UNIVERSIDAD CARLOS III DE MADRID working papers

Working Paper 10-27
Statistics and Econometrics Series 013
May 2010

Departamento de Estadística
Universidad Carlos III de Madrid
Calle Madrid, 126
28903 Getafe (Spain)
Fax (34) 91 624-98-49

# Representing Functional Data in Reproducing Kernel Hilbert Spaces with applications to clustering and classification

Javier González and Alberto Muñoz

**Abstract**
Functional data are difficult to manage for many traditional statistical techniques given their very high (or intrinsically infinite) dimensionality. The reason is that functional data are essentially functions and most algorithms are designed to work with (low) finite-dimensional vectors. Within this context we propose techniques to obtain finite-dimensional representations of functional data. The key idea is to consider each functional curve as a point in a general function space and then project these points onto a Reproducing Kernel Hilbert Space with the aid of Regularization theory. In this work we describe the projection method, analyze its theoretical properties and propose a model selection procedure to select appropriate Reproducing Kernel Hilbert spaces to project the functional data.

---

**Keywords:** Functional Data, Reproducing, Kernel Hilbert Spaces, Regularization Theory.

Universidad Carlos III de Madrid, c/ Madrid 126, 28903 Getafe (Madrid), e-mail adresses: (Alberto Muñoz) alberto.munoz@uc3m.es, (Javier González) javier.gonzalez@uc3m.es.

**Acknowledgement**s: The research of Alberto Muñoz and Javier González was supported by Spanish Goverment grants 2006-03563-001, 2004-02934-001/002 and Madrid Goverment grant 2007-04084-001.

## 1. Introduction

The field of Functional Data Analysis (FDA) [Ramsay and Silverman, 2006] [Ferraty and Vieu, 2006] deals naturally with data of very high (or intrinsically infinite) dimensionality. Typical examples are functions describing physical processes, genetic data, control quality charts or spectra of data in Chemometrics.

In practice each functional datum is given by a data set $f_n = \{(x_i, y_i) \in X \times Y\}_{i=1}^n$, where $X$ is the space of input variables and, in most cases, $Y = \mathbb{R}$. The first task in any FDA methodology is to transform the data set $f_n$ into a function $f : X \to Y$ and then to apply some generalized multivariate procedure able to cope with functions. Of course $n$, the number of data points which can be recorded, is finite while an accurate description of the underlying function would require an infinite number of observations. Therefore the choice of a particular $f$ will be done in general by selecting it from an infinite collection of alternative models. This is the typical context in which ill-posed problems arise [Tikhonov and Arsenin, 1977].

Most FDA approaches choose an orthogonal basis of functions $B = \{\phi_1, \dots \phi_d\}$ $(d \in \mathbb{N})$, where each $\phi_j$ belongs to a general function space (usually $L^2(X)$) and then represent each functional datum by means of a linear combination in $Span(B)$ [Ramsay and Silverman, 2006]. Usual choices for functions in $B$ are Fourier, Wavelets or B-splines functions.

Our approach in this work will be to evaluate the goodness of fit of a particular function to a given functional datum by means of some "loss function" $L(y, f(x))$. The seeked function will be the minimizer of the empirical error $\frac{1}{n}\sum_{i=1}^n L(y_i, f(x_i))$ in a hypothesis space $\mathcal{H}$. It is well known that to achieve well-posedness of this problem and uniform convergence of the empirical error to the generalization error defined by $\int_{X\times Y} L(y, f(x)) d_\nu(x, y)$ (where $\nu$ is some probability measure on $X \times Y$), imposing compactness in $\mathcal{H}$ is a sufficient condition [Cucker and Smale, 2001, Moguerza and Muñoz, 2006]. A way to achieve this is to use regularization theory and the natural function spaces to use are the Reproducing Kernel Hilbert Spaces (RKHSs). Following this approach we propose a finite-dimensional representation for functional data based on a particular projection of the original functions onto a Reproducing Kernel Hilbert Space (RKHS).

RKHSs [Cucker and Smale, 2001, Wahba, 2003] are characterized by a generalized covariance function called kernel and the approximating function will be a linear combination of its eigenfunctions. Under general rather conditions we can build kernels from orthonormal basis of functions [Rakotomamonjy and Canu, 2005]. In addition, we can directly choose

the kernel (see Section 4 for details); in Section 2 we propose a method to approximate the eigenfunctions of a given kernel as a previous step to obtain the proposed functional data representation. To focus on the kernel makes accessible a wider class of basis of functions to represent the functional data. In this sense our approach constitutes a generalization of the usual FDA setting.

The choice of the kernel in regularization methods is a relevant problem that has been extensively studied in the literature. We refer to [Keerthi and Lin, 2003, Lanckriet et al., 2004, Moguerza and Muñoz, 2006] for some references in the classification context and to [Cherkassky and Ma, 2004] regarding regression problems. In this paper we will make use of the Subspace Information Criterion (SIC) [Sugiyama and Ogawa, 2001, Sugiyama and Muller, 2002] to select the kernel that generates the RKHS. The SIC is designed to approximate the Generalization Error in general regularization methods and it has been proven to be very competitive as model selection criteria compared to other model selection criteria choices [Sugiyama and Ogawa, 2002]. In this work we will show how to adapt it to select the optimal space where project the curves.

This paper is organized as follows. In next section we show how to project functional data onto RKHSs. In Section 3 we study the metric for curves induced by the previous projection mrthodology. In Section 4 we describe how to use the SIC to select the space where project the curves. In Section 5 a wide variety of experimental results are shown and we conclude in Section 5 with some conclusions a future lines of research.

## 2. Representing Functional Data in Reproducing Kernel Hilbert Spaces

A Hilbert function space $H$ is a RKHS where all the (linear) evaluation functionals ($\mathcal{F}_x : H \to \mathbb{R}$ such that $\mathcal{F}_x(f) = f(x)$, where $x \in X$) are bounded (equivalently continuous). By the Riesz representation theorem, for each $x \in X$ there exists $h_x \in H$ such that for every $f \in H$ it holds that $f(x) = \langle h_x, f \rangle$, where $\langle , \rangle$ denotes the inner product in $H$. The RKHS $H$ is characterized by a continuous symmetric positive definite function $K : X \times X \to \mathbb{R}$ named Mercer Kernel or reproducing kernel for $H$ [Aroszajn, 1950]. The elements of $H$, $\mathcal{H}_K$ in the sequel, can be expressed as finite linear combinations of the form $h = \sum_s \lambda_s K(x_s, \cdot)$ where $\lambda_s \in \mathbb{R}$ and $x_s \in X$.

Consider the linear integral operator $L_K$ (associated to the kernel function $K$) defined by $L_K(f) = \int_X K(\cdot, s) f(s) ds$. When $X$ is compact and $K$ continuous, then $L_K$ has a countable sequence of eigenvalues $\{\lambda_j\}$ and (orthonormal) eigenfunctions $\{\phi_j\}$ and $K$ can be expressed by $K(x, y) = \sum_j \lambda_j \phi_j(x) \phi_j(y)$ where the convergence is absolute and uniform (Mercer's theorem [Mercer, 1909]).

### 2.1. Projecting functional data onto RKHSs

Let $X$ be a compact space or manifold in a Euclidean Space and $Y = \mathbb{R}$. Let $\nu$ be a Borel probability measure defined on $X \times Y$. In the sequel we will assume that $\nu$ is non degenerate.

Denote by $f_n$ a sample curve drawn form $\nu$ identified with a data set $\{(x_i, y_i) \in X \times Y\}_{i=1}^n$. Define $f_\nu : X \to Y$,

$$f_\nu = \int_X y d_\nu(y \,|\, x), \tag{1}$$

where $d_\nu(y \,|\, x)$ is the the conditional probability measure on $Y$. Thus $f_n$ is a sample version of size $n$ of $f_\nu$. In practice we are usually given a set of curves $\{f_{n,1}, \ldots, f_{n,m}\}$ where each sample curve$f_{n,l}$ is drawn, in the most general case, from a different measure $\nu_l$ and it is identified with a data set $\{(x_i, y_{il}) \in X \times Y\}_{i=1}^n$. For simplicity in notation we will assume that the vector $\mathbf{x} = (x_1, \ldots, x_n)^T$ is common for all the curves, as it is the habitual case in the literature [Ramsay and Silverman, 2006].

Next we develop a procedure to approximate $f_\nu$ using the associated $f_n$.

Definition 1. *Let $X$ be a compact space or manifold in and Euclidean Space, $Y = \mathbb{R}$ and $\nu$ a Borel probability measure defined on $X \times Y$. Let $f_n = \{(x_i, y_i) \in X \times Y\}_{i=1}^n$ be a sample curve drawn form $\nu$ and consider $f_\nu$ defined in eq. (1). Let $K : X \times X \to \mathbb{R}$ be a Mercer kernel and $\mathcal{H}_K$ its associated RKHS. Then we define the* ***Regularized $\gamma$-Projection*** *of $f_\nu$ onto $\mathcal{H}_K$ associated to the sample curve $f_n$ as*

$$f^*_{K,\gamma,n} = \Pi_{K,\gamma,n}(f_\nu) = \arg\min_{f \in \mathcal{H}_K} \frac{1}{n} \sum_{i=1}^n (f(x_i) - y_i)^2 + \gamma \|f\|_K^2 \,, \tag{2}$$

*where $\gamma > 0$ and $\|f\|_K$ represents the norm of the function $f$ in $\mathcal{H}_K$.*

Below, in Theorem 1, we show that $f^*_{K,\gamma,n} \in span\{K(x, x_i)\}$, then for every $x \in X$, $f(x) = \sum_{j=1}^n \alpha_j K(x_j, x)$, for appropriate $x_j \in X$ and $\alpha_j \in \mathbb{R}$. Thus, denoting $\alpha = (\alpha_1, \ldots, \alpha_n)^T$, $\mathbf{x} = (x_1, \ldots, x_n)^T$, and $K\big|_{\mathbf{X}}$ the matrix whose components are $(K\big|_{\mathbf{X}})_{ij} = K(x_i, x_j)$, we have $\|f^*_{K,\gamma,n}\|_K^2 = \sum_{i=1}^n \sum_{i=1}^n \alpha_i \alpha_j K(x_i, x_j) = \alpha^T K\big|_{\mathbf{X}} \alpha$. Eq. (2) quantifies the balance between the fitness of the function to the data (measured by $\frac{1}{n}\sum_{i=1}^n (f(x_i) - y_i)^2$) and the complexity of the solution (measured by $\|f\|_K^2$). Notice that that in eq. (2) we denote by $f^*_{K,\gamma,n}$ and $\Pi_{K,\gamma,n}(f_\nu)$ the estimated curve. While we will use the first notation in the sequel, we include the second to remark that the obtained curve is the result of projecting $f_\nu$ onto the $\mathcal{H}_K$ using $f_n$.

Definition 1 can be generalized in several directions. The first term can be replaced by a different loss function. For instance we could consider $L(x, y) = |x - y|$, or any linear convex combination of $L(x, y) = |x - y|^p$ loss functions. Other possible choice for the loss function in (2) is the so-called $\epsilon$-insensitive loss function, given by $L(y_i, f(x_i)) = (|f(x_i) - y_i| - \varepsilon)_+$, $\varepsilon \geq 0$ (used by the Support Vector Machine for regression [Smola and Schölkopf, 1998]). The conditions for a loss function $L : \mathbb{R} \times Y \to \mathbb{R}^+$ to guarantee uniform stability in the regularization approach are: 1) $L$ is a Lipschitz function, 2) There exists a constant $C$ such that $L(0, y) \leq C \; \forall y \in Y$. (see [Mukherjee et al., 2002] and [Bousquet and Elisseeff, 2002] for further details and implications).

