ESAIM: PS 21 (2017) 452–466
DOI: 10.1051/ps/2017002

ESAIM: Probability and Statistics
www.esaim-ps.org

ON NONPARAMETRIC CLASSIFICATION FOR WEAKLY DEPENDENT FUNCTIONAL PROCESSES

Ahmad Younso[1]

Abstract. The purpose of this paper is to investigate the moving window rule of classification to classify functions under mixing conditions. We consider a random variable X taking values in a metric space $(\mathcal{F}, \rho)$ with label $Y \in \{0, 1\}$. We extend some results on consistency and strong consistency of the moving window rule from the i.i.d. case to the weakly dependent case under mild assumptions. The practical use of the moving window rule will be illustrated through a simulation study. The performance of the moving window rule is investigated.

Mathematics Subject Classification. 62G08.

Received June 22, 2016. Revised January 6, 2017. Accepted February 15, 2017.

1. Introduction

In many experiments, the observations can be collected as elements in infinite dimensional spaces. This type of data, which we call functional data, arise naturally in many disciplines including medicine, economics, meteorology and many others. Statistics for functional random variables are becoming more and more important. The recent literature in this area shows the great potential of these new statistical methods for functional data. The most popular case of functional random variable corresponds to the situation when we observe random curves on different statistical units. Many multivariate statistical techniques have been extended to functional data and good overviews on this topic can be found in [1,7,8,15–18,20,25,30–33]. Nonparametric methods taking into account functional variables have been developed with very interesting practical motivations dealing with environmetrics (see [9]), chemometrics (see [15]), meteorological science (see [4]), speech recognition problems (see [16]), radar range profile (see [21]), medical data (see [19]), *etc.* Much of the early work on functional data analysis (FDA) focussed on i.i.d. functional random variables, but recently there has been heightened interest in dependent functional data. The need to take account of dependence is particularly evident in cases where functional data arise from segmenting a long time series into natural consecutive intervals (*e.g.* days, weeks, *etc.*) of equal length, as discussed by [22]. Electricity load curves, pollutant concentration curves and traffic volumes across the day are just a few examples of time series functional data studied in the literature (see [4, 9]). In this paper, we focus on the nonparametric classification for dependent functional random variables under weak dependence assumptions. Classification and regression estimation for functional data are of fundamental importance in the theory and practice of statistics. Various basic classification methods have been adapted to classify functional data. For example, under the assumption of independence, [20] adapt parametric multivariate

Keywords and phrases. Bayes rule, training data, moving window rule, mixing condition, consistency.

[1] Department of mathematical statistics, Faculty of sciences, Damascus University, Syria. ahyounso@yahoo.fr

Article published by EDP Sciences

regression models and [16] adapt the k-nearest neighbor method. We consider the problem of nonparametric classification by a kernel-based rule under α-and β-mixing conditions. The mixing conditions together with the functional approach allow us to classify segmented curves (*e.g.* electricity load curves) generated from continuous time processes. Let X be a random element with values in a metric space $(\mathcal{F}, \rho)$ where $\mathcal{F}$ is a function space and ρ denotes the metric on $\mathcal{F}$, and let Y be a random variable with values 0 or 1. The distribution of the pair (X, Y) is well defined by (μ, η) where $\mu(B) = \mathbb{P}(X \in B)$, for all Borel sets B on $\mathcal{F}$, and $\eta(x) = \mathbb{P}(Y = 1|X = x)$, for all $x \in \mathcal{F}$. In order to predict the unknown nature Y, called a class or label, of an observation $X = x$ with values in $\mathcal{F}$, the statistician creates a classifier $g : \mathcal{F} \longrightarrow \{0, 1\}$ which maps a new observation $x \in \mathcal{F}$ into its predicted label $g(x)$. It is certainly possible to wrongly specify its associated label y 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. An optimal classifier, called Bayes rule, is given by

$$g^*(x) = \mathbb{I}_{\{\eta(x) \geq 1/2\}},$$

where $\mathbb{I}_A$ denotes the indicator function of the set A. It is easy to see that the Bayes rule has the smallest probability of error, that is

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

We refer to Theorem 2.1 in [10] for the finite dimensional case. The Bayes rule depends on the distribution of (X, Y) which is generally unknown. But it is often possible to construct a classifier from a set of observations $D_n = \{(X_1, Y_1), \ldots, (X_n, Y_n)\}$ of (X, Y). The set D_n is called the training data. Among the various ways to define a classifier from a training data, one of the most popular is the moving window rule defined by

$$g_n(x) = \begin{cases} 0 & \text{if} \quad \sum_{i=1}^n \mathbb{I}_{\{Y_i=0, X_i \in B_{x,h}\}} \geq \sum_{i=1}^n \mathbb{I}_{\{Y_i=1, X_i \in B_{x,h}\}} \\ 1 & \text{otherwise,} \end{cases}$$

where $h = h(n)$ the smoothing factor, is a strictly positive number decreasing to 0 when $n \to \infty$ and $B_{x,h}$ denotes the closed ball of radius h centered at x. In order to establish the theoretical results, we write the moving window rule as follows

$$g_n(x) = \begin{cases} 0 & \text{if} \quad \eta_n(x) \leq \dfrac{\sum_{i=1}^n (1 - Y_i) \mathbb{I}_{\{X_i \in B_{x,h}\}}}{n\mu(B_{x,h})} \\ 1 & \text{otherwise,} \end{cases} \tag{1.1}$$

where

$$\eta_n(x) = \frac{\sum_{i=1}^n Y_i \mathbb{I}_{\{X_i \in B_{x,h}\}}}{n\mu(B_{x,h})}.$$

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, [27, 28, 39]. Let $L_n = L(g_n) = \mathbb{P}\{g_n(X) \neq Y\}$ be the error probability of $g_n(x)$. The classifier $g_n(x)$ is called consistent if

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

and called strongly consistent if

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

A classifier can be consistent for certain class of distribution of (X, Y), but not be consistent for others. The classifier $g_n(x)$ is called (strongly) universally consistent, if it is (strongly) consistent for all distributions 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 [11, 36]). [1] give some examples showing that the results of [11] on the consistency are no more valid in a general functional metric space $(\mathcal{F}, \rho)$ and they establish the consistency and the strong consistency under mild conditions on the distribution of (X, Y) and the metric space. Our aim in this paper is to establish the consistency and the strong consistency of the moving window rule for functional data under α- and β-mixing conditions.

2. Mixing conditions and preliminaries

Let $(\Omega, \mathcal{A}, \mathbb{P})$ be a probability space and let $\mathcal{B}$ and $\mathcal{C}$ be two sub σ-fields of $\mathcal{A}$. The α-mixing coefficient between $\mathcal{B}$ and $\mathcal{C}$ is defined by

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

and the β-mixing coefficient is defined by

$$\beta = \beta(\mathcal{B}, \mathcal{C}) = \mathbb{E} \sup_{C \in \mathcal{C}} |\mathbb{P}(C|\mathcal{B}) - \mathbb{P}(C)|.$$

Let $(Z_n)_{n \in \mathbb{Z}}$ be a stochastic process on $(\Omega, \mathcal{A}, \mathbb{P})$ and taking values in some space $(\Omega', \mathcal{A}')$. We denote the σ-fields generated by $(Z_i, i \leq s)$ and $(Z_i, i \geq s + t)$, respectively, by $\mathcal{B}_s$ and $\mathcal{C}_{s+t}$ for $s \in \mathbb{Z}$ and $t \in \mathbb{N}^*$.

