On the use of reproducing kernel Hilbert spaces in functional classification

José R. Berrendero[(1)], Antonio Cuevas[(1)], José L. Torrecilla[(1,2)]

[(1)] Departamento de Matemáticas, Universidad Autónoma de Madrid, Spain
[(2)] Institute BS-UC3M of Financial Big Data, Universidad Carlos III de Madrid, Spain

Abstract

The Hájek-Feldman dichotomy establishes that two Gaussian measures are either mutually absolutely continuous with respect to each other (and hence there is a Radon-Nikodym density for each measure with respect to the other one) or mutually singular. Unlike the case of finite dimensional Gaussian measures, there are non-trivial examples of both situations when dealing with Gaussian stochastic processes. This paper provides:

(a) Explicit expressions for the optimal (Bayes) rule and the minimal classification error probability in several relevant problems of supervised binary classification of mutually absolutely continuous Gaussian processes. The approach relies on some classical results in the theory of Reproducing Kernel Hilbert Spaces (RKHS).

(b) An interpretation, in terms of mutual singularity, for the "near perfect classification" phenomenon described by Delaigle and Hall (2012a). We show that the asymptotically optimal rule proposed by these authors can be identified with the sequence of optimal rules for an approximating sequence of classification problems in the absolutely continuous case.

(c) As an application, we discuss a natural variable selection method, which essentially consists of taking the original functional data $X(t)$, $t \in [0, 1]$ to a d-dimensional marginal $(X(t_1), \ldots, X(t_d))$ which is chosen in order to minimize the classification error of the corresponding Fisher's linear rule. We give precise conditions under which this discrimination method achieves the minimal classification error of the original functional problem.

Keywords: absolutely continuity, Radon–Nikodym derivatives, mutually singular processes, supervised functional classification, variable selection.
AMS 2010 subject classifications: Primary 62H30; secondary 62G99.

1 Introduction

In the booming field of statistics with functional data (often referred to as Functional Data Analysis, FDA) the computational and numerical aspects, as well as the real data applications, have understandably played a major role so far. However, the underlying probabilistic theory, connecting the models which generate the data (i.e., the stochastic processes) with the statistical functional methods is far less developed. The present work is an attempt to contribute to that connection. See, for example, Cuevas (2014) or Wang et al. (2016) for recent survey papers on FDA and Horváth and Kokoszka (2012) for a recent monograph.

Our conclusions here will present both theoretical and practical aspects. Roughly speaking, our aim is to prove that in the field of supervised functional classification, there are many useful underlying models (defined in terms of appropriate stochastic processes) for which the expression of the optimal rule can be explicitly given. This will also lead to a natural procedure for variable selection in some of these models. We are also able to shed some light on the interesting phenomenon of "near perfect classification", discussed by Delaigle and Hall (2012a). This phenomenon does not appear (except for trivial or artificial cases) in the classical finite-dimensional classification theory.

1.1 The framework: supervised classification and absolute continuity

We are concerned here with the problem of binary functional supervised classification. Throughout the paper $X = X(t) = X_t = X(t, \omega)$ will denote a stochastic process with $t \in I$, for some compact interval I. Unless otherwise specified we will assume $I = [0, T]$, with $T > 0$. This process can be observed in two populations identified by the random "label" variable Y; the conditional distributions of $X|Y = i$ for $i = 0, 1$, denoted by P_i, are assumed to be Gaussian.

As usual in the supervised classification setting, the aim is to classify an "unlabeled" observation X according to whether it comes from P_0 or from P_1. A classification rule is just a measurable function $g : \mathcal{X} \to \{0, 1\}$, where $\mathcal{X}$ is the space of trajectories of the process X.

The expression $P_1 << P_0$ indicates that P_1 is absolutely continuous with respect to P_0 (i.e. $P_0(A) = 0$ entails $P_1(A) = 0$). Note that, from the Hájek-Feldman dichotomy for Gaussian measures (Feldman, 1958), $P_1 << P_0$ implies also $P_0 << P_1$, so that both measures are in fact mutually absolutely continuous (or "equivalent"). This is often denoted $P_1 \sim P_0$.

When P_0 and P_1 are completely known in advance and $P_1 << P_0$, it can be shown that the optimal classification rule (often called *Bayes rule*) is

$$g^*(x) = \mathbb{I}_{\{\eta(x)>1/2\}} = \mathbb{I}_{\left\{\frac{dP_1(x)}{dP_0}>\frac{1-p}{p}\right\}}, \tag{1}$$

where $\mathbb{I}$ denotes the indicator function, $\eta(x) = \mathbb{P}(Y = 1|X = x) = \mathbb{E}(Y|X = x)$, $p = \mathbb{P}(Y = 1)$ and $\frac{dP_1(x)}{dP_0}$ is the Radon-Nikodym derivative of P_1 with respect to P_0. The corresponding minimal "classification error" (i.e., the misclassification probability) $L^* = \mathbb{P}(g^*(X) \neq Y)$ is called *Bayes error*; see, e.g., Devroye *et al.* (1996) for general background and Baíllo *et al.* (2011a) for additional details on the functional case. We assume throughout $0 < p < 1$.

If $\frac{dP_1(x)}{dP_0}$ is completely known, there is not much else to be said. However, in practice, this is not usually the case. For example if P_0 and P_1 are Gaussian measures, the general expression of $\frac{dP_1(x)}{dP_0}$ is sometimes available but it depends on the respective mean and covariance functions, which could be partially unknown.

The term "supervised" accounts for the fact that, in any case, a data set of "well-classified" independent observations $\mathcal{D}_n = ((X_1, Y_1), \ldots, (X_n, Y_n))$ from (X, Y) is assumed to be available beforehand. So, the classification rules are in fact constructed in terms of the sample data $\mathcal{D}_n$. Throughout the paper, the functional data $X = X(t)$ are supposed to be "densely observed"; see, e.g., Cuevas (2014, Sec 2.1). A common strategy is to use these data to estimate the optimal rule (1). This is the so-called *plug–in approach*. It is often implemented in a non-parametric way (e.g., estimating $\eta(x)$ by a nearest-neighbor estimator) which does not require much information on the precise structure of $\eta(x)$ or $\frac{dP_1(x)}{dP_0}$. However, in some other cases we have a quite precise information on the structure of $\frac{dP_1(x)}{dP_0}$, so that we can take advantage of this information to get better plug-in estimators of $g^*(x)$.

