On local times, density estimation and classification rules for functional data.

P. Llop[c,1,*], L. Forzani[a], R. Fraiman[b]

[a]*Instituto de Matemática Aplicada del Litoral - CONICET - Universidad Nacional del Litoral.*
[b]*Departamento de Matemática, Universidad de San Andrés and CMAT, Universidad de la República.*
[c]*Instituto de Matemática Aplicada del Litoral.*

Abstract

In this paper we define a $\sqrt{n}$–consistent nonparametric density estimator for functional data. Under mild conditions we obtain strong consistency, strong orders of convergence and derive the asymptotic distribution of the estimator. We propose a nonparametric classification rule based on local times (occupation measure) and include some simulations studies.

Key words: Functional data, Density estimation, Nearest neighbors

1. Introduction.

Recent advances in technology allow significantly more data to be recorded over a continuous period of time, where samples are trajectories of a stochastic process, for a given number of individuals. Such data are common in different fields, including health sciences, engineering, physical sciences, chemometrics, finance and social sciences. The set of statistical tools for analyzing such data is called functional data analysis, FDA from now on, an expression coined by Ramsay and Dalzell [22].

The books of Ramsay and Silverman [23, 24] describe in a systematic way some techniques for exploration of functional data. Further results can be found in Ferraty and Vieu [12].

*Corresponding author
Postal Adress : Güemes N^o 3450, CP 3000, Santa Fe, Argentina
Tel.: (+54)342-4559175 - Fax: (+54)-342-4550944
Email addresses: `lloppamela@gmail.com` (P. Llop), `liliana.forzani@gmail.com` (L. Forzani), `rfraiman@udesa.edu.ar` (R. Fraiman)

[1]This is a part of the author Doctoral Dissertation.

In general the data come from discretized functions. The most popular approaches to deal with these kind of data sets are to start by a “regularization procedure” (Hastie et al, 1995) (that tends to exclude from consideration the functions that are “too wiggly”) or by a “filtering method”, which leads to replacement of every function by its coefficients with respect to the basis of a suitable finite-dimensional subspace. On the contrary, in our setting “too wiggly” will be a good property under which we will be able to attain parametric rates of convergence, and so, we will focus on stochastic processes with irregular trajectories.

In this paper we consider a sample of curves $\{X_1(t), \ldots, X_n(t)\}$ generated from a stochastic process $X(t)$ given by

$$X(t) = \mu(t) + e(t), \quad t \in T, \tag{1}$$

where $\mu(t)$ stands for the mean function, and $e(t)$ is a zero mean, first-order stationary stochastic process with density unknown function f.

The problem we address is the nonparametric density estimation in the context of functional data, i.e. estimate the density of $X(t)$ when n independent trajectories of a continuous time stochastic processes verifying (1) are observed. When μ is constant, this problem has been considered by several authors in a different setup, mainly when a single sample path is observed over an increasing interval $[0, T]$ and T grows to infinity. This problem of estimating the marginal probability density function in dependent contexts have been studied starting in Rosenblatt [25], Nguyen [20], and mainly by Castellana and Leadbetter [6], where it is shown that for continuous time processes a parametric speed of convergence is attained by kernel type estimates. More recently, it has also been considered by Blanke and Bosq [4], Blanke [3], Kutoyants [16] among others. In particular, Labrador [17] propose a k–nearest neighbor type estimate using local time ideas.

We share with Labrador the use of local times (occupation measure) in order to define nearest neighbor estimates in two cases: when $\mu(t)$ is a constant independent of time and when it is not constant. In Section 2 we consider the first case and obtain strong consistency, strong order of convergence and the asymptotic distribution of the estimator under mild conditions. A parametric $\sqrt{n}$ speed of convergence is attained for the asymptotic distribution. In Section 3 we consider the second case and obtain strong consistency and strong orders of convergence. In Section 4 we applied our results to obtain a new classification rule for FDA. Moreover, some small simulations studies are given. All proofs are given at the Appendix.

2. Estimating the density function of stationary processes.

In this section we define a density estimator for stationary stochastic processes, i.e. in model (1) we consider $\mu(t) = \mu$ a constant independent of t so that $X(t)$ has the same properties as $e(t)$. Let $T \subset \mathbb{R}$ a finite interval, and $\{X(t) : t \in T\}$ a real–valued measurable continuous time stationary process with continuous trajectories which admits a local time. Let $\{X_1(t), \ldots, X_n(t)\}$ be a set of iid trajectories with the same distribution as $X(t)$. Assume that $X(t)$ has an unknown density function f.

Define $I_{(x,r)} = [x - r, x + r]$ and for $\{k_n\}$, $k_n < n$ for each n, a positive real numbers sequence converging to infinity, we define the random variable $H_n \doteq H_n(x)$ such that $\{X_1(t), \ldots, X_n(t)\}$ spend time k_n at $I_{(x,H_n(x))}$. That is,

$$k_n = \sum_{i=1}^{n} \int_T \mathbb{I}_{I_{(x,H_n(x))}}(X_i(t))\, dt = \sum_{i=1}^{n} \int_T \mathbb{I}_{\{|X_i(t)-x| \leq H_n(x)\}}(t)\, dt. \tag{2}$$

We define the estimator for the density f as

$$\widehat{f_n}(x) \doteq \frac{k_n}{2n|T|H_n(x)}.$$

2.1. Consistency, strong convergence rates and asymptotic distribution.

In Theorem 1 we prove, under mild conditions, the almost complete convergence of the estimator of the density f. Under some additional assumptions we obtain in Theorem 2 sharp bounds for strong rates of convergence and asymptotic normality with parametric rates of convergence in Theorem 3. The assumptions in these last theorems are closely related to those given in Castellana and Leadbetter [6] where it is shown that for continuous time processes a parametric speed of convergence is attained by kernel type density estimators.

The assumptions we consider are the following.

H1 The sequence $\{X_i(t), 1 \leq i \leq n, t \in T\}$ are iid random elements with the same distribution as $X(t)$, where $\{X(t) : t \in T\}$ is a stochastic process which admits a local time.

H2 $X(t)$ is a first order stationary stochastic process with an unknown density function f.

H3 $\{k_n\}$ is a positive real numbers sequence converging to infinity such that $k_n/n = o(1)$ and $\sum_{n=1}^{\infty} \exp(-ck_n) < \infty$ for each $c > 0$.

H4 The density f is a Lipschitz function.

H5 For each $c > 0$,

$$c^{-2}c_n^{-2}\int_T\int_T\int_{\{x:|u-x|\le cc_n\}}\int_{\{x:|v-x|\le cc_n\}}(f_{st}(u,v)-f(u)f(v))\,du\,dv\,ds\,dt$$

$$\to\int_{T\times T}(f_{st}(x,x)-f^2(x))ds\,dt \doteq c_0^2(x) > 0,$$

where $c_n = \frac{k_n}{n}$.

Theorem 1. Strong consistency. Under H1-H3, for almost all $x \in \mathbb{R}$,

$$\lim_{n\to\infty}\widehat{f_n}(x) = f(x) \quad \text{a.co.}$$

Here "a.co." stands for the almost complete convergence.

