Journal of Machine Learning Research 21 (2020) 1-33 Submitted 3/18; Revised 12/19; Published 1/20

On Mahalanobis Distance in Functional Settings

José R. Berrendero JOSER.BERRENDERO@UAM.ES
Beatriz Bueno-Larraz BEATRIZ.BUENO@UAM.ES
Antonio Cuevas ANTONIO.CUEVAS@UAM.ES

Department of Mathematics
Universidad Autónoma de Madrid
Madrid, Spain

Editor: Lorenzo Rosasco

Abstract

Mahalanobis distance is a classical tool in multivariate analysis. We suggest here an extension of this concept to the case of functional data. More precisely, the proposed definition concerns those statistical problems where the sample data are real functions defined on a compact interval of the real line. The obvious difficulty for such a functional extension is the non-invertibility of the covariance operator in infinite-dimensional cases. Unlike other recent proposals, our definition is suggested and motivated in terms of the Reproducing Kernel Hilbert Space (RKHS) associated with the stochastic process that generates the data. The proposed distance is a true metric; it depends on a unique real smoothing parameter which is fully motivated in RKHS terms. Moreover, it shares some properties of its finite dimensional counterpart: it is invariant under isometries, it can be consistently estimated from the data and its sampling distribution is known under Gaussian models. An empirical study for two statistical applications, outliers detection and binary classification, is included. The results are quite competitive when compared to other recent proposals in the literature.

Keywords: Functional data, Mahalanobis distance, reproducing kernel Hilbert spaces, kernel methods in statistics, square root operator

1. Introduction

The classical (finite-dimensional) Mahalanobis distance is a well-known tool in multivariate analysis with multiple applications. Let X be a random variable taking values in $\mathbb{R}^d$ with non-singular covariance matrix Σ. In many practical situations it is required to measure the distance between two points $x, y \in \mathbb{R}^d$ when considered as two possible observations drawn from X. Clearly, the usual (square) Euclidean distance $\|x-y\|^2 = (x-y)'(x-y)$ is not a suitable choice since it disregards the standard deviations and the covariances of the components of X (given a column vector $x \in \mathbb{R}^d$ we denote by x' the transpose of x). The most popular alternative is perhaps the classical Mahalanobis distance, $M(x,y)$, defined as

$$M(x,y) \;=\; \big((x-y)'\Sigma^{-1}(x-y)\big)^{1/2}. \tag{1}$$

Very often the interest is focused on studying "how extreme" a point x is within the distribution of X; this is typically evaluated in terms of $M(x, m)$, where m stands for the vector of means of X.

This distance is named after the Indian statistician P. C. Mahalanobis (1893-1972) who first proposed and analyzed this concept (Mahalanobis, 1936) in the setting of Gaussian distributions. Nowadays, some popular applications of the Mahalanobis distance are: supervised classification, outlier detection (Rousseeuw and van Zomeren, 1990 and Penny, 1996), multivariate depth measures (Zuo and Serfling, 2000), hypothesis testing (through Hotelling's statistic, Rencher, 2012, Ch. 5) or goodness of fit (Mardia, 1975). This list of references is far from exhaustive.

1.1. The infinite-dimensional (functional) case. The covariance operator

Our framework here is Functional Data Analysis (FDA); see, e.g., Cuevas (2014) for an overview. In other words, we deal with statistical problems involving functional data. Thus our sample is made of trajectories $X_1(t), \ldots, X_n(t)$ in $L^2[0,1]$ drawn from a second order stochastic process $X(t),\ t \in [0,1]$ with $m(t) = \mathbb{E}(X(t))$. The inner product and the norm in $L^2[0,1]$ will be denoted by $\langle \cdot, \cdot \rangle$ and $\| \cdot \|$, respectively.

We will henceforth assume that the covariance function $K(s,t) = \mathrm{Cov}(X(s), X(t))$ is continuous. The function K defines a linear operator $\mathcal{K} : L^2[0,1] \to L^2[0,1]$, called covariance operator, given by

$$\mathcal{K}f(t) \ = \ \int_0^1 K(t,s) f(s) \mathrm{d}s. \tag{2}$$

Note that the covariance operator is the infinite-dimensional version of the covariance transformation $x \mapsto \Sigma x$ in $\mathbb{R}^d$. Still, its properties are different in some important aspects as we will discuss below. We will define a Mahalanobis distance associated with $\mathcal{K}$ assuming that $\mathcal{K}$ is injective, that is, all the eigenvalues are strictly positive.

For later use, it will be useful to note that the assumption of continuity for $K(s,t)$ entails $m \in L^2[0,1]$ and $\mathbb{E}\|X\|^2 < \infty$. Indeed, since $\int_0^1 K(t,t)dt = \mathbb{E}\int_0^1 (X(t) - m(t))^2 \mathrm{d}t < \infty$, it holds $\int_0^1 (X(t) - m(t))^2 \mathrm{d}t < \infty$ almost surely (a.s.). Hence, as we are assuming $X(t) \in L^2[0,1]$ a.s., we also have $m \in L^2[0,1]$. Now, $\mathbb{E}\int_0^1 (X(t) - m(t))^2 \mathrm{d}t = \int_0^1 \mathbb{E}(X(t) - m(t))^2 \mathrm{d}t < \infty$ implies

$$\mathbb{E}\|X\|^2 = \mathbb{E}\int_0^1 X^2(t)\mathrm{d}t = \int_0^1 \mathbb{E}X^2(t)\mathrm{d}t = \int_0^1 \mathbb{E}(X(t) - m(t))^2 \mathrm{d}t + \int_0^1 m^2(t)\mathrm{d}t < \infty.$$

It can be noted that this reasoning does not require in fact the continuity of K (boundedness would suffice) but, for simplicity, we will keep the continuity assumption as it is quite common in functional data analysis and not very restrictive. Besides, continuity will be needed below to prove, for example, Theorem 14 on asymptotic distribution.

1.2. Some basic notions on linear operators. Functions of operators

Let $(\boldsymbol{\mathcal{H}}, \langle \cdot, \cdot \rangle)$ be a real, separable, infinite-dimensional Hilbert space. In what follows, we will mostly consider the case $\boldsymbol{\mathcal{H}} = L^2[0,1]$ but the basic ideas and definitions can be given in general. Denote by $\mathcal{C}$ the space of self-adjoint bounded linear operators on $\mathcal{H}$; see, e.g., Gohberg et al. (2003) for definitions and background.

Let us recall that for any linear bounded operator $\mathcal{A} : \mathcal{H} \to \mathcal{H}$ in $\mathcal{C}$, the operator norm of $\mathcal{A}$ is defined by

$$\|\mathcal{A}\|_{op} = \sup_{\|x\|=1} \|\mathcal{A}x\| = \sup_{\|x\|=1} |\langle \mathcal{A}x, x\rangle|.$$

If $\mathcal{K} \in \mathcal{C}$ is compact then, from the Spectral Theorem, $\mathcal{K}$ can be expressed in the form

$$\mathcal{K}x = \sum_j \lambda_j \langle x, e_j\rangle e_j, \tag{3}$$

$\{e_i\}$ being an orthonormal system of eigenfunctions of $\mathcal{K}$ and $\{\lambda_i\}$ the corresponding sequence of eigenvalues. Let $\mathcal{C}_+$ be the subset of positive operators in $\mathcal{C}$, that is the subset of those operators $\mathcal{A}$ in $\mathcal{C}$ such that $\langle \mathcal{A}x, x\rangle \geq 0$, for all x. Note that every compact operator in $\mathcal{C}_+$ has necessarily non-negative eigenvalues.

If $\mathcal{K} \in \mathcal{C}$ is compact, the spectrum of $\mathcal{K}$ is defined by

$$\sigma(\mathcal{K}) = \{\lambda_j : \lambda_j \text{ is an eigenvalue of } \mathcal{K}\} \cup \{0\}.$$

If $\mathcal{K}$ has a spectral representation (3) and f is a real function defined on a real interval containing $\sigma(K)$, bounded on $\sigma(\mathcal{K})$, we define the operator $f(\mathcal{K})$ by

$$f(\mathcal{K})(x) = f(0)P_0x + \sum_j f(\lambda_j)\langle x, e_j\rangle e_j,$$

where P_0x denotes the orthogonal projection of x on the kernel of $\mathcal{K}$. For these operators we have

$$\|f(\mathcal{K})\|_{op} = \sup_j |f(\lambda_j)|,$$

see, e.g., Gohberg et al. (2003, Ch. 8).

Some examples of functions $f(\mathcal{K})$ to be considered below are as follows.

(i) If $\mathcal{K} \in \mathcal{C}_+$ is compact and injective, then the square root of $\mathcal{K}$, denoted $\mathcal{K}^{1/2}$, is defined by $\mathcal{K}^{1/2}(x) = \sum_j \lambda_j^{1/2}\langle x, e_j\rangle e_j$. Note that the name "square root" is justified since $\mathcal{K}^{1/2}\mathcal{K}^{1/2} = \mathcal{K}$. In fact, $\mathcal{A}^{1/2}$ can be defined as well, in a natural way, even if $\mathcal{A}$ is not compact; see, for example, Debnath and Mikiusinski (2005, p. 173) for a definition of the square root operator when $\mathcal{A}$ is a bounded positive operator.

(ii) We will also need to use operator functions $f(\mathcal{A})$ in some examples where $\mathcal{A}$ is not necessarily compact but it can be expressed in the form $\mathcal{A}x = \sum_j \mu_j\langle x, e_j\rangle e_j$, for some (bounded) sequence of constants $\{\mu_j\}$ and some orthonormal system $\{e_j\}$. Then, we can define $f(\mathcal{A})$ (if f is a function bounded on $\{\mu_j\}$) by $f(\mathcal{A})(x) = \sum_j f(\mu_j)\langle x, e_j\rangle e_j$

(iii) A particular case arises when $\mathcal{A} = \mathcal{K} + h\mathbb{I}$, where $\mathcal{K}$ is a compact, injective operator, $\mathcal{K} \in \mathcal{C}_+$, h is a real number and $\mathbb{I}$ is the identity operator. Since $\mathcal{K}$ is injective there is an orthonormal basis $\{e_j\}$ of $\mathcal{H}$, made of eigenvectors of $\mathcal{K}$. Thus, if $\{\lambda_j\}$ are the corresponding eigenvalues, we have, for all $x \in \mathcal{H}$, $(\mathcal{K} + h\mathbb{I})(x) = \sum_j(\lambda_j + h)\langle x, e_j\rangle e_j$ and $(\mathcal{K} + h\mathbb{I})^{1/2}(x) = \sum_j(\lambda_j + h)^{1/2}\langle .x, e_j\rangle e_j$. Similarly, the positive part of $\mathcal{K} + h\mathbb{I}$ is defined by $(\mathcal{K} + h\mathbb{I})_+(x) = \sum_j(\lambda_j + h)^+\langle x, e_j\rangle e_j$, where $f(t) := t^+ = \max\{t, 0\}$ is the usual positive part function, defined for real numbers.

1.3. On the difficulties of defining a functional Mahalanobis-type distance

The aim of this paper is to extend the notion of the multivariate (finite-dimensional) Mahalanobis distance (1) to the functional case when $x, y \in L^2[0,1]$. Clearly, in view of (1), the inverse $\mathcal{K}^{-1}$ of the functional operator $\mathcal{K}$ should play some role in this extension if we want to keep a close analogy with the multivariate case. Unfortunately, such a direct approach utterly fails since, typically, $\mathcal{K}$ is not invertible in general as an operator, in the sense that there is no linear continuous operator $\mathcal{K}^{-1}$ such that $\mathcal{K}^{-1}\mathcal{K} = \mathcal{K}\mathcal{K}^{-1} = \mathbb{I}$, the identity operator. This is a well-known consequence of the fact that, when $K(s,t)$ is a continuous function, the operator $\mathcal{K}$ defined in (2) is compact, that is, it takes bounded sets to relative compact sets, which entails the non-invertibility.

It is interesting to see this from the spectral point of view, starting again with the finite-dimensional case. Indeed, some elementary linear algebra yields the following representations for Σx, $\Sigma^{-1}x$ and the squared Mahalanobis distance $M(x,y)^2$, for $x, y \in \mathbb{R}^d$:

$$\Sigma x = \sum_{j=1}^d \lambda_j (e_j' x) e_j, \quad \Sigma^{-1} x = \sum_{j=1}^d \frac{1}{\lambda_j} (e_j' x) e_j, \; \Sigma^{-1/2} x = \sum_{j=1}^d \frac{1}{\sqrt{\lambda_j}} (e_j' x) e_j,$$

$$M(x,y)^2 = (x-y)'\Sigma^{-1/2}\Sigma^{-1/2}(x-y) = \sum_{j=1}^d \frac{1}{\lambda_j} ((x-y)'e_j)^2, \tag{4}$$

where $\lambda_1, \ldots, \lambda_d$ are the, strictly positive, eigenvalues of Σ and $\{e_1, \ldots, e_d\}$ the corresponding orthonormal basis of eigenvectors.

In the functional case, the expression (3) of $\mathcal{K}$ in terms of the Spectral Theorem, would suggest an expression for the inverse operator of type

$$\mathcal{K}^{-1} x = \sum_{j=1}^\infty \frac{1}{\lambda_j} \langle x, e_j \rangle e_j. \tag{5}$$

However, it is easy to check that $\int_0^1 \int_0^1 K(s,t)^2 \mathrm{d}s\mathrm{d}t = \sum_{j=1}^\infty \lambda_j^2$ so that, in particular, the sequence $\{\lambda_j\}$ converges to zero very quickly. As a consequence, there is no hope of keeping a direct analogy with (4) and the very reason for this is the non-invertibility of the covariance operator. See more details on this in the next subsection.

Nevertheless, as we will see, the "failed" definition (5) gives some clues in order to extend Mahalanobis' idea to the infinite-dimensional case.

1.4. Some previous proposals

Motivated by the finite-dimensional spectral version (4) of the Mahalanobis distance, Galeano et al. (2015) have proposed the following definition of the functional Mahalanobis distance.

$$d_{FM}^k(x,y) \;=\; \left(\sum_{j=1}^k \frac{\langle x-y, e_j \rangle^2}{\lambda_j} \right)^{1/2}, \tag{6}$$

which is well-defined for any $x, y \in L^2[0,1]$ at the expense of introducing a sort of tuning parameter $k \in \mathbb{N}$. We keep the notation d_{FM}^k used in Galeano et al. (2015). Let us note that

$d^k_{FM}(x,y)$ is a semi-distance, since it lacks the identifiability condition $d^k_{FH}(x,y) = 0 \Rightarrow x = y$. The applications of d^k_{FM} considered by these authors focus mainly on supervised classification. While this proposal is quite simple and natural, its theoretical motivation presents some difficulties. The most important one is the fact that the series

$$\sum_{j=1}^{\infty} \frac{\langle x-y, e_j\rangle^2}{\lambda_j} \tag{7}$$

is divergent, with probability one, whenever x and y are independent trajectories from a Gaussian process with covariance function K and the mean function m belongs to the Reproducing Kernel Hilbert Space associated with K (see, e.g. Lukić and Beder, 2001, Cor. 7.1); the same holds if y is replaced with the mean function $m \in L^2[0,1]$. So, (6) is defined in terms of the k-th partial sum of a divergent series. As a consequence, one may expect that the definition might be strongly influenced by the choice of k. As we will discuss below, in practice this effect is not noticed if x and y are replaced with certain appropriate smoothed approximations but, in that case, the smoothing procedure should be incorporated to the definition.