1.2 Some especial characteristics of classification with functional data. The aims of this work

It can be seen from the above paragraphs that the supervised classification problem can be stated, with almost no formal difference, either in the ordinary finite-dimensional situation (where X takes values on the Euclidean space $\mathcal{X} = \mathbb{R}^d$) or in the functional case (where X is a stochastic process). In spite of these formal analogies, the passage to an infinite-dimensional (functional) sample space $\mathcal{X}$ entails some very important challenges. For example, the classical Fisher linear rule, which is still very popular in the finite-dimensional setting, cannot be easily adapted to the functional case (see, Baíllo *et al.* (2011b) for more details and references). However, we are more concerned here with another crucial difference, namely the lack of a natural "dominant" measure in functional spaces, playing a similar role to that of Lebesgue measure in $\mathbb{R}^d$. If we are working with Gaussian measures in $\mathbb{R}^d$, the optimal rule (1) can be established (using the chain rule for Radon-Nikodym derivatives) in terms of the ordinary (Lebesgue) densities of P_0 and P_1. In the functional case, we are forced to work with the "mutual" Radon-Nikodym derivatives dP_1/dP_0, provided that $P_1 << P_0$. Usually these derivatives are not easy to calculate or to work with. However, in some important examples they are explicitly known and reasonably easy to handle.

So first, we give and interpret explicit expressions for the optimal (Bayes) classification rule in some relevant cases with $P_1 << P_0$. Similar ideas are developed in Baíllo *et al.* (2011a) and Cadre (2013) but, unlike these references, our approach here relies heavily on the theory of Reproducing Kernel Hilbert Spaces (RKHS). See Sections 2 and 3 below.

In the second place, we consider the mutually singular case $P_1 \perp P_0$, i.e., when there exists a Borel set A such that $P_0(A) = 1$ and $P_1(A) = 0$. Note that this mutually singular (or "orthogonal") case is rarely found in the finite-dimensional classification setting, except in a few trivial or artificial cases. However, in the functional setting (that is, when P_1 and P_0 are distributions of stochastic processes) the singular case is an important, very common situation. As we argue in Section 4, this mutual singularity notion is behind the near perfect classification phenomenon described in Delaigle and Hall (2012a); see also Cuesta-Albertos and Dutta (2016). The point is to look at this phenomenon from a slightly different (coordinate free) RKHS perspective. We also show that an approximately optimal ("near perfect") classification rule to discriminate between P_0 and P_1 when $P_1 \perp P_0$, can be obtained in terms of the optimal rules of a sequence of problems (P_0^n, P_1^n) with $P_1^n << P_0^n$.

Third, in Section 5 we discuss a practical application of our results on the explicit calculation of optimal classification rules. We show that these results can be used to select the optimal projection of the sample trajectories $x(t)$, $t \in [0,1]$ into a d dimensional marginal $(X(t_1), \ldots, X(t_d))$ which is chosen in order to optimize the classification error based on such projection. The theoretical conditions for the optimality (in the original functional problem) of this method are also addressed. Section 6 is devoted to numerical experiments.

All the proofs, as well as some additional results, and some details about the simulation models we have used, can be found in the *Supplementary material* document.

2 Radon-Nikodym densities for Gaussian processes: some background

In the following paragraphs we review, for posterior use, some results regarding the explicit calculation of Radon-Nikodym derivatives of Gaussian processes in the convenient setting provided by the theory of Reproducing Kernel Hilbert Spaces. We first recall some very basic facts on the RKHS theory; see Berlinet and Thomas-Agnan (2004) for background.

Given a symmetric positive-semidefinite function $K(s,t)$, defined on $[0,T] \times [0,T]$ (in our case K will be the covariance function of a process), let us define the space $\mathcal{H}_0(K)$ of all real functions which can be expressed as finite linear combinations of type $\sum_i a_i K(\cdot, t_i)$ (i.e., the linear span of all functions $K(\cdot, t)$). In $\mathcal{H}_0(K)$ we consider the inner product $\langle f, g \rangle_K = \sum_{i,j} \alpha_i \beta_j K(s_j, t_i)$, where $f(x) = \sum_i \alpha_i K(x, t_i)$ and $g(x) = \sum_j \beta_j K(x, s_j)$.

Then, the RKHS associated with K, $\mathcal{H}(K)$, is defined as the completion of $\mathcal{H}_0(K)$. More precisely, $\mathcal{H}(K)$ is the set of functions $f : [0,T] \to \mathbb{R}$ which can be obtained as t pointwise limit of a Cauchy sequence $\{f_n\}$ of functions in $\mathcal{H}_0(K)$. The theoretical motivation for this definition is the well-known Moore-Aronszajn Theorem (see Berlinet and Thomas-Agnan (2004), p. 19). The functions in $\mathcal{H}(K)$ have the "reproducing property" $f(t) = \langle f, K(\cdot, t) \rangle_K$, where $\langle \cdot, \cdot \rangle$ denotes the natural extension to $\mathcal{H}(K)$ of the inner product in $\mathcal{H}_0(K)$.

Thus, in a very precise way, $\mathcal{H}(K)$ can be seen as the "natural Hilbert space" associated with a process $\{X(t), t \in [0,T]\}$. In fact, as we will next see, the space $\mathcal{H}(K)$ is deeply involved in some relevant probabilistic and statistical notions.

The following result is a slightly simplified version of Theorem 7A in Parzen (1961); see also Parzen (1962). It will be particularly useful in the rest of this paper.

Theorem 1. *(Parzen, 1961, Th. 7A). Let us denote by P_1 the distribution of a Gaussian process $\{X(t),\ t \in [0,T]\}$, with continuous trajectories, mean function denoted by $m = m(t) = \mathbb{E}(X(t))$ and continuous covariance function denoted by $K(s,t) = Cov(X(s), X(t))$. Let P_0 be the distribution of another Gaussian process with the same covariance function and with mean function identically 0. Then, $P_1 << P_0$ if and only if the mean function m belongs to the space $\mathcal{H}(K)$. In this case,*

$$\frac{dP_1(X)}{dP_0} = \exp\left(\langle X, m \rangle_K - \frac{1}{2} \langle m, m \rangle_K \right). \tag{2}$$

In the case $m \notin \mathcal{H}(K)$, we have $P_1 \perp P_0$.

Some remarks on this result.

(a) Except for trivial cases, the trajectories x of the process $X(t)$ are not included, with probability one, in $\mathcal{H}(K)$; see, e.g., Berlinet and Thomas-Agnan (2004, p. 66) for details. Thus, the expression $\langle X, m \rangle_K$ is employed, with some abuse of notation, to denote $\Psi^{-1}(m)$, where Ψ is a natural congruence $\Psi : \bar{\mathcal{L}}(X) \to \mathcal{H}(K)$ defined (up to completion) between the closed linear span $\bar{\mathcal{L}}(X)$ of the process X and $\mathcal{H}(K)$ by $\Psi(\sum_i a_i X_{t_i}) = \sum_i a_i K(\cdot, t_i)$. See Berlinet and Thomas-Agnan (2004, p. 65).

