Variable Selection in Functional Data Classification: a Maxima-Hunting Proposal

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

Universidad Autónoma de Madrid

Supplementary Material

S1 Introduction

This document includes extended versions of some sections of our paper "Variable selection in functional data classification: a maxima-hunting proposal". For the sake of brevity, many details about the empirical results (Sections 5 to 7) were omitted in that paper. Moreover, the proofs of all the theoretical results were also removed. The present document aims at giving the interested reader some design considerations, implementation aspects and empirical results omitted in the main paper together with further theoretical details and proofs. We refer to the original paper and references therein for additional details.

This document is organized as follows: in Section S2 we present the methods involved in the simulation study and the criteria for comparing them. Section S3 is devoted to general implementation details such as data representation and validation procedures. The simulated models are extensively described in Section S4 and some empirical results are presented in Section S5. Section S6 is devoted to real data and Section S7 contains some tentative rankings of the methods. Section S8 includes the proofs of all the theoretical results of the paper and a couple of auxiliary lemmas. An appendix contains the full list of the models considered in the simulations. Let us recall that the complete empirical outputs are collected in `www.uam.es/antonio.cuevas/exp/outputs.xlsx`.

S2 The variable selection methods under study. Criteria for comparisons

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

1. **Maxima hunting**. The methods based on the estimation of the maxima of $\mathcal{R}^2$ and $\mathcal{V}^2$ are implemented as follows. 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 $\mathcal{R}$ or $\mathcal{V}$. This relevance criterion and an alternative domain criterion (sorting by the length of the interval where the maximum is global maximum) are illustrated in Figure S1. Our empirical results (not included in this study) show that the use of this domain criterion does not lead, on average, to any improvement with respect to the relevance ordering.

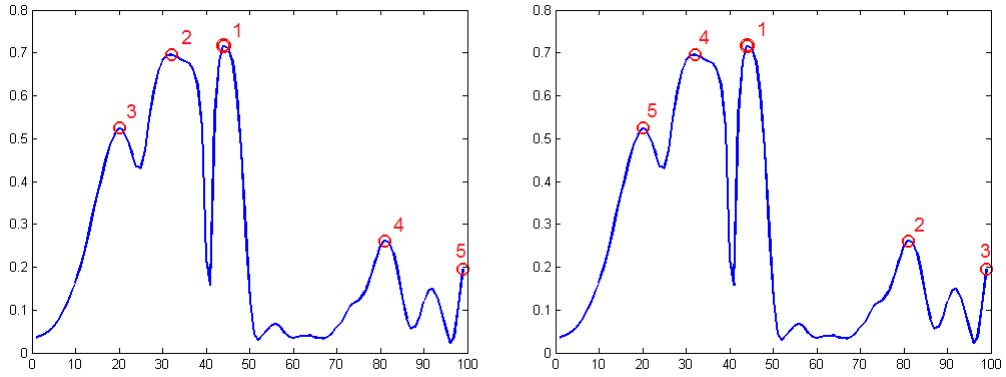

Figure S1: Blue line stands for $\mathcal{R}^2(X(t), Y)$ for the first derivative of the Tecator data. The first five selected maxima are marked in red and the selection order is indicated by the red number beside each maximum. On the left picture it is used the relevance criterion while the domain one is applied on the right graph. In this noiseless case identification by relevance is preferable.

2. **Univariate t-ranking method**, denoted by **T**, 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

$$T(X_t) = \frac{|\bar{X}_{1t} - \bar{X}_{0t}|}{\sqrt{s_{1t}^2/n_1 + s_{0t}^2/n_0}}.$$

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. It aims at maximizing the relevance of the selected variables avoiding an excess of redundancy which 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.
- A three-fold discretized version of the so-called *Mutual Information* measure for both relevance and redundancy (see equation (1) in Ding and Peng (2005)).

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 the results of **FCQ** (Fisher Correlation Quotient) and **MID** (Mutual Information Difference) which outperform on average their corresponding counterparts (FCD and MIQ). All four approaches outputs are shown on the web.