Another recent proposal is due to Ghiglietti et al. (2017). The idea is also to modify the template (7) to deal with the convergence issues. In this case, the suggested definition is

$$d_p(x,y) \;=\; \left(\int_0^{\infty} \sum_{j=1}^{\infty} \frac{\langle x-y, e_j\rangle^2}{e^{\lambda_j c}} g(c;p)\mathrm{d}c\right)^{1/2}, \tag{8}$$

where $p > 0$ and $g(c;p)$ is a weight function such that $g(0;p) = 1$, g is non-increasing and non-negative and $\int_0^\infty g(c;p)dc = p$. Moreover, for any $c > 0$, $g(c;p)$ is assumed to be non-decreasing in p with $\lim_{p\to\infty} g(c;p) = 1$. This definition does not suffer from any problem derived from divergence but, still, it depends on two smoothing functions: the exponential in the denominator of (8) and the weighting function $g(c;p)$. As pointed out also in Ghiglietti et al. (2017), a more convenient expression for (8) is given by the following weighted version of the template, formal definition (7),

$$d_p(x,y) \;=\; \left(\sum_{j=1}^{\infty} \frac{\langle x-y, e_j\rangle^2}{\lambda_j} h_j(p)\right)^{1/2},$$

where $h_j(p) = \int_0^\infty \lambda_j e^{-\lambda_j c} g(c;p)\mathrm{d}c$.

The applications of (8) offered in Ghiglietti et al. (2017) and Ghiglietti and Paganoni (2017) deal with hypotheses testing for two-sample problems of type $H_0 : m_1 = m_2$.

All in all, both proposals, Galeano et al. (2015) and Ghiglietti et al. (2017), are natural extensions to the functional case of the multivariate notion (1). Moreover, as suggested by the practical examples considered in both works, these options performed quite well in many cases. However, we believe that there is still some room to further explore the subject for the reasons we will explain below.

1.5. The organization on this work

In the next section some theory of RKHS and its connection with the Mahalanobis distance is introduced, together with the proposed definition. In Section 3 some properties of the proposed distance are presented and compared with those of the original multivariate definition. Then, a consistent estimator of the distance is analyzed in Section 4 and convergence rates are obtained. Finally, some numerical outputs corresponding to different statistical applications can be found in Section 5, along with some final remarks in Section 6.

2. A new definition of Mahalanobis distance for functional data

In this section we will propose a further definition of a Mahalanobis-type distance, denoted M_α (as it will depend on a tuning parameter α). Its most relevant features can be summarized as follows:

i) M_α is also inspired in the natural template (7). The serious convergence issues appearing in (7) are solved by smoothing.

ii) M_α depends on a single, real, easy to interpret smoothing parameter α whose choice is not critical, in the sense that the distance has some stability with respect to α. Hence, it is possible to think of a cross-validation or bootstrap-based choice of α. In particular, no auxiliary weight function is involved in the definition.

iii) $M_\alpha(x, y)$ is a true metric which is defined for any given pair x, y of functions in $L^2[0, 1]$. It shares some invariance properties with the finite-dimensional counterpart (1).

iv) If $m(t) = \mathbb{E}(X(t))$, the distribution of $M_\alpha(X, m)$ is explicitly known for Gaussian processes. In particular, $\mathbb{E}(M^2_\alpha(X, m))$ and $\mathrm{Var}(M^2_\alpha(X, m))$ have explicit, relatively simple expressions.

v) The distance $M_\alpha(X, m)$ can be consistently estimated from a random sample and convergence rates can be obtained under some mild conditions.

The main contribution of this paper is to show that the theory of Reproducing Kernel Hilbert Spaces (RKHS) provides a natural and useful framework in order to propose an extension of the Mahalanobis distance to the functional setting. We refer to Berlinet and Thomas-Agnan (2004), Appendix F in Janson (1997), Schölkopf and Smola (2002), Hsing and Eubank (2015) for background on RKHS's.

2.1. A few, very basic, ideas on RKHS's

There are several equivalent ways of approaching RKHS's. For our purposes, it is enough to recall that the RKHS associated with the covariance function $K = K(s, t)$ is defined by

$$\mathcal{H}(K) := \mathcal{K}^{1/2}(L^2[0, 1]) = \{f \in L^2[0, 1] : \sum_{j=1}^{\infty} \frac{\langle f, e_j\rangle^2}{\lambda_j} < \infty\}, \qquad (9)$$

(recall that we assume $\lambda_j > 0$ for all j) and the corresponding inner product and norm are given by

$$\langle f, g\rangle_K = \sum_{j=1}^{\infty} \frac{\langle f, e_j\rangle\langle g, e_j\rangle}{\lambda_j}, \quad \|f\|_K = \left(\sum_{j=1}^{\infty} \frac{\langle f, e_j\rangle^2}{\lambda_j}\right)^{1/2}. \tag{10}$$

It can be seen that all finite linear combinations of type $g(s) = \sum_{j=1}^{p} a_j K(s, t_j)$ are in $\mathcal{H}(K)$; in fact, $\mathcal{H}(K)$ could be defined as the family of such finite linear combinations plus the pointwise-limit of all Cauchy sequences made of functions of type $g(s) = \sum_{j=1}^{p} a_j K(s, t_j)$.

Note that, in particular, if $g(\cdot) = K(\cdot, t)$, $\langle g, e_j\rangle = \lambda_j e_j(t)$, so that $\langle f, g\rangle_K = f(t)$. This is the so-called "reproducing property" which accounts for the name of these spaces.

Also, $\sup_t \|\langle f_n, K(\cdot,t)\rangle_K\| = \sup_t |f_n(t)| \le \|f_n\|_K \sup_t \|K(\cdot,t)\|_K = \|f_n\|_K \sup_t K(t,t)^{1/2}$. So, $\|f_n\|_K \to 0$ implies $f_n(t) \to 0$ uniformly on t since K is assumed to be continuous.

2.2. The proposed definition

First note that in the finite dimensional case, the RKHS associated with a non-singular covariance matrix is just $\mathbb{R}^d$ and the square norm is $\|x\|_\Sigma^2 = \sum_{j=1}^{p} \frac{(x'e_j)^2}{\lambda_j}$. Therefore, the spectral expression (5) of the square Mahalanobis distance in the finite-dimensional case can be formulated in terms of the corresponding RKHS square norm, that is,

$$M^2(x, y) = \|x - y\|_\Sigma^2.$$

As discussed above, the difficulty to extend this idea to the functional case lies in the fact that in general, the RKHS space $\mathcal{H}(K)$ is much smaller than the ambient space $L^2[0,1]$ and, more importantly, the trajectories of a second-order stochastic process with continuous covariance function $K = K(s,t)$ do not belong, with probability one, to $\mathcal{H}(K)$.

This observation suggests us the simple strategy we will follow here: given two functions $x, y \in L^2[0,1]$, just approximate them by two other functions $x_\alpha, y_\alpha \in \mathcal{H}(K)$ and calculate the distance $\|x_\alpha - y_\alpha\|_K$. It only remains to decide how to obtain the RKHS approximations x_α and y_α. One could think of taking x_α as the "closest" function to x in $\mathcal{H}(K)$ but this approach also fails since $\mathcal{H}(K)$ is dense in $L^2[0,1]$ whenever all λ_j are strictly greater than zero (Remark 4.9 of Cucker and Zhou, 2007). Thus, every function $x \in L^2[0,1]$ can be arbitrarily well approximated by functions in $\mathcal{H}(K)$ in such a way that for any given function $x \in L^2[0,1]$ there is no function in $\mathcal{H}(K)$ "closest" to x. In other words, there is no projection of x over $\mathcal{H}(K)$ since this space is not closed.

The situation is reminiscent to that arising in the penalization methods in nonparametric functional estimation: we might look instead for the function x_α closest to x among those that are not "too rough". In more formal terms: let us fix a penalization parameter $\alpha > 0$. Given any $x \in L^2[0,1]$, define

$$x_\alpha = \operatorname*{argmin}_{f \in \mathcal{H}(K)} \|x - f\|^2 + \alpha\|f\|_K^2. \tag{11}$$

Whereas the term $\|x - f\|^2$ controls closeness to x, the term $\alpha\|f\|_K^2$ accounts for the roughness of the approximation. As we will see below, the "penalized projection" x_α is well-defined. In fact it admits a relatively simple closed form. Finally, the definition we propose for the functional α-Mahalanobis distance is as follows.

Definition 1 *Let $\mathcal{K}$ be the covariance operator associated with the continuous covariance function $K = K(s,t) = Cov(X(s), X(t))$ of a process $X = \{X(t),\ t \in [0,1]\}$ with trajectories in $L^2[0,1]$. Assume that $\mathcal{K}$ is injective. Thus, all the eigenvalues of $\mathcal{K}$ are strictly positive. Given a constant $\alpha > 0$ we define the α-Mahalanobis distance between two functions x and y in $L^2[0,1]$ by*

$$M_\alpha(x,y) = \|x_\alpha - y_\alpha\|_K, \tag{12}$$

where x_α is defined as the unique solution of (11) *(see Proposition 2 below) and $\|\cdot\|_K$ is the RKHS norm defined in* (10)*.*

Let us note that the original Mahalanobis distance as well as our adapted functional version (12) are both intrinsically dependent on the covariance operator $\mathcal{K}$. Therefore, (12) is naturally aimed at measuring either the "statistical distance" between two observations (trajectories) $x = x(t)$ and $y = y(t)$ of a process with covariance operator $\mathcal{K}$ or, more commonly, the distance between a trajectory $x = x(t)$ and the mean function $m = m(t)$. In the later case, the notation $M_\alpha(x, m)$ will be used.

As mentioned, given a realization x of the stochastic process, we have relatively simple expressions for both the smoothed trajectory x_α and the proposed distance. Such expressions are given in the following result.

Proposition 2 *In the setting of Definition 1 the smoothed trajectories x_α defined in (11) satisfy the following basic properties:*

(a) Let $\mathbb{I}$ be the identity operator on $L^2[0,1]$. Then, $\mathcal{K} + \alpha\mathbb{I}$ is invertible and

$$x_\alpha = (\mathcal{K} + \alpha\mathbb{I})^{-1}\mathcal{K}x = \sum_{j=1}^{\infty} \frac{\lambda_j}{\lambda_j + \alpha}\langle x, e_j\rangle\, e_j, \tag{13}$$

where λ_j, $j = 1, 2, \cdots$ are the eigenvalues of $\mathcal{K}$ and e_j stands for the unit eigenfunction of $\mathcal{K}$ corresponding to λ_j.

(b) Denoting as $\mathcal{K}^{1/2}$ the square root operator, the norm of x_α in $\mathcal{H}(K)$ satisfies

$$\|x_\alpha\|_K^2 = \sum_{j=1}^{\infty} \frac{\lambda_j}{(\lambda_j + \alpha)^2}\ \langle x, e_j\rangle^2 = \|\mathcal{K}^{1/2}(\mathcal{K} + \alpha\mathbb{I})^{-1}x\|^2, \tag{14}$$

and, therefore, $M_\alpha(x,y)^2 = \sum_{j=1}^{\infty} \frac{\lambda_j}{(\lambda_j+\alpha)^2}\ \langle x - y, e_j\rangle^2$.

Proof (a) From Spectral Theorem for compact and self-adjoint operators (see, e.g., Theorem 2 of Chapter 2 in Cucker and Smale, 2001), $\mathcal{K}x = \sum_j \lambda_j\langle x, e_j\rangle e_j$. Thus,

$$(\mathcal{K} + \alpha\mathbb{I})\, x = \sum_{j=1}^{\infty}(\lambda_j + \alpha)\langle y, e_j\rangle e_j \text{ and } (\mathcal{K} + \alpha\mathbb{I})^{-1}\, y = \sum_{j=1}^{\infty} \frac{1}{\lambda_j + \alpha}\langle y, e_j\rangle e_j. \tag{15}$$

The expression $x_\alpha = (\mathcal{K} + \alpha\mathbb{I})^{-1}\mathcal{K}x$ for the solution of the minimization problem (11) follows directly from Theorem 8.4 (i) in Cucker and Zhou (2007, p. 136); note that $\mathcal{K}^{-1/2}(f)$

is well defined for $f \in \mathcal{H}(K)$, and $\|f\|_K = \|\mathcal{K}^{-1/2}(f)\|$. Now, expression (13) readily follows by replacing y with $\mathcal{K}x$ in the second equation of (15).

(b) From (10) and (13),

$$\|x_\alpha\|_K^2 = \sum_{j=1}^{\infty} \frac{\lambda_j}{(\lambda_j + \alpha)^2} \langle x, e_j \rangle^2.$$

Moreover, from the Spectral Theorem $\mathcal{K}^{1/2}(x) = \sum_{i=1}^{\infty} \sqrt{\lambda_i} \langle x, e_i \rangle \, e_i$. Then, using (15)

$$\mathcal{K}^{1/2}(\mathcal{K} + \alpha \mathbb{I})^{-1} x = \sum_{j=1}^{\infty} \frac{\sqrt{\lambda_j}}{\lambda_j + \alpha} \langle x, e_j \rangle \, e_j.$$

By Parseval's identity we get $\|\mathcal{K}^{1/2}(\mathcal{K} + \alpha \mathbb{I})^{-1} x\|^2 = \|x_\alpha\|_K^2$, as desired.

■

Corollary 3 *The expression M_α given in* (12) *defines a metric in $L^2[0,1]$.*

Proof This result is a direct consequence of Proposition 2. Indeed, from expression (13), the transformation $x \mapsto x_\alpha$ form $L^2[0,1]$ to $\mathcal{H}(K)$ in injective (since the coefficients $\langle x, e_j \rangle$ completely determine x). Now, the result follows from the fact that $\| \cdot \|_K$ is a norm. ■

Remark 4 *The expression $x_\alpha = (\mathcal{K} + \alpha \mathbb{I})^{-1} \mathcal{K} x$ obtained in the first part of Proposition 2 has an interesting intuitive meaning: the transformation $x \mapsto \mathcal{K}x$ takes first the function $x \in L^2[0,1]$ to the space $\mathcal{H}(K)$, made of much nicer functions, with Fourier coefficients $\langle x, e_j \rangle$ converging quickly to zero, since we must have $\sum_{j=1}^{\infty} \langle x, e_j \rangle^2 / \lambda_j < \infty$; see* (10). *Then, after this "smoothing step", we perform an "approximation step" by applying the inverse operator $(\mathcal{K} + \alpha \mathbb{I})^{-1}$, in order the get, as a final output, a function x_α that is both, close to x and smoother than x.*

On the other hand, the expressions (11) *and* (13) *clearly relate our approach with the general methodology of Tikhonov regularization; see more details on this in Section 6 below.*

3. Some properties of the functional Mahalanobis distance

In this section we analyze in detail and prove some of the features of M_α we have anticipated above. We assume throughout that $X = X(t)$ fulfils the conditions indicated in the statement of Definition 1.

3.1. Invariance

In the finite dimensional case, one appealing property of the Mahalanobis distance is the fact that it does not change if we apply a non-singular linear transformation to the data. Then, the invariance for a large class of linear operators appears also as a desirable property for any extension of the Mahalanobis distance to the functional case. Here, we will prove invariance with respect to operators preserving the norm. We recall that a bounded linear operator L is an isometry if it maps $L^2[0,1]$ to $L^2[0,1]$ and $\|f\| = \|Lf\|$. In this case, it holds $L^*L = \mathbb{I}$, where L^* stands for the adjoint of L.

Theorem 5 *Let L be an isometry on $L^2[0,1]$. Then, $M_\alpha(x,y) = M_\alpha(Lx, Ly)$ for all $\alpha > 0$, $x, y \in L^2[0,1]$, where M_α was defined in* (12).