(b) Note that his definition of $\langle X, m \rangle_K$ in terms of a congruence, is strongly reminiscent of the definition of the Itô's stochastic integral. As a matter of fact, $\langle X, m \rangle_K$ can be seen itself as a stochastic integral. To see this consider the case where $X(t) = B(t)$ is the standard

Brownian Motion and $K(s,t) = \min(s,t)$. Then, it can be seen that $\mathcal{H}(K)$ coincides with the so-called Dirichlet space $\mathcal{D}[0,T]$ of those real functions g on $[0,T]$ such that there exists g' almost everywhere in $[0,T]$ with $g' \in L^2[0,T]$, and $g(t) = \int_0^t g'(s)ds$. The norm in $\mathcal{D}[0,T]$ is defined by $\|g\|_K = \left(\int_0^T g'^2(t)dt\right)^{1/2}$. Likewise, the inverse congruence $\langle X, m\rangle_K$ can also be expressed as the stochastic integral $\int_0^T m'(s)dB(s)$. Thus, Theorem 1 can be seen as an extension of the classical Cameron-Martin Theorem (Mörters and Peres, 2010, p. 24), which is stated for $X(t) = B(t)$. Some additional references on Radon-Nikodym derivatives in function spaces are Varberg (1961, 1964), Shepp (1966), Kailath (1971) and Segall and Kailath (1975), among others.

3 Classification of absolutely continuous Gaussian processes

In this section we consider the supervised classification problem, as stated in Subsection 1.1, under the following general model

$$P_i : m_i(t) + \epsilon_i(t), \text{ for } i = 0,1 \tag{3}$$

where, for $i = 0,1$, $\{\epsilon_i(t),\ t \in [0,T]\}$ are "noise processes" with mean 0 and continuous trajectories, and $m_i(t)$, $t \in [0,T]$ are some continuous functions defining the respective "trends" of P_0 and P_1.

The following result provides the expression of the Bayes (optimal) rule and the corresponding minimal error probability for this case, under the usual assumption of homoskedasticity. While the proof is a simple consequence of Theorem 1 and Theorem 1 in Baíllo *et al.* (2011a), this result will be essential in the rest of the paper.

Theorem 2. *In the classification problem under the model (3) assume*

(a) the noise processes ϵ_i are both Gaussian with continuous trajectories and common continuous covariance function $K(s,t)$.

(b) $m := m_1 - m_0 \in \mathcal{H}(K)$*, where* $(\mathcal{H}(K), \|\cdot\|_K)$ *denotes the RKHS associated with* K*.*

Then, the optimal Bayes rule is given by $g^*(X) = \mathbb{I}_{\{\eta^*(X)>0\}}$*, where*

$$\eta^*(X) = \langle X, m\rangle_K - \frac{1}{2}(\|m_1\|^2 - \|m_0\|^2) - \log\left(\frac{1-p}{p}\right). \tag{4}$$

Also, the corresponding optimal classification error $L^* = \mathbb{P}(g^*(X) \neq Y)$ *is*

$$L^* = (1-p)\Phi\left(-\frac{\| m \|_K}{2} - \frac{1}{\| m \|_K}\log\left(\frac{1-p}{p}\right)\right) + p\Phi\left(-\frac{\| m \|_K}{2} + \frac{1}{\| m \|_K}\log\left(\frac{1-p}{p}\right)\right),$$

where Φ *is the cumulative distribution function of a standard normal random variable and* $p = \mathbb{P}(Y=1)$ *is the prior probability of* P_1*. When* $p = 1/2$*, we have* $L^* = 1 - \Phi\left(\frac{\|m\|_K}{2}\right)$*.*

If we compare this result with the optimal rule given in Theorem 1 of the paper Delaigle and Hall (2012a), we see that (4) does not explicitly depends on the eigenvalues and eigenvectors of the covariance operator. Instead, the general expression (4) is given in terms of the "stochastic integral" $\langle X, m\rangle_K$. We will comment on this in more detail in Section 4.

3.1 The heteroskedastic case

General results such as those of Theorem 2 above, seem difficult to obtain when the covariance functions under P_0 and P_1 (denoted by K_0 and K_1, respectively) are different. Still, there exists a relatively rich literature about Gaussian measures giving characterizations of absolute continuity and expressions of Radon-Nikodym derivatives. These results can be applied to provide explicit expressions of optimal classifiers in some significant particular cases of the general problem (3), which include discrimination between non-homoskedastic models. Here, we use a result due to Shepp (1966), to obtain an expression of the Bayes rule when P_0 and P_1 are equivalent to the Wiener measure (the one corresponding to the Brownian motion).

In Theorem 3 below, $\varphi_{i,j}$, for $i = 0, 1$ and $j \geq 0$, stand for the unit eigenfunctions (with respect to the L^2 norm) of the integral operator defined by a kernel $\tilde{K}_i \in L^2([0,T] \times [0,T])$

such that

$$\int_0^s \int_0^t \tilde{K}_i(u,v)dudv = \min\{s,t\} - K_i(s,t). \tag{5}$$

We denote by $\lambda_{i,j}$ the corresponding eigenvalues, $x_{i,j} = \int_0^T \varphi_{i,j}(t)dX(t)$ and $\xi_{i,j} = \langle m_i', \varphi_{i,j}\rangle_{L^2}$. The existence of a kernel satisfying (5) and the a.e. existence of $m_i' \in L^2[0,T]$ is guaranteed by Theorem 1 in Shepp (1966). It can be also proved that $\lambda_{i,j} < 1$, for all i, j.

Theorem 3. *Let us consider the classification problem (3), where P_0 and P_1 are absolutely continuous with respect to the Wiener measure. Let $g^*(x) = \mathbb{I}_{\{\eta^*(x)>0\}}$ be the Bayes rule. Then, $\eta^*(X) = \sum_{j=1}^\infty \eta_j(X) - 2\log((1-p)/p)$, where*

$$\eta_j(X) := \log\left(\frac{1-\lambda_{0,j}}{1-\lambda_{1,j}}\right) + \frac{(x_{0,j}-\xi_{0,j})^2}{1-\lambda_{0,j}} - \frac{(x_{1,j}-\xi_{1,j})^2}{1-\lambda_{1,j}} - (x_{0,j}-x_{1,j})^2. \tag{6}$$

Observe that the Bayes rule in Theorem 3 is a quadratic classifier applied to the scores $x_{i,j}$, which in turn are defined as stochastic integrals with respect to $X(t)$.

