Supervised classification for a family of Gaussian functional models

Amparo Baíllo*, Juan Antonio Cuesta-Albertos†and Antonio Cuevas*

*Universidad Autónoma de Madrid and †Universidad de Cantabria

Abstract

In the framework of supervised classification (discrimination) for functional data, it is shown that the optimal classification rule can be explicitly obtained for a class of Gaussian processes with "triangular" covariance functions. This explicit knowledge has two practical consequences. First, the consistency of the well-known nearest neighbors classifier (which is not guaranteed in the problems with functional data) is established for the indicated class of processes. Second, and more important, parametric and nonparametric plug-in classifiers can be obtained by estimating the unknown elements in the optimal rule.

The performance of these new plug-in classifiers is checked, with positive results, through a simulation study and a real data example.

1 Introduction

Statement of the problem. Notation

Discrimination, also called "supervised classification" in modern terminology, is one of the oldest statistical problems in experimental science: the aim is to decide whether a random observation X (taking values in a "feature space" $\mathcal{F}$ endowed with a distance D) either belongs to the population P_0 or to P_1. For example, in a medical problem P_0 and P_1 could correspond to the group of "healthy" and "ill" individuals, respectively. The decision must be taken from the information provided by a "training sample" $\mathcal{X}_n = \{(X_i, Y_i), 1 \leq i \leq n\}$. Here X_i, $i = 1, \ldots, n$, are independent replications of X, measured on n randomly chosen individuals, and Y_i are the corresponding values of an indicator variable which takes values 0 or 1 according to the membership of the

*These authors have been partially supported by Spanish grant MTM2007-66632.

†This author have been partially supported by the Spanish grant MTM2008-0607-C02-02.

E-mail addresses: amparo.baillo@uam.es, cuestaj@unican.es, antonio.cuevas@uam.es

i-th individual to P_0 or P_1. The term "supervised" refers to the fact that the individuals in the training sample are supposed to be correctly classified, typically using "external" non statistical procedures, so that they provide a reliable basis for the assignation of the new observation. It is possible to consider the case where $K > 2$ populations, $P_0, \ldots, P_{K-1}$ are involved but, in what follows, we will restrict ourselves to the binary case $K = 2$.

The mathematical problem is to find a "classifier" (or "classification rule") $g_n(x) = g_n(x; \mathcal{X}_n)$, with $g_n : \mathcal{F} \to \{0, 1\}$, that minimizes the classification error $\mathbb{P}\{g_n(X) \neq Y\}$. It is not difficult to prove (e.g., Devroye *et al.*, 1996, p. 11) that the optimal classification rule (often called "Bayes rule") is

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

where $\eta(x) = \mathbb{E}(Y|X = x)$ and $\mathbb{I}_A$ stands for the indicator function of a set $A \subset \mathcal{F}$. Of course, since η is unknown the exact expression of this rule is usually unknown, and thus different procedures have been proposed to approximate g^* using the training data.

From now on we will use the following notation. Let μ_i be the distribution of X conditional on $Y = i$, that is, $\mu_i(B) = \mathbb{P}\{X \in B|Y = i\}$ for $B \in \mathcal{B}_{\mathcal{F}}$ (the Borel σ-algebra on $\mathcal{F}$) and $i = 0, 1$. We denote by $S_i \subset \mathcal{F}$ the support of μ_i, for $i = 0, 1$, $S = S_0 \cap S_1$ and $p = \mathbb{P}\{Y = 0\}$ (we assume $0 < p < 1$). Given two measures μ and ν, the expression $\mu << \nu$ denotes that μ is absolutely continuous with respect to ν (i.e., $\nu(B) = 0$ implies $\mu(B) = 0$).

The notation $C[0, 1]$ stands for the space of real continuous functions on the interval $[0, 1]$ endowed with the usual supremum norm, denoted by $\|\cdot\|$. The subspace of functions of class 2 (i.e. with two continuous derivatives) is denoted by $C^2[0, 1]$.

Finite dimensional spaces. Three classical discrimination procedures

The origin of the discrimination problem goes back to the classical work by Fisher (1936) where, in the d-variate framework $\mathcal{F} = \mathbb{R}^d$, a simple "linear classifier" of type $g_n(x) = \mathbb{I}_{\{w'x+w_0>0\}}$ was introduced for the case that both populations P_0 and P_1 are homoscedastic, that is, have a common covariance matrix Σ. Intuitively, $w'x + w_0 = 0$ is chosen as the affine hyperplane which provides the "maximum separation" between both populations. It is well-known (see, e.g., Duda *et al.* 2000 for details) that the the expression of Fisher's rule turns out to depend on the inverse Σ^{-1} of the covariance matrix. It is also known that Fisher's linear rule is in fact the optimal one (1) when the conditional distributions of $X|Y = 0$ and $X|Y = 1$ are homoscedastic normals and all the means and covariances are known. These conditions look quite restrictive but, as argued by Hand (2006) in a provocative paper, Fisher's rule (or rather its sampling approximation obtained by estimating the unknown parameters) is hard to beat in

practical examples. That is, while it is not difficult to construct examples where this rule outrageously fails, its performance is quite good in most cases found in real-life examples. For this reason, Fisher's linear rule is still the most popular classification tool among practitioners, in spite of the posterior intensive research on this topic. Thus, in a way, Fisher's rule represents a sort of "golden standard" in the multivariate statistical discrimination problem.

The books by Devroye *et al.* (1996), Duda *et al.* (2000) and Hastie *et al.* (2001) offer different interesting perspectives of the work done in discrimination theory since Fisher's pioneering paper. All of them focus on the standard multivariate case $\mathcal{F} = \mathbb{R}^d$. Many classifiers have been proposed as an alternative to Fisher's linear rule in this finite-dimensional setup. One of the simplest and easiest to motivate is the so-called k-nearest neighbors method. Fixed a positive integer value (or smoothing parameter) $k = k_n$ this rule simply classifies an incoming observation x in the population P_1 if the majority among the k training observations closest to x (with respect to the considered distance D) belong to P_1. More concretely the k-NN rule can be defined by

$$g_n(x) = \mathbb{I}_{\{\eta_n(x)>1/2\}}, \tag{2}$$

where

$$\eta_n(x) = \frac{1}{k}\sum_{i=1}^{n} \mathbb{I}_{\{X_i \in k(x)\}} Y_i \tag{3}$$

and "$X_i \in k(x)$" means that X_i is one of the k nearest neighbors of x.

In fact, the definition of the k-NN rule is extremely simple and can be introduced (in terms of "majority vote among the neighbors") with no explicit reference to any regression estimator. However, the idea of replacing the unknown regression function $\eta(x)$ in the optimal classifier (1) with a regression estimator (given by (3) in the case of the k-NN rule) is very natural. It suggests a general methodology to construct a wide class of classifiers by just plugging in different regression estimators η_n in (1) instead of the true regression function $\eta(x)$. In the finite dimensional case $\mathcal{F} = \mathbb{R}^d$ this is a particularly fruitful idea, as a wealth of different (parametric and nonparametric) estimators of $\eta(x)$ is available; see Audibert and Tsybakov (2007) for some reasons in favor of the plug-in methodology in classification. The main purpose of this work is to show that the *plug-in methodology* can be also successfully used for classification in some functional data models.

Discrimination of functional data. Differences with the finite-dimensional case

We are concerned here with the problem of (binary) supervised classification with functional data. That is, we assume throughout that the space $(\mathcal{F}, D)$ where the data

X_i live is a separable metric space (typically a space of functions). For some theoretical results, considered below, we will impose more specific assumptions on $\mathcal{F}$.

The study of discrimination techniques with functional data is not as developed as the corresponding finite-dimensional theory but, clearly, is one of the most active research topics in the booming field of functional data analysis (FDA). Two well-known books including broad overviews of FDA with interesting examples are Ferraty and Vieu (2006) and Ramsay and Silverman (2005). A recent survey on supervised and unsupervised classification with functional data can be found in Baíllo *et al.* (2009).

While the formal statement of the functional classification problem is very much the same as that indicated at the beginning of this section, there are some important differences with the classical finite-dimensional case.

(a) *Lack of a simple functional version of Fisher's linear rule*: As mentioned above, the idea behind Fisher's rule requires to invert the covariance operator. When $\mathcal{F} = \mathbb{R}^d$ this is increasingly difficult as the dimension d increases, but it becomes impossible in the functional framework where the operator is typically not invertible. Thus the applicability of Fisher's linear methodology to functional data is a non-trivial issue of current interest for research. See, for instance, James and Hastie (2001) and Shin (2008) for interesting adaptations of linear discrimination ideas to a functional setting.

(b) *Difficulty to implement the plug-in idea*: Unlike the finite-dimensional case, the plug-in methodology is not generally considered as a standard procedure to construct functional classifiers. When x is infinite-dimensional, there are yet few simple parametric models giving a good fit to the regression function and the structure of nonparametric estimators of η is relatively complicated.

(c) *The k-NN functional classifier is not universally consistent*: In the discrimination problem a sequence of classifiers $\{g_n\}$, based on samples of size n, is said to be "consistent" when the corresponding sequence of classification errors converges, as n tends to infinity, to the "lowest possible error" attained by the Bayes classifier (1); see Section 3 below for more details. It turns out (see Stone, 1977) that, in the case of finite-dimensional data $X_i \in \mathbb{R}^d$, any sequence of k-NN classifiers is consistent provided that $k_n \to \infty$ and $k_n/n \to 0$. Since such consistency holds irrespectively of the distribution of the data (X, Y), this property is called "universal consistency".

The definition of the k-NN classifier can be easily translated to the functional setup (by replacing the usual Euclidean distance in $\mathbb{R}^d$ with an appropriate functional metric D). However, the universal consistency is lost. Cérou and Guyader

(2006, Th. 2) have obtained sufficient conditions for consistency of the k-NN classifier when X takes values in a separable metric space. Nevertheless, the required assumptions are not always trivial to check. As the k-NN rule is a natural "default choice" in infinite-dimensional setups, an important issue is to ensure its consistency, at least for some functional models of practical interest.

The purpose and structure of this paper

This work aims to partially fill the gaps pointed out in the points (b) and (c) of the above paragraph. To this end, in Subsection 2.1 a simple expression is obtained for the Bayes (optimal) rule g^* in the case that both distributions, μ_0 and μ_1, are equivalent. However, g^* turns out to depend on the Radon-Nikodym derivative $d\mu_0/d\mu_1$ which is usually unknown, or has an extremely involved expression, even when μ_0 and μ_1 are completely known. An interesting exception is given by Gaussian processes with a specific type of covariance functions, called "triangular". For these processes the Radon-Nikodym derivative has been explicitly calculated by Varberg (1961) and Jørsboe (1968) whose results are collected and briefly commented in Subsection 2.2. In Subsection 2.3 parametric plug-in estimators for g^* are obtained by assuming that μ_0 and μ_1 are either (parametric) Brownian motions or Ornstein-Uhlenbeck processes. Non-parametric plug-in estimators for g^* are proposed and analyzed in Subsection 2.4, under the sole assumption that the covariance functions are triangular. Since the proofs of the results in this subsection are rather technical, they are deferred to a final appendix. This concludes our contributions regarding issue (b). Section 3 is devoted to the k-NN consistency problem introduced in (c): we use the above-mentioned result by Cérou and Guyader (2006) to show that the k-NN rule is consistent in functional classification problems where the data are generated by certain Gaussian triangular processes specified in Subsection 2.2.

Finally, in Section 4 the practical performance of the plug-in rules proposed in Section 2 is checked, and compared with the k-NN rule, through a simulation study and the analysis of a real data example.

2 The optimal classifier for a Gaussian family

2.1 A general expression based on Radon-Nikodym derivatives

When the distributions μ_0 and μ_1 of P_0 and P_1 are both absolutely continuous with respect to some common σ-finite measure μ, it is easy to see, as a consequence of Bayes formula, that the optimal rule is

$$g^*(x) = \mathbb{I}_{\{(1-p)f_1(x) > pf_0(x)\}}, \tag{4}$$

where $p = \mathbb{P}\{Y = 0\}$ and f_0, f_1 are the μ-densities of P_0 and P_1, respectively.