4. **DHB**. In the comparisons with real data sets we have incorporated the method recently proposed by Delaigle, Hall and Bathia (2012). We will 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 implement 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.

5. **PLS**. According to the available results (Preda, Saporta and Lévéder (2007); Delaigle and Hall (2012)) 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 do we loss by restricting ourselves to variable selection methods, instead of considering other more general linear projections procedures (as PLS) for dimension reduction.

6. **Base**. Variable selection methods are also compared with the functional k-NN classifier applied to the entire curves. In general, the Base performance can be seen as a reference to assess the usefulness of dimension reduction methods. Somewhat surprisingly, this Base procedure is often outperformed by variable selection methods.

The classifiers used in all cases are the nearest neighbor rule k-NN, based on the Euclidean distance, and the linear discriminant analysis (LDA). As motivated in the Introduction similar comparisons could be done with other classifiers, since the considered methods (except DHB) do not depend on the classifier. For comparing the different methods we use the classification accuracy, measured by the percentage of correct classification for the k-NN and LDA classifiers,

based only on the selected variables (or the PLS projections). Recall that LDA apply only for the reduced data since it typically fails with functional data.

S3 Some implementation details

Our empirical study required the implementation of all methods described above, including mRMR, DHB and PLS, as well as the classifiers k-NN and LDA. All algorithms have been implemented in MATLAB. The code is available upon request. We are also working on a user-friendly R library. Here are some algorithmic details:

- Our k-NN implementation allows for the use of different distances; we use the usual Euclidean distance. Also, the computation for different k's can be simultaneously made with no additional cost.

- We have implemented the minimum Redundancy Maximum Relevance algorithm in order to allow us to introduce different association measures (such as the distance correlation) in the definition of the method. The original version of mRMR (based on the mutual information measure) is available from http://penglab.janelia.org/proj/mRMR/. Also, a MATLAB/C++ function (not compatible with current MATLAB versions) can be also downloaded from that URL.

- We have implemented the original iterative PLS algorithm that can be found, e.g. in Helland (1988).

- We use the empirical estimators of distance correlation and distance covariance given by Székely, Rizzo and Bakirov (2007), which are also implemented in an efficient way. It can be seen that this estimator is asymptotically equivalent to that of Theorem 2. Therefore, the uniform convergence is also guaranteed.

- The DHB algorithm has been implemented according to the instructions given in Delaigle, Hall and Bathia (2012). We have also used the same parameters and the first stopping criterion proposed by these authors.

S4 The structure of the simulation study

Our simulation study consists of 400 experiments, aimed at comparing the practical performances of several intrinsic variable selection methods described in the previous subsection. These experiments are 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 take $p=\mathbb{P}(Y=1)=1/2$.

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

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

1. **Gaussian models**: they are denoted $G1, G1b, \ldots,$ $G8$. In all these models the distributions of $X(t)|Y_i$ are chosen among one of the following types: first, the **standard Brownian Motion**, B, in $[0,1]$, i.e., a Gaussian process with $\mathbb{E}(B(t)) = 0$ and covariance function $\gamma(s,t) = \min\{s,t\}$. Second, **Brownian Motion, BT, with a trend** $m(t)$, i.e., $BT(t) = B(t) + m(t)$; we have considered several choices for $m(t)$, a linear trend, $m(t) = ct$, a linear trend with random slope, i.e., $m(t) = \theta t$, where θ is a Gaussian r.v., and different members of two parametric families: the *peak* functions $\Phi_{m,k}$ and the *hillside* functions, defined by

$$\Phi_{m,k} = \int_0^t \varphi_{m,k}(s)ds \quad , \quad \text{hillside}_{t_0,b}(t) = b(t-t_0)\mathbb{I}_{[t_0,\infty)},$$

where, $\varphi_{m,k}(t) = \sqrt{2^{m-1}}\left[\mathbb{I}_{\left(\frac{2k-2}{2^m},\frac{2k-1}{2^m}\right)} - \mathbb{I}_{\left(\frac{2k-1}{2^m},\frac{2k}{2^m}\right)}\right]$ for $m \in \mathbb{N}$, $1 \leq k \leq 2^{m-1}$. Third, the **Brownian bridge**: $BB(t) = B(t) - tB(1)$. Our fourth class of Gaussian processes is the **OrnsteinUhlenbeck process**, with a covariance function of type $\gamma(s,t) = a\exp(-b|s-t|)$ and zero mean (OU) or different mean functions $m(t)$ (OUt). Finally smoother processes have been also computed by convolving Brownian trajectories with Gaussian kernels. We have considered two levels of smoothing denoted by sB and ssB.

