Statistica Sinica **26** (2016), 619-638
doi:http://dx.doi.org/10.5705/ss.202014.0014

VARIABLE SELECTION IN FUNCTIONAL DATA CLASSIFICATION: A MAXIMA-HUNTING PROPOSAL

José R. Berrendero, Antonio Cuevas and José L. Torrecilla

Universidad Autónoma de Madrid

Abstract: Variable selection is considered in the setting of supervised binary classification with functional data $\{X(t),\ t \in [0,1]\}$. By "variable selection" we mean any dimension-reduction method that leads to the replacement of the whole trajectory $\{X(t),\ t \in [0,1]\}$, with a low-dimensional vector $(X(t_1), \ldots, X(t_d))$ still keeping a similar classification error. Our proposal for variable selection is based on the idea of selecting the local maxima $(t_1, \ldots, t_d)$ of the function $\mathcal{V}^2_X(t) = \mathcal{V}^2(X(t), Y)$, where $\mathcal{V}$ denotes the "distance covariance" association measure for random variables due to Székely, Rizzo, and Bakirov (2007). This method provides a simple natural way to deal with the relevance vs. redundancy trade-off which typically appears in variable selection. A result of consistent estimation for the maxima of $\mathcal{V}^2_X$ is shown. We also show different models for the underlying process $X(t)$ under which the relevant information is concentrated on the maxima of $\mathcal{V}^2_X$. An extensive empirical study is presented, including about 400 simulated models and data examples aimed at comparing our variable selection method with other standard proposals for dimension reduction.

Key words and phrases: Distance correlation, functional data analysis, supervised classification, variable selection.

1. Introduction

When dealing with functional data, the use of dimension reduction techniques is a natural idea. Some of these techniques are based upon the use of general (linear) finite-dimensional projections. This is the case of functional principal component analysis (FPCA), see Li, Wang, and Carroll (2013), although the so-called functional partial least squares (PLS) methodology is in general preferable when a response variable is involved; see Delaigle and Hall (2012a) for a recent reference. Other common dimension reduction methods in the functional setting include sliced inverse regression (Hsing and Ren (2009); Jiang, Yu, and Wang (2013)) and additive models (Zhang, Park, and Wang (2013)). Also, the methods based on random projections can offer an interesting alternative. See, e.g., Cuevas (2014) for a short overview of dimension-reduction techniques together with additional references.

Our proposal here is for a more radical approach to dimension reduction using **variable selection methods**. The aim of variable selection, when applied to functional data, is to replace every infinite dimensional observation $\{x(t),\ t \in [0,1]\}$, with a finite dimensional vector $(x(t_1), \ldots, x(t_d))$. The selection of the "variables" $t_1, \ldots, t_d$ should be a consequence of a trade-off between two conflicting goals: representativeness and parsimony. We want to retain as much information as possible (thus selecting relevant variables) employing a small number of variables (thus avoiding redundancy).

It is clear that variable selection has, at least, an advantage when compared with other dimension reduction methods (PCA, PLS...) based on general projections: the output of any variable selection method is directly interpretable in terms of the original variables, provided the required number d of selected variables is not too large. As a matter of fact, variable selection is sometimes the main target itself where the focus is on model simplification.

We are especially interested in the "intrinsic" approaches to variable selection, in the sense that the final output should depend only on the data, not on any assumption on the underlying model (although the result should be interpretable in terms of the model). There is a vast literature on these topics published by researchers in machine learning and by mathematical statisticians. The approaches and the terminology used in these two communities are not always alike. Thus, in machine learning, variable selection is often referred to as *feature selection*. Also, the methods we have called "intrinsic" are often denoted as "filter methods" in machine learning. It is common (especially in the setting of regression models) to use the terms "sparse" or "sparsity" to describe situations in which variable selection is the first natural aim; see e.g., Gertheiss and Tutz (2010) and Rosasco et al. (2013). It has been also argued in Kneip and Sarda (2011) that the standard sparsity models are sometimes too restrictive so that it is advisable to combine them with other dimension reduction techniques. The "relevant" variables in a functional model are sometimes called "impact points" (McKeague and Sen (2010)) or "most predictive design points" (Ferraty, Hall, and Vieu (2010)). Also, the term "choice of components" has been used by Delaigle, Hall, and Bathia (2012) as a synonym for variable selection.

The recent literature in functional variable selection includes a version of the classical lasso procedure (Zhao, Chen, and Ogden (2014)), a study of consistency in the variable selection setup (Comminges and Dalalyan (2012)) and the use of inverse regression ideas in variable selection (Jiang and Liu (2014)). The monograph Guyon et al. (2006) contains a complete survey on feature extraction (including selection) from the point of view of machine learning. The overview paper by Fan and Lv (2010) has a more statistical orientation.

In what follows we focus on variable selection for the problem of supervised binary classification, with functional data. While the statement and basic ideas

behind the supervised classification (or discrimination) problem are widely known (see, e.g., Devroye, Györfi, and Lugosi (1996)), we need to briefly recall them for the sake of clarity and for notation purposes. Suppose that an explanatory random variable X, taking values in a *feature space* $\mathcal{F}$, can be observed in the individuals of two populations P_0 and P_1. Let Y denote a binary random variable, with values in $\{0,1\}$, indicating the membership to P_0 or P_1. On the basis of a data set $\mathcal{D}_n = ((X_1,Y_1),\ldots,(X_n,Y_n))$ of n independent observations drawn from (X,Y), the supervised classification problem aims at predicting the membership class Y of a new observation for which only the variable X is known.

A *classifier* or *classification rule* is just a measurable function $g : \mathcal{F} \to \{0,1\}$. It is natural to assess the performance of a classifier by the corresponding *classification error* $L = \mathbb{P}(g(X) \neq Y)$. It is well-known that the classification error $L = \mathbb{P}(g(X) \neq Y)$ is minimized by the so-called *Bayes classifier*, $g^*(x) = \mathbb{I}_{\{\eta(x)>1/2\}}$, where $\eta(x) = \mathbb{E}(Y|X=x) = \mathbb{P}(Y=1|X=x)$. Since g^* is in general unknown, it must be approximated, in different ways, by data-driven classifiers.

In our functional setting the feature space is (unless otherwise stated) $\mathcal{F} = \mathcal{C}[0,1]$, the space of real continuous functions defined on $[0,1]$, endowed with the supremum norm. Thus, our data will be of type $(X_1,Y_1),\ldots,(X_n,Y_n)$, where the X_i are iid trajectories in $\mathcal{C}[0,1]$ drawn from a stochastic process $X = X(t) = X(\omega,t)$. When no confusion is possible, we denote the whole process by X. When convenient, $X(t)$ is denoted X_t.