Proof Let $\mathcal{K}_L$ be the covariance operator of the process LX. The first step of the proof is to show that $\mathcal{K}_L = L\mathcal{K}L^*$. It is enough to prove that for all $f, g \in L^2[0,1]$, it holds $\langle \mathcal{K}_L f, g\rangle = \langle L\mathcal{K}L^* f, g\rangle$. Observe that

$$\langle \mathcal{K}_L f, g\rangle = \int_0^1 \mathcal{K}_L f(t) g(t) \mathrm{d}t = \int_0^1 \int_0^1 \mathbb{E}[(LX(s) - Lm(s))(LX(t) - Lm(t))] f(s) g(t) \mathrm{d}s \mathrm{d}t.$$

Then, using Fubini's theorem and the definition of the adjoint operator,

$$\langle \mathcal{K}_L f, g\rangle = \mathbb{E}\big[\langle L(X-m), f\rangle \cdot \langle L(X-m), g\rangle\big] = \mathbb{E}\big[\langle X - m, L^* f\rangle \cdot \langle X - m, L^* g\rangle\big].$$

Analogously, we also have

$$\langle L\mathcal{K}L^* f, g\rangle = \langle \mathcal{K}L^* f, L^* g\rangle = \mathbb{E}\big[\langle X - m, L^* f\rangle \cdot \langle X - m, L^* g\rangle\big].$$

From the last two equations we conclude $\mathcal{K}_L = L\mathcal{K}L^*$.

The second step of the proof is to observe that the eigenvalues λ_j of $\mathcal{K}_L$ are the same as those of $\mathcal{K}$, and the unit eigenfunction u_j of $\mathcal{K}_L$ for the eigenvalue λ_j is given by $v_j = Le_j$, where e_j is the unit eigenfunction corresponding to λ_j. Indeed, using $L^*L = \mathbb{I}$ we have

$$\mathcal{K}_L v_j = L\mathcal{K}L^* v_j = L\mathcal{K}L^* L e_j = \lambda_j L e_j = \lambda_j v_j, \quad j = 1, 2, \ldots$$

Then, by (14) and using that L is an isometry,

$$\begin{aligned} M_\alpha(Lx, Ly)^2 =& \|(Lx - Ly)_\alpha\|^2_{\mathcal{K}_L} = \sum_{j=1}^\infty \frac{\lambda_j}{(\lambda_j + \alpha)^2} \ \langle Lx - Ly, Le_j\rangle^2 \\ =& \sum_{j=1}^\infty \frac{\lambda_j}{(\lambda_j+\alpha)^2} \ \langle x - y, e_j\rangle^2 = M_\alpha(x,y)^2. \end{aligned}$$

■

The family of isometries on $L^2[0,1]$ contains some interesting examples such as symmetries and changes between orthonormal bases. Thus, this distance does not depend on the basis on which the data are represented.

3.1.1. On the effect of regularization on invariance properties.

We have shown that our proposed functional Mahalanobis distance is invariant under isometries. However, unlike the finite-dimensional case, it is not invariant for bounded operators in general. The reason for this different behavior is the need for regularization (which precludes general invariance) together with the fact that the class of bounded linear operators is too broad in the case of infinite-dimensional Hilbert spaces.

To shed some light on these issues, let us consider the Mahalanobis distance changes under a particular type of bounded operators, namely, those positive compact self-adjoint bounded operators that commute with $\mathcal{K}$. It turns out that these are precisely the operators that can be diagonalized simultaneously with $\mathcal{K}$ (see Theorem 1.1 in Gohberg et al., 2003, p. 243). This means that they have the same eigenfunctions as $\mathcal{K}$ with different positive eigenvalues. In the finite-dimensional case these transformations would correspond to non-uniform changes of scale along the principal directions (that is, eigenvectors).

To be more precise, let L be a positive semidefinite compact self-adjoint bounded operator that commutes with $\mathcal{K}$. Then, as in Theorem 1 we have that $\mathcal{K}_L = L\mathcal{K}L$, where $\mathcal{K}_L$ is the covariance operator of the process LX. Now, if $\mu_1 \geq \mu_2 \geq \cdots$ are the (non-negative) eigenvalues of L with unit eigenfunctions $e_1, e_2, \ldots$ (which are the same as those of $\mathcal{K}$), it holds $j = 1, 2, \ldots$ $\mathcal{K}_L e_j = L\mathcal{K}Le_j = \lambda_j \mu_j^2 e_j$. That is, e_j is a unit eigenfunction of $\mathcal{K}_L$ and its corresponding eigenvalue is $\lambda_j \mu_j^2$. Using Proposition 1(b) together with Spectral Theorem for compact self-adjoint operators we readily obtain

$$M_\alpha(Lx, Lm)^2 = \sum_{j=1}^{\infty} \frac{\lambda_j \mu_j^4}{(\lambda_j \mu_j^2 + \alpha)^2} \langle x - m, e_j \rangle^2,$$

It is very easy to see that, for all $j = 1, 2, \ldots$

$$\frac{\lambda_j \mu_j^4}{(\lambda_j \mu_j^2 + \alpha)^2} < \frac{\lambda_j}{(\lambda_j + \alpha)^2} \Leftrightarrow \mu_j < 1.$$

In particular, if L is such that it shrinks x along all the principal directions, then $M_\alpha(Lx, Lm) \leq M_\alpha(x, m)$, for all $\alpha > 0$.

The above discussion shows how regularization (associated with values $\alpha > 0$) is an obstacle for invariance: observe that in the finite-dimensional case, where the sum is finite and $\alpha = 0$, Mahalanobis distance would remain unchanged.

3.2. Distribution for Gaussian processes

It is a well known fact that the squared Mahalanobis distance to the mean for finite dimensional Gaussian data has a χ^2 distribution with d degrees of freedom, where d is the dimension of the data. In the functional case, the distribution of $M_\alpha(X, m)^2$ for a Gaussian process X equals that of an infinite linear combination of independent χ_1^2 random variables. We prove this fact in the following result and its corollary, and also give explicit expressions for the expectation and the variance of $M_\alpha(X, m)^2$.

Proposition 6 *Let $\{X(t),\ t \in [0,1]\}$ be an L^2-Gaussian process with continuous covariance function K such that the corresponding covariance operator $\mathcal{K}$ is injective. Let $\lambda_1, \lambda_2, \cdots$ be the eigenvalues of $\mathcal{K}$ and let $e_1, e_2, \ldots$ be the corresponding unit eigenfunctions.*

(a) *The squared Mahalanobis distance satisfies, for $X = X(t)$ and for all $f \in L^2[0,1]$,*

$$M_\alpha(X,f)^2 = \|(X-f)_\alpha\|_K^2 = \sum_{j=1}^{\infty} \beta_j Y_j, \tag{16}$$

where $\beta_j = \lambda_j^2(\lambda_j+\alpha)^{-2}$ and Y_j, $j = 1, 2, \cdots$, are non-central $\chi_1^2(\gamma_j)$ random variables with non-centrality parameter $\gamma_j = \mu_j^2/\lambda_j$, with $\mu_j := \langle m - f, e_j\rangle$.

(b) *We have for $X = X(t)$ and for all $f \in L^2[0,1]$,*

$$\mathbb{E}\left(M_\alpha(X,f)^2\right) = \sum_{j=1}^{\infty} \frac{\lambda_j^2}{(\lambda_j+\alpha)^2}\left(1 + \frac{\mu_j^2}{\lambda_j}\right),$$

and

$$\operatorname{Var}\left(M_\alpha(X,f)^2\right) = 2\sum_{j=1}^{\infty} \frac{\lambda_j^4}{(\lambda_j+\alpha)^4}\left(1 + \frac{2\mu_j^2}{\lambda_j}\right).$$

Proof (a) Using (14), $\|(X-f)_\alpha\|_K^2 = \sum_{j=1}^\infty \beta_j Y_j$, where $\beta_j = \lambda_j^2(\lambda_j+\alpha)^{-2}$ and $Y_j = \lambda_j^{-1}\langle X-f, e_j\rangle^2$. Since the process is Gaussian the variables $\lambda_j^{-1/2}\langle X-f, e_j\rangle$ are independent with normal distribution, mean $\lambda_j^{-1/2}\mu_j$ and variance 1 (see Ash and Gardner, 1975, p. 40). The result follows.

(b) It is easy to see that the partial sums in (16) form a sub-martingale with respect to the natural filtration $\sigma(Y_1,\ldots,Y_N)$. Indeed,

$$\mathbb{E}\left(\sum_{j=1}^{N+1} \beta_j Y_j \,\middle|\, Y_1,\ldots,Y_N\right) = \beta_{N+1}\mathbb{E}(Y_{N+1}) + \sum_{j=1}^{N} \beta_j Y_j \geq \sum_{j=1}^{N} \beta_j Y_j.$$

Moreover, if $\bar{\lambda} := \sup_j \lambda_j$, which is always finite,

$$\sup_N \mathbb{E}\Big(\sum_{j=1}^{N+1} \beta_j Y_j\Big) = \sum_{j=1}^{\infty} \frac{\lambda_j(\lambda_j+\mu_j^2)}{(\lambda_j+\alpha)^2} \leq \frac{\bar{\lambda}}{\alpha^2}\Big(\sum_{j=1}^{\infty}\lambda_j + \sum_{j=1}^{\infty}\mu_j^2\Big) < \infty,$$

because $m \in L^2[0,1]$ and $\sum_{j=1}^\infty \lambda_j = \int_0^1 K(t,t)\mathrm{d}t < \infty$ (see e.g. Cucker and Smale, 2001, Corollary 3, p. 34). Now, Doob's convergence Theorem implies $\sum_{j=1}^N \beta_j Y_j \to \sum_{j=1}^\infty \beta_j Y_j$ a.s. as $N \to \infty$, and Monotone Convergence Theorem yields the expression for the expectation of $M_\alpha(X,f)^2$.

The proof for the variance is fairly similar. Using Jensen's inequality, we deduce

$$\mathbb{E}\left[\left(\sum_{j=1}^{N+1} \beta_j\left(Y_j - \mathbb{E}Y_j\right)\right)^2 \,\middle|\, Y_1,\ldots,Y_N\right] \geq \left(\sum_{j=1}^{N} \beta_j\left(Y_j - \mathbb{E}Y_j\right)\right)^2.$$

Moreover, since the variables Y_j are independent,

$$\begin{aligned}\sup_N \mathbb{E}\left(\sum_{j=1}^{N}\beta_j\left(Y_j-\mathbb{E}Y_j\right)\right)^2 =&\sum_{j=1}^{\infty}\beta_j^2\mathrm{Var}(Y_j)=2\sum_{j=1}^{\infty}\frac{\lambda_j^3(\lambda_j+2\mu_j^2)}{(\lambda_j+\alpha)^4}\\ \leq&\frac{2\bar{\lambda}^3}{\alpha^4}\Big(\sum_{j=1}^{\infty}\lambda_j+2\sum_{j=1}^{\infty}\mu_j^2\Big)<\infty.\end{aligned}$$

Then, $\left(\sum_{j=1}^{N+1}\beta_j\left(Y_j-\mathbb{E}Y_j\right)\right)^2\to\left(\sum_{j=1}^{\infty}\beta_j\left(Y_j-\mathbb{E}Y_j\right)\right)^2$ a.s., as $N\to\infty$, and using Monotone Convergence Theorem,

$$\mathrm{Var}\left(M_\alpha(X,f)^2\right)=\lim_{N\to\infty}\mathrm{Var}\Big(\sum_{j=1}^{N}\beta_jY_j\Big)=2\sum_{j=1}^{\infty}\frac{\lambda_j^4}{(\lambda_j+\alpha)^4}\left(1+\frac{2\mu_j^2}{\lambda_j}\right).$$

■

When we compute the squared Mahalanobis distance from X to the mean the expressions above simplify because $\mu_j=0$ for each j, and then we have the following corollary.

Corollary 7 *Under the same assumptions of Proposition 6, if m denotes the mean function, we have $M_\alpha(X,m)^2=\sum_{j=1}^{\infty}\beta_jY_j$, where $\beta_j=\lambda_j^2(\lambda_j+\alpha)^{-2}$ and $Y_1,Y_2,\dots$ are independent χ_1^2 random variables. Moreover, $\mathbb{E}\left(M_\alpha(X,m)^2\right)=\sum_{j=1}^{\infty}\lambda_j^2(\lambda_j+\alpha)^{-2}$ and $\mathrm{Var}\left(M_\alpha(X,m)^2\right)=2\sum_{j=1}^{\infty}\lambda_j^4(\lambda_j+\alpha)^{-4}$.*

3.3. Stability with respect to the regularization parameter

Our definition of distance depends on a regularization parameter $\alpha>0$. In this subsection we prove the continuity of M_α with respect to this tuning parameter α. The proof of the main result requires the following auxiliary lemma, which has been adapted from Corollary 8.3 in Gohberg et al. (2003), p. 71.

Lemma 8 *Let $A_j:\mathcal{H}\to\mathcal{H}$, $j=1,2,\dots$, be a sequence of bounded invertible operators on a Hilbert space $\mathcal{H}$ which converges in norm $\|\cdot\|_{op}$ to another operator A, and such that $\sup_j\|A_j^{-1}\|_{op}<\infty$. Then A is also invertible, and $\|A_j^{-1}-A^{-1}\|_{op}\to 0$, as $j\to\infty$.*

We will apply the preceding lemma in the proof of the following result.

Proposition 9 *Let α_j be a sequence of positive real numbers such that $\alpha_j\to\alpha>0$, as $j\to\infty$. Then, $\|X_{\alpha_j}\|_K\to\|X_\alpha\|_K$ a.s. as $j\to\infty$.*

Proof Note that by Proposition 2(b), Eq (14), we have (by the sub-multiplicative property of the norm)

$$\begin{aligned}\Big|\|X_{\alpha_j}\|_K-\|X_\alpha\|_K\Big|\leq&\|\mathcal{K}^{1/2}(\mathcal{K}+\alpha_j\mathbb{I})^{-1}X-\mathcal{K}^{1/2}(\mathcal{K}+\alpha\mathbb{I})^{-1}X\|\\ \leq&\|\mathcal{K}^{1/2}\|_{op}\ \|(\mathcal{K}+\alpha_j\mathbb{I})^{-1}-(\mathcal{K}+\alpha\mathbb{I})^{-1}\|_{op}\ \|X\|.\end{aligned}$$

But $\|(\mathcal{K}+\alpha_j\mathbb{I})-(\mathcal{K}+\alpha\mathbb{I})\|_{op}=|\alpha_j-\alpha|\to 0,\quad$ as $j\to\infty$, and $\sup_j\|(\mathcal{K}+\alpha_j\mathbb{I})^{-1}\|_{op}\leq\sup_j\alpha_j^{-1}<\infty$ (see Gohberg et al., 2003, (1.14), p. 228). Therefore, $\|(\mathcal{K}+\alpha_j\mathbb{I})^{-1}-(\mathcal{K}+\alpha\mathbb{I})^{-1}\|_{op}\to 0$, as $j\to\infty$, by Lemma 8. ■

Observe that Proposition 9 implies the pointwise convergence of the sequence of distribution functions of $M_{\alpha_j}(X,m)$ to that of $M_\alpha(X,m)$. This fact in turn implies the pointwise convergence of the corresponding quantile functions.