Definition 2.1. The process $(Z_n)_{n \in \mathbb{Z}}$ is said to be strong mixing (or α-mixing) if

$$\alpha(t) = \sup_{s \in \mathbb{Z}} \alpha(\mathcal{B}_s, \mathcal{C}_{s+t}) \downarrow 0 \quad \text{as } t \to \infty.$$

The strong mixing coefficient is one of the most popular mixing coefficients. For more information on strong mixing processes, see for example, [34, 35]. If $\mathcal{B}_s = \sigma(Z_s)$ and $\mathcal{C}_{s+t} = \sigma(Z_{s+t})$, the process $(Z_n)_{n \in \mathbb{Z}}$ is called 2-α-mixing. The 2-α-mixing condition is weaker than strongly mixing (see [5]).

Definition 2.2. The process $(Z_n)_{n \in \mathbb{Z}}$ is said to be absolutely regular (or β-mixing) if

$$\beta(t) = \sup_{s \in \mathbb{Z}} \beta(\mathcal{B}_s, \mathcal{C}_{s+t}) \downarrow 0 \quad \text{as } t \to \infty.$$

Linear processes or more generally Markov chains may be absolutely regular (see [12]). The two mixing coefficients α and β are related by the inequality $2\alpha \leq \beta$ (see [34]). Consequently, any β-mixing process is α-mixing one. The following lemma (see [34]) is crucial to derive the consistency of the moving window rule. Let $\|.\|_\infty$ be the supremum norm.

Lemma 2.3. *If Z_1 and Z_2 are two $\mathbb{R}$-valued bounded random variables, then*

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

Now, let $(\Omega, \mathcal{A}, \mathbb{P})$ be a probability space and let $\mathcal{B}$ and $\mathcal{C}$ be two sub σ-fields of $\mathcal{A}$, we denote by $\mathcal{B} \vee \mathcal{C}$ the σ-field generated by $\mathcal{B} \cup \mathcal{C}$. The following coupling lemma (see [2]) will be needed to establish the strong consistency.

Lemma 2.4. *Let Z be a random variable on $(\Omega, \mathcal{A}, \mathbb{P})$ with values in some Polish space Ω' and $\mathcal{M}$ a sub σ-field of $\mathcal{A}$. Assume that there exists a random variable U uniformly distributed over $[0, 1]$, independent of $\sigma(Z) \vee \mathcal{M}$. Then, there exists a random variable Z^* measurable with respect to $\sigma(U) \vee \sigma(Z) \vee \mathcal{M}$, distributed as Z and independent of $\mathcal{M}$, such that*

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

Remark 2.5. A Polish space Ω' is a topological space which is separable and completely metrizable (see [23]). Most of the familiar objects of study in analysis involve Polish spaces. For example, $\mathbb{R}$ and $\mathbb{R}^d$ with the usual topology are Polish. For all $n \in \mathbb{N}^*$, $\{0, 1, \ldots, n-1\}$ is Polish with discrete topology. A countable product of Polish spaces is Polish, too.

3. Consistency in function space under mixing conditions

For convenience, we firstly introduce the notion of covering numbers (see [24]). For a given subset $\mathcal{G}$ of the metric space $(\mathcal{F}, \rho)$, the covering number is defined by

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

where $S_{x,\epsilon}$ is the open ball of radius $\epsilon > 0$ and center at $x \in \mathcal{F}$. The set $\mathcal{G}$ is said to be totally bounded if $\mathcal{N}(\epsilon, \mathcal{G}, \rho) < \infty$ for all $\epsilon > 0$. In particular, every relatively compact set is totally bounded and all totally bounded sets are bounded.

Assumption 3.1. There exists a sequence $(\mathcal{F}_k)_{k\geq 1}$ of totally bounded subsets of $\mathcal{F}$ such that $\mathcal{F}_k \subset \mathcal{F}_{k+1}$ for all $k \geq 1$ and $\mu\left(\bigcup_{k\geq 1} \mathcal{F}_k\right) = 1$.

Assumption 3.2. For any positive integer $k \geq 1$, any $i \neq j$ and $\epsilon_1 \in]0, 1]$, $\mathbb{P}((X_i, X_j) \in B_{x,h} \times B_{x,h}) \leq C[\mu(B_{x,h})]^{1+\epsilon_1}$, for all $x \in \mathcal{F}_k$, and some $C > 0$.

Assumption 3.3. The following Besicovich condition holds, for every $\epsilon > 0$,

$$\lim_{h \to 0+} \mu \left\{ x \in \mathcal{F} : \left| \frac{1}{\mu(B_{x,h})} \int_{B_{x,h}} \eta \mathrm{d}\mu - \eta(x) \right| > \epsilon \right\} = 0.$$

Remark 3.4. Note that Assumption 3.1 is always true whenever the space $(\mathcal{F}, \rho)$ is separable, see for example, [1, 26]. Regarding Assumption 3.2, we refer to [37] for the spatial version of this condition. This assumption can be linked with the classical local dependence condition met in the literature of the finite-dimensional case when X and (X_i, X_j) admit, respectively, the densities f and $f_{i,j}$ (see [5]). Assumption 3.3 holds for example if $\eta(x)$ is μ-continuous (see [7]).

We suppose that the training data D_n are observations of stationary 2-α-mixing functional process and there exist $C > 0$ and $\theta > 0$ such that

$$\alpha(t) \leq Ct^{-\theta} \text{ for all } t \in \mathbb{N}^*. \tag{3.1}$$

The hypothesis (3.1) means that the training data D_n are drawn from an arithmetically 2-α-mixing functional process. From now on, the notion $\mathcal{G}^c$ stands for the complement of any subset $\mathcal{G}$ of $\mathcal{F}$ and for simplicity of notation, we write $\mathcal{N}_k(\epsilon)$ instead of $\mathcal{N}(\epsilon, \mathcal{F}_k, \rho)$. The following two lemmas (see [1]) will be needed in the sequel.

Lemma 3.5. *Suppose that Assumption* 3.3 *holds. If $h \to 0$ as $n \to \infty$, then,*

$$\int_{\mathcal{F}} |\eta(x) - \mathbb{E}\eta_n(x)|\mu(\mathrm{d}x) = \int_{\mathcal{F}} \left| \eta(x) - \frac{\int_{B_{x,h}} \eta(t)\mu(\mathrm{d}t)}{\mu(B_{x,h})} \right| \mu(\mathrm{d}x) \longrightarrow 0.$$

Proof of Lemma 3.5 is a straightforward consequence of Assumption 3.3 and the Lebesgue dominated convergence theorem.

Lemma 3.6. *Suppose that $(\mathcal{F}_k)_{k\geq 1}$ is a sequence of totally bounded subsets of $\mathcal{F}$. Let k be a fixed positive integer. Then, for every $h > 0$,*

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

See [1] for the proof of Lemma 3.6.