The general expression of the optimal rule is substantially simpler in some particular cases. For example, for a Brownian motion versus a Brownian bridge we have:

Corollary 1. *Let us consider the classification problem (3), where $m_1 \equiv m_0 \equiv 0$, ϵ_0 is the standard Brownian motion on $[0,T]$, with $T < 1$, and ϵ_1 is the standard Brownian bridge on $[0,T]$. Let $g^*(x) = \mathbb{I}_{\{\eta^*(x)>0\}}$ be the Bayes rule. Then,*

$$\eta^*(X) = -\log(1-T) - \frac{X(T)^2}{(1-T)} - 2\log\left(\frac{1-p}{p}\right). \tag{7}$$

According to (7) we classify X as a Brownian bridge whenever $|X(T)|$ is below a certain threshold depending on T. Some authors (e.g., Lindquist and McKeague (2009)) call T an “impact point”, in the sense that the optimal classifier depends just on $X(T)$, so that there is no need of using the whole trajectory.

4 Another look at the "near perfect classification" phenomenon

The starting point in this section is again the classification problem between the Gaussian processes P_0 and P_1 defined in (3), where ϵ_0 and ϵ_1 are identically distributed according to the Gaussian process $\epsilon(t)$ with covariance function $K(s,t) = \mathbb{E}(\epsilon(s)\epsilon(t))$. The mean functions are $m_0(t) = 0$ and $m_1(t) = \sum_{j=1}^{\infty} \mu_j \phi_j(t)$, where the ϕ_j are the eigenfunctions of the Karhunen-Loève expansion of K, that is $K(s,t) = \sum_{j=1}^{\infty} \theta_j \phi_j(s)\phi_j(t)$.

Let us assume for simplicity that the prior probability is $\mathbb{P}(Y = 1) = 1/2$. This model has been considered by Delaigle and Hall (2012a). In short, these authors provide the explicit expression of the optimal rule under the assumption $\sum_{j=1}^{\infty} \theta_j^{-2}\mu_j^2 < \infty$. In addition, they find that, when $\sum_{j=1}^{\infty} \theta_j^{-1}\mu_j^2 = \infty$, the classification is "near perfect" in the sense that one may construct a rule with an arbitrarily small classification error. To be more specific, the classification rule they propose is the so-called "centroid classifier", T_n, defined by $T_n(X) = 1$ if and only if $D^2(X, \bar{X}_1) - D^2(X, \bar{X}_0) < 0$, where $\bar{X}_0$, $\bar{X}_1$ denote the sample means of the training data from P_0 and P_1 and $D(X, \bar{X}_j) = |\langle X, \psi\rangle_{L^2} - \langle \bar{X}_j, \psi\rangle_{L^2}|$, with $\langle X, \psi\rangle_{L^2} = \int_0^T X(t)\psi(t)dt$ and $\psi(t) = \sum_{j=1}^{\infty} \theta_j^{-1}\mu_j\phi_j(t)$. Of course, this requires $\psi \in L^2$ which (from Parseval's identity) amounts to $\sum_{j=1}^{\infty} \theta_j^{-2}\mu_j^2 < \infty$. Then, the asymptotic version of the classifier T_n under the assumed model is

$$T^0(X) = 1, \text{ if and only if } (\langle X, \psi\rangle_{L^2} - \langle m_1, \psi\rangle_{L^2})^2 - \langle X, \psi\rangle_{L^2}^2 < 0. \tag{8}$$

Now, a more precise summary of the above discussion is as follows.

Theorem 4. *(Delaigle and Hall, 2012a, Th.1). Let us consider the binary classification problem (3) under the Gaussian homoskedastic model with $m_0(t) = 0$ and continuous K.*

(a) If $\sum_{j\geq 1} \theta_j^{-1}\mu_j^2 < \infty$, the minimal (Bayes) misclassification probability is given by $err_0 = 1 - \Phi\left(\frac{1}{2}(\sum_{j\geq 1} \theta_j^{-1}\mu_j^2)^{1/2}\right)$. Moreover, under the extra assumption $\sum_{j\geq 1} \theta_j^{-2}\mu_j^2 < \infty$, the optimal classifier (that achieves this error) is the rule T^0 defined in (8).

(b) *If* $\sum_{j\geq 1}\theta_j^{-1}\mu_j^2 = \infty$*, the minimal misclassification probability is* $err_0 = 0$ *and it is achieved, in the limit, by a sequence of classifiers constructed from* T^0 *by replacing the function* ψ *with* $\psi^{(r)} = \sum_{j=1}^r \theta_j^{-1}\mu_j\phi_j(t)$*, with* $r \uparrow \infty$*.*

As pointed out in Delaigle and Hall (2012a), *"We argue that those [functional classification] problems have unusual, and fascinating, properties that set them apart from their finite dimensional counterparts. In particular we show that, in many quite standard settings, the performance of simple [linear] classifiers constructed from training samples becomes perfect as the sizes of those samples diverge [...]. That property never holds for finite dimensional data, except in pathological cases."* Our purpose here is to show that the setup of Theorem 4 (that is, Theorem 1 in Delaigle and Hall (2012a)) can be analyzed from the point of view of RKHS theory. We do this in Theorems 5 and 6 below.

Theorem 5. *In the framework of the classification problem considered in Theorem 4, with continuous trajectories and continuous common covariance function* K*, we have*

(a) $\sum_{j\geq 1}\theta_j^{-1}\mu_j^2 < \infty$ *if and only if* $P_1 \sim P_0$*. In that case, the Bayes rule* g^* *is*

$$g^*(X) = 1 \text{ if and only if } \langle X, m\rangle_K - \frac{1}{2}\parallel m \parallel_K^2 > 0, \tag{9}$$

with the notation of Equation (4). The corresponding optimal (Bayes) classification error is $L^* = 1 - \Phi\left(\parallel m \parallel_K /2\right)$*. Under the additional condition* $\sum_{j\geq 1}\theta_j^{-2}\mu_j^2 < \infty$*, the optimal rule given in Theorem 4 (a) provides an alternative expression of (9) based on the "coordinates"* θ_j *and* μ_j*.*

(b) $\sum_{j\geq 1}\theta_j^{-1}\mu_j^2 = \infty$ *if and only if* $P_1 \perp P_0$*. In this case the Bayes error is* $L^* = 0$*.*

We next make explicit the meaning of the near perfect classification phenomenon. The next theorem establishes that in the singular case (where the Bayes error is zero) we can construct a classification rule whose misclassification probability is arbitrarily small.