Remark 1. If f is Lipschitz the convergence is for all $x \in \mathbb{R}$.

Theorem 2. Strong rates of convergence. Assuming H1, H2 and H4, choose two sequences $\{k_n\}$ and $\{v_n\}$ of positive real numbers converging to infinity such that $(k_n/n)\,v_n = o(1)$, and that for each $c > 0$, $\sum_{n=1}^{\infty}\exp\left(-c\left(k_n/v_n\right)\right) < \infty$. For that k_n let us suppose that H5 holds. Then, for all $x \in \mathbb{R}$,

$$\lim_{n\to\infty} v_n(\widehat{f_n}(x) - f(x)) = 0 \quad \text{a.co.}$$

Remark 2. Suppose that $k_n = n^{\beta}$ and $v_n = n^{\gamma}$. Since the hypothesis $k_n/n = o(1)$ and $(k_n/n)\,v_n = o(1)$ have to be true we need $\beta < 1$ and $\beta - 1 + \gamma < 0$. Then, for any $\gamma < 1/2$, that is, $\gamma = 1/2 - \epsilon$ for some $0 < \epsilon < 1/2$, we can choose $\beta = 1/2 + \epsilon/2 < 1$ so that $\beta - 1 + \gamma < 0$.

Theorem 3. Asymptotic Normality. Assume H1, H2, H5 and that the density f has two bounded derivatives. If k_n is such that $\sqrt{n}/k_n = o(1)$, $k_n/n^{3/4} = o(1)$ then, for all $x \in \mathbb{R}$,

$$\sqrt{n}(\hat{f}_n(x) - f(x)) \to \mathcal{N}\left(0, \frac{2|T|}{c_0(x)}\right).$$

3. Estimating the density function of non–stationary processes.

In this section we define a density estimator for non–stationary stochastic processes. This means that, in model (1) $\mu(t)$ is deterministic but non constant with respect to t.

$$X(t) = \mu(t) + e(t),$$

where $\mu(t)$ is a continuous function and $e(t)$ is a zero mean, first-order stationary stochastic process with unknown density function f^e. The density function of $X(t)$ will be denoted $f^{X,t}$.

Let $\{X_1(t), \ldots, X_n(t)\}$ be independents trajectories with the same distribution as $X(t)$. We define the estimator of the density function $f^{X,t}$ as

$$\widehat{f}_n^{X,t}(x) = \widehat{f}_n^{u}(x - \overline{X}_n(t)), \tag{3}$$

where

$$\widehat{f}_n^{u}(x) \doteq \frac{k_n}{2n|T|H_n^u(x)}$$

with $u = \{U_{n1}, \ldots, U_{nn}\}$ given by

$$U_{ni}(t) = X_i(t) - \overline{X}_n(t) = e_i(t) - \overline{e}_n(t). \tag{4}$$

Here $\{e_1(t), \ldots, e_n(t)\}$ is a random sample of $e(t)$, $\overline{e}_n(t) = \dfrac{1}{n}\displaystyle\sum_{i=1}^{n} e_i(t)$ and H_n^u is defined as in (2) by replacing $\{X_1(t), \ldots, X_n(t)\}$ by u.

The sequence $\{U_{n1}(t), \ldots, U_{nn}(t)\}$ have the same density distribution for each t but they are not necessarily independent and therefore we can not use directly the theorems proved in Section 2.1. However we can still prove the complete convergence of the estimator of $f^{X,t}$, and obtain strong rates of convergence.

Theorem 4. *For fixed t, let $e(t)$ be a stochastic process satisfying the assumptions of Theorem 2, i.e. suppose H1, H2 and H4, holds with $e(t)$ instead of $X(t)$ and f^e instead of f. Choose two sequences $\{k_n\}$ and $\{v_n\}$ of positive real numbers converging to infinity such that $(k_n/n)\, v_n = o(1)$, for each $c > 0$, $\sum_{n=1}^{\infty} \exp\left(-c\,(k_n/v_n)\right) < \infty$ and $v_n(n/k_n)|\overline{e}_n(t)| \to 0$ a.co. For that k_n let us suppose that H5 holds with f^e instead of f. Then, for all $x \in \mathbb{R}$,*

$$\lim_{n\to\infty} v_n \left(\widehat{f}_n^{X,t}(x) - f^{X,t}(x)\right) = 0 \quad a.co.$$

Remark 3. Suppose that $k_n = n^{\beta}$ and $v_n = n^{\gamma}$. By Billingsley [2] we know that $\overline{e}_n(t) = o(n^{-\alpha})$ with $\alpha < 1/2$, on the other hand, since the hypothesis $\frac{k_n}{n} \to 0$ and $v_n \frac{n}{k_n}|\overline{e}_n(t)| \to 0$ have to be true we need $\beta < 1$ and $\gamma + 1 - \beta - \alpha < 0$. Then, for any $\gamma < 1/2$, that is, $\gamma = 1/2 - \epsilon$ for some $0 < \epsilon < 1/2$, we can choose $\beta = 1 - \epsilon/4 < 1$ and $\alpha = 1/2 - \epsilon/4 < 1/2$ so that $\gamma + 1 - \beta - \alpha < 0$.

3.1. k_n selection

In order to compute the number of "nearest-neighbors" k_n we use the least squared cross-validation method introduced by Rudemo [27] and Bowman [5]. They propose to compute the parameter k_n by minimizing an unbiased estimate $LS_n(k)$ of the mean integrated squared error $MISE_n(k) = \mathbb{E}(ISE_n(k))$ minus a quantity independent of the parameter k. The integrated squared error ISE_n is given by

$$\begin{aligned} ISE_n(k) &= \int_{-\infty}^{\infty} \left(\widehat{f_n}(x) - f(x)\right)^2 dx \\ &= \int_{-\infty}^{\infty} \widehat{f_n^2}(x)\, dx - 2\int_{-\infty}^{\infty} \widehat{f_n}(x) f(x)\, dx + \int_{-\infty}^{\infty} f^2(x)\, dx. \end{aligned}$$

Here $\int_{-\infty}^{\infty} f^2(x)\, dx$ does not depend on k and can be ignored as far as minimization over k is concerned. Therefore, instead of minimizing $ISE_n(k)$ we will minimize

$$LS_n(k) = \int_{-\infty}^{\infty} \widehat{f_n^2}(x)\, dx - 2\int_{-\infty}^{\infty} \widehat{f_n}(x) f(x)\, dx. \tag{5}$$

Let us note that $\int_{-\infty}^{\infty} \widehat{f_n^2}(x)\, dx$ can be calculated from the data. On the other hand, and since f in unknown, we will replace the second term

$$\int_{-\infty}^{\infty} \widehat{f_n}(x) f(x)\, dx = \mathbb{E}\left(\widehat{f_n}(X)\right),$$

by its leave-one-one estimator

$$\widehat{\mathbb{E}\left(\widehat{f_n}(X)\right)} = \frac{1}{nT}\sum_{i=1}^{n}\sum_{j=1}^{T} \widehat{f}_{n,-i}(X_i(t_j)),$$

whit

$$\widehat{f}_{n,-i}(x) = \frac{k}{2(n-1)|T|H_n(x)}$$

computed with all the elements in the sample except X_i. Therefore in (5) we get

$$LS_n(k) = \int_{-\infty}^{\infty} \widehat{f_n^2}(x)\, dx - \frac{2}{nT}\sum_{i=1}^{n}\sum_{j=1}^{T} \widehat{f}_{n,-i}(X_i(t_j)).$$

Is easy to prove that

$$\mathbb{E}(LS_n(k)) = \mathbb{E}(ISE_n(k)) - \int_{-\infty}^{\infty} f^2(x)\, dx,$$

so the value of k_n which minimizes $LS_n(k)$ indeed also minimizes an unbiased estimate of the mean integrated squared error of $\widehat{f_n}$.

