STATISTICS, OPTIMIZATION AND INFORMATION COMPUTING
Stat., Optim. Inf. Comput., Vol. **10**, June 2022, pp 789–814.
Published online in International Academic Press (www.IAPress.org)

Statistical Inference for Multivariate Conditional Cumulative Distribution Function Estimation By Stochastic Approximation Method

Sahar Slama [1,*], Yousri Slaoui [2], Hamdi Fathallah [1]

[1] *LAMMDA: Laboratoire de Mathématiques Modélisation Déterministe et Aléatoire, Sousse University and ESST Hammam Sousse, Tunisia*
[2] *Laboratoire de Mathématiques et Applications, Poitiers University and CNRS, France*

Abstract This paper handles non-parametric estimation of a conditional cumulative distribution function (CCDF). Using a recursive approach, we set forward a multivariate recursive estimator defined by stochastic approximation algorithm. Our basic objective is to investigate the statistical inference of our estimator and compare it with that of non-recursive Nadaraya-Watson's estimator. From this perspective, we first derive the asymptotic properties of the proposed estimator which highly depend on the choice of two parameters, the stepsize (γ_n) as well as the bandwidth (h_n). The second generation plug-in method, a method of bandwidth selection minimizing the Mean Weighted Integrated Squared Error $(MWISE)$ of the estimator in reference, entails the optimal choice of the bandwidth and therefore maintains an appropriate choice of the stepsize parameter. Basically, we demonstrate that, under some conditions, the Mean Squared Error (MSE) of the proposed estimator can be smaller than the one of Nadaraya Watson's estimator. We corroborate our theoretical results through simulation studies and two real dataset applications, namely the Insurance Company Benchmark (COIL 2000) dataset as well as the French Hospital Data of COVID-19 epidemic.

Keywords conditional cumulative distribution function (CCDF), kernel estimation method, stochastic approximation algorithm, plug-in principle

AMS 2010 subject classifications 62G20, 62H12, 62L20, 65D10.

DOI: 10.19139/soic-2310-5070-1416

1. Introduction

Assume that we observe independent identically distributed vectors $(X_1, Y_1), ..., (X_n, Y_n)$ of a bivariate random variable (X, Y) with common cumulative distribution function $\pi(x, y)$ where one is interested in modeling the functional dependence of the observation Y on the covariable X by the conditional cumulative distribution function (CCDF) of Y given $X = x$, denoted by for all real y and x

$$\pi(y|x) := \mathbb{P}\left[Y \leqslant y | X = x\right].$$

We shall also assume that the bivariate random variable (X, Y) (resp. the random variable X) has a density function $f_{(X,Y)}$ (resp. f_X) with respect to the Lebesgue measure. Recall that for all real y and x such that $f_X(x) \neq 0$, the CCDF of Y given $X = x$ is expressed by

$$\pi(y|x) = \frac{\int_{\mathbb{R}} \mathbb{1}_{\{u \leqslant y\}} f_{(X,Y)}(x, u) du}{f_X(x)}.$$

*Correspondence to: Sahar Slama (Email: sahar.slama@essths.u-sousse.tn). LAMMDA: Laboratoire de Mathématiques Modélisation Déterministe et Aléatoire, Sousse University and ESSTHS Rue Lamine El Abbessi, Hammam Sousse, 4011, Tunisia.

ISSN 2310-5070 (online) ISSN 2311-004X (print)

In a variety of non-parametric statistical problems, the estimation of a CCDF is a key aspect of inference. Remember that the CCDF has the merit of characterizing the conditional law of the considered bivariate random variables. Notably, the CCDF is often useful in reliability or survival analysis.
More specifically, the conditional survival function $S(y|x)$ defined by, for all real y and x, $S(y|x) := 1 - \pi(y|x)$ is of extreme interest, either by itself, or by its independence with the conditional hazard function $h(y|x)$ indicated by, for all real y and x, $h(y|x) := \dfrac{f(y|x)}{S(y|x)}$ where $f(y|x)$ denotes the conditional density of Y given $X = x$. Furthermore, conditional quantiles can also be deduced from the CCDF π by (pseudo)-inversion given x of the function $y \rightarrow \pi(y|x)$ and the same procedure may be applied to the estimator of CCDF to find conditional quantile estimators.
Several non-parametric estimators have been elaborated to estimate the CCDF. Many of them rely initially on estimating the $\int_{\mathbb{R}} \mathbb{1}_{\{u \leqslant y\}} f_{(X,Y)(x,u)} du$. The conditional cumulative distribution function was first extensively explored by [41] using a nearest-neighbor-type conditional empirical process. Subsequently, [17] motivated by the problem of setting prediction intervals in time series analysis, developed a new non-parametric method for CCDF estimation resting on an adjusted form of NadarayaWatson estimator. Afterwards, [28] established uniform asymptotic certainty bands for the CCDF using the same strategy.
For a general non-parametric regression model, [22] set up two estimators using a kernel approch, where the distribution of the error given the covariate is modeled by a CCDF provided by $\mathbb{P}(\epsilon \leq y|X = x)$.
On a given compact set, [9] constructed a minimax estimator of the CCDF. Thereafter, [43] built up a new estimator of CCDF investing a method of pre-adjusting the original observations non-parametrically. Recently, [8] introduced a new method to settle CCDF estimation problem based on local polynomial technique.
Many functional estimations are grounded on estimating the CCDF see e.g [23], [24] and [2]. The CCDF is involved in a wide range of applications, for instance, in medicine see [16], econometrics see [25] or machine learning domain see the recent work of [11].
In a broader context, extensive state of art works including various non-parametric approaches tackled the conditional estimation. We can state for example [14], [45], [5, 6], [18, 19] and [31]. For recent references see [4], [12], [1], [10] and [40].

Over the past decade, data streams have become an increasingly important area of research. Some of the most common data streams include Internet packet data, Twitter activity, Facebook newsfeed, credit card transactions and more recently COVID-19 epidemic data. In these situations, the data arrives regularly so that it is impossible to store it in a traditional database. In such a context, it is very interesting to build a recursive multivariate conditional cumulative distribution estimator that does not need to store all the data in memory and that can be easily updated to handle the online data. The basic target of this paper is to provide a non-parametric strategy to recursively estimate the CCDF.

The paper is organized as follows. In section 2 we introduce our method for the estimation procedure of CCDF. The main results of our recursive estimator are displayed in section 3. Section 4 exhibits the asymptotic properties of non-recursive Nadaraya-Watson's estimator. A second generation plug-in scheme for constructing data-driven bandwidth selection procedures is elaborated in section 5. Section 6 highlights the performances of our estimator on simulated data as well as a real dataset. Finally, the conclusion is drawn in section 7.

2. Preliminaries

Let $(X, Y) \in \mathbb{R}^d \times \mathbb{R}^q$ and $(X_1, Y_1), ..., (X_n, Y_n)$ be independent random vectors identically distributed as (X, Y) with joint density function $f_{(X,Y)}$ and let f_X denote the probability density of X.
In this paper, our central focus is upon the problem of estimating the CCDF of Y given $X = x$ provided by

$$\begin{aligned} \pi: \quad & \mathbb{R}^q \times \mathbb{R}^d \longrightarrow \mathbb{R} \\ & (y|x) \quad \longmapsto \mathbb{P}[Y \leqslant y|X = x] = \frac{\int_{\mathbb{R}^q} \mathbb{1}_{\{u \leqslant y\}} f_{(X,Y)}(x,u) du}{f_X(x)}. \end{aligned}$$

Here, we introduce our recursive estimator π_n specified by

$$\pi_n(y|x) = \begin{cases} \dfrac{a_n(x,y)}{f_n(x)} & \text{if} \quad f_n(x) \neq 0 \\ 0 & \text{otherwise} \end{cases}, \tag{1}$$

with
$a_n(x,y) = \Pi_n \sum\limits_{k=1}^{n} \Pi_k^{-1} \gamma_k \, \chi_k(y) h_k^{-d} \, \mathbf{K}\left(\dfrac{x - X_k}{h_k}\right)$ and $f_n(x) = \Pi_n \sum\limits_{k=1}^{n} \Pi_k^{-1} \gamma_k \, h_k^{-d} \, \mathbf{K}\left(\dfrac{x - X_k}{h_k}\right)$, where

- $\mathbf{K}$ is a multivariate kernel satisfying $\displaystyle\int_{\mathbb{R}^d} \mathbf{K}(t)dt = 1$.
- χ is a multivariate indicator function identified by $\chi_k : \mathbb{R}^q \longrightarrow \mathbb{R}, \; y \longmapsto \mathbb{1}_{\{Y_k \leqslant y\}}$.
- (h_n) is the bandwidth: a sequence of positive real numbers that tends to zero.
- (γ_n) is the stepsize: a positive sequence of real numbers decreasing towards zero and $\Pi_n = \prod\limits_{j=1}^{n} (1 - \gamma_j)$.

Our main purpose is to examine the asymptotic properties of the proposed multivariate estimator of the CCDF and to corroborate its performances. We shall compare our estimator to the generalized kernel CCDF estimator of Nadaraya-Watson [29,44] $\widetilde{\pi}_n$ expressed by

$$\widetilde{\pi}_n(y|x) = \begin{cases} \dfrac{\widetilde{a}_n(x,y)}{\widetilde{f}_n(x)} & \text{if} \quad \widetilde{f}_n(x) \neq 0 \\ 0 & \text{otherwise} \end{cases}, \tag{2}$$

with

$$\widetilde{a}_n(x,y) = \frac{1}{nh_n^d} \sum_{k=1}^{n} \chi_k(y) \, \mathbf{K}\left(\frac{x - X_k}{h_n}\right) \text{ and } \widetilde{f}_n(x) = \frac{1}{nh_n^d} \sum_{k=1}^{n} \mathbf{K}\left(\frac{x - X_k}{h_n}\right).$$

The recursive estimator was constructed based on dint of stochastic approximation method. Indeed, incorporating stochastic approximation algorithms in the context of non-parametric statistics dates back to the papers of [33] and [21]. Their research works have been extended in several directions. We refer the reader to [7], [29], [35] and [13].

Subsequently, in the paper [26], the multidimensional case was investigated in order to estimate a multivariate probability density using the estimation by confidence intervals. More recently, [39] developed a new recursive kernel estimator for regression function estimation. Additionally, [38] reused stochastic approximation methods to enhance the qualities of the univariate distribution function estimator and lately [36] elaborated the multivariate one. Following the same recursive approach, we intend to establish a multivariate conditional cumulative distribution function estimator.

To build up a stochastic algorithm, which approaches the function

$$a : (x,y) \longmapsto \int_{\mathbb{R}^q} \mathbb{1}_{\{u \leqslant y\}} f_{(X,Y)}(x,u) du$$

at a given couple of vectors (x,y), we define an algorithm of search of the zero function $\phi : z \longmapsto a(x,y) - z$ and we set:

$$(i) \; a_0(x,y) \in \mathbb{R} \qquad\qquad (ii) \text{ for all } n \geq 1, a_n(x,y) = a_{n-1}(x,y) + \gamma_n U_n(x,y),$$

where $U_n(x,y)$ corresponds to an observation of the function ϕ at the point $a_{n-1}(x,y)$.

Note that to define $U_n(x,y)$, we adopt the approach of [32] and [42].

By considering $U_n(x,y) = \chi_n(y)\, h_n^{-d}\, \mathbf{K}\left(\frac{x - X_n}{h_n}\right) - a_{n-1}(x,y)$, the stochastic approximation algorithm that is devoted to estimate recursively the function a at a couple of vectors (x,y) can be stated as follows :

$$a_n(x,y) = (1-\gamma_n)a_{n-1}(x,y) + \gamma_n\, \chi_n(y)\, h_n^{-d}\, \mathbf{K}\left(\frac{x - X_n}{h_n}\right). \tag{3}$$

Throughout this paper, we consider that $a_0(x,y) = 0$. Therefore, by recurrence, we get

$$a_n(x,y) = \Pi_n \sum_{k=1}^{n} \Pi_k^{-1}\gamma_k\, \chi_k(y) h_k^{-d}\, \mathbf{K}\left(\frac{x - X_k}{h_k}\right). \tag{4}$$

Within this framework, we use the recursive multivariate probability density estimator of the density function f_X noted f_n and defined in [26]. It was constructed with the same tools of stochastic approximation algorithm and under the condition that $f_0(x) = 0$, we have:

$$f_n(x) = \Pi_n \sum_{k=1}^{n} \Pi_k^{-1}\gamma_k\, h_k^{-d}\, \mathbf{K}\left(\frac{x - X_k}{h_k}\right). \tag{5}$$

The following notations are highly useful as they are invested throughout the whole paper.