4. A consistent estimator of the functional Mahalanobis distance

Again, throughout this section, we will assume that the process $\{X(t),\ t\in[0,1]\}$, with trajectories in $L^2[0,1]$, has a continuous covariance function $K=K(s,t)$ and that the covariance operator $\mathcal{K}$ is injective (that is, all the eigenvalues are strictly positive). As mentioned in the introduction, the continuity of K entails $\mathbb{E}\|X\|^2<\infty$ and $m\in L^2[0,1]$, where $m(t)=\mathbb{E}(X(t))$.

The problem considered in this section is as follows: given a sample $X_1(t),\ldots,X_n(t)$ of realizations of the stochastic process $X(t)$, we want to estimate the Mahalanobis distance between any trajectory of the process X and the mean function m in a consistent way. Let $\bar{X}(t)=n^{-1}\sum_{i=1}^n X_i(t)$ be the sample mean and let

$$\widehat{K}(s,t)=\frac{1}{n}\sum_{i=1}^n(X_i(s)-\bar{X}(s))(X_i(t)-\bar{X}(t))$$

be the sample covariance function. The function $\widehat{K}$ defines the sample covariance operator $\widehat{\mathcal{K}}f(\cdot)=\int_0^1\widehat{K}(\cdot,t)f(t)\mathrm{d}t$.

Define the following estimator for $M_\alpha(X,m)$:

$$\widehat{M_{\alpha,n}}(X,\bar{X}):=\|\widehat{X}_\alpha-\bar{X}_\alpha\|_{\widehat{K}},\tag{17}$$

where $\widehat{X}_\alpha=(\widehat{\mathcal{K}}+\alpha\mathbb{I})^{-1}\widehat{\mathcal{K}}X$ and $\bar{X}_\alpha=(\widehat{\mathcal{K}}+\alpha\mathbb{I})^{-1}\widehat{\mathcal{K}}\bar{X}$.

The following lemma collects, for posterior reference, some asymptotic results on the estimation of the mean function and covariance operator. Although similar results can be found at several places in the literature (see, e.g., Hsing and Eubank, 2015, Th. 8.1.2) we include them here for the sake of completeness.

Lemma 10 *Under the conditions on $\{X(t),\ t\in[0,1]\}$ stated at the beginning of this section, we have $\|\bar{X}-m\|\to 0$, $\|\widehat{K}-K\|_{L^2([0,1]\times[0,1])}\to 0$, and $\|\widehat{\mathcal{K}}-\mathcal{K}\|_{op}\to 0$, a.s., as $n\to\infty$.*

Proof Mourier's Strong Law of Large Numbers (see e.g. Theorem 4.5.2 in Laha and Rohatgi, 1979, p. 452) implies directly $\|\bar{X}-m\|\to 0$ since $(\mathbb{E}\|X\|)^2\leq\mathbb{E}\|X\|^2<\infty$ and $L^2[0,1]$ is a separable Banach space.

Consider the process $Z(s,t)=X(s)X(t)$. Then, $Z\in L^2([0,1]\times[0,1])$ and this is also a separable Banach space. Therefore, if $Z_i(s,t)=X_i(s)X_i(t)$, $\bar{Z}=n^{-1}\sum_{i=1}^n Z_i(s,t)$, and $m_z(s,t)=\mathbb{E}[X(s)X(t)]$, using again Mourier's SLLN we have

$$\|\bar{Z}-m_z\|_{L^2([0,1]\times[0,1])}\to 0,\quad\text{a.s.,}\quad n\to\infty,$$

and also, since $\widehat{K}(s,t) = \bar{Z}(s,t) - \bar{X}(s)\bar{X}(t)$, $\|\widehat{\mathcal{K}} - \mathcal{K}\|_{HS} \to 0$ a.s., where $\|\widehat{\mathcal{K}} - \mathcal{K}\|_{HS} = \|\widehat{K} - K\|_{L^2([0,1]\times[0,1])}$ stands for the Hilbert-Schmidt norm of the operator $\widehat{\mathcal{K}} - \mathcal{K}$.

Finally, for any $x \in L^2[0,1]$,

$$\|(\widehat{\mathcal{K}} - \mathcal{K})x\|^2 \ = \int_0^1 \langle \widehat{K}(t,\cdot) - K(t,\cdot), x\rangle^2 \mathrm{d}t \ \leq \ \|x\|^2 \ \|\widehat{K} - K\|^2_{L^2([0,1]\times[0,1])}.$$

Thus, in particular, the operator norm is smaller than the Hilbert-Schmidt norm and we have $\|\widehat{\mathcal{K}} - \mathcal{K}\|_{op} \leq \|\widehat{K} - K\|_{L^2([0,1]\times[0,1])} \to 0$ a.s. as $n \to \infty$. ■

Now, before going on with the proof of the consistency and convergence rates we need some, non-trivial, auxiliary results on the square root operator $\mathcal{K}^{1/2}$ which appears in the definition (9) of the RKHS, and will appear also in the proofs below.

4.1. On the square root operator

We refer here to the notation and definitions included in Subsection 1.2 above. In particular, recall that we denote by $\boldsymbol{\mathcal{C}}$ the family of bounder linear self-adjoint operators on $(\boldsymbol{\mathcal{H}}, \langle\cdot,\cdot\rangle)$. Given two operators $\mathcal{A}, \mathcal{B} \in \boldsymbol{\mathcal{C}}$ we will say that $\mathcal{A} \preceq \mathcal{B}$ if and only if $\mathcal{B} - \mathcal{A}$ is non-negative, that is $\mathcal{B} - \mathcal{A} \in \boldsymbol{\mathcal{C}}_+$.

The following proposition collects some, mostly elementary, operator inequalities which will be helpful below in the proofs of the results on consistency and convergence rates. The inequality in part (e) will be especially useful in what follows.

Proposition 11 *Let $\mathcal{K}_0, \mathcal{K} \in \boldsymbol{\mathcal{C}}_+$ be compact operators. Suppose that $\mathcal{K}$ is injective. Denote $h_0 = \|\mathcal{K}_0 - \mathcal{K}\|_{op}$. We have,*

(a) $(\mathcal{K} - \epsilon\mathbb{I})_+ \preceq \mathcal{K} \preceq (\mathcal{K} + \epsilon\mathbb{I})$, for all $\epsilon > 0$.

(b) $(\mathcal{K} - h_0\mathbb{I})_+ \preceq \mathcal{K}_0 \preceq (\mathcal{K} + h_0\mathbb{I})$.

(c) $(\mathcal{K} - h_0\mathbb{I})_+^{1/2} \preceq \mathcal{K}^{1/2} \preceq (\mathcal{K} + h_0\mathbb{I})^{1/2}$. Also, $(\mathcal{K} - h_0\mathbb{I})_+^{1/2} \preceq \mathcal{K}_0^{1/2} \preceq (\mathcal{K} + h_0\mathbb{I})^{1/2}$.

(d) Let $\mathcal{A}, \mathcal{B}, \mathcal{M} \in \boldsymbol{\mathcal{C}}_+$ be operators such that $\mathcal{A} \preceq \mathcal{M} \preceq \mathcal{B}$. Then,

$$\|\mathcal{B} - \mathcal{M}\|_{op} \leq \|\mathcal{B} - \mathcal{A}\|_{op}, \ \textit{and} \ \ \|\mathcal{M} - \mathcal{A}\|_{op} \leq \|\mathcal{B} - \mathcal{A}\|_{op}.$$

(e) $\|\mathcal{K}_0^{1/2} - \mathcal{K}^{1/2}\|_{op} \leq 4\|\mathcal{K}_0 - \mathcal{K}\|_{op}^{1/2}$.

Proof (a) The relation $\mathcal{K} \preceq (\mathcal{K} + \epsilon\mathbb{I})$ is obvious. To see the other inequality we will use (3), together with the fact that, as $\mathcal{K}$ is injective, Spectral Theorem implies that the eigenvectors $\{e_j\}$ corresponding to the positive eigenvalues $\{\lambda_j\}$ are in fact an orthonormal basis of $\boldsymbol{\mathcal{H}}$. Thus, for any $x = \sum_j \langle x, e_j\rangle e_j$, $\langle(\mathcal{K} - (\mathcal{K} - \epsilon\mathbb{I})_+)x, x\rangle = \sum_j (\lambda_j - (\lambda_j - \epsilon)^+) \langle x, e_j\rangle^2$, and $(\lambda_j - (\lambda_j - \epsilon)^+)$ is either $\lambda_j - (\lambda_j - \epsilon) = \epsilon > 0$, or λ_j, depending on whether $\lambda_j > \epsilon$ or $\lambda_j \leq \epsilon$, respectively.

(b) As $h_0 = \sup_{\|x\|=1} |\langle(\mathcal{K} - \mathcal{K}_0)x, x\rangle|$, we have for any x with $\|x\| = 1$,

$$\langle(\mathcal{K} + h_0\mathbb{I} - \mathcal{K}_0)x, x\rangle = \langle(\mathcal{K} - \mathcal{K}_0)x, x\rangle + h_0\langle\mathbb{I}x, x\rangle = \langle(\mathcal{K} - \mathcal{K}_0)x, x\rangle + h_0 \geq 0,$$

We thus conclude $\mathcal{K}_0 \preceq \mathcal{K} + h_0\mathbb{I}$.

Now, to prove the other inequality, let us first note that, we also have $\mathcal{K} - h_0\mathbb{I} \preceq \mathcal{K}_0$ (whose proof is similar to that of $\mathcal{K}_0 \preceq \mathcal{K} + h_0\mathbb{I}$). Now, the first inequality in (b) follows from the following general property:

$$\text{Let } \mathcal{A} \in \boldsymbol{\mathcal{C}}, \mathcal{B} \in \boldsymbol{\mathcal{C}}_+, \text{ with } \mathcal{A}x = \sum_j \lambda_j \langle x, e_j\rangle e_j \ (\{e_j\} \text{ being an orthonormal basis}).$$

$$\text{If } \mathcal{A} \preceq \mathcal{B}, \text{then } \mathcal{A}_+ \preceq \mathcal{B}.$$

To prove this, define $x^* = \sum_j x_j^* e_j$, where $x_j^* = \langle x, e_j\rangle$ when $\lambda_j > 0$, $x_j^* = 0$, otherwise and $x^{**} = x - x^*$. We have $\langle \mathcal{A}_+ x, x\rangle = \langle \mathcal{A}x^*, x^*\rangle$, so that

$$\langle \mathcal{B}x, x\rangle = \langle \mathcal{B}x^*, x^*\rangle + \langle \mathcal{B}x^{**}, x^{**}\rangle \geq \langle \mathcal{A}x^*, x^*\rangle = \langle \mathcal{A}_+ x, x\rangle.$$

We thus conclude $\mathcal{A}_+ \preceq \mathcal{B}$ and the inequality $(\mathcal{K} - h_0\mathbb{I})_+ \preceq \mathcal{K}_0$ follows from $(\mathcal{K} - h_0\mathbb{I}) \preceq \mathcal{K}_0$.

(c) This follows directly from (a) and (b), together with the operator monotonicity of the square root transformation for non-negative bounded self-adjoint operators; see Pedersen (1972).

(d) The result is immediate since $\langle \mathcal{B}x, x\rangle \geq \langle \mathcal{M}x, x\rangle \geq \langle \mathcal{A}x, x\rangle \geq 0$. and

$$\|\mathcal{B} - \mathcal{A}\|_{op} = \sup_{\|x\|=1} \left((\langle \mathcal{B}x, x\rangle - \langle \mathcal{A}x, x\rangle) \right).$$

(e) Using the precedent results (b), (c) and (d), together with triangular inequality, we get

$$\|\mathcal{K}_0^{1/2} - \mathcal{K}^{1/2}\|_{op} \leq 2\|(\mathcal{K} + h_0\mathbb{I})^{1/2} - (\mathcal{K} - h_0\mathbb{I})_+^{1/2}\|_{op}.$$

But, since

$$\langle(\mathcal{K} + h_0\mathbb{I})^{1/2}x - (\mathcal{K} - h_0\mathbb{I})_+^{1/2}x, x\rangle = \sum_j \left(\sqrt{(\lambda_j + h_0)} - \sqrt{(\lambda_j - h_0)^+} \right) \langle x, e_j\rangle^2,$$

and $\|x\|^2 = \sum_j \langle x, e_j\rangle^2 = 1$, we conclude

$$\|(\mathcal{K} + h_0\mathbb{I})^{1/2} - (\mathcal{K} - h_0\mathbb{I})_+^{1/2}\|_{op} \leq \sup_j \left(\sqrt{(\lambda_j + h_0)} - \sqrt{(\lambda_j - h_0)^+} \right)$$

$$= \sup_j \frac{(\lambda_j + h_0) - (\lambda_j - h_0)^+}{\sqrt{(\lambda_j + h_0)} + \sqrt{(\lambda_j - h_0)^+}} \leq \frac{2h_0}{h_0^{1/2}} = 2h_0^{1/2}.$$

■

4.2. Consistency

Let us first establish a consistency statement for (17). As we will see the square root operator, $\mathcal{K}^{1/2}$, will play a relevant role throughout this subsection.

Theorem 12 *Under the assumptions and notation established at the beginning of Section 4, denote, as usual, by m the mean function of X, $\bar{X}$ the sample mean trajectory and n the sample size. Then,*

$$\widehat{M}_{\alpha,n}(X,\bar{X}) \to M_\alpha(X,m) \quad \textit{a.s., as } n \to \infty.$$

In fact, the result is true when measuring the distance between the mean and any function f in $L^2[0,1]$, that is, $\widehat{M}_{\alpha,n}(f,\bar{X}) \to M_\alpha(f,m)$ a.s., as $n \to \infty$.

Proof First note that $\widehat{M}_{\alpha,n}(X,\bar{X}) = \|\widehat{\mathcal{K}}^{1/2}(\widehat{\mathcal{K}}+\alpha\mathbb{I})^{-1}(X-\bar{X})\|$. This follows from Proposition 2(b), Eq. (14). Therefore,

$$\begin{aligned}\left|\widehat{M}_{\alpha,n}(X,\bar{X}) - M_\alpha(X,m)\right| &\leq \|\widehat{\mathcal{K}}^{1/2}(\widehat{\mathcal{K}}+\alpha\mathbb{I})^{-1}(X-\bar{X}) - \mathcal{K}^{1/2}(\mathcal{K}+\alpha\mathbb{I})^{-1}(X-m)\| \\ &\leq \|\widehat{\mathcal{K}}^{1/2}(\widehat{\mathcal{K}}+\alpha\mathbb{I})^{-1}\|_{op}\ \|\bar{X}-m\| + \|\widehat{\mathcal{K}}^{1/2}(\widehat{\mathcal{K}}+\alpha\mathbb{I})^{-1} - \mathcal{K}^{1/2}(\mathcal{K}+\alpha\mathbb{I})^{-1}\|_{op}\ \|X-m\|.\end{aligned}$$

By Lemma 10, $\|\bar{X}-m\|$ goes to zero a.s. as $n \to \infty$. As a consequence, it is enough to show that $\|\widehat{\mathcal{K}}^{1/2}(\widehat{\mathcal{K}}+\alpha\mathbb{I})^{-1} - \mathcal{K}^{1/2}(\mathcal{K}+\alpha\mathbb{I})^{-1}\|_{op} \to 0$ a.s. For that purpose, observe that

$$\begin{aligned}&\|\widehat{\mathcal{K}}^{1/2}(\widehat{\mathcal{K}}+\alpha\mathbb{I})^{-1} - \mathcal{K}^{1/2}(\mathcal{K}+\alpha\mathbb{I})^{-1}\|_{op} \\ &\leq \|\widehat{\mathcal{K}}^{1/2}(\widehat{\mathcal{K}}+\alpha\mathbb{I})^{-1} - \widehat{\mathcal{K}}^{1/2}(\mathcal{K}+\alpha\mathbb{I})^{-1}\|_{op} + \|\widehat{\mathcal{K}}^{1/2}(\mathcal{K}+\alpha\mathbb{I})^{-1} - \mathcal{K}^{1/2}(\mathcal{K}+\alpha\mathbb{I})^{-1}\|_{op} \\ &\leq \|\widehat{\mathcal{K}}^{1/2}\|_{op}\ \|(\widehat{\mathcal{K}}+\alpha\mathbb{I})^{-1} - (\mathcal{K}+\alpha\mathbb{I})^{-1}\|_{op} + \|\widehat{\mathcal{K}}^{1/2} - \mathcal{K}^{1/2}\|_{op}\ \|(\mathcal{K}+\alpha\mathbb{I})^{-1}\|_{op}.\end{aligned}$$

Therefore, to end the proof we will show that $\|\widehat{\mathcal{K}}^{1/2} - \mathcal{K}^{1/2}\|_{op} \to 0$ a.s. as $n \to \infty$ and $\|(\widehat{\mathcal{K}}+\alpha\mathbb{I})^{-1} - (\mathcal{K}+\alpha\mathbb{I})^{-1}\|_{op} \to 0$ a.s. as $n \to \infty$. From Lemma 10, $\|\widehat{\mathcal{K}} - \mathcal{K}\|_{op} \to 0$ a.s. Moreover, using part (e) of Proposition 11 with $\mathcal{K}_0 = \widehat{\mathcal{K}}$ we also get $\|\widehat{\mathcal{K}}^{1/2} - \mathcal{K}^{1/2}\|_{op} \to 0$, a.s., as $n \to \infty$.