4. A new classification rule for functional data.

In this section, we apply our estimation results to obtain a new classification rule for functional data. The main goal in pattern recognition or classification problems is to classify individuals into groups. Information about these groups is provided by a training sample $\{(X_i, Y_i) : 1 \leq i \leq n\}$, where each curve X_i has a label Y_i attached, indicating which group it belongs to. A new observation X is given without its label and we want to predict the unknown label.

The classical books by Devroye et al. [10], Duda and Stork [11] and Hastie et al. [14] provide a broad coverage of these topics, for the standard multivariate case where the variable $\mathbf{X}$ takes values in $\mathbb{R}^d$.

However, the definitions are mainly the same for an arbitrary metric space E. Given a finite set $\{1, \ldots, m\}$ and a metric space E, an observation is a pair $(x, y) \in E \times \{1, \ldots, m\}$, where x is known and y is a class or label that denotes the unknown nature of the observation. A mapping $g : E \to \{1, \ldots, m\}$ is called a classifier and represents our guess of the class y given its associated element $x \in E$. The classification is wrong if given an observation (x, y), $g(x) \neq y$.

Let $(X, Y) \in E \times \{1, \ldots, m\}$ be a random pair. Since an error occurs if $g(X) \neq Y$, probability of misclassification for g is

$$L(g) = P(g(X) \neq Y). \tag{6}$$

Then the best possible classifier is the function g^* that minimizes (6). The minimum error probability (the Bayes error) is denoted by $L^* = L(g^*)$.

To obtain g^*, the distribution of (X, Y) should be known, but this is not typically the case. One must build up a classifier based on a training sample of independent pairs $\{(X_i, Y_i); 1 \leq i \leq n\}$, with the same distribution as the pair (X, Y) and known $Y_1, \ldots, Y_n$ values. Then a classifier is a function

$$g_n(\,\cdot\,; X_1, Y_1, \ldots, X_n, Y_n) : E \times (E \times \{1, \ldots, m\})^n \to \{1, \ldots, m\},$$

with probability of misclassification given by the conditional error probability

$$L_n(g_n) = P(g_n(X; X_1, Y_1, \ldots, X_n, Y_n) \neq Y | X_1, Y_1, \ldots, X_n, Y_n).$$

A sequence of classifiers $\{g_n; n \geq 1\}$ is called a rule.

In the finite dimensional case, there are several universally consistent classification rules. An important difference between the finite and the infinite-dimensional situations arises, however, with regard to consistency. Stone [29] provide general results for the universal consistency of a wide class of nonparametric classification rules, which in particular imply the consistency of most classical nonparametric rules. In particular, the most popular k-NN classifiers are *universally (weakly) consistent*, provided that $k = k_n \to \infty$ and $k/n \to 0$, as $n \to \infty$. This means that, for these k-NN classifiers, $L_n \to L^*$, in probability, as $n \to \infty$ (or equivalently $E(L_n) \to L^*$), *with no restriction at all on the underlying distribution* of $(\mathbf{X}_i, Y_i)$. In the infinite-dimensional case his result is no longer true. This has been pointed out by Cèrou and Guyader [7] who have studied the consistency of the k-NN classifier when E is a metric space. Some recent results regarding classification methods for functional data can be found in Cuevas et al. [9], Cuevas and Fraiman [8], Cèrou and Guyader [7], Devroye et al. [10], Duda and Stork [11], Ghosh and Chaudhuri [13], James and Hastie [15], Li and Yu [18], Lopez-Pintado and Romo [19], Preda et al. [21] and Rossi and Villa [26].

4.1. The rule

Let $E = C(T, \|\cdot\|)$ be the space of continuous functions on T, and $\|\cdot\|$ a norm on E. We want to classify a new data $X(t) \in E$ into one of the m classes $\mathcal{F}_j$, $j \in \mathcal{J} = \{1, \ldots, m\}$ using a training sample $\{(X_i(t), Y_i); 1 \leq i \leq n\}$ of iid random elements with the same distribution as the pair $(X(t), Y)$ and known $Y_1, \ldots, Y_n$ labels. We will assume that for each population $\mathcal{F}_j$ the model (1) holds; that is

$$X_{(j)}(t) = \mu_{(j)}(t) + e_{(j)}(t),$$

where $\mu_{(j)}(t)$ stands for the mean function of the population $\mathcal{F}_j$, $j \in \mathcal{J}$ and $e_{(j)}(t)$ is a zero mean, first-order stationary stochastic process with density unknown function $f_{(j)}$.

For each t fixed, the Bayes rule chooses the class $\mathcal{F}_j$ if and only if

$$f_{(j)}(X(t) - \mu_{(j)}(t)) > f_{(k)}(X(t) - \mu_{(k)}(t)), \quad \forall\, k \neq j,$$

This motives defining our classification rule in the following way: we will classify $X(t)$ into the class $\mathcal{F}_j$, $j \in \mathcal{J}$ (and define $\widehat{Y} = j$) if and only if

$$\lambda\left(\left\{t : \widehat{f}_{n,j}^{X,t}(X(t)) > \widehat{f}_{n,k}^{X,t}(X(t))\right\}\right) > \lambda\left(\left\{t : \widehat{f}_{n,j}^{X,t}(X(t)) \leq \widehat{f}_{n,k}^{X,t}(X(t))\right\}\right), \quad \forall\, k \neq j. \tag{7}$$

where $\widehat{f}_{n,(j)}^{X,t}$ is the estimator of $f_{(j)}^{X,t}$.

5. Algorithm and simulation studies.

In order to illustrate the use of our estimation method, in this section, we perform some simulation studies of nonparametric functional density estimation and functional discrimination.

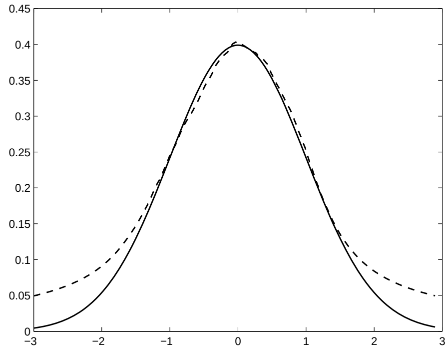

Figure 1: Estimated (dashed curve) and theoretical (solid curve) density functions of $X(t)$ for $k_n = 43.196$.

