Journal de la Société Française de Statistique
Vol. 159 No. 1 (2018)

On the Consistency of Kernel Classification Rule for Functional Random Field

Titre: Sur la consistance de la règle de classification du noyau pour le champ aléatoire fonctionnel

Ahmad Younso[1]

Abstract: We consider the classical moving window rule of classification for functional spatially dependent data. We investigate asymptotic properties of this nonparametric classification rule based on training data drawn from α or β-mixing random field taking values in infinite-dimensional space. We extend the results of Younso (2017a) concerning both the consistency and the strong consistency of the moving window classifier to the spatially dependent case under mild assumptions. We propose a method for bandwidth selection and we conduct some simulation studies.

Résumé : Nous considérons la règle de la fenêtre mobile pour classifier des données fonctionnelles spatialement dépendantes. Nous étudions les propriétés asymptotiques de cette règle de classification non paramétrique basée sur des données d'apprentissage tirées d'un champ aléatoire α ou $\beta-$mélangeant à valeurs en espace de dimension infinie. Nous étendons les résultats de Younso (2017a) concernant la consistance et la consistance forte au cas spatialement dépendant sous des hypothèses non restrictives. Nous proposons un critère pour choisir le paramètre de lissage et nous considérons l'application de notre approche sur des données simulées.

Keywords: Bayes rule, training data, moving window rule, random field, bandwidth, consistency.
Mots-clés : Règle de Bayes, données d'apprentissage, règle de fenêtre mobile, champ aléatoire, paramètre de lissage, consistance.
AMS 2000 subject classifications: 62M30, 62G20, 62H11

1. Introduction

In many studies, the observations can be collected as spatially dependent curves. This type of data arises in a variety of fields including econometrics, epidemiology, environmental sciences, image analysis, oceanography and many others. Many spatially dependent data can be represented by finite dimensional vectors and others may be represented by curves. For general applications, we refer the reader to Ramsay and Silverman (2002, 2005). The statistical treatment for spatially dependent data has received a lot of attention in recent years in finite or infinite dimensional space. Nonparametric approaches, including classification and estimation, have recently emerged as a flexible way to model spatial data. In some studies, it can be interesting to see spatio-temporal data as spatially dependent data. The need to classify observed functional data occurs in many scientific problems. For example, in medical imaging modalities, an important problem is how to classify image pixels into spatial regions in which the pixels exhibit similar temporal behavior. In this paper, we propose a nonparametric classification rule based on kernel method for classifying spatially dependent variables taking values in infinite dimensional space. For the spatially dependent case, most of existing theoretical nonparametric results concern the density and regression estimation in

[1] Department of mathematical statistics, Damascus university, Damascus, Syria.
E-mail: ahyounso@yahoo.fr

 ISSN: 2102-6238

finite dimensional space. For background material, the reader is referred to Tran (1990), Carbon et al. (1997), Biau and Cadre (2004) and Hallin et al. (2004, 2009). Despite the wide area of application, there is only a very few literature dedicated to models that take into account both the functional and spatial dependence features, see for example Ternynck (2014) and Dabo-Niang and Yao (2007, 2013). The literature dealing with nonparametric classification (kernel rule or nearest neighbor rule) when data are independent is extensive in finite or infinite dimensional spaces, see for example Devroye and Krzyżak (1989) and Devroye et al. (1996) for the finite dimensional case and Abraham et al. (2006), Ferraty et al. (2002, 2012) and Ferraty and Vieu (2006) for the infinite dimensional case. Younso (2017a) extends the results of Abraham et al. (2006) to the temporally dependent case. In the spatially dependent case, Younso (2017b) proposes a new kernel rule allowing for the classification of missing data in a finite-dimensional space and establishes the consistency of this new rule. In the functional case, the asymptotic properties of the kernel classification rule remains unexplored. In this paper, we investigate whether the moving window rule can be generalized to classify spatial functional data showing spatial dependence.

2. Moving window rule for functional random field

Let (E,d) be a metric space where E is a function space and d is the metric on E. Denote the integer lattice points in the N-dimensional Euclidean space by $\mathbb{Z}^N$, $N \geq 1$. Consider a strictly stationary random field $\{(X_{\mathbf{i}}, Y_{\mathbf{i}})\}_{\mathbf{i}\in\mathbb{Z}^N}$ defined on some probability space $(\Omega, \mathscr{F}, \mathbb{P})$ and taking values in $E \times \{0,1\}$. A point $\mathbf{i} = (i_1,...,i_N) \in \mathbb{Z}^N$ will be referred to as a site. For $\mathbf{n} = (n_1,...,n_N) \in (\mathbb{N}^*)^N$, we define the rectangular region $I_{\mathbf{n}}$ by

$$I_{\mathbf{n}} = \{\mathbf{i} \in \mathbb{Z}^N : \ 1 \leq i_k \leq n_k, \ \forall k = 1,...,N\}.$$

We will write $\mathbf{n} \to \infty$ if

$$\min_{k=1,...,N} n_k \to \infty.$$

Define $\hat{\mathbf{n}} = n_1 \times ... \times n_N = \text{Card}(I_{\mathbf{n}})$. For the sake of simplicity, we suppose that $(X_{\mathbf{i}}, Y_{\mathbf{i}})$ has the same distribution as the pair (X,Y) for all $\mathbf{i} \in \mathbb{Z}^N$. Observe that the distribution of (X,Y) may be well defined by (μ,η) where $\mu(B) = \mathbb{P}(X \in B)$, for all Borel sets B on $\mathscr{F}$, and $\eta(x) = \mathbb{P}(Y = 1|X = x)$, for all $x \in E$. Classical procedure of classification deals with predicting the unknown nature Y called a class (0 or 1) of an observation X with values in E. The statistician creates a classifier $g : E \longrightarrow \{0,1\}$ which maps a new observation $x \in E$ into its predicted label $g(x)$. It is certainly possible to wrongly specify the label y of a new observation x and an error occurs if $g(x) \neq y$. Let $L = L(g) = \mathbb{P}\{g(X) \neq Y\}$ denote the probability of error for the classifier g. There exists an optimal classifier, called Bayes rule, given by

$$g^*(x) = \begin{cases} 0 \text{ if } \mathbb{P}\{Y = 0|X = x\} \geq \mathbb{P}\{Y = 1|X = x\} \\ 1 \text{ otherwise.} \end{cases}$$

It is easy to see that the Bayes rule has the smallest probability of error, that is

$$L^* = L(g^*) = \inf_{g:E\to\{0,1\}} \mathbb{P}\{g(X) \neq Y\},$$

 ISSN: 2102-6238

(see Theorem 2.1 in Devroye et al., 1996 for the finite dimensional case). Unfortunately, the Bayes rule depends on the distribution of (X,Y) which is generally unknown to the statistician. But it is often possible to construct a classifier from a set of observations $D_{\mathbf{n}} = \{(X_{\mathbf{i}}, Y_{\mathbf{i}}),\ \mathbf{i} \in I_{\mathbf{n}}\}$. The set $D_{\mathbf{n}}$ is called the training data. Among the various ways to define a classifier from a training data, one of the most simple and popular is the moving window rule defined by

$$g_{\mathbf{n}}(x) = \begin{cases} 0 \text{ if } \displaystyle\sum_{\mathbf{i}\in I_{\mathbf{n}}} \mathbb{1}_{\{Y_{\mathbf{i}}=0, X_{\mathbf{i}}\in B_{x,b}\}} \geq \sum_{\mathbf{i}\in I_{\mathbf{n}}} \mathbb{1}_{\{Y_{\mathbf{i}}=1, X_{\mathbf{i}}\in B_{x,b}\}} \\ \\ 1 \text{ otherwise,} \end{cases}$$

where $\mathbb{1}_A$ denotes the indicator function of the set A, $b = b(\mathbf{n})$ the bandwidth, is a strictly positive number tending to 0 when $\mathbf{n} \to \infty$ and $B_{x,b}$ denotes the closed ball centered at x with radius b. In order to establish the theoretical results, we write the moving window rule as follows

$$g_{\mathbf{n}}(x) = \begin{cases} 0 \text{ if } \eta_{\mathbf{n}}(x) \leq \dfrac{\sum_{\mathbf{i}\in I_{\mathbf{n}}} (1-Y_{\mathbf{i}})\, \mathbb{1}_{\{X_{\mathbf{i}}\in B_{x,b}\}}}{\hat{\mathbf{n}}\mu(B_{x,b})} \\ \\ 1 \text{ otherwise,} \end{cases} \tag{1}$$

where

$$\eta_{\mathbf{n}}(x) = \frac{\sum_{\mathbf{i}\in I_{\mathbf{n}}} Y_{\mathbf{i}}\, \mathbb{1}_{\{X_{\mathbf{i}}\in B_{x,b}\}}}{\hat{\mathbf{n}}\mu(B_{x,b})}.$$

Clearly, the moving window rule is one of the kernel-based rules being derived from the kernel estimate in density and regression estimation (see for example Parzen, 1962, Nadaraya, 1989 and Watson, 1964). Let $L_{\mathbf{n}} = L(g_{\mathbf{n}}) = \mathbb{P}\{g_{\mathbf{n}}(X) \neq Y | D_{\mathbf{n}}\}$ be the error probability of $g_{\mathbf{n}}$. The classifier $g_{\mathbf{n}}$ is called consistent if

$$\mathbb{E}L_{\mathbf{n}} \longrightarrow L^* \text{ as } \mathbf{n} \to \infty$$

and called strongly consistent if

$$L_{\mathbf{n}} \longrightarrow L^* \text{ with probability one as } \mathbf{n} \to \infty.$$

A classifier can be consistent for certain class of distribution of (X,Y), but not be consistent for others. The classifier $g_{\mathbf{n}}$ is called universally (strongly) consistent, if it is (strongly) consistent for all distribution of (X,Y). Much of the existing theory on the consistency problems is based on the assumption that the available functional data are independent and identically distributed. In finite dimensional spaces, the moving window rule and the k-nearest neighbor rule are universally strongly consistent under classical conditions (see Devroye and Krzyżak, 1989 and Stone, 1977). Abraham et al. (2006) give some examples showing that the results of Devroye and Krzyżak (1989) on the consistency are no more valid in a general functional metric space and they establish the consistency and the strong consistency under mild conditions on the distribution of (X,Y) and the metric space. Younso (2017a) extends the results of Abraham et al. (2006) to the temporally dependent case. Our aim in this paper is to establish the consistency and the strong consistency of the moving window rule based on spatially dependent functional training data under some α- and β-mixing conditions.

