Brazilian Journal of Probability and Statistics
2020, Vol. 34, No. 1, 112–126
https://doi.org/10.1214/18-BJPS418

Nonparametric discrimination of areal functional data

Ahmad Younso
Damascus University

Abstract. We consider a new nonparametric rule of classification, inspired from the classical moving window rule, that allows for the classification of spatially dependent functional data containing some completely missing curves. We investigate the consistency of this classifier under mild conditions. The practical use of the classifier will be illustrated through simulation studies.

1 Introduction

The standard statistical techniques for modeling functional data are focused on independent functions, see for key references: Cardot and Sarda (2005), Biau, Bunea and Wegkamp (2005), Abraham, Biau and Cadre (2006), Cérou and Guyader (2006), Berlinet, Biau and Rouvière (2008), Chang, Chen and Ogden (2014) and Górecki, Krzysko and Wolynski (2015). However, in several disciplines of applied sciences there exists an increasing interest in modeling correlated functional data where the independence assumption does not hold: this is the case when samples of functions are observed over a discrete set of time points (temporally correlated functional data, see Younso (2017b)) or when these functions are observed in different sites of a region (spatially correlated functional data, see Nerini, Monestiez and Manté (2010), Delicado et al. (2010) and the references therein). The classical models based on the independence assumption are not appropriated for spatial functional data since they fail to capture locational information. For example, in the task of hyperspectral remote sensing labeling, a classifier could classify a pixel as *vegetation*, even if all adjacent pixels were classified as *non-vegetation*. It is then natural to assume that curves show certain spatial dependence. The spatial statistical literature distinguishes three types of spatial dependence that will be discussed in Section 3. In this paper, we focus on the model where, on the one hand, curves are indexed in the regular N-dimensional lattice $\mathbb{Z}^N$ and on the other hand, curves in nearby sites are more dependent than curves in sites far apart which motivates using the mixing condition that will be defined later. The need to classify spatial functional data occurs in many scientific problems. For example, in hyperspectral remote sensing, each pixel in the resulting image has its own spectral reflectance curve which is obtained by measuring the energy percentage emitted from the target pixel over a variety of different wavelengths. Since the spectral reflectance curves are obtained by physical measurements, they include different type of noise or errors. An important classification problem is how to use a set of identified spectral reflectance curves to classify a spectral reflectance curve (see, e.g., water or green vegetation) emitted from a new pixel's area captured on the Earth's surface. Many existing classification algorithms assume either certain parametric distributions for the data or certain forms of separating curves or surfaces, see Saltyte-Benth and Ducinskas (2005). These parametric classifiers become suboptimal and of limited use in practical applications when little or no information about the underlying distributions is available a priori. In comparison, nonparametric classifiers are usually more flexible in accommodating different data structures, and are hence more desirable, see Younso (2018). The

Key words and phrases. Bayes rule, spectral reflectance curve, functional data, moving window rule, random field, consistency.
Received June 2018; accepted September 2018.

major problem of the classification approach is the incomplete data when the curves in the set of sites to classify are corrupted or missing. For example, in sensing applications, missing observations occur when a subset of sensors may be absent or fail to operate at certain regions. In a parametric context, Saltyte-Benth and Ducinskas (2005) consider classification of the realization of a multivariate spatial–temporal Gaussian random field and Ruiz-Medina, Espejo and Romano (2014) propose a spatial functional normal mixed effect approach to classify of spatially dependent Gaussian curves. Romano, Balzanella and Verde (2010) deal with unsupervised classification problem and present a strategy for clustering spatio-functional data. Moughal (2013) compares the (SVM) classifier to the other two well known classifiers *that is*, Maximum likelihood (ML) and Spectral Angle Mapper (SAM). The results obtained from the sensor dataset of Moughal (2013) show that (SVM) classifier is much more effective than other conventional classifiers in term of classification accuracy. Lin and Yan (2016) construct a hyperspectral remote sensing imagery classifier by combining a polynomial kernel function with a radial basis kernel function to form a new kernel function model, and then a genetic algorithm is used to optimize the Support Vector Machine (SVM) parameters with ignoring the spatial dependence among observations. More recently, Carlo, Paolo and Roberto (2017) incorporate the spatial component as a non-stationary Markov random field conditioned to the k-nearest neighbourhood structure and propose a model based approach to the functional clustering when time-varying curves are spatially dependent. In the present paper, we propose and study a nonparametric kernel-based (supervised) classifier inspired by Younso (2017a) in the finite-dimensional case together with the classifier defined in Younso (2018) in the functional case. The proposed classifier, on the one hand, takes into account the spatial structure of data and on the other hand, it does not require any strict assumption about the statistical distribution of the data. Furthermore, the proposed method allows classification of unsampled site using information in its nearby sites. To the best of our knowledge, Younso (2018) is the first work dealing with the classification of spatial curves by kernel-based method. The literature dealing with the kernel-based rules and the nearest neighbor rule when data are independent is extensive in finite or infinite-dimensional spaces, for example Devroye, Györfi and Lugosi (1996) investigates the universal consistency of the kernel rule and the k-nearest neighbors rule in the finite-dimensional case (see also Devroye and Krzyżak (2013)). In the infinite-dimensional case, Abraham, Biau and Cadre (2006) and Ferraty, Van Keilegom and Vieu (2012) study the consistency of kernel rule and Cérou and Guyader (2006) study the consistency of the k-nearest neighbors rule. Abraham, Biau and Cadre (2006) show that the classical kernel rule is consistent and not universally consistent. They give sufficient conditions to extend the consistency result established by Devroye, Györfi and Lugosi (1996) to a function space. Younso (2017b) extends the result of Abraham, Biau and Cadre (2006) to temporally dependent case. The result of Younso (2017b) is extended to the functional random field by Younso (2018). Unfortunately, the classifier proposed by Younso (2018) can not be applicable to classify missing data. In this paper, we construct a kerne-based classifier that allows the classification of spatial observations even if they are completely missing. We study the consistency of this classifier under mild conditions. These results generalize the results of Younso (2017a) to the infinite-dimensional space.

The rest of the paper is organized as follows: Section 2 provides notations and definitions. In Section 3, some preliminaries are introduced and the main result on the consistency is stated. Section 4 gives insights on how to present numerical results or applications. Proofs of Lemma 3.4 and Theorem 3.1 are established in Section 5.

2 Moving window rule for discrimination of areal functional data

In this section, we introduce some notations and definitions. Classical supervised classification deals with predicting the unknown nature Y, called a *label*, of of an observation X with values in $\mathbb{R}^p$. Both X and Y are assumed to be random, and the distribution of (X, Y) just describes the frequency of encountering particular pairs in practice. In this paper, we deal with the case when X takes values in a function space instead of $\mathbb{R}^p$. Let $(\mathcal{E}, d)$ be a metric space where $\mathcal{E}$ is a function space and d is the metric on $\mathcal{E}$. Consider a strictly stationary functional random field $\{(X_{\mathbf{i}}, Y_{\mathbf{i}})\}_{\mathbf{i}\in\mathbb{Z}^N}$ defined on some probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and taking values in $\mathcal{E} \times \{0, 1\}$. In such context, the observed values of the functional random field are called *areal* (*lattice*) *functional data*. In the problem of classification in infinite-dimensional space, $X_{\mathbf{i}}$ is a curve and $Y_{\mathbf{i}}$ is the label (class) of $X_{\mathbf{i}}$, for each $\mathbf{i} \in \mathbb{Z}^N$. A point $\mathbf{i} = (i_1, \ldots, i_N) \in \mathbb{Z}^N$ will be referred to as a site. For $\mathbf{n} = (n_1, \ldots, n_N) \in (\mathbb{N}^*)^N$, we define the rectangular region $\mathcal{I}_{\mathbf{n}}$ by

$$\mathcal{I}_{\mathbf{n}} = \{\mathbf{i} \in \mathbb{Z}^N : 1 \le i_k \le n_k, \forall k = 1, \ldots, N\}.$$