5.1. Algorithm.

We built two samples from the original data set, the *learning sample* $(X_i(t), Y_i)_{i\in\mathcal{L}}$ and the *testing sample* $(X_j(t), Y_j)_{j\in\mathcal{T}}$. With the learning sample we compute the density estimator for each group $(\widehat{f}_{n,(1)}, \ldots, \widehat{f}_{n,(m)})$ using the crossvalidated value $\widehat{k}_{n,(j)}$, $j = 1, \ldots, m$. In order to measure the discriminant power of our method, we evaluate the estimators obtained with the learning samples at the testing sample and we classify it according to the rule given by (7). Finally, we compute the *Misclassification Rate* as

$$Misclas = \frac{1}{\sharp\mathcal{T}} \sum_{j\in\mathcal{T}} \mathbb{I}_{\{\widehat{Y}_j \neq Y_j\}}$$

Example 1. Let $X(t)$ be s stochastic process defined by

$$X(t) = \mu(t) + \sigma e(t), \qquad t \in T = [0, 1], \tag{8}$$

where

$$e(t) = \frac{w(t)}{\sqrt{t}}, \quad \text{with } w(t) \text{ the standard Wiener process.}$$

In a first stage we consider $\mu(t) = 0$ and $\sigma = 1$ so that $X(t)$ is stationary and, for each t, $X(t) \sim \mathcal{N}(0, 1)$. Figure 1 shows the theoretical density function of $X(t)$ and its estimator computed from a sample of size 200 measured at 100 equally spaced points on $[0, 1]$. As we can see in Figure 1 the estimator fits very well the true density except in the tails where, due to the nature of the processes, we have not enough data to perform a good estimation.

To assess the performance of our classification method, in a second stage we considered two classes under the model (8), both with $\sigma = 1$ but one of them with μ constant and equal to 0, and the another one with mean $\mu \neq 0$. In particular, we will consider the

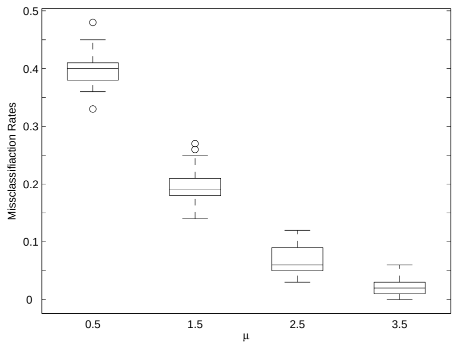

Figure 2: Boxplot of the misclassification error from 50 runs.

cases where $\mu = 0.5, 1.5, 2.5, 3.5$. We generate a learning sample of size 200 (100 of each class) measured at 150 instants of time in the interval $[0, 1]$ and a testing sample of the same size. With the learning sample we compute the estimators for each class and then, in order to obtain the misclassification error, we evaluate them in the testing sample. We repeat this procedure 50 times in order to obtain 50 misclassification rates for each case which are shown in Figure 2.

Note how the errors decrease exponentially: this is due to the fact that, since $\sigma = 1$, when we classify two populations which are very close in mean they are almost superposed. For instance, when we classify the population with mean 0 against the one with mean 0.5 (see Figure 3 (a)) we obtain big errors, as we can see in the first box of Figure 2. However as the means goes far apart (see Figure 3 (b)) our classification method has a good behavior as we can see in the last box of 2.

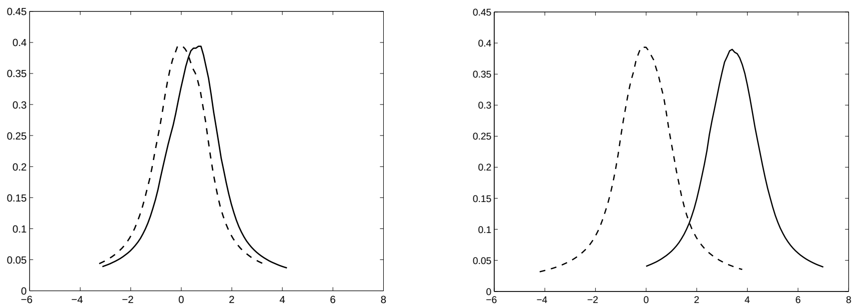

Figure 3: (a) Density Estimator for a $\mathcal{N}(0, 1)$ (dashed curve) and for a $\mathcal{N}(0.5, 1)$ (solid curve). (b) Density Estimator for a $\mathcal{N}(0, 1)$ (dashed curve) and for a $\mathcal{N}(3.5, 1)$ (solid curve).

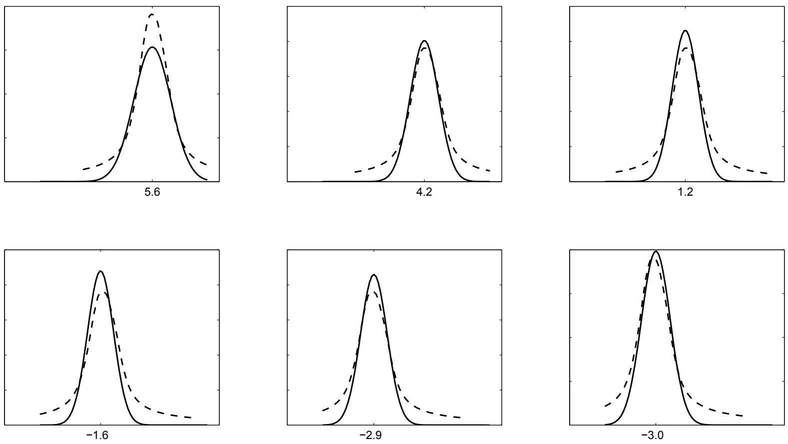

Figure 4: Estimated (dashed curve) and theoretical (solid curve) density functions of $X_1(t)$ for $t = 0.013, 0.180, 0.347, 0.513, 0.680, 0.847$.

Example 2. Let $X_1(t)$ and $X_2(t)$ as defined in Shin [28],

$$X_1(t) = 3\sqrt{2}\cos(\pi t) + \sqrt{2}cos(2\pi t) + S(t) \quad \text{and} \quad X_2(t) = \sqrt{2}cos(2\pi t) + S(t),$$

with

$$S(t) = \sum_{i=1}^{30} i^{-1/2} U_i \sqrt{2}\cos(i\pi t),$$

where U_i are iid standard normal random variables. Let us observe that for each t,

$$X_1(t) \sim \mathcal{N}\left(3\sqrt{2}\cos(\pi t) + \sqrt{2}\cos(2\pi t), 2\sum_{i=1}^{30}\cos^2(\pi t i)/i\right)$$

and

$$X_2(t) \sim \mathcal{N}\left(\sqrt{2}\cos(2\pi t), 2\sum_{i=1}^{30}\cos^2(\pi t i)/i\right),$$

For each class, we generate a learning sample of size 200 (100 of each class) measured at 150 instants of time in the interval [0, 1] and a testing sample of the same size. Figure 4 and Figure 5 shows, respectively, the density estimator and the theoretical density function of $X_1(t)$ and $X_2(t)$ for some instants of time, the dashed curves correspond to the density estimators and the solid curves correspond to the theoretical density function. As in the stationary case (Example 1) the density estimator fits very well the true density except in the tails where we do not have enough information.