 ISSN: 2102-6238

3. Mixing conditions

Let us first recall the definitions of α-mixing coefficients introduced by Rosenblatt (1956) and β-mixing coefficient introduced by Rozanov and Volkonskii (1959). Let $\mathscr{A}$ and $\mathscr{C}$ be two sub σ-algebras of $\mathscr{F}$. The α-mixing coefficient between $\mathscr{A}$ and $\mathscr{C}$ is defined by

$$\alpha = \alpha(\mathscr{A},\mathscr{C}) = \sup_{A\in\mathscr{A},C\in\mathscr{C}} |\mathbb{P}(A\cap C) - \mathbb{P}(A)\mathbb{P}(C)|$$

and the β-mixing coefficient is defined by

$$\beta = \beta(\mathscr{A},\mathscr{C}) = \mathbb{E}\sup_{A\in\mathscr{A}} |\mathbb{P}(A|\mathscr{C}) - \mathbb{P}(A)]|.$$

Let $\{Z_{\mathbf{i}}\}_{\mathbf{i}\in\mathbb{Z}^N}$ be a random field on $(\Omega,\mathscr{F},\mathbb{P})$ and taking values in some space $(\Omega',\mathscr{F}')$. For any $S,S'\subset\mathbb{Z}^N$ with finite cardinals, we denote by $\mathscr{B}(S)$ and $\mathscr{B}(S')$ the Borel σ-algebras generated by $\{Z_{\mathbf{i}}\}_{\mathbf{i}\in S}$ and $\{Z_{\mathbf{i}}\}_{\mathbf{i}\in S'}$ respectively.

Definition 3.1 The random field $\{Z_{\mathbf{i}}\}_{\mathbf{i}\in\mathbb{Z}^N}$ is said to be α-mixing or strongly mixing if

$$\alpha(t) = \sup_{\mathrm{dist}(S,S')\geq t} \alpha\big(\mathscr{B}(S),\mathscr{B}(S')\big) \downarrow 0 \text{ as } t\to\infty, \tag{2}$$

where

$$\mathrm{dist}(S,S') = \inf_{\mathbf{i}\in S,\mathbf{j}\in S'} \|\mathbf{i}-\mathbf{j}\|$$

and $\|.\|$ denotes the Euclidean norm.
Observe that α-mixing condition (2) is satisfied by many spatial models. Examples can be found in Neaderhouser (1980) and Rosenblatt (1985).

Definition 3.2 The random field $\{Z_{\mathbf{i}}\}_{\mathbf{i}\in\mathbb{Z}^N}$ is said to be β-mixing or absolutely regular if

$$\beta(t) = \sup_{\mathrm{dist}(S,S')\geq t} \beta\big(\mathscr{B}(S),\mathscr{B}(S')\big) \downarrow 0 \text{ as } t\to\infty.$$

The two mixing coefficients α and β are related by the inequality $2\alpha\leq\beta$ (see Rio, 2000). Consequently, any β-mixing random field is an α-mixing one. The following lemma, given in Rio (2000), is crucial in order to derive the consistency of the moving window rule.

Lemma 3.1. *Let Z_1 and Z_2 be two $\mathbb{R}$-valued bounded random variables. Then, we have*

$$|\mathrm{cov}(Z_1,Z_2)| \leq 4\|Z_1\|_\infty\|Z_2\|_\infty\alpha(\sigma(Z_1),\sigma(Z_2)),$$

where $\|.\|_\infty$ is the supremum norm and $\sigma(Z_i)$ is the σ-algebra generated by Z_i for $i=1,2$.

Now, let $\mathscr{A}$ and $\mathscr{C}$ be two sub σ-algebras of $\mathscr{F}$, we denote by $\mathscr{A}\vee\mathscr{C}$ the σ-algebra generated by $\mathscr{A}\cup\mathscr{C}$. The following coupling lemma (see Berbee, 1979) will be needed to establish the strong consistency.

Lemma 3.2. *Let Z be a random variable defined on $(\Omega,\mathscr{F},\mathbb{P})$ and taking values in some Polish space Ω' and $\mathscr{M}$ be a sub σ-algebra of $\mathscr{F}$. Assume that there exists a random variable U uniformly distributed over $[0,1]$, independent of $\sigma(Z)\vee\mathscr{M}$. Then, there exists a random variable Z^* measurable with respect to $\sigma(U)\vee\sigma(Z)\vee\mathscr{M}$, distributed as Z and independent of $\mathscr{M}$, such that*

$$\mathbb{P}(Z\neq Z^*) = \beta(\mathscr{M},\sigma(Z)).$$

 ISSN: 2102-6238

Remark 3.1 A Polish space Ω' is a topological space which is separable and completely metrizable (see Kechris, 1995). Most of the familiar objects of study in analysis involve Polish spaces. For example, $\mathbb{R}$ and $\mathbb{R}^p$ with the usual topology are Polish. For all $m \in \mathbb{N}^*$, $\{0,1,...,m-1\}$ is Polish with discrete topology. A countable product of Polish spaces is Polish, too.

4. Assumptions and preliminaries

For convenience, we firstly introduce the notion of covering numbers (see Kolmogorov and Tihomirov, 1961). For a given subset G of the metric space (E,d) , the covering number is defined by

$$\mathcal{N}(\varepsilon,G,d) = \inf\left\{k \geq 1 :\ \exists x_1,...,x_k \in E \text{ with } G \subset \bigcup_{i=1}^{k} S_{x_i,\varepsilon}\right\},$$

where $S_{x,\varepsilon}$ denotes the open ball centered at x with radius $\varepsilon > 0$. The set G is said to be totally bounded if $\mathcal{N}(\varepsilon,G,d) < \infty$ for all $\varepsilon > 0$. In particular, every relatively compact set is totally bounded and all totally bounded sets are bounded. Now, we introduced some assumptions.

Assumption 1 There exists a sequence $(E_k)_{k\geq 1}$ of totally bounded subsets of E such that $E_k \subset E_{k+1}$ for all $k \geq 1$ and $\mu\left(\bigcup_{k\geq 1} E_k\right) = 1$.

Assumption 2 For each integer $k \geq 1$, any $\mathbf{i} \neq \mathbf{j}$ and $\varepsilon_1 \in]0,1]$, there exists $C > 0$ such that $\mathbb{P}((X_\mathbf{i},X_\mathbf{j}) \in B_{x,b} \times B_{x,b}) \leq C[\mu(B_{x,b})]^{1+\varepsilon_1}$, for all $x \in E_k$.

Assumption 3 The following Besicovich condition holds, for every $\varepsilon > 0$,

$$\lim_{b\to 0+} \mu\left\{x \in E : \left|\frac{1}{\mu(B_{x,b})}\int_{B_{x,b}} \eta d\mu - \eta(x)\right| > \varepsilon\right\} = 0.$$

Remark 4.1 Note that Assumption 1 is always true whenever the space (E,d) is separable, see for example Abraham et al. (2006) and Kulkarni and Posner (1995). Assumption 2, used by Ternynck (2014), concerns the local dependency and a consequence is

$$|\mathbb{P}((X_\mathbf{i},X_\mathbf{j}) \in B_{x,b} \times B_{x,b}) - \mathbb{P}(X_\mathbf{i} \in B_{x,b})\mathbb{P}(X_\mathbf{j} \in B_{x,b})| \leq C',$$

for $C' = C + 1$. As noticed in Dabo-Niang et al. (2011), Assumption 2 can be linked with the classical local dependence condition met in the literature of the finite-dimensional case when X and $(X_\mathbf{i},X_\mathbf{j})$ admit, respectively, the densities f and $f_{\mathbf{i},\mathbf{j}}$ (see Tran, 1990). In the case $N = 1$, Assumption 2 has been used by (Bosq, 1998, page 54). Assumption 3 holds for example if $\eta(x)$ is μ-continuous (see Cérou and Guyader, 2006).
Now, we suppose that the random field $\{(X_\mathbf{i},Y_\mathbf{i})\}_{\mathbf{i}\in\mathbb{Z}^N}$ is arithmetically α-mixing in the sense that there exist $C > 0$ and $\theta > 0$ such that

$$\alpha(t) \leq Ct^{-\theta} \text{ for all } t \in \mathbb{R}_+^*. \tag{3}$$

From now on, G^c stands for the complement of any subset G of E and for simplicity of notation, we write $\mathcal{N}_k(\varepsilon)$ instead of $\mathcal{N}(\varepsilon,E_k,d)$. Before we state the main results, we introduce some lemmas that will be needed in the sequel. The following lemma is a direct consequence of Assumption 3.

 ISSN: 2102-6238

Lemma 4.1. *Assume that Assumption 3 holds. If $b \to 0$ as $\mathbf{n} \to \infty$, then,*

$$\int_E |\eta(x) - \mathbb{E}\eta_{\mathbf{n}}(x)|\mu(dx) = \int_E \left| \eta(x) - \frac{\int_{B_{x,b}} \eta(t)\mu(dt)}{\mu(B_{x,b})} \right| \mu(dx) \longrightarrow 0$$

as $\mathbf{n} \to \infty$.

For the proof of the following lemma, we refer to Abraham et al. (2006).