Theorem 3.7. *Suppose that $(\mathcal{F}_k)_{k\geq 1}$ is a sequence of totally bounded subsets of $\mathcal{F}$ and Assumption* 3.2 *and* (3.1) *hold with $\theta > 2$. If $h \to 0$ as $n \to \infty$, then for n sufficiently large and for every positive integer k,*

$$\mathbb{E}\int_{\mathcal{F}_k} |\eta_n(x) - \mathbb{E}\eta_n(x)|\mu(\mathrm{d}x) \leq C\left(\frac{1}{n}\mathcal{N}_k\left(\frac{h}{2}\right)\right)^{1/2}, \qquad \text{for some} \qquad C > 0.$$

Corollary 3.8. *Let $(\mathcal{F}_k)_{k\geq 1}$ be a sequence of totally bounded subsets of $\mathcal{F}$. Suppose that Assumptions* 3.2–3.3 *and* (3.1) *hold with $\theta > 2$. If $h \to 0$ and for every $k \geq 1$, $\mathcal{N}_k(h/2)/n \longrightarrow 0$ as $n \to \infty$, then*

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

Observe that Corollary 3.8 generalizes the consistency result of [1] to the weakly dependent case under the same assumptions on the smoothing factor h.

4. Strong consistency in function space under mixing conditions

In this section, we investigate the strong consistency of the moving window classifier under β-mixing condition. This mixing condition together with the coupling Lemma 2.4 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 [6]). In order to establish the strong consistency, we suppose that the training data D_n are observations of stationary and arithmetically β-mixing functional process, and that there exist $C > 0$ and $\theta > 0$ such that

$$\beta(t) \leq Ct^{-\theta} \qquad \text{for all} \qquad t \in \mathbb{N}^*. \tag{4.1}$$

The following theorem generalizes the strong consistency result of [1] to the β-mixing case.

Theorem 4.1. *Let $(\mathcal{F}_k)_{k\geq 1}$ be a sequence of totally bounded subsets of $\mathcal{F}$. Assume that the training data D_n are observations of β- mixing functional process and the metric space $(\mathcal{F}, \rho)$ is Polish. Assume that Assumptions* 3.1−3.2 *and* (4.1) *hold with $\theta > 2$. Let $(k_n)_{n\geq 1}$ be an increasing sequence of positive integers such that*

$$\sum_{n\geq 1} \mu(\mathcal{F}^c_{k_n}) < \infty \qquad \text{and} \qquad \sum_{n\geq 1} \mathcal{N}_{k_n}\left(\frac{h}{2}\right) p_n^{-\theta} < \infty,$$

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

$$\frac{n}{p_n \log(n) \mathcal{N}^2_{k_n}(h/2)} \longrightarrow \infty \quad \text{as} \quad n \to \infty,$$

then

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

Remark 4.2. For example, consider $\mathcal{N}_{k_n}(h/2) \simeq n^{\gamma_1}$ with $0 < \gamma_1 < 1$, and choose $p_n \simeq n^{\gamma_2}$ with $(1+\gamma_1)/\theta < \gamma_2 < 1$ and $\theta > 2$. Clearly, we have

$$\sum_{n\geq 1} \mathcal{N}_{k_n}\left(\frac{h}{2}\right) p_n^{-\theta} < \infty.$$

Furthermore, the condition

$$\frac{n}{p_n \log(n) \mathcal{N}^2_{k_n}(h/2)} \longrightarrow \infty \qquad \text{as} \qquad n \to \infty$$

may be satisfied if $\gamma_2 + 2\gamma_1 < 1$. The condition $\sum_{n\geq 1} \mu(\mathcal{F}^c_{k_n}) < \infty$ is used by [1] in order to obtain the strong consistency in the independent case.

5. Smoothing factor selection and simulation study

In practice, the choice of a smoothing parameter h is a crucial problem to the kernel classifier. A wrong value of h 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 h, 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 [14]), but the situation is slightly different for the nonparametric classification problem. However, taking

$$g_n(x) = \begin{cases} 0 & \text{if} \quad \sum_{i=1}^n Y_i \mathbb{1}_{\{\|X_i - x\| \leq h\}} \leq \sum_{i=1}^n (1-Y_i)\mathbb{1}_{\{\|X_i - x\| \leq h\}} \\ 1 & \text{otherwise,} \end{cases}$$

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

$$CV(h) = \frac{1}{n}\sum_{i=1}^n (Y_i - g_n^{-i}(X_i))^2 \omega(X_i),$$

where $g_n^{-i}(X_i)$ indicates the moving window rule based on leaving out the i pair (X_i, Y_i) and $\omega(X_i)$ is the weight of the element X_i. We assume that h belongs to some set $H_n \subset \mathbb{R}_+$ including $h_1^i, \ldots, h_k^i$ for all $i = 1, \ldots, n$ where h_j^i is the distance to the j^{th} neighbor of X_i with respect to the metric ρ 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 S having non-empty interior (see [29]). For the sake of simplicity, we will take $\omega(x)$ as a constant. Therefore, the cross-validated smoothing factor is given by

$$h_{\text{opt}} = \arg\min_{h \in H_n} CV(h).$$

Now, we use the R statistical programming environment to run a simulation study. We propose to investigate the performance of our method in the following simulated scenario. For each $i = 1, \ldots, n$ and $t \in [0, 1]$, we generate pairs $(X_i(t), Y_i)$ via the scheme (see [3]):

$$X_i(t) = \sin(F_i^1 \pi t) f_{\mu_i, \sigma_i}(t) + \sin(F_i^2 \pi t) f_{1-\mu_i, \sigma_i}(t) + \epsilon_t$$

where $f_{\mu,\sigma}$ stands for the normal density with mean μ and variance σ^2; F_i^1 and F_i^2 are independent uniform random variables on $[140, 150]$; μ_i is randomly uniform on $[0.1, 0.4]$; σ_i^2 is randomly uniform on $[0, 0.005]$ and the ϵ_t 's are dependent normal random variables with mean 0, variance 0.25 and covariance function $\mathrm{c}(k) = 0.5|k|^{-2.5}$ for all $k \neq 0$. It is important to mention that a gaussian process is α-mixing if and only if its covariance function $\mathrm{c}(k)$ converges to zero as $k \to \infty$. We suppose that the function space on the interval $[0, 1]$ is endowed with the norm defined by $\|x\| = \int_0^1 |x(t)| \mathrm{d}t$. This norm is used without discretizing the data. For example, if we have the following realization of $X(t)$:

$$x(t) = \sin(148.67\pi t) f_{0.18, 0.06}(t) + \sin(146\pi t) f_{0.82, 0.06}(t) - 0.39,$$

by using the function *integrate* in R statistical programming environment, we get $\|x\| = 1.43476$ with absolute error less than 0.0000021. For the norm $\|.\|$, we take the metric $\rho(X_i, X_j) = \|X_i - X_j\|$. Let the label Y_i associated to X_i be defined, for $i = 1, \ldots, n$, by

$$Y_i = \begin{cases} 0 & \text{if} \quad \mu_i \leq 0.25 \\ 1 & \text{otherwise.} \end{cases}$$