Regarding the second term in (2), we can replace $\|f\|_K^2$ with a general convex positive functional $\Omega(f)$. There are two frequent choices. In the first case, we consider $\|Lf\|^2$, where $L$ is a linear differential operator [Ramsay and Silverman, 2006, Chen and Haykin, 2002]. In particular the Green's function of the operator $L^*L$ ($L^*$ the adjoint operator to $L$) satisfies the condition of being a valid kernel and thus, this case may be seen as a particular case in the frame of the RKHS formalism. In the second case we consider $\|Pf\|^2$, where $P$ is a projection operator onto a finite dimensional subspace [Wahba, 1990]. The underlying idea is to choose two orthogonal sets of basis functions $\{\phi_k\}$ and $\{\psi_l\}$ in such a way that the $\{\phi_k\}$ (small in number) can provide a first approximation to the function, and the $\{\psi_l\}$ (usually much larger in number) are able to provide a larger accuracy in approximation. $P$ annihilates some of the $\{\psi_k\}$ when using $\|Pf\|^2$. For further details, see [Ramsay and Silverman, 2006], chapter 5. Notice that we need in every case to work with a bounded linear operator to guarantee that we can apply the Riesz representation theorem and be able to define a kernel in each case (see [Wahba, 2003] for additional possibilities).

THEOREM 1 (REPRESENTER THEOREM [CUCKER AND SMALE, 2001]). *Consider a sample curve $f_n$ defined by $\{(x_i, y_i) \in X \times Y\}_{i=1}^n$,then the minimizer $f^*_{K,\gamma,n}$ to the functional optimization problem in eq. (2) exists, is unique and admits a representation of the form*

$$f^*_{K,\gamma,n}(x) = \sum_{i=1}^{n} \alpha_i K(x_i, x), \quad \forall \boldsymbol{x} \in X\,, \tag{3}$$

*where now the $x_i$ points are the sample data (components of the vector $\boldsymbol{x}$) and the coefficients $\alpha_i \in \mathbb{R}$ are the solutions to the linear system:*

$$(\gamma n \boldsymbol{I}_n + K\big|_{\boldsymbol{x}})\alpha = \boldsymbol{y}, \tag{4}$$

*where $\boldsymbol{I}_n$ the identity matrix of dimension $n \times n$, $\alpha = (\alpha_1, \ldots, \alpha_n^T)$ and $\boldsymbol{y} = (y_1, \ldots, y_n)^T$.*

By solving eq. (4) we get a closed expression for $f^*_{K,\gamma,n}$, the minimizer of problem (2). When $\gamma = 0$ we can interpret eq. (2) as the orthogonal projection of $f_\nu$ onto $\mathcal{H}_K$ via $f_n$ as follows.

PROPOSITION 1. *Let $X$ be a compact space or manifold in a Euclidean Space, $Y = \mathbb{R}$ and $\nu$ a Borel probability measure defined on $X \times Y$. Let $f_n = \{(x_i, y_i) \in X \times Y\}_{i=1}^n$ be a sample curve drawn form $\nu$ and consider $f_\nu$ defined in eq. (1). Let $K : X \times X \longrightarrow \mathbb{R}$ be a continuous symmetric positive definite kernel with associated integral operator $L_K$ with eigenfunctions $\{\phi_1, \phi_2, \ldots\}$ and eigenvalues $\{\lambda_1, \lambda_2, \ldots\}$. Then, when $\gamma = 0$, the projected curve $f^*_{K,0,n}$ obtained by solving problem (2) can be written by*

$$f^*_{K,0,n} = \Pi_{K,0,n}(f_\nu) = \sum_{j=1} \lambda_j (\alpha^T \phi_{j,\boldsymbol{x}}) \phi_j(x), \tag{5}$$

*where $\alpha$ is the solution to eq. (4) and $\phi_{j,\boldsymbol{x}} = (\phi_j(x_1), \ldots, \phi_j(x_n))^T$. In addition*

$$f^*_{K,0,n} = \sum_j \lambda_j (\alpha^T \phi_{j,\boldsymbol{x}}) \phi_j \underset{n\to\infty}{\longrightarrow} \sum_j \lambda_j \langle f_\nu, \phi_j \rangle \phi_j, \tag{6}$$

*where the convergence is uniform in $X$.*

By the Spectral Theorem [Conway, 1990] $L_K(f_\nu) = \sum_j \lambda_j \langle f_\nu, \phi_j \rangle \phi_j$. Thus $f^*_{K,0,n}$ converges uniformly to $L_K(f_\nu)$ the orthogonal projection of $f_\nu$ onto $\mathcal{H}_K$. When $\gamma > 0$, $\Pi_{K,\gamma,n}$ can also be interpreted as a projection of $f_\nu$ onto $\mathcal{H}_K$ as it is shown in next proposition.

PROPOSITION 2. *Under the same assumptions as in Proposition 1, when $\gamma > 0$, the projected curve $f^*_{K,\gamma,n}$, given by the minimization of eq. (2) can also be interpreted as a projection of $f_\nu$ onto $\mathcal{H}_K$ and*

$$f^*_{K,\gamma,n} = \sum_j \lambda_j (\alpha^T \phi_{j,\boldsymbol{x}}) \phi_j \underset{n\to\infty}{\longrightarrow} \sum_j \lambda_j \langle f_\nu, \phi_j \rangle' \phi_j, \tag{7}$$

*where the convergence is uniform in $X$, $\alpha$ is the solution to eq. (4), $\{\lambda_j\}$ are the eigenvalues of $L_K$, $f_{\boldsymbol{x}} = (f(x_1), \ldots, f(x_n))^T$, $\phi_{j,\boldsymbol{x}} = (\phi_j(x_1), \ldots, \phi_j(x_n))^T$ and $\langle f, \phi_j \rangle' = \beta_j \langle f, \phi_j \rangle$ for appropriate $\beta_j \in \mathbb{R}$.*

Eq. (7) generalizes eq. (1) as the Ridge Regression generalizes the Least Squares regression (see [Swindel, 1981] for further details concerning the geometry of ridge regression).

In eq. (3) the projected curve $f^*_{K,\gamma,n}$ is expressed, via the vector $\alpha = (\alpha_1, \ldots, \alpha_n)^T$, as a linear combination in $Span\{K(x, x_i)\}$. In addition in eq. (7) the same curve can be seen as a linear combination of the eigenfunctions of $L_K$. Next theorem introduces a practical manner to estimate this representation, that is the weights $\lambda_j(\alpha^T \phi_{j,\mathbf{x}})$ in eq. (7).

THEOREM 2. *Let $X$ be a compact space or manifold in a Euclidean Space, $Y = \mathbb{R}$ and $\nu$ a Borel probability measure defined on $X \times Y$. Let $f_n = \{(\boldsymbol{x}_i, y_i) \in X \times Y\}_{i=1}^n$ be a sample curve drawn form $\nu$ and consider $f_\nu$ defined in eq. (1). Let $K : X \times X \longrightarrow \mathbb{R}$ be a continuous symmetric positive definite kernel with associated integral operator $L_K$ with eigenfunctions $\{\phi_1, \phi_2, \ldots\}$ and eigenvalues $\{\lambda_1, \lambda_2, \ldots\}$. Then, the projected curve $f^*_{K,\gamma,n}$, given by the minimization of (2), can be expressed as*

$$f^*_{K,\gamma,n}(\boldsymbol{x}) = \sum_j \lambda^*_j \phi_j(\boldsymbol{x}), \tag{8}$$

*where $\lambda^*_j$ are the weights of the projection of $f^*_{K,\gamma,n}(\boldsymbol{x})$ onto the function space generated by the eigenfunctions of $L_K$. In practice, when a finite sample is available, the first $d = rank(K\big|_{\boldsymbol{x}})$ weigths $\lambda^*_j$ can be estimated by*

$$\hat{\lambda}^*_j = \frac{l_j}{\sqrt{n}} (\alpha^T \boldsymbol{v}_j), \tag{9}$$

*for $l_j$ the j-th eigenvalue of the matrix $K\big|_{\boldsymbol{x}}$, $\boldsymbol{v}_j = (v_{j1}, \ldots, v_{jn})^T$, the j-th eigenvector and $\alpha$ the solution to eq. (4).*

Hence two possible finite representations are available for the projection of $f_\nu$ given $f_n$. The first one, in eq. (3) by the vector $\alpha = (\alpha_1, \ldots, \alpha_n)^T$, will be named as"Kernel Expansion". The second, given in eq. (8) by the vector $\hat{\lambda}^* = (\hat{\lambda}^*_1, \ldots, \hat{\lambda}^*_d)^T$ will be denominated as "RKHS representation". Next two remarks compare both representations in terms of their stability in the input variables.

DEFINITION 2. *Let $X$ be a compact space or manifold in a Euclidean Space, $Y = \mathbb{R}$ and $\nu$ a Borel probability measure defined on $X \times Y$. Let $f_n = \{(\boldsymbol{x}_i, y_i) \in X \times Y\}_{i=1}^n$ be a sample curve drawn form $\nu$. We say that $f_n^\epsilon = \{(x_i, y_i^\epsilon)\}_{i=1}^n$ is a $\epsilon$-perturbed curve of $f_n$ if*

$$\frac{|y_i - y_i^\epsilon|}{|y_i|} \leq \epsilon \ \ for \ \ all \ \ i = 1, \ldots, n. \tag{10}$$

DEFINITION 3. *Under the same assumptions as in Definition 2, consider a set of continuos functions $B = \{\varphi, \ldots, \varphi_q\}$ on $X$ where $q \leq n$. Let $f_n$ be a sample curve, $f_\nu$ defined in eq. (1) and $f_n^\epsilon$ an $\epsilon$-perturbed curve of $f_n$. Let $\Pi_{B,n} : L^2_\nu(X) \longrightarrow Span(B)$ be a general curves projection method onto $Span(B)$ using any sample curve of size $n$ and let*

$$\Pi_{B,n}(f_\nu) = \sum_j \beta_j \varphi_j \ \ and \ \ \Pi^\epsilon_{B,n}(f_\nu) = \sum_j \beta_j^\epsilon \varphi_j, \tag{11}$$

*be two projections of $f_\nu$ using $f_n$ and $f_n^\epsilon$ respectively. Then we say that the representation of $f_n$ given by $\beta = (\beta_1, \ldots, \beta_q)^T$ is $\epsilon$-stable in the input variables if*

$$\frac{|\beta_j - \beta_j^\epsilon|}{|\beta_j|} \leq \epsilon \ \ for \ \ all \ \ j = 1, \ldots, q. \tag{12}$$

THEOREM 3. *Under the conditions described in Theorem 2, the representation of $f^*_{K,\gamma,n}$ given by $\hat{\lambda}^* = (\hat{\lambda}_1^*, \ldots, \hat{\lambda}_d^*)^T$, where $\hat{\lambda}_j$ is estimated in eq. (9) and $d = rank(K\big|_{\boldsymbol{x}})$ is $\epsilon$-stable in the input variables.*

THEOREM 4. *Under the conditions described in Theorem 2 the representation of $f^*_{K,\gamma,n}$ in terms of the vector $\alpha = (\alpha_1, \ldots, \alpha_n)^T$, where $\alpha$ is the solution to eq. (4) is not $\epsilon$-stable in the input variables.*

Next we include and illustrative example to show the implications of the two previous theorems in a real example.