Lemma 4.2. *Assume that $(E_k)_{k\geq 1}$ is a sequence of totally bounded subsets of E. Let k be a fixed positive integer. Then, for every $b > 0$,*

$$\int_{E_k} \frac{1}{\mu(B_{x,b})} \mu(dx) \leq \mathscr{N}_k(b/2).$$

Lemma 4.3. *Let $(E_k)_{k\geq 1}$ be a sequence of totally bounded subsets of E. Assume that the training data $D_{\mathbf{n}}$ are observations of α-mixing functional random field such that (3) and that Assumption 2 is satisfied. Let k be a fixed positive integer. Then, for all $\mathbf{n} \in (\mathbb{N}^*)^N$,*

$$\mathbb{E}\int_{E_k} |\eta_{\mathbf{n}}(x) - \mathbb{E}\eta_{\mathbf{n}}(x)|\mu(dx) \leq C \left(\frac{1}{\hat{\mathbf{n}}} \mathscr{N}_k\left(\frac{b}{2}\right) \right)^{1/2}, \text{ for some } C > 0.$$

5. Main results

In this section, we establish the consistency and the strong consistency of the moving window rule.

Theorem 5.1 (Consistency)**.** *Let $(E_k)_{k\geq 1}$ be a sequence of totally bounded subsets of E. Assume that the training data $D_{\mathbf{n}}$ are observations of α-mixing functional random field such that (3) and that Assumption 1-3 hold. If $b \to 0$ and for every $k \geq 1$, $\frac{\mathscr{N}_k(b/2)}{\hat{\mathbf{n}}} \longrightarrow 0$ as $\mathbf{n} \to \infty$, then, for $\theta > 2N$,*

$$\mathbb{E}L_{\mathbf{n}} \longrightarrow L^* \text{ as } \mathbf{n} \to \infty.$$

where θ is the constant defined in (3).

Observe that, for $N = 1$, the same assumptions on the smoothing factor b are used by Abraham et al. (2006) and Younso (2017a) for the independent and the dependent case respectively.
Now, we investigate the strong consistency of the moving window classifier under β-mixing condition. This mixing condition together with the coupling Lemma 3.2 allow to generate independent and identically distributed random functional variables that we need to prove the strong consistency, while the more general mixing condition, the α-mixing condition, allows only to generate independent and identically distributed real-valued random variables (see Bradley, 1983). In order to establish the strong consistency, we suppose that $n_1 = n_2 = ... = n_N = n$. It means that if $n \to \infty$, the rectangular region $I_{\mathbf{n}}$ expands to infinity at the same rate along all directions. This isotropic assumption is used by El-Machkouri (2007). For the sake of simplicity, we will write

$$I_n = \{\mathbf{i} \in \mathbb{Z}^N : \ 1 \leq i_k \leq n, \ \forall k = 1, ..., N\}, \ \hat{\mathbf{n}} = \text{Card}(I_n) = n^N$$

 ISSN: 2102-6238

and

$$g_n(x) = \begin{cases} 0 \text{ if } \eta_n(x) \leq \frac{\sum_{\mathbf{i}\in I_n}(1-Y_{\mathbf{i}})\mathbb{1}_{\{X_{\mathbf{i}}\in B_{x,b}\}}}{n^N\mu(B_{x,b})} \\ \\ 1 \text{ otherwise,} \end{cases}$$

where

$$\eta_n(x) = \frac{\sum_{\mathbf{i}\in I_n} Y_{\mathbf{i}}\mathbb{1}_{\{X_{\mathbf{i}}\in B_{x,b}\}}}{n^N\mu(B_{x,b})},$$

$L_n = \mathbb{P}(g_n(X) \neq Y|D_n)$ and $b = b(n)$. Furthermore, the limit $\mathbf{n} \to \infty$ will be replaced by the limit $n \to \infty$. Before we formulate the result on the strong consistency, we suppose that the random field $\{(X_{\mathbf{i}}, Y_{\mathbf{i}})\}_{\mathbf{i}\in\mathbb{Z}^N}$ is arithmetically β-mixing in the sense that there exist $C_1 > 0$ and $\theta_1 > 0$ such that

$$\beta(t) \leq C_1 t^{-\theta_1} \text{ for all } t \in \mathbb{N}^*. \tag{4}$$

The following theorem generalizes the strong consistency result of Younso (2017a) to the spatial case.

Theorem 5.2 (Strong consistency)**.** *Let $(E_k)_{k\geq 1}$ be a sequence of totally bounded subsets of E. Assume that the training data D_n are observations of β-mixing functional random field such that (4) with $\theta_1 > 2N$ and that the metric space (E,d) is Polish. Assume also that Assumption 1-3 hold. Let $(k_n)_{n\geq 1}$ be an increasing sequence of positive integers such that*

$$\sum_{n\geq 1}\mu(E^c_{k_n}) < \infty \text{ and } \sum_{n\geq 1}\mathscr{N}_{k_n}\left(\frac{b}{2}\right)p_n^{-\theta_1} < \infty,$$

for some integer $p = p_n \in [1, n/2]$ with $p_n \to \infty$ as $n \to \infty$. If $b \to 0$ and

$$\frac{n^N}{p^N \log(n)\mathscr{N}^2_{k_n}(b/2)} \longrightarrow \infty \text{ as } n \to \infty,$$

then,

$$L_n \longrightarrow L^* \text{ as } n \to \infty \text{ with probability one.}$$

Remark 5.1 Observe that for $N = 1$, the assumptions on b are used by Younso (2017a) to obtain the strong consistency in the temporal case (see also Abraham et al. (2006) for similar assumptions in the independent case). Furthermore, consider $\mathscr{N}_{k_n}(b/2) \simeq n^{\gamma_1 N}$ with $0 < \gamma_1 < 1$, and choose $p_n \simeq n^{\gamma_2}$ with $(1+\gamma_1)N/\theta_1 < \gamma_2 < 1$ and $\theta_1 > 2N$. Clearly, we have

$$\sum_{n\geq 1}\mathscr{N}_{k_n}\left(\frac{b}{2}\right)p_n^{-\theta_1} < \infty.$$

The condition

$$\frac{n^N}{p^N \log(n)\mathscr{N}^2_{k_n}(b/2)} \longrightarrow \infty$$

may be satisfied if $\gamma_2 + 2\gamma_1 < 1$. The condition $\sum_{n\geq 1}\mu(E^c_{k_n}) < \infty$, used by Abraham et al. (2006) and Younso (2017a), is classical for this type of results.

 ISSN: 2102-6238

6. Smoothing factor selection and simulation study

In practice, the choice of a smoothing parameter b is a crucial problem to the kernel classifier. A wrong value of b may lead to catastrophic error rates. In principle, there is no universal criterion that would enable an optimal choice. Various techniques for the smoothing factor selection have been developed in the nonparametric kernel smoothing method. Among the different selection techniques to select the parameter b, one can propose the cross-validation criterion (CV). This technique, being widely used in statistics, is primarily a way of measuring the predictive performance of a statistical model. In the nonparametric functional regression, the (CV) criterion is implemented in R programming environment (see Febrero-Bande and de la Fuente, 2012), but the situation is slightly different for the nonparametric classification problem. However, taking

$$g_n(x) = \begin{cases} 0 \text{ if } \sum_{\mathbf{i}\in I_n} Y_{\mathbf{i}} \mathbb{1}_{\{d(X_{\mathbf{i}},x)\leq b\}} \leq \sum_{\mathbf{i}\in I_n} (1-Y_{\mathbf{i}}) \mathbb{1}_{\{d(X_{\mathbf{i}},x)\leq b\}} \\ \\ 1 \text{ otherwise,} \end{cases}$$

the (CV) criterion is based on minimizing, with respect to $b = b(n) \in \mathbb{R}_+$, the $CV(b)$ given by

$$CV(b) = \frac{1}{n^N} \sum_{\mathbf{i}\in I_n} (Y_{\mathbf{i}} - g_n^{-\mathbf{i}}(X_{\mathbf{i}}))^2 \omega(X_{\mathbf{i}}),$$

where $g_n^{-\mathbf{i}}(X_{\mathbf{i}})$ indicates the moving window rule based on leaving out the pair $(X_{\mathbf{i}}, Y_{\mathbf{i}})$ and $\omega(X_{\mathbf{i}})$ is the weight of the element $X_{\mathbf{i}}$. We assume that b belongs to some set $H_n \subset \mathbb{R}_+$ including $b_1^{\mathbf{i}}, \ldots, b_k^{\mathbf{i}}$ for all $\mathbf{i} \in I_n$ where $b_j^{\mathbf{i}}$ is the distance to the j^{th} neighbor of $X_{\mathbf{i}}$ with respect to the metric d and k is chosen depending on the size of training data set. The weight function $\omega(x)$ may be chosen as a bounded function with support on a bounded compact set having non-empty interior (see Rachdi and Vieu, 2007). For the sake of simplicity, we will take $\omega(x)$ as a constant. Therefore, the cross-validated smoothing factor is given by

$$b_{opt} = \arg\min_{b\in H_n} CV(b).$$

Now, we use the R statistical programming environment to run a simulation study for $N = 2$. We propose to investigate the performance of our method in the following simulated scenario. For each $\mathbf{i} \in I_n$ and $t \in [1,21]$, we generate pairs $(X_{\mathbf{i}}(t), Y_{\mathbf{i}})$ via the following model inspired by Ferraty and Vieu (2003) and Preda (2007) (see also Jacques and Preda, 2014):

$$\begin{aligned} &\text{Class}(Y_{\mathbf{i}} = 0) : X_{\mathbf{i}}(t) = U_{\mathbf{i}} h_1(t) + \varepsilon_{\mathbf{i}}(t) \\ &\text{Class}(Y_{\mathbf{i}} = 1) : X_{\mathbf{i}}(t) = U_{\mathbf{i}} h_1(t) + V_{\mathbf{i}} h_2(t) + \varepsilon_{\mathbf{i}}(t) \end{aligned}$$

where $U_{\mathbf{i}}$ and $V_{\mathbf{i}}$ are independent Gaussian variables such that $\mathbb{E}(U_{\mathbf{i}}) = \mathbb{E}(V_{\mathbf{i}}) = 0$, $\text{Var}(U_{\mathbf{i}}) = 1/2, \text{Var}(V_{\mathbf{i}}) = 1/12$ and $\varepsilon_{\mathbf{i}}(t)$ are dependent normal random variables with mean 0, variance 2 and covariance function $c(\|u\|) = 2\|u\|^{-5}$ for all $u \in \mathbb{R}^2$ with $u \neq 0$. It is also supposed that $\varepsilon_{\mathbf{i}}(t)$ are independent of both $U_{\mathbf{i}}$ and $V_{\mathbf{i}}$. The functions h_1 and h_2 (plotted on Figure 1) are defined for $t \in [1,21]$, by $h_1(t) = max(6 - |t-7|, 0)$ and $h_2(t) = max(6 - |t-15|, 0)$.