The expression (4) is particularly important in the finite dimensional problems with $\mathcal{F} = \mathbb{R}^d$, where the Lebesgue measure μ arises as the natural reference measure and the corresponding Lebesgue densities can be estimated in many ways. In the infinite-dimensional spaces there is no such obvious dominant measure. However if we assume that μ_0 and μ_1, with supports S_0 and S_1, are absolutely continuous with respect to each other on $S_0 \cap S_1$, the optimal rule can be also expressed in a simple way with respect to the Radon-Nikodym derivative $d\mu_0/d\mu_1$ as shown in the following result.

Theorem 1 *Assume that $\mu_0 << \mu_1$ and $\mu_1 << \mu_0$ on $S = S_0 \cap S_1$. Then*

$$\eta(x) \;=\; \begin{cases} 0 & \text{if } x \in S_0 \cap S^c \\ 1 & \text{if } x \in S_1 \cap S^c \\ \dfrac{1-p}{p\frac{d\mu_0}{d\mu_1}(x) + 1 - p} & \text{if } x \in S. \end{cases} \tag{5}$$

provides the expression for the optimal rule $g^*(x) = \mathbb{I}_{\{\eta(x) > 1/2\}}$.

Proof: Define $\mu = \mu_0 + \mu_1$. Then $\mu_i << \mu$, for $i = 0, 1$, and we can define the Radon-Nikodym derivatives $f_i = d\mu_i/d\mu$, for $i = 0, 1$. From the definition of the conditional expectation we know that $\eta(x) = E(Y|X = x) = P(Y = 1|X = x)$ can be expressed by

$$\eta(x) = \frac{f_1(x)(1-p)}{f_0(x)p + f_1(x)(1-p)}. \tag{6}$$

Observe that $\mu|_{S^c \cap S_i} = \mu_i|_{S^c \cap S_i}$ and thus $f_i|_{S^c \cap S_i} = \mathbb{I}_{S^c \cap S_i}$, for $i = 0, 1$. Since $\mu_0 << \mu_1$ and $\mu_1 << \mu_0$ on S then, on this set, there exists the Radon-Nikodym derivatives $d\mu_0/d\mu_1$ and $d\mu_1/d\mu_0$. In this case, it also holds that $\mu|_S << \mu_i|_S$, for both $i = 0, 1$ and

$$\frac{d\mu}{d\mu_i}(x) = 1 + \frac{d\mu_{1-i}}{d\mu_i}(x), \qquad \text{for any } x \in S.$$

Then (see, e.g., Folland 1999), for $i = 0, 1$ and for P_X-a.e. $x \in S$,

$$f_i(x) = \frac{d\mu_i}{d\mu}(x) = \left(\frac{d\mu}{d\mu_i}(x)\right)^{-1} = \frac{1}{1 + \frac{d\mu_{1-i}}{d\mu_i}(x)} \tag{7}$$

Substituting (7) into expression (6) we get (5). □

The mutual absolute continuity is not a very restrictive assumption if we deal with Gaussian measures. According to a well-known result by Feldman and Hájek (see Feldman, 1958) for any given pair of Gaussian processes, there is a dichotomy in such a way that they are either equivalent or mutually singular. In the first case both measures μ_0 and μ_1 have a common support S. As for the identification of the support, Vakhania

(1975) has proved that if a Gaussian process, with trajectories in a separable Banach space $\mathcal{F}$, is not degenerate (i.e., the distribution of any non-trivial linear continuous functional is not degenerate) then the support of such process is the whole space $\mathcal{F}$.

In any case, expression (5) would be of no practical use unless some expressions, reasonably easy to estimate, can be found for the Radon-Nikodym derivative $d\mu_0/d\mu_1$. This issue is considered in the next subsection.

2.2 Explicit expression for a family of Gaussian distributions

The best known Gaussian process is perhaps the standard Brownian motion $\{W(t),\ t \geq 0\}$, for which $\mathbb{E}(W(t)) = 0$ and the covariance function is $\mathrm{Cov}(W(s), W(t)) := \Gamma(s,t) = \min(s,t)$. A wide class of Brownian-type processes can be obtained by location and scale changes of type $m(t) + \sigma W(t)$, where $m(t)$ is a given mean function and $\sigma > 0$.

In fact, the covariance structure $\Gamma(s,t) = \min(s,t)$ can be generalized to define a much broader class of processes with $\Gamma(s,t) = u(\min(s,t))\, v(\max(s,t))$, where u and v denote suitable real functions. Covariance functions of this type are called *triangular*. They have received considerable attention in the literature. For example, Sacks and Ylvisaker (1966) use this condition in the study of optimal designs for regression problems where the errors are generated by a zero mean process with covariance function $\Gamma(s,t)$. It turns out that the Hilbert space with reproducing kernel K plays an important role in the results and, as these authors point out, the norm of this space is particularly easy to handle when Γ is triangular. On the other hand, Varberg (1964) has given an interesting representation of the processes $X(t),\ 0 \leq t < b$, with zero mean and triangular covariance function. This author proved that they can be expressed in the form $X(t) = \int_0^b W(u)\, d_u R(t,u)$, where W is the standard Wiener process and $R = R(t,u)$ is a function, of bounded variation with respect to u, defined in terms of Γ.

The so-called Ornstein-Uhlenbeck model, for which $\Gamma(s,t) = \sigma^2 \exp(-\beta|s-t|)$ $(\beta, \sigma > 0)$, provides another important class of processes with triangular covariance functions. They are widely used in physics and finance.

The following theorem is due to Varberg (1961, Th. 1) and Jørsboe (1968, p. 61). It shows that the Radon-Nikodym derivative can be expressed in a closed, relatively simple way for these special classes of Gaussian processes. For more information concerning explicit expressions of Radon-Nikodym derivatives for Gaussian processes see Segall and Kailath (1975) and references therein. From now on let us denote $m_i(t) = \mathbb{E}\,(X(t)|Y = i)$.

Theorem 2 *Let $(\mathcal{F}, D) = (C[0,1], \|\cdot\|)$. Assume that $X|Y = i$, for $i = 0, 1$, are Gaussian processes on $[0,1]$, with covariance functions $\Gamma_i(s,t) = u_i(\min(s,t))\, v_i(\max(s,t))$, for $s, t \in [0,1]$, where u_i, v_i, for $i = 0, 1$, are positive functions in $C^2[0,1]$. Assume*

also that v_i, for $i = 0, 1$, and $v_1 u_1' - u_1 v_1'$ are bounded away from zero on $[0, 1]$, that $u_1 v_1' - u_1' v_1 = u_0 v_0' - u_0' v_0$ and that $u_1(0) = 0$ if and only if $u_0(0) = 0$.

a) *Assume that $m_i \equiv 0$, for $i = 0, 1$. Then there exist some constants C_1, C_2, C_3 and a function F, whose expressions are given in the proof, such that*

$$\frac{d\mu_0}{d\mu_1}(x) = C_1 \exp\left[\frac{1}{2}\left(C_3 x^2(0) + C_2 x^2(1) - \int_0^1 \frac{x^2(t)}{v_0(t)v_1(t)} dF(t)\right)\right]. \tag{8}$$

b) *Assume now that the covariance functions are identical, i.e. $u_i = u$ and $v_i = v$ for $i = 0, 1$, that $m_1 \equiv 0$, m_0 is a function $m \in C^2[0, 1]$, such that $m(0) = 0$ whenever $u(0) = 0$. Then there exist some constants D_1, D_2 and a function G, whose expressions are given in the proof, such that*

$$\frac{d\mu_0}{d\mu_1}(x) = \exp\left\{D_1 + \left(D_2 - 2\,\frac{G(0)}{v(0)}\right) x(0) + 2\,\frac{G(1)}{v(1)}\, x(1) - 2\int_0^1 \frac{x(t)}{v(t)}\, dG(t)\right\}. \tag{9}$$

PROOF:

a) Varberg (1961, Th. 1) shows that, under the assumptions of (a), μ_0 and μ_1 are equivalent measures. The Radon-Nikodym derivative of μ_0 with respect to μ_1 is

$$\frac{d\mu_0}{d\mu_1}(x) = C_1 \exp\left\{\frac{1}{2}\left[C_4 x^2(0) + \int_0^1 F(t) d\left(\frac{x^2(t)}{v_0(t)v_1(t)}\right)\right]\right\}, \tag{10}$$

where

$$C_1 = \begin{cases} \left(\frac{v_0(0)v_1(1)}{v_0(1)v_1(0)}\right)^{1/2} & \text{if } u_0(0) = 0 \\ \left(\frac{u_1(0)v_1(1)}{v_0(1)u_0(0)}\right)^{1/2} & \text{if } u_0(0) \neq 0 \end{cases} \qquad C_4 = \begin{cases} 0 \text{ if } u_0(0) = 0 \\ \left(\frac{v_0(0)u_0(0) - u_1(0)v_1(0)}{v_1(0)v_0(0)u_0(0)u_1(0)}\right)^{1/2} & \text{if } u_0(0) \neq 0 \end{cases}$$

and $F = (v_1 v_0' - v_0 v_1')/(v_1 u_1' - u_1 v_1')$.

Observe that, by the assumptions of the theorem, F is differentiable with bounded derivative. Thus F is of bounded variation and it may be expressed as the difference of two bounded positive increasing functions. Therefore the stochastic integral (10) is well defined and it can be evaluated integrating by parts, leading to conclusion (8), with $C_3 = C_4 - F(0)/v_0(0)v_1(0)$ and $C_2 = F(1)/v_0(1)v_1(1)$.

b) In Jørsboe (1968), p. 61, it is proved that, under the indicated assumptions, μ_0 and μ_1 are equivalent measures with Radon-Nikodym derivative

$$\frac{d\mu_0}{d\mu_1}(x) = \exp\left\{D_3 + D_2\, x(0) + \frac{1}{2}\int_0^1 G(t) d\left(\frac{2x(t) - m(t)}{v(t)}\right)\right\},$$

with

$$D_3 = -\frac{m^2(0)}{2\,u(0)\,v(0)}\mathbb{I}_{\{u(0)>0\}}, \quad D_2 = \frac{m(0)}{u(0)\,v(0)}\mathbb{I}_{\{u(0)>0\}}$$

and $G = (vm' - mv')/(vu' - uv')$. Again, the integration by parts gives (9), where $D_1 = D_3 - \int_0^1 G\,d(m/v)$. □

In the general case where $m_0 \neq m_1$ and $\Gamma_0 \neq \Gamma_1$, let us denote by $P_{m,\Gamma}$ the distribution of the Gaussian process with mean m and covariance function Γ. Then, applying the chain rule for Radon-Nikodym derivatives (see, e.g., Folland, 1999) we get

$$\frac{d\mu_0}{d\mu_1}(x) = \frac{dP_{m_0,\Gamma_0}}{dP_{m_1,\Gamma_1}}(x) = \frac{dP_{m_0,\Gamma_0}}{dP_{0,\Gamma_0}}(x)\,\frac{dP_{0,\Gamma_0}}{dP_{0,\Gamma_1}}(x)\,\frac{dP_{0,\Gamma_1}}{dP_{m_1,\Gamma_1}}(x). \tag{11}$$

Under the appropriate assumptions the expressions of the Radon-Nikodym derivatives in the right-hand side of (11) are given in (8) and (9).