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

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

Define $\hat{\mathbf{n}} = n_1 \times \cdots \times n_N = \text{card}(\mathcal{I}_{\mathbf{n}})$ and assume that the functional random field is observed on a subset $\mathcal{S}_{\mathbf{n}} \subset \mathcal{I}_{\mathbf{n}}$ with $\mathcal{I}_{\mathbf{n}} - \mathcal{S}_{\mathbf{n}}$ is a bounded set for $\hat{\mathbf{n}}$ large enough. We propose a nonparametric discrimination technique consisting in predicting the label $Y_{\mathbf{j}}$ of a new observation $X_{\mathbf{j}}$ based on observations in a vicinity of $\mathbf{j}$, say $\nu_{\mathbf{j}} \subset \mathcal{S}_{\mathbf{n}}$, not containing $\mathbf{j}$. Obviously, the temporal context ($N = 1$) corresponds to the Markovian case. The strong mixing condition that we will define later permits to assume that observations in nearby sites are more dependent than observations in distant sites. Thus, we wish to predict the label $Y_{\mathbf{j}}$ of a new observation $X_{\mathbf{j}}$ based on observations in the vicinity $\nu_{\mathbf{j}} \subset \mathcal{S}_{\mathbf{n}}$ defined above. This technique is used by Younso (2017a) for classification and by Biau and Cadre (2004) and Carbon, Francq and Tran (2007) for estimating the regression function based on finite-dimensional areal (lattice) data. Let $\nu_{\mathbf{j}} = \mathbf{j} + \nu$, where $\nu \subset \mathbb{Z}^N$ is a fixed bounded set of sites not containing 0 with $\text{Card}(\nu) = l$ (l is also the cardinal of each $\nu_{\mathbf{j}}$). We assume that $X_{(\mathbf{j})} = \{X_{\mathbf{i}} : \mathbf{i} \in \nu_{\mathbf{j}}\}$ is a random element of $\mathcal{E}^l$ represented by a vector of l random curves. We assume also that the components of $X_{(\mathbf{j})}$ are ordered according to an arbitrary order on indices, for example the lexicographic order. The pair $(X_{(\mathbf{j})}, Y_{\mathbf{j}})$ may be completely described by μ, the probability measure for $X_{(\mathbf{j})}$, and $\eta(x)$, the regression of $Y_{\mathbf{j}}$ on $X_{(\mathbf{j})} = x$. We create a classifier $g : \mathcal{E}^l \to \{0, 1\}$ mapping $X_{(\mathbf{j})}$ into the predicted label of $X_{\mathbf{j}}$. Observe that a path of $X_{\mathbf{j}}$ is represent by a single curve while a path of $X_{(\mathbf{j})}$ is represented by a set of l curves. One finds in Ramsay and Silverman (2005) a brief example of bivariate functional data. For a more general case, we refer to Jacques and Preda (2014) and Górecki, Krzysko and Wolynski (2015). The error rate, or risk, of a rule g is $L(g) = \mathbb{P}\{g(X_{(\mathbf{j})}) \neq Y_{\mathbf{j}}\}$. This is minimized by the rule

$$g^*(\mathbf{x}) = \begin{cases} 0 & \text{if } \mathbb{P}\{Y_{\mathbf{j}} = 0 | X_{(\mathbf{j})} = \mathbf{x}\} \ge \mathbb{P}\{Y_{\mathbf{j}} = 1 | X_{(\mathbf{j})} = \mathbf{x}\}, \\ 1 & \text{otherwise,} \end{cases}$$

whose error rate $L^* = L(g^*)$ is called the Bayes-optimal risk and g^* is called the Bayes rule. This optimal rule depends on the distribution of $(X_{(\mathbf{j})}, Y_{\mathbf{j}})$ which is generally unknown. We take $\mathcal{J}_{\mathbf{n}} = \{\mathbf{i} \in \mathcal{S}_{\mathbf{n}} : \nu_{\mathbf{i}} \subset \mathcal{S}_{\mathbf{n}}\}$ and we use the data $D_{\mathbf{n}} = \{(X_{\mathbf{i}}, Y_{\mathbf{i}}) : \mathbf{i} \in \mathcal{J}_{\mathbf{n}}\}$ to construct a classifier $g_{\mathbf{n}}(\mathbf{x})$. A classifier could be constructed by combining the principle of kernel rule

defined in Younso (2017a) in the finite-dimensional case and the functional kernel-based rule defined in Younso (2018) as follows:

$$g_{\mathbf{n}}(\mathbf{x}) = \begin{cases} 0 & \text{if } \sum\limits_{\mathbf{i}\in\mathcal{J}_{\mathbf{n}}} \mathbb{1}_{\{Y_{\mathbf{i}}=0, X_{(\mathbf{i})}\in B_{\mathbf{x},b}\}} \geq \sum\limits_{\mathbf{i}\in\mathcal{J}_{\mathbf{n}}} \mathbb{1}_{\{Y_{\mathbf{i}}=1, X_{(\mathbf{i})}\in B_{\mathbf{x},b}\}}, \\ 1 & \text{otherwise,} \end{cases} \tag{2.1}$$

where $\mathbb{1}_A$ denotes the indicator function of the set A, $b = b(\mathbf{n})$ the smoothing factor, is a strictly positive number tending to 0 when $\mathbf{n} \to \infty$ and $B_{\mathbf{x},b} = \{\mathbf{z} = (z_1, \ldots, z_l) \in \mathcal{E}^l, \rho(\mathbf{z}, \mathbf{x}) \leq b\}$ denotes the closed ball centered at $\mathbf{x} = (x_1, \ldots, x_l) \in \mathcal{E}^l$ with radius b and ρ is the metric (product metric) on $\mathcal{E}^l$. The formula (2.1) is easy to apply, but in order to prove the theoretical results, we will use the following formula

$$g_{\mathbf{n}}(\mathbf{x}) = \begin{cases} 0 & \text{if } \eta_{\mathbf{n}}(\mathbf{x}) \leq \dfrac{\sum_{\mathbf{i}\in\mathcal{J}_{\mathbf{n}}}(1-Y_{\mathbf{i}})\mathbb{1}_{\{X_{(\mathbf{i})}\in B_{\mathbf{x},b}\}}}{\hat{\mathbf{n}}\mu(B_{\mathbf{x},b})}, \\ 1 & \text{otherwise,} \end{cases} \tag{2.2}$$

where

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

The best we can expect from $g_{\mathbf{n}}(\mathbf{x})$ is to achieve the Bayes risk. Let

$$L_{\mathbf{n}} = L(g_{\mathbf{n}}) = \mathbb{P}\{g_{\mathbf{n}}(X_{(\mathbf{j})}) \neq Y_{\mathbf{j}}\}$$

be the error probability of $g_{\mathbf{n}}(\mathbf{x})$. The classifier $g_{\mathbf{n}}(\mathbf{x})$ is said to be consistent if

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

Since $\mathcal{I}_{\mathbf{n}} - \mathcal{S}_{\mathbf{n}}$ is a bounded for $\hat{\mathbf{n}}$ large enough, $g_{\mathbf{n}}$ has the same asymptotic behavior on $\mathcal{J}_{\mathbf{n}}$ as on $\mathcal{I}_{\mathbf{n}}$.

With the same arguments as in (Abraham, Biau and Cadre (2006), Section 2), the classifier $g_{\mathbf{n}}(\mathbf{x})$ is not universally consistent in general metric spaces. In this paper, we investigate the consistency of this classifier under mild conditions.

3 General assumptions and main results

In this section, we propose a measure of spatial dependence. Then, we introduce some assumptions and preliminaries needed to prove the main result. Finally, we state the main result on consistency of the classifier.