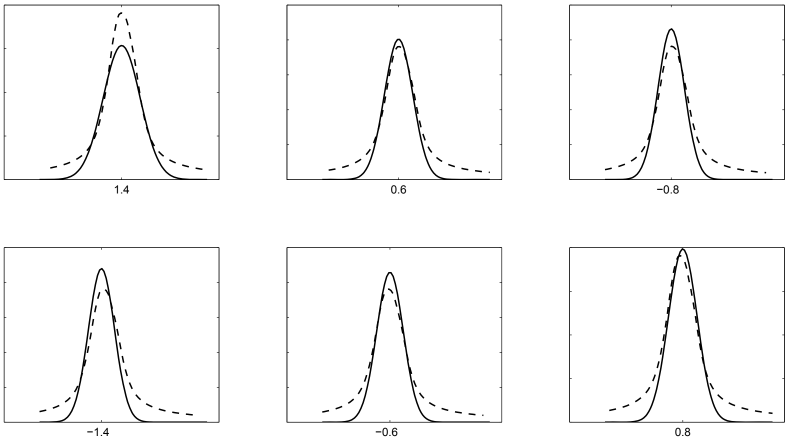

Figure 5: Estimated (dashed curve) and theoretical (solid curve) density functions of $X_2(t)$ for $t = 0.013, 0.180, 0.347, 0.513, 0.680, 0.847$.

Next, we evaluate this estimators in the testing sample in order to obtain the misclassification error. This procedure was replicated an additional 49 times by randomly building 49 learning samples and 49 testing samples. Finally, we get 50 misclassification rates. Figure 6 shows that our classification rule works as well as the classification defined in the original paper (Shin [28]).

As we can see in this example, the processes $e_1(t) = X_1(t) - 3\sqrt{2}\cos(\pi t) + \sqrt{2}\cos(2\pi t) = S(t)$ and $e_2(t) = X_2(t) - \sqrt{2}\cos(2\pi t) = S(t)$ are not stationaries since their variance (and consequently their distribution) depend on the time. This shows that our classification method is robust to the non stationarity.

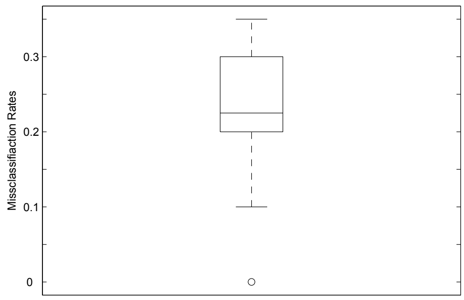

Figure 6: Boxplot of the misclassification error from 50 runs.

6. Conclusions.

In this paper we have proposed a nonparametric density estimation method for functional data following the model

$$X(t) = \mu(t) + e(t), \quad t \in T,$$

where $\mu(t)$ stands for the deterministic mean function, and $e(t)$ is a zero mean, first-order stationary stochastic process which admits a local time with density unknown function f.

When $\mu(t) = \mu$ is constant, $X(t)$ inherits the properties of $e(t)$ which means that it is stationary. In this context we obtained an estimator for the marginal density function of $X(t)$, which is the same for all t. We show that it is strongly consistent with rate of convergence n^{α}, for any $\alpha < 1/2$ and that it has asymptotic normal distribution with rate $\sqrt{n}$. Though this is not new in nonparametric setting (see Castellana and Leadbetter [6]), it is a surprising and desired property.

When $\mu(t)$ is nonconstant with respect to the time, $X(t)$ does not inherit the stationarity of $e(t)$ and it therefore has a different marginal density function for each t. In this context the estimator has shown to be strongly consistent for each t with the same rate as before.

In simulations studies, we computed the density estimator and we applied the estimation results to obtain a new classification rule for functional data.

Appendix A.

Proof of Theorem 1. Let

$$C_n = \left\{ \left| \widehat{f_n}(x) - f(x) \right| > \epsilon \right\}.$$

By definition of almost complete convergence we need to show that $\sum_{n=1}^{\infty} P(C_n) < \infty$ for all $\epsilon > 0$. Let fixed x so that the Lebesgue Differentiation Theorem holds for f in x. For this x we can write

$$C_n = A_n \cup B_n,$$

with

$$A_n = \left\{ H_n(x) < \frac{k_n}{2n|T|(f(x) + \epsilon)} \right\}$$

and

$$B_n = \begin{cases} \left\{ H_n(x) > \frac{k_n}{2n|T|(f(x)-\epsilon)} \right\} & \text{if } f(x) > \epsilon \\ \emptyset & \text{if } f(x) \leq \epsilon. \end{cases}$$

It will be sufficient to show that

$$\sum_{n=1}^{\infty} P(A_n) < \infty \quad \text{and} \quad \sum_{n=1}^{\infty} P(B_n) < \infty. \tag{A.1}$$

The proof of the right side inequality of (A.1) is similar to the one of the left side inequality and therefore it will be omitted. In order to prove the left side inequality of (A.1) let us define $a_n = \frac{k_n}{2n|T|(f(x)+\epsilon)}$. Then, we have

$$H_n < a_n \Leftrightarrow \sum_{i=1}^{n} \underbrace{\int_T \mathbb{I}_{I_{(x,a_n)}}(X_i(t))\, dt}_{\doteq Y_{ni}} > k_n.$$

where Y_{ni} are independent random variables and

$$P(A_n) = P\left(\sum_{i=1}^{n} Y_{ni} > k_n\right).$$

Let $p_n = \int_{\{u:|u-x|\le a_n\}} f(u)du$. Since the marginal distributions of X are the same for each t we get

$$\mathbb{E}\,(Y_{ni}) \;=\; \mathbb{E}\left(\int_T \mathbb{I}_{I_{(x,a_n)}}(X_i(t))\, dt\right) = \int_T P(X_i(t) \in I_{(x,a_n)})\, dt = |T| p_n \tag{A.2}$$

and using Cauchy-Schwartz inequality

$$\mathbb{E}\left(Y_{ni}^2\right) \;=\; \mathbb{E}\left(\left(\int_T \mathbb{I}_{I_{(x,a_n)}}(X_i(t))\, dt\right)^2\right) \le |T|\,\mathbb{E}\left(\int_T \left(\mathbb{I}_{I_{(x,a_n)}}(X_i(t))\right)^2\, dt\right) = |T|^2 p_n.$$