Finally, observe that $\|\widehat{\mathcal{K}} - \mathcal{K}\|_{op} = \|(\widehat{\mathcal{K}}+\alpha\mathbb{I}) - (\mathcal{K}+\alpha\mathbb{I})\|_{op} \to 0$ a.s., and we also have $\sup_n \|(\widehat{\mathcal{K}}+\alpha\mathbb{I})^{-1}\|_{op} \leq \alpha^{-1} < \infty$, see Gohberg et al. (2003, eq. (1.14), p. 228). Then, Lemma 8 implies $\|(\widehat{\mathcal{K}}+\alpha\mathbb{I})^{-1} - (\mathcal{K}+\alpha\mathbb{I})^{-1}\|_{op} \to 0$ a.s. as $n \to \infty$.

The statement for a given f readily follows along the same lines. ■

4.3. Asymptotic distribution

Putting together Theorem 12 and Proposition 6 we get the asymptotic distribution of $\widehat{M}_{\alpha,n}$.

Corollary 13 *Under the assumptions and notation of Theorem 12, $\widehat{M}_{\alpha,n}(X,\bar{X})$ converges in distribution to $\sum_{j=1}^\infty \beta_j Y_j$, where $\beta_j = \lambda_j^2(\lambda_j+\alpha)^{-2}$ and $Y_1, Y_2, \ldots$ are independent χ_1^2 random variables.*

We can also prove another asymptotic distribution result involving the distances between the sample and the population means, which could be useful to perform inference on the mean.

Theorem 14 *Under the assumptions and notation established at the beginning of Section 4, we have*

$$\sqrt{n}\ \widehat{M}_{\alpha,n}(\bar{X},m) \ \xrightarrow{\mathrm{d}}\ \Big(\sum_{j=1}^{\infty} \frac{\lambda_j^2}{(\lambda_j+\alpha)^2} Y_j\Big)^{\frac{1}{2}}, \tag{18}$$

where $Y_1, Y_2, \ldots$ are independent χ_1^2 random variables.

Proof We can rewrite the left-hand side of Equation (18) as,

$$\sqrt{n}\ \widehat{M}_{\alpha,n}(\bar{X},m) \ = \ \sqrt{n}(\widehat{M}_{\alpha,n}(\bar{X},m) - M_\alpha(\bar{X},m)) + \sqrt{n}M_\alpha(\bar{X},m). \tag{19}$$

Now, from Equation (12) and Proposition 2, we have

$$\begin{aligned}&\sqrt{n}|\widehat{M}_{\alpha,n}(\bar{X},m) - M_\alpha(\bar{X},m)| \leq \sqrt{n}\ \|\widehat{\mathcal{K}}^{1/2}(\widehat{\mathcal{K}}+\alpha\mathbb{I})^{-1}(\bar{X}-m) - \mathcal{K}^{1/2}(\mathcal{K}+\alpha\mathbb{I})^{-1}(\bar{X}-m)\| \\ &\leq \|\widehat{\mathcal{K}}^{1/2}(\widehat{\mathcal{K}}+\alpha\mathbb{I})^{-1} - \mathcal{K}^{1/2}(\mathcal{K}+\alpha\mathbb{I})^{-1}\|_{op}\ \|\sqrt{n}(\bar{X}-m)\|.\end{aligned}$$

As a part of the proof of Theorem 12 we have seen that the first norm in the right-hand side goes to zero a.s. as $n \to \infty$. From the Functional Central Limit Theorem (see e.g., Theorem 7.5.1 in Laha and Rohatgi, 1979), $\sqrt{n}(\bar{X}-m)$ converges in distribution in $L^2[0,1]$ to a Gaussian stochastic process Z with zero mean and covariance operator $\mathcal{K}$. Since the norm is a continuous function in this space, by the continuous mapping theorem the second term converges in distribution to the random variable $\|Z\|$. Thus, by Slutsky's theorem, the distribution of the product goes to zero, and this convergence holds also in probability since the limit is a constant.

We can rewrite the remaining term of Equation (19) as,

$$\sqrt{n}M_\alpha(\bar{X},m) \ = \ \sqrt{n}\ \|\mathcal{K}^{1/2}(\mathcal{K}+\alpha\mathbb{I})^{-1}(\bar{X}-m)\| \ = \ \Big\|\sqrt{n}\ \Big(\frac{1}{n}\sum_{i=1}^{n} \chi_{\alpha,i} - \mu_\alpha\Big)\Big\|,$$

where we denote $\chi_{\alpha,i} = \mathcal{K}^{1/2}(\mathcal{K}+\alpha\mathbb{I})^{-1}X_i$ and $\mu_\alpha = \mathcal{K}^{1/2}(\mathcal{K}+\alpha\mathbb{I})^{-1}m$. Since $\mathcal{K}^{1/2}(\mathcal{K}+\alpha\mathbb{I})^{-1}$ is a bounded linear operator and the process X is Bochner-integrable ($\mathbb{E}\|X\| < \infty$), the expectation and the operator commute, that is,

$$\mathbb{E}\chi_{\alpha,1} \ = \ \mathbb{E}[\mathcal{K}^{1/2}(\mathcal{K}+\alpha\mathbb{I})^{-1}X_1] \ = \ \mathcal{K}^{1/2}(\mathcal{K}+\alpha\mathbb{I})^{-1}\mathbb{E}X_1 \ = \ \mu_\alpha.$$

Therefore, we can use again the Functional Central Limit Theorem with $\chi_{\alpha,i}$ and μ_α, since

$$\mathbb{E}\|\chi_{\alpha,1}\|^2 \ \leq \ \|\mathcal{K}^{1/2}(\mathcal{K}+\alpha\mathbb{I})^{-1}\|_{op}^2\ \mathbb{E}\|X_1\|^2 \ < \ \infty,$$

which gives us that $\sqrt{n}M_\alpha(\bar{X},m)$ converges in distribution to $\|\xi\|$, ξ being a random element with zero mean and whose covariance operator is the same as that of $\chi_{\alpha,1}$.

Using the same reasoning as at the beginning of the proof of Theorem 5 and denoting as $\mathcal{A}^*$ the adjoint of the operator $\mathcal{A}$, the covariance operator of $\chi_{\alpha,1}$ is given by

$$\mathcal{K}^{1/2}(\mathcal{K}+\alpha\mathbb{I})^{-1}\mathcal{K}[\mathcal{K}^{1/2}(\mathcal{K}+\alpha\mathbb{I})^{-1}]^* \ = \ \mathcal{K}^{1/2}(\mathcal{K}+\alpha\mathbb{I})^{-1}\mathcal{K}(\mathcal{K}+\alpha\mathbb{I})^{-1}\mathcal{K}^{1/2},$$

since both $\mathcal{K}^{1/2}$ and $(\mathcal{K}+\alpha\mathbb{I})^{-1}$ are self-adjoint operators (for instance, Theorem 3.35 and Problem 3.32 of Kato, 2013, Chapter 5 and Proposition 2.4 of Conway, 1990, Chapter X).

Now ξ is a zero-mean Gaussian process with a compact covariance operator which has the same eigenfunctions as $\mathcal{K}$ and its eigenvalues are $\lambda_j^2(\lambda_j+\alpha)^{-2}$. Thus, using its Karhunen-Loève representation we get

$$\|\xi\| \;=\; \|\sum_{j=1}^{\infty} Z_j e_j\| \;=\; \Big(\sum_{j=1}^{\infty} Z_j^2\Big)^{\frac{1}{2}},$$

where e_j are the eigenfunctions of $\mathcal{K}$ and Z_j are independent Gaussian random variables with zero mean and variances $\lambda_j^2(\lambda_j+\alpha)^{-2}$ (the eigenvalues of the covariance operator of ξ). Then the result follows from the standardization of these Z_j, applying Slutsky's theorem to the sum of Equation (19). ■

4.4. Convergence rates

Convergence rates, in probability, can be obtained for the distance estimator using part (e) of Proposition 11.

Theorem 15 *Under the assumptions established at the beginning of Section 4 and assuming in addition that* $\mathbb{E}\|X\|^4<\infty$*, we have*

$$|\widehat{M}_{\alpha,n}(X,\bar{X})-M_\alpha(X,m)| = O_P(n^{-1/4}) \quad \text{as } n\to\infty. \tag{20}$$

where O_P *stands here for convergence order in probability.*

Proof At the beginning of the proof of Theorem 12, we got

$$\begin{aligned}&\left|\widehat{M}_{\alpha,n}(X,\bar{X})-M_\alpha(X,m)\right| \le \\ &\le \|\widehat{\mathcal{K}}^{1/2}(\widehat{\mathcal{K}}+\alpha\mathbb{I})^{-1}\|_{op}\,\|\bar{X}-m\| + \|\widehat{\mathcal{K}}^{1/2}(\widehat{\mathcal{K}}+\alpha\mathbb{I})^{-1}-\mathcal{K}^{1/2}(\mathcal{K}+\alpha\mathbb{I})^{-1}\|_{op}\,\|X-m\|.\end{aligned} \tag{21}$$

The first term in the right hand side of (21) is the product of two random variables, say U_1 and U_2. From the a.s. convergence of $\widehat{\mathcal{K}}$ towards $\mathcal{K}$, we know that U_1 converges a.s. to $\|\mathcal{K}^{1/2}(\mathcal{K}+\alpha\mathbb{I})^{-1}\|_{op}$. Regarding the second factor $U_2=\|\bar{X}-m\|$, we have $\mathbb{E}(U_2^2)=\frac{1}{n}\mathbb{E}(\|X-m\|^2)$; indeed, note that $\bar{X}-m=\sum_{i=1}^n Z_i/n$ where the $Z_i=X_i-m$ are iid random elements with the same distribution as $X-m$. Therefore, as we have $\mathbb{E}\|X\|^2<\infty$ (which entails $\mathbb{E}\|X-m\|^2<\infty$) we conclude that U_2 is $O_P(n^{-1/2})$ as $n\to\infty$. Therefore the global convergence rate for the first term in the right-hand side of (21) is of type $O_P(n^{-1/2})$.

Now, to handle the second term in the upper bound of (21) note that

$$\begin{aligned}&\|\widehat{\mathcal{K}}^{1/2}(\widehat{\mathcal{K}}+\alpha\mathbb{I})^{-1}-\mathcal{K}^{1/2}(\mathcal{K}+\alpha\mathbb{I})^{-1}\|_{op}\\ &\le \|\widehat{\mathcal{K}}^{1/2}\|_{op}\,\|(\widehat{\mathcal{K}}+\alpha\mathbb{I})^{-1}-(\mathcal{K}+\alpha\mathbb{I})^{-1}\|_{op}+\|\widehat{\mathcal{K}}^{1/2}-\mathcal{K}^{1/2}\|_{op}\,\|(\mathcal{K}+\alpha\mathbb{I})^{-1}\|_{op} := W_1+W_2.\end{aligned}$$

Regarding W_1, we first note $\|\widehat{\mathcal{K}}^{1/2}\|_{op}\to\|\mathcal{K}^{1/2}\|_{op}$, a.s., since $\|\widehat{\mathcal{K}}^{1/2}-\mathcal{K}^{1/2}\|_{op}\to 0$, a.s. Now, for $\|(\widehat{\mathcal{K}}+\alpha\mathbb{I})^{-1}-(\mathcal{K}+\alpha\mathbb{I})^{-1}\|_{op}$ we use the following inequality (see Gohberg et al., 2003, Cor. 8.2, p. 71): if $\mathcal{A}$ is an invertible bounded linear operator and $\mathcal{A}_n$ is a sequence of bounded linear operators such that $\|\mathcal{A}_n - \mathcal{A}\|_{op} \to 0$, then the $\mathcal{A}_n$ are also (eventually) invertible and

$$\|\mathcal{A}_n^{-1} - \mathcal{A}^{-1}\|_{op} \leq \frac{\|\mathcal{A}^{-1}\|_{op}^2 \|\mathcal{A}_n - \mathcal{A}\|_{op}}{1 - \|\mathcal{A}^{-1}\|_{op}\|\mathcal{A}_n - \mathcal{A}\|_{op}},$$

so that, in particular, $\|\mathcal{A}_n^{-1} - \mathcal{A}^{-1}\|_{op} = O\left(\|\mathcal{A}_n - \mathcal{A}\|_{op}\right)$, as $n \to \infty$. Using this result for $\mathcal{A} = \mathcal{K} + \alpha\mathbb{I}$, $\mathcal{A}_n = \widehat{\mathcal{K}} + \alpha\mathbb{I}$, we get $W_1 = O\left(\|\widehat{\mathcal{K}} - \mathcal{K}\|_{op}\right)$, a.s. But according to Theorem 8.1.2 in Hsing and Eubank (2015) we have, under the assumption $\mathbb{E}\|X\|^4 < \infty$, that $\sqrt{n}(\widehat{\mathcal{K}} - \mathcal{K})$ converges in law (with respect to the Hilbert-Schmidt norm, $\|\cdot\|_{HS}$) to a Gaussian element with mean 0. In particular, $\sqrt{n}(\widehat{\mathcal{K}} - \mathcal{K})$ must be bounded in probability (with respect to $\|\cdot\|_{HS}$) so that $\|\widehat{\mathcal{K}} - \mathcal{K}\|_{op} = O_P(n^{-1/2})$, since $\|\cdot\|_{op} \leq \|\cdot\|_{HS}$. So, we conclude that W_1 is $O_P(n^{-1/2})$, as $n \to \infty$.