 ISSN: 2102-6238

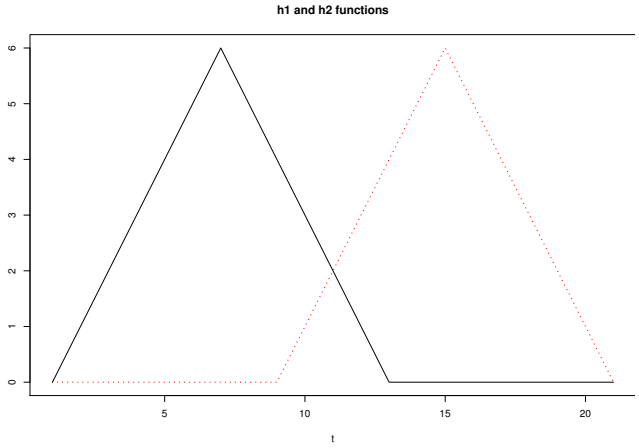

FIGURE 1. *Plots of the function $h_1(t)$ (solid line) and the function $h_2(t)$ (dashed line).*

It is important to mention that $\{\varepsilon_{\mathbf{i}}(t)\}_{\mathbf{i}\in I_n}$ are observations of an α-mixing random field since any Gaussian random field with covariance function $c(\|u\|)$ converges to zero as $\|u\| \to \infty$ is α-mixing. We suppose that the function space on the interval $[1,21]$ is endowed with the metric d (between x_1 and x_2) defined by $d(x_1,x_2) = \int_1^{21} |x_1(t) - x_2(t)|dt$. This metric is used without discretizing the data. The curve $X_{\mathbf{i}}(t)$ will be colored by *black* if $Y_{\mathbf{i}} = 1$ and by *red* if $Y_{\mathbf{i}} = 0$. Figure 2 displays two realizations of the $X_{\mathbf{i}}$'s and Figure 3 displays a sample of size $n^2 = 625$ observed on the two-dimensional grid $I_{25} = \{(i,j),\ 1 \leq i,j \leq 25\}$ (plotted on Figure 4).

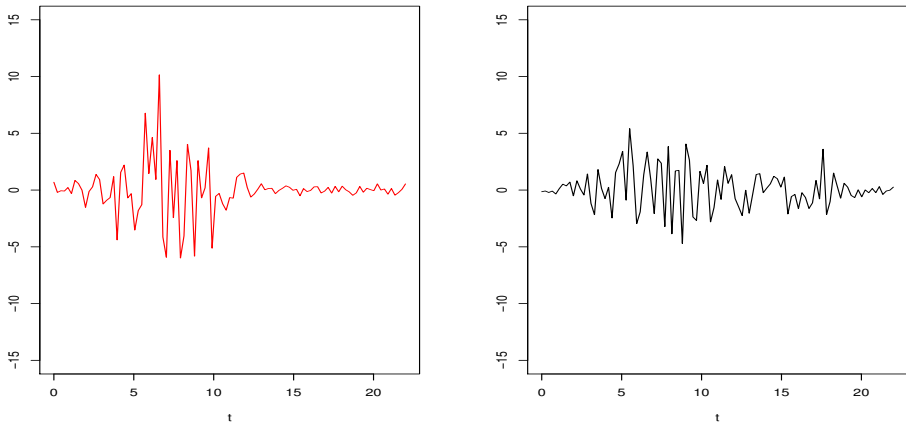

FIGURE 2. *Two realizations of simulated curves with label* 0 *(left) and label* 1 *(right).*

 ISSN: 2102-6238

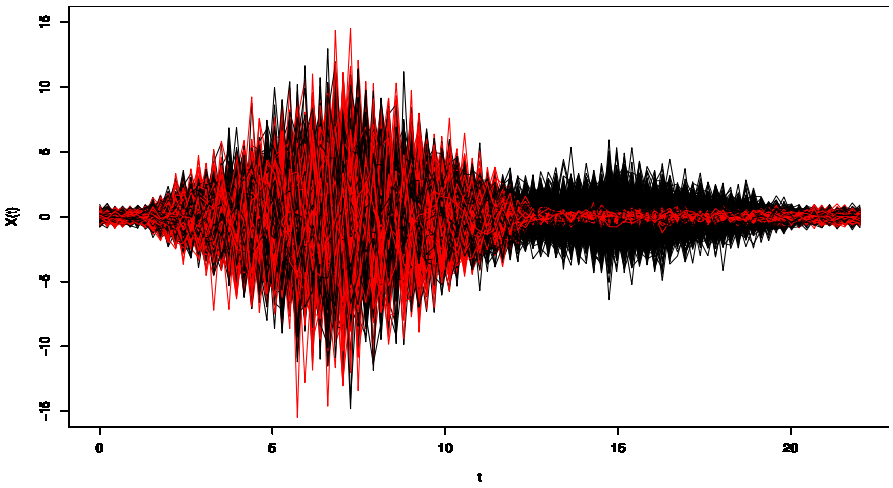

FIGURE 3. *Sample of 625 observed curves on the region* I_{25}.

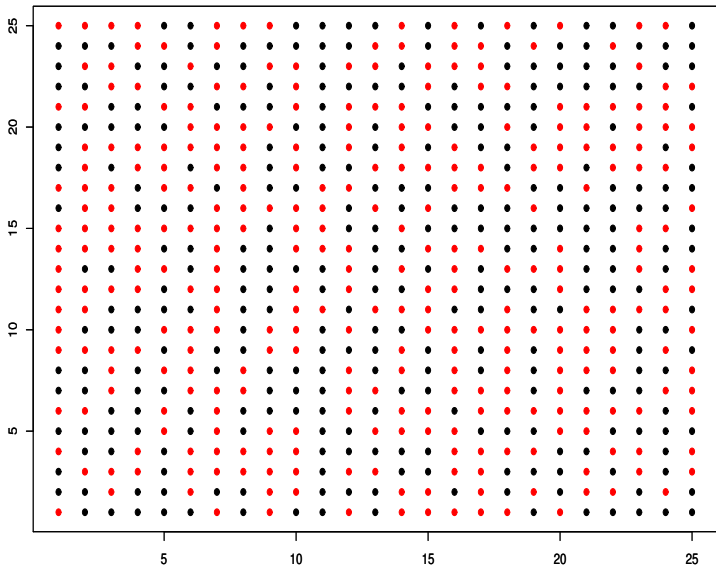

FIGURE 4. *The region* I_{25} *where the red sites are associated with curves from class* 0 *and the black sites are associated with curves from class* 1.

 ISSN: 2102-6238

Since the theoretical results of this paper are related to the consistency, it is natural to consider training samples with increasing sizes and to estimate the corresponding misclassification error rates. For this aim, for each $n = 25, 50, 75$, we generate 100 samples simulated on the rectangular region $C_{n,m} = \{(i,j) \in \mathbb{N}^2 : 1 \leq i \leq n+m, 1 \leq j \leq n\}$ which consists of $n(n+m)$ sites, for some positive integer $m \geq 1$. In each replication, the region $C_{n,m}$ is partitioned into two regions: the training region $I_n = \{(i,j) : 1 \leq i, j \leq n\}$ and the test region $J_{n,m} = \{(i,j) : n+1 \leq i \leq n+m, 1 \leq j \leq n\}$ where $\text{Card}(J_{n,m}) = nm$. We will choose m so that the size of the test sample is 150. For example, if $n = 25$, we choose $m = 6$. In this case, the training region is I_{25} and the test region is $J_{25,6} = \{(i,j) : 26 \leq i \leq 31, 1 \leq j \leq 25\}$. If $n = 50$, we choose $m = 3$. In this case, the training region is I_{50} and the test region is $J_{50,3} = \{(i,j) : 51 \leq i \leq 53, 1 \leq j \leq 50\}$. If $n = 75$, we choose $m = 2$. In this case, the training region is I_{75} and the test region is $J_{75,2} = \{(i,j) : 76 \leq i \leq 77, 1 \leq j \leq 75\}$. In each replication, the proposed classifier is determined on the basis of $D_n = \{(X_\mathbf{i}, Y_\mathbf{i}) : \mathbf{i} \in I_n\}$, the training sample, at hand (based on the optimal bandwidth minimizing the $CV(b)$) and the misclassification error rate (ER) is evaluated based on the associated test sample $D'_{n,m} = \{(X_\mathbf{i}, Y_\mathbf{i}) : \mathbf{i} \in J_{n,m}\}$, where $ER = \frac{1}{nm} \sum_{\mathbf{i} \in J_{n,m}} \mathbb{1}_{\{Y_\mathbf{i} \neq g(X_\mathbf{i})\}}$. Table 3 reports the average error rate (AER), obtained by averaging the error rates associated with the corresponding 100 test samples.