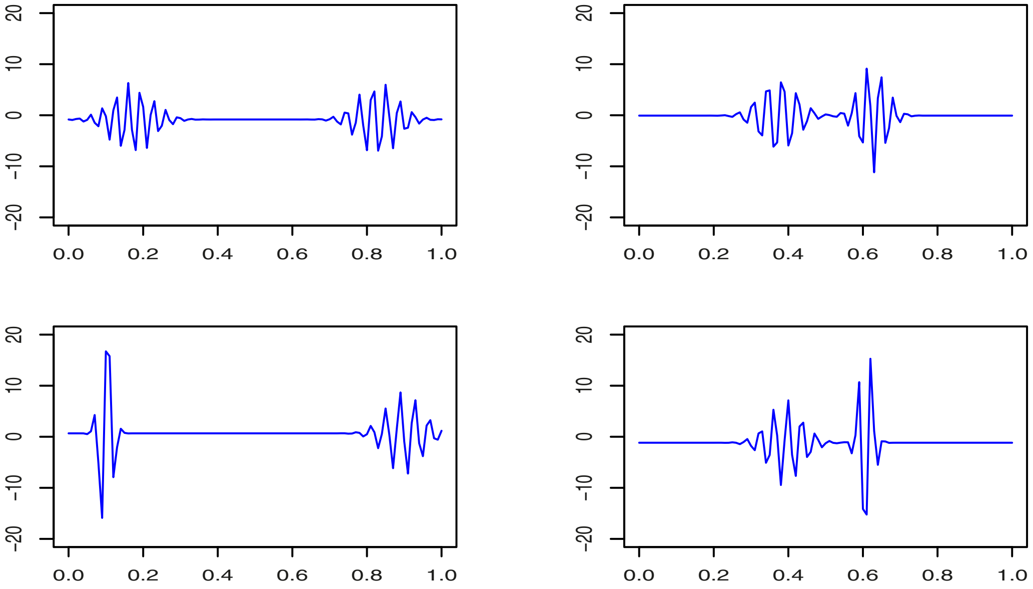

FIGURE 1. Four typical realizations of simulated curves with label 0 (*left*) and label 1 (*right*).

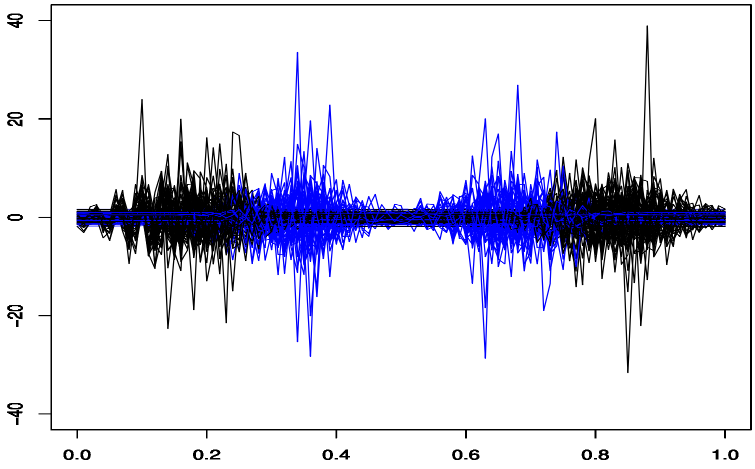

FIGURE 2. Sample of size $n = 100$ with labels 1 (*blue*) and labels 0 (*black*). (color online)

We firstly simulate a training sample of size $n = 100$ for $(X(t), Y)$ using the above scenario. Figure 1 displays four typical realizations of the X_i 's and Figure 2 displays plots for the training sample. We have for all $i, j = 1, \dots, 100$,

$$\min_{i \neq j} \rho(X_i, X_j) = 0.88, \ \max_{i \neq j} \rho(X_i, X_j) = 6.30.$$

For the sake of simplicity, the distances between the simulated curves are rounded off to two decimal digits. We estimate $CV(h)$ at different values of $h \in [0.88, 6.30]$ as in the following table:

Table 1. Some estimated values of $CV(h)$ at different values of h.

h	0.88	1.80	2.20	2.40	2.50	3.00	3.50	4.50	5.50	6.30
$CV(h)$	49%	40%	4%	1%	2%	27%	63%	49%	49%	49%

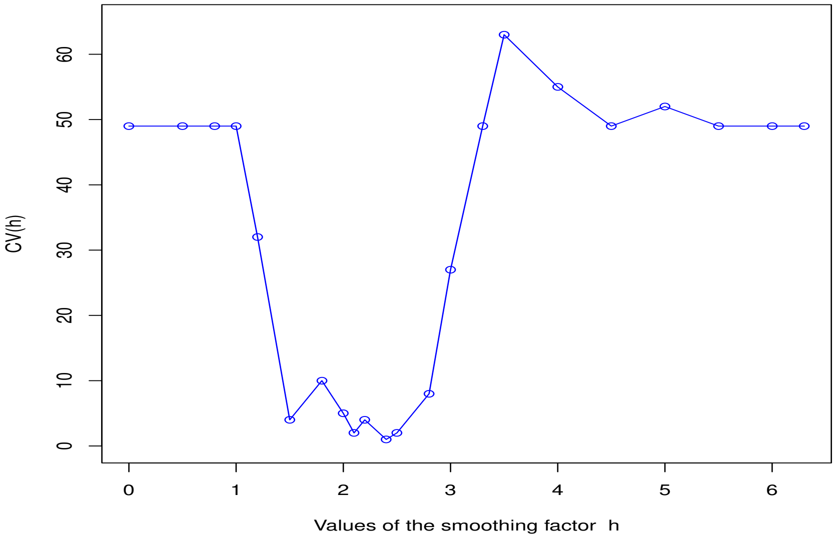

Figure 3. Variation of $CV(h)$ as a function of the smoothing factor h.

Table 2. Some estimated values of ER at different values of h.

h	0.88	1.80	2.20	2.40	2.50	3.00	3.50	4.50	5.50	6.30
ER	60%	48%	10%	4%	14%	40%	42%	42%	42%	50%

Figure 3 displays the variation of $CV(h)$ as a function of the smoothing factor h. It is evident that the function $CV(h)$ has a relative minimum value at $h \approx 2.4$. This allows to take $\widehat{h}_{\text{opt}} = 2.4$ as an optimal value of h.

Then, we simulate a testing sample of size $m = 50$ which is used to look at the behaviour of our method and we estimate the error rate of classification (ER) corresponding to the different values of h in Table 1. We have the following table:

Table 2 shows that the moving window rule is very sensitive to the choice of the optimal smoothing factor. The lowest possible error rate is at $\widehat{h}_{\text{opt}} = 2.40$.

Now, since the theoretical results of this paper are related to the consistency, it is natural to consider training samples with increasing sizes. For this aim, we generated, for each sample size, 100 training samples of size n and 100 corresponding test samples of size 100. In each replication, the proposed classifier was determined on the basis of the training sample at hand (based on the optimal bandwidth minimizing the $CV(h)$) and the misclassification error rate (ER) was evaluated based on the associated test sample. Table 3 then reports the average error rate (AER), obtained by averaging the error rates associated with the corresponding 100 test samples.

Table 3 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.

Table 3. Estimated optimal bandwidths and average error rates corresponding to training samples of different sizes.

n	25	50	100	150	200	250
$\widehat{h}_{\text{opt}}$	2.37	2.14	2.11	2.09	1.98	1.97
AER	11.79%	7.87%	5.23%	4.10%	3.85%	3.63%

6. Proofs

In order to establish the proofs in this section, we use $g_n(x)$ given by (1.1).