* $\forall\, i,j \in \{1,\ldots,d\}$,

$$a_i^{(1)}(.,y) := \frac{\partial a}{\partial x_i}(.,y), \qquad \pi_i^{(1)}(y|.) := \frac{\partial \pi}{\partial x_i}(y|.), \qquad f_{X_i}^{(1)}(.) := \frac{\partial f_X}{\partial x_i}(.),$$
$$a_{ij}^{(2)}(.,y) := \frac{\partial^2 a}{\partial x_i \partial x_j}(.,y), \qquad \pi_{ij}^{(2)}(y|.) := \frac{\partial^2 \pi}{\partial x_i \partial x_j}(y|.), \qquad f_{X_{ij}}^{(2)}(.) := \frac{\partial^2 f_X}{\partial x_i \partial x_j}(.),$$
$$\mu_i(\mathbf{K}) := \int_{\mathbb{R}^d} z_i^2\, \mathbf{K}(z) dz.$$

* $\xi := \lim_{n\to+\infty} (n\gamma_n)^{-1}$.

* $R(\mathbf{K}) := \int_{\mathbb{R}^d} \mathbf{K}^2\,(z)\, dz.$

* $I_1 := \int_{\mathbb{R}^{d+q}} \left(\sum_{j=1}^{d} \mu_j(\mathbf{K}) \left[\pi_{jj}^{(2)}(y|x) f_X(x) + 2\pi_j^{(1)}(y|x) f_{X_j}^{(1)}(x)\right]\right)^2 f_{X,Y}(x,y) dx dy.$

* $I_2 := \int_{\mathbb{R}^{d+q}} \pi(y|x)\,(1-\pi(y|x))\, f_X(x) f_{X,Y}(x,y) dx dy.$

* $\forall\, x \in \mathbb{R}^d, y \in \mathbb{R}^q$,

$$\mathcal{Z}_n(x,y) := h_n^{-d}\chi_n(y)\, \mathbf{K}\left(\frac{x - X_n}{h_n}\right) \qquad \text{and} \qquad \mathcal{W}_n(x) := h_n^{-d}\, \mathbf{K}\left(\frac{x - X_n}{h_n}\right).$$

In the sequel, let us present the following definition of class of regularly varying sequences introduced by Galambos and Seneta in [15].

Definition 1
Let (v_n) be a nonrandom positive sequence and $\gamma \in \mathbb{R}$. We say that

$$(v_n) \in \mathcal{GS}(\gamma) \text{ if } \lim_{n\to+\infty} n\left[1 - \frac{v_{n-1}}{v_n}\right] = \gamma.$$

In what follows, we introduce a lemma that will be widely invested throughout the study of our estimator π_n. It is worth noting that the proof of this lemma was recorded in [26].

Lemma 1
Let $(v_n) \in \mathcal{GS}(v^*), (\gamma_n) \in \mathcal{GS}(-\alpha)$ and let $m > 0$ such that $m - v^*\xi > 0$. Then,

$$\lim_{n\to+\infty} v_n \Pi_n^m \sum_{k=1}^{n} \Pi_k^{-m} \frac{\gamma_k}{v_k} = \frac{1}{m - v^*\xi}.$$

Moreover, for any positive sequence (α_n) such that $\lim_{n\to+\infty} \alpha_n = 0$ and all $C \in \mathbb{R}$,

$$\lim_{n\to+\infty} v_n \Pi_n^m \left[\sum_{k=1}^{n} \Pi_k^{-m} \frac{\gamma_k}{v_k} \alpha_k + C\right] = 0.$$

In order to introduce our theoretical main results, we need the following technical assumptions.

Assumptions:

(A_1) $\mathbf{K} : \mathbb{R}^d \longrightarrow \mathbb{R}$ is a continuous bounded function satisfying:

$$\int_{\mathbb{R}^d} \mathbf{K}(u)du = 1 \ , \ \forall j \in \{1, \dots, d\}, \int_{\mathbb{R}} u_j \, \mathbf{K}(u)du_j = 0 \ \text{ and } \int_{\mathbb{R}^d} u_j^2 \, |\mathbf{K}(u)|du < \infty.$$

(A_2) (i) $(\gamma_n) \in \mathcal{GS}(-\alpha)$, with $\alpha \in \left(\frac{1}{2}, 1\right]$.

(ii) $(h_n) \in \mathcal{GS}(-a)$, with $a \in (0, 1]$.

(iii) $\lim_{n\to+\infty} (n\gamma_n) \in \left(\min\{2a, \frac{\alpha - ad}{2}\}, \infty\right]$.

(A_3) (i) f_X is bounded, twice differentiable and for all $i, j \in \{1, \dots, d\}$, $f^{(2)}_{X_{ij}}$ is bounded.

(ii) a is bounded, twice differentiable with respect to x and for all $i, j \in \{1, \dots, d\}$, $a^{(2)}_{ij}$ is bounded.

(iii) $f_{(X,Y)}$ is bounded and twice continuously differentiable with respect to x.

Discussion of the assumptions:
All these assumptions are standard and are generally assumed within the context of non-parametric estimation. Classical assumption (A_1) provides regularity conditions on the kernel density estimator introduced by [34] and [30]. It is widely used in the non-parametric framework for the functional estimation. Assumption (A_2) on the stepsize and the bandwidth was used in the recursive framework for the estimation of the density function in [26], [37] and for the distribution function estimation in [38]. Furthermore, assumption $(A_2)(iii)$ on the limit of $(n\gamma_n)$ as n goes to infinity is frequent in the framework of stochastic approximation algorithms. It implies, in particular, that the limit of $(n\gamma_n)^{-1}$ is finite. Moreover, assumptions in $(A3)$ are technical conditions imposed in order to ensure the reliability of proofs. Those conditions on the density of the couple (X, Y) were applied in the non-recursive framework for the estimation of the regression function [29,44] and in the recursive framework [27], [39].
In this section, we need to recall the following proposition which introduces the bias and the variance of f_n. The proof of this result was depicted in [26].

Bias and variance of f_n:

Proposition 1
Under assumptions $(A_1)-(A_3)$, and assuming that, for all $i,j\in\{1,\dots,d\}$, $f^{(2)}_{X_{ij}}$ is continuous at x, we obtain

1. If $a\in\left(0,\frac{\alpha}{d+4}\right]$, then
$$\mathbb{E}[f_n(x)]-f_X(x)=\frac{1}{2(1-2a\xi)}\left(\sum_{j=1}^{d}\mu_j(\mathbf{K})f_{X^{(2)}_{jj}}(x)\right)h_n^2+\mathbf{o}\left(h_n^2\right). \tag{6}$$
If $a\in\left(\frac{\alpha}{d+4},1\right)$, then
$$\mathbb{E}[f_n(x)]-f_X(x)=\mathbf{o}\left(\sqrt{\gamma_n h_n^{-d}}\right). \tag{7}$$
2. If $a\in\left(0,\frac{\alpha}{d+4}\right)$, then
$$Var[f_n(x)]=\mathbf{o}\left(h_n^4\right). \tag{8}$$
If $a\in\left[\frac{\alpha}{d+4},1\right)$, then
$$Var[f_n(x)]=\frac{1}{2-(\alpha-ad)\xi}f_X(x)R(\mathbf{K})\frac{\gamma_n}{h_n^d}+\mathbf{o}\left(\gamma_n h_n^{-d}\right). \tag{9}$$

3. Main results

In order to explore the asymptotic properties of our estimator π_n, we need first to introduce the following proposition which provides the bias and the variance of a_n.

3.1. Bias and variance of a_n:

Proposition 2
Under assumptions $(A_1)-(A_3)$, and assuming that, for all $i,j\in\{1,\dots,d\}$, $a^{(2)}_{ij}$ is continuous at x, we obtain

1. If $a\in\left(0,\frac{\alpha}{d+4}\right]$, then
$$\mathbb{E}[a_n(x,y)]-a(x,y)=\frac{1}{2(1-2a\xi)}\left(\sum_{j=1}^{d}\mu_j(\mathbf{K})a^{(2)}_{jj}(x,y)\right)h_n^2+\mathbf{o}\left(h_n^2\right). \tag{10}$$
If $a\in\left(\frac{\alpha}{d+4},1\right)$, then
$$\mathbb{E}[a_n(x,y)]-a(x,y)=\mathbf{o}\left(\sqrt{\gamma_n h_n^{-d}}\right). \tag{11}$$
2. If $a\in\left(0,\frac{\alpha}{d+4}\right)$, then
$$Var[a_n(x,y)]=\mathbf{o}\left(h_n^4\right). \tag{12}$$
If $a\in\left[\frac{\alpha}{d+4},1\right)$, then
$$Var[a_n(x,y)]=\frac{1}{2-(\alpha-ad)\xi}a(x,y)R(\mathbf{K})\frac{\gamma_n}{h_n^d}+\mathbf{o}\left(\gamma_n h_n^{-d}\right). \tag{13}$$

Proof
We have

$$\mathbb{E}[a_n(x,y)] - a(x,y) = \Pi_n \sum_{k=1}^{n} \Pi_k^{-1}\gamma_k(\mathbb{E}[\mathcal{Z}_k(x,y)] - a(x,y)) + \Pi_n[a_0(x,y) - a(x,y)].$$

Relying upon the assumption (A_1), we have $\int_{\mathbb{R}^d} \mathbf{K}(z)dz = 1$. Hence, it follows that

$$\begin{aligned}
\mathbb{E}[\mathcal{Z}_k(x,y)] - a(x,y) &= \int_{\mathbb{R}^d} h_k^{-d}\mathbf{K}\left(\frac{x-t}{h_k}\right)\mathbb{E}[\chi_k(y)|X=t]f_X(t)dt - \int_{\mathbb{R}^d} \mathbf{K}(t)\, a(x,y)dt \\
&= \int_{\mathbb{R}^d} h_k^{-d}\mathbf{K}\left(\frac{x-t}{h_k}\right)a(t,y)dt - \int_{\mathbb{R}^d} \mathbf{K}(t)\, a(x,y)dt \\
&= \int_{\mathbb{R}^d} \mathbf{K}(z)\left[a(x - zh_k, y) - a(x,y)\right]dz.
\end{aligned}$$

Moreover, Taylor's expansion with integral remainder ensures that

$$\begin{aligned}
\mathbb{E}[\mathcal{Z}_k(x,y)] - a(x,y) &= \int_{\mathbb{R}^d} \mathbf{K}(z)\left[\sum_{i=1}^{d} \frac{\partial a}{\partial x_i}(x,y)z_i h_k + \int_0^1 (1-t)\sum_{i,j=1}^{d}\frac{\partial^2 a}{\partial x_i \partial x_j}(x - tzh_k, y)z_i z_j h_k^2 dt\right]dz \\
&= \frac{h_k^2}{2}\sum_{j=1}^{d}\mu_j(\mathbf{K})a_{jj}^{(2)}(x,y) + h_k^2\eta_k(x).
\end{aligned}$$

where, $\eta_k(x) := \sum_{i,j=1}^{d}\int_{\mathbb{R}^d}\int_0^1 (1-t)\left[a_{ij}^{(2)}(x - tzh_k, y) - a_{ij}^{(2)}(x,y)\right] z_i z_j\, \mathbf{K}(z)dtdz.$
Hence,

$$\begin{aligned}
\mathbb{E}[a_n(x,y)] - a(x,y) &= \frac{1}{2}\sum_{j=1}^{d}\mu_j(\mathbf{K})a_{jj}^{(2)}(x,y)\Pi_n\sum_{k=1}^{n}\Pi_k^{-1}\gamma_k h_k^2 + \Pi_n\sum_{k=1}^{n}\Pi_k^{-1}\gamma_k h_k^2\eta_k(x) \\
&\quad + \Pi_n[a_0(x,y) - a(x,y)].
\end{aligned}$$

Since $a_{ij}^{(2)}$ is bounded and continuous at x for all $i,j \in \{1,\dots,d\}$, we obtain $\lim_{k\to+\infty}\eta_k(x) = 0$.
− For the case $a \leqslant \alpha/(d+4)$, we have $\lim_{n\to+\infty}(n\gamma_n) > 2a$ and then $1 - 2a\xi > 0$. The application of lemma 1 enables us to write

$$\mathbb{E}[a_n(x,y)] - a(x,y) = \frac{1}{2(1-2a\xi)}\left(\sum_{j=1}^{d}\mu_j(\mathbf{K})a_{jj}^{(2)}(x,y)\right)h_n^2 + \mathbf{o}\left(h_n^2\right).$$

− For the case $a > \alpha/(d+4)$, we have $\lim_{n\to+\infty}(n\gamma_n) > \frac{\alpha-a}{2}$ which yields that $h_n^2 = \mathbf{o}\left(\sqrt{\gamma_n h_n^{-d}}\right)$. Then, the use of lemma 1 leads to

$$\mathbb{E}[a_n(x,y)] - a(x,y) = \mathbf{o}\left(\sqrt{\gamma_n h_n^{-d}}\right).$$

Therefore, the claimed result (11) is established.