Several functional classifiers have been considered in the literature (see, e.g., Baíllo, Cuevas, and Fraiman (2011) for a survey). Among them, maybe the simplest one is the so-called k-nearest neighbors rule (k-NN). Additionally, we consider, as a standard choice, the classical linear Fisher's classifier (henceforth LDA) applied to the selected variables.

In Section 3 we propose a "maxima hunting" (MH) method for variable selection. It is essentially based on the idea of selecting the local maxima $(t_1,\ldots,t_d)$ of the function $\mathcal{V}^2(t) = \mathcal{V}^2(X(t),Y)$, where $\mathcal{V}^2$ denotes the "distance covariance" association measure for random variables due to Székely, Rizzo, and Bakirov (2007). An alternative version of the MH procedure can be obtained by replacing $\mathcal{V}^2(t)$ by the "distance correlation" $\mathcal{R}^2(t)$. See Section 2 for a short review of the definitions and properties of $\mathcal{V}^2$ and $\mathcal{R}^2$.

Some simplified versions for $\mathcal{V}^2$ are obtained in Theorem 1 of Section 3, for the particular case where Y is a binary variable. A result of consistent estimation (Theorem 2) for the maxima of $\mathcal{V}^2$ is also proved in that section.

In Section 4 we give several models (identified in terms of the conditional distributions $X(t)|Y=j$) in which the optimal classification rule depends only on a finite number of variables. We also show that in some of these models the variables to be selected coincide with the maxima of $\mathcal{V}^2$. These results provide a

theoretical basis for the techniques of variable selection in functional classification models. Usually these techniques are considered from an exclusively algorithmic or computational point of view. It is therefore of some interest to motivate them in "population terms", by identifying some specific models where these techniques make sense. As pointed out by Biau, Cadre, and Paris (2015), "Curiously, despite a huge research activity in this area, few attempts have been made to connect the rich theory of stochastic processes with functional data analysis". So the present paper can be seen as a contribution to partially filling this gap.

An extensive simulation study, comparing our variable selection methods with other dimension reduction procedures (as well as with the "baseline option" of doing no variable selection at all) is included in Section 5. Three data examples are discussed in Section 6. Section 7 includes some final conclusions as well as a ranking of the considered methods.

Proofs are included in the Supplementary Material document.

2. An Auxiliary Tool: The Distance Covariance

The problem of finding appropriate association measures between random variables (beyond the standard linear correlation coefficient) has received attention in recent years; see for instance Hall and Miller (2011). We use here the association measure proposed by Székely, Rizzo, and Bakirov (2007), see also Székely and Rizzo (2009). It is called *distance covariance* (or *distance correlation* in the standardized version). It has a number of valuable properties: it can be used to define the association between two random variables X and Y of arbitrary (possibly different) dimensions; it characterizes independence in the sense that the distance covariance between X and Y is zero if and only if X and Y are independent; the distance correlation can be easily estimated in a natural plug-in way, with no need for smoothing or discretization.

Definition 1. Given two random variables X and Y taking values in $\mathbb{R}^p$ and $\mathbb{R}^q$, respectively, let $\varphi_{X,Y}$, φ_X and φ_Y be the characteristic functions of (X, Y), X and Y, respectively. Assume that the components of X and Y have finite first-order moments. The distance covariance between X and Y is the non-negative square root of

$$\mathcal{V}^2(X,Y) = \int_{\mathbb{R}^{p+q}} \mid \varphi_{X,Y}(u,v) - \varphi_X(u)\varphi_Y(v) \mid^2 w(u,v)dudv, \qquad (2.1)$$

with $w(u,v) = (c_p c_q |u|_p^{1+p} |v|_q^{1+q})^{-1}$, where $c_d = \pi^{(1+d)/2}/\Gamma((1+d)/2)$ is half the surface area of the unit sphere in $\mathbb{R}^{d+1}$ and $|\cdot|_d$ stands for the Euclidean norm in $\mathbb{R}^d$. With $\mathcal{V}^2(X) = \mathcal{V}^2(X,X)$, the (square) distance correlation is $\mathcal{R}^2(X,Y) = \mathcal{V}^2(X,Y)/\sqrt{\mathcal{V}^2(X)\mathcal{V}^2(Y)}$ if $\mathcal{V}^2(X)\mathcal{V}^2(Y) > 0$, $\mathcal{R}^2(X,Y) = 0$ otherwise.

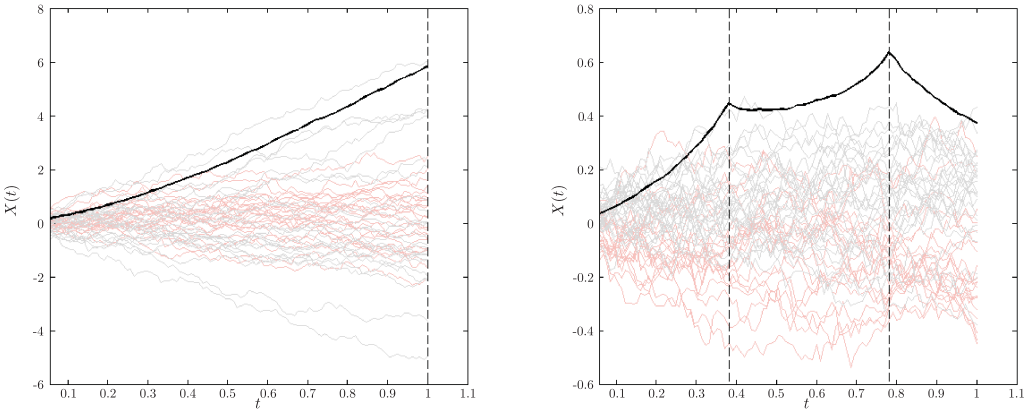

Figure 1. Left: 50 trajectories of model in Proposition 1. Right: Logistic model L11 (explained in Subsection 5.2) with 50 Ornstein-Uhlenbeck trajectories. $\mathcal{V}^2(X_t, Y)$ (scaled) is in black and the relevant variables are marked by vertical dashed lines.

These definitions make sense even if X and Y have different dimensions. The association measure $\mathcal{V}^2(X, Y)$ can be consistently estimated through a relatively simple average of products calculated in terms of the mutual pairwise distances $|X_i - X_j|_p$ and $|Y_i - Y_j|_q$ between the sample values X_i and the Y_j; see Székely and Rizzo (2009, expression (2.8)). See also Li, Zhong, and Zhu (2012) for a different use of the correlation distance in variable selection.