EXAMPLE 1. *We consider two similar functional data curves to illustrate the behavior of the Kernel expansion (given in (3)) and the RKHS representation system (given in (8)). The two curves are temperatures curves corresponding to daily series averaged over the period from 1960 to 1994 in Canada ([Ramsay and Silverman, 2006], Chapter 1, and correspond to the cities "St. Johns" and "Halifax". We consider the Gaussian kernel $K(\boldsymbol{x}, \boldsymbol{y}) = e^{-\rho\|x-y\|^2}$ with $\rho = 10^{-4}$ and $\gamma = 1$) and obtain the kernel expansion and the RKHS representation for both curves. In the experimental section we will detail how to fix the pairs of parameters $(\rho, \gamma)$ in a wide variety of experients.*

*Figure 1, left (upper and lower), shows the curves and their projections onto the function space $\mathcal{H}_K$ generated by the eigenfunctions of $K$. The two central plots in Figure 1 show the kernel expansion representation for both curves and it is apparent they are quite different, despite the fact the two curves are similar. Figure 1, right, shows the RKHS representations for both curves and now they look similar, in agreement with Theorem 3. In addition, we*

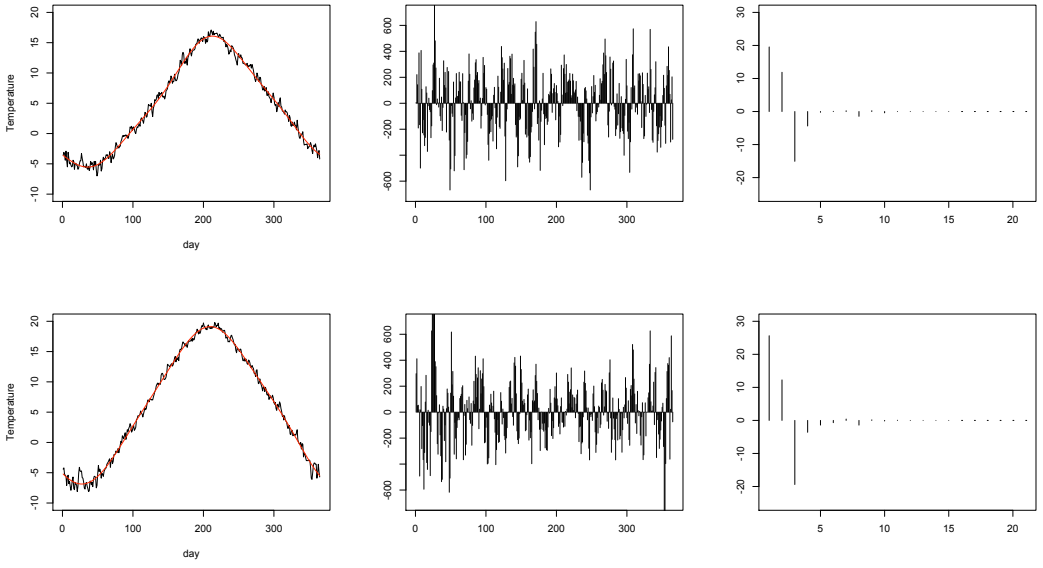



**Fig. 1.** Two Canadian curves, and their Kernel Expansion and RKHS representations.

*can see that the RKHS representations is representing the curves in a no more than a 10-dimensional space (essentilly 4), which agrees with the result obtained by the dimensionality test proposed in [Hall and Vial, 2006] for this set of curves. We can therefore conclude that the RKHS representation is robust against the presence of noise in the data in agreement with Theorem 3.*

## 3. Distance measures induced by the projections of functional data onto RKHSs

In Functional Data Analysis we are generally given a set of curves $\{f_{n,1}, \ldots, f_{n,m}\}$ where each sample curve$f_{n,l}$ is identified with a data set $\{(x_i, y_{il}) \in X \times Y\}_{i=1}^{n}$. In practical cluster and classification problems $n$ is generally very large. This makes the functional data to be not tractable for most algorithms that are commonly designed to work either with (small) finite-dimensional vectors or with distances matrices. In this context, to determine an appropriate distance matrix between the curves (with dimensions $m \times m$ where $m \ll n$) makes the problem solvable in practice.

Several methods have been proposed in the literature to define distances between curves. For instance the Dynamic Time Warping [Sakoe and Chiba, 1978] calculates the dissimilarity between two series by warping them before calcuating its Euclidean distance. Other approach followed in [Ferraty and Vieu, 2006] is to define some semi-metric as measure of similarity for the curves. In any of the previous approaches similarities/disimilarities can be transformed to distances. See [Gower, 1986] for details.

In this section we study the metric for curves induced by the projection defined in eq. (1). The proposed metric will be determined by $K$ and $\gamma$ and we will be the input of classification and cluster procedures. Notice that many kernels can determine the same metric [Burges, 1998] which in practice is not a problem for our purposes.

DEFINITION 4. *Let $X$ be a compact space or manifold in and Euclidean Space, $Y = \mathbb{R}$ and $\nu, \mu$ two Borel probability measures defined on $X \times Y$. Let $f_n = \{(x_i, y_i) \in X \times Y\}_{i=1}^n$ and $g_n = \{(x_i, y_i') \in X \times Y\}_{i=1}^n$ two sample curves drawn form $\nu$ and $\mu$ respectively and let $f_\nu$, $g_\mu$ defined following eq. (1). Let $K : X \times X \to \mathbb{R}$ be a Mercer kernel and $\mathcal{H}_K$ its associated RKHS. Then we define the* ***Empirical Regularized $\gamma$ Inner Product*** *between $f_\nu$, $g_\mu$ as*

$$\langle f_\nu, g_\mu \rangle_{K,\gamma,n} = \langle \Pi_{K,\gamma,n}(f_\nu), \Pi_{K,\gamma,n}(g_\mu) \rangle_K \tag{13}$$

*where given $h_1$ and $h_2$, $\langle h_1, h_2 \rangle_K = \sum_j \lambda_j^{-1} a_j b_j$ for $h_1 \in \sum_j a_j \phi_j \in \mathcal{H}_K$ and $h_2 = \sum_j b_j \phi_j \in \mathcal{H}_K$ being $\{\lambda_j\}$ the eigenvalues of $L_K$.*

Notice that, given a kernel $K$, we define the inner product of $f_\nu$ and $g_\mu$ as the inner product of their projections onto $\mathcal{H}_K$. In practice, using eq. (8) and the definition of $\langle \cdot, \cdot \rangle_K$ it is straightforward to check that an estimator of $\langle f_\nu, g_\mu \rangle_{K,\gamma,n}$ is given by

$$\sum_{j=1}^{n} l_j^{-1} (\hat{\lambda}_{fj}^* \hat{\lambda}_{gj}^*), \tag{14}$$

where $l_j$ is the j-th eigenvalue of $K\big|_{\mathbf{X}}$ and $\hat{\lambda}_{fj}^*, \hat{\lambda}_{gj}^*$, the components of the "RKHS" representation of $f_\nu$ and $g_\mu$, are given by eq. (9).

DEFINITION 5. *Given the elements of Definition 4 we define the* ***Empirical Regularized $\gamma$ Distance*** *for two curves $f_\nu$, $g_\mu$ as*

$$d_{K,\gamma,n}(f_\nu, g_\mu) = \langle f_\nu, g_\mu \rangle_{K,\lambda,n} + \langle f_\nu, g_\mu \rangle_{K,\lambda,n} - 2\langle f_\nu, g_\mu \rangle_{K,\lambda,n} \tag{15}$$

This distance can be estimated by replacing eq. (14) in eq. (15). Hence given a set of curves, the distance defined in eq. (15) can be estimated for each pair of curves obtaining a distance matrix $\mathbf{D}$ that can be used as the input of cluster or classification algorithms.

## 4. Model selection in functional data regularization

A central problem in statistics is the selection of appropriate models for the data. In our context, to select a model for a sample curve $f_n$ means to find appropriate $K$ and $\gamma$ in eq. (2).

A typical manner to afford the model selection problem is to minimize some measure of the predictive error, for instance, the averaged difference between the estimated and the true values of some test points contained in the data: the traditional cross validation (CV), its

generalized version (GCV) [Craven and Wahba, 1979] or the $Cp$ measure [Mallows, 1973] constitute some examples. However the optimality of this approach is not guaranteed since the real generalization capacity of the models is not estimated. Instead, model selection criteria that deals with the generalization error have also been proposed: from the point of view of the information theory the Akaike information criterion (AIC) [Akaike, 1974] and its corrected modification (cAIC) [Sugiura, 1978] are the most representative cases. From the Bayesian perspective the bayesian information criterion (BIC) [Schwarz, 1978] is a well known example. Other approaches different form the two previous points of view are the structural risk minimization (SRM) [Vapnik, 1995] or the Vapnik measure (VM) [Cherkassky et al., 1999].

In [Sugiyama and Ogawa, 2001] the Subspace information Criterion (SIC) is proposed as a new alternative of model selection. It is very competitive [Sugiyama and Muller, 2002] compared to previous measures and it represents a natural framework for model selection in regularization methods. In this section we will particularize it to select the appropriate model in eq. (2).

### *4.1. Model selection problem*

Let $X$ a compact space or manifold, $Y = \mathbb{R}$ and $\nu$ a probability measure over $X \times Y$. Let $f_n$ be a sample curve drawn form $\nu$ identified with a data set $\{(x_i, y_i) \in X \times Y\}_{i=1}^n$ and define the target function $f_\nu : X \to Y$ as $f_\nu = \int_X y d_\nu(y \,|\, x)$ for $\nu(y \,|x)$ the conditional measure on $Y$. In the sequel we will assume that $f_\nu$ belongs to $L^2_\nu(X)$ and that $f_\nu$ is a bounded function.

Define $\epsilon = y - f_\nu(x)$. Then

$$\mathsf{E}_\nu(\epsilon) = \int_Y (f_\nu(x) - y) d_\nu(y \,|\, x), \tag{16}$$

where $\mathsf{E}_\nu$ denote the expectation over the measure $\nu$. It is straightforward to check that $\mathsf{E}_\nu(\epsilon) = 0$ and hence the variance of $\epsilon$ is given by

$$\mathsf{Var}_\nu(\epsilon) = \int_Y (f_\nu(x) - y)^2 d_\nu(y \,|\, x), \tag{17}$$

where $\mathsf{Var}_\nu(\epsilon)$ denotes the variance over $\nu$. Using the definition of $f_\nu$ and because $\mathsf{E}_\nu(\epsilon) = 0$, given the sample points $\{(x_i, y_i) \in X \times Y\}_{i=1}^n$ we have that

$$y_i = f_\nu(x_i) + \epsilon_i, \tag{18}$$

where the $\epsilon_i$ are unknown additive independent noise components from a distribution with zero mean and variance $\mathsf{Var}_\nu(\epsilon)$. Notice that both, $f_\nu$ and $\mathsf{Var}_\nu(\epsilon)$ are totally determined by $\nu$.

Given $f_n$ consider a set of pairs $\{(K, \gamma)\}$, where each $K : X \times X \longrightarrow \mathbb{R}$ is a Mercer kernel function and $\gamma > 0$. This set can be either finite or infinite. In this last case $K$ is commonly defined as a parameter dependent kernel, for instance, a Gaussian kernel.

Let $f^*_{K,\gamma,n}$ be the projected curve obtained via eq. (2) using the sample $f_n$ and some $\gamma$ and $K$. The model selection problem is stated as finding, for a fixed sample curve $f_n$, the pair $(K^*, \gamma^*)$ that minimizes the generalization error defined as

$$\mathsf{E}_\epsilon \left( \int_X (f^*_{K,\gamma,n} - f_\nu)^2 dx \right) = \mathsf{E}\|f^*_{K,\gamma,n} - f_\nu\|^2, \tag{19}$$

where $\mathsf{E}_\epsilon$ denotes the expectation over the noise $\epsilon$. For simplicity in notation in the sequel we will write $\mathsf{E}$ instead of $\mathsf{E}_\epsilon$. Notice that $f^*_{K,\gamma,n}$ belongs to $\mathcal{H}_{K,n} = Span(\{K(x, x_i)\})$ while, in general, it is common to assume that the function $f_\nu$ belongs to $L^2_\nu(X)$.

### 4.2. Subspace Information Criterion (SIC) for functional data regularization

The Subspace Information Criteria (SIC) [Sugiyama and Ogawa, 2001] was proposed as a procedure to give an unbiased estimator of the generalization error in eq. (19) in general regularization methods. In this section we follow the general model selection approach described above and we will describe the SIC, adapting it to our particular problem in eq. (2).

Let $K$ be a Mercer kernel function $L_K$ its associated integral operator and $\mathcal{H}_K$ its corresponding RKHS. We first decompose the target function $f_\nu$ as follows. Let $f_{\nu,\mathcal{H}_K}$ be the orthogonal projection of $f_\nu$ onto $\mathcal{H}_K$ ($f_{\nu,\mathcal{H}_K} = L_K(f_\nu)$, see Proposition 1 for details) and define $f^\perp_{\nu,\mathcal{H}_K}$ as

$$f^\perp_{\nu,\mathcal{H}_K} = f_\nu - f_{\nu,\mathcal{H}_K}, \tag{20}$$

the orthogonal complement of $f_{\nu,\mathcal{H}_K}$. In this context, to define the SIC proposed in [Sugiyama and Muller, 2002] we assume that $f^\perp_{\nu,\mathcal{H}_K} = 0$ or equivalently when $f_\nu$ is assumed to belong to $\mathcal{H}_K$.