For the variance, and owing to the independence of X_i, for $i = 1, \dots, n$, it's obvious that

$$\begin{aligned} Var[a_n(x,y)] &= \Pi_n^2 \sum_{k=1}^n \Pi_k^{-2} \gamma_k^2 Var[\mathcal{Z}_k(x,y)] \\ &= \Pi_n^2 \sum_{k=1}^n \Pi_k^{-2} \gamma_k^2 \left(\mathbb{E}[\mathcal{Z}_k^2(x,y)] - \mathbb{E}[\mathcal{Z}_k(x,y)]^2 \right) \\ &= \Pi_n^2 \sum_{k=1}^n \Pi_k^{-2} \gamma_k^2 \left(\int_{\mathbb{R}^d} h_k^{-d} \mathbf{K}^2(z) a(x - zh_k, y) dz - \left(\int_{\mathbb{R}^d} \mathbf{K}(z) a(x - zh_k, y) dz \right)^2 \right). \end{aligned}$$

As matter of fact, the Taylor's expansions theorem ensures that

$$Var[a_n(x,y)] = \Pi_n^2 \sum_{k=1}^n \Pi_k^{-2} \gamma_k^2 h_k^{-d} \left[a(x,y) \int_{\mathbb{R}^d} \mathbf{K}^2(z)\, dz + \nu_k(x) - h_k^d \eta_k(x) \right],$$

where,

$$\nu_k(x) = \int_{\mathbb{R}^d} \mathbf{K}^2(z) [a(x - zh_k, y) - a(x,y)] dz \text{ and } \eta_k(x) = \left(\int_{\mathbb{R}^d} \mathbf{K}(z) a(x - zh_k, y) dz \right)^2.$$

− For the case $a \geqslant \alpha/(d+4)$, we have $\lim_{n \to +\infty} (n\gamma_n) > \frac{\alpha - ad}{2}$ and then $1 - 2a\xi > 0$. Since a is bounded and continuous at x, we have $\lim_{k \to +\infty} \nu_k(x) = 0$ and $\lim_{k \to +\infty} h_k \eta_k(x) = 0$. Therefore, the application of lemma 1 ensures that

$$\begin{aligned} Var[a_n(x,y)] &= \Pi_n^2 \sum_{k=1}^n \Pi_k^{-2} \gamma_k^2 h_k^{-d} \left[a(x,y) R(\mathbf{K}) + \nu_k(x) - h_k^d \eta_k(x) \right] \\ &= \frac{1}{2 - (\alpha - ad)\xi} \frac{\gamma_n}{h_n^d} [a(x,y) R(\mathbf{K}) + \mathbf{o}(1)]. \end{aligned}$$

Thus, this leads to the result displayed in (13).
− For the case $a < \alpha/(d+4)$, we have $\lim_{n \to +\infty} (n\gamma_n) > 2a$ which provides that $\gamma_n h_n^{-d} = \mathbf{o}\left(h_n^4\right)$. By applying lemma 1, we infer that

$$\begin{aligned} Var[a_n(x,y)] &= \Pi_n^2 \sum_{k=1}^n \Pi_k^{-2} \gamma_k \mathbf{o}\left(h_k^4\right) \\ &= \mathbf{o}\left(h_n^4\right). \end{aligned}$$

□

In the following theorem, we introduce our main result which provides the bias and the variance of our CCDF multivariate estimator π_n.

3.2. Bias and variance of π_n:

Theorem 1
Let assumptions $(A_1) - (A_3)$ hold and assume that, for all $i, j \in \{1, \dots, d\}$, $a_{ij}^{(2)}$ and $f_{X_{ij}}^{(2)}$ are continuous at x, we obtain

1. If $a \in \left(0, \frac{\alpha}{d+4}\right]$, then

$$\mathbb{E}[\pi_n(y|x)] - \pi(y|x) = \frac{1}{2(1-2a\xi)} \frac{1}{f_X(x)} \sum_{j=1}^{d} \mu_j(\mathbf{K}) \Big[\pi_{jj}^{(2)}(y|x) f_X(x) + 2\pi_j^{(1)}(y|x) f_{X_j}^{(1)}(x)\Big] h_n^2 + \mathbf{o}\left(h_n^2\right). \tag{14}$$

If $a \in \left(\frac{\alpha}{d+4}, 1\right)$, then

$$\mathbb{E}[\pi_n(y|x)] - \pi(y|x) = \mathbf{o}\left(\sqrt{\gamma_n h_n^{-d}}\right). \tag{15}$$

2. If $a \in \left(0, \frac{\alpha}{d+4}\right)$, then

$$Var[\pi_n(y|x)] = \mathbf{o}\left(h_n^4\right). \tag{16}$$

If $a \in \left[\frac{\alpha}{d+4}, 1\right)$, then

$$Var[\pi_n(y|x)] = \frac{R(\mathbf{K})}{2-(\alpha-ad)\xi} \frac{\pi(y|x)(1-\pi(y|x))}{f_X(x)} \frac{\gamma_n}{h_n^d} + \mathbf{o}\left(\frac{\gamma_n}{h_n^d}\right). \tag{17}$$

It is noteworthy that the bias and the variance of the estimator π_n defined by the stochastic approximation algorithm (1) mainly depend on the choice of the stepsize (γ_n).

Proof
Our proof rests upon the following decomposition, for $f_n \neq 0$

$$\pi_n(y|x) - \pi(y|x) = A_n(x,y) \frac{f_X(x)}{f_n(x)}, \tag{18}$$

with

$$A_n(x,y) = \frac{1}{f_X(x)} \left(a_n(x,y) - a(x,y)\right) - \frac{\pi(y|x)}{f_X(x)} \left(f_n(x) - f_X(x)\right).$$

It follows from (18) that the asymptotic behavior of $\pi_n(y|x) - \pi(y|x)$ can be deduced from the one of $A_n(x,y)$. For the bias of π_n, we can state

$$\mathbb{E}[A_n(x,y)] = \frac{1}{f_X(x)} \left(\mathbb{E}[a_n(x,y)] - a(x,y)\right) - \frac{\pi(y|x)}{f_X(x)} \left(\mathbb{E}[f_n(x)] - f_X(x)\right).$$

Now, using the first bias part of proposition 1 and proposition 2 and considering the fact that $a(x,y) = \pi(y|x) f_X(x)$, then by combining the assertions (6), (7), (10) and (11); we obtain the relations (14) and (15).
In order to confirm the variance of π_n statement, we have

$$Var[A_n(x,y)] = \frac{1}{(f_X(x))^2} Var[a_n(x,y)] + \frac{(\pi(y|x))^2}{(f_X(x))^2} Var[f_n(x)] - 2\frac{\pi(y|x)}{(f_X(x))^2} Cov(a_n(x,y), f_n(x)). \tag{19}$$

Given that the X_k's are independent, then for all $i = 1, \dots, n$, $k = 1, \dots, n$ and $i \neq k$, we have $Cov(\mathcal{Z}_k(x,y), \mathcal{W}_i(x)) = 0$. Using Taylor's expansion with integral remainder and lemma 1, classical computations entail

$$Cov(a_n(x,y), f_n(x)) = \frac{R(\mathbf{K})}{2-(\alpha-ad)\xi} \pi(y|x) f_X(x) \frac{\gamma_n}{h_n^d} + \mathbf{o}\left(\frac{\gamma_n}{h_n^d}\right). \tag{20}$$

Consequently, relations (16) and (17) follow from the combination of assertions (8), (9), (12), (13) and (20). For the case $a \geq \alpha/(d+4)$, we deduce with (19) that

$$Var[\pi_n(y|x)] = \frac{R(\mathbf{K})}{2-(\alpha-ad)\xi}\frac{\pi(y|x)(1-\pi(y|x))}{f_X(x)}\frac{\gamma_n}{h_n^d} + \mathbf{o}\left(\frac{\gamma_n}{h_n^d}\right).$$

Proceeding with the same reasoning applied for the case $a < \alpha/(d+4)$, we obtain the desired result (16). $\square$

In the sequel, let us present the following theorem which identifies the asymptotic normality of our recursive estimator π_n. Throughout this paper, we shall denote convergence in probability, convergence in distribution and the Gaussian distribution by $\xrightarrow[n\to+\infty]{\mathbb{P}}$, $\xrightarrow[n\to+\infty]{\mathcal{D}}$ and $\mathcal{N}$, respectively.

3.3. *Weak pointwise convergence rate of* π_n

Theorem 2

Let assumptions $(A_1)-(A_3)$ hold.

1. If there exists a non-negative real c such that $\gamma_n^{-1}h_n^{d+4} \xrightarrow[n\to+\infty]{} c$, then

$$\sqrt{\gamma_n^{-1}h_n^d}\,(\pi_n(y|x)-\pi(y|x)) \xrightarrow[n\to+\infty]{\mathcal{D}} \mathcal{N}\left(\sqrt{c}\,m(x,y)\ ,\ \sigma^2(x,y)\right). \tag{21}$$

   with

$$m(x,y) = \frac{1}{2(1-2a\xi)}\frac{1}{f_X(x)}\sum_{j=1}^{d}\mu_j(\mathbf{K})\left[\pi_{jj}^{(2)}(y|x)f_X(x) + 2\pi_j^{(1)}(y|x)f_{X_j}^{(1)}(x)\right]$$

   and

$$\sigma^2(x,y) = \frac{R(\mathbf{K})}{2-(\alpha-ad)\xi}\frac{\pi(y|x)(1-\pi(y|x))}{f_X(x)}.$$

2. If $\gamma_n^{-1}h_n^{d+4} \xrightarrow[n\to+\infty]{} \infty$, then

$$\frac{1}{h_n^2}\left(\pi_n(y|x)-\pi(y|x)\right) \xrightarrow[n\to+\infty]{\mathbb{P}} m(x,y).$$

Proof

We can write, for all $n \geq 0$, $x \in \mathbb{R}^d, y \in \mathbb{R}^q$,

$$\begin{aligned} A_n(x,y) - \mathbb{E}[A_n(x,y)] &= \frac{1}{f_X(x)}[a_n(x,y) - \mathbb{E}[a_n(x,y)]] - \frac{\pi(y|x)}{f_X(x)}[f_n(x) - \mathbb{E}[f_n(x)]] \\ &= \frac{1}{f_X(x)}\Pi_n\sum_{k=1}^{n}S_k(x,y), \end{aligned} \tag{22}$$

where

$$S_k(x,y) := \Pi_k^{-1}\gamma_k\left(T_k(x,y) - \mathbb{E}[T_k(x,y)]\right) \quad \text{and} \quad T_k(x,y) = \mathcal{Z}_k(x,y) - \pi(y|x)\mathcal{W}_k(x).$$

This proof falls naturally into the application of Lyapunov's theorem for $S_k(x,y)$.

On the one hand, we can write

$$\begin{aligned} u_n^2 := \sum_{k=1}^{n}Var[S_k(x,y)] = \sum_{k=1}^{n}\Pi_k^{-2}\gamma_k^2 Var\left[\mathcal{Z}_k(x,y) - \pi(y|x)\mathcal{W}_k(x)\right] \\ - 2\pi(y|x)\,Cov\left(\mathcal{Z}_k(x,y), \mathcal{W}_k(x)\right)\Big). \end{aligned}$$

Therefore, by applying lemma 1, it can be deduced that

$$\begin{aligned} u_n^2 &= \sum_{k=1}^{n} \Pi_k^{-2}\gamma_k^2 h_k^{-d}\Big(R(\mathbf{K})f_X(x)\pi(y|x)\left(1-\pi(y|x)\right)+\mathbf{o}(1)\Big) \\ &= \frac{f_X(x)^2}{\Pi_n^2}\frac{\gamma_n}{h_n^d}[\sigma^2(x,y)+\mathbf{o}\,(1)]. \end{aligned} \tag{23}$$

On the other hand, we can write

$$\mathbb{E}[|T_k(x,y)|^{2+p}] = \mathbf{O}\left(\frac{1}{h_k^{d(1+p)}}\right), \quad \forall p>0.$$

This yields,

$$\begin{aligned} \sum_{k=1}^{n}\mathbb{E}[|S_k(x,y)|^{2+p}] &= \mathbf{O}\left(\sum_{k=1}^{n}\Pi_k^{-(2+p)}\gamma_k^{2+p}\mathbb{E}\left[|T_k(x,y)|^{2+p}\right]\right) \\ &= \mathbf{O}\left(\sum_{k=1}^{n}\Pi_k^{-(2+p)}\gamma_k^{2+p}\frac{1}{h_k^{d(1+p)}}\right). \end{aligned}$$