TABLE 1. *Table 1: Estimated optimal bandwidths and average error rates corresponding to training samples of different sizes.*

n	25	50	75
h_{opt}	3.3	2.8	2.0
AER	13.2%	10.3%	8.4%

Table 1 shows that the estimated optimal bandwidth and the error rate decrease when the training sample size increases. This means that the practical results in the simulation study are in line with the theoretical results.

7. Proofs

We recall that we have by (1)

$$g_\mathbf{n}(x) = \begin{cases} 0 \text{ if } \eta_\mathbf{n}(x) \leq \dfrac{\sum_{\mathbf{i} \in I_\mathbf{n}} (1 - Y_\mathbf{i}) \mathbb{1}_{\{X_\mathbf{i} \in B_{x,b}\}}}{\hat{\mathbf{n}} \mu(B_{x,b})} \\ \\ 1 \text{ otherwise,} \end{cases}$$

where

$$\eta_\mathbf{n}(x) = \frac{\sum_{\mathbf{i} \in I_\mathbf{n}} Y_\mathbf{i} \mathbb{1}_{\{X_\mathbf{i} \in B_{x,b}\}}}{\hat{\mathbf{n}} \mu(B_{x,b})}.$$

 ISSN: 2102-6238

Proof of Lemma 4.3 By Cauchy-Schwartz inequality, we have

$$\begin{aligned}\mathbb{E}|\eta_{\mathbf{n}}(x)-\mathbb{E}\eta_{\mathbf{n}}(x)| &\leq (\mathrm{var}(\eta_{\mathbf{n}}(x)))^{1/2}\\ &\leq \left(\frac{\mathrm{var}(Y\mathbb{1}_{\{X\in B_{x,b}\}})}{\hat{\mathbf{n}}(\mu(B_{x,b}))^2}+S_{\mathbf{n}}(x)\right)^{1/2}\\ &\leq \left(\frac{\mathbb{E}(Y\mathbb{1}_{\{X\in B_{x,b}\}})^2}{\hat{\mathbf{n}}(\mu(B_{x,b}))^2}+S_{\mathbf{n}}(x)\right)^{1/2},\end{aligned}$$

where

$$S_{\mathbf{n}}(x)=\frac{1}{(\hat{\mathbf{n}}\mu(B_{x,b}))^2}\sum_{(\mathbf{i},\mathbf{j})\in I_{\mathbf{n}}\times I_{\mathbf{n}}:\mathbf{i}\neq\mathbf{j}}|\mathrm{cov}(\Delta_{\mathbf{i}},\Delta_{\mathbf{j}})|$$

and $\Delta_{\mathbf{i}}=Y_{\mathbf{i}}\mathbb{1}_{\{X_{\mathbf{i}}\in B_{x,b}\}}$ for all $\mathbf{i}\in I_{\mathbf{n}}$. Now, since $|Y|\leq 1$, we obtain

$$\mathbb{E}|\eta_{\mathbf{n}}(x)-\mathbb{E}\eta_{\mathbf{n}}(x)|\leq\left(\frac{1}{\hat{\mathbf{n}}\mu(B_{x,b})}+S_{\mathbf{n}}(x)\right)^{1/2}. \tag{5}$$

Let us first deal with the cross term $S_{\mathbf{n}}(x)$. Let $u_{\mathbf{n}}$ a sequence of positive numbers such that $u_{\mathbf{n}}\to\infty$ as $\mathbf{n}\to\infty$. Then, we can write

$$\begin{aligned}S_{\mathbf{n}}(x)=&\frac{1}{(\hat{\mathbf{n}}\mu(B_{x,b}))^2}\sum_{(\mathbf{i},\mathbf{j})\in I_{\mathbf{n}}\times I_{\mathbf{n}}:0<\|\mathbf{i}-\mathbf{j}\|\leq u_{\mathbf{n}}}|\mathrm{cov}(\Delta_{\mathbf{i}},\Delta_{\mathbf{j}})|\\ &+\frac{1}{(\hat{\mathbf{n}}\mu(B_{x,b}))^2}\sum_{(\mathbf{i},\mathbf{j})\in I_{\mathbf{n}}\times I_{\mathbf{n}}:\|\mathbf{i}-\mathbf{j}\|>u_{\mathbf{n}}}|\mathrm{cov}(\Delta_{\mathbf{i}},\Delta_{\mathbf{j}})|.\end{aligned} \tag{6}$$

Now, for $0<\|\mathbf{i}-\mathbf{j}\|\leq u_{\mathbf{n}}$, according to Assumption 2, we have

$$\begin{aligned}|\mathrm{cov}(\Delta_{\mathbf{i}},\Delta_{\mathbf{j}})|&\leq\mathbb{E}(\Delta_{\mathbf{i}}\Delta_{\mathbf{j}})+\mathbb{E}(\Delta_{\mathbf{i}})\mathbb{E}(\Delta_{\mathbf{j}})\\ &\leq\mathbb{P}((X_{\mathbf{i}},X_{\mathbf{j}})\in B_{x,b}\times B_{x,b})+\{\mathbb{P}(X\in B_{x,b})\}^2\\ &\leq C\{\mu(B_{x,b})\}^{1+\varepsilon_1}+\{\mu(B_{x,b})\}^2,\end{aligned}$$

where $0<\varepsilon_1\leq 1$ is the constant defined in Assumption 2 and C is a generic positive constant, independent of both x and $\mathbf{n}$, whose value may vary from line to line. Since $\mu(B_{x,b})\leq 1$ and

$$Card\{(\mathbf{i},\mathbf{j})\in I_{\mathbf{n}}\times I_{\mathbf{n}}:0<\|\mathbf{i}-\mathbf{j}\|\leq u_{\mathbf{n}}\}\leq\sum_{\mathbf{i}\in I_{\mathbf{n}}}Card\{\mathbf{j}\in I_{\mathbf{n}}:0<\|\mathbf{i}-\mathbf{j}\|\leq u_{\mathbf{n}}\}\leq\hat{\mathbf{n}}(2u_{\mathbf{n}})^N,$$

then, $\{\mu(B_{x,b})\}^2\leq\{\mu(B_{x,b})\}^{1+\varepsilon_1}$ and

$$\sum_{(\mathbf{i},\mathbf{j})\in I_{\mathbf{n}}\times I_{\mathbf{n}}:0<\|\mathbf{i}-\mathbf{j}\|\leq u_{\mathbf{n}}}|\mathrm{cov}(\Delta_{\mathbf{i}},\Delta_{\mathbf{j}})|\leq C\hat{\mathbf{n}}u_{\mathbf{n}}^N\{\mu(B_{x,b})\}^{1+\varepsilon_1}. \tag{7}$$

If $\|\mathbf{i}-\mathbf{j}\|>u_{\mathbf{n}}$, by Lemma 3.1 and the fact that $|Y|\leq 1$, we get

$$|\mathrm{cov}(\Delta_{\mathbf{i}},\Delta_{\mathbf{j}})|\leq 4\alpha(\|\mathbf{i}-\mathbf{j}\|). \tag{8}$$

 ISSN: 2102-6238

From (6), (7) and (8), we can write

$$\begin{aligned} S_{\mathbf{n}}(x) \leq & \frac{Cu_{\mathbf{n}}^{N}}{\hat{\mathbf{n}}(\mu(B_{x,b}))^{1-\varepsilon_1}} + \frac{4}{(\hat{\mathbf{n}}\mu(B_{x,b}))^{2}} \sum_{(\mathbf{i},\mathbf{j})\in I_{\mathbf{n}}\times I_{\mathbf{n}}:\|\mathbf{i}-\mathbf{j}\|\geq u_{\mathbf{n}}} \alpha(\|\mathbf{i}-\mathbf{j}\|) \\ & \leq \frac{Cu_{\mathbf{n}}^{N}}{\hat{\mathbf{n}}(\mu(B_{x,b}))^{1-\varepsilon_1}} + \frac{4}{\hat{\mathbf{n}}(\mu(B_{x,b}))^{2}} \sum_{i\geq u_{\mathbf{n}}} i^{N-1}\alpha(i). \end{aligned} \tag{9}$$

Since by assumption $\alpha(i) \leq Ci^{-\theta}$ for some $\theta > 2N$, it follows that

$$\sum_{i\geq u_{\mathbf{n}}} i^{N-1}\alpha(i) \leq C\int_{u_{\mathbf{n}}-1}^{\infty} t^{N-\theta-1}dt. \tag{10}$$

Now, since $u_{\mathbf{n}} \to \infty$, then $u_{\mathbf{n}} - 1 \geq u_{\mathbf{n}}/2$ for $\hat{\mathbf{n}}$ sufficiently large. Then, we have

$$\int_{u_{\mathbf{n}}-1}^{\infty} t^{N-\theta-1}dt \leq \frac{Cu_{\mathbf{n}}{}^{N-\theta}}{\theta - N}.$$

Consequently, by (9) and (10), we obtain

$$S_n(x) \leq \frac{Cu_{\mathbf{n}}^{N}}{\hat{\mathbf{n}}(\mu(B_{x,b}))^{1-\varepsilon_1}} + \frac{Cu_{\mathbf{n}}^{N-\theta}}{\hat{\mathbf{n}}(\mu(B_{x,b}))^{2}}. \tag{11}$$

If we choose $u_{\mathbf{n}} = \{\mu(B_{x,b})\}^{-\varepsilon_1/N}$ and $N/(\theta - N) < \varepsilon_1 \leq 1$, from Assumption 2 and since $\theta > 2N$, we get

$$S_{\mathbf{n}}(x) \leq \frac{C}{\hat{\mathbf{n}}\mu(B_{x,b})}. \tag{12}$$

Thus, form (5) and (12), it follows that

$$\mathbb{E}|\eta_{\mathbf{n}}(x) - \mathbb{E}\eta_{\mathbf{n}}(x)| \leq \frac{C}{\sqrt{\hat{\mathbf{n}}\mu(B_{x,b})}}.$$