The three classic types of spatial data structures (geostatistical data, point patterns, and areal data) can be combined with functional data as it is shown in the examples of each situation provided in Delicado et al. (2010). In the present paper, we focus on the case of areal (lattice) data when functional random fields are observed on a regular grid containing a finite number of sites whose whole constitutes the entire study region. There are many ways to model spatial dependence. For example, it may be modeled through the covariance function (geostatistical models, see Romano, Balzanella and Verde (2010)), through the set of conditional distributions of each spatial observation given all others (spatial Markovian models, see Carlo, Paolo and Roberto (2017)) or through mixing conditions (weak dependence models, see Younso (2018)). For spatial processes in a regular lattice, mixing conditions (like α, β and ϕ-mixing) have become usual structures for modelling spatial dependence in nonparametric estimation due to their ease of use to obtain asymptotic results, see for example Dabo-Niang and Yao (2007, 2013), Ternynck (2014) and Biau and Cadre (2004) for the multivariate

case. We focus on the α-mixing (or strong mixing) notion, which is one of the most general among the different mixing structures introduced in the literature, see Rosenblatt (1956) and Ibragimov (1962). Thus, in order to obtain asymptotic results, we will assume throughout the paper, the functional random field $\{(X_{\mathbf{i}}, Y_{\mathbf{i}})\}_{\mathbf{i}\in\mathbb{Z}^N}$ satisfies the following mixing condition: there exists a function $\varphi : \mathbb{R} \to \mathbb{R}^+$ with $\varphi(t) \searrow 0$ as $t \to \infty$, such that whenever $E, E' \subset \mathbb{Z}^N$ with finite cardinals,

$$\begin{aligned}\alpha\big(\mathcal{B}(E), \mathcal{B}(E')\big) &:= \sup\big\{\big|\mathbb{P}(A \cap C) - \mathbb{P}(A)\mathbb{P}(C)\big|, A \in \mathcal{B}(E), C \in \mathcal{B}(E')\big\} \\ &\leq \psi\big(\mathrm{card}(E), \mathrm{card}(E')\big)\varphi\big(\mathrm{dist}(E, E')\big),\end{aligned}$$

where $\mathcal{B}(E)$ (*resp.* $\mathcal{B}(E')$) denotes the Borel σ-field generated by $(X_{\mathbf{i}}, Y_{\mathbf{i}})_{\mathbf{i}\in E}$ (*resp.* $(X_{\mathbf{i}}, Y_{\mathbf{i}})_{\mathbf{i}\in E'}$), $\mathrm{card}(E)$ (*resp.* $\mathrm{card}(E')$) the cardinality of E (*resp.* E'), $\mathrm{dist}(E, E')$ the Euclidean distance between E and E', and $\psi : \mathbb{N}^2 \to \mathbb{R}_+$ is a symmetric positive function which is nondecreasing in each variable. Throughout the paper, it will be also assumed for simplicity that ψ satisfies

$$\psi(n, m) \leq \min\{n, m\}, \quad \forall n, m \in \mathbb{N}. \tag{3.1}$$

If $h \equiv 1$, the random field is called strongly mixing. They are satisfied by many spatial models. The above mixing condition is used for example by Ternynck (2014) to estimate the regression function for functional data.

Now, for convenience, we introduce the notion of covering numbers (see Kolmogorov and Tihomirov (1961)). For a given subset $\mathcal{G}$ of the metric space $(\mathcal{E}^l, \rho)$, the ε-covering number is defined by

$$\mathcal{N}(\varepsilon, \mathcal{G}, \rho) = \inf\left\{k \geq 1 : \exists \mathbf{x}_1, \ldots, \mathbf{x}_k \in \mathcal{E}^l \text{ with } \mathcal{G} \subset \bigcup_{i=1}^{k} S_{\mathbf{x}_i,\varepsilon}\right\},$$

where $S_{\mathbf{x},\varepsilon} = \{\mathbf{z} \in \mathcal{E}^l,\ \rho(\mathbf{z}, \mathbf{x}) < \varepsilon\}$ is the open ball centered at $\mathbf{x} = (x_1, \ldots, x_l) \in \mathcal{E}^l$ with radius $\varepsilon > 0$. Note that as a function of ε, $\mathcal{N}(\varepsilon, \mathcal{G}, \rho)$ is nonincreasing, piecewise-constant and right-continuous. The set $\mathcal{G}$ is called totally bounded if $\mathcal{N}(\varepsilon, \mathcal{G}, \rho) < \infty$ for all $\varepsilon > 0$. In particular, every relatively compact set is totally bounded and all totally bounded sets are bounded, see Biau, Cérou and Guyader (2010) for further details. Now, we introduce the following assumptions.

Assumption 1. There exists a sequence $(\mathcal{F}_k)_{k\geq 1}$ of totally bounded subsets of $\mathcal{E}^l$ such that $\mathcal{F}_k \subset \mathcal{F}_{k+1}$ for all $k \geq 1$ and $\mu(\bigcup_{k\geq 1} \mathcal{F}_k) = 1$.

Assumption 2. For all $\varepsilon_1 \in]0, 1]$ and $\mathbf{i} \neq \mathbf{j}$ with $\nu_{\mathbf{i}} \cap \nu_{\mathbf{j}} = \phi$, $\mathbb{P}((X_{(\mathbf{i})}, X_{(\mathbf{j})}) \in B_{\mathbf{x},b} \times B_{\mathbf{x},b}) \leq C[\mu(B_{\mathbf{x},b})]^{1+\varepsilon_1}$, for all $\mathbf{x} \in \mathcal{E}^l$ and some $C > 0$.

Assumption 3. The Lebesgue differentiation theorem holds, that is, for every $\varepsilon > 0$,

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

where $d\mathbf{x} = dx_1 \cdots dx_l$.

Remark 3.1. Note that Assumption 1 is always true whenever the space $(\mathcal{E}, d)$ is separable (since the finite product of separable spaces is separable), see Abraham, Biau and Cadre (2006) and Kulkarni and Posner (1995) for various examples in the univariate case. Assumption 2, used by Ternynck (2014), can be linked with the classical local dependence condition

met in the literature of the finite-dimensional case when $X_{(\mathbf{i})}$ and $(X_{(\mathbf{i})}, X_{(\mathbf{j})})$ admit, respectively, the densities f and $f_{\mathbf{i},\mathbf{j}}$ (see Carbon, Francq and Tran (2007)). Assumption 3 automatically holds for any integrable function η in finite-dimensional spaces. In a general setting, it holds if, for instance, η is μ-continuous (see Cérou and Guyader (2006) for the univariate functional case).

Suppose that the training data $D_{\mathbf{n}}$ is drawn from an arithmetically α-mixing functional random field in the sense that there exist $C > 0$ and $\theta > 0$ such that

$$\varphi(t) \leq Ct^{-\theta} \quad \text{for all } t \in \mathbb{R}_+^*. \tag{3.2}$$

The notion $\mathcal{G}^c$ stands for the complement of any subset $\mathcal{G}$ of $\mathcal{E}^l$. For simplicity of notation, we write $\mathcal{N}_k(\varepsilon)$ instead of $\mathcal{N}(\varepsilon, \mathcal{E}_k, \rho)$. Before we state the main result, we introduce the following useful lemmas.

Lemma 3.1. *Let Z_1 and Z_2 be two $\mathbb{R}$-valued bounded random variables. Then, we have $|\operatorname{cov}(Z_1, Z_2)| \leq 4\|Z_1\|_\infty \|Z_2\|_\infty \alpha(\sigma(Z_1), \sigma(Z_2))$, where $\|.\|_\infty$ denotes the supremum norm and $\sigma(Z_i)$ denotes the Borel σ-field generated by Z_i for $i = 1, 2$.*

The proof of Lemma 3.1 is given in Rio (2000). The following lemma is a consequence of Assumption 3 and the Lebesgue dominated convergence theorem.