We first decompose eq. (19) in a sum. Let $f^*_{K,\gamma} = \mathsf{E}(f^*_{K,\gamma,n})$ (see Proposition 2). Then the generalization error of $f^*_{K,\gamma,n}$ is given by:

$$\begin{aligned} G(f^*_{K,\gamma,n}) &= \mathsf{E}\|f^*_{K,\gamma,n} - f_{\nu,\mathcal{H}_K}\|^2 \\ &= \mathsf{E}\|f^*_{K,\gamma,n} - f^*_{K,\gamma} + f^*_{K,\gamma} - f_{\nu,\mathcal{H}_K}\|^2 \\ &= \mathsf{E}\|f^*_{K,\gamma,n} - f^*_{K,\gamma}\|^2 + \mathsf{E}\|f^*_{K,\gamma} - f_{\nu,\mathcal{H}_K}\|^2 \\ &+ 2\mathsf{E}\langle f^*_{K,\gamma,n} - f^*_{K,\gamma}, f^*_{K,\gamma} - f_{\nu,\mathcal{H}_K}\rangle, \end{aligned}$$

where the last term equals zero since $(f^*_{K,\gamma,n} - f^*_{K,\gamma})$ and $(f^*_{K,\gamma} - f_{\nu,\mathcal{H}_K})$ are orthogonal functions. Therefore $G(f^*_{K,\gamma,n})$ can be decomposed as

$$G(f^*_{K,\gamma,n}) = \mathsf{Var}(f^*_{K,\gamma,n}) + Bias^2(f^*_{K,\gamma,n}, f_{\nu,\mathcal{H}_K}), \tag{21}$$

where

$$\mathsf{Var}(f^*_{K,\gamma,n}) = \mathsf{E}\|f^*_{K,\gamma,n} - f^*_{K,\gamma}\|^2 \tag{22}$$

and

$$Bias^2(f^*_{K,\gamma,n}, f_{\nu,\mathcal{H}_K}) = \mathsf{E}\|f^*_{K,\gamma} - f_{\nu,\mathcal{H}_K}\|^2. \tag{23}$$

Eq. (19) assesses the quality of $f^*_{K,\gamma,n}$ in terms of its bias and variance. In practice the functions $f_{\nu,\mathcal{H}_K}$ and $f_{K,\gamma}$ are obviously unknown and therefore eq. (19) cannot be directly estimated. The key idea of the SIC is to replace $f_{\nu,\mathcal{H}_K}$ by an unbiased estimator $f_u$ ($\mathsf{E}(f_u) = f_{\nu,\mathcal{H}_K}$) to roughly approximate $\mathsf{E}\|f^*_{K,\gamma,n} - f_{\nu,\mathcal{H}_K}\|^2$ by $\mathsf{E}\|f^*_{K,\gamma,n} - f_u\|^2$. Next we introduce a formal definition of the SIC adapted to our problem in eq. (1).

DEFINITION 6. *The Subspace Infomation Criterion of the projected curve $f^*_{K,\gamma,n}$ is defined as*

$$SIC(f^*_{K,\gamma,n}) = \mathsf{E}\|f^*_{K,\gamma,n} - f_u\|^2, \tag{24}$$

*where $f_u = f^*_{K,0,n}$.*

The projection $f_u = f^*_{K,0,n}$ is an unbiased estimator of $f_{\nu,\mathcal{H}_K}$ (see Proposition 1 where, by hypothesis, $f_\nu = f_{\nu,\mathcal{H}_K} = L_K(f_\nu)$). Remark that while $f^*_{K,0,n}$ is estimated using the sample $f_n$ and therefore it is a random variable, $f_u$ in eq. (26) is considered to be a fixed function. In addition, both $f^*_{K,\gamma,n}$ and $f_u$ belong to $\mathcal{H}_{K,n}$ and therefore the properties of the RKHSs can be used to estimate eq. (21) by eq. (24).

Denote by $\mathbf{K}$ the matrix defined by $(\mathbf{K})_{ij} = K(x_i, x_j)$. Let $\alpha = (\alpha_1, \ldots, \alpha_n)^T$ and $\alpha^0 = (\alpha^0_1, \ldots, \alpha^0_n)^T$ the kernel expansion representations (see eq. (3)) of $f^*_{K,\gamma,n}$ and $f^*_{K,0,n}$ respectively. In practice, $\alpha = \mathbf{H}_\gamma \mathbf{y}$ where $\mathbf{H}_\gamma = (\gamma n \mathbf{I}_n + \mathbf{K})^{-1}$ and $\alpha^0 = \mathbf{H}_0 \mathbf{y}$ (that is $\mathbf{H}_0 = \mathbf{K}^{-1}$). When $\mathbf{K}$ is not invertible then $\mathbf{H}_0 = \mathbf{K}^+$ is the Moore-Penrose pseudoinverse of $\mathbf{K}$. Then, it holds that

$$f^*_{K,\gamma,n} = \sum_{i=1}^n \alpha_i K(x, x_i) \;\; and \;\; f_u = \sum_{i=1}^n \alpha^0_i K(x_i, x). \tag{25}$$

Operating from eq. (21) we can rewrite the $SIC(f^*_{K,\gamma,n})$ as

$$SIC(f^*_{K,\gamma,n}) = \mathsf{E}_\epsilon \|\alpha - \mathsf{E}(\alpha)\|^2_{\mathbf{K}} + \|\mathsf{E}_\epsilon \alpha - \alpha^0\|^2_{\mathbf{K}}, \tag{26}$$

where $\|\mathbf{a}\|_{\mathbf{K}} = \mathbf{a}^T \mathbf{K} \mathbf{a}$. See [Sugiyama and Ogawa, 2001, Sugiyama and Muller, 2002] for further details. Notice that the first term estimates the variance of $f^*_{K,\gamma,n}$ while the second estimates its squared bias. In particular the variance term can be calculated as follows:

$$\widehat{Var}(f_{K,\gamma,n}) = \sigma^2 tr(\mathbf{K}\mathbf{H}_\gamma \mathbf{H}^T_\gamma), \tag{27}$$

where following [Wahba, 1990] an estimator of $\sigma^2$ is given by

$$\hat{\sigma}^2 = \frac{\|\mathbf{K}\alpha - \mathbf{y}\|^2}{n - tr(\mathbf{K}\mathbf{H}_\gamma)}. \tag{28}$$

Regarding the bias term, it can be estimated by

$$\widehat{Bias}^2(f_{K,\gamma,n}) = \|\alpha - \alpha^0\|^2_{\mathbf{K}} - \hat{\sigma}^2 tr\left(\mathbf{K}(\mathbf{H}_\gamma - \mathbf{H}_0)(\mathbf{H}_\gamma - \mathbf{H}_0)^T\right), \tag{29}$$

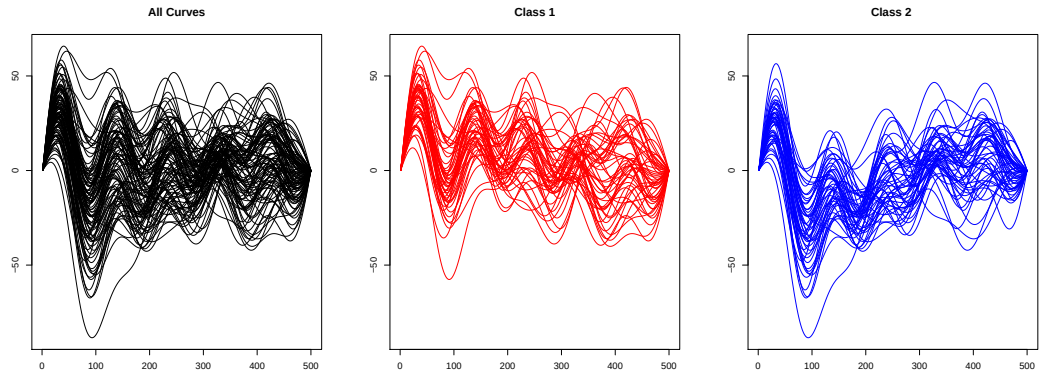



**Fig. 2.** Left: all curves together. Center: Class 1 curves. Right Class 2 curves.

See [Sugiyama and Ogawa, 2001] for details. Finally using eqs. (27) and (29) the SIC can be finally estimated by

$$\begin{aligned} SIC(f_{K,\gamma,n}) &= \|\alpha - \alpha^0\|^2_{\mathbf{K}} - \hat{\sigma}^2 tr\left(\mathbf{K}(\mathbf{H}_\gamma - \mathbf{H}_0)(\mathbf{H}_\gamma - \mathbf{H}_0)^T\right) \\ &+ \hat{\sigma}^2 tr(\mathbf{K}\mathbf{H}_\gamma\mathbf{H}_\gamma^T) \end{aligned}$$

where an estimator of $\sigma^2$ is given in eq. (28).

## 5. Experiments

In this section we apply the previous methodology to several classification and cluster examples. First, we motivate the necessity of using general kernel functions in cases where the use of the data covariance matrix is inappropriate. In the second experiment, we check the performance or our functional data representation with several simulated and real data sets. Finally, we conclude with a real example where the data are a set of temperature series.

### *5.1. RKHS projections versus PCA projections*

In Statistics it is usual to reduce the dimension of high dimensional data before affording cluster or classification tasks. In FDA this is achieved by using the Functional Principal Components (FPCA) [Ramsay and Silverman, 2006, Hall and Vial, 2006]. As in the multivariate case, this technique make use of the data covariance function to determine the subspace in which the data are projected. This subspace is spanned by the data covariance eigenfunctions and it is always a RKHS (see [Rakotomamonjy and Canu, 2005]). Within this setting, FPCA can be considered a particular case of our methodology.

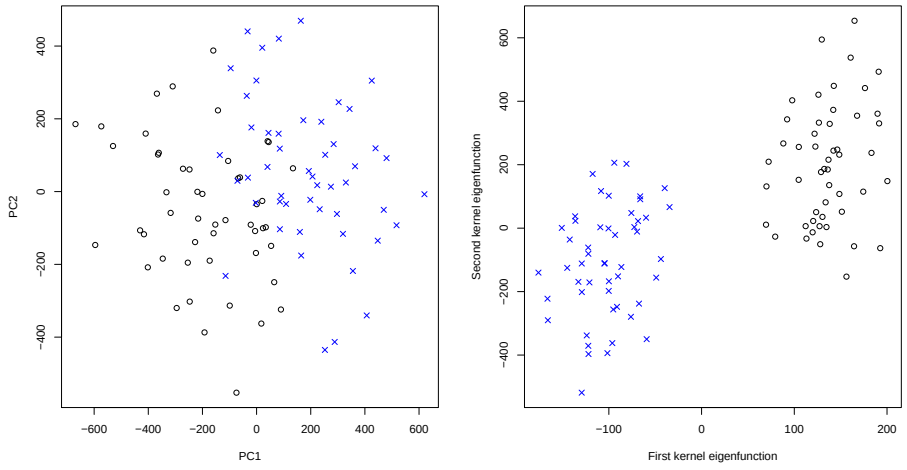



**Fig. 3.** Two first FPCA projection (left) and RKHS projections (right).

To choose the data covariance $S$ as kernel $K$ in eq. (2) is justified in certain theoretical cases (see [James and Sugar, 2003]). In practice, more general kernels can be considered. The next example illustrates this situation in a clustering problem.

Consider two families of 10 dimensional curves sampled at 500 points:

- Class 1: $c(x) = \sum_{j=1}^{10} a_j \phi_j(x) = sin(j\pi x)$, where $a_i \sim N_{10}(\mu_1, \Sigma)$
- Class 2: $c(x) = \sum_{j=1}^{10} b_j \phi_j(x) = sin(j\pi x)$, where $b_j \sim N_{10}(\mu_2, \Sigma)$

with $x \in [0, 1]$ and for $\mu_1 = (8, 8, 1, 2, 3, 4, 5, 6, 7, 8)$, $\mu_2 = (-8, -8, 1, 2, 3, 4, 5, 6, 7, 8)$, and $\Sigma = diag(1, 150, 150, 10, 10, 10, 10, 10, 10, 10)$. For our experiment, we generated 50 curves of each family (see Figure 2).

We compare the RKHS representation system ($\lambda^*$) using the data covariance and a generalized covariance: a Gaussian kernel. To this aim we first separate (automatically) the curves using row the data. We performe 10 runs of a k-means algorithm (with 2 centroids) and a hierarchical cluster by using the Ward method. The misclassification errors are 25.2% and 24% respectively.