Let us define $\widetilde{Y}_{ni} \doteq Y_{ni} - \mathbb{E}\,(Y_{ni})$. Then, using (A.2), we get

$$\begin{aligned} \sum_{n=1}^{\infty} P(A_n) &= \sum_{n=1}^{\infty} P\left(\sum_{i=1}^{n} \widetilde{Y}_{ni} > k_n - n|T|p_n\right) \\ &= \sum_{n=1}^{\infty} P\left(\sum_{i=1}^{n} \widetilde{Y}_{ni} > k_n\left(1 - \frac{p_n}{2a_n(f(x)+\epsilon)}\right)\right); \end{aligned} \tag{A.3}$$

where in the last equality we have used the relation between a_n and k_n. Since $a_n \to 0$ by the Lebesgue Differentiation Theorem we have that $p_n/2a_n \to f(x)$. Therefore, there exists $N_1 = N_1(x)$, such that if $n \ge N_1(x)$

$$\left|\frac{p_n}{2a_n} - f(x)\right| < \epsilon/2 \Rightarrow 1 - \frac{p_n}{2a_n(f(x)+\epsilon)} > \frac{\epsilon}{2(f(x)+\epsilon)} = C(x,\epsilon) := C. \tag{A.4}$$

By Bernstein inequality (Bernstein [1]), using this last statement and the fact $\left|\widetilde{Y}_{ni}\right| \leq 2|T|$, $\operatorname{var}\left(\widetilde{Y}_{ni}\right) \leq |T|^2 p_n$ we conclude that for $n \geq N_1(x)$,

$$\begin{aligned} P\left(\sum_{i=1}^{n} \widetilde{Y}_{ni} > k_n \left(1 - \frac{p_n}{2a_n(f(x)+\epsilon)}\right)\right) &\leq P\left(\sum_{i=1}^{n} \widetilde{Y}_{ni} > k_n C\right) \\ &\leq 2\exp\left(-\frac{k_n^2 C^2}{2n|T|^2 p_n + 4|T|k_n C}\right). \end{aligned}$$

In order to bound the exponent we use that (A.4) implies $p_n < (f(x)+\epsilon)2a_n = k_n/n|T|$ and as a consequence

$$\frac{k_n^2 C^2}{2n|T|^2 p_n + 4|T|k_n C} > \frac{k_n^2 C^2}{2|T|k_n + 4|T|k_n C} = k_n \frac{C^2}{2|T|(1+2C)}.$$

Replacing this bound into (A.3) we get, for $n \geq N_1(x)$, that

$$P\left(\sum_{i=1}^{n} \widetilde{Y}_{ni} > k_n \left(1 - \frac{p_n}{2a_n(f(x)+\epsilon)}\right)\right) \leq 2\exp(-ck_n)$$

with $c = \frac{C^2}{2|T|(1+2C)}$. Finally hypothesis H3 implies $\sum_{n=N_1(x)}^{\infty} \exp(-k_n c) < \infty$ and the Theorem follows.

Proof of Theorem 2. Let

$$C_n = \left\{ v_n \left| \widehat{f}_n(x) - f(x) \right| > \epsilon \right\},$$

we need to prove that $\sum_{n=1}^{\infty} P(C_n) < \infty$ for all $\epsilon > 0$. We do the analysis analogous that in Theorem 1, replacing ϵ by $\epsilon_n = \frac{\epsilon}{v_n}$ and we get (A.3) for ϵ_n. This is

$$\sum_{n=1}^{\infty} P(A_n) = \sum_{n=1}^{\infty} P\left(\sum_{i=1}^{n} \widetilde{Y}_{ni} > k_n \left(1 - \frac{p_n}{2a_n(f(x)+\epsilon_n)}\right)\right); \tag{A.5}$$

The Mean Value Theorem and the Lipschitz Condition for f ensure the existence of $x_n \in I(x, a_n)$ for which $p_n/2a_n = f(x_n)$. Using this and the Lipschitz Condition again we get

$$\left|\frac{p_n}{2a_n} - f(x)\right| = |f(x_n) - f(x)| \leq K|x_n - x| \leq K a_n. \tag{A.6}$$

Now, by the definition of a_n, the fact that $\epsilon_n \to 0$ and the hypothesis $(k_n/n)v_n = o(1)$, for all $\epsilon > 0$ there exists N_1 such that for $n \geq N_1(x)$

$$Ka_n \leq C_1(x)\frac{k_n}{n} \leq \frac{1}{2}\frac{\epsilon}{v_n} = \frac{\epsilon_n}{2}. \tag{A.7}$$

Therefore from (A.6) and (A.7) we have

$$1 - \frac{1}{f(x)+\epsilon_n}\frac{p_n}{2a_n} \geq \frac{\epsilon_n - Ka_n}{f(x)+\epsilon_n} \geq \frac{C_2}{v_n}.$$

With this in (A.5),

$$\sum_{n=1}^{\infty} P(A_n) \leq \sum_{n=1}^{\infty} P\left(\sum_{i=1}^{n} \widetilde{Y}_{ni} > C_2\frac{k_n}{v_n}\right).$$

Now, since $a_n \sim k_n/n$, from H5 we get

$$\frac{1}{a_n^2}\text{var}\left(\widetilde{Y}_{ni}\right) = \tag{A.8}$$

$$= \frac{1}{a_n^2}\int_T\int_T\int_{\{u:|u-x|\leq a_n\}}\int_{\{v:|v-x|\leq a_n\}} \left(f_{st}(u,v) - f(u)f(v)\right)\,du\,dv\,dt\,ds \to c_0^2 \geq 0,$$

then for $n \geq N_2(x)$, $\text{var}\left(\widetilde{Y}_{ni}\right) \leq C_3 a_n^2$. Applying Bernstein inequality with $\left|\widetilde{Y}_{ni}\right| \leq 2|T|$, $\text{var}\left(\widetilde{Y}_{ni}\right) \leq C_3 a_n^2$, we get for $n \geq N_3(x) = \max\{N_1(x), N_2(x)\}$

$$\begin{aligned}\sum_{n=1}^{\infty} P(A_n) &\leq \sum_{n=1}^{\infty} P\left(\sum_{i=1}^{n}\widetilde{Y}_{ni} > C_2\frac{k_n}{v_n}\right) \\ &\leq 2\exp\left\{-\left(\frac{C_3}{C_4\frac{k_n}{n}+C_5}\right)\frac{k_n}{v_n}\right\}.\end{aligned} \tag{A.9}$$

In order to bound the exponent we use the fact that $k_n/n \to 0$ and then we get

$$\sum_{n=1}^{\infty} P(A_n) \leq 2\sum_{n=1}^{\infty}\exp\left\{-C_6\frac{k_n}{v_n}\right\} < \infty,$$

Finally using that $\sum_{n=1}^{\infty}\exp\left(-c\left(k_n/v_n\right)\right) < \infty$, for each $c > 0$ we get the Theorem.