As for the term W_2 we use part (e) in Proposition 11, with $\mathcal{K}_0 = \widehat{\mathcal{K}}$. This gives $\|\widehat{\mathcal{K}}^{1/2} - \mathcal{K}^{1/2}\|_{op} \leq 4\|\widehat{\mathcal{K}} - \mathcal{K}\|_{op}^{1/2}$, so that, finally (using again Theorem 8.1.2 in Hsing and Eubank, 2015) we get that W_2 is $O_P(n^{-1/4})$. Collecting the convergence orders obtained for U_2, W_1 and W_2 we get (20). ■

5. Statistical applications

The purpose of this section is to give a general overview of possible applications of the proposed distance by analyzing its practical performance under various simulation scenarios and real data examples . The selected models and examples have been mostly chosen among those previously proposed in the literature. However, as usual in empirical studies, many other meaningful scenarios could be considered. Thus we make no attempt to reach any definitive conclusion. Only the long term practitioners' experience will lead to a safer judgment.

5.1. Exploratory analysis

The Mahalanobis distance can be used to analyze and summarize some interesting features of the data. Typically this is done, by example, via outlier detection and boxplots-type summaries of the data. We follow here the experimental setting proposed in Arribas-Gil and Romo (2014), where some real and simulated data sets are used for outliers detection and functional boxplots.

5.1.1. Outliers detection

Our aim. We will measure the number of true positives and false positives of different methods when detecting functional outliers. We ran 100 simulations of each model with different contamination rates $c = 0$, 0.05, 0.1, 0.15 and 0.2.

Methods. We replicate the simulation study proposed in Arribas-Gil and Romo (2014) with a selection of the methods presented there (we have excluded those methods that assumed some prior knowledge of the number of outliers). The details about the implementations of each method can be found in that paper. We have adapted the original code provided by the authors to include our method.

In order to formally define what we exactly mean by "an outlier" with our method, we have approximated the distribution of the random variable $\|X_\alpha - m_\alpha\|_K$ given in Corollary 7 for Gaussian processes through a Monte Carlo sample of size 2000 where the Monte Carlo observations are generated using the covariance structure of the original data. Note that this procedure may present some drawbacks when the process is far from being Gaussian, as we will see when generating the boxplots.

Then we mark as outliers the curves whose distance to the mean is greater that the 95% of the distances for the simulated data. A weak point of this method would be that the distribution of Corollary 7 is computed using the covariance structure of the data. Therefore, if the number of outliers is large compared with the sample size, this estimation is biased. In order to partially overcome this problem, we compute the covariance function using the robust minimum covariance determinant (MCD) estimator, see Rousseeuw and Driessen (1999).

Regarding our proposal, we have noticed that the choice of α does not affect the number of selected outliers significantly. We have chosen $\alpha = 0.01$, but an automatic technique (as the one proposed in Arribas-Gil and Romo, 2014 for the choice of the factor of the adjusted outliergram) could be used as well.

Simulation models. The curves are generated using three different combinations of the main process (from which most trajectories are drawn) and the contamination one (from which the outliers come from). Given a contamination rate c, $n - \lceil c \cdot n \rceil$ curves are drawn from the main process and $\lceil c \cdot n \rceil$ from the contamination one (we denote as $\lceil x \rceil$ the smallest integer not smaller than x).

- The first model is defined by,

  $$\begin{aligned} \text{main process: } X(t) &= 30t(1-t)^{3/2} + \varepsilon(t), \\ \text{contamination process: } X(t) &= 30t^{3/2}(1-t) + \varepsilon(t), \end{aligned}$$

  for $t \in [0,1]$, where ε is a Gaussian process with zero mean and covariance function $K(s,t) = 0.3\ \exp(-|s-t|/0.3)$.

- The second model is given by,

  $$\begin{aligned} \text{main process: } X(t) &= 4t + \varepsilon(t), \\ \text{contamination process: } X(t) &= 4t + (-1)^u 1.8 + (0.02\pi)^{-1/2}\ e^{\frac{-(t-\mu)^2}{0.02}} + \varepsilon(t), \end{aligned}$$

  where ε is a Gaussian process with zero mean and covariance function $K(s,t) = \exp(-|s-t|)$, u follows a Bernoulli distribution with parameter 0.5 and μ is uniformly distributed over $[0.25, 0.75]$.

- Finally, using the same definitions for ε and μ, the third model is given by,

  $$\begin{aligned} \text{main process: } X(t) &= 4t + \varepsilon(t), \\ \text{contamination process: } X(t) &= 4t + 2\sin(4(t+\mu)\pi) + \varepsilon(t). \end{aligned}$$

The sample size for each simulation was 100 and the curves are simulated in a discretized fashion over a grid of 50 equidistant points in $[0,1]$.

Results. The proportions (p_c and p_f) of correctly and falsely detected outliers for each method on the different settings can be found in Table 1. We can see that the Mahalanobis-based method proposed in this paper (denoted *Mah RKHS* in the table) is quite competitive.

5.1.2. Boxplots

Our aim. As a part of the exploratory analysis of the data, we include the functional boxplots of two real data sets used also in Arribas-Gil and Romo (2014). We do not include here the boxplots obtained with the other methods, which can be found in that paper.

Our methodology. We use the proposed Mahalanobis distance to define a depth measure by $(1 + M_\alpha^2(x,m))^{-1}$, for a realization x of the process. Using this depth, we mark as the functional median the deepest curve of the set. The central band of the boxplot is built as the envelope of the 50% deepest curves, and the "whiskers" are constructed as the envelope of all the curves that are not marked as outliers. In order to detect the outliers we use the same procedure as before. However, the sample sizes now are too small to estimate robustly the covariance matrix over the grid, so we use the standard empirical covariance matrix.

For the mortality data set we choose again the value $\alpha = 0.01$. However, for the Berkeley data set the distribution of the distances seems far from Gaussian. Then, in this case the value of α has a greater effect on the number of curves marked as outliers with our procedure. In an attempt to overcome this problem, the parameter α is adjusted automatically in order to minimize an estimate of the KL divergence between the empirical distribution and the distribution for Gaussian processes. The selected values of α with this procedure are 0.089 for the female set and 0.1 for the male set.

Data sets. The two data sets under study are:

- Male mortality rates in Australia 1901-2003: this data set can be found in the R package "fds". It contains Australia male log mortality rates between 1901 and 2003, provided by the Australian Demographic Data Bank.
- Berkeley growth: this data set is available in the R package "fda". It contains height measures of 54 girls and 39 boys, under the age of 18, at 31 fixed points.

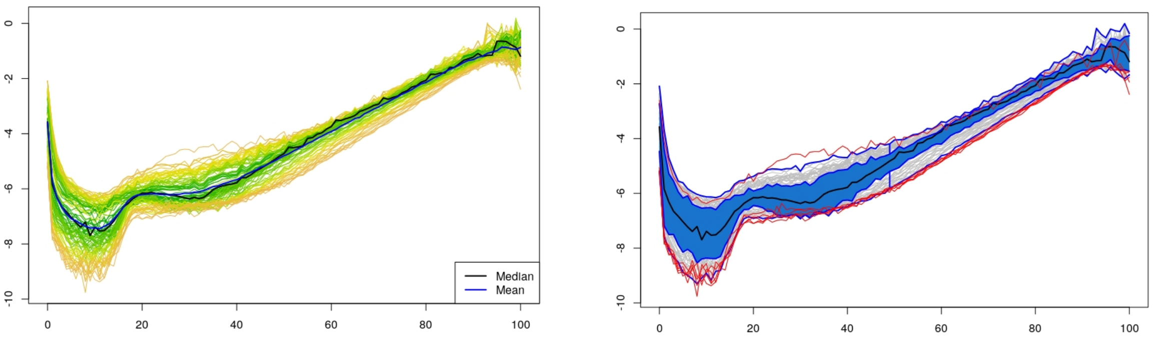
In Arribas-Gil and Romo (2014) the authors suggest to smooth the data, since the curves in the first set are very irregular. However, the distance proposed here has an intrinsic smoothing procedure, so we work directly with the original curves.

Results. The curves marked as outliers for the male mortality set correspond to years 1919 (influenza epidemic) and 1999-2003, which are among the curves detected using other different proposals in Arribas-Gil and Romo (2014). The resulting boxplot for this data set can be found in Figure 1b, where the outliers are plotted in red. This figure includes also (on the left) a graphic representation of the depths: from green, the deepest curves, to ochre, the outer ones.

The boxplots for the Berkely growth sets, female and male, are shown in Figure 2. Although now the smoothing parameter is adapted to the data set, the number of outliers detected is quite large when compared to the sample size.

c= 0	Model 1		Model 2		Model 3	
Method	p_c	p_f	p_c	p_f	p_c	p_f
Fun. BP	-	**0.001** (0.003)	-	**0.001** (0.002)	-	**0.000** (0.002)
Adj. Fun. BP	-	0.007 (0.010)	-	0.006 (0.010)	-	0.007 (0.012)
Rob. Mah. Dist.	-	0.016 (0.014)	-	0.015 (0.013)	-	0.015 (0.015)
ISE	-	0.038 (0.020)	-	0.032 (0.021)	-	0.033 (0.021)
DB weighting	-	0.016 (0.012)	-	0.015 (0.011)	-	0.014 (0.011)
Outliergram	-	0.054 (0.025)	-	0.057 (0.027)	-	0.058 (0.022)
Adj. Ourliergram	-	0.012 (0.012)	-	0.011 (0.013)	-	0.011 (0.014)
Mah. RKHS	-	0.037 (0.015)	-	0.033 (0.018)	-	0.035 (0.016)
c= 0.05	Model 1		Model 2		Model 3	
Method	p_c	p_f	p_c	p_f	p_c	p_f
Fun. BP	0.186 (0.193)	**0.001** (0.003)	0.208 (0.220)	**0.000** (0.001)	0.184 (0.179)	**0.000** (0.002)
Adj. Fun. BP	0.576 (0.282)	0.008 (0.012)	0.551 (0.330)	0.006 (0.010)	0.588 (0.344)	0.008 (0.012)
Rob. Mah. Dist.	0.976 (0.096)	0.008 (0.009)	0.361 (0.250)	0.008 (0.010)	0.104 (0.153)	0.015 (0.013)
ISE	0.865 (0.313)	0.033 (0.020)	**1.000** (0.000)	0.038 (0.026)	**1.000** (0.000)	0.033 (0.021)
DB weighting	0.894 (0.259)	0.008 (0.009)	0.941 (0.203)	0.012 (0.011)	0.957 (0.168)	0.011 (0.009)
Outliergram	**0.998** (0.020)	0.038 (0.022)	0.998 (0.020)	0.033 (0.021)	**1.000** (0.000)	0.036 (0.023)
Adj. Ourliergram	0.994 (0.035)	0.006 (0.008)	0.978 (0.070)	0.006 (0.009)	0.998 (0.020)	0.012 (0.014)
Mah. RKHS	**0.998** (0.020)	0.022 (0.016)	**1.000** (0.000)	0.027 (0.014)	**1.000** (0.000)	0.031 (0.016)
c= 0.1	Model 1		Model 2		Model 3	
Method	p_c	p_f	p_c	p_f	p_c	p_f
Fun. BP	0.139 (0.123)	**0.000** (0.001)	0.158 (0.151)	**0.000** (0.002)	0.134 (0.128)	**0.000** (0.002)
Adj. Fun. BP	0.549 (0.239)	0.005 (0.008)	0.593 (0.268)	0.008 (0.010)	0.632 (0.248)	0.008 (0.012)
Rob. Mah. Dist.	0.961 (0.105)	0.004 (0.007)	0.373 (0.170)	0.007 (0.009)	0.104 (0.108)	0.011 (0.014)
ISE	0.790 (0.335)	0.027 (0.017)	**1.000** (0.000)	0.036 (0.021)	**1.000** (0.000)	0.033 (0.022)
DB weighting	0.176 (0.247)	0.001 (0.004)	0.910 (0.232)	0.005 (0.008)	0.922 (0.258)	0.006 (0.008)
Outliergram	**0.981** (0.040)	0.020 (0.014)	0.998 (0.014)	0.018 (0.012)	**1.000** (0.000)	0.020 (0.016)
Adj. Ourliergram	0.897 (0.118)	0.006 (0.009)	0.971 (0.076)	0.006 (0.009)	**1.000** (0.000)	0.007 (0.011)
Mah. RKHS	0.767 (0.148)	0.014 (0.012)	**1.000** (0.000)	0.014 (0.011)	0.995 (0.030)	0.015 (0.013)
c= 0.15	Model 1		Model 2		Model 3	
Method	p_c	p_f	p_c	p_f	p_c	p_f
Fun. BP	0.098 (0.105)	**0.000** (0.002)	0.114 (0.101)	**0.000** (0.002)	0.134 (0.130)	**0.000** (0.001)
Adj. Fun. BP	0.494 (0.215)	0.006 (0.010)	0.550 (0.242)	0.006 (0.009)	0.584 (0.247)	0.006 (0.009)
Rob. Mah. Dist.	**0.927** (0.098)	0.001 (0.003)	0.324 (0.184)	0.004 (0.007)	0.152 (0.175)	0.005 (0.008)
ISE	0.778 (0.349)	0.027 (0.018)	**0.999** (0.007)	0.040 (0.029)	**1.000** (0.000)	0.034 (0.023)
DB weighting	0.020 (0.039)	0.001 (0.003)	0.659 (0.329)	0.002 (0.005)	0.634 (0.391)	0.002 (0.005)
Outliergram	0.879 (0.137)	0.011 (0.012)	0.984 (0.043)	0.008 (0.011)	0.999 (0.013)	0.008 (0.009)
Adj. Ourliergram	0.616 (0.220)	0.003 (0.007)	0.969 (0.099)	0.006 (0.008)	0.996 (0.019)	0.007 (0.010)
Mah. RKHS	0.295 (0.122)	0.013 (0.011)	0.988 (0.052)	0.008 (0.009)	0.941 (0.167)	0.007 (0.009)
c= 0.2	Model 1		Model 2		Model 3	
Method	p_c	p_f	p_c	p_f	p_c	p_f
Fun. BP	0.060 (0.090)	**0.000** (0.001)	0.098 (0.104)	**0.000** (0.001)	0.102 (0.094)	**0.000** (0.000)
Adj. Fun. BP	0.376 (0.226)	0.003 (0.006)	0.509 (0.205)	0.005 (0.009)	0.540 (0.227)	0.003 (0.006)
Rob. Mah. Dist.	**0.866** (0.167)	0.000 (0.002)	0.304 (0.171)	0.002 (0.005)	0.111 (0.118)	0.004 (0.007)
ISE	0.513 (0.396)	0.031 (0.023)	**0.997** (0.018)	0.047 (0.031)	**0.999** (0.010)	0.028 (0.023)
DB weighting	0.015 (0.025)	0.001 (0.004)	0.216 (0.228)	0.001 (0.003)	0.111 (0.179)	**0.000** (0.002)
Outliergram	0.356 (0.202)	0.002 (0.005)	0.894 (0.158)	0.001 (0.004)	0.959 (0.146)	0.001 (0.004)
Adj. Ourliergram	0.248 (0.179)	0.001 (0.003)	0.959 (0.074)	0.004 (0.008)	**0.999** (0.007)	0.008 (0.011)
Mah. RKHS	0.141 (0.089)	0.012 (0.011)	0.945 (0.127)	0.005 (0.007)	0.749 (0.232)	0.006 (0.009)