2.3 Parametric plug-in rules

The aim of this subsection is twofold. First and foremost, we show how the theoretical results of Subsections 2.1 and 2.2 become useful in practice. To this end, we consider examples of well-known Gaussian processes that fulfill the requirements of Theorems 1 and 2, namely Brownian motions with drift and Ornstein-Uhlenbeck processes. We derive the expressions of the Radon-Nikodym derivatives $d\mu_0/d\mu_1$ for these examples. Then, it is straightforward to compute the Bayes rule g^* for classification between two elements of one of these families. In these particular examples the mean and variance of the Gaussian process $X|Y = i$ have known parametric expressions (up to a finite number of parameters). Thus g^* is completely specified as long as the parameters have known values. When this is not the case, we can substitute each unknown parameter in g^* by some estimate. The resulting discrimination procedure is called the *parametric plug-in* rule. In particular, for the Bayes rules given in (12), (13), (14) and (15) below the explicit expression of the parameter estimates is given in the appendix.

The second objective of Subsection 2.3 is to obtain the expressions of the Bayes rules for the models used in Section 4 and to derive the corresponding parametric plug-in versions.

Two Brownian motions

Let us denote $X(t;i) = (X(t)|Y = i)$. In the Brownian case, using the standard notation in stochastic differential equations, $X(t;i)$ is just the solution of $dX(t;i) = m_i(t)\,dt + \sigma_i W_i(t)\,dt$, for $i = 0, 1$ and $t \in [0,1]$. Here $m_1 \equiv 0$, $m_0(t) = ct$, $0 < c < \infty$ is a constant, W_0 and W_1 are two uncorrelated Brownian motions and $(X(0;i) \sim N(0, \theta_i^2)$.

Then, if $\sigma_0 = \sigma_1 = \sigma$, the conditions of Theorem 2 are satisfied with $u_i(t) = \theta_i^2 + \sigma^2 t$ and $v_i \equiv 1$, for $i = 0, 1$.

When $\theta_0 = \theta_1 = 0$, we have $X(0; i) \equiv 0$ and, for any $x \in S$,

$$\frac{d\mu_0}{d\mu_1}(x) = \exp\left\{\frac{c}{\sigma^2}(2\,x(1) - c)\right\}.$$

Thus the Bayes rule is

$$g^*(x) = \mathbb{I}_{\{x(1)<c/2\}}. \tag{12}$$

If $\theta_i \neq 0$ for $i = 0, 1$, then $X(0; i)$ is random and a similar calculation yields that the Bayes rule classifies x in population P_1 whenever

$$\frac{c}{\sigma^2}\left[2(x(1) - x(0)) - c\right] + \frac{1}{2}\left(\frac{1}{\theta_1^2} - \frac{1}{\theta_0^2}\right) x^2(0) < \log\left(\frac{\theta_0}{\theta_1}\right). \tag{13}$$

Replacing the unknown parameters, c, σ and θ_i in (12) and (13) by estimates, we obtain the corresponding *parametric plug-in* rules.

When $\sigma_0 \neq \sigma_1$, then $u_i(t) = \theta_i^2 + \sigma_i^2 t$, $v_i \equiv 1$, for $i = 0, 1$, and the hypothesis $u_1 v_1' - u_1' v_1 = u_0 v_0' - u_0' v_0$ in Theorem 2 is not satisfied. In fact, if this last equality does not hold, by Theorem 1 in Varberg (1961) we know that μ_0 and μ_1 are mutually singular.

Two Ornstein-Uhlenbeck processes

Let $X|Y = i$, for $i = 0, 1$, be Ornstein-Uhlenbeck processes given by

$$dX(t; i) = -\,\beta_i\,(X(t; i) - \eta_i)\,dt + \sqrt{2\beta_i}\,\sigma_i\,dW_i(t),$$

where W_0 and W_1 are two independent Brownian motions and $\beta_i > 0$, $\sigma_i > 0$, η_i are constants.

If $X(0; i)$ is equal to a constant c_i, we have that $m_i(t) = \eta_i + (c_i - \eta_i)e^{-\beta_i t}$ and $\Gamma_i(s, t) = \sigma_i^2\left(e^{-\beta_i|s-t|} - e^{-\beta_i|s+t|}\right)$. Fixing $v_i(1) = 1$, we get $u_i(t) = \sigma_i^2 e^{-\beta_i}(e^{\beta_i t} - e^{-\beta_i t})$ and $v_i(t) = e^{\beta_i(1-t)}$ for $i = 0, 1$. The condition $u_1 v_1' - u_1' v_1 = u_0 v_0' - u_0' v_0$ in Theorem 2 is fulfilled if and only if $\beta_0\sigma_0^2 = \beta_1\sigma_1^2$. Also, since $u_i(0) = 0$, then $m_i(0) = c_i$ has to be 0 for $i = 0, 1$. Then it is straightforward to check that the Bayes rule g^* classifies x in population P_1 if

$$\begin{aligned} 0 > 2\left(\beta_0^2(\sigma_0^2 - \eta_0^2) - \beta_1^2(\sigma_1^2 - \eta_1^2)\right) + 4\,x(1)(\eta_0\beta_0 - \eta_1\beta_1) + (\beta_1 - \beta_0)\,x^2(1) \\ + 4\,(\eta_0\beta_0^2 - \eta_1\beta_1^2)\int_0^1 x(t)\,dt + (\beta_1^2 - \beta_0^2)\int_0^1 x^2(t)\,dt. \end{aligned} \tag{14}$$

When $X(0; i)$ is random, it follows a normal distribution with mean η_i and variance σ_i^2. Then $m_i(t) = \eta_i$, for all $t \in [0, 1]$, and $\Gamma_i(s, t) = \sigma_i^2 e^{-\beta_i|s-t|}$, $u_i(t) = \sigma_i^2 e^{-\beta_i(1-t)}$ and

$v_i(t) = e^{\beta_i(1-t)}$. Consequently, the Bayes rule assigns x to population P_1 if

$$2\beta_1\sigma_1^2(\log(\beta_1) - \log(\beta_0)) > 2\left[\beta_0^2\sigma_0^2 - \beta_1^2\sigma_1^2 + \beta_1\eta_1^2(1+\beta_1) - \beta_0\eta_0^2(1+\beta_0)\right]$$
$$+4\,x(1)(\eta_0\beta_0 - \eta_1\beta_1) + 4\,(\eta_0\beta_0^2 - \eta_1\beta_1^2)\int_0^1 x(t)\,dt$$
$$+(\beta_1 - \beta_0)\left[x^2(0) + x^2(1) + (\beta_1+\beta_0)\int_0^1 x^2(t)\,dt\right]. \quad (15)$$

The parametric plug-in classification rule is derived by substituting the unknown parameters β_i, η_i and σ_i, $i = 0, 1$, in (14) and (15) with their corresponding estimators.

2.4 Nonparametric plug-in rules

In this section we analyze the situation in which the processes ultimately belong to the Gaussian family fulfilling the conditions of Theorem 2, but we do not place any parametric assumption on the mean and the covariance functions. However, let us note that, until we get to the estimation of the Radon-Nikodym derivatives, the Gaussianiaty assumption is not needed. Specifically, we only assume that the covariance functions of the involved processes are of type $\Gamma(s,t) = u(\min(s,t))v(\max(s,t))$, for some (unknown) real functions u, v where v is bounded away from 0 on the interval $[0,1]$.

Observe that, in order to use a plug-in version of the optimal classification rule along the lines of Theorems 1 and 2, we need to estimate the functions m, u and v as well as their first and second derivatives. Since these estimation problems have some independent interest, in this subsection we consider them in a general setup, not necessarily linked to the classification problem. Thus we use the ordinary iid sampling model with a fixed sample size denoted, for simplicity, by n in all cases.

Regarding u and v, let us note that the condition $\Gamma(s,t) = u(\min(s,t))v(\max(s,t))$, for $s,t \in [0,1]$, entails $u(s) = \Gamma(s,1)/v(1)$ and $v(t) = \Gamma(0,t)/u(0)$ if $u(0) > 0$. However, it is clear that these conditions only determine u and v up to multiplicative constants so that one can impose (without loss of generality) the additional assumption $v(1) = 1$. Thus, it turns out that u and v can be uniquely determined in terms of $\Gamma(0,t)$ and $\Gamma(s,1)$. Our study will require three steps: first, the estimation of the mean function m and its derivatives, then the analogous study for $\Gamma(0,t)$, $\Gamma(s,1)$ and $\sigma^2(t) := \Gamma(t,t)$ and, finally, the analysis of more involved functions defined in terms of these.

In Propositions 1 to 3 below we assume that the sample data are $X_1, \ldots, X_n$, iid trajectories of a process X in the space $C[0,1]$, endowed with the supremum norm, $\|\cdot\|$.

Estimation of the mean and covariance functions and their derivatives

To estimate the mean function $m(t) = \mathbb{E}[X(t)]$ and its derivatives, we will only need to assume that $\{X_n\}$ satisfies that $\mathbb{E}\|X_1\|^2 < \infty$, which (see p. 172 in Araujo and Giné,

1980) implies that the distribution of X_1 satisfies the Central Limit Theorem (CLT) in $(C[0,1], \|\cdot\|)$.

The natural estimator of m is the sample mean, denoted by $\hat{m}_n(t) = \sum_{i=1}^n X_i(t)/n$. Since the derivatives of m are also involved in the expressions of the Radon-Nikodym derivatives obtained in Theorem 2, we will also need to consider the estimation of m' and m''. Our estimators will depend on a given sequence $h_n \downarrow 0$ of smoothing parameters. Given $t \in [h_n, 1-h_n]$, define

$$\hat{m}_n'(t) := \frac{\hat{m}_n(t+h_n) - \hat{m}_n(t-h_n)}{2h_n}, \qquad \hat{m}_n''(t) := \frac{\hat{m}_n(t+h_n) + \hat{m}_n(t-h_n) - 2m_n(t)}{h_n^2}.$$

For $t \in [0, h_n)$, we define

$$\hat{m}_n'(t) := \frac{\hat{m}_n(t+h_n) - \hat{m}_n(0)}{h_n + t}, \qquad \hat{m}_n''(t) := \frac{\hat{m}_n(t+h_n) + \hat{m}_n(0) - 2\hat{m}_n(\gamma_n)}{\gamma_n^2}.$$

where $\gamma_n = (t+h_n)/2$. The definition of $\hat{m}_n'$ and $\hat{m}_n''$ on $(1-h_n, 1]$ is similar. These definitions allow us to handle analogously the extreme points and the inner ones. Thus we will not pay special attention to the extreme points in the proofs.

There is a slight notational abuse in these definitions as, for example, $\hat{m}_n'(t)$ is not the derivative of $\hat{m}_n(t)$ but an estimator of $m'(t)$. We keep this notation throughout the manuscript for simplicity.

As mentioned at the beginning of this section, due to the triangular structure of Γ, in principle we should only concentrate on the estimation of the functions $s \mapsto \Gamma(s,1)$ and $t \mapsto \Gamma(0,t)$ and their derivatives. However, due to technical reasons we will also need to consider the function $\sigma^2(t) = \Gamma(t,t)$ and its derivatives. Natural nonparametric estimators of these functions can be given in terms of the empirical covariance

$$\hat{\Gamma}_n(s,t) := \frac{1}{n}\sum_i \left(X_i(s) - \hat{m}_n(s)\right)\left(X_i(t) - \hat{m}_n(t)\right), \ s,t \in [0,1].$$

The estimation of the required derivatives is carried out in an analogous way as we did with the mean function. Observe finally that, since $v(1) = 1$, we can estimate $u(t) = \Gamma(t,1)$ by $\hat{u}_n(t) := \hat{\Gamma}_n(t,1)$ for any $t \in [0,1]$ and similarly for its first two derivatives. Regarding the function σ^2, we estimate $\sigma^2(t)$ by $\hat{\sigma}_n^2(t) := \hat{\Gamma}_n(t,t)$.