Proof of Theorem 3. We do the analysis analogous to Theorem 1 where we replace ϵ by $\epsilon_n = \frac{t}{\sqrt{n}}$ and we get calling $S_n = \sum_{i=1}^{n} Y_{ni}$ and $s_n^2 = \mathrm{var}(S_n)$

$$P\left(\sqrt{n}\left(\hat{f}_n(x) - f(x)\right) \le t\right) \le P\left(\frac{S_n - E(S_n)}{s_n} \le \frac{k_n - n|T|p_n}{s_n}\right) \tag{A.10}$$

Now, since by (A.8) $s_n^2 = O(na_n^2)$, by Lindenberg Theorem

$$\frac{S_n - E(S_n)}{s_n} \to N(0,1), \tag{A.11}$$

where the convergence is in distribution. On the other hand,

$$\begin{aligned} \frac{k_n - n|T|p_n}{s_n} &= \frac{2na_n|T|(f(x) + t/\sqrt{n}) - n|T|\int_{x-a_n}^{x+a_n} f(u)du}{s_n} \\ &= s_n^{-1}n|T|\int_{x-a_n}^{x+a_n}(f(x) - f(u))du + \frac{2na_n|T|t}{s_n\sqrt{n}}. \end{aligned} \tag{A.12}$$

By Taylor Theorem, there exists a number x^* between x and u such that

$$\int_{x-a_n}^{x+a_n}(f(x) - f(u))\,du = -\frac{1}{2}\int_{x-a_n}^{x+a_n} f''(x^*)(u-x)^2\,du.$$

Since f has two derivatives bounded

$$\left|s_n^{-1}n|T|\int_{x-a_n}^{x+a_n}(f(x) - f(u))du\right| \le Cs_n^{-1}na_n^3.$$

Therefore, in (A.12) we have

$$\frac{k_n - n|T|p_n}{s_n} = O(s_n^{-1}na_n^3) + \frac{2na_n|T|t}{s_n\sqrt{n}}.$$

Since $s_n^{-1}na_n^3 \to 0$ and by (A.8) $s_n^2/(na_n^2) \to c_o^2$,

$$\lim_{n\to\infty}\frac{k_n - E(S_n)}{s_n} = \frac{2|T|}{c_0}t.$$

Finally from this, (A.11) and (A.10) we get the result.

In order to prove Theorem 4 we need to define the quantities

$$\widehat{f_n^e}(x) = \frac{k_n}{2n|T|H_n^e(x)} \quad \text{with} \quad k_n = \sum_{i=1}^{n} \int_T \mathbb{I}_{\{|e_i(t)-x|\leq H_n^e(x)\}}(t)\, dt. \tag{A.13}$$

Observe that these are theoretical quantities since the $e_i(t)$ are not observable and they can not be computed.

Lemma 5. *Under H1-H4 with $e(t)$ instead of $X(t)$ and f^e instead of f, for fixed t, for all $x \in \mathbb{R}$*

$$\lim_{n\to\infty} v_n \left(\widehat{f_n^u}(x - \overline{X}_n(t)) - \widehat{f_n^e}(x - \mu(t)) \right) = 0, \qquad a.co.$$

In order to prove this Lemma we need an auxiliary result:

Lemma 6. *For fixed t, let H_n^u and $\overline{e}_n(t)$ as defined in (4) where $u = \{U_{n1}, \ldots, U_{nn}\}$ with $U_{ni}(t) = X_i(t) - \overline{X}_n(t) = e_i(t) - \overline{e}_n(t)$ and H_n^e as defined in (A.13). Then, for each n, x,*

$$\left| H_n^u(x - \overline{X}_n(t)) - H_n^e(x - \mu(t)) \right| \leq 2|\overline{e}_n(t)|.$$

Proof of Lemma 6. It is an immediate consequence of

(i) $\left| H_n^u(x - \overline{X}_n(t)) - H_n^u(x - \mu(t)) \right| \leq |\overline{e}_n(t)|.$

(ii) $\left| H_n^u(x - \mu(t)) - H_n^e(x - \mu(t)) \right| \leq |\overline{e}_n(t)|.$

We will prove only (i) since the proof of (ii) is analogous. Let x fixed, using that $\overline{X}_n(t) = \mu(t) + \overline{e}_n(t)$ we get

$$\left\{ t : \left| U_{ni}(t) - (x - \overline{X}_n(t)) \right| \leq H_n^u(x - \overline{X}_n(t)) \right\} \subset \left\{ t : |U_{ni}(t) - (x - \mu(t))| \leq H_n^u(x - \overline{X}_n(t)) + |\overline{e}_n(t)| \right\},$$

therefore

$$k_n = \sum_{i=1}^{n} \int_T \mathbb{I}_{\{|U_{ni}(t)-(x-\overline{X}_n(t))|\leq H_n^u(x-\overline{X}_n(t))\}}(t)\, dt \leq \sum_{i=1}^{n} \int_T \mathbb{I}_{\{|U_{ni}(t)-(x-\mu(t))|\leq H_n^u(x-\overline{X}_n(t))+|\overline{e}_n(t)|\}}(t)\, dt,$$

and for the definition of k_n we get

$$H_n^u(x - \mu(t)) \leq H_n^u(x - \overline{X}_n(t)) + |\overline{e}_n(t)|. \tag{A.14}$$

In the same way we can prove that

$$H_n^u(x - \overline{X}_n(t)) \leq H_n^u(x - \mu(t)) + |\overline{e}_n(t)|. \tag{A.15}$$

And from (A.14) and (A.15) we have

$$\left| H_n^u(x - \overline{X}_n(t)) - H_n^u(x - \mu(t)) \right| \leq |\overline{e}_n(t)|.$$

Proof of Lemma 5. Let $\epsilon > 0$, x, t fixed, $\widehat{f_n^e}(\cdot)$ as defined in (A.13), then

$$\begin{aligned} v_n \left| \widehat{f_n^u}(x - \overline{X}_n(t)) - \widehat{f_n^e}(x - \mu(t)) \right| &= \frac{k_n v_n}{2n|T|} \frac{\left| H_n^e(x - \mu(t)) - H_n^u(x - \overline{X}_n(t)) \right|}{H_n^u(x - \overline{X}_n(t)) H_n^e(x - \mu(t))} \\ &\leq \frac{k_n v_n}{2n|T|} \frac{2|\overline{e}_n(t)|}{H_n^u(x - \overline{X}_n(t)) H_n^e(x - \mu(t))}. \end{aligned}$$

Where in the last inequality we have used Lemma 6. By Theorem 1, $\widehat{f_n^e}(x - \mu(t)) < f^e(x - \mu(t)) + \epsilon$ for all $n \geq N_1(x, \epsilon, t)$, this implies that

$$H_n^e(x - \mu(t)) > C_1(x, t, \epsilon) \frac{k_n}{n} = C_1 \frac{k_n}{n}. \tag{A.16}$$

By Lemma 6 and (A.16)

$$H_n^u(x - \overline{X}_n(t)) + 2|\overline{e}_n(t)| \geq C_1 \frac{k_n}{n}. \tag{A.17}$$

Since by hypothesis $v_n \frac{n}{k_n} |\overline{e}_n(t)| \to 0$, for all $n \geq N_2(x, \epsilon, t)$ we have that

$$\frac{n}{k_n} |\overline{e}_n(t)| \leq \frac{1}{4} C_1 \quad \text{and therefore} \quad C_1 \frac{k_n}{n} - 2|\overline{e}_n(t)| \geq \frac{1}{2} C_1 \frac{k_n}{n}.$$