Theorem 6. *Consider the singular case analyzed in Theorem 5. Then, there is a sequence of approximating classification problems, of type P_{0n} vs. P_{1n}, of the absolutely continuous type $P_{0n} \sim P_{1n}$, such that $P_{in} \underset{n\to\infty}{\overset{weakly}{\longrightarrow}} P_i$, for $i = 0, 1$ and the misclassification probabilities of the respective optimal rules (whose expressions are explicitly known) tend to zero.*

Now, we are in position to comment the contributions of the above Theorems 5 and 6, from the perspective of Theorem 1 in Delaigle and Hall (2012a) (see Theorem 4 above for a slightly simplified version). First, Theorem 5 is, in some sense, analogue to the Delaigle-Hall's result. In the absolutely continuous case, Theorem 5 (a) provides a completely general, coordinate-free expression for the Bayes rule. It only requires the condition $\sum_{j\geq 1} \theta_j^{-1}\mu_j^2 < \infty$ which is minimal in the sense that it amounts to $P_0 \sim P_1$. Moreover, under the Delaigle-Hall's assumption $\sum_{j\geq 1} \theta_j^{-2}\mu_j^2 < \infty$, such Bayes rule can be expressed in "elementary terms" with no resort to the stochastic integral $\langle X, m\rangle_K$ which appears in (9). This highlights an interesting contribution of Theorem 1 in Delaigle and Hall (2012a) which remains "hidden" unless the whole problem is considered from the RKHS point of view.

Theorems 5 (b) and 6 shed some light on the "near-perfect" classification phenomenon in two specific aspects. First, Theorem 5 (b) shows that Delaigle-Hall‘s condition $\sum_{j\geq 1} \theta_j^{-1}\mu_j^2 = \infty$ has a probabilistic interpretation in terms of mutual singularity of measures. Second, Theorem 6 shows that the classification problem in this singular case can be arbitrarily approximated by a sequence of problems in the absolutely continuous case for which the Bayes rules are explicitly known. This establishes an useful link between the dual cases of singularity and absolutely continuity.

5 An application to dimension reduction via variable selection

Variable selection (VS) is becoming a popular methodology for dimension reduction in FDA, alongside with the classical Principal Component Analysis (PCA, see Hall, P. (2011) for a survey) or Partial Least Squares (PLS, see Delaigle and Hall (2012b)). In general, all dimension reduction methods aim at transforming, via projections, the original functional data into a finite-dimensional counterpart; see also Mosler and Mozharovskyi (2015); Bongiorno and Goia (2016) for recent related references. The underlying idea is to simplify the handling of the original data by replacing them with carefully chosen finite-dimensional projections. This usually simplifies the statistical treatment, at the cost of some loss of information.

Unlike PCA and PLS, the result of using a VS methodology is directly interpretable in terms of the original data since the final output of VS is just a part of the trajectories $\{x(t),\ t \in [0,1]\}$, instead of a general linear transformation of them (which is the case in PCA or PLS). More precisely, VS aims at defining a suitable finite-dimensional projection

$$\{x(t),\ t \in [0,1]\} \mapsto (x(t_1), \ldots, x(t_d)). \tag{10}$$

This approach is well established in functional regression where a wide range of VS methodologies, mostly dealing with penalized linear models, have been proposed (Aneiros and Vieu, 2014). In the functional supervised classification setting, the specific nature of this problem has motivated different VS strategies, not necessarily relying on linear assumptions (which are common in their regression counterparts). Let us cite, for example, Delaigle, *et al.* (2012) who consider a VS procedure based on the minimization of the cross-validation error. A logistic regression model for binary classification, based on impact points with Brownian-like predictors, is provided in Lindquist and McKeague (2009). In Berrendero *et al.* (2016b), variables are selected according to the local maxima of a relevance function in an intrinsic (model-free) way. The same authors (Berrendero *et al.*, 2016a) study the performance of a modified version of the so-called mRMR (minimum Redundancy Maximum Relevance) methodology, a VS strategy quite popular in the machine learning community.

In what follows we consider the variable selection methodology in our RKHS framework for the purpose of binary functional classification. Our goal in the rest of this section is to answer two natural questions:

(1) How to select (in an asymptotically optimal way) the "impact points" $t_1, \dots, t_d$ in (10)?

(2) In which cases the proposed variable selection methodology does not entail any loss with respect to the optimal solution achieved using the whole data trajectories (with no dimension reduction involved)?

The following subsections 5.1 and 5.2 are devoted to answer to the questions (1) and (2), respectively. As we will see, the theoretical results in Section 3 will play a relevant role here. In a way, the practical use of variable selection methodologies for functional data relies on the typical "smoothness" properties of functional data: one expects that the choice of $x(t_0)$ will be nearly equivalent to that of $x(t_0 + \epsilon)$. As pointed out by Müller (2009), *"(...) functional data are inherently infinite-dimensional and smoothness often is a central assumption. Smoothness has no meaning for multivariate data analysis, which in contrast to FDA is permutation invariant"*

5.1 A natural method for variable selection in the Gaussian case

We deal here with the functional supervised classification problem under the model (3), assuming that the error processes ϵ_0 and ϵ_1 are Gaussian and homoskedastic. We want to use a variable selection methodology, as indicated in (10), where d is fixed. We assume throughout that all the marginals $(X(t_1), \dots, X(t_d))$ are non-degenerate (that is, the corresponding covariance matrix is not singular).

The goal is to properly select (using the whole trajectories) the impact points $(t_1, \dots, t_d)$ so that the original functional data would be reduced to $(X(t_1), \dots, X(t_d))$. Then, our classification problem boils down to discriminate between to d-dimensional homoskedastic

Gaussian distributions with mean vectors $(m_i(t_1), \ldots, m_i(t_d))$, $i = 0, 1$ and common covariance matrix, denoted by $K_{t_1,\ldots,t_d}$, whose i, j entry is $K(t_i, t_j)$. It is well-known (see, e.g., Izenman (2008, p. 244)) that the optimal misclassification probability in such a discrimination problem is a function of the Mahalanobis distance between both mean vectors. The square of such distance is given by

$$\psi(t_1, \ldots, t_d) := m_{t_1,\ldots,t_d}^{\top} K_{t_1,\ldots,t_d}^{-1} m_{t_1,\ldots,t_d},$$

where $u^{\top}$ denotes the transpose of u and $m_{t_1,\ldots,t_d} = (m_1(t_1), \ldots, m_1(t_d)) - (m_0(t_1), \ldots, m_0(t_d))$ is the difference between both mean vectors.

It is intuitively obvious that the misclassification probability should be a decreasing function of this Mahalanobis distance (in fact, this will be formally established in the proof of Theorem 7 (c) below). Hence, a natural criterion for variable selection in this setup is as follows: take $(t_1^*, \ldots, t_d^*) \in \Theta$ such that

$$\psi(t_1^*, \ldots, t_d^*) \geq \psi(t_1, \ldots, t_d), \text{ for all } (t_1, \ldots, t_d) \in \Theta, \tag{11}$$

where $\Theta \subset [0, 1]^d$ is a compact set with $t_i \leq t_{i+1}$ for all $(t_1, \ldots, t_d) \in \Theta$.