2. **Logistic models**: they are defined through the general pattern (M2): the process $X = X(t)$ follows one of the above mentioned distributions and $Y \sim \text{Binom}(1, \eta(X))$ with $\eta(x) = (1 + e^{-\Psi(x(t_1),\cdots,x(t_d))})^{-1}$, a function of the relevant variables $x(t_1), \cdots, x(t_d)$. We have 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, all the selection methods are checked for four sample sizes (n = 30, 50, 100, 200). In this way we get $100 \times 4 = 400$ experiments.

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

The parameters involved 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 are 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)]. In the case of the simulation study, the validation and test samples are randomly generated. In the real data sets we proceed by cross-validation.

In summary, the methodology used in the simulation study is as follows (see also the flowchart in Figure S2):

1. Each output in the tables is based on an average over 200 runs.
2. In each run of the simulation experiments three samples are generated: the training sample of size n (= 30, 50, 100, 200), a validation sample of size 200 and a test sample of size 200.
3. The relevant variables are selected using the training sample.
4. The validation sample is used for the choice of the parameters.
5. The "accuracy" outputs in the tables correspond to the percentages of correct classification obtained for the test samples. In all cases the classification is done either via a k-NN classifier, applied to the "reduced data" (i.e., to the selected variables or the PLS projections), or via the linear classifier LDA.

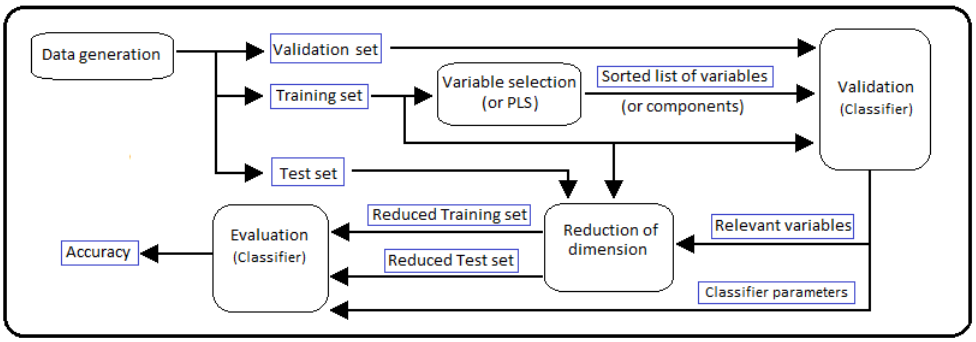

Figure S2: Methodology flowchart for simulations. This process is repeated 200 times for each experiment.

S5 A few numerical outputs from the simulations

First Table S1 shows the number of selected variables (or PLS components) associated with Table S1 in the paper. Models are in rows and methods in columns. Let us recall that the

first two columns correspond to the mRMR method implemented with different association measures. T denotes the univariate t-ranking method and, as usual, PLS stands for the partial least squares dimension reduction method. Besides, MHR and MHV stand for the maxima-hunting methods described above, based on the distance correlation and the distance covariance, respectively. The method called "Base" refers to the classification success obtained with the complete curves (with a k-NN functional classifier). The marked outputs correspond to the winner and second best method in each row. Next, we present a summary of some results grouping the 400 experiments (averaged over 200 runs) by sample size. The outputs obtained with k-NN are shown in Table S2 while those corresponding to LDA are in Table S3. Each row contains the averages on 100 models with a specific sample size except the last block which summary the results of the 400 experiments. Then, in both tables, blocks 1 to 4 corresponds to $n = 30$ $n = 50$, $n = 100$ and $n = 200$ respectively, and the last block has de average of the 400 examples. Different measures are presented in rows and methods in columns. The measures considered are the average accuracy, the average standard deviation of the classification accuracy and the number of variables (or components) selected in average. Additionally, the proportion of models in which the methods beats the "Base" results, and the average of this gain in terms of accuracy are presented only in Table S2 since the "Base" method cannot be computed with LDA. Columns are organized as in the previous table.