For the application of lemma 1, let us assume that there exists a positive real p such that

$$\lim_{n\to+\infty}(n\gamma_n) > \frac{1+p}{2+p}(\alpha - ad).$$

Then, we obtain

$$\sum_{k=1}^{n}\mathbb{E}[|S_k(x,y)|^{2+p}] = \mathbf{O}\left(\frac{\gamma_n^{1+p}}{\Pi_n^{2+p}h_n^{d(1+p)}}\right).$$

It follows that

$$\frac{1}{u_n^{2+p}}\sum_{k=1}^{n}\mathbb{E}[|S_k(x,y)|^{2+p}] = \mathbf{O}\left(\frac{\gamma_n^{1+p}}{u_n^{2+p}\Pi_n^{2+p}h_n^{d(1+p)}}\right).$$

As a sequel, with the assertion (23), we can write

$$\frac{1}{u_n^{2+p}}\sum_{k=1}^{n}\mathbb{E}[|S_k(x,y)|^{2+p}] = \mathbf{O}\left(\left(\frac{\gamma_n}{h_n^d}\right)^{p/2}\right) = \mathbf{o}\,(1)\,.$$

Moreover, since we have

$$\lim_{n\to+\infty}\frac{1}{u_n^{2+p}}\sum_{k=1}^{n}\mathbb{E}\left[|S_k(x,y)-\mathbb{E}[S_k(x,y)]|^{2+p}\right] = \lim_{n\to+\infty}\frac{1}{u_n^{2+p}}\sum_{k=1}^{n}\mathbb{E}[|S_k(x,y)|^{2+p}] = 0,$$

then, by applying the Lyapunov theorem, we get

$$\frac{1}{u_n}\sum_{k=1}^{n}(S_k(x,y)-\mathbb{E}[S_k(x,y)]) \xrightarrow[n\to+\infty]{\mathcal{D}} \mathcal{N}\,(0\,,1)\,.$$

This implies

$$\frac{1}{u_n}\sum_{k=1}^{n}S_k(x,y) \xrightarrow[n\to+\infty]{\mathcal{D}} \mathcal{N}\,(0\,,1)\,.$$

Additionally, the relations (18) and (22) ensure that

$$\frac{1}{u_n \Pi_n} f_X(x) \left(\pi_n(y|x) - \mathbb{E}[\pi_n(y|x)]\right) \underset{n\to+\infty}{\xrightarrow{\mathcal{D}}} \mathcal{N}(0\,,1)\,. \tag{24}$$

Given that

$$u_n^2 = \frac{f_X(x)^2}{\Pi_n^2}\frac{\gamma_n}{h_n^d}[\sigma^2(x,y) + \mathbf{o}(1)],$$

and by replacing u_n with its value in relation (24), we deduce that

$$\sqrt{\gamma_n^{-1} h_n^d}\left(\pi_n(y|x) - \mathbb{E}[\pi_n(y|x)]\right) \underset{n\to+\infty}{\xrightarrow{\mathcal{D}}} \mathcal{N}\left(0\,,\sigma^2(x,y)\right). \tag{25}$$

Since we have $\sqrt{\gamma_n^{-1} h_n^{d+4}} \underset{n\to+\infty}{\longrightarrow} \sqrt{c}$, then the convergence (21) follows from the combination of relations (14), (15) and convergence (25). □

In order to assess the asymptotic quality of the CCDF recursive estimator π_n, we set up the Mean Weighted Integrated Squared Error ($MWISE$).

3.4. Asymptotic expressions of $MWISE[\pi_n]$

We first introduce the $MWISE$ expression:

$$MWISE[\pi_n] = \int_{\mathbb{R}^{d+q}} \left[\left(\mathbb{E}[\pi_n(y|x)] - \pi(y|x)\right)^2 + Var[\pi_n(y|x)]\right] {f_X}^2(x) f_{X,Y}(x,y)dxdy. \tag{26}$$

Proposition 3
The $MWISE$ of the estimator π_n is maintained as follows.
If $a \in \left(0, \frac{\alpha}{d+4}\right)$, then

$$MWISE[\pi_n] = \frac{1}{4}\frac{I_1}{(1-2a\xi)^2} h_n^4 + \mathbf{o}(h_n^4).$$

If $a = \frac{\alpha}{d+4}$, then

$$MWISE[\pi_n] = \frac{I_2}{2-(\alpha - ad)\xi} R(\mathbf{K})\gamma_n h_n^{-d} + \frac{1}{4}\frac{I_1}{(1-2a\xi)^2} h_n^4 + \mathbf{o}(h_n^4).$$

If $a \in \left(\frac{\alpha}{d+4}, 1\right)$, then

$$MWISE[\pi_n] = \frac{I_2}{2-(\alpha - ad)\xi} R(\mathbf{K})\gamma_n h_n^{-d} + \mathbf{o}\left(\gamma_n h_n^{-d}\right).$$

Proof
Based on the relation (26) and by distinguishing the different possible cases according to the expressions of the Bias ((14) and (15)) as well as the Variance ((16) and (17)), one can prove this proposition and find the required result. □

The following corollary ensures that the bandwidth which minimizes the $MWISE$ of π_n depends on the choice of the stepsize (γ_n). As a matter of fact, the corresponding $MWISE$ depends also on (γ_n).

Corollary 1
Let assumptions $(A_1)-(A_3)$ hold. To minimize the $MWISE$ of π_n, the bandwidth (h_n) must be equal to

$$\left(\left(\frac{d\,(1-2a\xi)^2}{2-(\alpha-ad)\xi}\frac{I_2}{I_1}R(\mathbf{K})\right)^{\frac{1}{d+4}}\gamma_n^{\frac{1}{d+4}}\right).$$

Hence, the corresponding $MWISE$ is determined by

$$MWISE[\pi_n]=\frac{d+4}{4d^{\frac{d}{d+4}}}\left(\frac{I_1}{(1-2a\xi)^2}\right)^{\frac{d}{d+4}}\left(\frac{I_2}{2-(\alpha-ad)\xi}\right)^{\frac{4}{d+4}}R(\mathbf{K})^{\frac{4}{d+4}}\gamma_n^{\frac{4}{d+4}}+\mathbf{o}\left(\gamma_n^{\frac{4}{d+4}}\right).$$

The following corollary holds in the special case where (γ_n) is chosen as $(\gamma_n)=(\gamma_0 n^{-1})$ in order to minimize the $MWISE[\pi_n]$.

Corollary 2
Let assumptions $(A_1)-(A_3)$ hold. To minimize the $MWISE$ of π_n, we need to opt for the stepsize (γ_n) in $\mathcal{GS}(-1)$ such that $\lim_{n\to\infty}(n\gamma_n)=\gamma_0$. Then the bandwidth (h_n) must be equal to

$$\left(\left(\frac{d\,(\gamma_0(d+4)-2)}{2(d+4)}\frac{I_2}{I_1}R(\mathbf{K})\right)^{\frac{1}{d+4}}n^{-\frac{1}{d+4}}\right).$$

Consequently, the corresponding $MWISE$ is identified by

$$MWISE[\pi_n]=\frac{(d+4)^{\frac{3d+8}{d+4}}}{4^{\frac{d+6}{d+4}}d^{\frac{d}{d+4}}}\gamma_0^2\left(\gamma_0(d+4)-2\right)^{-\frac{(2d+4)}{d+4}}I_1^{\frac{d}{d+4}}I_2^{\frac{4}{d+4}}R(\mathbf{K})^{\frac{4}{d+4}}n^{-\frac{4}{d+4}}+\mathbf{o}\left(n^{-\frac{4}{d+4}}\right).$$

In order to get the optimal choice of (γ_n), we deduce that the minimum of $MWISE[\pi_n]$ is achieved at $\gamma_0=1$. Hence, we introduce the following corollary.

Corollary 3
Let assumptions $(A_1)-(A_3)$ hold. To minimize the $MWISE$ of π_n, we must select the stepsize (γ_n) in $\mathcal{GS}(-1)$ such that $\lim_{n\to\infty}(n\gamma_n)=1$. Therefore, the optimal bandwidth (h_n) must equal

$$\left(\left(\frac{d(d+2)}{2(d+4)}\frac{I_2}{I_1}R(\mathbf{K})\right)^{\frac{1}{d+4}}n^{-\frac{1}{d+4}}\right). \tag{27}$$

As a result, the corresponding $MWISE$ is expressed by

$$MWISE[\pi_n]=\frac{(d+4)^{\frac{3d+8}{d+4}}}{4^{\frac{d+6}{d+4}}d^{\frac{d}{d+4}}(d+2)^{\frac{2d+4}{d+4}}}I_1^{\frac{d}{d+4}}I_2^{\frac{4}{d+4}}R(\mathbf{K})^{\frac{4}{d+4}}n^{-\frac{4}{d+4}}+\mathbf{o}\left(n^{-\frac{4}{d+4}}\right).$$

Remark 1
Note that, for the particular case where the stepsize (γ_n) is in $\mathcal{GS}(-1)$ such that $\lim_{n\to\infty}(n\gamma_n)=1$ and the bandwidth (h_n) is chosen such that $\lim_{n\to\infty}nh_n^{d+4}=0$ (which corresponds to undersmoothing), the asymptotic normality of our proposed estimator is indicated as follows

$$\sqrt{nh_n^d}\left(\pi_n(y|x)-\pi(y|x)\right)\xrightarrow[n\to+\infty]{\mathcal{D}}\mathcal{N}\left(0\,,\,\frac{d+4}{2(d+2)}R(\mathbf{K})\frac{\pi(y|x)(1-\pi(y|x))}{f_X(x)}\right). \tag{28}$$

The statistical inference of the CCDF multivariate non-recursive estimator $\widetilde{\pi}_n$ is addressed in our next section. The following results can be handled in nearly the same way as π_n. The unique difference lies in the fact that it pertains to a non-recursive case. (See [17] for more details of the univariate case.)

4. Asymptotic properties of $\widetilde{\pi}_n$

In order to tackle the asymptotic properties of our estimator π_n, we need first to introduce the following proposition which provides the bias and the variance of $\widetilde{\pi}_n$.

4.1. Bias and variance of $\widetilde{\pi}_n$

Proposition 4
Let assumptions (A_1) and (A_3) hold. Then the bias and variance of Nadaraya-Watson's estimator are displayed as follows.

1. The bias of $\widetilde{\pi}_n$:

$$\mathbb{E}[\widetilde{\pi}_n(y|x)] - \pi(y|x) = \frac{1}{2f_X(x)} \left(\sum_{j=1}^{d} \mu_j(\mathbf{K}) \left[\pi_{jj}^{(2)}(y|x) f_X(x) + 2\pi_j^{(1)}(y|x) {f_X}_j^{(1)}(x) \right] \right) h_n^2 + \mathbf{o}\left(h_n^2\right).$$

2. The variance of $\widetilde{\pi}_n$:

$$Var[\widetilde{\pi}_n(x)] = R(\mathbf{K}) \frac{\pi(y|x)(1-\pi(y|x))}{f_X(x)} \frac{1}{nh_n^d} + \mathbf{o}\left(\frac{1}{nh_n^d}\right).$$

The following proposition yields the distribution convergence rate of the non-recursive estimator.

4.2. Asymptotic normality of $\widetilde{\pi}_n$

Theorem 3
Let assumptions $(A_1) - (A_3)$ hold and suppose that $nh_n^{d+4} \underset{n\to+\infty}{\longrightarrow} 0$. Then,

$$\sqrt{nh_n^d}\left(\widetilde{\pi}_n(y|x) - \pi(y|x)\right) \underset{n\to+\infty}{\overset{\mathcal{D}}{\longrightarrow}} \mathcal{N}\left(0\ ,\ R(\mathbf{K}) \frac{\pi(y|x)(1-\pi(y|x))}{f_X(x)}\right). \tag{29}$$

Remark 2
It is obvious to infer from the expressions (28) and (29) that our CCDF proposed estimator is better than non-recursive one in terms of variance.

In the next subsection, we exhibit the expression of the Mean Weighted Integrated Squared Error of Nadaraya-Watson's estimator.