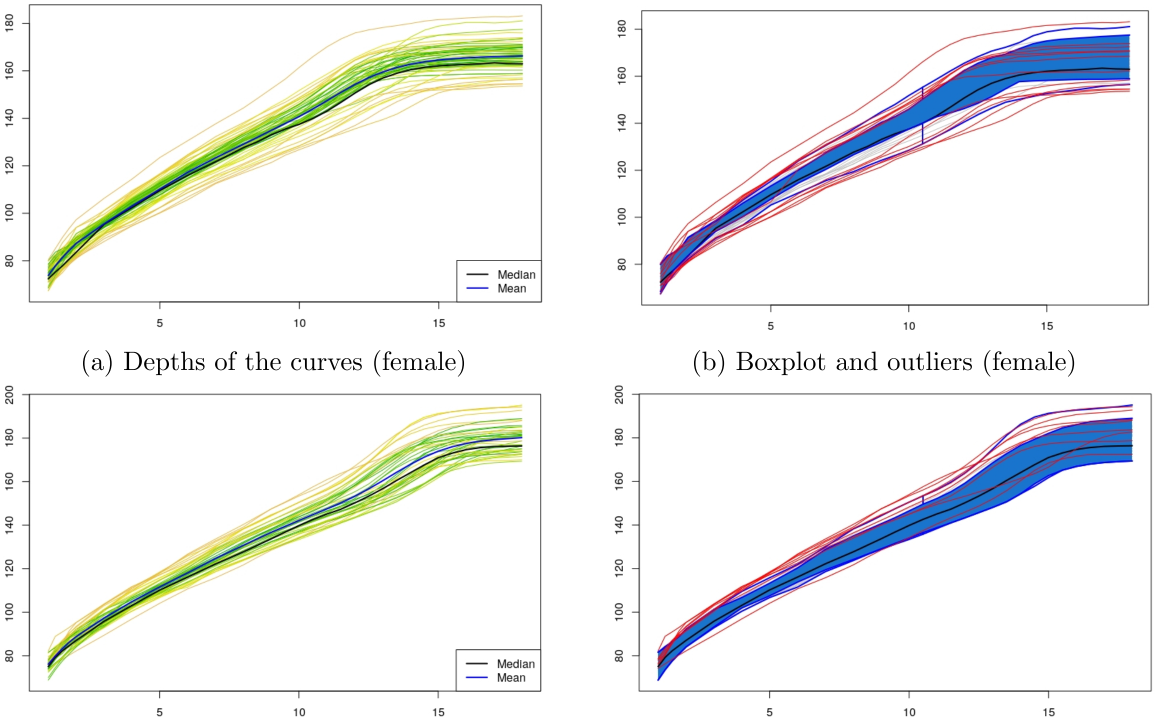
Table 1: Proportions of correctly and falsely detected outliers (standard errors appear in brackets)



(a) Depths of the curves

(b) Boxplot and outliers

Figure 1: Male mortality rates in Australia 1901-2003



(a) Depths of the curves (female)

(b) Boxplot and outliers (female)

(c) Depths of the curves (male)

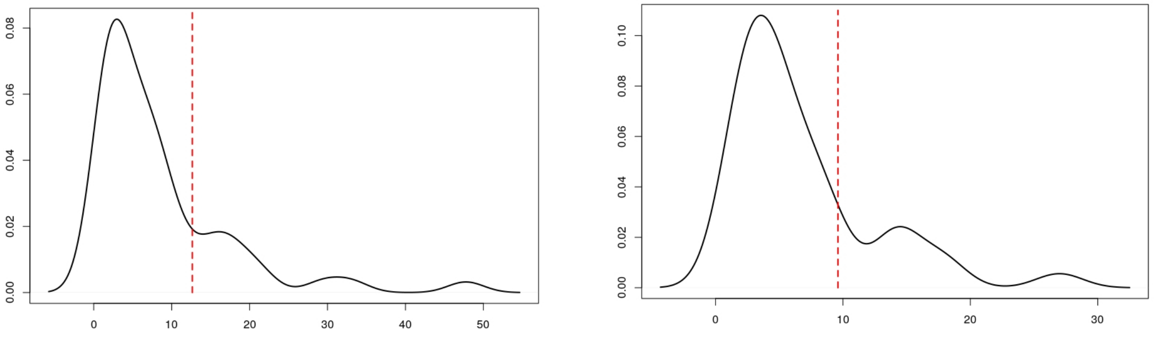
(d) Boxplot and outliers (male)

Figure 2: Berkeley growth

Female: 3, 8, 10, 13, 15, 18, 26, 29, 42, 43, 48 and 53.

Male: 5, 10, 15, 27, 29, 32, 35 and 37.

But if we look at the estimated density functions corresponding to the distribution of M_α^2 on each set (Figure 3), we can see that these distributions have two modes. In fact, all the curves marked as outliers are the ones that fall into the second mode (whose distance to the mean is greater that the red dotted line). This behavior is similar to that of the Integrated Squared Error shown in Arribas-Gil and Romo (2014).



(a) Female

(b) Male

Figure 3: Estimated density functions of the distributions of M_α^2 for Berkeley growth.

5.2. Supervised classification

The goals of the study. Mahalanobis distance can be also used as a tool for supervised classification. Mimicking a common criterion (see, e.g., Galeano et al., 2015, Section 3.3) we would classify the curve x to the population $j \in \{1, \ldots, k\}$ whenever

$$M_\alpha^2(x, m_j) - 2\log\pi_j = \min_{1\leq i\leq k}\left(M_\alpha^2(x, m_i) - 2\log\pi_i\right),$$

where $\pi_1, \ldots, \pi_k$ denote the prior probabilities of the classes, which are assumed to be homoskedastic, and m_j stands for the j-th mean function. In the examples below we will just consider the case of equal probabilities, which amounts to assign x to the population for which the Mahalanobis distance to the respective mean is smaller. The heteroskedastic case is also considered later on.

Here we present two different examples of binary classification with same prior probabilities, for which the percentage of misclassified curves is measured.

In order to check the performance of our proposal, we compare it with other classifiers presented below. The name used on the tables for each method is shown between brackets.

- Optimal Bayes classifier proposed in Dai et al. (2017) ("OB"). This is a functional extension of the classical multivariate Bayes classifier based on nonparametric estimators of the density functions corresponding to the main coefficients in Karhunen-Loève expansions. Here the curves are projected onto a common sequence of eigenfunctions, and the quotient in Bayes classifier is taken using the densities of these projections. The authors propose three approaches to estimate these densities. We have chosen the implementation which assumes that these densities are Gaussian since, according to their results, it seems to slightly outperform the others. The number of eigenfunctions used for the projections is fixed by cross-validation.

- Mahalanobis-based semidistance of Equation (6) proposed in Galeano et al. (2015) ("d_{FM}^k").

- k-nearest neighbours with 3 and 5 neighbours ("knn3" and "knn5"). In spite of its simplicity, this method tends to show a good performance when dealing with functional data.

Our methodology. Our proposal is denoted as "M_α". Now the parameter α is fixed by cross-validation, for $\alpha \in [10^{-4}, 10^{-1}]$. For heterokedastic problems, we have implemented our binary classifier mimicking an improvement that is usually made in the multivariate context. In that finite setting, given two equiprobable populations with covariance matrices Σ_0, Σ_1, a curve x is assigned to class 1, according to the Quadratic Discriminant classifier, whenever

$$M^2_{\Sigma_0}(x, x_0) - M^2_{\Sigma_1}(x, x_1) > \log \frac{|\Sigma_1|}{|\Sigma_0|},$$

where $M_{\Sigma_0}, M_{\Sigma_1}$ denote the finite dimensional Mahalanobis distances using the covariance matrices Σ_0 and Σ_1 respectively. The finite dimensional distance M is defined in (1) (see, for instance, Section 8.3.7 of Izenman, 2008). In most cases with multivariate data, classifying with this rule gives better results than merely classifying to class 1 when $M^2(x, x_0) > M^2(x, x_1)$.

In the case of functional data we will consider the following heuristic version of this idea. If m_0, K_0 and m_1, K_1 are the mean and covariance functions of each class, we classify x to class 1 if $M^2_{\alpha,K_0}(x, m_0) - M^2_{\alpha,K_1}(x, m_1) > C$, and to 0 otherwise. These are the distances between the projections into $\mathcal{H}(K_0)$ and $\mathcal{H}(K_1)$, respectively. The constant C is computed as $\log((\lambda^1_1 \cdot \ldots \cdot \lambda^1_{10})/(\lambda^0_1 \cdot \ldots \cdot \lambda^0_{10}))$, where λ^0_j, λ^1_j, $j = 1, \ldots, 10$, are the ten larger eigenvalues of $\mathcal{K}_0$ and $\mathcal{K}_1$, respectively.

First case study: cut Brownian Motion vs. Brownian Bridge. The first problem under consideration is to distinguish between two "cut" versions of a standard Brownian Motion and a Brownian Bridge. By "cut" we mean to take the process $X(t)$ on the interval $t \in [0, T]$, $T < 1$. We know an explicit expression for the Bayes error of this problem, which depends on the cut point T. For the case of equal prior probabilities of the classes, which will be the case here, this Bayes error is given by,

$$L^* \;=\; \frac{1}{2} - \Phi\left(\frac{(-(1-T)\log(1-T))^{1/2}}{(T(1-T))^{1/2}}\right) + \Phi\left(\frac{(-(1-T)\log(1-T))^{1/2}}{T^{1/2}}\right),$$

where Φ stands for the distribution function of a standard Gaussian random variable. Since both processes are almost indistinguishable around zero, $L^* \to 0.5$ when $T \to 0$. Also $L^* \to 0$ when $T \to 1$, since then one can decide the class with no error just looking at the last point of the curve.

The trajectories of both processes are shown in Figure 4 and the cut points considered, $0.75, 0.8125, 0.875, 0.9375$ and 1, are marked with vertical dotted lines. For each class, 50 samples are drawn for training and 250 for test. The experiment is run 500 times for each cut point, and the trajectories are sampled over an equidistant grid in $[0, 1]$ of size 50.

Second case study: simulated data as in Dai et al. (2017). We have implemented also the experimental setting proposed in Dai et al. (2017). The authors consider three different scenarios. For the first two, the curves of both classes $X^{(0)}$ and $X^{(1)}$, are drawn from processes

$$X^{(i)}(t) \;=\; m_i(t) + \sum_{j=1}^{50} A_{j,i}\phi_j(t) + \varepsilon, \quad i = 0, 1,$$

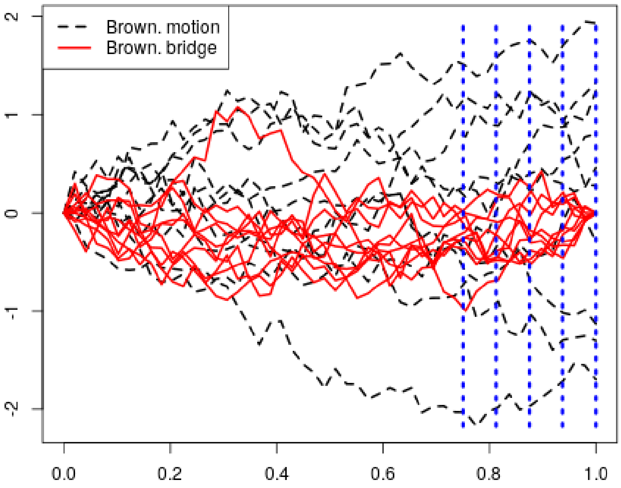


Figure 4: Trajectories of Brownian Motion and Bridge with cut points (vertical)

where ε is a Gaussian variable with zero mean and variance 0.01. Function ϕ_j is the jth element in the Fourier basis, starting with

$$\phi_1(t) = 1, \quad \phi_2(t) = \sqrt{2}\cos(2\pi t), \quad \phi_3(t) = \sqrt{2}\sin(2\pi t).$$

For Scenario A, the coefficients $A_{j,0}, A_{j,1}$ are independent centered Gaussian variables. For Scenario B they are independent centered exponential random variables. Finally, in Scenario C the processes are

$$X^{(i)}(t) \;=\; m_i(t) + \sum_{j=1}^{50} \frac{A_{j,i}}{B_i}\phi_j(t), \;\; i = 0, 1,$$

where $A_{j,0}, A_{j,1}$ are the same as in Scenario B and B_0, B_1 are independent variables with common distribution $\chi^2_{30}/30$. Thus, in this latter case the coefficients of the basis expansion are dependent but uncorrelated. The means and the variances of the coefficients $A_{j,i}$, $i = 0, 1$, are changed in order to check the "same" and "different" scenarios for mean and variances. Then $m_0(t) = 0$ always, and $m_1(t)$ is either 0 or t. In the same way, the variance of $A_{j,0}$ is always $\exp(-j/3)$ and the variance of $A_{j,1}$ is either $\exp(-j/3)$, or $\exp(-j/2)$. The curves are sampled on 51 equidistant points in $[0, 1]$.

The prior probabilities of both classes are set to 0.5 and two sample sizes, 50 and 100, are tested for training. For test we use 500 realizations of the processes. Each experiment is repeated 500 times.

Results. Table 2 shows the percentages of misclassified curves, as well as the Bayes errors for the first data set. Our proposal and knn with 5 neighbors seem to outperform the other methods for this problem.

The misclassification percentages for all the different scenarios of the second data set are shown in Table 3. Our proposal is mainly the winner, although in Scenario A it is overtaken by the Optimal Bayes classifier in the case of equal means and different variances. Also knn with 5 neighbors performs better sometimes in the case of different means and equal variances.

6. Some final remarks

In this section we include some final remarks and additional comments.

6.1. Sensitivity of the distance with respect to α and λ_i

In Section 3.3 we proved that the proposed distance is continuous with respect to the smoothing parameter α. However the choice of this value may be of importance in some cases, specially in those cases where cross-validation procedures can not be applied. Then we will try to shed some light on the role this parameter plays in the distance.

To simplify matters in the following discussion we consider the distance from a function x to the mean m. Let us begin with a simple, derivative-based, analysis of the squared distance, when considered as a function of α and λ_j, as defined in (12),

$$M_\alpha(x,m)^2 = \sum_{j=1}^{\infty} \frac{\lambda_j}{(\lambda_j+\alpha)^2} \langle x-m, e_j\rangle^2. \tag{22}$$

Some informal conclusions can be drawn from the above formula as well as from the definition of x_α:

(a) In general, larger values of α generate smoother curves f_α, for a given original function f. In order to gain some intuition on this, we show in Figure 5, the aspect of a smoothed Brownian trajectory x_α for different values of α.

(b) Of course, $M_\alpha(x,m)^2$ is a decreasing function of α. This means that when α gets larger, so providing increasingly smoother versions of x and m, the distance decreases to 0. Clearly a very large value of α would provide functions x_α and m_α too far away from the original unsmoothed versions x and m. On the opposite side, the choice $\alpha = 0$ leads to the divergence problems addressed in Subsection 1.4.

(c) The derivative (with respect to α) of the j-th coefficient $g(\lambda_j,\alpha) = \frac{\lambda_j}{(\lambda_j+\alpha)^2}$ in (22) is $\frac{\partial}{\partial\alpha} g(\lambda_j,\alpha) = \frac{-2\lambda_j}{(\lambda_j+\alpha)^3}$ whose absolute value is again a decreasing function of α.