Lemma 3.2. *Assume that Assumption 3 holds. If $b \to 0$ as $\mathbf{n} \to \infty$, then,*

$$\int_{\mathcal{E}^l} |\eta(\mathbf{x}) - \mathbb{E}\eta_{\mathbf{n}}(\mathbf{x})| \mu(d\mathbf{x}) = \int_{\mathcal{E}^l} \left| \eta(\mathbf{x}) - \frac{\int_{B_{\mathbf{x},b}} \eta(\mathbf{t}) \mu(d\mathbf{t})}{\mu(B_{\mathbf{x},b})} \right| \mu(d\mathbf{x}) \longrightarrow 0.$$

For the proof of the following lemma, we refer to Abraham, Biau and Cadre (2006).

Lemma 3.3. *Assume that $(\mathcal{F}_k)_{k \geq 1}$ is a sequence of totally bounded subsets of $\mathcal{E}^l$. Let k be a fixed positive integer. Then, for every $b > 0$,*

$$\int_{\mathcal{F}_k} \frac{1}{\mu(B_{\mathbf{x},b})} \mu(d\mathbf{x}) \leq \mathcal{N}_k(b/2).$$

Lemma 3.4. *Assume that Assumption 3 holds and $D_{\mathbf{n}}$ is drawn from an α-mixing functional random field such that (3.1) and (3.2) with $\theta > 2N$. Let k be a fixed positive integer. Then, for all $\mathbf{n} \in (\mathbb{N}^*)^N$,*

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

As mentioned in Section 2, a desirable property for classifiers is consistency because it shows the performance of the classifier comparing to the optimal Bayes classifier. The main result on the consistency is stated in the following theorem. This theorem is an extension to functional data of the result of Younso (2017a) in finite-dimensional setting.

Theorem 3.1 (Consistency). *Assume that $D_{\mathbf{n}}$ is is drawn from an α-mixing functional random field such that (3.1) and (3.2), and $(\mathcal{F}_k)_{k \geq 1}$ is a sequence of totally bounded subsets of $\mathcal{E}^l$. If Assumptions 1–3 are satisfied, and $b \to 0$ and for every $k \geq 1$, $\frac{\mathcal{N}_k(b/2)}{\hat{\mathbf{n}}} \longrightarrow 0$ as $\mathbf{n} \to \infty$, then, for $\theta > 2N$,*

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

Observe that, for $N = 1$, similar assumptions on the smoothing factor b are used by Abraham, Biau and Cadre (2006) in the classical independent case.

4 Simulation studies

In this section, we make two empirical comparisons. In order to make the required comparisons, we will conduct two simulation studies.

Comparison with classifier of Younso (2018)

The moving window classifier either in the classical spatial case studied by Younso (2018) or in the new version (2.1) can be extended without difficulty from the binary case to the multiclass case when $Y_{\mathbf{j}}$ takes values in the set of labels $\{0, \ldots, M\}$ where $M \geq 1$. In the classical multiclass case, the Bayes rule can be computed by

$$g^*(x) = \arg \max_{0 \leq k \leq M} \mathbb{P}(Y_{\mathbf{j}} = k | X_{\mathbf{j}} = x)$$

and the moving window rule is given by

$$g_{\mathbf{n}}(x) = \arg \max_{0 \leq k \leq M} \sum_{\mathbf{i} \in \mathcal{I}_{\mathbf{n}}} \mathbb{1}_{\{Y_{\mathbf{i}} = k, X_{\mathbf{i}} \in B_{x,h}\}} \tag{4.1}$$

where $x \in \mathcal{E}$ and $B_{x,h}$ is the closed ball in $\mathcal{E}$ centered at x with radius $h > 0$. In the new multiclass case, the Bayes rule can be computed by

$$g^*(\mathbf{x}) = \arg \max_{0 \leq k \leq M} \mathbb{P}(Y_{\mathbf{j}} = k | X_{(\mathbf{j})} = \mathbf{x})$$

and the moving window rule is given by

$$g_{\mathbf{n}}(\mathbf{x}) = \arg \max_{0 \leq k \leq M} \sum_{\mathbf{i} \in \mathcal{J}_{\mathbf{n}}} \mathbb{1}_{\{Y_{\mathbf{i}} = k, X_{(\mathbf{i})} \in B_{\mathbf{x},b}\}}. \tag{4.2}$$

Our aim in this section is to compare the performance of the two classifiers (4.1) and (4.2) based on simulated samples. We use the R statistical programming environment to run a simulation studies for $N = 2$. The simulated data are located on the area $\mathcal{I}_{(n,n)} = \{\mathbf{i} = (i, j) \in \mathbb{Z}^2 : 1 \leq i, j \leq n\}$ and let $\{(X_{(i,j)}, Y_{(i,j)})\}$ the field of interest. We propose to investigate the performance of the two methods in the following simulated scenario. For each $\mathbf{i} \in \mathcal{I}_{\mathbf{n}}$ and $t \in [0, 1]$, we generate pairs $(X_{\mathbf{i}}(t), Y_{\mathbf{i}})$ via the following model inspired by Jiang and Serban (2012):

$$\begin{aligned}
\text{class}(Y_{\mathbf{i}} = 0) &: X_{\mathbf{i}}(t) = f_0(t) + \varepsilon_{\mathbf{i}}, \\
\text{class}(Y_{\mathbf{i}} = 1) &: X_{\mathbf{i}}(t) = f_1(t) + \varepsilon_{\mathbf{i}}, \\
\text{class}(Y_{\mathbf{i}} = 2) &: X_{\mathbf{i}}(t) = f_2(t) + \varepsilon_{\mathbf{i}},
\end{aligned}$$

where $\varepsilon_{\mathbf{i}}$'s are standard Gaussian and spatially correlated according to the covariance structure $\mathrm{c}(\|\mathbf{u}\|) \sim \|\mathbf{u}\|^{-5}$ for all $\mathbf{u} \in \mathbb{Z}^2$ with $\mathbf{u} \neq 0$, and $\|.\|$ is the Euclidean norm on $\mathbb{Z}^2$, and $f_0(t) = e^t \cos(t)$, $f_1(t) = \cos(5\pi t/2)$ and $f_2(t) = -(f_0(t) + f_1(t))/2$. Figure 1 shows the plots of the functions f_0, f_1 and f_2. It is important to mention that $\{\varepsilon_{\mathbf{i}}\}_{\mathbf{i} \in \mathcal{I}_{\mathbf{n}}}$ are observations of an α-mixing random field since any Gaussian random field with covariance function $\mathrm{c}(\|\mathbf{u}\|)$ converges to zero as $\|\mathbf{u}\| \to \infty$ is α-mixing. We suppose that the function space $\mathcal{E}$ on the interval $[0, 1]$ is endowed with the metric d (between z_1 and z_2) defined by

$$d(z_1, z_2) = \int_0^1 |z_1(t) - z_2(t)| \, dt.$$

Let

$$\rho(\mathbf{z}, \mathbf{z}') = \max_{1 \leq i \leq l} d(z_i, z_i'),$$

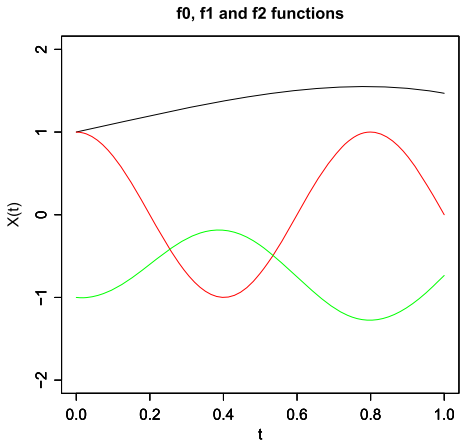

Figure 1 *Plots of the function* $f_0(t)$ (*black line*), *the function* $f_1(t)$ (*red line*) *and the function* $f_2(t)$ (*green line*).