By using FPCA, the first two principal components explain over 80% of the variability. This two components are plotted in Figure 3 (left). Applying the two previous cluster procedures over this new projection we obtain misclassification errors of 15% (for the k-means) and 18% (for the hierarchical cluster). The dimension reduction improves the results but a large number of curves is still assigned to wrong families. On the other hand, if the two first

projections are achieved by using regularization with the kernel $K(x,y) = e^{-\rho\|x-y\|^2}$ with $\rho = 10$ and regularization parameter $\gamma = 1$ (see Figure 3, 0% of errors are obtained with both cluster algorithms, what justify the use of a generalized covariance function.

### 5.2. Waveform data

We consider, in this experiment a modified version of the thee class waveform data [Breiman et al., 1984]. In this example we consider 400 predictors for each curve, instead the 21 of the original case. The three classes of the problem are defined by:

$$x(t) = uh_1(t) + (1-u)h_2(t) + \varepsilon(t) \text{ for class 1;}$$

$$x(t) = uh_1(t) + (1-u)h_3(t) + \varepsilon(t) \text{ for class 2;}$$

$$x(t) = uh_2(t) + (1-u)h_3(t) + \varepsilon(t) \text{ for class 3;}$$

for $u$ a uniform random variable on $(0,1)$, $\varepsilon(t)$ standard normal variables, and the $h_k$ function the shifted waveforms for $t \in [1,21]$:

$$h_1(t) = max(6 - |t-11|, 0),\ h_2(t) = h_1(t-4) \text{ and } h_3(t) = h_1(t+4).$$

We generated 1200 observations of the model (400 of each class), and we considered 450 for training the models and 750 as test sample. A plot of the three classes of the problem is shown in Figure 4. The objective of this example is to illustrate that an effective foregoing reduction of the dimension of the curves (projecting them onto certain RKHSs via eq. (2)) improves the classification errors of a variety of classification algorithms compared to the case when they trained over the raw data. To this aim we consider the following classification procedures:

- SVM, the Support Vector Machines [Boser et al., 1992, Moguerza and Muñoz, 2006]. In our experiments we use the linear kernel and we fix regularization parameter $C = 100$.
- FDA, the Flexible Discriminant Analysis [Hastie et al., 1994]. We use two variants: FDA/BRUTO which is based on Additive Models and spline smoothing parameters and FDA/MARS which make use of the Multivariate Adaptative Regression Splines [Friedman, 1991].
- PLSR/LDA, classification method described in [Boulesteix, 2004] which consists in Partial Least Squares dimension reduction and Linear Discriminant Analysis applied on the PLS components.

We consider five different RHKSs where project the data. First, we use two basis of functions, both of dimension 10, to construct two kernel functions via eq. (31): one of P-splines and other of B-splines (see [Pearce and Wand, 2006] to see how to construct kernels form basis of splines). We also consider the data covariance function and two generalized covariance functions: a Gaussian kernel given by $K(\mathbf{x},\mathbf{y}) = \exp\{-\rho\|\mathbf{x}-\mathbf{y}\|^2\}$ and a Laplace kernel

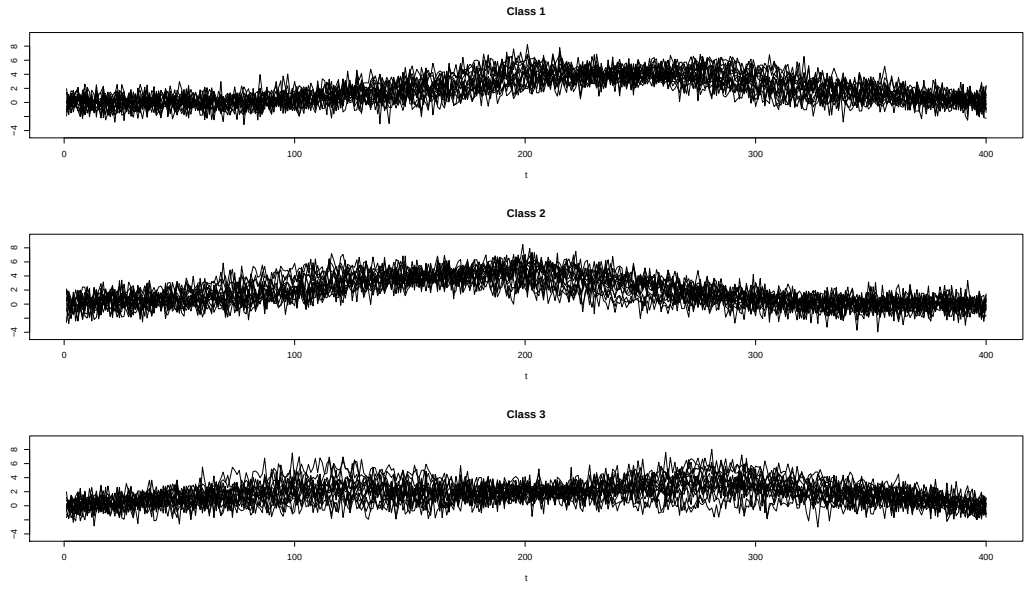



**Fig. 4.** Three classes of the waveform data set.

$K(\mathbf{x}, \mathbf{y}) = \exp\{-\rho\|\mathbf{x} - \mathbf{y}\|\}$ where, in both cases, $\rho = 1$. We project the data onto the RKHSs induced by the previous kernels using eq. (2) for $\gamma = 10^{-3}$.

We classify the curves applying the four classification procedures, $SVM$, $FDA_{bruto}$, $FDA_{mars}$ and $LDA/PLS$ described above using the five estimated projections and also using the raw data. In Table 1 we show the final averaged errors after 100 runs of the experiment. In the projections we decide the number of components to retain by cross validations over the errors. This means that the errors in the table are selected as the best classification result when the only first $d$ eigenfunctions of the proposed kernels for $d = 1, \ldots, 10$ are taken into account to represent the curves.

Results are shown in Table 1. It is clear that reduce the dimension of the curves by projecting them onto the proposed RKHSs always improves significantly the classification errors of the techniques compared the situation in which the raw data are used. To illustrate better these differences we include Figure 5 where the confidence intervals of the errors are shown. The best algorithm-projection combination is a Laplace kernel with the $FDA_{bruto}$ algorithm. It is also remarkable the good performance of the SVM trained over the raw data compared with the rest of the techniques.

### *5.3. Real classification examples*

In this section we analyze three real data sets of functional data:

**Table 1.** Comparative of the the averaged errors for the four classifications algorithms and 5 curves representations (+ the raw data) in the Waveform data. In italic letters the best technique of each row is remarked. In bold letters the best technique of each table. Results are obtained after 100 runs.

| Method | B-Splines | P- Splines | Cov. | RBF | Laplace | Raw data |
|---|---|---|---|---|---|---|
| $SVM$ | 0.0491 (0.0011) | 0.0485 (0.0013) | 0.0519 (0.0012) | 0.0470 (0.0012) | *0.0467 (0.0012)* | 0.0628 (0.0014) |
| $FDA_{bruto}$ | 0.0293 (0.0010) | 0.0353 (0.0010) | 0.0313 (0.0009) | 0.0289 (0.0010) | ***0.0288 (0.0010)*** | 0.0839 (0.0017) |
| $FDA_{mars}$ | 0.0399 (0.0014) | 0.0362 (0.0013) | 0.0413 (0.0014) | 0.0449 (0.0014) | *0.0395 (0.0013)* | 0.1091 (0.0020) |
| $LDA/PLS$ | 0.0610 (0.0017) | 0.0665 (0.0018) | *0.0606 (0.0019)* | 0.0612 (0.0018) | 0.0613 (0.0018) | 0.1675 (0.0026) |

- *Growth data*: This data set consists on 93 growth curves for a sample of 54 boys and 39 girls [Ramsay and Silverman, 2006] (see Figure 6). The observations were measured at a set of twenty nine ages from one to eighteen years old. The data were originally smoothed by using a spline basis.

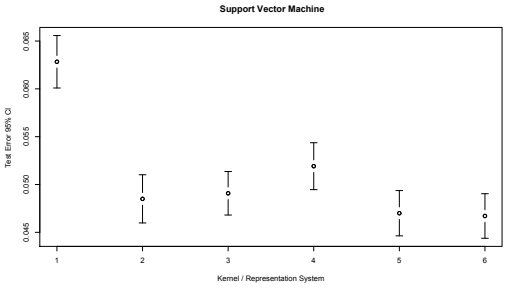
- *Phoneme data*: The third data set correspond to 800 discretized log-periodograms of the phonemes "aa" and "ao". Each phoneme is associted with a class of the experiment. A plot of 25 series of each class is shown in Figure 7.

- *Spectrometric data*. This data set is made of 215 observation is the near infrared absorbance spectrum of a meat sample. Each observation consists in a 100 channel spectrum of absorbance in the wavelength range from 850 to 1050 nm recorded on a Tecator Infratec Food and Feed Analyzer. The two classes are determined by those samples with more (class 1) or less (class 2) than a 20% of fat content. In Figure 8 we show the original curves.

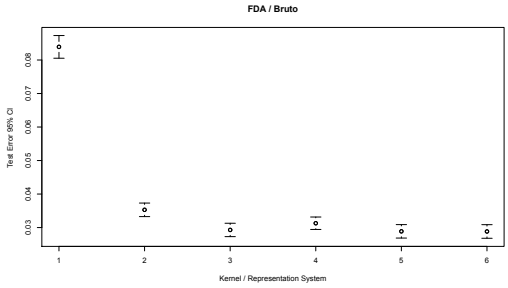
To test our methodology, we follow the the same comparative scheme used in the previous section. However, in this case we optimize the parameters of the Gaussian and Laplace kernels by means of the SIC described in Section 4.2. We fix the penalization parameter $\gamma = 10^{-3}$ and we search the $\rho$ parameter value (in both kernels) in a grid of 100 values in the interval $[10^{-4}, 10^{-1}]$. The optimal $\rho^*$ is fixed as the value that minimizes the avaraged SIC for each set of sample curves $\{f_{n,1}, \ldots, f_{n,m}\}$. Denote by $f^{*l}_{K_{\rho_i},\gamma,n}$ the projection of $f_\nu$ onto the RKHS associated to the parameter dependent kernel $K_{\rho_i}$ (Gausian or Laplace in this example) using $f_{n,l}$. Then the optimal $\rho^*$ is given by

$$\rho^* = arg\min_{\rho_i} \frac{1}{m} \sum_{l=1}^{m} SIC(f^{*l}_{K_{\rho_i},\gamma,n}) \ for \ i = 1, \ldots, 100, \tag{30}$$

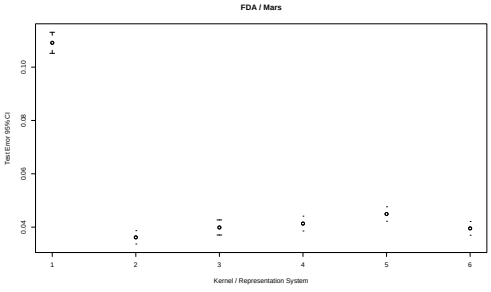
See eq. (30) for details.



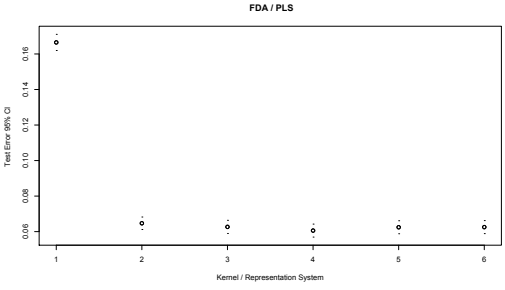

(a) Support Vector Machine




(b) Flexible Discriminant Analysis, bruto.




(c) Flexible Discriminant Analysis, mars.




(d) Linear Discriminant Analysis, PLS

**Fig. 5.** Confidence Intervals (95%) for the errors of the 5 representation and the row data in four classification techniques. The representation systems are: (1) Raw data, (2) B-splines, (3) P-splines, (4) Data covariance, (5) Gaussian Kernel and (6) Laplace Kernel.