Proof of Theorem 3.7. Let $x \in \mathcal{F}$ be a fixed element. By Cauchy–Schwartz inequality, we have

$$\mathbb{E}|\eta_n(x) - \mathbb{E}\eta_n(x)| \leq (\text{var}(\eta_n(x)))^{1/2} \leq \left(\frac{\mathbb{E}(Y\mathbb{1}_{\{X\in B_{x,h}\}})^2}{n(\mu(B_{x,h}))^2} + S_n(x)\right)^{1/2},$$

where

$$S_n(x) = \frac{1}{(n\mu(B_{x,h}))^2}\sum_{i\neq j}|\text{cov}(\Delta_{\text{i}}, \Delta_{\text{j}})|$$

and $\varDelta_i = Y_i\mathbb{1}_{\{X_i\in B_{x,h}\}}$ for $i = 1, \ldots, n$. Now, since $|Y| \leq 1$, we obtain

$$\mathbb{E}|\eta_n(x) - \mathbb{E}\eta_n(x)| \leq \left(\frac{1}{n\mu(B_{x,h})} + S_n(x)\right)^{1/2}. \tag{6.1}$$

Let us first deal with the cross term $S_n(x)$. Choose u_n a sequence of increasing positives such that $u_n \to \infty$ as $n \to \infty$. Then

$$\begin{aligned} S_n(x) =& \frac{1}{(n\mu(B_{x,h}))^2}\sum_{0<|i-j|\leq u_n}|\text{cov}(\Delta_{\text{i}}, \Delta_{\text{j}})| \\ &+ \frac{1}{(n\mu(B_{x,h}))^2}\sum_{|i-j|>u_n}|\text{cov}(\Delta_{\text{i}}, \Delta_{\text{j}})|. \end{aligned} \tag{6.2}$$

Now, for $0 < |i-j| \leq u_n$, Assumption 2.2 implies that

$$\begin{aligned} |\text{cov}(\Delta_{\text{i}}, \Delta_{\text{j}})| &\leq \mathbb{E}(\varDelta_i\varDelta_j) + \mathbb{E}(\varDelta_i)\mathbb{E}(\varDelta_j) \\ &\leq \mathbb{P}((X_i, X_j) \in B_{x,h} \times B_{x,h}) + \{\mathbb{P}(X \in B_{x,h})\}^2 \\ &\leq C\{\mu(B_{x,h})\}^{1+\epsilon_1} + \{\mu(B_{x,h})\}^2, \end{aligned}$$

where $0 < \epsilon_1 \leq 1$ is the constant defined in Assumption 2.2 and C is a generic positive constant, independent of both x and n, whose value may vary from line to line. Since $\mu(B_{x,h}) \leq 1$, we have $\{\mu(B_{x,h})\}^2 \leq \{\mu(B_{x,h})\}^{1+\epsilon_1}$ and then

$$|\text{cov}(\varDelta_i, \varDelta_j)| \leq C\{\mu(B_{x,h})\}^{1+\epsilon_1}. \tag{6.3}$$

If $|i-j| > u_n$, by Lemma 2.2 and the fact that $|Y| \leq 1$, we get

$$|\text{cov}(\Delta_{\text{i}}, \Delta_{\text{j}})| \leq 4\alpha(|i-j|). \tag{6.4}$$

From (6.2), (6.3) and (6.4), we can write

$$\begin{aligned} S_n(x) &\leq \frac{Cu_n}{n(\mu(B_{x,h}))^{1-\epsilon_1}} + \frac{4}{(n\mu(B_{x,h}))^2} \sum_{|i-j|\geq u_n} \alpha(|i-j|) \\ &\leq \frac{Cu_n}{n(\mu(B_{x,h}))^{1-\epsilon_1}} + \frac{4}{n(\mu(B_{x,h}))^2} \sum_{i\geq u_n} \alpha(i). \end{aligned} \tag{6.5}$$

Since $u_n > 1$ and $u_n - 1 \geq u_n/2$ for n sufficiently large, it follows from (3.1) that

$$\sum_{i\geq u_n} \alpha(i) \leq C \int_{u_n-1}^{\infty} t^{-\theta}\mathrm{d}t \leq C \int_{u_n/2}^{\infty} t^{-\theta}\mathrm{d}t \leq \frac{Cu_n^{1-\theta}}{\theta-1}\cdot \tag{6.6}$$

Consequently, by (6.5) and (6.6), we obtain

$$S_n(x) \leq \frac{Cu_n}{n(\mu(B_{x,h}))^{1-\epsilon_1}} + \frac{Cu_n^{1-\theta}}{n(\mu(B_{x,h}))^2}\cdot$$

Choosing $u_n = 1/(\mu(B_{x,h}))^{\epsilon_1}$ and $1/(\theta-1) < \epsilon_1 \leq 1$, where $\theta > 2$, we get for n sufficiently large $\{\mu(B_{x,h})\}^{(\theta-1)\epsilon_1} \leq \mu(B_{x,h})$ and

$$S_n(x) \leq \frac{C}{n\mu(B_{x,h})}\cdot \tag{6.7}$$

Thus, from (6.1) and (6.7), it follows that

$$\mathbb{E}|\eta_n(x) - \mathbb{E}\eta_n(x)| \leq \frac{C}{\sqrt{n\mu(B_{x,h})}}\cdot \tag{6.8}$$

By Fubini's theorem, Jensens's inequality and Lemma 3.6, we get

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

□

Proof of Corollary 3.8. By Theorem 2.3 in [10], whose extention to the infinite dimensional setting is straightforward, the corollary will be proved if we show that

$$\mathbb{E}\int_{\mathcal{F}} |\eta(x) - \eta_n(x)|\mu(\mathrm{d}x) \longrightarrow 0 \text{ as } n \to \infty.$$

Since $\eta(x) \leq 1$ and $\mathbb{E}\eta_n(x) \leq 1$, we have, for any integer $k \geq 1$,

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

Consequently, according to Theorem 3.7, for n sufficiently large, we get the following inequality

$$\mathbb{E}\int_{\mathcal{F}} |\eta(x) - \eta_n(x)|\mu(\mathrm{d}x) \leq \int_{\mathcal{F}} |\eta(x) - \mathbb{E}\eta_n(x)|\mu(\mathrm{d}x) + C\left(\frac{1}{n}\mathcal{N}_k\left(\frac{h}{2}\right)\right)^{1/2} + 2\mu(\mathcal{F}_k^c).$$