Proposition 1 *Let $\{X_n\}$ be iid trajectories in $C[0,1]$ of a process such that $\mathbb{E}\|X_1\|^2 < \infty$ and whose mean function $m : [0,1] \to \mathbb{R}$ has a Lipschitz second derivative.*

a) *For the mean estimation problem we have,*

$$\|m - \hat{m}_n\| = O_P(n^{-1/2}) \tag{16}$$

$$\|m' - \hat{m}_n'\| = O_P\left((n^{1/2}h_n)^{-1}\right) + O(h_n^2) \tag{17}$$

$$\|m'' - \hat{m}_n''\| = O_P\left((n^{1/2}h_n^2)^{-1}\right) + O(h_n) \tag{18}$$

b) *Assume that* $\mathbb{E}\|X_1\|^4 < \infty$ *and that the functions* $t \to \Gamma(t,1)$, $t \to \Gamma(0,t)$ *and* σ^2 *admit Lipschitz second order derivatives. Then, we have*

$$\|\hat{\Gamma}_n(\cdot,1) - \Gamma(\cdot,1)\| = \|\hat{u}_n - u\| = O_P(n^{-1/2}), \tag{19}$$

$$\|\hat{\Gamma}'_n(\cdot,1) - \Gamma'(\cdot,1)\| = \|\hat{u}'_n - u'\| = O_P\left(\left(n^{1/2}h_n\right)^{-1}\right) + O(h_n^2), \tag{20}$$

$$\|\hat{\Gamma}''_n(\cdot,1) - \Gamma''(\cdot,1)\| = \|\hat{u}''_n - u''\| = O_P\left(\left(n^{1/2}h_n^2\right)^{-1}\right) + O(h_n), \tag{21}$$

Similar results also hold for $\hat{\Gamma}_n(0,\cdot)$ *and* $\hat{\sigma}_n^2$.

From the proof of this proposition (see the Appendix) it can be checked that the assumption $\mathbb{E}\|X_1\|^4 < \infty$ can be replaced with $\mathbb{E}\|X_1\|^{2+\delta} < \infty$, for some $\delta > 0$, and $\mathbb{E}(X^r(1)) < \infty$ for any $r > 0$.

Estimation of v

The estimation of v is harder than that of u. It will be useful to distinguish two cases, where the estimators must be defined in different ways. In the case $u(0) > 0$ (corresponding to the case $\sigma^2(0) > 0$) we have $v(t) = \Gamma(0,t)/u(0)$ which is estimated by

$$\hat{v}_n(t) := \frac{1}{\hat{u}_n(0)}\hat{\Gamma}_n(0,t), t \in [0,1]. \tag{22}$$

When $u(0) = 0$ (which implies that $\sigma^2(0) = 0$), the estimator proposed in (22) is, at best, highly unstable. This case is not unusual: see, for instance, the examples introduced in Subsection 2.3 when $X(0)/Y = i$ is constant. For the sake of simplicity from now on assume that $\sigma^2(t) > 0$ for $t \in (0,1)$.

The first step is to define $\hat{v}_n(t) = \hat{\sigma}_n^2(t)/\hat{u}_n(t)$ for $t \in [\delta_n, 1]$, where δ_n is a sequence of positive numbers converging to zero (whose rate will be determined later). Then we define estimates for the first and the second derivatives of v on the same interval. The structure of v_n as a quotient suggests defining, on $[\delta_n, 1]$,

$$\begin{aligned}
\hat{v}'_n &:= \frac{1}{\hat{u}_n^2}\left((\hat{\sigma}_n^2)'\hat{u}_n - \hat{u}'_n\hat{\sigma}_n^2\right), \\
\hat{v}''_n &:= \frac{1}{\hat{u}_n^3}\Big(\hat{u}_n\left((\hat{\sigma}_n^2)''\hat{u}_n - \hat{u}''_n\hat{\sigma}_n^2\right) - 2\hat{u}'_n((\hat{\sigma}_n^2)'\hat{u}_n - \hat{u}'_n\hat{\sigma}_n^2)\Big),
\end{aligned}$$

where $(\hat{\sigma}_n^2)'(t) = \hat{\Gamma}'_n(t,t)$, $(\hat{\sigma}_n^2)''(t) = \hat{\Gamma}''_n(t,t)$

Now we complete the definition of our estimator of v on the whole interval by using a Taylor-kind expansion on $[0, \delta_n)$,

$$\hat{v}_n(t) = \hat{v}_n(\delta_n) + (t-\delta_n)\hat{v}'_n(\delta_n) + \frac{1}{2}(t-\delta_n)^2\hat{v}''_n(\delta_n), \quad \text{if } t \in [0,\delta_n). \tag{23}$$

Finally, take

$$\begin{aligned}
\hat{v}'_n(t) &:= \hat{v}'_n(\delta_n) + (t-\delta_n)\hat{v}''_n(\delta_n), \quad \text{if } t \in [0,\delta_n). \\
\hat{v}''_n(t) &:= \hat{v}''_n(\delta_n), \quad \text{if } t \in [0,\delta_n).
\end{aligned}$$

Proposition 2 *Let the assumptions of Proposition 1 (b) hold.*

a) *If $u(0) > 0$ then the rate of convergence of $\|\hat{v}_n - v\|$, $\|\hat{v}_n' - v'\|$ and $\|\hat{v}_n'' - v''\|$ are the same as those of (19), (20) and (21), respectively.*

b) *If $u(0) = 0$ assume that $\inf_t u'(t) > 0$ and $\inf_{t\in[\delta,1]} \sigma^2(t) > 0$ for every $\delta > 0$. Let $\{\delta_n\} \downarrow 0$ be such that $\sup(n^{-1/2}, h_n) = o(\delta_n)$. Then*

$$\begin{aligned}
\|\hat{v}_n - v\| &= O_P\left(\frac{\delta_n}{h_n^2\sqrt{n}}\right) + O(h_n) + O(\delta_n^3) \\
\|\hat{v}_n' - v'\| &= O_P\left(\frac{1}{h_n^2\sqrt{n}}\right) + O\left(\frac{h_n}{\delta_n}\right) + O(\delta_n^2) \\
\|\hat{v}_n'' - v''\| &= O_P\left(\frac{1}{\delta_n h_n^2\sqrt{n}}\right) + O\left(\frac{h_n}{\delta_n^2}\right) + O(\delta_n).
\end{aligned}$$

Estimation of the Radon-Nikodym derivatives

Here we plug-in the estimates of m, u, v and their derivatives obtained above in the Radon-Nikodym derivatives $f = d\mu_0/d\mu_1$ obtained above in Theorem 2. Denote by $\hat{f}_n$ the resulting estimate. Then, we compute the convergence rate to the Bayes risk of the error attained by the corresponding nonparametric plug-in classification procedure.

According to Theorem 2 the Radon-Nikodym densities of interest are the exponential of some integrals, ratios, products or square roots of functions estimated with orders of convergence appearing in Propositions 1 and 2. The final rate will be that of the worst estimate handled, which corresponds to the second order derivatives. As with the estimation of v, there is some difference in the orders depending on whether $\sigma^2(0)$ is strictly positive or not.

The main conclusions are summarized in the following result.

Theorem 3 *Let us assume that conditions in Proposition 1 (b) and Theorem 2 hold.*

a) *If $u_i(0) > 0$ for $i = 0, 1$, then for $h_n = O(n^{-1/6})$ we get*

$$\log \hat{f}_n(x) - \log \frac{d\mu_0}{d\mu_1}(x) = O_P\left(n^{-1/6}\right), \ x \in \mathcal{C}[0,1].$$

b) *If $u_i(0) = 0$ for $i = 0, 1$ and $\inf_t u'(t) > 0$ and $\inf_{t\in[\delta,1]} \sigma^2(t) > 0$ for every $\delta > 0$, then, for $h_n = O(n^{-9/50})$ we have*

$$\mathbb{E}\left(\log \hat{f}_n(X) - \log \frac{d\mu_0}{d\mu_1}(X) \middle| X_1, \ldots, X_n\right) = O_P\left(n^{-1/10}\right).$$

Let us note that, in any case, our nonparametric estimator $\hat{f}_n(x) = dP_{\hat{m}_0\hat{\Gamma}_0}/dP_{\hat{m}_1\hat{\Gamma}_1}$ is constructed, using (11), under the sole assumption that the covariance function has a triangular structure. So, the estimator is formally the same in both cases a) and b) of Theorem 2. If we knew that $m_i = 0$ for $i = 0, 1$ then we could employ $\hat{f}_n(x) = dP_{\hat{m}_0\hat{\Gamma}_0}/dP_{\hat{m}_0\hat{\Gamma}_1}$ and the rates of Theorem 3 would improve, under the assumptions of Theorem 3 b), to $O_P(n^{-3/28})$.

Using higher order derivatives

The proof of Theorem 3 was based on the use of Taylor expansions of order two. Next we show how the existence of higher order derivatives improves the estimation process.

Proposition 3 *Under the assumptions of Theorem 3 suppose further that the mean function $m : [0, 1] \to \mathbb{R}$ as well as the functions $t \to \Gamma(t, 1)$, $t \to \Gamma(0, t)$ and σ^2 admit Lipschitz third order derivatives. Then the rates in Theorem 3 a) and b) are improved to $O_P(n^{-1/4})$ and $O_P(n^{-5/32})$, respectively.*

A remark similar to that made after Theorem 3 applies here. If we incorporate the information $m_i = 0$ to the estimator, the convergence rate in Proposition 3 b) slightly improves to $O_P(n^{-1/6})$.

The convergence orders may be further improved by assuming additional smoothness orders and taking advantage of numerical differentiation techniques (see, for instance, p. 146 in Gautschi, 1997). We will not develop this idea in the present work. However, let us observe that in the estimation of functions with infinite derivatives it is possible to obtain orders as close to $O_P(n^{-1/2})$ as desired by choosing k large enough in the k-point rule (see, for instance, Herzeg and Cvetkovic, 1986).

Estimation of the probability of misclassification

We denote by $\hat{L}_n := L(\hat{g}_n) = \mathbb{P}\{\hat{g}_n(X) \neq Y | \mathcal{X}_n\}$ the classification error associated with the nonparametric plug-in rule $\hat{g}_n(x) = \mathbb{I}_{\{\hat{\eta}_n(x) > 1/2\}}$. Here $\hat{\eta}_n$ is obtained by substituting the Radon-Nikodym derivative $f = d\mu_0/d\mu_1$ in (5) with the estimator $\hat{f}_n$ obtained by replacing m, u, v and their derivatives with the corresponding nonparametric estimators obtained along this subsection. The following result is an example of how the convergence rates for the difference between the logarithms of the Radon-Nikodym derivatives $\hat{f}_n(x)$ and $f(x)$ can be translated into convergence rates of $\hat{L}_n$ to the Bayes error L^*.

Theorem 4 *Let the assumptions of Proposition 1 (b) and Theorem 2 hold. If $u_i(0) > 0$ for $i = 0, 1$, then taking $h_n = O(n^{-1/6})$ we get $\hat{L}_n - L^* = O_P\left(n^{-1/6}\right)$.*

In the case when $u_i(0) = 0$, for $i = 0, 1$, we can prove that $\hat{L}_n - L^*$ is $O_P(n^{-1/10})$ under the assumptions that $\inf_t u'(t) > 0$ and $\inf_{t\in[\delta,1]} \sigma^2(t) > 0$ for every $\delta > 0$. The idea is to follow the same steps as in the proof of Theorem 4, but bounding the integrals in (38) and (42) as we did along the proof of Theorem 3.

3 Consistency of the k-NN functional rules

As stated in the introduction, the k-NN classifier is not universally consistent in the functional setting. However, Cérou and Guyader (2006) provide sufficient conditions for the consistency $L_n \to L^*$ in probability (or, equivalently, $\mathbb{E}(L_n) \to L^*$), where L_n is the conditional classification error of the k-NN rule. In this section we show that these conditions are fulfilled by the Gaussian processes introduced in Section 2.2 and, in consequence, that the k-NN is consistent in probability for them.