This is a summary of the complete results that allow us to draw some general considerations about the performance of the methods. The reader can consult the Excel tables available online with the entire results, if interested on some particular model. Although tables seem clear enough, let us point out some details. In general terms, Tables S2 and S3 show that MHR is the overall winner on average with a slight advantage, followed by MHV and PLS. PLS and the maxima-hunting methods (MHR and MHV) obtain similar scores and clearly outperform the other methods. Note that they also beat (often very clearly) the Base method in almost all cases using just a few variables. This suggests that dimension reduction would be, in fact, "mandatory" in many cases. Note that these methods obtain improvements close to 2% of the total accuracy with just the $5-6\%$ of original variables. In terms of number of variables, MHR and MHV often need less variables to achieve better results than th e rest of variable selection methods.

Table S2 also shows that the benefits of reducing the dimension (compared to the Base approach) are higher when lower sample sizes are considered. This is a relevant fact since in many practical cases (e.g. in biomedical studies) only small samples are available.

S6 Real data examples

We have chosen three examples due to their popularity in FDA. Here we give a broader description than in the main paper. We start with a summary of some basic features in Table S4. Figure S3 shows the trajectories $X(t)$ and mean functions for each set and each class.

Berkeley Growth Study. These data have been thoroughly analyzed in the monograph by Ramsay and Silverman (2005) and are available in the *fda* package of R. It contains the

Table S1: Average number of selected variables (or components) over 200 runs of the considered methods with sample size $n = 50$

				k-NN outputs			
Models	FCQ	MID	T	PLS	MHR	MHV	Base
L2_OUt	9.96	8.71	11.56	3.41	7.19	7.32	100
L6_OU	12.86	10.84	14.23	3.54	10.77	10.70	100
L10_B	7.79	10.64	8.76	4.96	8.23	7.64	100
L11_ssB	8.52	6.03	9.61	4.06	2.54	2.52	100
L12_sB	7.96	6.67	8.18	4.43	4.51	5.64	100
G1	11.08	7.75	11.91	3.62	5.46	4.49	100
G3	9.69	5.13	11.46	5.84	1.77	1.97	100
G6	11.46	11.53	12.94	5.55	5.22	4.59	100
M2	9.90	12.14	13.46	7.49	6.79	6.15	100
M6	8.73	9.90	9.51	5.51	6.25	6.50	100
M10	8.88	11.43	9.56	4.96	6.82	6.17	100

				LDA outputs			
Models	FCQ	MID	T	PLS	MHR	MHV	Base
L2_OUt	5.64	5.91	6.64	2.00	6.49	6.17	-
L6_OU	6.45	5.63	6.90	1.38	6.40	6.54	-
L10_B	5.10	7.74	4.89	4.47	7.39	6.53	-
L11_ssB	3.00	3.66	3.25	2.28	2.15	2.29	-
L12_sB	3.40	3.48	3.86	1.72	3.44	4.29	-
G1	8.03	6.08	7.87	4.83	6.40	6.69	-
G3	9.07	7.05	8.97	6.52	5.43	5.50	-
G6	13.93	10.74	14.07	4.41	7.62	6.84	-
M2	11.89	13.97	11.80	4.96	11.11	10.58	-
M6	4.66	6.39	4.13	3.19	6.40	6.86	-
M10	9.72	7.71	10.60	3.33	6.59	6.25	-