3. Variable Selection Based on Maxima Hunting

Our proposal is to select the values of t corresponding to local maxima of the distance-covariance function $\mathcal{V}^2(X_t, Y)$ or, alternatively, of the distance correlation function $\mathcal{R}^2(X_t, Y)$. This method is a natural way to deal with the relevance vs. redundancy trade-off: the selected values must carry a large amount of information on Y, which takes into account the *relevance* of the selected variables; considering local maxima automatically takes care of the *redundancy* problem, since the highly relevant points close to the local maxima are automatically excluded from consideration. The results of Section 5 have the practical performance of the maxima-hunting method as quite satisfactory. Figure 1 shows how the function $\mathcal{V}^2(X_t, Y)$ looks in two different examples.

The extreme flexibility of these association measures allows us to consider multivariate responses Y so one can apply the same ideas for multiple classification or even to a regression problem. However, we limit ourselves here to the problem of binary classification. In this case we can derive simplified expressions for $\mathcal{V}^2(X, Y)$ which are particularly convenient for empirical approximations.

The results of this section are obtained for the d-variate case, although we use them just for d=1. Here, d denotes a natural number and t stands for a vector $t = (t_1, \ldots, t_d) \in [0,1]^d$. For a given process X, we abbreviate $X(t) = (X(t_1), \ldots, X(t_d))$ by X_t and Z' denotes an independent copy of a random variable Z. We write $u^\top$ and $|u|_d$ to denote the transposed and the Euclidean norm of a vector $u \in \mathbb{R}^d$. Let $\eta(x) = \mathbb{P}(Y = 1|X = x)$ so that $Y|X \sim \text{Binomial}(1, \eta(X))$. Observe that $p = \mathbb{P}(Y = 1) = \mathbb{E}(\mathbb{P}(Y = 1|X)) = \mathbb{E}(\eta(X))$.

Theorem 1. *In our setting, the function $\mathcal{V}^2(X_t, Y)$ at (2.1) can be alternatively calculated as*

(a)

$$\mathcal{V}^2(X_t, Y) = \frac{2}{c_d} \int_{\mathbb{R}^d} \frac{|\zeta(u,t)|^2}{|u|_d^{d+1}} du, \tag{3.1}$$

where $\zeta(u,t) = \mathbb{E}\left[(\eta(X) - p)\, e^{iu^\top X_t}\right]$ and c_d is given in Definition 1.

(b)

$$\begin{aligned} \mathcal{V}^2(X_t, Y) =& -2\mathbb{E}\left[(\eta(X) - p)(\eta(X') - p)|X_t - X'_t|_d\right] \\ =& -2\mathbb{E}\left[(Y - p)(Y' - p)|X_t - X'_t|_d\right], \end{aligned} \tag{3.2}$$

where (X', Y') denotes an independent copy of (X, Y);

(c)

$$\mathcal{V}^2(X_t, Y) = 4p^2(1-p)^2 \left[I_{01}(t) - \frac{I_{00}(t) + I_{11}(t)}{2}\right], \tag{3.3}$$

where $I_{ij}(t) = \mathbb{E}\left(|X_t - X'_t|_d \,|\, Y = i, Y' = j\right)$.

In a training sample $\{(X_i, Y_i),\ i = 1, \ldots, n\}$ denote by $X_1^{(0)}, \ldots, X_{n_0}^{(0)}$ and $X_1^{(1)}, \ldots, X_{n_1}^{(1)}$ the X-observations corresponding to values $Y_i = 0$ and $Y_i = 1$, respectively. We use these data to obtain an estimator of $\mathcal{V}^2(X_t, Y)$ that is uniformly consistent in t. As a consequence, we can estimate the local maxima of $\mathcal{V}^2(X_t, Y)$: using part (c) of Theorem 1, a natural estimator for $\mathcal{V}^2(X_t, Y)$ is

$$\mathcal{V}_n^2(X_t, Y) = 4\hat{p}^2(1-\hat{p})^2 \left[\hat{I}_{01}(t) - \frac{\hat{I}_{00}(t) + \hat{I}_{11}(t)}{2}\right],$$

where $\hat{p} = n_1/(n_0 + n_1)$, $\hat{I}_{rr}(t) = (2/n_r(n_r - 1)) \sum_{i<j} |X_i^{(r)}(t) - X_j^{(r)}(t)|_d$, for $r = 0, 1$, and $\hat{I}_{01}(t) = (1/n_0 n_1) \sum_{i=1}^{n_0} \sum_{j=1}^{n_1} |X_i^{(0)}(t) - X_j^{(1)}(t)|_d$.

Theorem 2. *Let $X = X_t$, with $t \in [0,1]^d$, be a process with continuous trajectories almost surely, such that $\mathbb{E}(\|X\|_\infty \log^+ \|X\|_\infty) < \infty$. Then $\mathcal{V}_n^2(X_t, Y)$ is continuous in t and*

$$\sup_{t \in [0,1]^d} |\mathcal{V}_n^2(X_t, Y) - \mathcal{V}^2(X_t, Y)| \to 0 \quad \textit{a.s., as } n \to \infty.$$

Accordingly, if $\mathcal{V}^2(X_t, Y)$ has exactly m local maxima at $t_1, \ldots, t_m$, then $\mathcal{V}_n^2(X_t, Y)$ eventually has at least m maxima at $t_{1n}, \ldots, t_{mn}$ with $t_{jn} \to t_j$, as $n \to \infty$, a.s., for $j = 1, \ldots, m$.

4. Some Theoretical, Model-oriented Motivation for Variable Selection and Maxima-hunting

For the binary functional classification problem we aim at selecting *a finite number of variables*. One might think that this is a "too coarse" approach for functional data, but we provide some theoretical motivation; in some relevant models, variable selection is "the best we can do" in the sense that, the optimal classifier depends on only a finite (typically small) number of variables. In many situations a proper variable selection leads to an improvement in efficiency (with respect to the baseline option of using the full sample curves), due to the gains associated with a smaller noise level.

The distribution of $X(t)|Y = i$, is denoted by μ_i for $i = 0, 1$. In all the examples the processes are Gaussian. Many of these models have non-smooth, Brownian-like trajectories. These play a role in statistical applications, in particular in the classification problem; see, e.g., Lindquist and McKeague (2009).

Basic notations and results are used throughout (see, e.g., Athreya and Lahiri (2006, chap. 4)). Thus, (see Feldman (1958)) if μ_0 and μ_1 are Gaussian, they are either equivalent or mutually singular, and we write the Radon-Nikodym derivative of μ_1 which respect to μ_0 as $f = d\mu_1/d\mu_0$.