Throughout this section the feature space where the variable X takes values is a separable metric space $(\mathcal{F}, D)$. As usual, we will denote by P_X the distribution of X defined by $P_X(B) = \mathbb{P}\{X \in B\}$ for $B \in \mathcal{B}_{\mathcal{F}}$, where $\mathcal{B}_{\mathcal{F}}$ are the Borel sets of $\mathcal{F}$.

The key assumption is a regularity condition on the regression function $\eta(x) = \mathbb{E}(Y|X = x)$ which is called *Besicovich condition* **(BC)**. The function η is said to fulfill **(BC)** if

$$\lim_{\delta\to 0} \frac{1}{P_X(B_{X,\delta})} \int_{B_{X,\delta}} \eta(z)\, dP_X(z) = \eta(X) \quad \text{in probability,}$$

where $B_{x,\delta} := \{z \in \mathcal{F} : D(x, z) \le \delta\}$ is the closed ball with center x and radius δ. Besicovich condition plays, for instance, an important role in the consistency of kernel rules (see Abraham *et al.* 2006).

Cérou and Guyader (2006, Th. 2) have proved that, if $(\mathcal{F}, D)$ is separable and condition **(BC)** is fulfilled, then the k-NN classifier defined by (2) and (3) is consistent in probability provided that $k_n \to \infty$ and $k_n/n \to 0$. In order to apply this result in our case, it will be sufficient to observe that the continuity (P_X-a.e.) of $\eta(x)$ implies also **(BC)**. Consequently we can establish the following result, whose proof is immediate from Theorems 1 and 2.

Proposition 4 *Under the assumptions of Theorem 1 suppose that $P_X(\partial S) = 0$. Then for P_X-a.e. x, z in the topological interior of S,*

$$|\eta(z) - \eta(x)| = \left| \frac{1-p}{p\frac{d\mu_0}{d\mu_1}(z) + 1 - p} - \frac{1-p}{p\frac{d\mu_0}{d\mu_1}(x) + 1 - p} \right| \le \frac{p}{1-p} \left| \frac{d\mu_0}{d\mu_1}(x) - \frac{d\mu_0}{d\mu_1}(z) \right|. \quad (24)$$

As a consequence, for both cases a) and b) considered in Theorem 2 the k-NN functional classifier is consistent in probability, provided that $k_n \to \infty$ and $k_n/n \to 0$.

Of course, the point is that the Radon-Nikodym derivatives given in Theorem 2 are continuous on $C[0,1]$. So (24) would imply also the continuity of $\eta(x)$ which in turn entails the Besicovich condition **(BC)** and the consistency.

4 Empirical results

In this section we compare the performance of the k-NN classification procedure with the plug-in one for infinite-dimensional data. First (Subsection 4.1) we describe the results of a simulation study carried out with processes from the two Gaussian families specified in Subsection 2.3. Afterwards (Subsection 4.2) we focus on a real-data set.

4.1 Monte Carlo study

The observations will be realizations of two Ornstein-Uhlenbeck processes and two Brownian motions as described in Subsection 2.3. The parameters chosen for the pairs of processes are specified in Table 1 (in Figure 1 we have depicted some trajectories of the processes used in the simulations).

Figure 1 here.

We assume that $p = \mathbb{P}\{Y = 0\}$, the proportion of observations coming from P_0, is $1/2$ and is known in advance. For each $i = 0, 1$ we take a training sample with size $n_i = 100$ and a test sample with size 50 from P_i. The processes are observed at equidistant times of the interval $[0,1]$, $t_0 = 0, t_1, \ldots, t_N = 1$, with $N = 50$. We denote by $\Delta = t_j - t_{j-1}$ the internodal distance. The number of Monte Carlo runs is 1000. In each run we use the training sample to construct four classifiers: k-NN with the supremum norm and with a PLS-based semimetric (see e.g. Ferraty and Vieu, 2006, p. 30), parametric and nonparametric plug-in as introduced in Subsections 2.3 and 2.4 respectively. The performance of these classifiers is assessed by the proportion of correctly classified observations in the test samples. We also compute this proportion for the Bayes rule associated to each model. The number k of neighbours and the number of PLS directions for projection are chosen via cross-validation from a maximum of 10 neighbours and 5 PLS directions respectively.

When applying the nonparametric plug-in method to the data functions evaluated on the whole interval $[0,1]$ we observed a noticeable boundary effect near 0, especially in the estimation of v and its derivatives. This made the nonparametric plug-in method perform poorly. In order to avoid this, the Radon-Nikodym derivative for the nonparametric plug-in rule has been evaluated on the trajectories restricted to the interval $[h_n, 1]$, where h_n is the same (and unique) smoothing parameter used in the estimation of the derivatives of u_i and v_i. The value of h_n has been chosen among $\{2\Delta, 4\Delta, \ldots, 20\Delta\}$ via cross-validation: for each $h_n = k\Delta$ we compute the corresponding estimated classifica-

tion error with the usual leave-one-out device (every training observation is classified, as if it were a new incoming observation, using the remaining data as a training sample).

In Table 1 we display the mean and the standard deviation (between parentheses) of the proportion of correct classifications over the 1000 Monte Carlo samples. We see that the parametric plug-in procedure is the one performing best: it is very near the optimum.

As it could be expected, the nonparametric plug-in behaves worse than the parametric one. Its best performance corresponds to the random start cases $u_i(0) > 0$ for $i = 0, 1$. In these situations, it is the second better classifier. When $u_i(0) = 0$, the parametric plug-in is still the winner, the k-NN with PLS is the second and the k-NN with the supremum metric and the nonparametric plug-in perform similarly.

It is interesting to note that the k-NN classification method is always reliable (even with the supremum metric, although PLS semimetric yields better results). Thus one of the conclusions of the study is that, when classifying functional data, the k-NN procedure is generally a safe choice, free of model assumptions.

Table 1 here.

4.2 A real data set

We compare the performance of the k-NN classification procedure with the nonparametric plug-in one in the analysis of data from research in experimental cardiology. The experiment was conducted at the Vall d'Hebron Hospital (Barcelona, Spain). See Ruiz-Meana *et al.* (2003) for biochemical and medical details on the data and Cuevas, Febrero and Fraiman (2004, 2006) for previous analysis of these observations.

The variable under study is the mitochondrial calcium overload (MCO), which measures the level of the mitochondrial calcium ion (Ca2+). This variable was observed every 10 seconds during an hour in isolated mouse cardiac cells. The aim of the study was to assess whether a drug called Cariporide increased the MCO level. The data we analyze here consist of two samples of functions with sizes $n_0 = 45$ (control group) and $n_1 = 44$ (treatment group with Cariporide). In Figure 2 we display (a) all the data and (b) the group means.

Figure 2 here.

In many cases the first three minutes each curve shows oscillations which correspond to normal contractions of the cells. This first part of the curves has been eliminated (as in the original experiments with these data) because it has high variability and depends on uncontrolled factors.

To obtain a better approach of the distributions to normality, we have considered a transformation of the data, $X = \log(\text{MCO} - 85)$. The performance of any of the classification procedures considered is described by the probability of correctly classifying one of the transformed observations, approximated via cross-validation.

Obviously, in this case, we do not have enough information to consider using the parametric plug-in classifier. Consequently we only employ the k-NN (with uniform metric and PLS-based semimetric) and the nonparametric plug-in discrimination rules. The results appear in Table 2. It is interesting to notice that the results in this case, in some sense, are the opposite to those obtained with the simulations. The nonparametric plug-in clearly outperforms the other two and the k-NN with the supremum metric does better than the k-NN with PLS.

Table 2 here.

Acknowledgement. The authors want to thank Javier Segura for bringing to our knowledge some numerical differentiation techniques (in particular, the k-point rule).

5 Appendix

A.1 Parameter estimation for the models of Subsection 2.3

Two Brownian motions

In the simulations of Section 4 the estimator of c is $\hat{c} = \arg\min_c \sum_{j=1}^{N} (\hat{m}_0(t_j) - c\, t_j)^2$, where m_i is the sample mean of the observations coming from P_i. The parameters θ_i and σ^2 are respectively estimated by $\hat{\theta}_i = \sum_{j=1}^{n_i} \left(X_j(0;i) - \hat{m}_i(0)\right)^2 / (n_i - 1)$ and $\hat{\sigma}^2 = \sum_{i=0,1} \sum_{j=1}^{n_i} \left(X_j(1;i) - \hat{m}_i(1) - X_i(0;i) + \hat{m}_i(0)\right)^2 / (n_0 + n_1 - 1)$.

Two Ornstein-Uhlenbeck processes

The estimation of the unknown parameters (β_i, η_i and σ_i, $i = 0, 1$) is carried out via linear least-squares regression between the realizations of the process at consecutive time points. The main idea is that, for $i = 0, 1$ and for any $0 \le s < t \le 1$, we have

$$X(t;i) \;=\; X(s;i)\, e^{-\beta_i (t-s)} + \eta_i \left(1 - e^{-\beta_i (t-s)}\right) + \sigma_i \sqrt{1 - e^{-2\beta_i (t-s)}}\, Z, \tag{25}$$

where Z is $N(0,1)$. The updating formula (25) is valid when $X(0;i)$ is either deterministic or random. In particular, for $i = 0, 1$, $k = 1, \ldots, n_i$ and $j = 0, \ldots, N-1$,

$$X_k(t_{j+1};i) \;=\; a_i X_k(t_j;i) + b_i + \sigma_i \sqrt{1 - e^{-2\beta_i \Delta}}\, Z_{kj}, \tag{26}$$

where $a_i := e^{-\beta_i \Delta}$, $b_i := \eta_i \left(1 - e^{-\beta_i \Delta}\right)$ and Z_{kj} are i.i.d. variables $N(0,1)$.

Observe that, by estimating the parameters of the simple linear regression equation (26), we can construct estimators of β_i, η_i and σ_i. When $X(0;i)$ is deterministic, we compute the least-squares estimators of a_i and b_i, that is, the values $\hat{a}_i$ and $\hat{b}_i$ minimizing $\sum_{k=1}^{n_i} \sum_{j=0}^{N-1} u_{kj}^2$, where $u_{kj} := X_k(t_{j+1};i) - (\hat{a}_i X_k(t_j;i) + \hat{b}_i)$ are the residuals. Then

$$\hat{\beta}_i = -\frac{\log(\hat{a}_i)}{\Delta}, \qquad \hat{\eta}_i = \frac{\hat{b}_i}{1 - \hat{a}_i}, \qquad \hat{\sigma}_i^2 = \frac{1}{(1 - \hat{a}_i^2)(n_i N - 2)} \sum_{k=1}^{n_i} \sum_{j=0}^{N-1} u_{kj}^2. \tag{27}$$

When $X(0;i)$ is random, we can compute $\hat{\beta}_i$ and $\hat{\sigma}_i^2$ as in (27), but η_i is better estimated by $\hat{\eta}_i = \sum_{j=1}^{n_i}\sum_{k=0}^{N} X_{ij}(t_k)/(n_i\,(N+1))$.

A.2 Proofs of the results in 2.4

Proof of Proposition 1

(a) By the functional CLT in $(C[0,1], \|\cdot\|)$ (see p. 172 in Araujo and Giné, 1980) the sequence $\sqrt{n}(\hat{m}_n - m)$ converges weakly. This entails that the sequence $\|\sqrt{n}(\hat{m}_n - m)\|$ is bounded in probability which in turn implies (16). Concerning (17) and (18), let us denote $X_i^*(t) = X_i(t) - m(t),\ t \in [0,1], i = 1,2,\ldots$. Note that, for $t \in [h_n, 1-h_n]$,

$$\begin{aligned} |m'(t) - \hat{m}_n'(t)| &\leq \left| m'(t) - \frac{m(t+h_n) - m(t-h_n)}{2h_n} \right| \\ &\quad + \left| \frac{1}{2h_n n}\sum_{i=1}^{n} X_i^*(t+h_n) \right| + \left| \frac{1}{2h_n n}\sum_{i=1}^{n} X_i^*(t-h_n) \right| \\ &\leq \left| m'(t) - \frac{m(t+h_n) - m(t-h_n)}{2h_n} \right| + \left\| \frac{1}{h_n n}\sum_{i=1}^{n} X_i^* \right\|. \end{aligned} \tag{28}$$

The CLT applied to the sequence $\{X_n^*\}$ allows us to conclude that the second term in the right-hand side of (28) is $O_P\left((n^{1/2}h_n)^{-1}\right)$. A second order Taylor expansion of the first term implies that there exist $\psi_n^{(1)} \in (t - h_n t)$ and $\psi_n^{(2)} \in (t, t+h_n)$ such that

$$\left| m'(t) - \frac{m(t+h_n) - m(t-h_n)}{2h_n} \right| = \frac{h_n}{4}\left| m''(\psi_n^{(1)}) - m''(\psi_n^{(2)}) \right| \leq \frac{Lh_n^2}{4} = O(h_n^2),$$

where L is the Lipschitz constant associated with m''.