We will denote this criterion RK-VS. Here RK is after Reproducing Kernel since, as we will see in Subsection 5.2, the RKHS theory provides a simple functional interpretation for this method. Some remarks are in order:

(a) *On the meaning of* (11): by construction, the choice $(t_1^*, \ldots, t_d^*)$ is optimal in the sense that the discrimination rule based on the data $(X(t_1^*), \ldots, X(t_d^*))$ has the minimal probability of classification error among all the rules based on the projections $(X(t_1), \ldots, X(t_d))$.

(b) *Regarding the choice of* Θ: in principle, the method (11) could work for any choice of the compact set $\Theta \subset \Theta_0$, where $\Theta_0 = \{(t_1, \ldots, t_d) \in [0, 1]^d : t_1 \leq \ldots \leq t_d\}$. However, we must take care of the degeneracy issues, appearing when $K_{t_1,\ldots,t_d}$ is a singular matrix. From our non-degeneracy assumptions, this would only happen when some coordinates among the

t_i are identical. A simple way to cope with this restriction is just taking

$$\Theta = \Theta(\delta) = \{(t_1, \ldots, t_d) \in [0, T]^d : \, t_{(i)} + \delta \leq t_{(i+1)}, \text{ for } i = 0, \ldots, d-1\},$$

where $t_{(i)}$, $i = 1, \ldots, d$, denote the ordered values (with $t_{(0)} := 0$), and $\delta > 0$ is a constant small enough. We will assume henceforth this choice for Θ but the whole approach would also work if we just take $\Theta = \Theta_0$, as long as we understand that if the optimum in (11) is achieved at a point with ties in the coordinates, we should identify all coordinates i, j for which $t_i = t_j$ and to reduce the dimension accordingly. This would mean that the optimum would correspond to some point in $[0, 1]^{d'}$ for some $d' < d$.

(c) *Estimation and consistency.* Since m and K are usually unknown, we propose to replace them by appropriate estimators $\hat{m}_{t_1,\ldots,t_d}(t)$ and $\hat{K}_{t_1,\ldots,t_d}$ (more on this below). So the criterion we suggest for variable selection in practice is to choose points such that

$$\begin{aligned} &\hat{\psi}(\hat{t}_1, \ldots, \hat{t}_d) \geq \hat{\psi}(t_1, \ldots, t_d) \text{ for all } (t_1, \ldots, t_d) \in \Theta, \\ &\text{where } \hat{\psi}(t_1, \ldots, t_d) := \hat{m}^\top_{t_1,\ldots,t_d} \hat{K}^{-1}_{t_1,\ldots,t_d} \hat{m}_{t_1,\ldots,t_d}. \end{aligned} \tag{12}$$

The simplest estimators for the mean function m and the covariance operator kernel K are their corresponding empirical counterparts (some additional discussion on this can be found in the Supplementary Material document). Assuming this, we have proved (see Theorem S1 in the Supplementary Material) the consistency of the empirical RK-VS procedure defined in (12), combined with the classical Fisher's linear rule for the discrimination problem based on the selected variables $(X(\hat{t}_1), \ldots, X(\hat{t}_d))$. This classifier will be denoted RK-C. In more precise terms, this consistency means that the misclassification probability L_n of the linear rule corresponding to the variable choice $(X(\hat{t}_1), \ldots, X(\hat{t}_d))$ converges almost surely, as $n \to \infty$ to the minimal misclassification probability L^* corresponding to the "population-based" optimal choice $(X(t_1^*), \ldots, X(t_d^*))$.

(d) *On the effective implementation of the variable selection method*: the practical calculation of the optimum in (12) might be difficult in practice when $d > 3$. The non-concavity of $\hat{\psi}(t_1, \ldots, t_d)$ could lead to situations with potentially many local maxima. We have used an iterative, greedy algorithm to deal with this problem. This and other issues associated with the use of RK-VS are addressed in Subsections S2.1 and S2.2 of the Supplementary Material document. A toy example illustrating how RK-methods work is also included in that document (at S2.2).

5.2 An interpretation in functional terms. Optimality

In this subsection we discuss the conceptual link between the contents of Subsection 5.1 and the rest of the paper. In other words, we address the functional interpretation of the RK-C classification method which results from the combination of the variable selection procedure (11) with the use of Fisher's linear classification rule applied to the selected variables.

In principle, we should recall that variable selection, as any other dimension reduction method, might entail some loss of efficiency, which could be seen as the cost to be paid for the simplification in the data. But in some particular situations the reduced data (via variable selection in our case) could be used instead of the original ones with no loss of efficiency at all. The identification of these situations is the purpose of the following theorem which also provides an interpretation, in RKHS terms, of the finite dimensional Mahalanobis distance. The proof (which is a simple application of Theorem 2) can be found in the Supplementary Material Document.

Theorem 7. *Let us consider the functional classification problem of discriminating between the processes P_0 and P_1 with continuous mean functions m_i and continuous trajectories of type $X(t) := m_i(t) + \epsilon_i(t)$, $t \in [0, T]$, the ϵ_i are independent Gaussian processes with mean 0 and common continuous covariance function $K(s, t)$. Then,*

(a) the d-dimensional classification problem of discriminating between P_0 and P_1 on the

sole basis of the projections $(X(t_1),\ldots,X(t_d))$ *at given points* $t_1,\ldots,t_d$ *is equivalent (in the sense of having the same optimal rule and Bayes error) to the corresponding functional problem whenever* $m := m_1 - m_0$ *has the form* $m(\cdot) = \sum_{i=1}^d \alpha_i K(\cdot, t_i)$.

(b) Denote by $K_{t_1,\ldots,t_d}$ *the covariance matrix of* $(X(t_1),\ldots,X(t_d))$ *and let* $m_{t_1,\ldots,t_d}$ *be the difference between both mean vectors (under* P_1 *and* P_0*). Then, the square Mahalanobis distance, given by* $m_{t_1,\ldots,t_d}^\top K_{t_1,\ldots,t_d}^{-1} m_{t_1,\ldots,t_d}$*, between the distributions of* $(X(t_1),\ldots,X(t_d))|Y = i$ *for* $i = 0, 1$*, coincides with* $\|m\|_K^2$*, the square norm of* m *in the RKHS induced by* K*, provided again that* $m(\cdot) = \sum_{i=1}^d \alpha_i K(\cdot, t_i)$.

(c) The optimal choice for $(t_1,\ldots,t_d)$*, in the sense of minimizing the classification error, is obtained by maximizing* $\|m\|_K^2$ *among all functions* m *in the RKHS space having an expression of type* $m(\cdot) = \sum_{i=1}^d \alpha_i K(\cdot, t_i)$.