Our results in this section use Baíllo, Cuesta-Albertos, and Cuevas (2011, Thm. 1):

$$\eta(x) = \left[\frac{1-p}{p}\frac{d\mu_0}{d\mu_1}(x) + 1\right]^{-1}, \quad \text{for } x \in \mathcal{S}, \tag{4.1}$$

where $\mathcal{S}$ is the common support of μ_0 and μ_1, and $p = \mathbb{P}(Y = 1)$. This provides expression to the optimal rule $g^*(x) = \mathbb{I}_{\{\eta(x)>1/2\}}$ in some important cases where the Radon-Nikodym derivative is explicitly known.

Some examples. Two situations in which the Radon-Nikodym derivatives can be explicitly calculated arise when μ_0 is the standard Brownian motion $B(t)$, and μ_1 corresponds to $B(t)$ plus a stochastic or a linear trend. In these cases the Bayes rule has the form $g^*(X) = h(X(1))$. We state this formally, proofs can be found in the Supplementary Material.

Proposition 1. *If μ_0 is the distribution of a standard Brownian motion $B(t)$, $t \in [0, 1]$ and μ_1 is the distribution of $B(t) + \theta t$, where θ is $N(0, 1)$, independent from B, then the Bayes rule is given by $g^*(x) = \mathbb{I}_{\left\{x_1^2 > 4\log\left(\frac{\sqrt{2}(1-p)}{p}\right)\right\}}(x)$, for all $x \in \mathcal{C}[0, 1]$.*

In particular, when the prior probabilities of the groups are equal, $p = 1/2$, we get $g^*(x) = 1$ if and only if $|x_1| > 2\sqrt{\log\sqrt{2}} \approx 1.77$.

Proposition 2. *If μ_0 is the distribution of a standard Brownian motion $B(t)$, $t \in [0,1]$ and μ_1 is the distribution of $B(t) + ct$, where $c \neq 0$ is a constant, then for $x \in \mathcal{C}[0,1]$ the Bayes rule is given by*

$$g^*(x) = \begin{cases} \mathbb{I}_{\left\{x_1 > \frac{c}{2} - \frac{1}{c}\log\left(\frac{p}{1-p}\right)\right\}}(x), & \text{if } c > 0, \\ \mathbb{I}_{\left\{x_1 < \frac{c}{2} - \frac{1}{c}\log\left(\frac{p}{1-p}\right)\right\}}(x), & \text{if } c < 0. \end{cases}$$

Consider the countable family of *Haar functions*, $\varphi_{m,k} = \sqrt{2^{m-1}}\left[\mathbb{I}_{\left(\frac{2k-2}{2^m}, \frac{2k-1}{2^m}\right)} - \mathbb{I}_{\left(\frac{2k-1}{2^m}, \frac{2k}{2^m}\right)}\right]$, for $m, k \in \mathbb{N}$, $1 \leq k \leq 2^{m-1}$. They form an orthonormal basis in $L^2[0,1]$. Take the "peak" functions $\Phi_{m,k}$ to be

$$\Phi_{m,k}(t) = \int_0^t \varphi_{m,k}(s)ds. \tag{4.2}$$

We want to use the peak functions to define the trend of the μ_1 distribution in another model of type "Brownian versus Brownian plus trend". In this case the Bayes rule depends just on three points.

Proposition 3. *Let μ_0 be the distribution of a standard Brownian motion $B(t)$, $t \in [0,1]$ and μ_1 be the distribution of $B(t) + \Phi_{m,k}(t)$, where $\Phi_{m,k}$ is one of the peak functions at (4.2). Then, for $x \in \mathcal{C}[0,1]$ the regression function $\eta(x) = \mathbb{E}(Y|X = x)$ is*

$$\eta(x) = \left\{ \frac{1-p}{p} \exp\left(\frac{1}{2} - 2^{\frac{m-1}{2}} \left[\left(x_{\frac{2k-1}{2^m}} - x_{\frac{2k-2}{2^m}} \right) + \left(x_{\frac{2k-1}{2^m}} - x_{\frac{2k}{2^m}} \right) \right] \right) + 1 \right\}^{-1} \tag{4.3}$$

and the Bayes rule $g^*(x) = \mathbb{I}_{\{\eta(x) > 1/2\}}$ has $g^*(x) = 1$ if and only if

$$\left(x_{\frac{2k-1}{2^m}} - x_{\frac{2k-2}{2^m}} \right) + \left(x_{\frac{2k-1}{2^m}} - x_{\frac{2k}{2^m}} \right) > \frac{1}{\sqrt{2^{m+1}}} - \frac{1}{\sqrt{2^{m-1}}} \log\left(\frac{p}{1-p} \right). \tag{4.4}$$

According to the Cameron-Martin Theorem (see Mörters and Peres (2010)), for the equivalence of μ_1 and μ_0 the trend function must belong to the Dirichlet space $\mathcal{D}[0,1]$ of real functions F defined in $[0,1]$ which have a derivative F' in $L^2[0,1]$ with $F(t) = \int_0^t F'(s)ds$. It can be seen (Mörters and Peres (2010, p. 28)) that $\{\Phi_{m,k}\}$ is an orthonormal basis for $\mathcal{D}[0,1]$.

Remark 1. Analogous calculations can be performed (still obtaining explicit expressions for the Bayes rule of type $g^*(x) = g(x(t_1), \ldots, x(t_d)))$, using a rescaled Brownian motion $\sigma B(t)$ or the Brownian Bridge instead of $B(t)$, or a piecewise linear trend instead of these. Likewise, other models could be obtained by linear combinations in the trend functions or by finite mixtures of other simpler models. Many of them have been included in the simulation study of Section 5.

We provide some theoretical support for the maxima-hunting method by showing that, in some useful models, the optimal classification rule depends on the maxima of the distance covariance function $\mathcal{V}^2(X_t, Y)$; in some particular examples, other points (closely linked to the maxima) are also relevant.

Proposition 4. *For the models of Propositions* 1 *and* 2*, the distance covariance functions* $\mathcal{V}^2(X_t, Y)$ *both have a unique relative maximum at the point* $t = 1$.

Remark 2. Similar results can be obtained for the model considered in Proposition 3, as well as for the Brownian bridge vs. Brownian motion model.

The model considered in Proposition 1 provides a clear example of the advantages of using the distance covariance measure $\mathcal{V}^2(X_t, Y)$ rather than the ordinary covariance $Cov^2(X_t, Y)$ in the maxima-hunting procedure. Here $Cov^2(X_t, Y) = p^2(1-p)^2(\mathbb{E}(X(t)|Y=0) - \mathbb{E}(X(t)|Y=1))^2 = 0$, for all $t \in [0,1]$, so that the ordinary covariance is useless to detect any difference between the values of t.

5. A Simulation Study

We describe here in detail the methods under study and the models to be considered together with a summary of the results. The full outputs can be found in `www.uam.es/antonio.cuevas/exp/outputs.xlsx`.