Applying a similar reasoning to (18), we obtain that, if $t \in [h_n, 1-h_n]$, then,

$$|m''(t) - \hat{m}_n''(t)| \leq \left| m''(t) - \frac{m(t+h_n) + m(t-h_n) - 2m(t)}{h_n^2} \right| + 4\left\| \frac{1}{h_n^2 n}\sum_{i=1}^{n} Y_i \right\|. \tag{29}$$

The CLT implies that the order of the second term in (29) is $O_P\left((n^{1/2}h_n^2)^{-1}\right)$. A second order Taylor's expansion on t again gives that

$$\left| m''(t) - \frac{m(t+h_n) + m(t-h_n) - 2m(t)}{h_n^2} \right| = \left| m''(t) - \frac{1}{2}\left(m''(\psi_n^{(1)}) + m''(\psi_n^{(2)}) \right) \right| \leq Lh_n.$$

(b) Since

$$\begin{aligned} \hat{\Gamma}(t,1) - \Gamma(t,1) &= \frac{1}{n}\sum_i \left((X_i^*(t) + m(t) - \hat{m}_n(t))(X_i^*(1) + m(1) - \hat{m}_n(1)) \right) - \Gamma(t,1) \\ &= \frac{1}{n}\sum_i \left(X_i^*(t)X_i^*(1) - \Gamma(t,1) \right) + (m(t) - \hat{m}_n(t))\frac{1}{n}\sum_i X_i^*(1) \\ &\quad + (m(1) - \hat{m}_n(1))\frac{1}{n}\sum_i X_i^*(t) + (m(t) - \hat{m}_n(t))(m(1) - \hat{m}_n(1)), \end{aligned}$$

then

$$\begin{aligned}\|\hat{\Gamma}(\cdot,1)-\Gamma(\cdot,1)\| &\leq \left\|\frac{1}{n}\sum_i (X_i^*X_i^*(1)-\Gamma(\cdot,1))\right\| + \|m-\hat{m}_n\|\left|\frac{1}{n}\sum_i X_i^*(1)\right| \\ &\quad +|m(1)-\hat{m}_n(1)|\left\|\frac{1}{n}\sum_i X_i^*\right\| + \|m-\hat{m}_n\|\,|m(1)-\hat{m}_n(1)| \\ &=: T_n^{(1)}+T_n^{(2)}+T_n^{(3)}+T_n^{(4)}.\end{aligned}$$

The assumption $\mathbb{E}\|X_1\|^4 < \infty$ implies $\mathbb{E}\|X_i^*X_i^*(1)\|^2 < \infty$ and thus the sequence $\{X_i^*X_i^*(1)\}$ satisfies the CLT in the supremum norm. Then, since $E[X_i^*X_i^*(1)] = \Gamma(\cdot,1)$, we have that $T_n^{(1)} = O_P(n^{-1/2})$. Also $T_n^{(2)} = O_P(n^{-1})$ because the CLT (real case) implies that $\sum_i X_i^*(1)/n = O_P(n^{-1/2})$ and, according to Proposition 1 (a), $\|m-\hat{m}_n\| = O_P(n^{-1/2})$.

The CLT applied to $\{X_i^*\}$ and Proposition 1 (a) yield that $T_n^{(3)}$ and $T_n^{(4)}$ are $O_P(n^{-1})$. This allows us to conclude (19). The derivatives of $\Gamma(\cdot,1)$ are handled as those of m. The estimators of $\Gamma(0,\cdot)$ and $\sigma(\cdot)$ behave analogously to $\Gamma(\cdot,1)$. □

Proof of Proposition 2

a) According to expression (22) for $\hat{v}_n(t)$, this estimator is a quotient of two convergent sequences. As that in the denominator, $\hat{u}_n(0)$, converges to $u(0) > 0$, an upper bound for the overall rate of the quotient is the slowest rate between $\hat{\Gamma}_n(0,t)$ and $\hat{u}_n(0)$. Similar arguments apply for the first and second derivatives.

b) Let $t \in [\delta_n, 1]$. The hypothesis on u' implies that $\inf_{t\geq\delta_n} u(t) \geq O(\delta_n)$. Since $n^{-1/2} = o(\delta_n)$, from (19) we obtain that $\inf_{t\geq\delta_n} \hat{u}_n(t) \geq O_P(\delta_n)$. Therefore, a direct calculation based on the expression of $\hat{v}_n$ together with Proposition 1 b) leads to

$$\sup_{t\in[\delta_n,1]} |\hat{v}_n(t)-v(t)| = O_P\left(\frac{1}{\delta_n\sqrt{n}}\right). \tag{30}$$

The same reasoning, taking into account the relative orders between δ_n and h_n leads to

$$\sup_{t\in[\delta_n,1]} |\hat{v}_n'(t)-v'(t)| = O_P\left(\frac{1}{\delta_n h_n\sqrt{n}}\right) + O\left(\frac{h_n^2}{\delta_n}\right) \tag{31}$$

$$\sup_{t\in[\delta_n,1]} |\hat{v}_n''(t)-v''(t)| = O_P\left(\frac{1}{\delta_n h_n^2\sqrt{n}}\right) + O\left(\frac{h_n}{\delta_n^2}\right). \tag{32}$$

Now, let $t \in [0, \delta_n]$. Using the second-order Taylor expansion of v at δ_n, together

with the definition (23) of $\hat{v}_n$, we obtain that there exists $\psi_n \in (t, \delta_n)$ such that

$$\begin{aligned}
|\hat{v}_n(t) - v(t)| &\leq |\hat{v}_n(\delta_n) - v(\delta_n)| + (\delta_n - t)|\hat{v}_n'(\delta_n) - v'(\delta_n)| \\
&\quad + \frac{1}{2}(t-\delta_n)^2|\hat{v}_n''(\delta_n) - v''(\delta_n)| + \frac{1}{2}(t-\delta_n)^2|v''(\delta_n) - v''(\psi_n)| \\
&\leq O_P\left(\frac{1}{\delta_n\sqrt{n}}\right) + O_P\left(\frac{1}{h_n\sqrt{n}}\right) + O\left(h_n^2\right) + O_P\left(\frac{\delta_n}{h_n^2\sqrt{n}}\right) + O\left(h_n\right) + O(\delta_n^3) \\
&= O_P\left(\frac{\delta_n}{h_n^2\sqrt{n}}\right) + O\left(h_n\right) + O(\delta_n^3),
\end{aligned}$$

where we have applied (30), (31) and (32) and the fact that v'' is Lipschitz. Then the first statement in Proposition 2 b) is deduced from here and (30). The remaining two statements are proved similarly. □

Next we state a technical lemma which will be employed to prove Theorem 3.

Lemma 1 *Let $\{Y(t), t \in [0,1]\}$ be a stochastic process whose mean function $m(t)$ and variance function $\sigma^2(t)$ satisfy that $m(0) = \sigma(0) = 0$ and both have a bounded derivative. Let $\{\delta_n\}$ be positive numbers which converge to zero. Then*

$$\mathbb{E}\int_0^{\delta_n} |Y(t)|dt = O(\delta_n^{3/2}) \quad \textit{and} \quad \mathbb{E}\int_0^{\delta_n} Y^2(t)dt = O(\delta_n^2).$$

Proof: Let H be a common upper bound for the derivatives of m^2 and σ^2.

$$\begin{aligned}
\int_0^{\delta_n} \mathbb{E}|Y(t)|dt &\leq \int_0^{\delta_n} \mathbb{E}^{1/2}(Y^2(t))dt = \int_0^{\delta_n} (m(t)^2 + \sigma^2(t))^{1/2}dt \\
&\leq (2H)^{1/2}\int_0^{\delta_n} t^{1/2}dt = O(\delta_n^{3/2}).
\end{aligned}$$

The second statement in the lemma follows analogously. □

Proof of Theorem 3: From expressions (8) and (9) we see that $f = d\mu_0/d\mu_1$ is a function of m_i, u_i, v_i and their derivatives. Statement a) corresponds to the simplest case in which $u_i(0) > 0$. In this situation, the simple structure of the estimators shows that an upper bound for the convergence rate for $\log f_n(x)$ is the worst rate for the estimators involved in its definition, namely that of the estimators v_0'' and v_1''.

Hence, we concentrate on part b). For simplicity we will omit the sub-index in v_i for the rest of the proof. First notice that in the expressions for $d\mu_0/d\mu_1$ which we obtained in Theorem 2 the second derivatives of v only appear inside integrals. In other words, we only need to care about differences of the type

$$\int_0^1 X^r(t)(\hat{k}_n(t)\hat{v}_n''(t) - k(t)v''(t))\,dt = O_P\left(\int_0^1 X^r(t)k(t)(\hat{v}_n''(t) - v''(t))dt\right), \qquad (33)$$

for $r = 1, 2$. Here k is a function depending on u, v, u', v', m and m' and X is a mixture of the Brownian motions under consideration. Let us analyze the case in Theorem 2 b) for which $r = 1$ and the function k can be expressed as $k = k_1/\left(v((vu' - uv')^2\right)$, where k_1 is a function which can be written in terms of u, v, u', v', m and m'. Therefore, the assumptions in Theorem 2, imply that k is bounded. Let K be an upper bound of k.

We split in two the integral in the right-hand side of (33), over the intervals $[0, \delta_n]$ and $[\delta_n, 1]$. Now, from (32) in the proof of Proposition 2, we have that

$$\mathbb{E}\left(\left| \int_{\delta_n}^{1} X(t)k(t)(\hat{v}_n''(t) - v''(t))dt \right| \Big| X_1, \ldots, X_n \right) \leq \left(O_P\left(\frac{1}{\delta_n h_n^2 \sqrt{n}} \right) + O\left(\frac{h_n}{\delta_n^2} \right) \right) \left(\int_{\delta_n}^{1} \mathbb{E}(X^2(t))dt \right)^{1/2}. \quad (34)$$

With respect to the other integral, we have that

$$\mathbb{E}\left(\left| \int_{0}^{\delta_n} X(t)k(t)(\hat{v}_n''(t) - v''(t))dt \right| \Big| X_1, \ldots, X_n \right) \leq K \left\| \hat{v}_n'' - v'' \right\| \mathbb{E} \int_0^{\delta_n} |X(t)|dt = O_P\left(\frac{\delta_n^{1/2}}{\sqrt{n}} \right) + O\left(\frac{h_n}{\delta_n^{1/2}} \right) + O(\delta_n^{5/2}), \quad (35)$$

where the last equality comes from Lemma 1 and Proposition 2 b). Equations (34) and (35) give

$$\mathbb{E}\left(\left| \int_{0}^{1} X(t)k(t)(\hat{v}_n''(t) - v''(t))dt \right| \Big| X_1, \ldots, X_n \right) \leq O_P\left(\frac{1}{\delta_n h_n^2 \sqrt{n}} \right) + O\left(\frac{h_n}{\delta_n^2} \right) + O(\delta_n^{5/2}).$$

Taking $h_n = \delta_n^{9/2}$ and $\delta_n = n^{-1/25}$ equates the three terms and yields the result. □

Proof of Proposition 3: It follows the same steps as the proof of Proposition 1, the only difference being that if we apply a third order Taylor expansion in (29), we obtain

$$\left| m''(t) - \frac{m(t + h_n) + m(t - h_n) - 2m(t)}{h_n^2} \right| = \frac{h_n}{3!} \left| \left(m'''(\psi_n^1) - m'''(\psi_n^2) \right) \right| \leq \frac{L h_n^2}{3!},$$

and the result follows. □

Proof of Theorem 4: Let us use the following inequality (see, e.g., Devroye *et al.*, 1996, p. 93)

$$\hat{L}_n - L^* \leq 2\, \mathbb{E}\left(|\eta(X) - \eta_n(X)| \mid \mathcal{X}_n \right),$$

where η is given in (5) and η_n is obtained substituting $f = d\mu_0/d\mu_1$ by $\hat{f}_n$ in (5). Without loss of generality in this proof we consider $p = \mathbb{P}\{Y = 0\} = 1/2$.