At this point, one might wonder about the role of the assumption $m(\cdot) = \sum_{i=1}^d \alpha_i K(\cdot, t_i)$. The natural question is: to what extent such condition is needed in our approach to variable selection? In this respect, it is particularly important to note that the method defined in (12), *still makes sense even if such assumption is not fulfilled*; in that case, the method provides (asymptotically, if m and K must be estimated) the best choice $(X(t_1^*),\ldots,X(t_d^*))$ of variables (for a given d) in order to obtain a maximal separation in the Mahalanobis distance for the mean vectors under P_0 and P_1. Note that, in principle, this idea could be considered without any assumption on the functional model (except, perhaps, homoskedasticity). The contribution of Theorem 7 is just to establish in precise terms the conditions on the *functional* classification model under which the proposed variable selection procedure is optimal.

6 Experiments

In Section 5 we proposed a variable selection method (denoted RK-VS) which was associated, in a natural way, with a classifier (that we have denoted by RK-C). Such classifier is nothing but the standard Fisher's linear rule applied to the selected variables. For conceptual purposes is convenient to distinguish between both methods (RK-VS and RK-C) since, in principle, they could work separately, as we could use different classifiers on the variables selected by RK-VS. Still, the RK-C method is, as seen above, the natural choice under a Gaussian homoskedastic model. We have carried out extensive simulations in order to check the performances of both (RK-VS and RK-C) methods when compared with other alternatives for dimension reduction+classification. Due to space limitations, the corresponding outputs can be mostly found in the Supplementary Material document; although they are briefly commented in Subsection 6.2 below. We have just kept in the next subsection some results concerning classification in functional real data examples.

6.1 Real data examples

We now study the RK-C performance in two real data examples. We have chosen the "easiest" and the "hardest" data sets (from the classification point of view) of those considered in Delaigle and Hall (2012a). Given the close connections between our theoretical setting and that of these authors, this partial coincidence of data sets seems pertinent.

Thus, we follow the same methodology as in the cited paper, that is, we divide the data set randomly in a training sample of size n ($n = 30, 50, 100$) and a test sample with the remaining observations. Then, the RK-C classifier is constructed from the training set and it is used to classify the test data. The misclassification error rate is estimated through 200 runs of the whole process. The number of variables selected by RK-C is fixed by a standard leave-one-out cross-validation procedure over the training data.

We consider two data sets. *Wheat* data correspond to 100 near infrared spectra of wheat samples measured from 1100nm to 2500nm in 2nm intervals. Following Delaigle and Hall (2012a) we divide the data in two populations according to the protein content (more or less than 15) and use the derivative curves obtained with splines. For this data set near perfect classification is achieved. *Phoneme* is another popular data set in functional data

Table 1: Misclassification percentages (and standard deviations) for the classification methods considered in Table 2 of Delaigle and Hall (2012a), the new RK-C method and DHB

Data	n	Classification rules					
		CENT$_{PC1}$	CENT$_{PLS}$	NP	CENT$_{PCp}$	**RK-C**	DHB
Wheat	30	0.89 (2.49)	0.46 (1.24)	0.49 (1.29)	15.0 (1.25)	**0.08 (0.70)**	1.78 (2.97)
	50	0.22 (1.09)	0.06 (0.63)	0.01 (0.14)	14.4 (5.52)	**0.02 (0.13)**	0.75 (2.26)
Phoneme	30	22.5 (3.59)	24.2 (5.37)	24.4 (5.31)	23.7 (2.37)	**22.5 (3.34)**	23.0 (5.25)
	50	20.8 (2.08)	21.5 (3.02)	21.9 (2.91)	23.4 (1.80)	**21.3 (2.08)**	21.8 (3.95)
	100	20.0 (1.09)	20.1 (1.12)	20.1 (1.37)	23.4 (1.36)	**20.1 (1.13)**	20.4 (1.61)

analysis. It consists of log-periodograms obtained from the pronunciation of the phonemes "aa" (695 curves) and "ao" (1022 curves) recorded in 256 equispaced points. This is not an easy problem. As in Delaigle and Hall (2012a), we make the trajectories continuous with a local linear smoother and remove the noisiest part keeping the first 50 variables.

We have also included in the comparison the "componentwise" alternative proposed in Delaigle, *et al.* (2012), denoted DHB. This method also applies Fisher's linear rule (often denoted LDA) after a variable selection procedure. In this case the authors follow a wrapper approach where the variable selection is carried out by minimizing the leave-one-out classification error. Further details can be found in the Supplementary Material document.

Table 1 shows exactly the same results of Table 2 in Delaigle and Hall (2012a) plus two extra columns for our RK-C method (in boldface) and DHB. Since we have followed the same methodology, the results are completely comparable despite the minimum differences due to the randomness. CENT$_{PC1}$ and CENT$_{PLS}$ stand for the centroid classifier (8), where the function ψ is estimated via principal components or PLS components, respectively. NP refers to the classifier based in the non-parametric functional regression method proposed by Ferraty and Vieu (2006) and CENT$_{PCp}$ denotes the usual centroid classifier applied to the multivariate principal component projections. The outputs correspond to the average (over 200 runs) percentages of misclassification obtained for each method, sample size and data set. The values in parentheses correspond to the standard deviations of these errors.

The results show that the RK-C classifier is clearly competitive against the remaining methods. In addition, there is perhaps some interpretability advantage in the use of RK-C, as this method is based in dimension reduction via variable selection so that the "reduced data" are directly interpretable in terms of the original variables. Let us finally point out that the variable selection process is quite efficient: in the wheat example, near perfect classification is achieved using just one variable; in the much harder phoneme example, the average number of selected variables is around 2.5 for RK-C and 2 for DHB.

6.2 Some comments on the simulations included in the Supplementary Material document

The degree of success and efficiency of the RK-VS procedure is assessed by means of an extensive benchmark of 94 simulation models, four sample sizes and three classifiers. These experiments are specially designed to test variable selection in the functional classification setting since in all cases the corresponding optimal classifiers depend on the trajectories through their values at a finite number of points. The models include different relationships between distributions, several stochastic processes (such as Brownian motion, Brownian Bridge, Ornstein-Uhlenbeck and smoothed Brownian Motions) and differ in complexity and number of relevant variables. This benchmark was previously used in Berrendero *et al.* (2016a,b) so the results are fully comparable. Here, the RK-VS method is compared with the winners in these papers, which include both variable selection procedures and projection-based techniques for dimensionality reduction.

In general terms, RK-VS is the overall winner. It outperforms in average the other competitors for dimensionality reduction, with better results for bigger sample sizes. It achieves a good performance in terms of both classification accuracy and number d of variables (in the simulations d is chosen by standard validation procedures). RK-VS is also one of the (computationally) faster alternatives and its cost in computer time remains nearly constant with the sample size. The RK-VS results are very similar for the three considered classifiers.