Results are shown in Table 2. In agreement with the previous experiment, to project the curves onto the proposed RKHSs improves the results achieved by the classification procedures using the raw data. Just one exception appears, the Phoneme data using the LDA/PLS procure where non effective improvement is obtained. The best projection has an error of 19.35% misclassified curves (using $S$) while the error for the raw data is 19.13%.

Regarding the Growth data, the best result corresponds to the LDA/PLS technique combined with a P-splines kernel. It is remarkable that for this data set the projection using the P-Splines kernels achieve the minimum error in the four classification procedures.

The Support Vector Machine combined with Laplace and Gaussian kernels obtains the lower errors in the Spectrometric and Phoneme data. This is a clear example of how the use

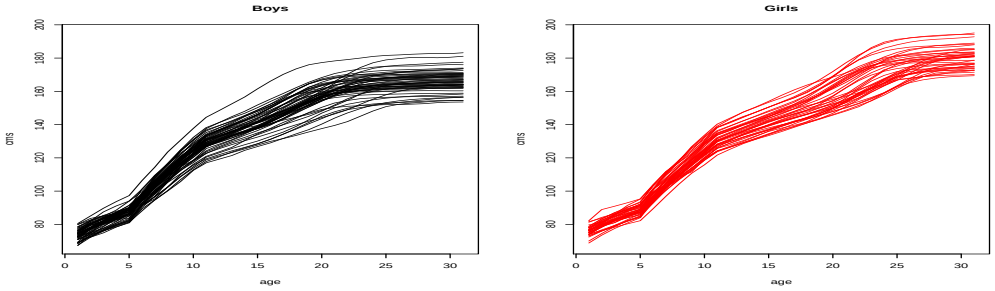



**Fig. 6.** Growth data.

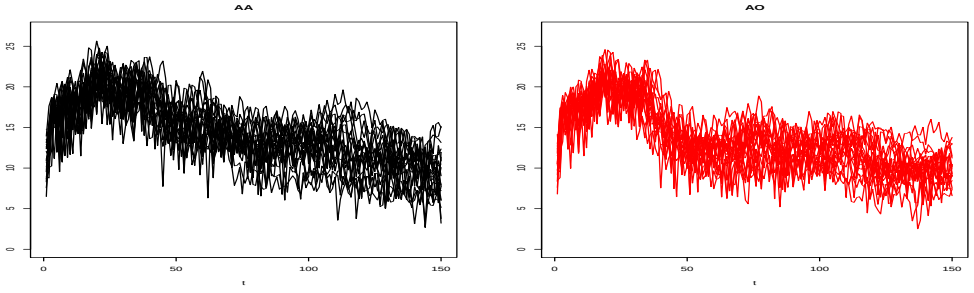



**Fig. 7.** Phoneme curves by classes. The projection of the curves onto the two first supervised Fisher discriminant components is also shown. There is a clear overlapping of the classes.

of generalized covariance functions is useful to improve the classical FDA approach (that focuses on specific basis of functions instead of generalized covariances).

To conclude the analysis we check the accuracy of the previous results by comparing the errors in Table 2 with those achieved by two techniques specificaly designed to deal with functional data:

- The P-spline Signal Regression (PSR), developed by Marx and Eilers [Marx and Eilers, 1999].
- The Non Parametric Curves Discrimination (NPCD) developoed by Ferraty and Vieu [Ferraty and Vieu, 2003]. This procedure uses a semi-metric to obtain the distance between the curves. We select the optimal metric between a set of alternatives. In particular, we consider the Partial Least Squares (for a number of components fixed by cross validation for $p = 1, \ldots, 10$) and the derivative semi-metrics ($d_2$).

In Table 3 we compare the best results from Table 2 (for each data base) with the results obtained by previous techniques. It is clear that we are able to outperform their classification

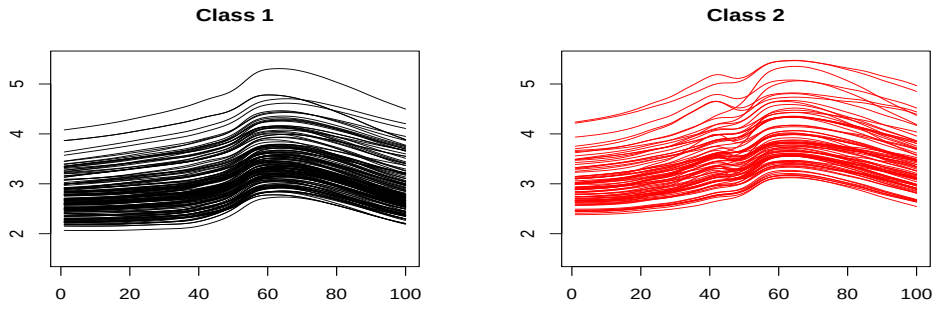



**Fig. 8.** Spectrometric data.

errors in the three cases specially for the growth data set. In this case the PSR misclassified the 5.21% of the curves, the MPLSR with a derivative semi-metric the 4.49% while we obtain an error of 1.16% using the LDA/PLS procedure combined with the projection induced by the P-splines kernel.

### *5.4. Cluster of temperature series and model selection criteria*

In this example we analyze the whole set of temperature curves described in Example 1. See Figure 9 a). The objective is to find the hidden cluster structure of the curves and to study it in terms of climate regions in Canada. To this aim we proceed in two steps: (1) we project the time series onto certain RKHS and (2) we apply a cluster procedure over the projections.

To select the RKHS where project the curves we use the SIC criteria described in Section 4.2. We optimize the parameter $\rho$ of the Gaussian kernel from a set of 50 equally spaced values in the interval $[10^{-7}, 10^{-1}]$ and we fix $\gamma = 1$.

In this case the value of $\rho$ that minimizes the averaged SIC for the set of series is 0.0791. See Figure 9 b). We project the series using a Gaussian kernel with this parameter and we apply a hierarchical cluster method over the projections (Ward method). Using a priori information about the climate in Canada (www.nrcan.gc.ca), we know that in this country there exist four climate zones (see Figure 11 b). Therefore we decide to retain 4 clusters. The series corresponding to each one of the obtained clusters are drawn in Figure 10. In addition, in Figure 11 a) we show the location of each series and we point out the clusters they belong to. The cities assigned to each cluster are detailed as follows:

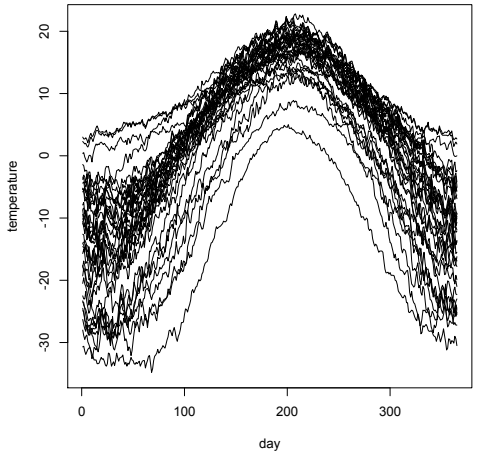
- Zone A (Cirles with horizontal line): Scheffervll, Churchill, Uranium, Cty. Dawson, Yellowknife, Iqaluit, Inuvik, Resolute.

**Table 2.** Comparative of the the averaged errors for the four classifications algorithms and 5 curves representations (+ the raw data) in the three real data sets. In italic letters the best technique of each row is remarked. In bold letters the best technique of each table. Results are obtained after 100 runs.

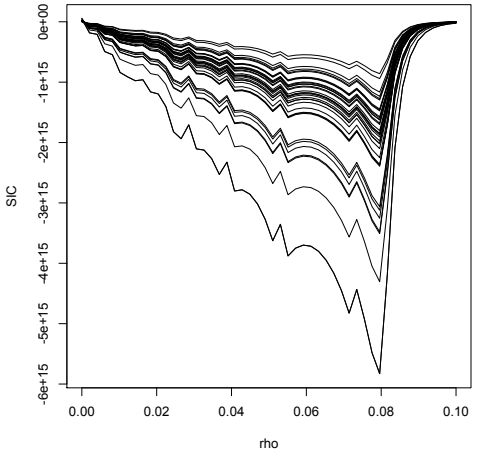
| **Growth data** Method/RKHS | B-Splines | P- Splines | Cov. | RBF | Laplace | Raw data |
|---|---|---|---|---|---|---|
| $SVM$ | 0.0600 (0.0075) | *0.0158 (0.0042)* | 0.0326 (0.0052) | 0.0568 (0.0071) | 0.0400 (0.0059) | 0.0811 (0.0076) |
| $FDA_{bruto}$ | 0.0368 (0.0055) | *0.0316 (0.0048)* | 0.0347 (0.0050) | 0.0368 (0.0056) | 0.0516 (0.0054) | 0.3695 (0.0163) |
| $FDA_{mars}$ | 0.0463 (0.0058) | *0.0442 (0.0049)* | 0.0579 (0.0062) | 0.0484 (0.0058) | 0.0684 (0.0066) | 0.0832 (0.0084) |
| $LDA/PLS$ | 0.0200 (0.0048) | **0.0116 (0.0042)** | 0.0211 (0.0040) | 0.0200 (0.0048) | 0.0305 (0.0047) | 0.0379 (0.0056) |
| **Spectrometric data** Method /RKHS | B-Splines | P- Splines | Cov. | RBF | Laplace | Raw data |
| $SVM$ | 0.0179 (0.0025) | 0.0833 (0.0051) | 0.0162 (0.0027) | 0.0183 (0.0025) | **0.0154 (0.0024)** | 0.0200 (0.0023) |
| $FDA_{bruto}$ | 0.0675 (0.0079) | 0.0600 (0.0043) | 0.0621 (0.0090) | *0.0571 (0.0079)* | 0.0617 (0.0096) | 0.2979 (0.0176) |
| $FDA_{mars}$ | 0.0371 (0.0038) | 0.0554 (0.0043) | *0.0296 (0.0030)* | 0.0358 (0.0031) | 0.0325 (0.0031) | 0.0671 (0.0052) |
| $LDA/PLS$ | 0.0896 (0.0053) | 0.0925 (0.0061) | 0.0871 (0.0049) | 0.1075 (0.0055) | 0.0879 (0.0060) | *0.0762 (0.0070)* |
| **Phoneme data** Method /RKHS | B-Splines | P- Splines | Cov. | RBF | Laplace | Raw data |
| $SVM$ | 0.1835 (0.0036) | 0.1842 (0.0033) | 0.1924 (0.0035) | **0.1814 (0.0036)** | 0.1830 (0.0036) | 0.2328 (0.0053) |
| $FDA_{bruto}$ | 0.1867 (0.0036) | 0.1849 (0.0037) | 0.1958 (0.0034) | *0.1831* (0.0035) | 0.1872 *(0.0034)* | 0.2087 (0.0037) |
| $FDA_{mars}$ | 0.1926 (0.0036) | 0.2019 (0.0038) | 0.2041 (0.0039) | *0.1918* (0.0033) | 0.1964 *(0.0034)* | 0.2695 (0.0050) |
| $LDA/PLS$ | 0.1990 (0.0039) | 0.2263 (0.0036) | 0.1935 (0.0035) | 0.1966 (0.0037) | 0.2006 (0.0037) | *0.1913 (0.0039)* |

**Table 3.** Comparative for the Growth data. For the NPCD method only the best results among the set of tested semi-metrics is shown. In bold, the best results for each data set is remarked.

| **Growth data** | $PSR$ | $NPCD_{d^2}$ | Best Regularization |
|---|---|---|---|
| Test Error | 0.0521 | 0.0494 | **0.0116** |
| | (0.0045) | (0.0400) | **(0.0042)** |
| **Tecator data** | $PSR$ | $NPCD_{d^2}$ | Best Regularization |
| Test Error | 0.0736 | 0.0218 | **0.0154** |
| | (0.0039) | (0.0021) | **(0.0027)** |
| **Phoneme data** | $PSR$ | $NPCD_{p=5}$ | Best Regularization |
| Test Error | 0.1866 | 0.1928 | **0.1814** |
| | (0.0085) | (0.0031) | **0.0036** |




(a) Series of averaged daily temperature in 35 cities of Canada




(b) Value of the SIC for the 35 projected series using a Gaussian kernel for different values of $\rho$.