5.1. The variable selection methods under study. Criteria for comparisons

These are the methods, and their corresponding notations as they appear in the tables and figures below.

1. **Maxima-hunting**. The functional data $x(t)$, $t \in [0,1]$ are discretized to $(x(t_1), \ldots, x(t_N))$, so a non-trivial practical problem is to decide which points in the grid are the local maxima: a point t_i is declared to be a local maximum when it is the highest local maximum on the sub-grid $\{t_j\}$, $j = i - h \ldots, i + h$. The proper choice of h depends on the nature and discretization pattern of the data at hand. Thus, h could be considered as a smoothing parameter to be selected in an approximately optimal way. In our experiments h is chosen by a validation step explained in next section.

Then, we sort the maxima t_i by **relevance** (the value of the function at t_i). This seems to be the natural order and it produces better results than other simple sorting strategies. We denote these maxima-hunting methods by **MHR** and **MHV** depending on the use of $\mathcal{R}^2$ or $\mathcal{V}^2$.

2. **Univariate t-ranking method.** Denoted by **T**, it is frequently used when selecting relevant variables (see e.g. the review by Fan and Lv (2010)). It is

based on the simple idea of selecting the variables X_t with highest Student's t two-sample scores $T(X_t) = |\bar{X}_{1t} - \bar{X}_{0t}|/\sqrt{s_{1t}^2/n_1 + s_{0t}^2/n_0}$.

3. **mRMR**. The minimum Redundancy Maximum Relevance algorithm, proposed in Ding and Peng (2005) and Peng, Long, and Ding (2005), is a relevant intrinsic variable selection method; see Berrendero, Cuevas, and Torrecilla (2015) for a recent contribution. It aims at maximizing the relevance of the selected variables avoiding an excess of redundancy what seems particularly suitable for functional data. Denoting the set of selected variables by S, the variables are sequentially incorporated to S with the criterion of maximizing the difference $Relevance(S) - Redundancy(S)$ (or alternatively the quotient $Relevance(S)/Redundancy(S)$). Two ways of measuring relevance and redundancy have been proposed: the Fisher statistic for relevance and the standard correlation for redundancy; three-fold discretized version of the so-called *Mutual Information* measure for both relevance and redundancy (see Ding and Peng (2005, equation (1))).

In principle these two approaches are intended for continuous and discrete variables respectively. However, Ding and Peng (2005) report a good performance for the second one even in the continuous case. We have considered mRMR as a natural competitor for our maxima-hunting approximation. We have computed both Fisher-Correlation and Mutual Information approaches with both difference and quotient criteria. For the sake of clarity we only show here the results of **FCQ** (Fisher Correlation Quotient) and **MID** (Mutual Information Difference) that outperform, on average, their corresponding counterparts.

4. **PLS**. According to the available results (Preda, Saporta, and Lévéder (2007); Delaigle and Hall (2012a)) PLS is the "'method of choice" for dimension reduction in functional classification. Note however that PLS is not a variable selection procedure; in particular it lacks the interpretability of variable selection. In some sense, the motivation for including PLS is to check how much we lose by restricting ourselves to variable selection methods, instead of considering other more general linear projections procedures (as PLS) for dimension reduction.

5. **Base**. The k-NN classifier is applied to the entire curves. The Base performance can be seen as a reference to assess the usefulness of dimension reduction methods. Somewhat surprisingly, Base is often outperformed. The Base method cannot be implemented with LDA since this classifier typically fails with infinite or high-dimensional data; see, e.g. Cuevas (2014, Sec. 6.1), for some insights and references.

The **classifiers** used in all cases are either k-NN, based on the Euclidean distance or LDA (applied to the selected variables). Similar comparisons could be done with other classifiers, since the considered methods do not depend on the classifier. For comparing the different methods we use the natural accuracy measure, defined by the percentage of correct classification.

5.2. The structure of the simulation study

Our simulation study consisted of 400 experiments, aimed at comparing the practical performances of several intrinsic variable selection methods described in the previous subsection. These experiments were obtained by considering 100 different underlying models and 4 sample sizes, where by "model" we mean either,

(M1) a pair of distributions for $X|Y=0$ and $X|Y=1$ (corresponding to P_0 and P_1, respectively); in all cases, we took $p=\mathbb{P}(Y=1)=1/2$.

(M2) The marginal distribution of X plus $\eta(x)=\mathbb{P}(Y=1|X=x)$.

Models varied in difficulty and number of relevant variables. In all of them the optimal Bayes rule turned out to depend on a finite number of relevant variables, see Section 3. The processes involved included also different levels of smoothing. The full list of considered models is available in the Supplementary Material document. All of them belong to one of the following classes:

1. **Gaussian models**. They are denoted $G1, G1b, \ldots, G8$. All of them were generated according to the general pattern (M1). The distributions of $X(t)|Y=i$ were chosen among the following: **standard Brownian Motion**, B, in $[0,1]$; **Brownian Motion, BT, with a trend** $m(t)$, $BT(t)=B(t)+m(t)$ (we have considered several choices for $m(t)$); **Brownian bridge**: $BB(t)=B(t)-tB(1)$; the **Ornstein-Uhlenbeck process**, with covariance function $\gamma(s,t)=a\exp(-b|s-t|)$ and zero mean (OU) or different mean functions $m(t)$ (OUt). Smoother processes have been also computed by convolving Brownian trajectories with Gaussian kernels. We considered two levels of smoothing, denoted by sB and ssB.

2. **Logistic models**. They are defined through the general pattern (M2) with the process $X=X(t)$ having one of the above mentioned distributions and $Y\sim \text{Binom}(1,\eta(X))$ with $\eta(x)=(1+e^{-\Psi(x(t_1),\ldots,x(t_d))})^{-1}$, a function of the relevant variables $x(t_1),\ldots,x(t_d)$. We considered 15 versions of this model and a few variants, denoted $L1, L2, L3, L3b, \ldots, L15$. They correspond to different choices for the link function Ψ (most of them linear or polynomial) and for the distribution of X. For example, in the models L2 and L8 we have $\Psi(x)=10x_{30}+10x_{70}$ and $\Psi(x)=10x_{50}^4+50x_{80}^3+20x_{30}^2$, respectively.

3. **Mixtures**. They are obtained by combining (via mixtures) in several ways the above mentioned Gaussian distributions assumed for $X|Y=0$ and $X|Y=1$. These models are denoted M1, ..., M11 in the output tables.

For each model, the variable selection methods (as well as PLS) were checked for sample sizes $n=30$, 50, 100, 200. Thus $100\times 4=400$ experiments.