4.3. Asymptotic expression of $MWISE[\widetilde{\pi}_n]$

Corollary 4
The $MWISE$ expression of the non-recursive CCDF estimator is given by

$$MWISE[\widetilde{\pi}_n] = \frac{1}{4} I_1 h_n^4 + I_2 R(\mathbf{K}) \frac{1}{nh_n^d} + \mathbf{o}\left(h_n^4 + \frac{1}{nh_n^d}\right).$$

Proposition 5
Let assumptions (A_1) and (A_3) hold. To minimize the $MWISE$ of $\widetilde{\pi}_n$, the bandwidth (h_n) must be equal to

$$\left(\left(d\,\frac{I_2}{I_1} R(\mathbf{K})\right)^{\frac{1}{d+4}} n^{-\frac{1}{d+4}}\right). \tag{30}$$

Hence, the corresponding $MWISE$ is determined by

$$MWISE[\widetilde{\pi}_n] = \frac{d+4}{4d^{\frac{d}{d+4}}} I_2^{\frac{4}{d+4}} I_1^{\frac{d}{d+4}} R(\mathbf{K})^{\frac{4}{d+4}} n^{-\frac{4}{d+4}} + \mathbf{o}\left(n^{-\frac{4}{d+4}}\right).$$

5. Bandwidth selection

Although theoretical asymptotic study yields the optimal bandwidth, the fact that we do not know the density function makes it hard to interpret it in practice. Hence, kernel smoothing in non-parametric statistics requires the choice of a bandwidth parameter. This choice is crucial to obtain a good rate of consistency of the kernel estimators. It has a significant influence on the size of the bias. One has to find an appropriate bandwidth that produces an estimator which has a good balance between the bias and the variance of the estimator of the function $a(\cdot,\cdot)$ as well as $f(\cdot)$. It is worth noticing that the bandwidth selection methods reported in the literature can be divided into three broad classes: the cross-validation techniques, the plug-in ideas and the bootstrap procedure.
In this investigation, we are basically interested in the plug-in method. Altman and Leger developed an efficient method of bandwidth selection in [3], which minimizes an estimate of the mean weighted integrated squared error, using the density function as a weight function. For this reason, we followed the work of [37].

5.1. Plug-in bandwidth selection

As a result to the plug-in procedure, based on the expression of the $MWISE$, we estimate the unknown quantities I_1 and I_2 by elaborating asymptotic unbiased estimators. This process is known as a plug-in estimate. Basically, we introduce $(b_n) \in \mathcal{GS}(-\delta),\ \delta \in (0,1)$. In practice, [3] set $b_n = n^{-\delta} \min\left\{\widehat{s}, \frac{Q_3 - Q_1}{1.349}\right\}$, with $\widehat{s}$ being the sample standard deviation and Q_1, Q_3 being the first and third quartiles.
In the following and for the sake of simplicity, the kernel $\mathbf{K}$ we shall use is considered as a product of univariate kernels K satisfying $\int_{\mathbb{R}} K(x)dx = 1$.

In addition, we note:

$$I_1 = \mu^2(K)\left(J_1 - 2J_2 + J_3\right),$$

where

$$J_1 = \int_{\mathbb{R}^{d+q}} \left(\sum_{j=1}^{d} a_{jj}^{(2)}(x,y)\right)^2 f_{X,Y}(x,y)dxdy,\ J_3 = \int_{\mathbb{R}^{d+q}} \left(\sum_{j=1}^{d} f_X{}_{jj}^{(2)}(x)\right)^2 \pi^2(y|x) f_{X,Y}(x,y)dxdy,$$
$$J_2 = \int_{\mathbb{R}^{d+q}} \left(\sum_{j=1}^{d} a_{jj}^{(2)}(x,y)\right)\left(\sum_{j=1}^{d} f_X{}_{jj}^{(2)}(x)\right) \pi(y|x) f_{X,Y}(x,y)dxdy, \qquad \mu(K) = \int_{\mathbb{R}} z^2 K(z)dz.$$

5.2. Recursive estimator π_n:

To estimate the optimal bandwidth (27), we need to estimate I_1 and I_2. Here we can write

$$a_n(x,y) = \Pi_n \sum_{k=1}^{n} \Pi_k^{-1}\gamma_k\, h_k^{-d}\chi_k(y)\, \mathbf{K}\left(\frac{x - X_k}{h_k}\right) = \Pi_n \sum_{k=1}^{n} \Pi_k^{-1}\gamma_k h_k^{-d} \prod_{i=1}^{d} K\left(\frac{x - X_{k_i}}{h_k}\right)\chi_{k_i}(y)$$

and

$$f_n(x) = \Pi_n \sum_{k=1}^{n} \Pi_k^{-1}\gamma_k\, h_k^{-d}\, \mathbf{K}\left(\frac{x - X_k}{h_k}\right) = \Pi_n \sum_{k=1}^{n} \Pi_k^{-1}\gamma_k h_k^{-d} \prod_{i=1}^{d} K\left(\frac{x_i - X_{k_i}}{h_k}\right).$$

Estimation of I_1:

$$\widehat{J}_1 = \frac{\Pi_n^2}{n}\sum_{\substack{i,j,k=1\\ i\neq j\neq k}}^{n} \Pi_j^{-1}\Pi_k^{-1}\gamma_j\gamma_k {b'}_j^{-(d+2)}{b'}_k^{-(d+2)}\left[\sum_{v=1}^{d} K_{b'}^{(2)}\left(\frac{X_{i_v}-X_{j_v}}{b_j^{'}}\right)\prod_{\substack{l=1\\ l\neq v}}^{d} K_b\left(\frac{X_{i_l}-X_{j_l}}{b_j}\right)\right]$$
$$\times\left[\sum_{v=1}^{d} K_{b'}^{(2)}\left(\frac{X_{i_v}-X_{k_v}}{b_k^{'}}\right)\prod_{\substack{l=1\\ l\neq v}}^{d} K_b\left(\frac{X_{i_l}-X_{k_l}}{b_k}\right)\right]\prod_{s=1}^{q}\chi_{(j+1)_s}(Y_{i_s})\chi_{(k+1)_s}(Y_{i_s}),$$

$$\widehat{J}_2 = \frac{\Pi_n^2}{n}\sum_{\substack{i,j,k=1\\ i\neq j\neq k}}^{n} \Pi_j^{-1}\Pi_k^{-1}\gamma_j\gamma_k {b'}_j^{-(d+2)}{b'}_k^{-(d+2)}\left[\sum_{v=1}^{d} K_{b'}^{(2)}\left(\frac{X_{i_v}-X_{j_v}}{b_j^{'}}\right)\prod_{\substack{l=1\\ l\neq v}}^{d} K_b\left(\frac{X_{i_l}-X_{j_l}}{b_j}\right)\right]$$
$$\times\left[\sum_{v=1}^{d} K_{b'}^{(2)}\left(\frac{X_{i_v}-X_{k_v}}{b_k^{'}}\right)\prod_{\substack{l=1\\ l\neq v}}^{d} K_b\left(\frac{X_{i_l}-X_{k_l}}{b_k}\right)\right]\prod_{s=1}^{q}\chi_{(i+1)_s}(Y_{i_s})\chi_{(j+1)_s}(Y_{i_s}),$$

$$\widehat{J}_3 = \frac{\Pi_n^2}{n}\sum_{\substack{i,j,k,m=1\\ i\neq j\neq k\neq m}}^{n} \Pi_j^{-1}\Pi_k^{-1}\gamma_j\gamma_k {b'}_j^{-(d+2)}{b'}_k^{-(d+2)}\left[\sum_{v=1}^{d} K_{b'}^{(2)}\left(\frac{X_{i_v}-X_{j_v}}{b_j^{'}}\right)\prod_{\substack{l=1\\ l\neq v}}^{d} K_b\left(\frac{X_{i_l}-X_{j_l}}{b_j}\right)\right]$$
$$\times\left[\sum_{v=1}^{d} K_{b'}^{(2)}\left(\frac{X_{i_v}-X_{k_v}}{b_k^{'}}\right)\prod_{\substack{l=1\\ l\neq v}}^{d} K_b\left(\frac{X_{i_l}-X_{k_l}}{b_k}\right)\right]\prod_{s=1}^{q}\chi_{(i+1)_s}(Y_{i_s})\chi_{(m+1)_s}(Y_{i_s}),$$

where K_b stands for a kernel with bandwidth b_n such that $\delta = -\frac{2}{5}$ and $K_{b'}^{(2)}$ corresponds to the second derivative of a kernel $K_{b'}$ with the associated bandwidth $b_n^{'}$ such that $\delta = -\frac{3}{14}$. Note that our choice of the parameter δ is based on the work of [37].
At this stage, we obtain

$$\widehat{I}_1 = \mu^2(K)\left(\widehat{J}_1 - 2\widehat{J}_2 + \widehat{J}_3\right).$$

Estimation of I_2:

$$\widehat{I}_2 = \frac{\Pi_n}{n}\sum_{\substack{i,k=1\\ i\neq k}}^{n}\Pi_k^{-1}\gamma_k b_k^{-1}\prod_{l=1}^{d} K_b\left(\frac{X_{i_l}-X_{k_l}}{b_k}\right)\prod_{s=1}^{q}\chi_{(i+1)_s}(Y_{i_s})\left(1-\chi_{(k+1)_s}(Y_{i_s})\right),$$

where K_b is a kernel with bandwidth b_n such that $\delta = -\dfrac{2}{5}$.

As a result, the plug-in estimator of (27) is determined by

$$(h_n) = \left(\left(\frac{d(d+2)}{2(d+4)}\right)^{\frac{1}{d+4}}\left(\frac{\widehat{I}_2}{\widehat{I}_1}\right)^{\frac{1}{d+4}} R(\mathbf{K})^{\frac{1}{d+4}} n^{-\frac{1}{d+4}}\right), \tag{31}$$

Eventually, an estimator of $MWISE[\pi_n]$ is specified by

$$\widehat{MWISE}[\pi_n] = \frac{(d+4)^{\frac{3d+8}{d+4}}}{4^{\frac{d+6}{d+4}} d^{\frac{d}{d+4}} (d+2)^{\frac{d+6}{d+4}}}\left(\widehat{I}_1\right)^{\frac{d}{d+4}}\left(\widehat{I}_2\right)^{\frac{4}{d+4}} R(\mathbf{K})^{\frac{1}{d+4}} n^{-\frac{4}{d+4}} + \mathbf{o}\left(n^{-\frac{4}{d+4}}\right).$$

5.3. Non-Recursive estimator $\widetilde{\pi}_n$:

To estimate the optimal bandwidth (30), we need to estimate I_1 and I_2.Therefore, we can state

$$\widetilde{a}_n(x,y)=\frac{1}{nh_n^d}\sum_{k=1}^{n}\chi_k(y)\,\mathbf{K}\left(\frac{x-X_k}{h_n}\right)=\frac{1}{nh_n^d}\sum_{k=1}^{n}\prod_{i=1}^{d}K\left(\frac{x-X_{k_i}}{h_n}\right)\chi_{k_i}(y)$$

and

$$\widetilde{f}_n(x)=\frac{1}{nh_n^d}\sum_{k=1}^{n}\mathbf{K}\left(\frac{x-X_k}{h_n}\right)=\frac{1}{nh_n^d}\sum_{k=1}^{n}\prod_{i=1}^{d}K\left(\frac{x_i-X_{k_i}}{h_n}\right).$$

Estimation of I_1:

$$\begin{aligned}\widetilde{J}_1=\frac{1}{n^3b_n^6}\sum_{\substack{i,j,k=1\\ i\neq j\neq k}}^{n}&\left[\sum_{v=1}^{d}K_{b'}^{(2)}\left(\frac{X_{i_v}-X_{j_v}}{b_n'}\right)\prod_{\substack{l=1\\ l\neq v}}^{d}K_b\left(\frac{X_{i_l}-X_{j_l}}{b_n}\right)\right]\\ &\times\left[\sum_{v=1}^{d}K_{b'}^{(2)}\left(\frac{X_{i_v}-X_{k_v}}{b_n'}\right)\prod_{\substack{l=1\\ l\neq v}}^{d}K_b\left(\frac{X_{i_l}-X_{k_l}}{b_n}\right)\right]\prod_{s=1}^{q}\chi_{(j+1)_s}(Y_{i_s})\chi_{(k+1)_s}(Y_{i_s}),\end{aligned}$$

$$\begin{aligned}\widetilde{J}_2=\frac{1}{n^3b_n^6}\sum_{\substack{i,j,k=1\\ i\neq j\neq k}}^{n}&\left[\sum_{v=1}^{d}K_{b'}^{(2)}\left(\frac{X_{i_v}-X_{j_v}}{b_n'}\right)\prod_{\substack{l=1\\ l\neq v}}^{d}K_b\left(\frac{X_{i_l}-X_{j_l}}{b_n}\right)\right]\\ &\times\left[\sum_{v=1}^{d}K_{b'}^{(2)}\left(\frac{X_{i_v}-X_{k_v}}{b_n'}\right)\prod_{\substack{l=1\\ l\neq v}}^{d}K_b\left(\frac{X_{i_l}-X_{k_l}}{b_n}\right)\right]\prod_{s=1}^{q}\chi_{(i+1)_s}(Y_{i_s})\chi_{(j+1)_s}(Y_{i_s}),\end{aligned}$$

$$\begin{aligned}\widetilde{J}_3=\frac{1}{n^4b_n^6}\sum_{\substack{i,j,k,m=1\\ i\neq j\neq k\neq m}}^{n}&\left[\sum_{v=1}^{d}K_{b'}^{(2)}\left(\frac{X_{i_v}-X_{j_v}}{b_n'}\right)\prod_{\substack{l=1\\ l\neq v}}^{d}K_b\left(\frac{X_{i_l}-X_{j_l}}{b_n}\right)\right]\\ &\times\left[\sum_{v=1}^{d}K_{b'}^{(2)}\left(\frac{X_{i_v}-X_{k_v}}{b_n'}\right)\prod_{\substack{l=1\\ l\neq v}}^{d}K_b\left(\frac{X_{i_l}-X_{k_l}}{b_n}\right)\right]\prod_{s=1}^{q}\chi_{(i+1)_s}(Y_{i_s})\chi_{(m+1)_s}(Y_{i_s}),\end{aligned}$$

where K_b is a kernel with bandwidth b_n such that $\delta=-\frac{2}{5}$ and $K_{b'}^{(2)}$ is the second derivative of a kernel $K_{b'}$ with the associated bandwidth b_n' such that $\delta=-\frac{3}{14}$.
Then, we obtain

$$\widetilde{I}_1=\mu^2(K)\left(\widetilde{J}_1-2\widetilde{J}_2+\widetilde{J}_3\right).$$