**Fig. 9.** Set of temperature series and results of the MSC for different values of $\rho$.

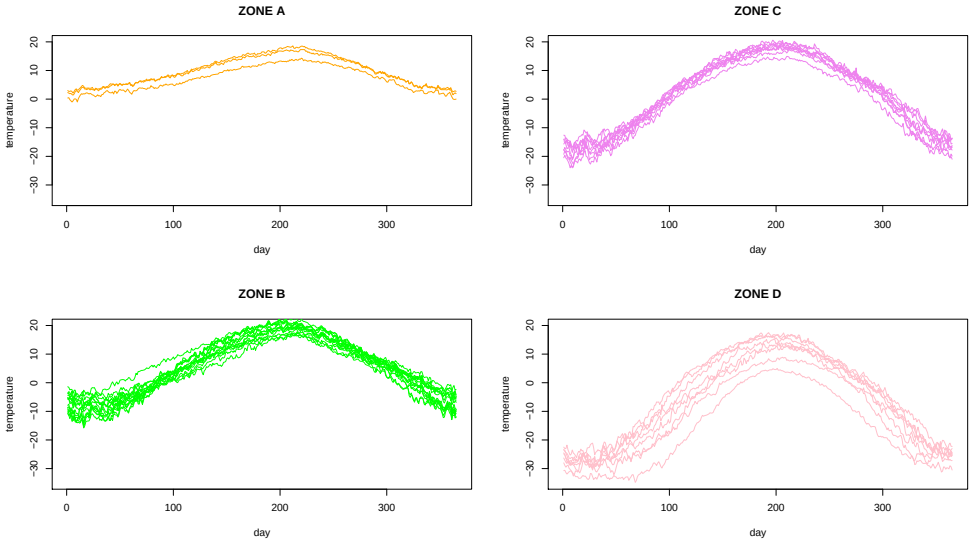



**Fig. 10.** Clusters of the Canadian temperature data set.

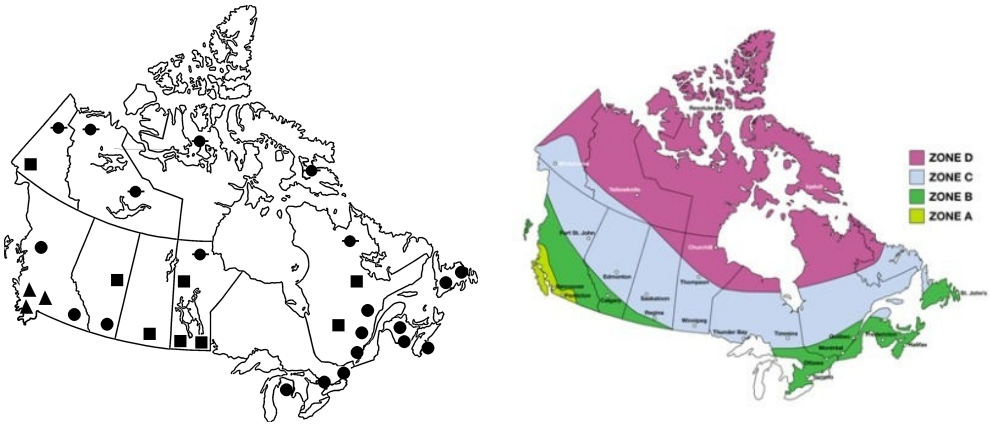
- Zone B (Squares): Arvida, Bagottville, Thunderbay, Winnipeg, The Pas, Regina, Pr. Albert, Edmonton, Whitehorse.
- Zone C (Triangles): Vancouver, Victoria, Pr. Rupert.
- Zone D (Circles): St. Johns, Halifax, Sydney, Yarmouth, Charlottvl, Fredericton, Quebec, Sherbrooke, Montreal, Ottawa, Toronto, London, Calgary, Kamloops, Pr. George.

As can be seen in Figure 10, the 4 zones are perfectly discovered existing a perfect identification between the four clusters and the four climate regions.

To validate the effectiveness of the selected projection we repeat the previous analysis with a different value of $\rho$ within the interval $[10^{-7}, 10^{-1}]$. In particular we fix $\rho = 10^{-5}$ and we show in Figure 12 the new clusters obtained this way. It is clear that the four climate regions are not properly revealed showing the utility of our methodology.

## 6. Conclusions

In this work we have proposed a methodology to represent functional data via their projections onto Reproducing Kernel Hilbert spaces with the aid of Regularization theory. Two representation systems for functional data naturally appear: the RKHS and the Kernel expansion representation. In Theorems 3 and 4 we have studied their stability properties




(a) Stations and clusters (b) Climate models

**Fig. 11.** Map of the stations locations and map of the four spacial climate models in Canada. Image Source: Office of Energy Efficiency Canada

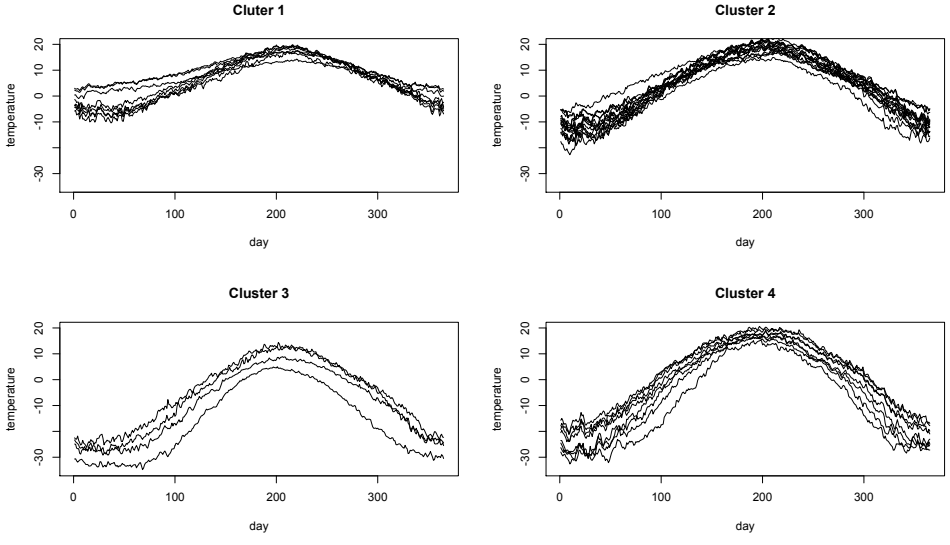



**Fig. 12.** Cluster analysis of the Canadian temperature data set using the Ward method over the obtained projections for $\rho = 10^{-5}$.

concluding that the RKHS Representation (Theorem 2), in contrast to the Kernel Expansion, is $\epsilon$-stable in the input variables and therefore adequate to represent functional data. In addition the RKHS Representation allows to evaluate the dimension of the curves (Example 1) and it enables to reinterpret the regularization process like a curve projection mechanism (Propositions 1 and 2).

Another contribution of this work is the generalization of the classical FDA representation techniques. Any orthogonal basis $B = \{\varphi_1, \ldots, \varphi_d\}$ for $d \in \mathbb{N}$ of continuos functions on $X$, for instance B-splines, fourier basis, P-splines, defines a kernel (and therefore an RKHS) given by

$$K(x, y) = \sum_{j=1}^{p} \varphi_j(x)\varphi_j(y). \tag{31}$$

See [Rakotomamonjy and Canu, 2005] for details. However, in this paper we have shown how to select generalized covariance functions appropriate for functional data and how it is possible to work directly with their eigenfunctions (basis of the RKHS). This makes accessible a larger class of basis of functions where represent the functional data, constituting this methodology a generalization of the classical FDA formalism.

Regarding future work, we want to investigate the choice of kernels appropriate for preespecified tasks or data sets. The idea is to specify objective functions in terms of distance criteria (as it happens, for instance, for principal component analysis). Given the direct relationship existing between kernel functions and distance functions, this gives as a methodology to specify optimal kernels in advance and to obtain, in consequence, optimal representation systems for given tasks.

## A. Proofs

Proof (Proof Proposition 1). First, operating from eq. (3), we have that

$$\begin{aligned} f^*_{K,0,n} = \Pi_{K,0,n}(f_\nu) &= \sum_{i=1}^{n} \alpha_i K(x_i, x) = \sum_{i=1}^{n} \alpha_i \left( \sum_{j=1} \lambda_j \phi_j(x_i)\phi_j(x) \right) \\ &= \sum_{j=1} \lambda_j \left( \sum_{i=1}^{n} \alpha_i \phi_j(x_i) \right) \phi_j(x) \\ &= \sum_{j=1} \lambda_j (\alpha^T \phi_{j,\mathbf{x}})\phi_j(x), \end{aligned} \tag{32}$$

To check the uniform convergence of $\Pi_{K,0,n}(f_\nu)$ to $L_K(f_\nu)$ we have to prove that for every $\epsilon > 0$ there exists a $N \in \mathbb{N}$ such that for all $x \in X$ and all $n \leq N$, then $|f^*_{K,0,n}(x) - L_K(f_\nu)(x)| < \epsilon$. To this aim, consider the sequence

$$a_n = \sup |f^*_{K,0,n}(x) - L_K(f_\nu)(x)|, \tag{33}$$

where the supremum is taken over all $x \in X$. Then $\Pi_{K,0,n}(f_\nu)$ converges to $L_K(f_\nu)$ uniformly if and only if $a_n$ goes to 0 when $n \longrightarrow \infty$.

Let $f_{\nu,\mathcal{H}_K} = L_K(f_\nu)$ be the orthogonal projection of $f_\nu$ onto $\mathcal{H}_K$. By the spectral theorem

$$f_{\nu,\mathcal{H}_K} = L_K(f_\nu) = \sum_j \lambda_j \langle f_\nu, \phi_j \rangle \phi_j, \tag{34}$$

When $n \longrightarrow \infty$ the problem in eq. (2) tends to

$$f^*_{K,\gamma} = \Pi_{K,\gamma,\infty}(f) = \arg\min_{f \in \mathcal{H}_K} \int_{X \times Y} (y - f(x))^2 d_\nu(x,y) + \gamma \|f\|^2_K \,, \tag{35}$$

which unique minimizer [Cucker and Smale, 2001] is given by

$$f^*_{K,\gamma} = (Id + \gamma L_K)^{-1} f_{\nu,\mathcal{H}_K}. \tag{36}$$

Since $\gamma = 0$, is direct to see (from eq. (36)) that $f^*_{K,0} = f_{\nu,\mathcal{H}_K}$. Then when $n \longrightarrow \infty$, $f^*_{K,0}$ the unique solution to eq. (2) tends to $f_{\nu,\mathcal{H}_K}$ the unique solution of eq. (35) and therefore $a_n \longrightarrow 0$. Then

$$\Pi_{K,0,n}(f) = \sum_j \lambda_j (\alpha^T \phi_{j,\mathbf{x}}) \phi_j \underset{n \to \infty}{\longrightarrow} L_K(f) = \sum_j \lambda_j \langle f, \phi_j \rangle \phi_j \tag{37}$$

uniformly in $X$, what concludes the proof.

Proof (Proof Proposition 2). By Proposition 1 we now that, for $\alpha$ the solution to eq. (4), then $\Pi_{K,\gamma,n}(f) = \sum_j \lambda_j (\alpha^T \phi_{j,\mathbf{x}}) \phi_j$. In addition the unique solution for problem in eq. (2) when $n \longrightarrow \infty$ is given by $f^*_{K,\gamma} = (Id + \gamma L_K)^{-1} f_{\nu,\mathcal{H}_K}$ for $f_{\nu,\mathcal{H}_K} = L_K(f_\nu)$ the orthogonal projection of $f_\nu$ onto $\mathcal{H}_K$.

Since $f^*_{K,\gamma} \in H_K$ the we can write $f^*_{K,\gamma} = \sum_j \beta'_j \phi_j$ for appropriate $\beta' \in \mathbb{R}$ and for $\phi_1, \phi_2 \ldots$ the eigenfunctions of $K$. Without loss of generally we can rewrite $f^*_{K,\gamma}$ as

$$f^*_{K,\gamma} = \sum_j \lambda_j \beta_j \langle f_\nu, \phi_j \rangle \phi_j \tag{38}$$

since $\langle \cdot, \cdot \rangle$ is well defined and $\lambda_j$, the eigenvalues of $L_K$, are all real. Denote $\beta_j = {\beta'}_j (\lambda_j \langle f, \phi_j \rangle)^{-1}$ and define $\langle f, \phi_j \rangle'$ such that $\langle f, \phi_j \rangle' = \beta_j \langle f, \phi_j \rangle$. Then we have that $f^*_{K,\gamma} = \sum_j \lambda_j \langle f, \phi_j \rangle' \phi_j$.