Functional simulated data were **discretized** to $(x(t_1), \ldots, x(t_{100}))$, where the t_i are equispaced points in $[0, 1]$. To avoid the degeneracy $x(t_0) = 0$ in the Brownian-like models we took $t_1 = 6/105$. Similarly, for the case of the Brownian bridge, we truncated as well at the end of the interval.

The involved parameters are the number k of nearest neighbors in the k-NN classifier, the dimension of the reduced space (number of variables or PLS components), and the smoothing parameter h in maxima-hunting methods. These were set by standard data-based validation procedures. Parameter validation can be carried out mainly through a validation set or by cross-validation on the training set (see e.g. Guyon et al. (2006)). For the simulation study, validation and test samples of size 200 were randomly generated. In the data sets we proceed by cross-validation.

5.3. A few numerical outputs from the simulations

We have selected (with no particular criterion in mind) a sampling of just a few examples among the 400 experiments. The complete simulation outputs can be downloaded from `www.uam.es/antonio.cuevas/exp/outputs.xlsx`. Table 1 provides the performance (averaged on 200 runs) measured in terms of classification accuracy (percentages of correct classification). Models are presented in rows and methods in columns. The marked outputs correspond to the winner and second best method in each row.

The outputs of Table 1 are more or less representative of the overall conclusions of the entire study. For instance, MHR appears as the overall winner on average, with a slight advantage. PLS and the maxima-hunting methods (MHR and MHV) obtain similar scores and clearly outperform the other benchmark methods. They also beat (often very clearly) the Base method in almost all cases using just a few variables. This shows that dimension reduction is, in fact, "mandatory" in many cases. Regarding the comparison of k-NN and LDA in the second stage (after dimension reduction) the results show a slight advantage for k-NN (on average). The complete failure of LDA in models G1 and G3 was to be expected since in these cases the mean functions are identical in the populations. In terms of number of variables, when k-NN is used, MHR and MHV need less variables to achieve better results than the rest of variable selection methods. When LDA is used, the number of required variables is quite similar in all methods; see the Supplementary Material, Section S4.

6. Data Examples

We have chosen three examples due to their popularity in FDA. There are many references on these datasets so we just give brief descriptions of them;

Table 1. Average correct classification outputs, over 200 runs, with $n = 50$.

				k-**NN outputs**			
Models	FCQ	MID	T	PLS	MHR	MHV	Base
L2_OUt	82.47	82.11	81.68	83.27	83.22	83.23	82.60
L6_OU	88.41	89.81	86.19	90.93	90.75	90.83	90.56
L10_B	81.09	85.02	81.13	85.90	87.27	87.42	85.46
L11_ssB	82.31	80.85	82.28	78.81	83.10	82.81	79.89
L12_sB	77.24	75.83	77.41	74.92	78.57	76.62	74.78
G1	65.86	70.70	65.57	66.95	71.59	71.80	70.10
G3	63.09	73.39	60.57	60.56	77.47	77.06	65.26
G6	84.27	91.95	84.14	93.67	93.38	93.71	92.19
M2	70.77	69.82	69.16	78.16	74.76	75.68	71.14
M6	81.15	83.08	79.73	83.47	83.32	83.35	80.99
M10	64.93	68.33	64.58	68.25	70.66	70.94	68.95
				LDA outputs			
Models	FCQ	MID	T	PLS	MHR	MHV	Base
L2_OUt	79.80	78.95	78.23	80.07	80.24	80.14	-
L6_OU	87.79	88.91	84.46	91.01	89.44	89.35	-
L10_B	75.97	75.44	76.04	77.60	77.63	77.76	-
L11_ssB	80.95	80.09	80.81	79.39	81.88	81.63	-
L12_sB	76.39	75.20	76.40	75.02	77.38	75.96	-
G1	51.27	51.24	51.20	51.44	51.55	51.70	-
G3	51.09	52.26	50.96	50.35	52.95	52.69	-
G6	87.72	95.28	87.80	97.77	96.54	96.85	-
M2	67.44	76.51	66.81	84.38	82.24	83.06	-
M6	79.99	79.92	79.63	81.39	81.08	81.38	-
M10	60.03	65.61	59.24	67.49	67.25	67.99	

additional details can be found in the Supplementary Material document. Figure 2 shows the trajectories $X(t)$ and mean functions for each set and each class.

Berkeley Growth Data. The heights of 54 girls and 39 boys measured at 31 non-equidistant time points. See, e.g., Ramsay and Silverman (2005).

Tecator. 215 near-infrared absorbance spectra (100 grid points each) of finely chopped meat, obtained using a Tecator Infratec Food & Feed Analyzer. The sample is separated in two classes according to the fat content (smaller or larger than 20%). Tecator curves are often used in a differentiated version. We use here the second derivatives. See Ferraty and Vieu (2006) for details.

Phoneme. As in Delaigle, Hall, and Bathia (2012) we use the "binary" version of these data corresponding to log-periodograms constructed from 32 ms long recordings of males pronouncing the phonemes "aa" and "ao". The sample size

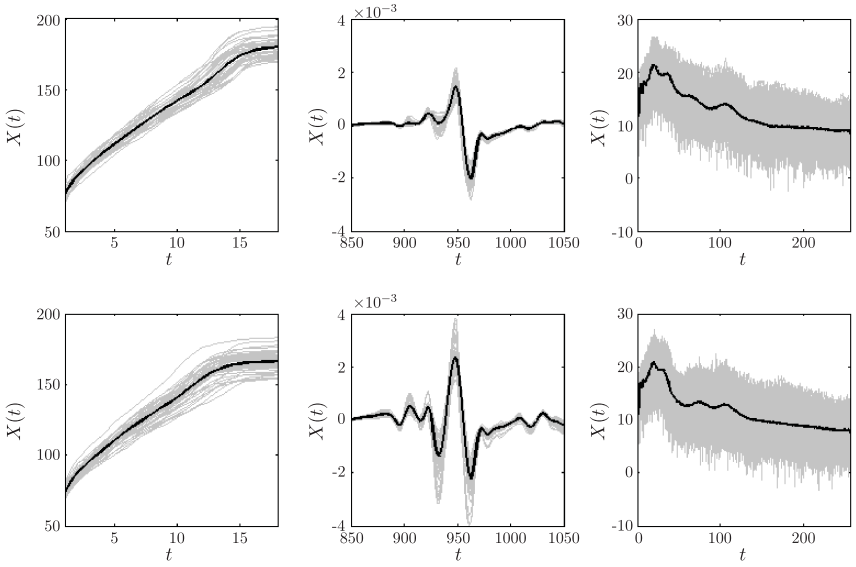

Figure 2. Data trajectories and mean functions from class 0 (first row) and class 1 (second row). Columns correspond to growth, Tecator and phoneme data from left to right.

is $n = 1,717$ (695 from "aa" and $1,022$ from "ao"). Each curve was observed at 256 equispaced points.