Therefore, by Lemma 3.5 and the assumptions on h, we get for every $k \geq 1$,

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

If we let k go to infinity, Assumption 2.1 yields

$$\limsup_{n\to\infty} \mathbb{E} \int_{\mathcal{F}} |\eta(x) - \eta_n(x)|\mu(\mathrm{d}x) = 0, \tag{6.9}$$

and the proof of the corollary is completed. □

Proof of Theorem 4.1. We set $Z = (X, Y)$ and $Z_i = (X_i, Y_i)$ for $i = 1, \ldots, n$. By assumptions, X and X_i take value in the Polish metric space $\mathcal{F}$, so, $Z = (X, Y)$ and $Z_i = (X_i, Y_i)$ take values in the product Polish space $\mathcal{F} \times \{0, 1\}$. We will now use the blocks decomposition introduced by [13] (see also [38]) which will be useful afterwards. Without loss of generality, let $n = 2pq$ for $p = p_n, q = q_n \in [1, n/2]$ such that $p_n \to \infty$ as $n \to \infty$ and let us define blocks as follow

$$\begin{array}{ll} W_1 = (Z_1, \ldots, Z_p), & V_1 = (Z_{p+1}, \ldots, Z_{2p}) \\ W_2 = (Z_{2p+1}, \ldots, Z_{3p}), & V_2^{=}(Z_{3p+1}, \ldots, Z_{4p}) \\ \quad\cdots & \quad\cdots \\ W_q = (Z_{2(q-1)p+1}, \ldots, Z_{(2q-1)p}), & V_q = (Z_{(2q-1)p+1}, \ldots, Z_{2pq}). \end{array}$$

Observe that W_i and V_i are $\sigma(Z_j, j \in I_i)$-measurable and $\sigma(Z_j, j \in \tilde{I}_i)$-measurable respectively, where $I_i = \{j :\ 2(i-1)p + 1 \leq j \leq (2i-1)p\}$ and $\tilde{I}_i = \{j :\ (2i-1)p + 1 \leq j \leq 2ip\}$ for all $i = 1, \ldots, q$. Furthermore, we have $|j - j^{'}| > p$ for any $j \in I_i$ and $j^{'} \in I_{i^{'}}$ if $i \neq i^{'}$. In the same way, one can show that $|j - j^{'}| > p$ for any $j \in \tilde{I}_i$ and $j^{'} \in \tilde{I}_{i^{'}}$ if $i \neq i^{'}$. Now, according to Lemma 2.4, we can find mutually independent random vectors

$$W_1^* = (Z_1^*, \ldots, Z_p^*), \ldots, W_q^* = \left(Z_{2(q-1)p+1}^*, \ldots, Z_{(2q-1)p}^*\right)$$

such that for all $i = 1, \ldots, q$, W_i^* has the same probability distribution as W_i and $\mathbb{P}(W_i \neq W_i^*) \leq \beta(p)$. We can find also mutually independent random vectors

$$V_1^* = \left(Z_{p+1}^*, \ldots, Z_{2p}^*\right), \ldots, V_q^* = \left(Z_{(2q-1)p+1}^*, \ldots, Z_{2qp}^*\right)$$

such that for all $i = 1, \ldots, q$, V_i^* has the same probability distribution as V_i and $\mathbb{P}(V_i \neq V_i^*) \leq \beta(p)$. As a consequence, for all $i = 1, \ldots, n$,

$$\mathbb{P}(Z_i \neq Z_i^*) = \mathbb{P}((X_i, Y_i) \neq (X_i^*, Y_i^*)) \leq \beta(p). \tag{6.10}$$

By definition of blocks, the variables $\sum_{j=2(i-1)p+1}^{(2i-1)p} Y_i^* \mathbb{1}_{\{X_i^* \in B_{x,h}\}}$ are mutually independent and the variables $\sum_{j=(2i-1)p}^{2ip} Y_i^* \mathbb{1}_{\{X_i^* \in B_{x,h}\}}$ are also mutually independent for all $i = 1, \ldots, q$. If we denote

$$\eta_n^*(x) = \frac{\sum_{i=1}^n Y_i^* \mathbb{1}_{\{X_i^* \in B_{x,h}\}}}{n\mu(B_{x,h})},\ \eta_{w,n}^*(x) = \frac{\displaystyle\sum_{i=1}^{q} \sum_{j=2(i-1)p+1}^{(2i-1)p} Y_i^* \mathbb{1}_{\{X_i^* \in B_{x,h}\}}}{n\mu(B_{x,h})}$$

and

$$\eta_{v,n}^*(x) = \frac{\displaystyle\sum_{i=1}^{q} \sum_{j=(2i-1)p+1}^{2ip} Y_i^* \mathbb{1}_{\{X_i^* \in B_{x,h}\}}}{n\mu(B_{x,h})},$$

then,

$$\eta_n^*(x) = \eta_{w,n}^*(x) + \eta_{v,n}^*(x). \tag{6.11}$$

Let $(k_n)_{n\geq 1}$ be the increasing positive sequence defined in the statement of Theorem 4.1. By Theorem 2.3 in [10], the theorem will be proved if we show that

$$\int_{\mathcal{F}} |\eta(x) - \eta_n(x)|\mu(\mathrm{d}x) \longrightarrow 0 \quad \text{as } n \to \infty \ \text{ with probability one.} \tag{6.12}$$

We first proceed to show that

$$\int_{\mathcal{F}_{k_n}} |\eta(x) - \eta_n(x)|\mu(\mathrm{d}x) \to 0 \text{ with probability one as } \ n \to \infty. \tag{6.13}$$

By Fubini's theorem, we have

$$\begin{aligned}\int_{\mathcal{F}_{k_n}} |\eta(x) - \eta_n(x)|\mu(\mathrm{d}x) &\leq \int_{\mathcal{F}_{k_n}} |\eta(x) - \mathbb{E}\eta_n(x)|\mu(\mathrm{d}x) + \int_{\mathcal{F}_{k_n}} |\eta_n(x) - \mathbb{E}\eta_n(x)|\mu(\mathrm{d}x).\\ &\leq \mathbb{E}\int_{\mathcal{F}_{k_n}} |\eta(x) - \eta_n(x)|\mu(\mathrm{d}x) + \int_{\mathcal{F}_{k_n}} |\eta_n(x) - \mathbb{E}\eta_n(x)|\mu(\mathrm{d}x). \end{aligned} \tag{6.14}$$

According to (6.9), we have

$$\mathbb{E}\int_{\mathcal{F}_{k_n}} |\eta(x) - \eta_n(x)|\mu(\mathrm{d}x) \leq \mathbb{E}\int_{\mathcal{F}} |\eta(x) - \eta_n(x)|\mu(\mathrm{d}x) \longrightarrow 0 \text{ as } n \to \infty.$$

Thus, by (6.14), it suffices to show that

$$\int_{\mathcal{F}_{k_n}} |\eta_n(x) - \mathbb{E}\eta_n(x)|\mu(\mathrm{d}x) \to 0 \text{ with probability one as } \ n \to \infty. \tag{6.15}$$