Since λ_j are usually unknown, and must be estimated, it is also interesting to consider the sensitivity of the coefficients $g(\lambda_j,\alpha)$ with respect to λ_j. Note that $\frac{\partial}{\partial\lambda_j} g(\lambda_j,\alpha) = \frac{\alpha-\lambda_j}{(\alpha+\lambda_j)^3}$ so that the speed of variation (in absolute value) is minimized for $\lambda_j = \alpha$.

t	Bayes	M_α	OB	d_{FM}^k	knn3	knn5
0.75	33.9	42.5 (3.5)	43.5 (2.5)	46.4 (3.2)	43.2 (2.8)	**42.4** (2.8)
0.8125	30.8	**40.0** (3.7)	41.9 (2.6)	44.8 (3.3)	41.0 (2.8)	40.1 (3.0)
0.875	26.9	**36.1** (3.6)	40.2 (2.6)	42.6 (3.7)	38.0 (3.0)	36.9 (3.0)
0.9375	20.9	**32.3** (3.1)	38.0 (2.8)	39.9 (3.5)	33.7 (2.7)	32.5 (2.7)
1	0.0	**26.5** (2.8)	35.9 (2.9)	36.0 (3.5)	28.4 (2.7)	27.6 (2.7)

Table 2: Percentage of misclassification for the cut versions of Brownian Motion Brownian Bridge

Scenario A (Gaussian)							
n	mean	sd	M_α	OB	d_{FM}^k	knn3	knn5
50	same	diff	35.9 (3.5)	**19.0** (4.0)	47.0 (3.1)	45.6 (2.2)	46.2 (2.0)
	diff	same	42.3 (3.8)	47.3 (6.8)	43.7 (3.7)	42.9 (3.6)	**42.0** (3.6)
	diff	diff	**29.1** (5.0)	36.4 (10.1)	40.0 (5.4)	39.7 (3.0)	40.0 (3.1)
100	same	diff	34.2 (3.0)	**9.3** (2.1)	45.8 (3.5)	44.6 (1.9)	45.4 (1.8)
	diff	same	**34.6** (4.5)	45.1 (8.2)	37.0 (4.4)	42.1 (3.0)	41.0 (3.0)
	diff	diff	**22.0** (4.9)	35.7 (11.3)	34.2 (6.2)	38.3 (2.4)	38.6 (2.5)
Scenario B (exponential)							
n	mean	sd	M_α	OB	d_{FM}^k	knn3	knn5
50	same	diff	**24.2** (5.2)	30.2 (10.4)	37.0 (6.6)	37.6 (2.6)	38.0 (2.7)
	diff	same	41.8 (3.9)	49.1 (5.5)	42.3 (4.1)	38.0 (3.4)	**37.2** (3.6)
	diff	diff	**14.3** (4.8)	31.8 (12.8)	25.1 (9.0)	24.7 (3.1)	25.1 (3.5)
100	same	diff	**16.9** (3.1)	24.0 (9.6)	28.2 (6.1)	35.3 (2.4)	35.7 (2.3)
	diff	same	**34.5** (4.6)	48.3 (5.9)	36.7 (4.2)	36.5 (2.8)	35.6 (2.7)
	diff	diff	**7.7** (2.9)	30.1 (13.4)	17.8 (6.3)	21.6 (2.4)	21.8 (2.6)
Scenario C (dependent)							
n	mean	sd	M_α	OB	d_{FM}^k	knn3	knn5
50	same	diff	**30.0** (5.4)	33.3 (8.1)	40.1 (5.9)	39.9 (2.7)	39.9 (2.7)
	diff	same	43.6 (4.1)	48.8 (4.8)	42.9 (4.2)	38.1 (3.6)	**37.5** (3.8)
	diff	diff	**19.9** (4.9)	36.2 (11.0)	30.3 (7.7)	26.4 (3.1)	26.6 (3.3)
100	same	diff	**21.7** (3.0)	28.0 (7.5)	29.4 (5.7)	37.6 (2.4)	37.5 (2.4)
	diff	same	38.0 (4.3)	48.8 (5.0)	38.9 (3.8)	36.5 (2.7)	**35.6** (2.8)
	diff	diff	**13.3** (3.2)	34.6 (11.0)	23.2 (6.1)	23.4 (2.4)	23.3 (2.4)

Table 3: Percentage of misclassification for the experimental setting of Dai et al. (2017).

6.2. Mahalanobis-based classifiers and optimality

Our distance-based classifiers might be seen as a geometrically oriented proposal to functional classification. One might wonder under which circumstances these classifiers are close to the optimal (Bayes) classifier. It might be seen (see, e.g, Baíllo et al., 2011) that, under very general conditions, when the distribution P_1 of $X|Y=1$ is absolutely continuous with respect to that of $X|Y=0$, denoted by P_0, the Bayes rule is just $g(x) = \mathbf{1}(\{\frac{dP_1(x)}{dP_0} > \log\frac{1-p}{p}\})$, $\frac{dP_1(x)}{dP_0}$ being the Radon-Nikodym derivative of P_1 with respect to P_0, $p=\mathbb{P}(Y=1)$ and $\mathbf{1}(A)$ the indicator function of the event A . Now, taking into account the results by Parzen (1961) concerning the explicit expression of $\frac{dP_1(x)}{dP_0}$ in the Gaussian homoskedastic case (see Berrendero et al., 2018 for additional details) one might see that the optimal rule amounts to assign the observation x to population 1 whenever

$$\langle x-m_1, x-m_1\rangle_K - \langle x-m_2, x-m_2\rangle_K > \log\frac{1-p}{p}, \tag{23}$$

Here we must assume that P_1 and P_0 are Gaussian processes with continuous trajectories and a continuous covariance function K and the mean functions m_0 and m_1 belong both to the RKHS, $\mathcal{H}(K)$, associated with K. Also, it is important to note that the notation

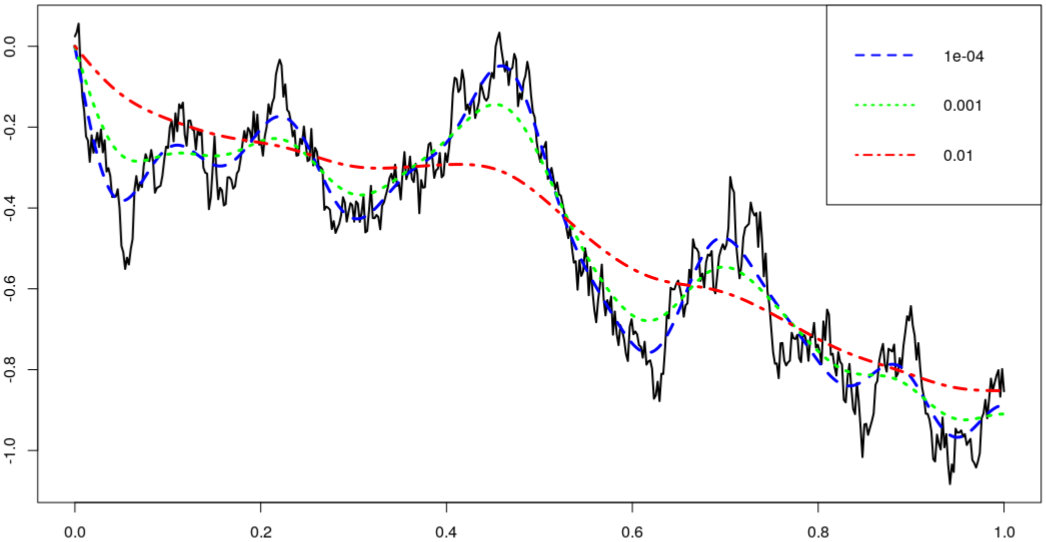


Figure 5: Smoothed versions of a Brownian trajectory for different values of α

$\langle x-m_1, x-m_1\rangle_K$ must be carefully interpreted in terms of the so-called Loève isometry since, with probability one, the trajectories x do not belong to $\mathcal{H}(K)$ so that $\langle x,\cdot\rangle_K$ is not directly defined and, in fact, (23) is defined in terms of such isometry.

As a consequence,

$$\langle x_\alpha - m_1, x_\alpha - m_1\rangle_K - \langle x_\alpha - m_2, x_\alpha - m_2\rangle_K > \log\frac{1-p}{p},$$

(where $\langle\cdot,\cdot\rangle_K$ is now the true inner product in $\mathcal{H}(K)$) can be seen as an approximation to the optimal rule (23) when we replace the Loéve isometry $\langle x-m_1, x-m_1\rangle_K$ with the smoothed approximation given by our estimated Mahalanobis-type distance. Note that, strictly speaking, $M_\alpha(x,m_i)^2 = \langle x_\alpha - m_{i\alpha}, x_\alpha - m_{i\alpha}\rangle_K = \|x_\alpha - m_{i\alpha}\|_K^2$ but, if we assume m_1, $m_2 \in \mathcal{H}(K)$, there is no need of considering the smoothed versions of these functions.

6.3. Connections with Tikhonov's regularization methodology

Our proposal is closely related to standard regularization results developed to deal with ill-posed equations in functional analysis; see e.g. Kress (1989). Indeed, observe that we can write $x_\alpha = (\mathcal{K}+\alpha\mathbb{I})^{-1}\mathcal{K}x = \mathcal{K}^{1/2}(\mathcal{K}+\alpha\mathbb{I})^{-1}\mathcal{K}^{1/2}x = \mathcal{K}^{1/2}R_\alpha x$, where $R_\alpha x := (\mathcal{K}+\alpha\mathbb{I})^{-1}\mathcal{K}^{1/2}x$ is a Thikonov regularization operator (which yields a continuous approximation to the Moore-Penrose pseudoinverse of $\mathcal{K}^{1/2}$).

Generally speaking, each regularization operator R_α is associated with a function $q(\alpha,\lambda_j)$ that downweights the effect of $1/\sqrt{\lambda_j}$ in the spectral representation of x_α. In the case of Thikonov regularization, $q(\alpha,\lambda_j) = \lambda_j/(\lambda_j+\alpha)$. Thus,

$$Rx_\alpha = \sum_{j=1}^\infty \frac{\sqrt{\lambda_j}}{\lambda_j+\alpha}\langle x, e_j\rangle e_j = \sum_{j=1}^\infty \frac{q(\alpha,\lambda_j)}{\sqrt{\lambda_j}}\langle x, e_j\rangle e_j,$$

and therefore $\|x_\alpha\|_K^2 = \|\mathcal{K}^{1/2}R_\alpha x\|_K^2 = \|R_\alpha x\|^2 = \sum_{j=1}^\infty q(\alpha,\lambda_j)^2/\lambda_j$.

In principle, to define the Mahalanobis distance one could select other regularization methods. For instance, the cut-off operator is given by

$$R_\alpha x = \sum_{\lambda_j \geq \alpha} \frac{1}{\sqrt{\lambda_j}} \langle x, e_j \rangle e_j,$$

which corresponds to $q(\alpha, \lambda_j) = 1$, when $\lambda_j \geq \alpha$, and $q(\alpha, \lambda_j) = 0$, when $\lambda_j < \alpha$. For this choice, the regularization parameter determines the terms of the series that are retained. Other possibility is the so-called Landweber regularization operator, given by $q(m, a, \lambda_j) = 1 - (1 - a\lambda_j)^{m+1}$, for $0 < a < \min\{\lambda_j^{-1} : \lambda_j > 0\}$. In this case regularization depends on two parameters, a and m.

We believe Tikhonov scheme provides a natural choice with good properties since it depends on just one regularization parameter (unlike Landweber method) and defines a metric in $L^2[0,1]$ (unlike the cut-off approach).

Acknowledgments

This work has been partially supported by Spanish Grant MTM2016-78751-P. The authors are grateful to Daniel Estévez and Dmitry Yakubovich for their help with operator theory. The constructive comments and suggestions from three anonymous reviewers are also gratefully acknowledged.

References

Ana Arribas-Gil and Juan Romo. Shape outlier detection and visualization for functional data: the outliergram. *Biostatistics*, 15(4):603–619, 2014.

Robert B. Ash and Melvin F. Gardner. *Topics in Stochastic Processes.* Academic Press, 1975.

Amparo Baíllo, Antonio Cuevas, and Juan Antonio Cuesta-Albertos. Supervised classification for a family of Gaussian functional models. *Scandinavian Journal of Statistics*, 38 (3):480–498, 2011.

Alain Berlinet and Christine Thomas-Agnan. *Reproducing Kernel Hilbert Spaces in Probability and Statistics.* Kluwer Academic, 2004.

José R. Berrendero, Antonio Cuevas, and José L. Torrecilla. On the use of reproducing kernel Hilbert spaces in functional classification. *Journal of the American Statistical Association*, 113(3):1210–1218, 2018.

John B. Conway. *A course in functional analysis.* Springer, 1990.

Felipe Cucker and Steve Smale. On the mathematical foundations of learning. *Bulletin of the American Mathematical Society*, 39(1):1–49, 2001.

Felipe Cucker and Ding Xuan Zhou. *Learning Theory: an Approximation Theory Viewpoint.* Cambridge University Press, 2007.

Antonio Cuevas. A partial overview of the theory of statistics with functional data. *Journal of Statistical Planning and Inference*, 147:1–23, 2014.

Xiongtao Dai, Hans-Georg Müller, and Fang Yao. Optimal Bayes classifiers for functional data and density ratios. *Biometrika*, 104(3):545–560, 2017.

Lokenath Debnath and Piotr Mikiusinski. *Introduction to Hilbert Spaces with Applications (3rd Ed.)*. Elsevier, 2005.

Pedro Galeano, Esdras Joseph, and Rosa E Lillo. The Mahalanobis distance for functional data with applications to classification. *Technometrics*, 57:281–291, 2015.

Andrea Ghiglietti and Anna Maria Paganoni. Exact tests for the means of Gaussian stochastic processes. *Statistics & Probability Letters*, 131:102–107, 2017.

Andrea Ghiglietti, Francesca Ieva, and Anna Maria Paganoni. Statistical inference for stochastic processes: two-sample hypothesis tests. *Journal of Statistical Planning and Inference*, 180:49–68, 2017.

Israel Gohberg, Seymour Goldberg, and Marinus Kaashoek. *Basic Classes of Linear Operators*. Birkhäuser, 2003.

Tailen Hsing and Randall Eubank. *Theoretical Foundations of Functional Data Analysis, with an Introduction to Linear Operators*. John Wiley & Sons, 2015.

Alan J. Izenman. *Modern Multivariate Statistical Techniques*. Springer, 2008.

Svante Janson. *Gaussian Hilbert Spaces*. Cambridge University Press, 1997.

Tosio Kato. *Perturbation Theory for Linear Operators*. Springer, 2013.

Rainer Kress. *Linear Integral Equations*. Springer, 1989.

Radha G. Laha and Vijay K. Rohatgi. *Probability Theory*. John Wiley & Sons, 1979.

Milan N. Lukić and Jay H. Beder. Stochastic processes with sample paths in reproducing kernel Hilbert spaces. *Transactions of the American Mathematical Society*, 353:3945–3969, 2001.

Prasanta C. Mahalanobis. On the generalized distance in statistics. *Proceedings of the National Institute of Sciences (Calcutta)*, 2:49–55, 1936.

Kanti V. Mardia. Assessment of multinormality and the robustness of Hotelling's t2-test. *Applied Statistics*, pages 163–171, 1975.

Emanuel Parzen. An approach to time series analysis. *Journal of the American Statistical Association*, 32:951–989, 1961.

Gert K. Pedersen. Some operator monotone functions. *Proceedings of the American Mathematical Society*, 36:309–310, 1972.

Kay I. Penny. Appropriate critical values when testing for a single multivariate outlier by using the Mahalanobis distance. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 45:73–81, 1996.

Alvin C. Rencher. *Methods of Multivariate Analysis*, volume 3 ed. John Wiley & Sons, 2012.

Peter J. Rousseeuw and Katrien van Driessen. A fast algorithm for the minimum covariance determinant estimator. *Technometrics*, 41:212–223, 1999.

Peter J. Rousseeuw and Bert C. van Zomeren. Unmasking multivariate outliers and leverage points. *Journal of the American Statistical Association*, 85(411):633–639, 1990.

Bernhard Schölkopf and Alexander J. Smola. *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond.* MIT Press, 2002.

Yijun Zuo and Robert Serfling. General notions of statistical depth function. *Annals of Statistics*, 28:461–482, 2000.