Therefore, according to Fubini's theorem, Jensens's inequality and Lemma 4.2, we conclude that

$$\begin{aligned} \mathbb{E}\int_{E_k} |\eta_{\mathbf{n}}(x) - \mathbb{E}\eta_{\mathbf{n}}(x)|\mu(dx) &\leq C\int_{E_k} \frac{1}{\sqrt{\hat{\mathbf{n}}\mu(B_{x,b})}}\mu(dx) \\ &\leq C\left(\int_{E_k} \frac{1}{\hat{\mathbf{n}}\mu(B_{x,b})}\mu(dx)\right)^{1/2} \\ &\leq C\left(\frac{1}{\hat{\mathbf{n}}}\mathscr{N}_k\left(\frac{b}{2}\right)\right)^{1/2} \end{aligned}$$

and the proof is completed. □

 ISSN: 2102-6238

Proof of Theorem 5.1 Since the extension of Theorem 2.3 in Devroye et al. (1996) to the infinite dimensional setting is straightforward, thus, the consistency will be proved if we show that

$$\mathbb{E}\int_E |\eta(x)-\eta_{\mathbf{n}}(x)|\mu(dx) \longrightarrow 0 \text{ as } \mathbf{n}\to\infty. \tag{13}$$

Since $\eta(x)\leq 1$ and $\mathbb{E}\eta_{\mathbf{n}}(x)\leq 1$ for every $x\in E$ and $\mathbf{n}\in(\mathbb{N}^*)^N$, we have for each $k\geq 1$,

$$\begin{aligned}
&\mathbb{E}\int_E |\eta(x)-\eta_{\mathbf{n}}(x)|\mu(dx)\\
&\quad=\mathbb{E}\int_{E_k} |\eta(x)-\eta_{\mathbf{n}}(x)|\mu(dx)+\mathbb{E}\int_{E_k^c} |\eta(x)-\eta_{\mathbf{n}}(x)|\mu(dx)\\
&\quad\leq\int_{E_k} |\eta(x)-\mathbb{E}\eta_{\mathbf{n}}(x)|\mu(dx)+\mathbb{E}\int_{E_k} |\eta_{\mathbf{n}}(x)-\mathbb{E}\eta_{\mathbf{n}}(x)|\mu(dx)+2\mu(E_k^c).
\end{aligned}$$

Consequently, according to Lemma 4.3, we get the following inequality

$$\begin{aligned}
&\mathbb{E}\int_E |\eta(x)-\eta_{\mathbf{n}}(x)|\mu(dx)\\
&\qquad\leq\int_E |\eta(x)-\mathbb{E}\eta_{\mathbf{n}}(x)|\mu(dx)+C\left(\frac{1}{\hat{\mathbf{n}}}\mathscr{N}_k\left(\frac{b}{2}\right)\right)^{1/2}+2\mu(E_k^c).
\end{aligned} \tag{14}$$

Therefore, by using Lemma 4.1 and the assumptions on b, we obtain for each $k\geq 1$,

$$\limsup_{\mathbf{n}\to\infty}\mathbb{E}\int_E |\eta(x)-\eta_{\mathbf{n}}(x)|\mu(dx)\leq 2\mu(E_k^c).$$

If we let k go to infinity, Assumption 1 yields

$$\limsup_{\mathbf{n}\to\infty}\mathbb{E}\int_E |\eta(x)-\eta_{\mathbf{n}}(x)|\mu(dx)=0 \tag{15}$$

and the proof of the theorem is completed. □

Proof of Theorem 5.2 The strong consistency will be proved if we show that

$$\int_E |\eta(x)-\eta_n(x)|\mu(dx)\longrightarrow 0 \quad \text{as } n\to\infty \text{ with probability one.} \tag{16}$$

We set $Z=(X,Y)$ and $Z_{\mathbf{i}}=(X_{\mathbf{i}},Y_{\mathbf{i}})$ for each $\mathbf{i}\in I_n$. By assumptions, X and $X_{\mathbf{i}}$ take values in the Polish metric space E, so, $Z=(X,Y)$ and $Z_{\mathbf{i}}=(X_{\mathbf{i}},Y_{\mathbf{i}})$ take values in the product space $E\times\{0,1\}$ which is also Polish. Without loss of generality, let $n=2pq$ for $p=p_n,q=q_n\in[1,n/2]$ two positive integers where p is define in Theorem 5.2 such that $p_n\to\infty$ as $n\to\infty$, and let

$$J_q=\{\mathbf{j}=(j_1,...,j_N)\in\mathbb{N}^N:\ 0\leq j_k\leq q-1,\ \forall k=1,...,N\}.$$

 ISSN: 2102-6238

We define blocks, inspired by Tran (1990) (see also Carbon et al., 1997), as follow, for each $\mathbf{j} \in J_q$,

$$
\begin{aligned}
S_{\mathbf{j}}^{(1)} &= \{\mathbf{i} \in I_n : 2j_k p + 1 \leq i_k \leq (2j_k+1)p,\ k = 1,\dots,N\} \\
S_{\mathbf{j}}^{(2)} &= \{\mathbf{i} \in I_n : 2j_k p + 1 \leq i_k \leq (2j_k+1)p,\ k = 1,\dots,N-1 \\
&\qquad \textit{and}\ (2j_N+1)p + 1 \leq i_N \leq 2(j_N+1)p\} \\
&\qquad \dots \\
S_{\mathbf{j}}^{(2^N-1)} &= \{\mathbf{i} \in I_n : (2j_k+1)p + 1 \leq i_k \leq 2(j_k+1)p,\ k = 1,\dots,N-1 \\
&\qquad \textit{and}\ 2j_N p + 1 \leq i_N \leq (2j_N+1)p\} \\
S_{\mathbf{j}}^{(2^N)} &= \{\mathbf{i} \in I_n : (2j_k+1)p + 1 \leq i_k \leq 2(j_k+1)p,\ k = 1,\dots,N\}.
\end{aligned}
$$

We have

$$
I_n = \bigcup_{i=1}^{2^N} \bigcup_{\mathbf{j} \in J_q} S_{\mathbf{j}}^{(i)} \tag{17}
$$

and one can easily prove that for all $\mathbf{j} \in J_q$, $\mathrm{Card}(S_{\mathbf{j}}^{(i)}) = p^N$ and for all $\mathbf{j} \neq \mathbf{j}'$, $\mathrm{dist}(S_{\mathbf{j}}^{(i)}, S_{\mathbf{j}'}^{(i)}) \geq p$. For each $i = 1,\dots,2^N$ and $\mathbf{j} \in J_q$, let $W_{\mathbf{j}}^{(i)} = (Z_{\mathbf{i}},\ \mathbf{i} \in S_{\mathbf{j}}^{(i)})$ and let $\psi : \{1,\dots,q^N\} \to J_q$ be a bijection. We can define a lexicographic order relation $\leq_{lex}$ on J_q as follows: for all $l, l' \in \{1,\dots,q^N\}$, we have $\psi(l) \leq_{lex} \psi(l')$ if $l \leq l'$. For any $\mathbf{j} \in J_q$, we can find $l \in \{1,\dots,q^N\}$ with $\psi(l) = \mathbf{j}$. Now, for each $i = 1,\dots,2^N$, by applying Lemma 3.2 together with a decomposition in blocks similar to that introduced by Doukhan et al. (1995) (see also Viennet, 1967) on the family of vectors $\{W_{\psi(l)}^{(i)},\ l = 1,\dots,q^N\}$, we can generate independent copies $\{\tilde{W}_{\psi(l)}^{(i)},\ l = 1,\dots,q^N\}$ such that: they are mutually independent, for all $l \in \{1,\dots,q^N\}$, $\tilde{W}_{\psi(l)}^{(i)} = (\tilde{Z}_{\mathbf{i}},\ \mathbf{i} \in S_{\psi(l)}^{(i)})$ has the same distribution as $W_{\psi(l)}^{(i)} = (Z_{\mathbf{i}},\ \mathbf{i} \in S_{\psi(l)}^{(i)})$ and $\mathbb{P}(W_{\psi(l)}^{(i)} \neq \tilde{W}_{\psi(l)}^{(i)}) \leq \beta(p)$ because $\mathrm{dist}(S_{\psi(l)}^{(i)}, S_{\psi(l')}^{(i)}) \geq p$ for any $l \neq l'$. As a consequence, we have

$$
\mathbb{P}(Z_{\mathbf{i}} \neq \tilde{Z}_{\mathbf{i}}) = \mathbb{P}((X_{\mathbf{i}}, Y_{\mathbf{i}}) \neq (\tilde{X}_{\mathbf{i}}, \tilde{Y}_{\mathbf{i}})) \leq \beta(p) \text{ for all } \mathbf{i} \in I_n. \tag{18}
$$

By (17), we can write

$$
\sum_{\mathbf{i} \in I_n} \tilde{Y}_{\mathbf{i}} \mathbb{I}_{\{\tilde{X}_{\mathbf{i}} \in B_{x,b}\}} = \sum_{i=1}^{2^N} \sum_{\mathbf{j} \in J_q} \sum_{\mathbf{i} \in S_{\mathbf{j}}^{(i)}} \tilde{Y}_{\mathbf{i}} \mathbb{I}_{\{\tilde{X}_{\mathbf{i}} \in B_{x,b}\}}
$$

where for each $i = 1,\dots,2^N$, the variables $\left\{ \sum_{\mathbf{i} \in S_{\mathbf{j}}^{(i)}} \tilde{Y}_{\mathbf{i}} \mathbb{I}_{\{\tilde{X}_{\mathbf{i}} \in B_{x,b}\}},\ \mathbf{j} \in J_q \right\}$ are mutually independent. If we denote

$$
\tilde{\eta}_n(x) = \frac{\sum_{\mathbf{i} \in I_n} \tilde{Y}_{\mathbf{i}} \mathbb{I}_{\{\tilde{X}_{\mathbf{i}} \in B_{x,b}\}}}{n^N \mu(B_{x,b})} \text{ and } \tilde{\eta}_{n,i}(x) = \frac{\sum_{\mathbf{j} \in J_q} \sum_{\mathbf{i} \in S_{\mathbf{j}}^{(i)}} \tilde{Y}_{\mathbf{i}} \mathbb{I}_{\{\tilde{X}_{\mathbf{i}} \in B_{x,b}\}}}{n^N \mu(B_{x,b})}
$$

then,

$$
\tilde{\eta}_n(x) = \sum_{i=1}^{2^N} \tilde{\eta}_{n,i}(x). \tag{19}
$$

 ISSN: 2102-6238

Let $(k_n)_{n\geq 1}$ be the sequence of increasing positive integers defined in Theorem 5.2. We first proceed to show that

$$\int_{E_{k_n}} |\eta(x)-\eta_n(x)|\mu(dx)\to 0 \text{ with probability one as } n\to\infty. \tag{20}$$