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

$$\begin{aligned}&\mathbb{P}\left(\left|\int_{\mathcal{F}_{k_n}} |\eta_n(x) - \mathbb{E}\eta_n(x)|\mu(\mathrm{d}x) - \int_{\mathcal{F}_{k_n}} |\eta_n^*(x) - \mathbb{E}\eta_n^*(x)|\mu(\mathrm{d}x)\right| > \epsilon\right)\\ &\leq \epsilon^{-1}\mathbb{E}\left|\int_{\mathcal{F}_{k_n}} |\eta_n(x) - \mathbb{E}\eta_n(x)|\mu(\mathrm{d}x) - \int_{\mathcal{F}_{k_n}} |\eta_n^*(x) - \mathbb{E}\eta_n^*(x)|\mu(\mathrm{d}x)\right|\\ &\leq 2\epsilon^{-1}\mathbb{E}\int_{\mathcal{F}_{k_n}} |\eta_n^*(x) - \eta_n(x)|\mu(\mathrm{d}x)\\ &= 2\epsilon^{-1}\mathbb{E}\int_{\mathcal{F}_{k_n}} \left|\frac{\sum_{i=1}^n Y_i^*\mathbb{1}_{\{X_i^*\in B_{x,h}\}}}{n\mu(B_{x,h})} - \frac{\sum_{i=1}^n Y_i\mathbb{1}_{\{X_i\in B_{x,h}\}}}{n\mu(B_{x,h})}\right|\mu(\mathrm{d}x)\\ &\leq 4\epsilon^{-1}\sum_{i=1}^n \mathbb{E}\mathbb{1}_{\{(X_i^*,Y_i^*)\neq(X_i,Y_i)\}}\int_{\mathcal{F}_{k_n}} \frac{1}{n\mu(B_{x,h})}\mu(\mathrm{d}x).\end{aligned}$$

As a consequence, by Lemma 3.6, (6.10) and (4.1), we have

$$\begin{aligned}&\mathbb{P}\left(\left|\int_{\mathcal{F}_{k_n}} |\eta_n(x) - \mathbb{E}\eta_n(x)|\mu(\mathrm{d}x) - \int_{\mathcal{F}_{k_n}} |\eta_n^*(x) - \mathbb{E}\eta_n^*(x)|\mu(\mathrm{d}x)\right| > \epsilon\right)\\ &\leq 4\epsilon^{-1}\sum_{i=1}^n \mathbb{P}\left((X_i^*,Y_i^*)\neq(X_i,Y_i)\right)\int_{\mathcal{F}_{k_n}} \frac{1}{n\mu(B_{x,h})}\mu(\mathrm{d}x)\\ &\leq C\epsilon^{-1}\mathcal{N}_{k_n}\left(\frac{h}{2}\right)\beta(p) \leq C\epsilon^{-1}\mathcal{N}_{k_n}\left(\frac{h}{2}\right)p^{-\theta}\end{aligned}$$

for some generic constant $C > 0$. Thus, by the assumptions on h and the Borel−Cantelli lemma, we get

$$\int_{\mathcal{F}_{k_n}} |\eta_n(x) - \mathbb{E}\eta_n(x)|\mu(\mathrm{d}x) - \int_{\mathcal{F}_{k_n}} |\eta_n^*(x) - \mathbb{E}\eta_n^*(x)|\mu(\mathrm{d}x) \longrightarrow 0$$

with probability one as $n \to \infty$. So, (6.13) will be proved if we show that

$$\int_{\mathcal{F}_{k_n}} |\eta_n^*(x) - \mathbb{E}\eta_n^*(x)|\mu(\mathrm{d}x) \longrightarrow 0 \text{ with probability one as } n \to \infty. \tag{6.16}$$

To do that, by (6.11) , we have

$$\int_{\mathcal{F}_{k_n}} |\eta_n^*(x) - \mathbb{E}\eta_n^*(x)|\mu(\mathrm{d}x) \leq \int_{\mathcal{F}_{k_n}} |\eta_{w,n}^*(x) - \mathbb{E}\eta_{w,n}^*(x)|\mu(\mathrm{d}x) + \int_{\mathcal{F}_{k_n}} |\eta_{v,n}^*(x) - \mathbb{E}\eta_{v,n}^*(x)|\mu(\mathrm{d}x). \tag{6.17}$$

Therefore, we have to prove that the two terms on the right hand side of the inequality (6.17) tend to zero as $n \to \infty$. Let $F : ((\mathcal{F} \times \{0,1\})^p)^q \to \mathbb{R}$ a real function defined as follows

$$F(W_1^*, \dots, W_q^*) = \int_{\mathcal{F}_{k_n}} |\eta_{w,n}^*(x) - \mathbb{E}\eta_{w,n}^*(x)|\mu(\mathrm{d}x).$$

For $w_i \neq w_i^{'}$ where $w_i,\ w_i^{'} \in (\mathcal{F} \times \{0,1\})^p$, by Lemma 3.6, we have

$$\begin{aligned} |F(W_1^*, \dots w_i, \dots, W_q^*) - F(W_1^*, \dots w_i^{'}, \dots, W_q^*)| &\leq \frac{2p}{n} \int_{\mathcal{F}_{k_n}} \frac{1}{\mu(B_{x,h})}\mu(\mathrm{d}x) \\ &\leq \frac{Cp}{n}\mathcal{N}_{k_n}\left(\frac{h}{2}\right) \cdot \end{aligned}$$

By McDiarmid's inequality, for every $\epsilon > 0$,

$$\mathbb{P}\left(\left|F(W_1^*, \dots, W_q^*) - \mathbb{E}(F(W_1^*, \dots, W_q^*))\right| > \epsilon\right) \leq 2\exp\left(-\frac{\epsilon^2 n}{C^2 p \mathcal{N}_{k_n}^2\left(\frac{h}{2}\right)}\right) \cdot$$

With the help of the Borel−Cantelli lemma and the assumption on h, we get

$$\int_{\mathcal{F}_{k_n}} |\eta_{w,n}^*(x) - \mathbb{E}\eta_{w,n}^*(x)|\mu(\mathrm{d}x) - \mathbb{E}\int_{\mathcal{F}_{k_n}} |\eta_{w,n}^*(x) - \mathbb{E}\eta_{w,n}^*(x)|\mu(\mathrm{d}x) \longrightarrow 0$$

with probability one as $n \to \infty$. Similar arguments can be used to prove

$$\int_{\mathcal{F}_{k_n}} |\eta_{v,n}^*(x) - \mathbb{E}\eta_{v,n}^*(x)|\mu(\mathrm{d}x) - \mathbb{E}\int_{\mathcal{F}_{k_n}} |\eta_{v,n}^*(x) - \mathbb{E}\eta_{v,n}^*(x)|\mu(\mathrm{d}x) \longrightarrow 0$$

with probability one as $n \to \infty$. So, (6.16) will be proved if we show that

$$\mathbb{E}\int_{\mathcal{F}_{k_n}} |\eta_{w,n}^*(x) - \mathbb{E}\eta_{w,n}^*(x)|\mu(\mathrm{d}x) \longrightarrow 0 \quad \text{as} \quad n \to 0 \tag{6.18}$$

and

$$\mathbb{E}\int_{\mathcal{F}_{k_n}} |\eta_{v,n}^*(x) - \mathbb{E}\eta_{v,n}^*(x)|\mu(\mathrm{d}x) \longrightarrow 0 \quad \text{as} \quad n \to 0. \tag{6.19}$$

Since $2\alpha(t) \leq \beta(t) \leq Ct^{-\theta}$ for each $t \in \mathbb{N}^*$, with a straightforward adaptation of the proof of Theorem 3.7, one can easily prove (6.18) and (6.19). As a consequence, the proof of (6.16) is completed and then, the proof of (6.13) is also completed. To finish the proof of the theorem, let us denote for all $n \geq 1$ and $i = 1, \ldots, n$,

$$Z_i^n = \int_{\mathcal{F}_{k_n}^c} \frac{\mathbb{1}_{\{X_i \in B_{x,h}\}}}{\mu(B_{x,h})} \mu(\mathrm{d}x).$$