To end the proof, we only have to check the uniform convergence in $X$ of $f^*_{K,\gamma,n}$ to $f^*_{K,\gamma}$. Following the same reasoning that in proof 1 we define the sequence

$$b_n = \sup |f^*_{K,\gamma,n}(x) - f^*_{K,\gamma}(x)|, \tag{39}$$

where the supremum is taken over all $x \in X$. Then $b_n$ goes to 0 when $n \longrightarrow \infty$ by the same reason that $a_n$ goes to 0 in proof 1 and therefore

$$f^*_{K,\gamma,n} = \sum_j \lambda_j(\alpha^T\phi_{j,\mathbf{X}})\phi_j \underset{n\to\infty}{\longrightarrow} f^*_{K,\gamma} = \sum_j \lambda_j\langle f,\phi_j\rangle'\phi_j, \tag{40}$$

uniformly in $X$ what concludes the proof.

Proof (Proof Theorem 2). Operating from eq. (32)

$$f^*_{K,\gamma,n}(\mathbf{x}) = \sum_{j=1} \lambda_j(\alpha^T\phi_{j,\mathbf{X}})\phi_j(\mathbf{x}) = \sum_{j=1} \lambda^*_j\phi_j(\mathbf{x}), \tag{41}$$

for $\lambda^*_j = \lambda_j(\alpha^T\phi_{j,\mathbf{X}})$.

Following [Smale and Zhou, 2007] the eigenvalues and eigenvectors of $K\big|_{\mathbf{X}}/n$ converge, to the eigenvalues and eigenfunctions of $L_K$. In addition each $\phi_j(x_i)$ and $\lambda_j$ can be estimated by $\sqrt{n}\mathbf{v}_{ji}$ and $\hat{\lambda}_j = l_j/n$ respectively. Therefore replacing in $\lambda^*_j = \lambda_j(\alpha^T\phi_{j,\mathbf{X}})$ each $\lambda_j$ and $\phi_j(x_i)$ by its estimators

$$\hat{\lambda}^*_j = \hat{\lambda}_j(\alpha^T\hat{\phi}_{j,\mathbf{X}}) = \frac{l_j}{n}(\alpha^T\sqrt{n}\mathbf{v}_j) = \frac{l_j}{\sqrt{n}}\alpha^T\mathbf{v}_j \tag{42}$$

what concludes the proof.

Proof (Proof Theorem 3). Consider a sample curve $f_n$ and an $\epsilon$-perturbed curve $f^\epsilon_n \equiv \{(x_i, y^\epsilon_i) \in X \times Y\}^n_{i=1}$. Then $f^*_{K,\gamma,n}(\mathbf{x}) \simeq f^{*\epsilon}_{K,\gamma,n}(\mathbf{x})$ and given that the $\phi_j$ are a basis for $\mathcal{H}_K$, it must happen that $\lambda^*_j \simeq \lambda^{*\epsilon}_j$ and therefore $\hat{\lambda}^*_j \simeq \hat{\lambda}^{*\epsilon}_j$. Hence

$$\frac{|\hat{\lambda}^*_j - \hat{\lambda}^{*\epsilon}_j|}{|\hat{\lambda}_j|} \leq \epsilon, \tag{43}$$

for $j = 1, \ldots, d$ and the representation system is $\epsilon$-stable. Notice that the truth of this statement relies in the fact that the eigenvalues and eigenvectors of $K\big|_{\mathbf{X}}$ converge, respectively, to the eigenvalues and eigenfunctions of $L_K$ and therefore $\hat{\lambda}^*_j \longrightarrow \lambda^*_j$. See Theorem 3 for details.

Proof (Proof Theorem 4). By Theorem 3 we know that we can write $f^*_{K,\gamma,n}(x) = \sum_{j=1} \lambda_j(\alpha^T\phi_{j,\mathbf{X}})\phi_j(x)$. In addition, since $\{\phi_j\}$ is a basis for $\mathcal{H}_K$, then $\alpha^T\phi_{j,\mathbf{X}} \longrightarrow \langle f_\nu, \phi_j\rangle$ (see Theorem 3). Therefore, for any set $\alpha' = (\alpha_1{}', \ldots, \alpha_n{}')^T$ such that $(\alpha')^T\phi_{j,\mathbf{X}} \longrightarrow \langle f_\nu, \phi_j\rangle$ we will have that $\sum^n_{i=1} \alpha^{*\prime}_i K(x_i, x) = f^*_{K,\gamma,n}(x)$. Now, given the sample curve $f_n \equiv \{(x_i, y_i) \in X \times Y\}^n_{i=1}$, consider an $\epsilon$-perturbed curve $f^\epsilon_n \equiv \{(x_i, y^\epsilon_i) \in X \times Y\}^n_{i=1}$, such that

$$\frac{|y_i - y^\epsilon_i|}{|y_i|} \leq \epsilon, \tag{44}$$

Denote by $(\alpha^\epsilon)$ the representation corresponding to $f_n^\epsilon$. Given that $f^{*\epsilon}_{K,\gamma,n}(x) \simeq f^*_{K,\gamma,n}(x)$ (because of the continuity of $f_\nu$), it will happen that $(\alpha^\epsilon)^T \phi_{j,\mathbf{X}} \simeq \alpha^T, \phi_{j,\mathbf{X}^\epsilon}$ and, nevertheless, by the previous reasoning, $\alpha^\epsilon$ and $\alpha$ can be quite different. Therefore is not guaranteed that

$$\frac{|\alpha_i - \alpha_i^\epsilon|}{|\alpha_i|} \leq \epsilon. \tag{45}$$

for all $i = 1, \ldots, n$ and therefore the representation is not $\epsilon$-stable.

## References


H. Akaike. A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19(6):716–723, 1974.

N. Aroszajn. Theory of reproducing kernels. *Transactions of the American Mathematical Society*, 68(3):337–404, 1950.

B. E. Boser, I. Guyon, and V. Vapnik. A training algorithm for optimal margin classifiers. *Proc. Fifth ACM Workshop on Computational Learning Theory (COLT) ACM Press, New York*, pages 144–152, 1992.

A. L. Boulesteix. Pls dimension reduction for classification with microarray data. *Statistical Applications in Genetics and Molecular Biology*, 3(1):article 33, 2004.

O. Bousquet and A. Elisseeff. Stability and generalization. *Journal of Machine Learning Research*, 2:499–526, 2002.

L. Breiman, J. Friedman, R. Olshen, and C. Stone. Classification and regression trees. *Wadsworth, Belmont, CA*, 1984.

C. Burges. Geometry and invariance in kernel based methods. *Advances in Kernel Methods. Support Vector Learning. MIT Press Cambridge USA*, 1998.

Z. Chen and S. Haykin. On different facets of regularization theory. *Neural Computation*, 14:2791–2846, 2002.

V. Cherkassky and Y. Ma. Practical selection of svm parameters and noise estimation for svm regression. *Neural Networks*, 17:113–126, 2004.

V. Cherkassky, X. Shao, F. M. Mulier, and V. N. Vapnik. Model complexity control for regression using vc generalization bounds. *IEEE Transactions on Neural Networks*, 10(5): 1075–1089, 1999.

F. Conway. A course in functional analysis. *Springer-Verlag*, 1990.

P. Craven and G. Wahba. Smoothing noisy data with spline functions: Estimating the correct degree of smoothing by the method of geeralized cross validation. *Numerische Mathemetik*, 31:377–403, 1979.

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

F. Ferraty and P. Vieu. Curves discrimination: a nonparametric functional approach. *Computational Statistics & Data Analysis*, 44:161–173, 2003.

F Ferraty and P. Vieu. *Nonparametric Functional Data Analysis*. Springer, 2006.

J. Friedman. Multivariate adaptive regression splines. *Annals of Statistics*, 19:11–41, 1991.

J. C. Gower. Metric and euclidean properties of dissimilarities coefficients. *Journal of Classification*, 3:5–48, 1986.

P Hall and C. Vial. Assessing the finite dimensionality of functional data. *Journal of the Royal Statistical Society: Series B*, 68(4):689–705, 2006.

T. Hastie, A. Buja, and R. Tibshirani. Flexible discriminant analysis by optimal scoring. *Journal of the American Statistical Association*, 89:1255–1270, 1994.

G. M. James and C. A. Sugar. Clustering for sparsely sampled functional data. *Journal of the American Statistical Association*, 98, 2003.

S. Sathiya Keerthi and Chih-Jen Lin. Asymptotic behaviors of support vectormachines with gaussian kernel. *Neural Computation*, 15:1667–1689, 2003.

G.R.G. Lanckriet, N. Cristianini, L. Bartlett, P.and El Ghaoui, and M.I. Jordan. Learning the kernel matrix with semidefinite programming. *Journal of Machine Learning Research*, 5:27–72, 2004.

C. L. Mallows. Some comments on cp. *Technometrics*, 15(4):661–675, 1973.

B. Marx and P. Eilers. Generalized linear regression on sampled signals and curves: a p-spline approach. *Technometrics*, 41:1–13, 1999.

J. Mercer. Functions of positive and negative type and their connection with the theory of integral equations. *Philos. Trans. Royal Soc. London*, 209:415–446, 1909.

J.M. Moguerza and A. Muñoz. Support vector machines with applications. *Statistical Science*, 21(4), 2006.

S. Mukherjee, P. Rifkin, and T. Poggio. Regression and classification with regularization. *Nonlinear Estimation and Classification', Lecture Notes in Statistics*, 171:107–124, 2002.

N. D. Pearce and M. P. Wand. Penalized splines and reproducing kernel methods. *The American Statistician*, 60(3):233–240, 2006.

A. Rakotomamonjy and S. Canu. Frames, reproducing kernels, regularization and learning. *Journal of Machine Learning Research*, 6:1485–1515, 2005.

J. O. Ramsay and B. W. Silverman. Functional data analysis, 2nd ed. *Springer, New York*, 2006.

H. Sakoe and S. Chiba. Dynamic programming algorithm optimization for spoken word recognition. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 26(1):43–49, 1978.

G. Schwarz. Estimating the dimension of a model. *Annals of Statistics*, 6:461–464, 1978.

S. Smale and D. X. Zhou. Geometry on probability spaces. *Working Paper*, 2007.

A Smola and B. Schölkopf. A tutorial on support vector regression. 1998.

P. T. Spellman, G. Sherlock, M. Q. Zhang, V. R. Iyer, K. Anders, M. B. Eisen, P. O. Brown, D. Bostein, and B. Futcher. Comprehensive identification of cell cycle-regulated genes of the yeast saccharomyces cerevisiae by microarray hybridization. *Mol. Biol. Cell.*, 9:3273–3297, 1998.

N. Sugiura. Further analysis of the data by akaikes information criterion and the finite corrections. *Communications in Statistics. Theory and Methods*, 7(1):13–26, 1978.

M. Sugiyama and K. R. Muller. The subspace information criterion for infite dimensional hypothesis spaces. *Journal of Machine Learning Research*, 3:323–359, 2002.

M. Sugiyama and H. Ogawa. Subspace information criterion for model selection. *Neural Computation*, 2001.

M Sugiyama and H. Ogawa. Theoretical and experimental evaluation of the subspace information criterion. *Machine Learning*, 48:25–50, 2002.

F. B. Swindel. Geometry of ridge regression illustrated. *The American Statistician*, 35(1): 12–15, 1981.

A.N. Tikhonov and V.Y. Arsenin. Solutions of ill-posed problems. *John Wiley and Sons, New York*, 1977.

V. Vapnik. The nature of statistical learning theory. *Springer, New York*, 1995.

G. Wahba. Reproducing kernel hilbert spaces - two brief reviews. *Proceedings of the 13th IFAC Symposium on System Identification*, pages 549–559, 2003.

G. Wahba. Spline models for observational data. *Series in Applied Mathematics, SIAM. Philadelphia*, 59, 1990.