In the comparisons with data sets we have incorporated the method recently proposed by Delaigle, Hall, and Bathia (2012). We denote it by DHB. Given a classifier, the DHB method proposes a leave-one-out choice of the best variables for the considered classification problem. While this is a worthwhile natural idea, it is computationally intensive. So the authors implemented a slightly modified version, which we have closely followed. It is based on a sort of trade-off between full and sequential search, together with some additional computational savings. Let us note, as an important difference with our maxima-hunting method, that the DHB procedure is a "wrapper" method, in the sense that it depends on the chosen classifier. Following Delaigle, Hall, and Bathia (2012), we have only implemented the DHB method with the LDA classifier.

Apart from that, we proceeded as in the simulation study except for the generation of the training, validation and test samples. Here we considered the usual cross-validation procedure which avoids splitting the sample (sometimes small) into three different sets. Each output was obtained by standard leave-one-out cross-validation. The only exception was the phoneme data set for which this procedure is extremely time-consuming (due to the large sample size); we used instead ten-fold cross-validation (10CV). The respective validation steps

Table 2. Classification accuracy (in %) for the data with both classifiers.

				k-NN outputs				
Data	FCQ	MID	T	PLS	MHR	MHV	DHB	Base
Growth	83.87	95.70	83.87	94.62	95.70	94.62	-	96.77
Tecator	99.07	99.07	99.07	97.21	99.53	99.53	-	98.60
Phoneme	80.43	79.62	80.43	82.53	80.20	78.86	-	78.97
				LDA outputs				
Data	FCQ	MID	T	PLS	MHR	MHV	DHB	Base
Growth	91.40	94.62	91.40	95.70	95.70	96.77	96.77	-
Tecator	94.42	95.81	94.42	94.42	95.35	94.88	95.35	-
Phoneme	79.38	80.37	79.09	80.60	80.20	78.92	77.34	-

Table 3. Average number of variables (or components) selected for the data sets.

				k-NN outputs				
Data	FCQ	MID	T	PLS	MHR	MHV	DHB	Base
Growth	1.0	3.5	1.0	2.8	4.0	4.0	-	31
Tecator	3.0	5.7	3.0	2.7	1.0	1.0	-	100
Phoneme	10.7	15.3	12.3	12.9	10.2	12.3	-	256
				LDA outputs				
Data	FCQ	MID	T	PLS	MHR	MHV	DHB	Base
Growth	5.0	3.4	5.0	2.0	4.0	4.0	2.3	-
Tecator	8.4	2.6	3.1	9.7	1.7	1.8	3.0	-
Phoneme	8.5	17.1	7.9	15.5	16.1	11.0	2.0	-

were done with the same resampling schemes within the training samples. This is a usual way to proceed when working with data; see Hastie, Tibshirani, and Friedman (2009, Sec. 7.10). Several outputs are given in Tables 2 (accuracy) and 3 (number of variables). The complete results can be found in `www.uam.es/antonio.cuevas/exp/outputs.xlsx`.

These results are similar to those obtained in the simulation study. While (as expected) there is no clear global winner, maxima-hunting method looks to be a very competitive choice. In particular, Tecator outputs are striking, since MHR and MHV achieve (with k-NN) a near perfect classification with just one variable. Maxima-hunting methods (particularly MHR) outperform, or are very close to, the Base outputs (which use the entire curves). PLS is overcome by our methods in two of the three problems but it is the clear winner in the phoneme example. In any case, it should be kept in mind, as a counterpart, the ease of interpretability of the variable selection methods.

The DHB method performs well in the first two considered examples but (relatively) fails in the phoneme case. There is maybe some room for improve-

Table 4. Average ranking scores over the 400 experiments.

k-NN rankings							
Ranking criterion	FCQ	MID	T	PLS	MHR	MHV	Base
Relative	4.42	5.80	2.93	6.99	8.42	7.35	3.64
Positional	6.44	6.71	5.50	7.96	8.68	7.84	5.89
F1	11.62	12.04	9.46	17.39	17.96	15.41	10.15
LDA rankings							
Ranking criterion	FCQ	MID	T	PLS	MHR	MHV	Base
Relative	3.76	5.19	1.96	6.90	8.62	8.07	-
Positional	6.70	6.99	5.92	8.13	8.79	8.49	-
F1	11.95	12.52	10.22	17.49	18.41	17.47	-

ment in the stopping criterion (we have used the same parameters as in Delaigle, Hall, and Bathia (2012)). By construction, this is (in the machine learning terminology) a "wrapper" method. This means that the variables selected by DHB are specific for the LDA classifier (and might dramatically change with other classification rules). The use of the LDA classifier did not lead to any significant gain; in fact, the results are globally worse than those of k-NN, except for a few particular cases.

Although our methodology is not primarily targeted to the best classification rate, but to the choice of the most representative variables, we can conclude that MH procedures combined with the simple k-NN are competitive when compared with PLS and other successful and sophisticated methods in the literature: see Galeano, Joseph, and Lillo (2014) for Tecator data, Mosler and Mozharovskyi (2014) for growth data, and Delaigle, Hall, and Bathia (2012) for phoneme data.

7. Overall Conclusions: A Tentative Global Ranking of Methods

We have summarized the conclusions of our 400 simulation experiments in three rankings, prepared with different criteria, according to **classification accuracy**. With the **relative ranking** criterion, the winner method (with performance W) in each of the 400 experiments gets 10 score points, and the method with the worst performance (say w) gets 0 points. The score of any other method, with performance u is just assigned in a proportional way: $10(u-w)/(W-w)$. The **positional ranking** scoring criterion just gives 10 points to the winner in every experiment, 9 points to the second one, etc. The **F1 ranking** strongly rewards the winner. For each experiment, points are divided as in an F1 Grand Prix: the winner gets 25 points and the rest 18, 15, 10, 8, 6 and 4 successively. The final average scores are given in Table 4. The winner and the second best methods in each category appear marked.

1. The maxima-hunting methods are the global winners (in particular when using the distance correlation measure), even if there is still room for improvement in the maxima identification. In fact, the maxima-hunting procedures result in accuracy improvements (when using the whole trajectories) in 88.00% of the considered experiments. Overall, the gain of accuracy associated with **MHR** variable selection is relevant (2.41%).

2. The univariate ranking methods, such as the t ranking (which ignore the dependence between the involved variables), are clearly outperformed by the "functional" procedures. The superiority of the maxima-hunting methods over the rest of variable selection procedures, while requiring often a lesser number of variables, is remarkable.