Observe that, f and $\hat{f}_n$ are always positive since they are Radon-Nikodym derivatives of one probability measure with respect to another. Thus, for any x, we have

$$|\eta(x) - \eta_n(x)| = \frac{|f(x) - \hat{f}_n(x)|}{(1 + \hat{f}_n(x)(1 + f(x))} \leq |f(x) - \hat{f}_n(x)|,$$

which implies that

$$\hat{L}_n - L^* \leq 2\, \mathbb{E}\left(|f(x) - \hat{f}_n(x)| \Big| \mathcal{X}_n \right). \tag{36}$$

We obtain convergence rates (in probability) for the conditional expectation in the right of (36). Since all the cases are similar, let us consider the simple situation in which $m_0 \neq m_1$ and $\Gamma_0 = \Gamma_1 = \Gamma$ with $\Gamma(s,t) = u(\min(s,t))\, v(\max(s,t))$. Then

$$f - \hat{f}_n = \frac{dP_{m_0,\Gamma}}{dP_{m_1,\Gamma}} - \frac{dP_{\hat{m}_0,\hat{\Gamma}_0}}{dP_{\hat{m}_1,\hat{\Gamma}_1}} = \frac{dP_{m_0,\Gamma}}{dP_{m_1,\Gamma}} - \frac{dP_{\hat{m}_0,\hat{\Gamma}_0}}{dP_{\hat{m}_1,\hat{\Gamma}_0}} + \frac{dP_{\hat{m}_0,\hat{\Gamma}_0}}{dP_{\hat{m}_1,\hat{\Gamma}_0}} \left(1 - \frac{dP_{\hat{m}_1,\hat{\Gamma}_0}}{dP_{\hat{m}_1,\hat{\Gamma}_1}} \right). \tag{37}$$

By Theorem 2 (b) and the mean value theorem we have that, for any x,

$$\frac{dP_{m_0,\Gamma}}{dP_{m_1,\Gamma}}(x) - \frac{dP_{\hat{m}_0,\hat{\Gamma}_0}}{dP_{\hat{m}_1,\hat{\Gamma}_0}}(x) = e^z (z_1 - z_2),$$

where (using the notation of Theorem 2)

$$\begin{aligned}
z_1 &= D_1 + \left(D_2 - 2\,\frac{G(0)}{v(0)} \right) x(0) + 2\,\frac{G(1)}{v(1)}\, x(1) - 2 \int_0^1 \frac{x(t)}{v(t)}\, G'(t)\, dt, \\
z_2 &= \hat{D}_{1;0} + \left(\hat{D}_{2;0} - 2\,\frac{\hat{G}(0)}{\hat{v}_0(0)} \right) x(0) + 2\,\frac{\hat{G}_0(1)}{\hat{v}_0(1)}\, x(1) - 2 \int_0^1 \frac{x(t)}{\hat{v}_0(t)}\, \hat{G}_0'(t)\, dt
\end{aligned}$$

and $z = \lambda\, z_1 + (1-\lambda) z_2$ for some $\lambda \in [0,1]$. The subscripts 0 in the expression of z_2 mean that the estimation is carried out only with the sample from P_0.

Consequently,

$$\begin{aligned}
&\mathbb{E}\left(\left| \frac{dP_{m_0,\Gamma}}{dP_{m_1,\Gamma}}(X) - \frac{dP_{\hat{m}_0,\hat{\Gamma}_0}}{dP_{\hat{m}_1,\hat{\Gamma}_0}}(X) \right| \Big| \mathcal{X}_n \right) \\
&\quad \leq \mathbb{E}\left\{ e^{|Z_1|+|Z_2|} \left[|D_1 - \hat{D}_{1;0}| + \left(|D_2 - \hat{D}_{2;0}| + 2 \left| \frac{G(0)}{v(0)} - \frac{\hat{G}(0)}{\hat{v}_0(0)} \right| \right) |X(0)| \right.\right. \\
&\qquad \left.\left. + 2 \left| \frac{G(1)}{v(1)} - \frac{\hat{G}_0(1)}{\hat{v}_0(1)} \right| |X(1)| + 2 \int_0^1 |X(t)| \left| \frac{G'(t)}{v(t)} - \frac{\hat{G}_0'(t)}{\hat{v}_0(t)} \right| dt \right] \Big| \mathcal{X}_n \right\} \qquad (38) \\
&\quad \leq \kappa \left\{ |D_1 - \hat{D}_{1;0}|\, \mathbb{E}\left(e^{A\|X\|} \,|\mathcal{X}_n \right) + \left(|D_2 - \hat{D}_{2;0}| + 2 \max_{t=0,1} \left| \frac{G(t)}{v(t)} - \frac{\hat{G}_0(t)}{\hat{v}_0(t)} \right| \right. \right. \qquad (39) \\
&\qquad \left.\left. + 2 \int_0^1 \left| \frac{G'(t)}{v(t)} - \frac{\hat{G}_0'(t)}{\hat{v}_0(t)} \right| dt \right) \mathbb{E}\left(\|X\| e^{A\|X\|} | \mathcal{X}_n \right) \right\} \qquad (40)
\end{aligned}$$

where $\kappa = \exp(|D_1| + |\hat{D}_{1;0}|)$ and

$$A = \max\left(|D_2| + |\hat{D}_{2;0}|, \left\|\frac{G}{v} + \frac{\hat{G}_0}{\hat{v}_0}\right\|, \left\|\frac{G'}{v} + \frac{\hat{G}_0'}{\hat{v}_0}\right\|\right).$$

Using Propositions 1 and 2 we obtain that the conditional expectations appearing in (39) and (40) are bounded in probability. Then

$$\begin{aligned}\mathbb{E}\left(|\frac{dP_{m_0,\Gamma}}{dP_{m_1,\Gamma}}(X) - \frac{dP_{\hat{m}_0,\hat{\Gamma}_0}}{dP_{\hat{m}_1,\hat{\Gamma}_0}}(X)|\middle|\mathcal{X}_n\right) &= O_P\left(\max_{j=1,2}|D_j - \hat{D}_{j;0}|\right)\\ &+ O_P\left(\max_{t=0,1}\left|\frac{G(t)}{v(t)} - \frac{\hat{G}_0(t)}{\hat{v}_0(t)}\right|\right) + O_P\left(\int_0^1\left|\frac{G'(t)}{v'(t)} - \frac{\hat{G}_0'(t)}{\hat{v}_0'(t)}\right|dt\right).\end{aligned}$$

To find the convergence rates to 0 of these last three terms we use the expressions of D_1, D_2 and G appearing in Theorem 2. Some straighforward computations yield $|D_1 - \hat{D}_{1;0}| = O_P(\|\hat{v}_0' - v'\|)$, $|D_2 - \hat{D}_{2;0}| = O_P(\|\hat{v}_0 - v\|)$,

$$\max_{t=0,1}\left|\frac{G(t)}{v(t)} - \frac{\hat{G}_0(t)}{\hat{v}_0(t)}\right| = O_P(\|\hat{v}_0' - v'\|) \quad \text{and} \quad \int_0^1\left|\frac{G'(t)}{v'(t)} - \frac{\hat{G}_0'(t)}{\hat{v}_0'(t)}\right|dt = O_P(\|\hat{v}_0'' - v''\|).$$

Thus we get

$$\mathbb{E}\left(|\frac{dP_{m_0,\Gamma}}{dP_{m_1,\Gamma}}(X) - \frac{dP_{\hat{m}_0,\hat{\Gamma}_0}}{dP_{\hat{m}_1,\hat{\Gamma}_0}}(X)|\middle|\mathcal{X}_n\right) = O_P(\|\hat{v}_0'' - v''\|). \tag{41}$$

Let us now focus on the last term of (37). The analysis is similar to the one carried out above. On the one hand, for any x it holds that

$$\frac{dP_{\hat{m}_0,\hat{\Gamma}_0}}{dP_{\hat{m}_1,\hat{\Gamma}_0}}(x) \leq \kappa\, e^{2B\|x\|},$$

where $B = \max(|\hat{D}_{2;0}|, \|\hat{G}_0/\hat{v}_0\|, \|\hat{G}_0'/\hat{v}_0\|)$. On the other hand, for any x it also holds that

$$\begin{aligned}&\left|1 - \frac{dP_{\hat{m}_1,\hat{\Gamma}_0}}{dP_{\hat{m}_1,\hat{\Gamma}_1}}(x)\right|\\ &\quad\leq |C_1 - \hat{C}_1| + \frac{1}{2}\,\hat{C}_1\, e^{\Lambda\|x\|^2}\left(|\hat{C}_3|x^2(0) + |\hat{C}_2|x^2(1) + \int_0^1 x^2(t)\frac{|\hat{F}'(t)|}{\hat{v}_0(t)\hat{v}_1(t)}\,dt\right) \quad (42)\\ &\quad\leq |C_1 - \hat{C}_1| + \hat{C}_1\,\Lambda\, e^{\Lambda\|x\|^2}\|x\|^2,\end{aligned}$$

where $\Lambda = (|\hat{C}_3| + |\hat{C}_2| + \int_0^1 |\hat{F}'|/(\hat{v}_0\hat{v}_1))/2$. Consequently

$$\begin{aligned}&\mathbb{E}\left(\frac{dP_{\hat{m}_0,\hat{\Gamma}_0}}{dP_{\hat{m}_1,\hat{\Gamma}_0}}(X)\;|1 - \frac{dP_{\hat{m}_1,\hat{\Gamma}_0}}{dP_{\hat{m}_1,\hat{\Gamma}_1}}(X)|\middle|\mathcal{X}_n\right)\\ &\quad\leq \kappa\left\{|C_1 - \hat{C}_1|\;\mathbb{E}\left(e^{2B\|X\|}|\mathcal{X}_n\right) + \hat{C}_1\,\Lambda\,\mathbb{E}\left(\|X\|^2\, e^{2B\|X\| + \Lambda\|X\|^2}\middle|\mathcal{X}_n\right)\right\}. \quad (43)\end{aligned}$$

The conditional expectations in (43) and $\hat{C}_1$ are $O_P(1)$. The term Λ is $O_P(\max_{j=0,1} \|\hat{v}_j'' - v''\|)$. The difference $|C_1 - \hat{C}_1|$ is $O_P(\max_{j=0,1} \|\hat{v}_j - v\|)$. Thus the term in (43) is $O_P(\max_{j=0,1} \|\hat{v}_j'' - v''\|)$. This, together with (41) and Proposition 2 (a), yield the desired result. □

References

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

Araujo A. and Giné, E. (1980). *The Central Limit Theorem for Real and Banach Valued Random Variables.* Wiley.

Audibert, J.Y. and Tsybakov, A.B. (2007). Fast learning rates for plug-in classifiers. *Ann. Statist.* **35**, 608–633.