Estimation of I_2:

$$\widetilde{I}_2=\frac{1}{n^2b_n}\sum_{\substack{i,k=1\\ i\neq k}}^{n}\prod_{l=1}^{d}K_b\left(\frac{X_{i_l}-X_{k_l}}{b_n}\right)\prod_{s=1}^{q}\chi_{(i+1)_s}(Y_{i_s})\left(1-\chi_{(k+1)_s}(Y_{i_s})\right),$$

where K_b is a kernel with bandwidth b_n such that $\delta = -\dfrac{2}{5}$.
As a result, the plug-in estimator of (30) is denoted by

$$(h_n) = \left(\left(\frac{\widetilde{I}_2}{\widetilde{I}_1} \right)^{\frac{1}{d+4}} R(\mathbf{K})^{\frac{1}{d+4}} n^{-\frac{1}{d+4}} \right), \tag{32}$$

Finally, a non-recursive estimator of $MWISE[\pi_n]$ is provided by

$$\widetilde{MWISE}[\widetilde{\pi}_n] = \frac{5}{4} \left(\widetilde{I}_2 \right)^{\frac{4}{d+4}} \left(\widetilde{I}_1 \right)^{\frac{1}{d+4}} R(\mathbf{K})^{\frac{1}{d+4}} n^{-\frac{4}{d+4}} + \mathbf{o} \left(n^{-\frac{4}{d+4}} \right).$$

The major aim of our next section lies in comparing the performance of our recursive estimator (1) with that of non-recursive Nadaraya-Watson one (2).

6. Numerical applications

Let's start our numerical studies with some simulations with different dimensions Models.

6.1. Simulation studies

In order to compare the proposed recursive estimator with the Nadaraya-Watson non-recursive one, we consider three sample sizes: $n = 100$, 200 and 500, a fixed number of simulations: $N = 500$ and four distribution models:

- Model 1: $(X, Y) \in \mathbb{R} \times \mathbb{R}$:
  $Y = 2\sin(\pi X) + \epsilon$, where X follows the binomial distribution $\mathcal{B}(2, 1/3)$ and ϵ follows the normal distribution $\mathcal{N}(0, 1)$.

- Model 2: $(X, Y) \in \mathbb{R}^2 \times \mathbb{R}$:
  $Y = \exp(-X/2) + \epsilon$, where X follows the exponential distribution $\mathcal{P}(1/2)$ and ϵ follows the normal distribution $\mathcal{N}(0, 1/2)$.

- Model 3: $(X, Y) \in \mathbb{R}^3 \times \mathbb{R}^2$:
  $Y = AX + \epsilon$ with $A = \begin{pmatrix} 1 & 1 & 1 \\ 1 & -2 & 1 \end{pmatrix}$, $X = 0 \times \mathbb{1}_{Z<=0.5} + \mathbb{1}_{Z>0.5}$ where Z follows the uniform distribution $\mathcal{U}\left(\begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} \right)$ and ϵ follows the normal distribution $\mathcal{N}(0, 1/2)$.

- Model 4: $(X, Y) \in \mathbb{R}^{10} \times \mathbb{R}^{10}$:
  $Y = \exp(X) + \epsilon$, with $X = \lfloor Z \rfloor$ where Z follows the 10-dimensional normal distribution $\mathcal{N}(0_{10}, I_{10})$ and ϵ follows the normal distribution $\mathcal{N}(0, 1)$.

We denote by π_i^* the reference CCDF and by π_i the test CCDF. Then, we calculate the following two measures:
· Mean squared error: $MSE = \dfrac{1}{n} \sum\limits_{i=1}^{n} (\pi_i - \pi_i^*)^2$.
· The linear correlation: $Cor = \dfrac{Cov(\pi_i, \pi_i^*)}{\sigma(\pi_i)\sigma(\pi_i^*)}$.

In what follows, we portray the different steps of the simulation algorithm in the multivariate case.

6.2. Simulation Algorithm

Algorithm 1 K is the Gaussian kernel, d the dimension size, n the simple size, Np the number of observations and N the number of iterations.

Input: K, d, n, Np and N.

1: A random initialization of $\widehat{\Pi}^{(0)}$ (resp. $\widetilde{\Pi}^{(0)}$.)

2: **for** $l = 1, \ldots, N$ **do**

3: A random sample vectors $X_1, \ldots, X_d$ and Y of length n.

4: A choice value for the recursive bandwidth vectors $h_1, \ldots, h_n$. (resp. the non-recursive bandwidth values h_n) using the plug-in approach given in (31) (resp. (32)).

5: The choice of the stepsize $(\gamma_n) = \left(n^{-1}\right)$.

6: We fix $x_1, \ldots, x_d$ and consider an arbitrary sampling vector T of y of length N_p.

7: $\widehat{\pi}_l(y|x) = \dfrac{\sum\limits_{k=1}^{n} k\gamma_k \mathbb{1}_{\{Y_k \leqslant y\}} h_k^{-d} \prod\limits_{i=1}^{d} K\left(\dfrac{x_i - X_{k_i}}{h_k}\right)}{\sum\limits_{k=1}^{n} k\gamma_k h_k^{-d} \prod\limits_{i=1}^{d} K\left(\dfrac{x_i - X_{k_i}}{h_k}\right)}$ for the multivariate recursive CCDF estimator. (resp.

$\widetilde{\pi}_l(y|x) = \dfrac{\sum\limits_{k=1}^{n} \mathbb{1}_{\{Y_k \leqslant y\}} \prod\limits_{i=1}^{d} K\left(\dfrac{x_i - X_{k_i}}{h_k}\right)}{\sum\limits_{k=1}^{n} \prod\limits_{i=1}^{d} K\left(\dfrac{x_i - X_{k_i}}{h_k}\right)}$ for the multivariate non-recursive CCDF estimator).

$\widehat{\Pi}^{(l)} = \widehat{\pi}_l(T|x)$. (resp. $\widetilde{\Pi}^{(l)} = \widetilde{\pi}_l(T|x)$.)

8: **end for**

9: $\widehat{\pi} = N^{-1} \sum\limits_{l=1}^{N} \widehat{\Pi}^{(l)}$ (resp. $\widetilde{\pi} = N^{-1} \sum\limits_{l=1}^{N} \widetilde{\Pi}^{(l)}$.)

output: The vectors $\widehat{\pi}$ and $\widetilde{\pi}$.

			$x=0$		$x=2$	
Model	*MSE*/*Cor*	n	Nadaraya's estimator	Recursive estimator	Nadaraya's estimator	Recursive estimator
Model 1	*MSE*	100	3.671766e-07	**1.333452e-07**	1.342202e-05	**8.228683e-06**
		200	3.377698e-07	**1.185438e-07**	3.434738e-06	**2.050298e-06**
		500	1.748553e-07	**4.915888e-08**	1.407363e-06	**6.641233e-07**
	Cor	100	9.999992e-01	**9.999997e-01**	9.999773e-01	**9.999849e-01**
		200	9.999992e-01	**9.999997e-01**	9.999914e-01	**9.999947e-01**
		500	9.999997e-01	**9.999999e-01**	9.999962e-01	**9.999981e-01**

Table 1. Quantitative comparison between the recursive estimator and the non-recursive one with stepsize $(\gamma_n) = (n^{-1})$ through a plug-in method for Model 1.

			$x=(0,0)$		$x=(1,1)$	
Model	*MSE*/*Cor*	n	Nadaraya's estimator	Recursive estimator	Nadaraya's estimator	Recursive estimator
Model 2	*MSE*	100	0.005299303	**0.002345599**	0.011464770	**0.005503603**
		200	0.004444445	**0.001976603**	0.009766589	**0.004850061**
		500	0.003609582	**0.001795659**	0.006988903	**0.004007790**
	Cor	100	0.993278442	**0.997042336**	0.985998000	**0.993297472**
		200	0.993932848	**0.997313718**	0.987516900	**0.993740343**
		500	0.994773451	**0.997402840**	0.990016792	**0.994209320**

Table 2. Quantitative comparison between the recursive estimator and the non-recursive one with stepsize $(\gamma_n) = (n^{-1})$ through a plug-in method for Model 2.

			$x = (0, 0, 0)$		$x = (1, 1, 1)$	
Model	MSE/Cor	n	Nadaraya's estimator	Recursive estimator	Nadaraya's estimator	Recursive estimator
Model 3	MSE	100	0.005170397	**0.003118756**	0.007481887	**0.002756530**
		200	0.004968340	**0.002747727**	0.007215537	**0.002449973**
		500	0.004639918	**0.002444590**	0.007200206	**0.002260629**
	Cor	100	0.989849852	**0.993964835**	0.989210840	**0.996122260**
		200	0.990247670	**0.994651835**	0.989322443	**0.996398260**
		500	0.990928518	**0.995271000**	0.989458730	**0.996602105**

Table 3. Quantitative comparison between the recursive estimator and the non-recursive one with stepsize $(\gamma_n) = (n^{-1})$ through a plug-in method for Model 3.

			$x = (0, 0, 0, 0, 0, 0, 0, 0, 0, 0)$	
Model	MSE/Cor	n	Nadaraya's estimator	Recursive estimator
Model 4	MSE	100	0.012132600	**0.008651937**
		200	0.012042460	**0.007971624**
		500	0.009422018	**0.006973041**
	Cor	100	0.982886350	**0.987784007**
		200	0.984891065	**0.987888565**
		500	0.985115900	**0.989107240**

Table 4. Quantitative comparison between the recursive estimator and the non-recursive one with stepsize $(\gamma_n) = (n^{-1})$ through a plug-in method for Model 4.

Departing from Tables 1, 2, 3 and 4, we conclude that:

1. The MSE of the proposed recursive estimator with stepsize $(\gamma_n) = (n^{-1})$ through a plug-in method is smaller than that of Nadaraya-Watson's non-recursive estimator.

2. The estimators get closer to the true CCDF function as sample size increases, i.e., the MSE decreases as the simple size increases and therefore the Cor increases as the sample size increases.

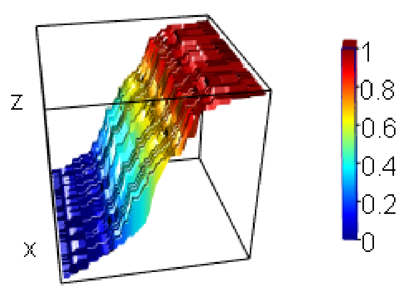

Figure 1. The reference CCDF for Model 1 for one simple simulation with $n = 200$.

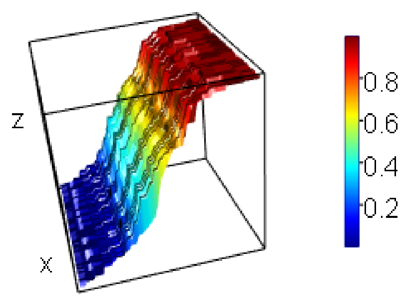

Figure 2. The recursive CCDF estimator for Model 1 for one simple simulation with $n = 200$.

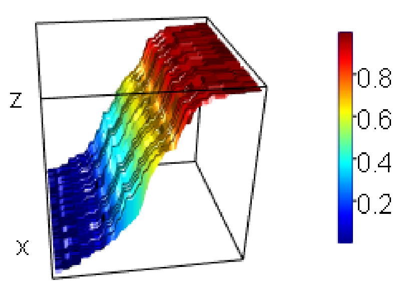

Figure 3. The non-recursive CCDF estimator for Model 1 for one simple simulation with $n = 200$.

