spaces*, Volume 149 of Lecture Notes in Statistics.

Carroll, R.J. and Wand, M.P. (1991). Semiparametric estimation in logistic measurement error modelss, Journal of the Royal Statistical Society: Series B (Methodological), Vol. 53, No. 3, pp. 573–585.

Chen, J.H. and Shao, J. (2000). Nearest neighbor imputation for survey data, Journal of Official Statistics, Vol. 16, No. 2, pp. 113.

Cheng, P.E. (1994). Nonparametric estimation of mean functionals with data missing at random, Journal of the American Statistical Association, Vol. 89, No. 425, pp. 81–87.

Dabo-Niang, S. (2002). Estimation de la densité dans un espace de dimension infinie: Application aux diffusions, Comptes Rendus Mathematique, Vol. 334, No. 3, pp. 213–216.

Duncan, G.J. and Hill, D. H. (1985). An investigation of the extent and consequences of measurement error in labor-economic survey data, Journal of Labor Economics, Vol. 3, No. 4, pp. 508–532.

Ferraty, F., Goia, A. and Vieu, P. (2002). Functional nonparametric model for time series: A fractal approach for dimension reduction, Test, Vol. 11, No. 2, pp. 317–344.

Ferraty, F., Laksaci, A. and Vieu, P. (2006). Estimating some characteristics of the conditional distribution in nonparametric functional models, Statistical Inference for Stochastic Processes, Vol. 9, No. 1, pp. 47–76.

Ferraty, F., Rabhi, A. and Vieu, P. (2005). Conditional quantiles for dependent functional data with application to the climatic El Niño Phenomenon, Sankhyā, Vol.67, No. 2, pp. 378–398.

Ferraty, F., Sued, M. and Vieu, P. (2013). Mean estimation with data missing at random for functional covariables, Statistics, Vol. 47, No. 4, pp. 688–706.

Ferraty, F. and Vieu, P. (2002). The functional nonparametric model and application to spectrometric data, Computational Statistics, Vol. 17, No. 4, pp. 545–564.

Ferraty, F. and Vieu, P. (2003). Curves discrimination: a nonparametric functional approach, Computational Statistics & Data Analysis, Vol. 44, No. (1-2), pp. 161–173.

Ferraty, F., and Vieu, P. (2006). *Nonparametric Functional Data Analysis: Theory and Practice*. Springer Science & Business Media.

Ibrahim, F., Hajj Hassan, A., Demongeot, J. and Rachdi, M. (2020). Regression model for surrogate data in high dimensional statistics, Communications in Statistics-Theory and Methods, Vol. 49, No. 13, pp. 3206–3227.

Ling, N., Liang, L. and Vieu, P. (2015). Nonparametric regression estimation for functional stationary ergodic data with missing at random, Journal of Statistical Planning and Inference,

Vol. 162, pp. 75–87.

Pepe, M. S. (1992). Inference using surrogate outcome data and validation sample, Biometrika, Vol. 79, No. 2, pp. 355–65.

Pérez-González, A., Vilar-Fernández, J. M. and González-Manteiga, W. (2009). Asymptotic properties of local polynomial regression with missing data and correlated errors, Annals of the Institute of Statistical Mathematics, Vol. 61, No. 1, pp. 85–109.

Ramsay, J. and Silverman, B. (1997). *Functional Data Analysis*, Springer-Verlag.

Ramsay, J. and Silverman, B. (2002). *Appied Functional Data Analysis*, Springer-Verlag.

Roussas, G. G. (1968). On some properties of nonparametric estimates of probability density functions, Bull. Soc. Math. Greece (N.S.), Vol. 9, No. 9, pp. 29–43.

Wang, Q. (2006). Nonparametric regression function estimation with surrogate data and validation sampling, Journal of Multivariate Analysis, Vol. 97, No. 5, pp. 1142–1161.

Wittes, J., Lakatos, E. and Probstfield, J. (1989). Surrogate endpoints in clinical trails: Cardiovascular diseases, Statistics in Medicine, Vol. 8, No. 4, pp. 415-25.

Yates, F. (1933). The analysis of replicated experiments when the field results are incomplete, Empire Journal of Experimental Agriculture, Vol. 1, No. 2, pp. 129–142.