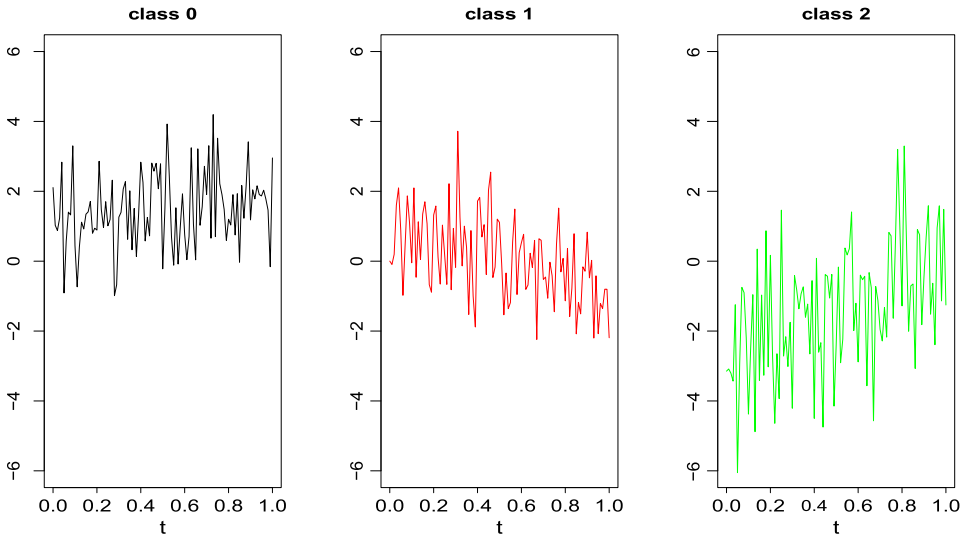

Figure 2 *Three typical curves with label* 0 (*black*), *label* 1 (*red*) *and label* 2 (*green*).

for any $\mathbf{z} = (z_1, \ldots, z_l)$, $\mathbf{z}' = (z'_1, \ldots, z'_l) \in \mathcal{E}^l$. The curve $X_{\mathbf{i}}(t)$ will be colored by *black* if $Y_{\mathbf{i}} = 0$, by *red* if $Y_{\mathbf{i}} = 1$ and by *green* if $Y_{\mathbf{i}} = 2$. We generate simulations of the random field $\{X_{\mathbf{i}}(t)\}_{\mathbf{i}\in\mathbb{Z}^2}$ on the rectangular region $\mathcal{I}_{\mathbf{n}}$ where $\mathbf{n} = (n, n)$ for $n \geq 1$. Figure 2 displays three typical realizations of the $X_{\mathbf{i}}$'s and Figure 3 displays a sample of size 625 simulated on $\mathcal{I}_{(25,25)} = \{(i, j), 1 \leq i, j \leq 25\}$. Since the theoretical results are related to the consistency, it is natural to consider training samples with increasing sizes. For this aim, we generate, for each sample size n^2 ($n = 50, 80, 100$), 100 training samples on $\mathcal{J}_{\mathbf{n}}$ and 100 corresponding test samples of size 100 on

$$\mathcal{M} = \{(4k, 4l), 1 \leq k, l \leq 10\}.$$

In the classical case, the set $\mathcal{M}$ contains observed sites being not classified. In the new case, we suppose that $\mathcal{M}$ contains non observed sites. Thus,

$$\mathcal{J}_{\mathbf{n}} = \mathcal{I}_{\mathbf{n}} - \big\{\{\nu_{\mathbf{j}} \cup \{\mathbf{j}\}, \mathbf{j} \in \mathcal{M}\} \cup \{(1, j), (k, 1), (n, l), (m, n) : 1 \leq j, k, l, m \leq n\}\big\},$$

where $\nu_{\mathbf{j}}$ is the vicinity of $\mathbf{j}$ including only the eight sites surrounding $\mathbf{j}$. Figure 4 shows the results of one replication for $n = 100$. In order for the classifier (4.1) to be usable in our case,

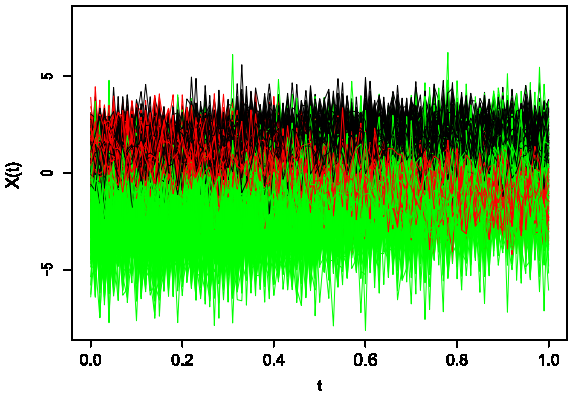

Figure 3 *Sample of* 625 *simulated curves on the region* $\mathcal{I}_{(25,25)}$.

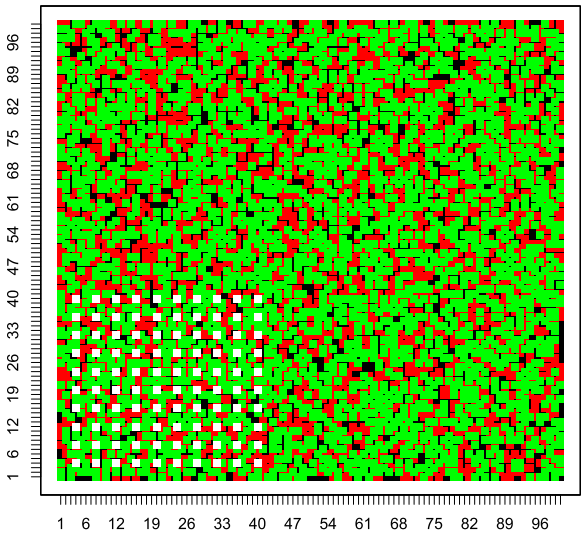

Figure 4 *The non-observed sites are blank and the observed sites are colored black, red or green.*

we have to adjust it slightly where it becomes

$$g_{\mathbf{n}}(x) = \arg \max_{0 \le k \le M} \sum_{\mathbf{i} \in \mathcal{I}_{\mathbf{n}} - \mathcal{M}} \mathbb{1}_{\{Y_{\mathbf{i}} = k, X_{\mathbf{i}} \in B_{x,b}\}}. \tag{4.3}$$

Since $\mathcal{M}$ is a bounded, $g_{\mathbf{n}}$ has the same asymptotic behavior on $\mathcal{I}_{\mathbf{n}}$ as on $\mathcal{I}_{\mathbf{n}} - \mathcal{M}$. We use the classifiers (4.2) and (4.3), independently of each other, to predict the label of each $X_{\mathbf{j}}$ for $\mathbf{j} \in \mathcal{M}$. In the classical case, we use the observation $X_{\mathbf{j}}$ itself to predict its class while in the new case, we use observations in nearby sites. In each replication, the proposed classifiers are determined on the basis of the training sample at hand. For each classifier, the optimal bandwidth $\widehat{h}_{\text{opt}}$ ($\widehat{b}_{\text{opt}}$) is chosen by minimizing the cross-validation criterion. The misclassification error rate (ER) is evaluated based on the associated test sample for each classifier. Table 1 and Table 2 report the average error rate (AER), obtained by averaging the error rates associated with the corresponding 100 test samples for the classifier (4.2) and (4.3). Table 1 shows that the estimated optimal bandwidth and the average 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. In comparing the above two tables, we observe that the corresponding error values in the two tables begin to be close as n increases. This supports

Table 1 *Estimated optimal bandwidths and average error rates corresponding to the classifier* (4.2) *with samples of different sizes*

n	50	80	100
$\widehat{b}_{\text{opt}}$	2.9	2.77	2.15
AER	29.4%	18.5%	10.6%

Table 2 *Estimated optimal bandwidths and average error rates corresponding to the classifier* (4.3) *with samples of different sizes*

n	50	80	100
$\widehat{h}_{\text{opt}}$	2.2	1.44	1.42
AER	18.22%	13.23%	6.13%

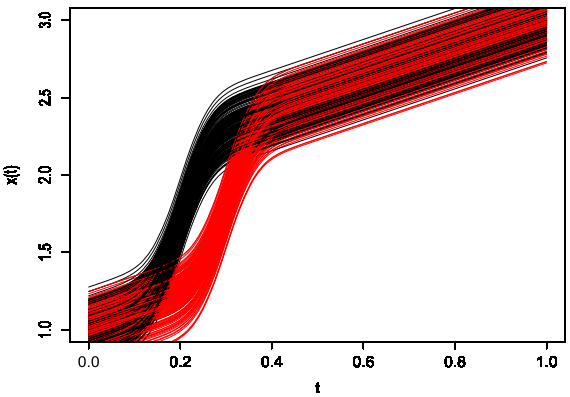

Figure 5 *Sample of* 400 *simulated curves on the region* $\mathcal{I}_{(20,20)}$ *with label* 0 (*black*) *and label* 1 (*red*).

the possibility of using the classifier (4.2) as an alternative to the classifier (4.3) when there are missing observations.