One can easily prove that

$$\begin{aligned}&\int_{E_{k_n}} |\eta(x)-\eta_n(x)|\mu(dx)\\ &\quad\leq \mathbb{E}\int_{E_{k_n}} |\eta(x)-\eta_n(x)|\mu(dx)+\int_{E_{k_n}} |\eta_n(x)-\mathbb{E}\eta_n(x)|\mu(dx).\end{aligned} \tag{21}$$

According to (15), we have

$$\mathbb{E}\int_{E_{k_n}} |\eta(x)-\eta_n(x)|\mu(dx)\leq \mathbb{E}\int_{E} |\eta(x)-\eta_n(x)|\mu(dx)\longrightarrow 0 \text{ as } n\to\infty.$$

Thus, by (21), in order to prove (20), it suffices to show that

$$\int_{E_{k_n}} |\eta_n(x)-\mathbb{E}\eta_n(x)|\mu(dx)\to 0 \text{ with probability one as } n\to\infty. \tag{22}$$

Using Markov's inequality, we have for any $\varepsilon>0$,

$$\begin{aligned}&\mathbb{P}\left(\left|\int_{E_{k_n}} |\eta_n(x)-\mathbb{E}\eta_n(x)|\mu(dx)-\int_{E_{k_n}} |\tilde{\eta}_n(x)-\mathbb{E}\tilde{\eta}_n(x)|\mu(dx)\right|>\varepsilon\right)\\
&\leq \varepsilon^{-1}\mathbb{E}\left|\int_{E_{k_n}} |\eta_n(x)-\mathbb{E}\eta_n(x)|\mu(dx)-\int_{E_{k_n}} |\tilde{\eta}_n(x)-\mathbb{E}\tilde{\eta}_n(x)|\mu(dx)\right|\\
&\leq \varepsilon^{-1}\mathbb{E}\int_{E_{k_n}} \big||\eta_n(x)-\mathbb{E}\eta_n(x)|-|\tilde{\eta}_n(x)-\mathbb{E}\tilde{\eta}_n(x)|\big|\,\mu(dx)\\
&\leq \varepsilon^{-1}\mathbb{E}\left(\int_{E_{k_n}} |\tilde{\eta}_n(x)-\eta_n(x)|\mu(dx)+\mathbb{E}\int_{E_{k_n}} |\tilde{\eta}_n(x)-\eta_n(x)|\mu(dx)\right)\\
&= 2\varepsilon^{-1}\mathbb{E}\int_{E_{k_n}} |\tilde{\eta}_n(x)-\eta_n(x)|\mu(dx)\\
&= 2\varepsilon^{-1}\mathbb{E}\int_{E_{k_n}} \left|\frac{\sum_{\mathbf{i}\in I_n}\tilde{Y}_{\mathbf{i}}\mathbb{I}_{\{\tilde{X}_{\mathbf{i}}\in B_{x,b}\}}}{n^N\mu(B_{x,b})}-\frac{\sum_{\mathbf{i}\in I_n}Y_{\mathbf{i}}\mathbb{I}_{\{X_{\mathbf{i}}\in B_{x,b}\}}}{n^N\mu(B_{x,b})}\right|\mu(dx)\\
&\leq 2\varepsilon^{-1}\sum_{\mathbf{i}\in I_n}\mathbb{E}\mathbb{I}_{\{(\tilde{X}_{\mathbf{i}},\tilde{Y}_{\mathbf{i}})\neq(X_{\mathbf{i}},Y_{\mathbf{i}})\}}\int_{E_{k_n}}\left|\frac{\tilde{Y}_{\mathbf{i}}\mathbb{I}_{\{\tilde{X}_{\mathbf{i}}\in B_{x,b}\}}}{n^N\mu(B_{x,b})}-\frac{Y_{\mathbf{i}}\mathbb{I}_{\{X_{\mathbf{i}}\in B_{x,b}\}}}{n^N\mu(B_{x,b})}\right|\mu(dx)\\
&\leq 4\varepsilon^{-1}\sum_{\mathbf{i}\in I_n}\mathbb{E}\mathbb{I}_{\{(\tilde{X}_{\mathbf{i}},\tilde{Y}_{\mathbf{i}})\neq(X_{\mathbf{i}},Y_{\mathbf{i}})\}}\int_{E_{k_n}}\frac{1}{n^N\mu(B_{x,b})}\mu(dx)\end{aligned}$$

As a consequence, by Lemma 4.2, (18) and (4), we have

$$\mathbb{P}\left(\left|\int_{E_{k_n}} |\eta_n(x)-\mathbb{E}\eta_n(x)|\mu(dx)-\int_{E_{k_n}} |\tilde{\eta}_n(x)-\mathbb{E}\tilde{\eta}_n(x)|\mu(dx)\right|>\varepsilon\right)$$

 ISSN: 2102-6238

$$\leq 4\varepsilon^{-1}\sum_{\mathbf{i}\in I_n}\mathbb{P}\left((\tilde{X}_{\mathbf{i}},\tilde{Y}_{\mathbf{i}})\neq(X_{\mathbf{i}},Y_{\mathbf{i}})\right)\int_{E_{k_n}}\frac{1}{n^N\mu(B_{x,b})}\mu(dx)$$
$$\leq C\varepsilon^{-1}\mathscr{N}_{k_n}\left(\frac{b}{2}\right)\beta(p)\leq C\varepsilon^{-1}\mathscr{N}_{k_n}\left(\frac{b}{2}\right)p^{-\theta_1},$$

for some generic constant $C>0$ independent of both x and n. Thus, by the assumptions on p and the Borel-Cantelli lemma, we conclude that

$$\int_{E_{k_n}}|\eta_n(x)-\mathbb{E}\eta_n(x)|\mu(dx)-\int_{E_{k_n}}|\tilde{\eta}_n(x)-\mathbb{E}\tilde{\eta}_n(x)|\mu(dx)\longrightarrow 0$$

with probability one as $n\to\infty$. Consequently, (22) will be proved if we show that

$$\int_{E_{k_n}}|\tilde{\eta}_n(x)-\mathbb{E}\tilde{\eta}_n(x)|\mu(dx)\longrightarrow 0 \text{ with probability one as } n\to\infty. \tag{23}$$

To do that, by (19), we have

$$\int_{E_{k_n}}|\tilde{\eta}_n(x)-\mathbb{E}\tilde{\eta}_n(x)|\mu(dx)\leq\sum_{i=1}^{2^N}\int_{E_{k_n}}|\tilde{\eta}_{n,i}(x)-\mathbb{E}\tilde{\eta}_{n,i}(x)|\mu(dx). \tag{24}$$

To establish (23), by (24) it is sufficient to show that

$$\int_{E_{k_n}}|\tilde{\eta}_{n,i}(x)-\mathbb{E}\tilde{\eta}_{n,i}(x)|\mu(dx)\to 0 \text{ as } n\to\infty \text{ with probability one,} \tag{25}$$

for each $1\leq i\leq 2^N$. Without loss of generality, we show (25) for $i=1$. To do that, we denote for each $l=1,...,q^N$, $\tilde{W}_l:=\tilde{W}^{(1)}_{\psi(l)}=((\tilde{X}_{\mathbf{i}},\tilde{Y}_{\mathbf{i}}),\ \mathbf{i}\in S^{(i)}_{\psi(l)})$. Let $F:((E\times\{0,1\})^{p^N})^{q^N}\to\mathbb{R}$ a real function defined as follows

$$F(\tilde{W}_1,...,\tilde{W}_{q^N})=\int_{E_{k_n}}|\tilde{\eta}_{n,1}(x)-\mathbb{E}\tilde{\eta}_{n,1}(x)|\mu(dx)$$
$$=\int_{E_{k_n}}\left|\sum_{\mathbf{j}\in J_q}\sum_{\mathbf{i}\in S^{(1)}_{\mathbf{j}}}\left(\frac{\tilde{Y}_{\mathbf{i}}\mathbb{1}_{\{\tilde{X}_{\mathbf{i}}\in B_{x,b}\}}}{n^N\mu(B_{x,b})}-\frac{\mathbb{E}(\tilde{Y}\mathbb{1}_{\{\tilde{X}\in B_{x,b}\}})}{n^N\mu(B_{x,b})}\right)\right|\mu(dx).$$

For $\tilde{w}_l\neq\tilde{w}_l'$ where $\tilde{w}_l,\ \tilde{w}_l'\in(E\times\{0,1\})^{p^N}$, using Lemma 4.2, we have

$$|F(\tilde{W}_1,...\tilde{w}_l,...,\tilde{W}_{q^N})-F(\tilde{W}_1,...\tilde{w}_l',...,\tilde{W}_{q^N})|\leq\frac{2p^N}{n^N}\int_{E_{k_n}}\frac{1}{\mu(B_{x,b})}\mu(dx)$$
$$\leq\frac{Cp^N}{n^N}\mathscr{N}_{k_n}\left(\frac{b}{2}\right),$$

where $C>0$ is the generic constant. Hence, by McDiarmid's inequality (see McDiarmid, 1989), we have for every $\varepsilon>0$,

$$\mathbb{P}\left(\left|F(\tilde{W}_1,...,\tilde{W}_{q^N})-\mathbb{E}(F(\tilde{W}_1,...,\tilde{W}_{q^N}))\right|>\varepsilon\right)\leq 2\exp\left(-\frac{\varepsilon^2n^N}{C^2p^N\mathscr{N}^2_{k_n}\left(\frac{b}{2}\right)}\right).$$