3. As an overall conclusion, variable selection appears as a **highly competitive alternative to PLS**, which is so far the standard dimension reduction method in high-dimensional and functional statistics (whenever a response variable is involved). The results of the above rankings show that variable selection offers a better balance in terms of both accuracy and interpretability.

4. On average, the use of the classical Fisher's discriminant rule LDA (after dimension reduction) provides worse results than the nonparametric k-NN rule. An example of superiority of a linear classifier is shown in Delaigle and Hall (2012b) where an asymptotic optimality result is provided. In addition, under some conditions, the proposed classifier turns out to be "near-perfect" (in the sense that the probability of classification error can be made arbitrarily small) to discriminate between two Gaussian processes. This is an interesting phenomenon which does not appear in the finite dimensional case. However, it requires that the Gaussian measures under discrimination are mutually singular (note that this situation cannot happen with two non-degenerate Gaussian measures in $\mathbb{R}^d$). This topic will be considered in a forthcoming manuscript by the authors.

A final remark. The present study shows that there are several quite natural models in which the maxima-hunting method is definitely to be recommended. The data results are also encouraging. Our results suggest that, even when there is no clear, well-founded guess on the nature of the underlying model, the idea of selecting the maxima of the distance correlation is a suitable choice; it always allows for a direct interpretation. It is natural to ask what type of models would typically be less favorable for the maxima-hunting approach. As a rough, practical guide, we might say that adverse situations might typically arise in those cases where the trajectories are extremely smooth, or when they are very wiggly, with many noisy abrupt peaks which tend to mislead the calculation of the maxima in the distance correlation function.

Supplementary Materials

Some further methodological and technical details are explained in the Supplementary Materials document. It also includes the proofs as well as some extra simulation outputs and the list of the 100 considered models. The full simulation outputs are included in an Excel file downloadable from `www.uam.es/antonio.cuevas/exp/outputs.xlsx`.

Acknowledgement

This research has been supported by Spanish grant MTM2013-44045-P. The useful comments from two referees are gratefully acknowledged.

References

Athreya, K. B. and Lahiri, S. N. (2006). *Measure Theory and Probability Theory*. Springer.

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

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

Berrendero, J. R., Cuevas, A. and Torrecilla, J. L. (2015). The mRMR variable selection method: a comparative study for functional data. To appear in *J. Statist. Comput. Simulation* DOI: 10.1080/00949655.2015.1042378.

Biau, G., Cadre, B. and Paris, Q. (2015). Cox process functional learning. *Stat. Inference Stoch. Process.* **18**, 257-277.

Comminges L. and Dalalyan, A. S. (2012). Tight conditions for consistency of variable selection in the context of high dimensionality. *Ann. Statist.* **40**, 2667-2696.

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

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

Delaigle, A. and Hall, P. (2012b). Achieving near perfect classification for functional data. *J. Roy. Statist. Soc. Ser. B* **74**, 267-286.

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

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

Ding, C. and Peng, H. (2005). Minimum redundancy feature selection from microarray gene expression data. *J. Bioinform. Comput. Biol.* **3**, 185-205.

Fan, J. and Lv, J. (2010). A selective overview of variable selection in high dimensional feature space. *Statist. Sinica* **20**, 101-148.

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

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

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

Galeano, P. and Joseph, E. and Lillo, R. E. (2014). The Mahalanobis distance for functional data with applications to classification. To appear in *Technometrics*.

Gertheiss, J. and Tutz, G. (2010). Sparse modeling of categorical explanatory variables. *Ann. Appl. Statist.* **4**, 2150-2180.

Guyon, I., Gunn, S., Nikravesh, M. and Zadeh, L. A. (2006). *Feature Extraction: Foundations and Applications*. Springer-Verlag, Berlin Heidelberg.

Hall, P. and Miller, H. (2011). Determining and depicting relationships among components in high-dimensional variable selection. *J. Comput. Graph. Statist.* **20** 988-1006.

Hastie, T. and Tibshirani, R. and Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference and Prediction*. Springer, New York.

Hsing, T. and Ren, H. (2009). An RKHS formulation of the inverse regression dimension reduction problem. *Ann. Statist.* **37**, 726-755.

Jiang, B and Liu, J. S. (2014). Variable selection for general index models via sliced inverse regression. *Ann. Statist.* **42**, 1751-1786.

Jiang, C.R., Yu, W. and Wang, J. L. (2013). Inverse regression for longitudinal data, *Ann. Statist.* **42**, 563-591.

Kneip, A. and Sarda, P. (2011). Factor models and variable selection in high-dimensional regression analysis. *Ann. Statist.* **39**, 2410-2447.

Li, R., Zhong, W. and Zhu, L. (2012). Feature screening via distance correlation learning. *J. Am. Statist. Assoc.* **107**, 1129-1139.

Li, Y., Wang, N. and Carroll, R. J. (2013). Selecting the number of principal components in functional data. *J. Amer. Statist. Assoc.* **108**, 1284-1294.

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

McKeague, I. W. and Sen, B. (2010). Fractals with point impact in functional linear regression. *Ann. Statist.* **38**, 2559-2586.

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

Mosler, K. and Mozharovskyi, P. (2014). Fast DD-classification of functional data. arXiv preprint arXiv:1403.1158.

Peng, H., Long, F. and Ding, C. (2005). Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Trans. Pattern Anal. Mach. Intell.* **27**, 1226-1238.

Preda, C. and Saporta, G. and Lévéder, C. (2007). PLS classification of functional data. *Comput. Statist.* **22**, 223-235.

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

Rosasco, L., Villa, S., Mosci, S. and Santoro, M. (2013). Nonparametric sparsity and regularization *J. Mach. Learn. Res.* **14**, 1665-1714.

Székely, G. J. and Rizzo, M. L. (2009). Brownian distance covariance. *Ann. Appl. Statist.* **3**, 1236-1265.

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

Zhang, X., Park, B. U. and Wang, J. L. (2013). Time-varying additive models for longitudinal data. *J. Amer. Statist. Assoc.* **108**, 983-998.

Zhao, Y., Chen, H. and Ogden, R. T. (2014). Wavelet-based weighted LASSO and screening approaches in functional linear regression. To appear in *J. Comput. Graph. Statist.*

Departamento de Matemáticas, Universidad Autónoma de Madrid, 28049-Madrid, Spain.
E-mail: joser.berrendero@uam.es

Departamento de Matemáticas, Universidad Autónoma de Madrid, 28049-Madrid, Spain.
E-mail: antonio.cuevas@uam.es

Departamento de Matemáticas, Universidad Autónoma de Madrid, 28049-Madrid, Spain.
E-mail: joseluis.torrecilla@uam.es

(Received August 2014; accepted May 2015)