Comparison with (SVM) classifier

We simulate random curves from the model

$$\text{class}(Y_{\mathbf{i}} = 0) : X_{\mathbf{i}}(t) = \Phi\big(20(t - 0.2)\big) + \varepsilon_{\mathbf{i}},$$
$$\text{class}(Y_{\mathbf{i}} = 1) : X_{\mathbf{i}}(t) = \Phi\big(20(t - 0.3)\big) + \varepsilon_{\mathbf{i}}$$

with Φ is the standard normal distribution function and $\varepsilon_{\mathbf{i}}$'s are standard Gaussian and spatially correlated according to the covariance structure defined above. Figure 5 shows a sample of size 400 on the region $\mathcal{I}_{(20,20)}$. Firstly, for the above set of simulated curves, we use the classifier (2.1) to re-classify the set of curves located on the region test $\mathcal{M} = \{(4k, 4l), 1 \leq k, l \leq 10\}$ using the remaining data located on $\mathcal{I}_{(20,20)} - \mathcal{M}$. Secondly, we use the (SVM) classifier to re-classify the curves issued from the same region test $\mathcal{M}$ using the remaining data. For Implementing support vector machine in R programming language, we can use the package $e1071$. For the classifier (2.1), we have $\widehat{b}_{\text{opt}} = 1.7$ with the misclassification error rate ER $= 14\%$. On the other hand, by implementation of SVM in R programming language, we have ER $= 17\%$. Finally, it results that the classifier (2.1) proceeds well comparing to the (SVM) classifier.

5 Proofs

Proof of Lemma 3.4. By Cauchy–Schwarz inequality, for a fixed $\mathbf{x} \in \mathcal{E}^l$,

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

where $S_{\mathbf{n}}(\mathbf{x}) = \frac{1}{(\hat{\mathbf{n}}\mu(B_{\mathbf{x},b}))^2}\sum_{\mathbf{i}\neq\mathbf{j}} |\mathrm{cov}(\Delta_{\mathbf{i}}, \Delta_{\mathbf{j}})|$ and $\Delta_{\mathbf{i}} = Y_{\mathbf{i}}\mathbb{1}_{\{X_{(\mathbf{i})}\in B_{\mathbf{x},b}\}}$ for all $\mathbf{i} \in \mathcal{J}_{\mathbf{n}}$. Now, since $|Y_{\mathbf{i}}| \leq 1$, we obtain

$$\begin{aligned}\mathbb{E}|\eta_{\mathbf{n}}(\mathbf{x}) - \mathbb{E}\eta_{\mathbf{n}}(\mathbf{x})| &\leq \left(\frac{\mathbb{E}(\mathbb{1}_{\{X_{(\mathbf{1})}\in B_{\mathbf{x},b}\}})}{\hat{\mathbf{n}}(\mu(B_{\mathbf{x},b}))^2} + S_{\mathbf{n}}(\mathbf{x})\right)^{1/2} \\ &\leq \left(\frac{1}{\hat{\mathbf{n}}\mu(B_{\mathbf{x},b})} + S_{\mathbf{n}}(\mathbf{x})\right)^{1/2}. \qquad (5.1)\end{aligned}$$

Let $u_{\mathbf{n}}$ a sequence of positives such that $u_{\mathbf{n}} \to \infty$ as $\mathbf{n} \to \infty$. We denote

$$\begin{aligned}\mathcal{S}_1 &= \big\{(\mathbf{i},\mathbf{j}) \in \mathcal{J}_{\mathbf{n}}^2 : \nu_{\mathbf{i}} \cap \nu_{\mathbf{j}} \neq \phi\big\}, \\ \mathcal{S}_2 &= \big\{(\mathbf{i},\mathbf{j}) \in \mathcal{J}_{\mathbf{n}}^2 : \nu_{\mathbf{i}} \cap \nu_{\mathbf{j}} = \phi \text{ and } 0 < \|\mathbf{i}-\mathbf{j}\| \leq u_{\mathbf{n}}\big\}\}, \\ \mathcal{S}_3 &= \big\{(\mathbf{i},\mathbf{j}) \in \mathcal{J}_{\mathbf{n}}^2 : \nu_{\mathbf{i}} \cap \nu_{\mathbf{j}} = \phi \text{ and } \|\mathbf{i}-\mathbf{j}\| > u_{\mathbf{n}}\big\}\}.\end{aligned}$$

Since $l = \mathrm{card}(\nu_{\mathbf{j}})$, we have

$$\begin{aligned}\mathrm{card}(\mathcal{S}_1) &\leq \sum_{\mathbf{i}\in\mathcal{J}_{\mathbf{n}}} \mathrm{card}\{\mathbf{j} : \nu_{\mathbf{i}} \cap \nu_{\mathbf{j}} \neq \phi\} \leq \hat{\mathbf{n}} \sum_{\mathbf{k}\in\nu_{\mathbf{i}}} \mathrm{card}\{\mathbf{j} : \mathbf{k} \in \nu_{\mathbf{j}} \neq \phi\} \\ &\leq \hat{\mathbf{n}} l^2. \qquad (5.2)\end{aligned}$$

Furthermore, we have

$$\mathrm{card}(\mathcal{S}_2) \leq \sum_{\mathbf{i}\in\mathcal{J}_{\mathbf{n}}} \mathrm{card}\big\{\mathbf{j} \in \mathcal{J}_{\mathbf{n}} : 0 < \|\mathbf{i}-\mathbf{j}\| \leq u_{\mathbf{n}}\big\} \leq \hat{\mathbf{n}}(2u_{\mathbf{n}})^N. \qquad (5.3)$$

We can write

$$S_{\mathbf{n}}(\mathbf{x}) = J_{\mathbf{n},1}(\mathbf{x}) + J_{\mathbf{n},2}(\mathbf{x}) + J_{\mathbf{n},3}(\mathbf{x}),$$

where

$$J_{\mathbf{n},k}(\mathbf{x}) = \frac{1}{(\hat{\mathbf{n}}\mu(B_{\mathbf{x},b}))^2} \sum_{(\mathbf{i},\mathbf{j})\in\mathcal{S}_k} \big|\mathrm{cov}(\Delta_{\mathbf{i}}, \Delta_{\mathbf{j}})\big|,$$

for $k = 1, 2, 3$. Since $|Y_{\mathbf{1}}| \leq 1$, using Cauchy–Schwarz inequality together with (5.2), we have

$$J_{\mathbf{n},1}(\mathbf{x}) \leq \frac{\mathrm{card}(\mathcal{S}_1)\mathbb{E}(Y_{\mathbf{1}}\mathbb{1}_{\{X_{(\mathbf{1})}\in B_{\mathbf{x},b}\}})^2}{(\hat{\mathbf{n}}\mu(B_{\mathbf{x},b}))^2} \leq \frac{l^2}{\hat{\mathbf{n}}\mu(B_{\mathbf{x},b})}. \qquad (5.4)$$

Now, for $0 < \|\mathbf{i}-\mathbf{j}\| \leq u_{\mathbf{n}}$ with $\nu_{\mathbf{i}} \cap \nu_{\mathbf{j}} = \phi$, Assumption 2 yields