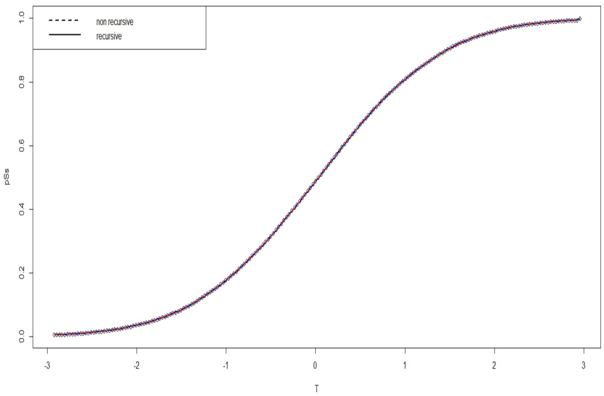

Figure 4. Qualitative comparison between the recursive estimator and the non-recursive one for Model 1 with $n = 200$, $N = 500$ and $x = 0$.

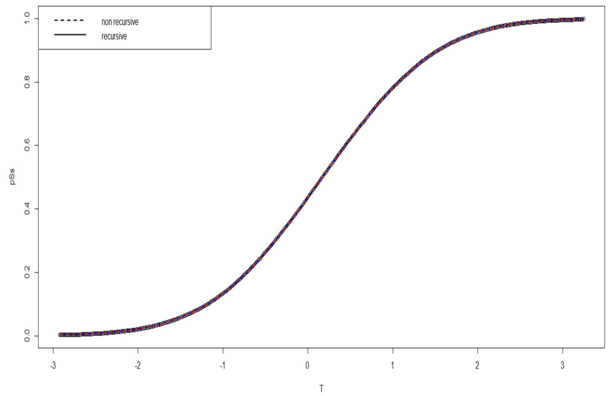

Figure 5. Qualitative comparison between the recursive estimator and the non-recursive one for Model 1 with $n = 500$, $N = 500$ and $x = 0$.

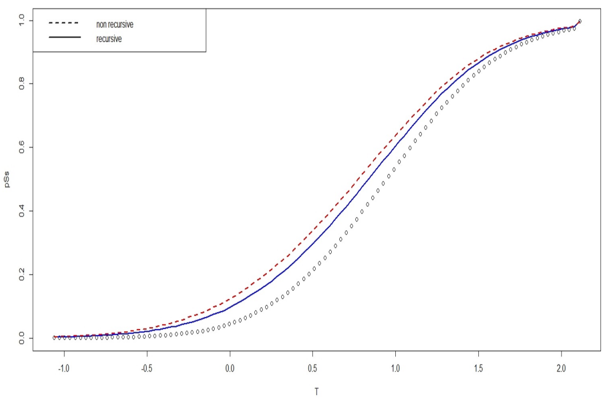

Figure 6. Qualitative comparison between the recursive estimator and the non-recursive one for Model 2 with $n = 100$, $N = 500$ and $x = (0, 0)$.

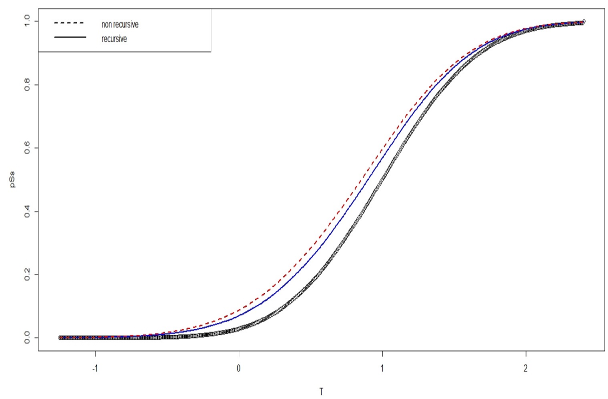

Figure 7. Qualitative comparison between the recursive estimator and the non-recursive one for Model 2 with $n = 500$, $N = 500$ and $x = (0, 0)$.

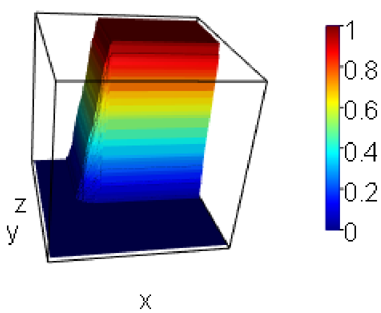

Figure 8. The reference CCDF for Model 3 for one simple simulation with $n = 500$ and $x = (1, 1, 1)$.

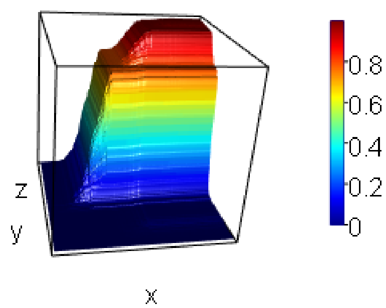

Figure 9. The recursive CCDF estimator for Model 3 for one simple simulation with $n = 500$ and $x = (1, 1, 1)$.

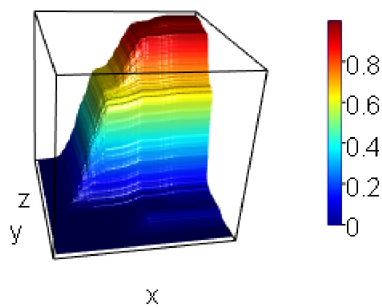

Figure 10. The non-recursive CCDF estimator for Model 3 for one simple simulation with $n = 500$ and $x = (1, 1, 1)$.

6.3. Real Datasets:

In this section, our focal point is to examine two real datasets Models, namely the Insurance Company Benchmark (COIL 2000) dataset as well as the French Hospital Data of COVID-19.

6.3.1. Application 1: The Insurance Company Benchmark (COIL 2000) dataset The (COIL 2000) dataset is found in data.world website
https://data.world/uci/insurance-company-benchmark-coil-2000).
Information about customers consists of 86 variables and includes product usage data and sociodemographic data derived from zip area codes. The data are supplied by the Dutch data mining company Sentient Machine Research and rest on a real world business problem. The training set involves over 5000 descriptions of customers, including the information of whether or not they have a caravan insurance policy. A test set includes 4000 customers whom only the organizers know if they have a caravan insurance policy. This corresponds to a Dataset to train and validate prediction models and build up a description (5822 customer records). Each record consists of 86 attributes, incorporating sociodemographic data (attribute 1-43) and product ownership (attributes 44-86).
The sociodemographic data are derived from zip codes. All customers living in areas with the same zip code have the same sociodemographic attributes.
As far as our application is concerned, we shall consider the following two models:

- Model 1: X corresponds to the sociodemographic data attribute number 16 and Y stands for the whole 5822 observations of customer records.
- Model 2: X corresponds to the sociodemographic 5-dimensional data attributes number 6,8,11,12 and 13 and Y stands for the whole 5822 observations of customer records.

		x=0		x=1	
		Nadaraya's estimator	Recursive estimator	Nadaraya's estimator	Recursive estimator
Model 1	MSE	0.006979541	0.004748803	0.001999743	0.001390595
	Cor	0.987112335	0.990653245	0.995770041	0.996574025

		x=(0,0,0,0,0)		x=(1,1,1,2,2)	
		Nadaraya's estimator	Recursive estimator	Nadaraya's estimator	Recursive estimator
Model 2	MSE	0.000708417	0.0005969373	0.00414241	0.003705582
	Cor	0.996664739	0.9979940979	0.97540614	0.976614363

Table 5. Quantitative comparison between Nadaraya-Watson estimator and the proposed estimator with stepsizes $(\gamma_n) = (n^{-1})$ through a plug-in method for the Insurance Company Benchmark (COIL 2000) dataset case.

6.3.2. Application 2: French Hospital Data of COVID19
The French Hospital data of the COVID-19 epidemic are extracted from
https://www.data.gouv.fr/fr/datasets/donnees-hospitalieres-relatives-a-lepidemie-de-covid-19/.
The **Santé publique France**'s mission is to improve and protect the health of populations. During the health crisis linked to the COVID-19 epidemic, **Santé publique France** has taken in charge monitoring and understanding the dynamics of the epidemic, anticipating the different scenarios and implementing actions to prevent and limit the transmission of this virus on the national territory.

Description of the dataset
This dataset provides information on the hospital situation regarding the COVID-19 epidemic. We have opted for the first proposed file:

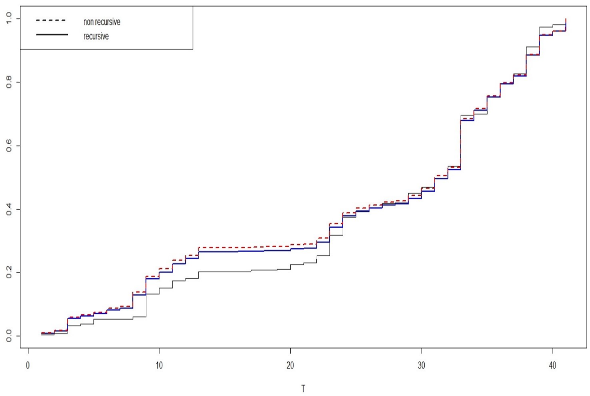

Figure 11. Qualitative comparison between the recursive estimator and Nadara Watson's non-recursive one for the dataset Model 1 with $x = 1$.

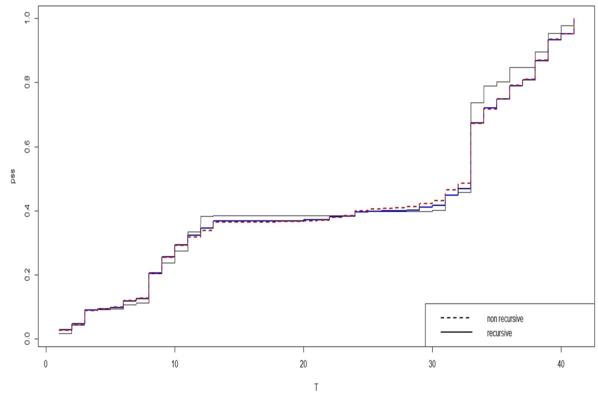

Figure 12. Qualitative comparison between the recursive estimator and the non-recursive one for the dataset Model 2 with $x = (0, 0, 0, 0, 0)$.

Hospital data related to the COVID-19 epidemic by department (dep), sex of the patient (sex), number of hospitalized patients (hosp), number of persons currently in intensive care or resuscitation (rea), number of persons currently in follow-up and rehabilitation care (SSR) or long-term care units (USLD), number of persons currently in conventional hospitalization (HospConv), number of persons currently hospitalized in another type of service (autres), cumulative number of persons returning home (rad) or cumulative number of dead persons (dc).
The data have been daily updated. For the current application, we considered the data of $28/07/2021$, with a total of 150894 observations. For simplicity reason, we have chosen to just study the department of 'Vienne' database. Therefore, for our application, we served of a dataframe of 1494 observations and 6 variables. Hence, we shall consider the following three models:

- Model 1: $X = dc$ and $Y = hosp$.
- Model 2: $X1 = sex$, $X2 = rea$, $X3 = dc$ and $Y = hosp$.
- Model 3: $X1 = sex$, $X2 = rea$, $X3 = dc$, $Y1 = hosp$ and $Y2 = rad$.

		x=119		x=17	
		Nadaraya's estimator	Recursive estimator	Nadaraya's estimator	Recursive estimator
Model 1	MSE	0.02538461	0.01900265	0.008079359	0.007509241
	Cor	0.86514927	0.88894736	0.781752320	0.789593904

		$x = (2, 0, 17)$		$x = (1, 0, 0)$	
		Nadaraya's estimator	Recursive estimator	Nadaraya's estimator	Recursive estimator
Model 2	MSE	0.02176777	0.01416798	0.01026319	0.005681196
	Cor	0.55644221	0.63197054	0.74621610	0.819493783
Model 3	MSE	0.02107916	0.009991914	0.04130757	0.03504694
	Cor	0.85362436	0.920385555	0.61896604	0.64659332

Table 6. Quantitative comparison between Nadaraya-Watson estimator and the proposed estimator with stepsizes $(\gamma_n) = (n^{-1})$ through a plug-in method for the COVID-19 epidemic dataset case.

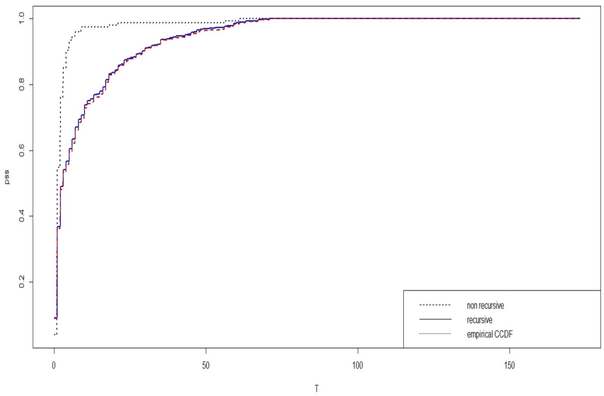

Figure 13. Qualitative comparison between the recursive estimator and the non-recursive one for the COVID-19 epidemic dataset Model 1 with $x = 17$.