It follows that

$$\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^n Z_i^n\right] = \mu(\mathcal{F}_{k_n}^c).$$

By assumption and the Borel−Cantelli lemma, we have

$$\frac{1}{n}\sum_{i=1}^n Z_i^n \longrightarrow 0 \text{ with probability one as } n \to \infty. \tag{6.20}$$

Hence, we can write

$$\begin{aligned}\int_{\mathcal{F}} |\eta(x) - \eta_n(x)|\mu(\mathrm{d}x) &= \int_{\mathcal{F}_{k_n}} |\eta(x) - \eta_n(x)|\mu(\mathrm{d}x) + \int_{\mathcal{F}_{k_n}^c} |\eta(x) - \eta_n(x)|\mu(\mathrm{d}x) \\ &\leq \int_{\mathcal{F}_{k_n}} |\eta(x) - \eta_n(x)|\mu(\mathrm{d}x) + \mu(\mathcal{F}_{k_n}^c) + \frac{1}{n}\sum_{i=1}^n Z_i^n.\end{aligned}$$

Finally, by Assumption 3.1, (6.13) and (6.20), all terms on the right hand side of the last inequality tend to 0 as $n \to \infty$, and the proof is completed. □

7. Conclusion

In this paper, we have assessed the performance of the moving window rule based on weakly dependent functional data by studying the (strong) consistency of this classifier under mild assumptions. For the practical use of the moving window rule, we have used the cross-validation method to choose the smoothing parameter. As a sequel of the simulation study, we have shown that the optimal smoothing factor obtained by the cross-validation criterion implies the smallest estimated error rate. We have also shown that the average rate of misclassification decreases as the training sample size increases which is in line with the theoretical results on the consistency.

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

References

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

[2] H.C.P. Berbee, Random walks with stationary increments and renewal theory. *Math. Cent. Tract.* Amsterdam (1979).

[3] A. Berlinet, G. Biau and L. Rouviére, Functional supervised classification with wavelets. *Ann. l'ISUP* **52** (2008) 61–80.

[4] P. Besse, H. Cardot and D. Stephenson, Autoregressive forecasting of some functional climatic variations. *Scand. J. Statist.* **27** (2000) 673–687.

[5] D. Bosq, Nonparametric Statistics for Stochastic Processes. Springer Verlag, New York (1998).

[6] R.C. Bradley, Approximations theorems for strongly mixing random variables. *Michigan Math. J.* **30** (1983) 69–81.

[7] F. Cérou and A. Guyader, Nearest neighbor classification in infinite dimension. *ESAIM: PS* **10** (2006) 340–355.

[8] S. Dabo-Niang and F. Ferraty, Functional and Operatorial Statistics. Physica Verlag, Springer, Heidelberg (2008).

[9] J. Damon and S. Guillas, The inclusion of exogenous variables in functional autoregressive ozone forecasting. *Environmetrics* **13** (2002) 759–774.

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

[11] L. Devroye and A. Krzyzak, An equivalence theorem for L_1 convergence of the kernel regression estimate. *J. Stat. Plan. Inference* **23** (1989) 71–82.

[12] P. Doukhan, P. Massart and E. Rio, The functional central limit theorem for strongly mixing processes. *Ann. Inst. Henri Poincaré Probab. Statist.* **30** (1994) 63–82.

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

[14] M. Febrero−Bande and M. Oviedo de la Fuente, Statistical computing in functional data analysis: The R package fda.usc. *J. Statist. Software* **51** (2012).

[15] F. Ferraty and P. Vieu, The functional nonparametric model and application to spectrometric data. *Comput. Statist.* (2002) 17–564.

[16] F. Ferraty and P. Vieu, Curves discrimination: A nonparametric functional approach. *Comput. Statist. Data Anal.* **44** (2003) 161–173.

[17] F. Ferraty and P. Vieu, Nonparametric Functional Data Analysis. Springer Verlag, New York (2006).

[18] F. Ferraty and P. Vieu, Additive prediction and boosting for functional data. *Comput. Statist. Data Anal.* **53** (2009) 1400–1413.

[19] T. Gasser, P. Hall and B. Presnell, Nonparametric estimation of the mode of a distribution of random curves. *J. Roy. Statist. Soc.* **60** (1998) 681–691.

[20] T. Górecki, M. Krzyśko and W. Wolyński, Classification problems based on regression models for multi-dimensional functional data. *Stat. Trans. New Series* **16** (2015) 97–110.

[21] P. Hall, P. Poskitt and D. Presnell, A functional data-analytic approach to signal discrimination. *Technometrics* **43** (2001) 140–143.

[22] S. Hörmann and P. Kokoszka, Weakly dependent functional data. *Ann. Statist.* **38** (2010) 1845–1884.

[23] A.S. Kechris, Classical descriptive set theory. Springer Verlag, New York (1995).

[24] A.N. Kolmogorov and V.M. Tihomirov, ϵ-entropy and ϵ-capacity of sets in functional spaces. *Amer. Math. Soc. Transl.* **17** (1961) 277–364.

[25] D. Kosiorowski, Functional regression in short-term prediction of economic time series. *Stat. Trans. New Series* **15** (2014) 611–626.

[26] S.R. Kulkarni and S.E. Posner, Rate of convergence of nearest neighbor estimation under arbitrary sampling. *IEEE Trans. Inform. Theory* **41** (1995) 1028–1039.

[27] E.A. Nadaraya, On estimating regression. *Theory Probab. Appl.* **9** (1964) 141–142.

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

[29] M. Rachdi and P. Vieu, Nonparametric regression for functional data: automatic smoothing parameter selection. *J. Statist. Plan. Inference* **137** (2007) 2784–2801.

[30] J.O. Ramsay, G. Hooker and S. Graves, Functional Data Analysis with R and Matlab. Springer Verlag, New York (2009).

[31] J.O. Ramsay and B.W. Silverman, Applied Functional Data Analysis. Methods and Case Sudies. Springer Verlag, New York (2002).

[32] J.O. Ramsay and B.W. Silverman, Functional Data Analysis. Springer Verlag, New York (1997).

[33] J.A. Rice and B.W. Silverman, Estimating the mean and covariance structure nonparametrically when the data are curves. *Roy. Statist. Soc.* **53** (1991) 233–243.

[34] E. Rio, Théorie asymptotique des processus aléatoires faiblement dépendants. Springer Verlag, Berlin Heidelberg (2000).

[35] M. Rosenblatt, A central limit theorem and a strong mixing condition. *Proc. Nat. Acad. Sci., USA* **42** (1956) 43–47.

[36] C.J. Stone, Consistent nonparametric regression. *Ann. Statist.* **5** (1977) 595–620.

[37] C. Ternynck, Spatial regression estimation for functional data with spatial dependency. *J. Soc. Française Statist.* **155** (2014) 138–160.

[38] G. Viennet, Inequalities for absolutely sequence. Application to density estimation. *Probab. Theory Relat. Fields* **107** (1967) 467–492.

[39] G.S. Watson, Smooth regression analysis. *Sankhya Ser. A* **26** (1964) 359–372.