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

where $C > 0$ is a generic constant and $0 < \varepsilon_1 \leq 1$. Since $\mu(B_{\mathbf{x},b}) \leq 1$, then $|\operatorname{cov}(\Delta_{\mathbf{i}}, \Delta_{\mathbf{j}})| \leq C\{\mu(B_{\mathbf{x},b})\}^{1+\varepsilon_1}$. Thus, by (5.3), we have

$$J_{\mathbf{n},2}(\mathbf{x}) = \frac{1}{(\hat{\mathbf{n}}\mu(B_{\mathbf{x},b}))^2} \sum_{(\mathbf{i},\mathbf{j})\in\mathcal{S}_2} |\operatorname{cov}(\Delta_{\mathbf{i}}, \Delta_{\mathbf{j}})|$$
$$\leq C\frac{\operatorname{card}(\mathcal{S}_2)}{\hat{\mathbf{n}}^2(\mu(B_{\mathbf{x},b}))^{1-\varepsilon_1}} \leq \frac{Cu_{\mathbf{n}}^N}{\hat{\mathbf{n}}(\mu(B_{\mathbf{x},b}))^{1-\varepsilon_1}}.$$

Choosing $u_{\mathbf{n}} = \{\mu(B_{\mathbf{x},b})\}^{-\varepsilon_1/N}$, we get

$$J_{\mathbf{n},2}(\mathbf{x}) \leq \frac{C}{\hat{\mathbf{n}}\mu(B_{\mathbf{x},b})}. \tag{5.5}$$

Let $\tilde{r} = \max\{\|\mathbf{i}-\mathbf{j}\|, \mathbf{i}, \mathbf{j} \in \nu\}$ be the diameter of $\nu \subset \mathbb{Z}^N$. By Lemma 3.1 in Younso (2017b), we have for each $\mathbf{i}, \mathbf{j} \in \mathcal{J}_{\mathbf{n}}$,

$$\operatorname{dist}(\nu_{\mathbf{i}}, \nu_{\mathbf{j}}) \geq \max\{\|\mathbf{i}-\mathbf{j}\| - \tilde{r}, 0\}.$$

As a consequence, if $\|\mathbf{i}-\mathbf{j}\| > u_{\mathbf{n}}$ with $\nu_{\mathbf{i}} \cap \nu_{\mathbf{j}} = \phi$, since $|Y_{\mathbf{1}}| \leq 1$, by Lemmas 3.1 and (3.1), we get

$$|\operatorname{cov}(\Delta_{\mathbf{i}}, \Delta_{\mathbf{j}})| \leq 4\alpha(\operatorname{dist}(\nu_{\mathbf{i}}, \nu_{\mathbf{j}})) \leq 4\min\{l, l\}\varphi(\operatorname{dist}(\nu_{\mathbf{i}}, \nu_{\mathbf{j}}))$$
$$\leq 4l\varphi(\max\{\|\mathbf{i}-\mathbf{j}\| - \tilde{r}, 0\}). \tag{5.6}$$

Using the convention $\varphi(t) = \varphi(0)$ for $t < 0$ together with (5.6), we have

$$J_{\mathbf{n},3}(\mathbf{x}) = \frac{1}{(\hat{\mathbf{n}}\mu(B_{\mathbf{x},b}))^2} \sum_{(\mathbf{i},\mathbf{j})\in\mathcal{S}_3} |\operatorname{cov}(\Delta_{\mathbf{i}}, \Delta_{\mathbf{j}})|$$
$$\leq \frac{4l}{(\hat{\mathbf{n}}\mu(B_{\mathbf{x},b}))^2} \sum_{\|\mathbf{i}-\mathbf{j}\|\geq u_{\mathbf{n}}} \varphi(\|\mathbf{i}-\mathbf{j}\| - \tilde{r}) \leq \frac{4l}{\hat{\mathbf{n}}(\mu(B_{\mathbf{x},b}))^2} \sum_{i\geq u_{\mathbf{n}}} i^{N-1}\varphi(i-\tilde{r}).$$

Since $u_{\mathbf{n}} - \tilde{r} > u_{\mathbf{n}}/2$ for $\hat{\mathbf{n}}$ large enough, we can write

$$J_{\mathbf{n},3}(\mathbf{x}) \leq \frac{4l}{\hat{\mathbf{n}}(\mu(B_{\mathbf{x},b}))^2} \sum_{i\geq u_{\mathbf{n}}/2} i^{N-1}\varphi(i) \leq C\int_{u_{\mathbf{n}}/2-1}^{\infty} t^{N-\theta-1}\,dt \leq Cu_{\mathbf{n}}^{N-\theta} \tag{5.7}$$

since by assumption $\varphi(i) \leq Ci^{-\theta}$ for some $\theta > 2N$. If we choose $u_{\mathbf{n}} = \{\mu(B_{x,b})\}^{-\varepsilon_1/N}$ and $N/(\theta - N) < \varepsilon_1 \leq 1$, the inequality (5.7) yields

$$J_{\mathbf{n},3}(\mathbf{x}) \leq \frac{Cu_{\mathbf{n}}^{N-\theta}}{\hat{\mathbf{n}}(\mu(B_{\mathbf{x},b}))^2} \leq \frac{C}{\hat{\mathbf{n}}\mu(B_{\mathbf{x},b})}. \tag{5.8}$$

Using (5.4), (5.5) and (5.8), we get $S_{\mathbf{n}}(\mathbf{x}) \leq \frac{C}{\hat{\mathbf{n}}\mu(B_{\mathbf{x},b})}$. Thus, by the inequality (5.1), we obtain

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

According to Fubini's theorem, Jensens's inequality and Lemma 3.4, we get

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

□

Proof of Theorem 3.1. By Theorem 2.3 in Devroye, Györfi and Lugosi (1996) whose extension to our infinite dimensional case is straightforward, Theorem 3.1 will be proved if we show that

$$\mathbb{E}\int_{\mathcal{E}^l}|\eta(\mathbf{x})-\eta_{\mathbf{n}}(\mathbf{x})|\mu(d\mathbf{x})\to 0\quad\text{as }\mathbf{n}\to\infty.$$

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

$$\begin{aligned}
&\mathbb{E}\int_{\mathcal{E}^l}|\eta(\mathbf{x})-\eta_{\mathbf{n}}(\mathbf{x})|\mu(d\mathbf{x})\\
&=\mathbb{E}\int_{\mathcal{F}_k}|\eta(\mathbf{x})-\eta_{\mathbf{n}}(\mathbf{x})|\mu(d\mathbf{x})+\mathbb{E}\int_{\mathcal{F}_k^c}|\eta(\mathbf{x})-\eta_{\mathbf{n}}(\mathbf{x})|\mu(d\mathbf{x})\\
&\leq\int_{\mathcal{F}_k}|\eta(\mathbf{x})-\mathbb{E}\eta_{\mathbf{n}}(\mathbf{x})|\mu(d\mathbf{x})+\mathbb{E}\int_{\mathcal{F}_k}|\eta_{\mathbf{n}}\mathbf{x})-\mathbb{E}\eta_{\mathbf{n}}(\mathbf{x})|\mu(d\mathbf{x})+2\mu(\mathcal{F}_k^c).
\end{aligned}$$

Consequently, according to Lemma 3.4, we get the following inequality

$$\begin{aligned}
&\mathbb{E}\int_{\mathcal{E}^l}|\eta(\mathbf{x})-\eta_{\mathbf{n}}(\mathbf{x})|\mu(d\mathbf{x})\\
&\leq\int_{\mathcal{E}^l}|\eta(\mathbf{x})-\mathbb{E}\eta_{\mathbf{n}}(\mathbf{x})|\mu(dx)+C\left(\frac{1}{\hat{\mathbf{n}}}\mathcal{N}_k\left(\frac{b}{2}\right)\right)^{1/2}+2\mu(\mathcal{F}_k^c).
\end{aligned}$$

By Lemma 3.2 and the assumptions on b, we obtain for every $k\geq 1$,

$$\limsup_{\mathbf{n}\to\infty}\mathbb{E}\int_{\mathcal{E}^l}|\eta(\mathbf{x})-\eta_{\mathbf{n}}(\mathbf{x})|\mu(d\mathbf{x})\leq 2\mu(\mathcal{F}_k^c).$$

If we let k go to infinity, Assumption 1 yields the proof of the theorem. □

Acknowledgments

The author thanks the anonymous referees and the editor for their valuable suggestions that helped to improve this paper.