 ISSN: 2102-6238

Now, by the Borel-Cantelli lemma and the assumption on b, we conclude that

$$\int_{E_{k_n}} |\tilde{\eta}_{n,1}(x) - \mathbb{E}\tilde{\eta}_{n,1}(x)|\mu(dx) - \mathbb{E}\int_{E_{k_n}} |\tilde{\eta}_{n,1}(x) - \mathbb{E}\tilde{\eta}_{n,1}(x)|\mu(dx) \longrightarrow 0$$

with probability one as $n \to \infty$. As a consequence, by (23), the proof of (22) will be completed if we prove that

$$\mathbb{E}\int_{E_{k_n}} |\tilde{\eta}_{n,1}(x) - \mathbb{E}\tilde{\eta}_{n,1}(x)|\mu(dx) \longrightarrow 0 \text{ as } n \to 0. \tag{26}$$

Since $2\alpha(t) \leq \beta(t) \leq Ct^{-\theta_1}$ for each $t \in \mathbb{N}^*$, with a straightforward adaptation of the proof of Lemma 4.3, one can easily prove (26). So, the proof of (25) is completed and then, the proof of (23) is also completed. To finish the proof of the theorem, let us denote for each $n \geq 1$ and $\mathbf{i} \in I_n$,

$$Z_{\mathbf{i}}^n = \int_{E_{k_n}^c} \frac{\mathbb{1}_{\{X_{\mathbf{i}} \in B_{x,b}\}}}{\mu(B_{x,b})}\mu(dx).$$

Consequently,

$$\mathbb{E}\left[\frac{1}{n^N}\sum_{\mathbf{i}\in I_n} Z_{\mathbf{i}}^n\right] = \mu(E_{k_n}^c).$$

By the assumption $\sum_{n\geq 1}\mu(E_{k_n}^c) < \infty$ together with the Borel-Cantelli lemma, we have

$$\frac{1}{n^N}\sum_{\mathbf{i}\in I_n} Z_{\mathbf{i}}^n \longrightarrow 0 \text{ with probability one as } n \to \infty. \tag{27}$$

Hence, we can write

$$\begin{aligned}\int_E |\eta(x) - \eta_n(x)|\mu(dx) &= \int_{E_{k_n}} |\eta(x) - \eta_n(x)|\mu(dx) + \int_{E_{k_n}^c} |\eta(x) - \eta_n(x)|\mu(dx) \\ &\leq \int_{E_{k_n}} |\eta(x) - \eta_n(x)|\mu(dx) + \mu(E_{k_n}^c) + \frac{1}{n^N}\sum_{\mathbf{i}\in I_n} Z_{\mathbf{i}}^n.\end{aligned}$$

Finally, according to Assumption 1, (20) and (27), the three terms on the right hand side of the last inequality tend to 0 as $n \to \infty$, hence (16) tends to 0 and the proof of the theorem is completed. □

Acknowledgments

The author would like to thank the anonymous referees whose valuable comments led to an improved version of the paper.

References

Abraham, C., Biau, G., and Cadre, B. (2006). On the kernel rule for function classification. *Ann. Inst. Statist. Math.*, 58:619–633.

Berbee, H. (1979). Random walks with stationary increments and renewal theory. *Math. Cent. Tracts. Amsterdam*, 58.

 ISSN: 2102-6238

Biau, G. and Cadre, B. (2004). Nonparametric spatial prediction. *Statistical Inference for Stochastic Processes*, 7:327–349.

Bosq, D. (1998). *Nonparametric Statistics for Stochastic Processes*. Springer-Verlag, New York.

Bradley, R. C. (1983). Approximation theorems for strongly mixing random variables. *Michigan Math. J.*, 30:69–81.

Carbon, M., Tran, L., and Wu, B. (1997). Kernel density estimation for random fields (density estimation for random fields). *Statistics & Probability Letters*, 36:115–125.

Cérou, F. and Guyader, A. (2006). Nearest neighbor classification in infinite dimension. *ESAIM: Probability and Statistics*, 10:340–355.

Dabo-Niang, S., Rachdi, M., and Yao, A. F. (2011). Kernel regression estimation for spatial functional random variables. *Far East Journal of Theoretical Statistics*, 32:77–113.

Dabo-Niang, S. and Yao, A. F. (2007). Kernel regression estimation for continuous spatial processes. *Mathematical Methods of Statistics*, 16:298–317.

Dabo-Niang, S. and Yao, A. F. (2013). Kernel spatial density estimation in infinite dimension space. *Metrika*, 76:19–52.

Devroye, L., Györfi, L., and Lugosi, G. (1996). *A probabilitic Theory of Pattern Recognition*. Springer-Verlag, New York.

Devroye, L. and Krzyżak, A. (1989). An equivalence theorem for L_1 convergence of the kernel regression estimate. *Journal of statistical planning and Inference*, 23:71–82.

Doukhan, P., Massart, P., and Rio, E. (1995). Invariance principles for absolutely regular empirical processes. *Ann. Inst. H. Poincaré Probab. Statist.*, 31:393–427.

El-Machkouri, M. (2007). Nonparametric regression estimation for random fields in a fixed-design. *Statistical Inference for Stochastic Processes*, 10:29–47.

Febrero-Bande, M. and de la Fuente, M. O. (2012). Statistical computing in functional data analysis: The R package fda.usc. *Journal of statistical Software*, 51.

Ferraty, F., Keilegom, I. V., and Vieu, P. (2002). The functional nonparametric model and application to spectrometric data. *Comput. Statist.*, 17:54–564.

Ferraty, F., Keilegom, I. V., and Vieu, P. (2012). Regression when both response and predictor are functions. *J. Multivariate Anal.*, 109:10–28.

Ferraty, F. and Vieu, P. (2003). Curves discrimination : a nonparametric approach. *Computational Statistics and Data Analysis*, 44:161–173.

Ferraty, F. and Vieu, P. (2006). *Nonparametric functional data analysis*. Springer-Verlag, New York.

Hallin, M., Lu, Z., and Tran, L. (2004). Local linear spatial regression. *The Annals of Statistics*, 32:2469–2500.

Hallin, M., Lu, Z., and Yu, K. (2009). Local linear spatial quantile regression. *The Annals of Statistics*, 15:659–686.

Jacques, J. and Preda, C. (2014). Model-based clustering for multivariate functional data. *Computational Statistics and Data Analysis*, 71:92–106.

Kechris, A. S. (1995). *Classical descriptive set theory*. Springer-Verlag, New York.

Kolmogorov, A. N. and Tihomirov, V. M. (1961). ε-entropy and ε-capacity of sets in functional spaces. *American Mathematical Society Translations*, 17:277–364.

Kulkarni, S. R. and Posner, S. E. (1995). Rate of convergence of nearest neighbor estimation under arbitrary sampling. *IEEE Transactions on Information Theory*, 41:1028–1039.

McDiarmid, C. (1989). On the method of bounded differences, in surveys in combinatorics. *Cambridge University Press, Cambridge*, 794:261–283.

Nadaraya, E. (1989). On estimating regression. *Theory of probability and its applications*, 9:141–142.

Neaderhouser, C. C. (1980). Convergence of block spins defined on random fields. *J. Statist. Phys.*, 22:673–684.

Parzen, E. (1962). On estimation of a probability density function and mode. *Ann. Math. Statist.*, 33:1065–1076.

Preda, C. (2007). Regression models for functional data by reproducing kernel hilbert spaces methods. *Journal of Statistical Planning and Inference*, 137::829–840.

Rachdi, M. and Vieu, P. (2007). Nonparametric regression for functional data: automatic smoothing parameter selection. *Journal of Statistical Planning and Inference*, 137:2784–2801.

Ramsay, J. and Silverman, B. (2002). *Applied Functional Data Analysis, Methods and Case Studies*. Springer-Verlag, New York.

Ramsay, J. and Silverman, B. (2005). *Functional Data Analysis*. Springer-Verlag, Springer Series in Statistics, second edition.

Rio, E. (2000). *Théorie asymptotique des processus aléatoires faiblement dépendants. Mathématiques et Applications*. Springer, Berlin.

 ISSN: 2102-6238

Rosenblatt, M. (1956). A central limit theorem and a strong mixing condition. *Proc. Nat. Acad. Sci., USA*, 42:43–47.
Rosenblatt, M. (1985). *Stationary sequences and random fields*. Birkhauser, Boston.
Rozanov, Y. A. and Volkonskii, V. (1959). Some limit theorems for random functions. *I. Teor. Veroyatn. Primen.*, 4:186–207.
Stone, C. J. (1977). Consistent nonparametric regression. *The Annals of Statistics*, 5:595–620.
Ternynck, C. (2014). Spatial regression estimation for functional data with spatial dependency. *Journal de la Société Française de Statistique*, 155:138–160.
Tran, L. (1990). Kernel density estimation on random fields. *Journal of Multivariate Analysis*, 34:37–53.
Viennet, G. (1967). Inequalities for absolutely sequence. *Application to density estimation. Probability Theory and Related Fields*, 107:467–492.
Watson, G. (1964). Smooth regression analysis. *Sankhy a Ser. A*, 26:359–372.
Younso, A. (2017a). On nonparametric classification for weakly dependent functional processes. *ESAIM: Probability and Statistics*. https://doi.org/10.1051/ps/2017002.
Younso, A. (2017b). On the consistency of a new kernel rule for spatially dependent data. *Statistics & Probability Letters*, 131:64–71.

 ISSN: 2102-6238