These outputs are remarkable and encouraging especially taking into account that only 7 out of 94 models under study fulfill all the regularity conditions required for the best performance of RK-VS. A summary of the outputs and further details can be found in the Supplementary Material document. The full outputs are available in the auxiliary file *outputs.xlsx*.

7 Conclusions

(a) A formalized theory for binary classification in FDA can be built relying on the explicit calculations of the Radon-Nikodym derivatives between absolutely continuous Gaussian measures. The RKHS theory provides a particularly useful tool in this regard.

(b) The problem of discriminating between mutually singular measures can be also completely clarified in this setup, including approximate expressions for the optimal rule. Interestingly, these discrimination problems involving mutually singular distributions are rather artificial (and very uncommon) in the multivariate, finite-dimensional framework but become very significant in the functional setting. We provide some additional insights and theoretical motivations for the results in Delaigle and Hall (2012a) which is, to our knowledge, the first study of these "near-perfect" functional classification problems.

(c) Relying on the work by Shepp (1966), some hints can be also given on the harder functional problem of discriminating between two Gaussian heteroskedastic processes.

(d) The optimal (Bayes) classification rules and the corresponding minimal probabilities of classification error depend typically on unknown quantities and functions. Their estimation is a natural challenge for future work. There is also considerable room for the analysis of discrimination problems under non-Gaussian and/or multiclass problems.

(e) As a practical application of our results we have considered the following natural idea: in the setting of a functional binary discrimination problem, we look for the "best projection" of dimension d that minimizes the classification error of the usual d-dimensional Fisher's linear rule based on the selected projection. The results obtained in Sections 3 and

4 allow us to identify the cases in which such “best Fisher rule” is indeed optimal from the functional point of view. On the practical side, the proposed method for variable selection + classification shows a good performance in our empirical studies.

Acknowledgements. Research partially supported by Spanish grant MTM2013-44045-P. The comments from two referees and an Associate Editor are gratefully acknowledged.

Supplementary materials. The file *Supplementary.pdf* includes the proofs as well as some additional details on the experiments. All the simulation outputs are given in *Outputs.xlsx*.

References

Aneiros, G. and Vieu, P. (2014) Variable selection in infinite-dimensional problems. *Probability Letters*. **94**, 12–20.

Baíllo, A., Cuevas, A. and Cuesta-Albertos, J.A. (2011a) Supervised classification for a family of Gaussian functional models. *Scand. J. Stat.* **38**, 480–498.

Baíllo, A., Cuevas, A. and Fraiman, R. (2011b) Classification methods with functional data. In *Oxford Handbook of Functional Data Analysis*, pp. 259–297. F. Ferraty and Y. Romain, eds. Oxford University Press.

Bongiorno, E. G., and Goia, A. (2016) Classification methods for Hilbert data based on surrogate density. *Computational Statistics and Data Analysis*. **99**, 204–222.

Berlinet, A. and Thomas-Agnan, C. (2011) *Reproducing Kernel Hilbert Spaces in Probability and Statistics*. Springer.

Berrendero, J.R., Cuevas, A. and Torrecilla, J.L. (2016a) The mRMR variable selection method: a comparative study for functional data. *Journal of Statistical Computation and Simulation* **86**, 891–907.

Berrendero, J.R., Cuevas, A. and Torrecilla, J.L. (2016b) Variable selection in functional data analysis: a maxima-hunting proposal. *Statistica Sinica*, **26**, 619–638.

Cadre, B. (2013). Supervised classification of diffusion paths. *Math. Methods Statist.* **22**, 213-235.

Cuevas, A. (2014). A partial overview of the theory of statistics with functional data. *J. of Statist. Plann. Inf.* **147**, 1–23.

Cuesta-Albertos, J.A. and Dutta, S. (2016). On perfect classification for Gaussian processes. Manuscript, arxiv: 1602.04941v1.

Delaigle, A. and Hall, P. (2012a). Achieving near perfect classification for functional data. *J. R. Statist. Soc. B* **74**, 267–286.

Delaigle, A. and Hall, P. (2012b). Methodology and theory for partial least squares applied to functional data. *Ann. Statist.* **40**, 322–352.

Delaigle, A., Hall, P., and Bathia, N. (2012). Componentwise classification and clustering of functional data. *Biometrika* **99**, 299–313.

Devroye, L., Györfi, L. and Lugosi, G. (1996). *A probabilistic theory of pattern recognition.* Springer–Verlag.

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

Ferraty, F., Hall, P., and Vieu, P. (2010). Most-predictive design points for functional data predictors. *Biometrika* **97**, 807–824.

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

Hall, P. (2011) Principal component analysis for functional data. methodology, theory, and discussion. In *Oxford Handbook of Functional Data Analysis*, pp. 210–234. F. Ferraty and Y. Romain, eds. Oxford University Press.

Horváth, L. and Kokoszka, P. (2012). *Inference for Functional Data with Applications.* Springer, New York.

Izenman, A.J. (2008). *Modern Multivariate Statistical Techniques.* Springer, New York.

Kailath, T. (1971). RKHS approach to detection and estimation problems. I. Deterministic signals in Gaussian noise. *IEEE Trans. Information Theory.* **IT-17**, 530–549.

Lindquist, M.A. and McKeague, I.W.(2009). Logistic regression with Brownian-like predictors. *J. Am. Statist. Assoc.* **104** , 1575–1585.

Mörters, P. and Peres, Y. (2010). *Brownian Motion.* Cambridge University Press.

Mosler, K. and Mozharovskyi, P. (2015). Fast DD-classification of functional data, *Statistical Papers*, 1-35.

Müller, H.G. (2009). Functional data analysis. *Encyclopedia of Mathematics.* `https://www.encyclopediaofmath.org/index.php/Functional_data_analysis`

Parzen, E. (1961). An approach to time series analysis. *Ann. Math. Statist.* **32**, 951–989.

Parzen, E. (1962). Extraction and detection problems and reproducing kernel Hilbert space. *J. SIAM Control Ser. A.* **1**, 35–62.

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

Shepp, L.A. (1966). Radon-Nikodym derivatives of Gaussian measures. *Ann. Math. Statist.*. **37** 321–354.

Székely, G. J., Rizzo, M. L. and Bakirov, N. K. (2007). Measuring and testing dependence by correlation of distances. *Ann. Statist.* **35**, 2769–2794.

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.

Wang, J.-L., Chiou, J.-M. and Müller, H.-G. (2016). Review of functional data analysis. *Ann. Rev. Statist.* **3**, 257–295.