References

Abraham, C., Biau, G. and Cadre, B. (2006). On the kernel rule for function classification. *Annals of the Institute of Statistical Mathematics* **58**, 619–633. MR2327897 https://doi.org/10.1007/s10463-006-0032-1

Berlinet, A., Biau, G. and Rouvière, L. (2008). Functional supervised classification with wavelets. *Annales de l'ISUP* **52**, 61–80. MR2435041

Biau, G., Bunea, F. and Wegkamp, M. H. (2005). Functional classification in Hilbert spaces. *IEEE Transactions on Information Theory* **51**, 2163–2172. MR2235289 https://doi.org/10.1109/TIT.2005.847705

Biau, G. and Cadre, B. (2004). Nonparametric spatial prediction. *Statistical Inference for Stochastic Processes* **3**, 327–349. MR2111294 https://doi.org/10.1023/B:SISP.0000049116.23705.88

Biau, G., Cérou, F. and Guyader, A. (2010). Rates of convergence of the functional k-nearest neighbor estimate. *IEEE Transactions on Information Theory* **56**, 2034–2040. MR2654492 https://doi.org/10.1109/TIT.2010.2040857

Carbon, M., Francq, C. and Tran, L. T. (2007). Kernel regression estimation for random fields. *Journal of Statistical Planning and Inference* **137**, 778–798. MR2301715 https://doi.org/10.1016/j.jspi.2006.06.008

Cardot, H. and Sarda, P. (2005). Estimation in generalized linear models for functional data via penalized likelihood. *Journal of Multivariate Analysis* **92**, 24–41. MR2102242 https://doi.org/10.1016/j.jmva.2003.08.008

Carlo, G., Paolo, G. and Roberto, P. (2017). Spatial clustering of curves with an application of satellite data. *Spatial Statistics* **20**, 110–124. MR3654006 https://doi.org/10.1016/j.spasta.2017.01.006

Cérou, F. and Guyader, A. (2006). Nearest neighbor classification in infinite dimension. *ESAIM Probabilités Et Statistique* **10**, 340–355. MR2247925 https://doi.org/10.1051/ps:2006014

Chang, C., Chen, Y. and Ogden, R. T. (2014). Functional data classification: A wavelet approach. *Computational Statistics* **29**, 1497–1513. MR3279004 https://doi.org/10.1007/s00180-014-0503-4

Dabo-Niang, S. and Yao, A. F. (2007). Kernel regression estimation for continuous spatial processes. *Mathematical Methods of Statistics* **16**, 298–317. MR2378278 https://doi.org/10.3103/S1066530707040023

Dabo-Niang, S. and Yao, A. F. (2013). Kernel spatial density estimation in infinite dimension space. *Metrika* **76**, 19–52. MR3018356 https://doi.org/10.1007/s00184-011-0374-4

Delicado, P., Giraldo, R., Comas, C. and Mateu, J. (2010). Statistics for spatial functional data: Some recent contributions. *EnvironMetrics* **21**, 224–239. MR2842240 https://doi.org/10.1002/env.1003

Devroye, L., Györfi, L. and Lugosi, G. (1996). *A Probabilitic Theory of Pattern Recognition*. New York: Sprinер-Verlag. MR1383093 https://doi.org/10.1007/978-1-4612-0711-5

Devroye, L. and Krzyżak, A. (2013). An equivalence theorem for L1 convergence of the kernel regression estimate. *Journal of Statistical Planning and Inference* **23**, 71–82. MR1029241 https://doi.org/10.1016/0378-3758(89)90040-2

Ferraty, F., Van Keilegom, I. and Vieu, P. (2012). Regression when both response and predictor are functions. *Journal of Multivariate Analysis* **109**, 10–28. MR2922850 https://doi.org/10.1016/j.jmva.2012.02.008

Górecki, T., Krzysko, M. and Wolynski (2015). Classification problems based on regression models for multidimensional functional data. *Statistics in Transition New Series* **16**, 97–110.

Ibragimov, I. A. (1962). Some limit theorems for stationary processes. *Teoriâ Veroâtnostej I Ee Primeneniâ* **7**, 361–392. MR0148125

Jacques, J. and Preda, C. (2014). Model-based clustering for multivariate functional data. *Computational Statistics & Data Analysis* **71**, 92–106. MR3131956 https://doi.org/10.1016/j.csda.2012.12.004

Jiang, H. and Serban, N. (2012). Clustering random curves under spatial interdependence with application to service accessibility. *Technometrics* **54**, 108–119. MR2929427 https://doi.org/10.1080/00401706.2012.657106

Kolmogorov, A. N. and Tihomirov, V. M. (1961). ϵ-entropy and ϵ-capacity of sets in functional spaces. *Translations - American Mathematical Society* **17**, 277–364. MR0124720

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. MR1366756 https://doi.org/10.1109/18.391248

Lin, Z. and Yan, L. (2016). A support vector machine classifier based on a new kernel function model for hyperspectral data. *GIScience and Remote Sensing* **53**, 85–101.

Moughal, T. A. (2013). Hyperspectral image classification using support vector machine. *Journal of Physics Conference Series* **439**.

Nerini, D., Monestiez, P. and Manté, C. (2010). Cokriging for spatial functional data. *Journal of Multivariate Analysis* **101**, 409–418. MR2564350 https://doi.org/10.1016/j.jmva.2009.03.005

Ramsay, J. O. and Silverman, B. W. (2005). *Functional Data Analysis*, 2nd ed. *Springer Series in Statistics*. New York: Springer. MR2168993

Rio, E. (2000). *Théorie asymptotique des processus aléatoires faiblement dépendants*. *Mathématiques et Applications*. Berlin: Spriner. MR2117923

Romano, E., Balzanella, A. and Verde, R. (2010). Clustering spatio-functional data: A model based approach. In *Classification as a Tool for Research*, 167–175. Berlin: Springer. MR2722134 https://doi.org/10.1007/978-3-642-10745-0_17

Rosenblatt, M. (1956). A central limit theorem and a strong mixing condition. *Proceedings of the National Academy of Sciences of the United States of America* **42**, 43–47. MR0074711 https://doi.org/10.1073/pnas.42.1.43

Ruiz-Medina, M. D., Espejo, R. and Romano, E. (2014). Spatial functional normal mixed effect approach for curve classification. *Advances in Data Analysis and Classification* **8**, 257–285. MR3253860 https://doi.org/10.1007/s11634-014-0174-6

Saltyte-Benth, J. and Ducinskas, K. (2005). Linear discriminant analysis of multivariate spatial–temporal regressions. *Scandinavian Journal of Statistics* **32**, 281–294. MR2188674 https://doi.org/10.1111/j.1467-9469.2005.00421.x

Ternynck, C. (2014). Spatial regression estimation for functional data with spatial dependency. *Journal de la Société Française de Statistique* **155**, 673–684. MR3211759

Younso, A. (2017a). On the consistency of a new kernel rule for spatially dependent data. *Statistics & Probability Letters* **131**, 64–71. MR3706697 https://doi.org/10.1016/j.spl.2017.08.008

Younso, A. (2017b). On nonparametric classification for weakly dependent functional processes. *ESAIM Probabilités Et Statistique* **21**, 452–466. MR3743922 https://doi.org/10.1051/ps/2017002

Younso, A. (2018). On the consistency of kernel classification rule for functional random field. *Journal de la Société Française de Statistique* **159**, 68–87. MR3803124

Department of Mathematical Statistics
Faculty of Sciences
Damascus University
Damascus
Syria
E-mail: ahyounso@yahoo.fr