Replacing in (A.17) we obtain

$$H_n^u(x - \overline{X}_n(t)) \geq C_1 \frac{k_n}{n} - 2|\overline{e}_n(t)| \geq \frac{C_1}{2} \frac{k_n}{n}. \tag{A.18}$$

So from (A.16) and (A.18), for all $n \geq \max\{N_1, N_2\}$ we get

$$\begin{aligned} v_n \left| \widehat{f_n^u}(x - \overline{X}_n(t)) - \widehat{f_n^e}(x - \mu(t)) \right| &< \frac{1}{|T|} \frac{k_n v_n}{n} \frac{|\overline{e}_n(t)|}{\frac{C_1}{2} \frac{k_n}{n} C_1 \frac{k_n}{n}} \\ &= C_3 v_n \frac{n}{k_n} |\overline{e}_n(t)|. \end{aligned}$$

Finally,

$$\begin{aligned} \sum_{n=1}^{\infty} P\left(v_n \left| \widehat{f_n^u}(x - \overline{X}_n(t)) - \widehat{f_n^e}(x - \mu(t)) \right| \geq \epsilon \right) &\leq \sum_{n=1}^{\infty} P\left(C_3 v_n \frac{n}{k_n} |\overline{e}_n(t)| \geq \epsilon \right) \\ &= \sum_{n=1}^{\infty} P\left(v_n \frac{n}{k_n} |\overline{e}_n(t)| \geq \frac{\epsilon}{C_3} \right) \\ &= \sum_{n=1}^{\infty} P\left(v_n \frac{n}{k_n} |\overline{e}_n(t)| \geq \epsilon_0 \right) < \infty, \end{aligned}$$

which concludes the proof.

PROOF OF THEOREM 4. By definition (3) and since $X(t)$ satisfies the model (1) we need to prove

$$\lim_{n\to\infty} v_n \left(\widehat{f}_n^u(x - \bar{X}_n(t)) - f^e(x - \mu(t)) \right) = 0, \qquad a.co.$$

Let $\epsilon > 0$ and x, t fixed.

$$\begin{aligned} \left\{ v_n \left| \widehat{f}_n^u(x - \bar{X}_n(t)) - f^e(x - \mu(t)) \right| \geq \epsilon \right\} \quad \subset \quad & \left\{ v_n \left| \widehat{f}_n^u(x - \bar{X}_n(t)) - \widehat{f}_n^e(x - \mu(t)) \right| \geq \epsilon/2 \right\} \\ & \cup \left\{ v_n \left| \widehat{f}_n^e(x - \mu(t)) - f^e(x - \mu(t)) \right| \geq \epsilon/2 \right\} = I \cup II. \end{aligned}$$

The result follows applying Lemma 5 and Theorem 2.

References

[1] S. Bernstein, Probability Theory, M.L.Gostechizdat, 1946.

[2] P. Billingsley, Convergence of Probability masures, John Willey, 1968.

[3] D. Blanke, Adaptive sampling schemes for density estimation, Journal of Statistical Planning and Inference 136 (9) (2004) 2898–2917.

[4] D. Blanke, D. Bosq, Accurate rates of density estimators for continuous-time processes, Statist. Probab. Lett. 33 (2) (1997) 185–191.

[5] A. Bowman, An alternative method of cross-validation for the smoothing of density estimates, Biometrika 71 (2) (1984) 353–360.

[6] J. V. Castellana, M. R. Leadbetter, On smoothed probability density estimation for stationary processes, Stochastic Process. Appl. 21 (2) (1986) 179–193.

[7] F. Cèrou, A. Guyader, Nearest neighbor classification in infinite dimension, ESAIM Probab. Stat. 10 (2006) 340–355.

[8] A. Cuevas, R. Fraiman, On depth measures and dual statistics. A methodology for dealing with general data, J. Multivariate Anal. 100 (4) (2009) 753–766.

[9] A. Cuevas, M. Febrero, R. Fraiman, Robust estimation and classification for functional data via projection-based depth notions, Comput. Statist. 22 (3) (2007) 481–496.

[10] L. Devroye, L. Gyorfi, G. Lugosi, A Probabilistic Theory of Pattern Recognition, Springer-Verlag, 1996.

[11] P. Duda, R.O. Hart, D. Stork, Pattern Classification, 2nd edition, Wiley, 2000.

[12] F. Ferraty, P. Vieu, Nonparametric Functional Data Analysis. Theory and Practice, Springer, New York, 2006.

[13] A. K. Ghosh, P. Chaudhuri, On maximal depth and related classifiers, Scand. J. Statist. 32 (2005) 327–350.

[14] T. Hastie, R. Tibshirani, J. Friedman, The Elements of Statistical Learning, Springer-Verlag, New York, 2001.

[15] G. M. James, T. J. Hastie, Functional linear discriminant analysis for irregularly sampled curves, J. Roy. Statist. Soc. Ser. B 63 (2001) 533–550.

[16] Y. Kutoyants, On invariant density estimation for ergodic diffusion processes, SORT. 28 (2) (2004) 111–124.

[17] B. Labrador, Strong pointwise consistency of the k_T-occupation time density estimator, Statist. Probab. Lett. 78 (9) (2008) 1128–1137.

[18] B. Li, Q. Yu, Classification of functional data: A segmentation approach, Comput. Stat. Data Anal. 52 (10) (2008) 4790–4800.

[19] S. Lopez-Pintado, J. Romo, Depth based classification for functional data, DIMACS Series in Discrete Mathematics 72 (2006) 103–120.

[20] H. Nguyen, Density Estimation in a Continuous-Time Stationary Markov Process, Ann. Statist. 7 (2) (1979) 341–348.

[21] C. Preda, G. Saporta, C. Lévéder, PLS classification of functional data, Comput. Statist. 22 (2) (2007) 223–235.

[22] J. Ramsay, C. Dalzell, Some tools for functional data analysis, Journal of the Royal Statistical Society B 53 (1991) 539–572.

[23] J. Ramsay, B. Silverman, Applied Functional Data Analysis, Springer Series in Statistics, Springer, New York, 2002.

[24] J. Ramsay, B. Silverman, Functional Data Analysis, Springer Series in Statistics, Springer, New York, 2005.

[25] M. Rosenblatt, Density estimates and Markov Sequences, Nonparametric Techniques in Statistical Inference, Cambridge Univ. Press. Mathematical Reviews (1970) 199–210.

[26] F. Rossi, N. Villa, Support vector machine for functional data classification, Neurocomputing 69 (2006) 730–742.

[27] M. Rudemo, Empirical choice of histograms and kernel density estimators, Scand. J. Statist. 9 (2) (1982) 65–78.

[28] H. Shin, An extension of Fisher's discriminant analysis for stochastic processes, J. Multivariate Anal. 99 (2008) 1191–1216.

[29] C. Stone, Consistent non-parametric regression., Ann. Statist. 8 (1977) 348–360.