Baíllo, A., Cuevas, A. and Fraiman, R. (2009). Classification methods for functional data. To appear in *Oxford Handbook on Statistics and FDA*, F. Ferraty and Y. Romain, eds. Oxford University Press.

Cérou, F. and Guyader, A. (2006). Nearest neighbor classification in infinite dimension. *ESAIM Probab. Stat.* **10**, 340-355.

Cuevas, A., Febrero, M. and Fraiman, R. (2004). An anova test for functional data. *Comput. Statist. Data Anal.* **47**, 111–122.

Cuevas, A., Febrero, M. and Fraiman, R. (2006). On the use of the bootstrap for estimating functions with functional data. *Comput. Statist. Data Anal.* **51**, 1063–1074.

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

Duda, R.O., Hart, P.E., Stork, D.G. (2000). *Pattern Classification, 2nd edition*. Wiley.

Ferraty, F. and Vieu, P. (2006). *Nonparametric Functional Data Analysis: Theory and Practice*. Springer.

Feldman, J. (1958). Equivalence and perpendicularity of Gaussian processes. *Pacific J. Math.* **8**, 699–708.

Fisher, R.A. (1936). The use of multiple measurements in taxonomic problems. *Annals of Eugenics* **7**, 179–188.

Folland, G. B. (1999). *Real Analysis Modern Techniques and their Applications*. Wiley, New York.

Gautschi, W. (1997). *Numerical Analysis. An Introduction*. Birkhäuser. Boston.

Hand, D. (2006). Classifier technology and the illusion of progress. *Statist. Sci.* **21**, 1–34.

Hastie, T., Tibshirani, R. and Friedman, J. (2001). *The Elements of Statistical Learning.* Springer. New York.

Herzeg, D. and Cvetkovic, L. (1986). On a numerical differentiation. *SIAM J. Numer. Anal.* **23**, 686–691.

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

Jørsboe, O. G. (1968). *Equivalence or Singularity of Gaussian Measures on Function Spaces.* Various Publications Series, No. 4, Matematisk Institut, Aarhus Universitet, Aarhus.

Ramsay, J.O. and Silverman, B.W. (2005). *Functional Data Analysis.* Second edition. Springer.

Ruiz-Meana, M., García-Dorado, D., Pina, P., Inserte, J., Agulló, L. and Soler-Soler, J. (2003). Cariporide preserves mitochondrial proton gradient and delays ATP depletion in cardiomyocites during ischemic conditions. *Am. J. Physiol. Heart Circ. Physiol.* **285**, H999-H1006.

Sacks, J. and Ylvisaker, N.D. (1966). Designs for regression problems with correlated errors. *Ann. Math. Statist.* **37**, 66–89.

Segall, A. and Kailath, T. (1975). Radon-Nikodym derivatives with respect to measures induced by discontinuous independent-increment processes. *Ann. Probab.* **3**, 449–464.

Shin, J. (2008). An extension of Fisher's discriminant analysis for stochastic processes. *J. Multiv. Anal.* **99**, 1191–1216.

Stone, C. J. (1977). Consistent nonparametric regression. *Ann. Statist.* **5**, 595–645.

Vakhania, N.N. (1975). The topological support of Gaussian measure in Banach space. *Nagoya Math. J.* **57**, 59–63.

Varberg, D.E. (1961). On equivalence of Gaussian measures. *Pacific J. Math.* **11**, 751–762.

Varberg, D.E. (1964). On Gaussian measures equivalent to Wiener measure. *Trans. Amer. Math. Soc.* **113**, 262–273.

			k-NN $\|\ \|_\infty$	k-NN PLS	Nonpar. plug-in	Param. plug-in	Bayes rule
Two Brownian motions	Deterministic at $t=0$ $(\theta_0=\theta_1=0)$	$c=1.5$, $\sigma=1$	0.68	0.73	0.71	0.77	0.77
			(0.07)	(0.07)	(0.16)	(0.06)	(0.06)
		$c=3$, $\sigma=1$	0.90	0.91	0.86	0.93	0.93
			(0.05)	(0.05)	(0.16)	(0.04)	(0.03)
		$c=2$, $\sigma=2$	0.60	0.64	0.64	0.69	0.69
			(0.08)	(0.08)	(0.16)	(0.07)	(0.06)
	Random at $t=0$ $(\theta_0,\theta_1\neq 0)$	$c=1.5$, $\sigma=1$ $\theta_0=\theta_1=1$	0.67	0.66	0.71	0.77	0.77
			(0.07)	(0.08)	(0.08)	(0.07)	(0.06)
		$c=1.5$, $\sigma=1$ $\theta_0=\theta_1=0.5$	0.67	0.70	0.72	0.77	0.77
			(0.07)	(0.08)	(0.08)	(0.06)	(0.06)
Two Ornstein-Uhlenbeck processes	Deterministic at $t=0$	$\beta_0=1$, $\eta_0=0$, $\sigma_0=1$ $\beta_1=1$, $\eta_1=1$	0.54	0.58	0.60	0.63	0.62
			(0.08)	(0.08)	(0.14)	(0.07)	(0.07)
		$\beta_0=0.4$, $\eta_0=0$, $\sigma_0=0.4$ $\beta_1=1$, $\eta_1=1$	0.83	0.86	0.82	0.88	0.88
			(0.09)	(0.06)	(0.16)	(0.05)	(0.05)
	Random at $t=0$	$\beta_0=0.5$, $\eta_0=0$, $\sigma_0=1$ $\beta_1=1$, $\eta_1=0.5$	0.59	0.60	0.63	0.63	0.64
			(0.13)	(0.11)	(0.14)	(0.07)	(0.14)
		$\beta_0=0.5$, $\eta_0=0$, $\sigma_0=2$ $\beta_1=1$, $\eta_1=2$	0.69	0.72	0.74	0.74	0.74
			(0.11)	(0.10)	(0.11)	(0.07)	(0.09)

Table 1: Results of the Monte Carlo study

k-NN $\|\ \|_\infty$	k-NN PLS	Nonpar. plug-in
0.79	0.66	0.85

Table 2: Proportion of correctly classified for the transformed cell data.

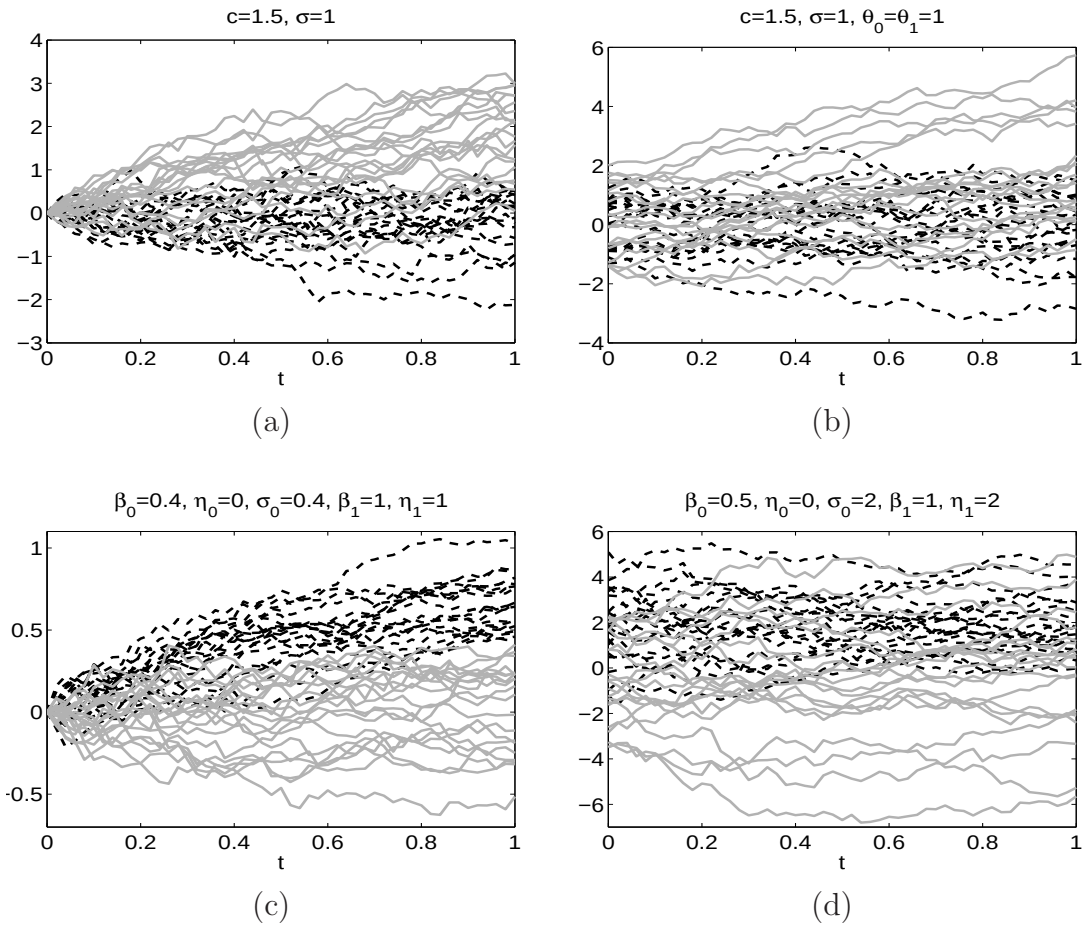
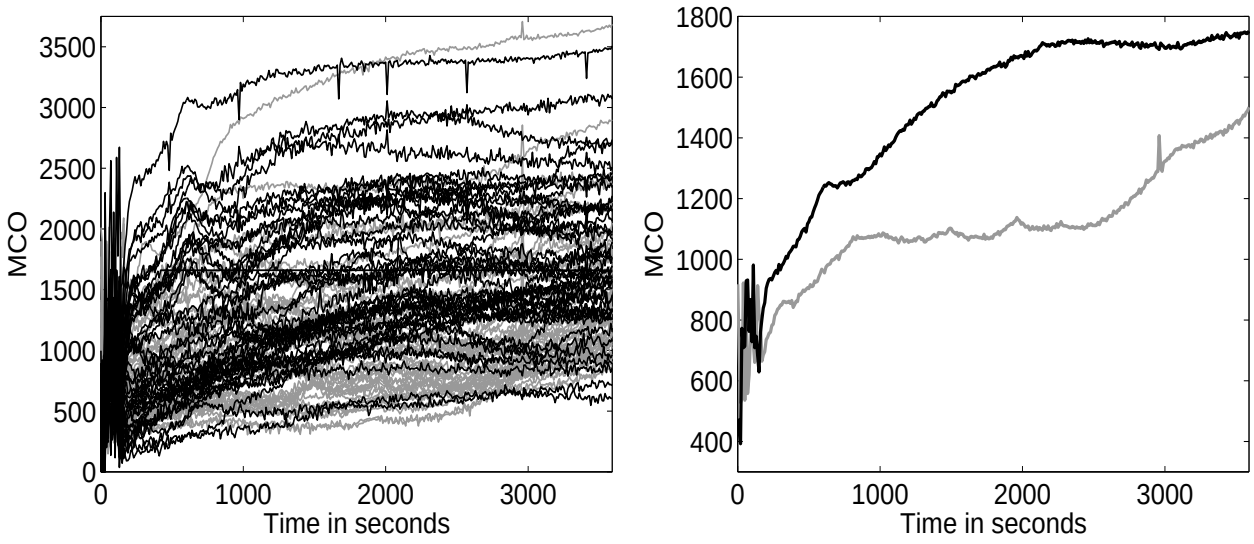


Figure 1: Some trajectories (P_0 in gray and P_1 in dotted black) of the processes used in the Monte Carlo study. In (a) and (b) we have two Brownian motions and in (c) and (d) the processes are Ornstein-Uhlenbeck. In (a) and (c) $X(0)|Y = i$ is 0 and in (b) and (d) it is random.



(a) (b)

Figure 2: Cell data (control group in grey and treatment group in black): (a) all the original observations; (b) sample means.