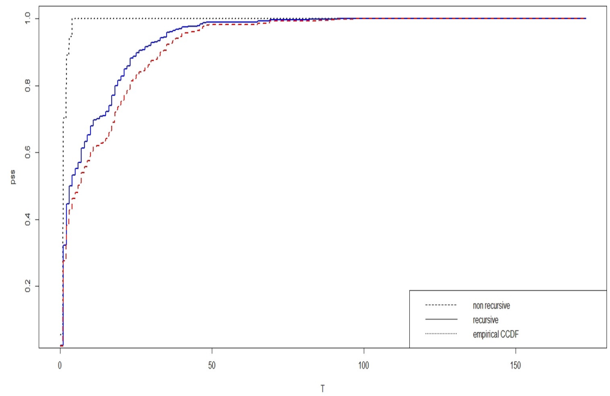

Figure 14. Qualitative comparison between the recursive estimator and the non-recursive one for the COVID-19 dataset Model 2 with $x = (2, 0, 17)$.

Data interpretation:

Referring to Tables 5, 6 and Figures 11, 12, 13 and 14 , we conclude that:

1. For all considered Models, the proposed recursive estimator with stepsize $(\gamma_n) = (n^{-1})$ through a plug-in method outperformed the non-recursive one in terms of estimation error MSE and Cor.

2. The proposed recursive estimator is closer to the true CCDF function, compared with Nadaraya-Watson's non-recursive estimator.

Concerning the COVID-19 epidemic Model 1, we can infer that, for a fixed number of deaths $x = 17$, the proportion of hospitalized cases less then 20 is 50% and the proportion of hospitalized cases less then 50 is 85%. Moreover, 99% of the population have a number of hospitalized cases less than 100.
Likewise, the COVID-19 epidemic Model 2 yielded the same results as model 1. Indeed, we recorded for the women gender a fixed value of sex $x_1 = 2$, a fixed number of REA persons $x_2 = 0$ and deaths $x_3 = 17$.

7. Conclusion

In this work, we elaborated a multivariate recursive CCDF estimator. We tackled the asymptotic properties of the proposed estimator by providing the bias as well as the variance in order to demonstrate that our estimator asymptotically follows a normal distribution. Subsequently, we revealed that the use of our recursive estimator with an appropriate choice of the bandwidth and the stepsize enables us to get closer to the true conditional cumulative distribution function rather than non-recursive one. The basic merit of recursive estimators resides in the fact that one can update the estimation with each additional new observation. Therefore, instead of re-running the data each time, it is possible to rewrite our considered estimator as a combination of two (or more) estimators, where each estimator is based on separate datasets, which can be very interesting to keep the computational cost reasonably low. It is noteworthy that all computation and simulation have been done using the **R** statistical software.
A future research direction would be to extend our findings to the setting of serially dependent observations, α-mixing framework like in [20]. Another outstanding direction lies in recursively estimating the modal regression, which requires non trivial mathematics. This would go well beyond the scope of the present paper.

Acknowledgement

This work benefited from the financial support of the PHC-Utique project no.17G1505 as well as the project CPER E-Data funded by the Nouvelle-Aquitaine Region. The Region and the European Union are also supporting this project within the framework of the Programme Opérationnel FEDER/FSE 2014-2020.

References

1. F. A. Al-Awadhi, Z. Kaid, A. Laksaci, I. Ouassou, and M. Rachdi, *Functional data analysis: local linear estimation of the L_1-conditional quantiles*, Statistical Methods & Applications. Journal of the Italian Statistical Society, vol. 28, no. 2, pp. 217–240, 2019.
2. I. M. Almanjahie, Z. Chikr Elmezouar, A. Laksaci, and M. Rachdi, *kNN local linear estimation of the conditional cumulative distribution function: Dependent functional data case*, Comptes Rendus Mathématique. Académie des Sciences. Paris, vol. 356, no. 10, pp. 1036–1039, 2018.
3. N. Altman and C. Léger, *Bandwidth selection for kernel distribution function estimation*, Journal of Statistical Planning and Inference, vol. 46, no. 2, pp. 195–214, 1995.
4. F. Benziadi, A. Laksaci, and F. Tebboune, *Recursive kernel estimate of the conditional quantile for functional ergodic data*, Communications in Statistics. Theory and Methods, vol. 45, no. 11, pp. 3097–3113, 2016.
5. A. Berlinet, A. Gannoun, and E. Matzner-Lober, *Propriétés asymptotiques d'estimateurs convergents des quantiles conditionnels*, Comptes Rendus de l'Académie des Sciences. Série I. Mathématique, vol. 326, no. 5, pp. 611–614, 1998.
6. A. Berlinet, A. Gannoun, and E. Matzner-Lø ber, *Normalité asymptotique d'estimateurs convergents du mode conditionnel*, The Canadian Journal of Statistics, vol. 26, no. 2, pp. 365–380, 1998.
7. J. R. Blum, *Multidimensional stochastic approximation methods*, Annals of Mathematical Statistics, vol. 25, pp. 737–744, 1954.
8. O. Bouanani, S. Rahmani, and L. Ait-Hennani, *Local linear conditional cumulative distribution function with mixing data*, Arabian Journal of Mathematics, vol. 9, no. 2, pp. 289–307, 2020.
9. E. Brunel, F. Comte, and C. Lacour, *Minimax estimation of the conditional cumulative distribution function*, Sankhya A. The Indian Journal of Statistics, vol. 72, no. 2, pp. 293–330, 2010.
10. Z. Chikr-Elmezouar, I. M. Almanjahie, A. Laksaci, and M. Rachdi, *FDA: strong consistency of the kNN local linear estimation of the functional conditional density and mode*, Journal of Nonparametric Statistics, vol. 31, no. 1, pp. 175–195, 2019.
11. P. Chilinski and R. Silva, *Neural likelihoods via cumulative distribution function*, Proceedings of Machine Learning Research, vol. 124, pp. 420–429, 2020.
12. M. Chowdhury, C. Wu, and R. Modarres, *Nonparametric estimation of conditional distribution functions with longitudinal data and time-varying parametric models*, Metrika. International Journal for Theoretical and Applied Statistics, vol. 81, no. 1, pp. 61–83, 2018.
13. M. Duflo, *Random iterative models*, vol. 34 of *Applications of Mathematics (New York)*. Springer-Verlag, Berlin, 1997. Translated from the 1990 French original by Stephen S. Wilson and revised by the author.
14. J. Fan, Q. Yao, and H. Tong, *Estimation of conditional densities and sensitivity measures in nonlinear dynamical systems*, Biometrika, vol. 83, no. 1, pp. 189–206, 1996.
15. J. Galambos and E. Seneta, *Regularly varying sequences*, Proceedings of the American Mathematical Society, vol. 41, pp. 110–116, 1973.
16. A. Gannoun, S. Girard, C. Guinot, and J. Saracco, *Reference curves based on nonparametric quantile regression*, Statistics in Medicine, vol. 21, no. 9, pp. 3119–3135, 2002.
17. P. Hall, R. C. L. Wolff, and Q. Yao, *Methods for estimating a conditional distribution function*, Journal of the American Statistical Association, vol. 94, no. 445, pp. 154–163, 1999.
18. T. Honda, *Nonparametric estimation of the conditional median function for long-range dependent processes*, Journal of the Japan Statistical Society (Nihon Tôkei Gakkai Kaihô), vol. 30, no. 2, pp. 129–142, 2000.
19. T. Honda, *Nonparametric estimation of a conditional quantile for α-mixing processes*, Annals of the Institute of Statistical Mathematics, vol. 52, no. 3, pp. 459–470, 2000.
20. S. Khardani and Y. Slaoui, *Recursive kernel density estimation and optimal bandwidth selection under alpha-mixing data*, Journal of Statistical Theory and Practice, vol. 13, no. 36, 2019.
21. J. Kiefer and J. Wolfowitz, *Stochastic estimation of the maximum of a regression function*, Annals of Mathematical Statistics, vol. 23, pp. 462–466, 1952.
22. S. Kiwitt and N. Neumeyer, *Estimating the conditional error distribution in non-parametric regression*, Scandinavian Journal of Statistics. Theory and Applications, vol. 39, no. 2, pp. 259–281, 2012.
23. A. Laksaci and F. Maref, *Conditional cumulative distribution estimation and its applications*, JPSS. Journal of Probability and Statistical Science, vol. 7, no. 1, pp. 57–69, 2009.
24. A. Laksaci and N. Hachemi, *Note on the functional linear estimate of conditional cumulative distribution function*, JPSS. Journal of Probability and Statistical Science, vol. 10, no. 2, pp. 153–160, 2012.
25. Q. Li, J. Lin, and J. S. Racine, *Optimal bandwidth selection for nonparametric conditional distribution and quantile functions*, Journal of Business & Economic Statistics, vol. 31, no. 1, pp. 57–65, 2013.
26. A. Mokkadem, M. Pelletier, and Y. Slaoui, *The stochastic approximation method for the estimation of a multivariate probability density*, Journal of Statistical Planning and Inference, vol. 139, no. 7, pp. 2459–2478, 2009.
27. A. Mokkadem, M. Pelletier, and Y. Slaoui, *Revisiting Révész's stochastic approximation method for the estimation of a regression function*, ALEA. Latin American Journal of Probability and Mathematical Statistics, vol. 6, pp. 63–114, 2009.

28. A. Muller-Gueudin, S. Ferrigno, and M. Maumy-Bertrand, *Certainty bands for the conditional cumulative distribution function and applications*, 47èmes Journées de Statistique de la SFdS, hal-01154624, June, 2015.
29. E. A. Nadaraya, *On estimating regression*, Theory of Probability and Its Applications, vol. 9, pp. 141–142, 1964.
30. E. Parzen, *On estimation of a probability density function and mode*, Ann. Math. Statist., vol. 33, pp. 1065–1076, 09 1962.
31. S. Plancade, *Adaptive estimation of the conditional cumulative distribution function from current status data*, Journal of Statistical Planning and Inference, vol. 143, no. 9, pp. 1466–1485, 2013.
32. P. Révész, *How to apply the method of stochastic approximation in the non-parametric estimation of a regression function*, Mathematische Operationsforschung und Statistik Series Statistics, vol. 8, no. 1, pp. 119–126, 1977.
33. H. Robbins and S. Monro, *A stochastic approximation method*, Annals of Mathematical Statistics, vol. 22, pp. 400–407, 1951.
34. M. Rosenblatt, *Remarks on some nonparametric estimates of a density function*, Ann. Math. Statist., vol. 27, pp. 832–837, 09 1956.
35. D. Ruppert, *Almost sure approximations to the Robbins-Monro and Kiefer-Wolfowitz processes with dependent noise*, The Annals of Probability, vol. 10, no. 1, pp. 178–187, 1982.
36. S. Slama and Y. Slaoui, *Multivariate distribution function estimation using stochastic approximation method*, International Journal of Mathematics and Statistics, vol. 22, no. 2, pp. 31–59, 2021.
37. Y. Slaoui, *Bandwidth selection for recursive kernel density estimators defined by stochastic approximation method*, Journal of Probability and Statistics, pp. Art. ID 739640, 11, 2014.
38. Y. Slaoui, *The stochastic approximation method for estimation of a distribution function*, Mathematical Methods of Statistics, vol. 23, no. 4, pp. 306–325, 2014.
39. Y. Slaoui, *Large and moderate deviation principles for averaged stochastic approximation method for the estimation of a regression function*, Serdica. Mathematical Journal. Serdika. Matematichesko Spisanie, vol. 41, no. 2-3, pp. 307–328, 2015.
40. Y. Slaoui and S. Khardani, *Adaptive recursive kernel conditional density estimators under censoring data*, ALEA. Latin American Journal of Probability and Mathematical Statistics, vol. 17, no. 1, pp. 389–417, 2020.
41. W. Stute, *Conditional empirical processes*, Ann. Statist., vol. 14, pp. 638–647, 06 1986.
42. A. B. Tsybakov, *Recurrent estimation of the mode of a multidimensional distribution*, Akademiya Nauk SSSR. Institut Problem Peredachi Informatsii Akademii Nauk SSSR. Problemy Peredachi Informatsii, vol. 26, no. 1, pp. 31–37, 1990.
43. N. Veraverbeke, I. Gijbels, and M. Omelka, *Preadjusted non-parametric estimation of a conditional distribution function*, Journal of the Royal Statistical Society. Series B. Statistical Methodology, vol. 76, no. 2, pp. 399–438, 2014.
44. G. S. Watson, *Smooth regression analysis*, Sankhyā (Statistics). The Indian Journal of Statistics. Series A, vol. 26, pp. 359–372, 1964.
45. K. Yu and M. C. Jones, *Local linear quantile regression*, Journal of the American Statistical Association, vol. 93, no. 441, pp. 228–237, 1998.