Table S2: The first four blocks are organized as follows. Row 1: average accuracy (proportion of correct classification over 200 runs) over 100 models. Row 2: average standard deviation of the classification accuracy (in 200 runs) over 100 simulation models. Row 3: average number of variables (or components) used by the method. Row 4: proportion of experiments where the method outperforms the Base. Row 5: average gain over the Base method. The training sample size is indicated at the headboard of each block. The last block ("All simulations") provides the global averages. All outputs correspond to the k-NN classifier.

k-NN outputs

$n = 30$	FCQ	MID	T	PLS	MHR	MHV	Base
Avg. acc.(%)	79.65	80.09	79.16	81.42	81.87	81.53	78.98
Avg. acc. std.(%)	4.18	4.20	4.36	3.52	3.68	3.77	3.84
Avg. dim. red.	9.50	9.20	9.90	4.28	6.21	6.26	100
Victories over Base(%)	58.00	71.00	51.00	77.00	95.00	89.00	
Avg. gain over Base(%)	0.67	1.11	0.18	2.44	2.89	2.55	
$n = 50$	FCQ	MID	T	PLS	MHR	MHV	Base
Avg. acc.(%)	80.40	81.43	79.84	82.48	82.89	82.59	80.34
Avg. acc. std.(%)	3.93	3.67	4.04	3.16	3.41	3.45	3.35
Avg. dim. red.	9.63	9.28	10.07	4.77	6.15	6.24	100
Victories over Base(%)	53.00	71.00	46.00	76.00	91.00	89.00	
Avg. gain over Base(%)	0.06	1.09	-0.50	2.14	2.55	2.25	
$n = 100$	FCQ	MID	T	PLS	MHR	MHV	Base
Avg. acc.(%)	81.34	83.01	80.71	83.79	84.21	83.87	81.99
Avg. acc. std.(%)	3.64	3.23	3.74	2.84	3.16	3.25	2.95
Avg. dim. red.	9.89	9.58	10.25	5.50	6.10	6.11	100
Victories over Base(%)	49.00	71.00	38.00	77.00	86.00	81.00	
Avg. gain over Base(%)	-0.65	1.02	-1.28	1.80	2.22	1.88	
$n = 200$	FCQ	MID	T	PLS	MHR	MHV	Base
Avg. acc.(%)	82.09	84.28	81.27	84.84	85.37	84.96	83.38
Avg. acc. std.(%)	3.47	2.93	3.57	2.63	3.05	3.09	2.69
Avg. dim. red.	10.07	9.75	10.43	6.22	5.81	5.76	100
Victories over Base(%)	42.00	73.00	33.00	72.00	80.00	75.00	
Avg. gain over Base(%)	-1.28	0.90	-2.11	1.46	1.99	1.58	
All simulations	FCQ	MID	T	PLS	MHR	MHV	Base
Avg. acc.(%)	80.87	82.20	80.24	83.13	83.58	83.23	81.17
Avg. acc. std.(%)	3.81	3.51	3.93	3.04	3.33	3.39	3.21
Avg. dim. red.	9.77	9.45	10.16	5.19	6.07	6.09	100
Victories over Base(%)	50.50	71.50	42.00	75.50	88.00	83.50	
Avg. gain over Base(%)	-0.30	1.03	-0.93	1.96	2.41	2.06	

Table S3: The first four blocks are organized as follows. Row 1: average accuracy (proportion of correct classification over 200 runs) over 100 models. Row 2: average standard deviation of the classification accuracy (in 200 runs) over 100 simulation models. Row 3: average number of variables (or components) used by the method. The training sample size is indicated at the headboard of each block. The last block ("All simulations") provides the global averages. All outputs correspond to the LDA classifier.

LDA outputs

$n = 30$	FCQ	MID	T	PLS	MHR	MHV	Base
Avg. acc.(%)	77.58	78.72	76.77	81.04	80.66	80.71	-
Avg. acc. std.(%)	4.37	4.47	4.48	3.58	4.08	4.06	-
Avg. dim. red.	4.71	5.61	4.92	2.73	5.51	5.43	-
$n = 50$	FCQ	MID	T	PLS	MHR	MHV	Base
Avg. acc.(%)	78.53	80.28	77.77	81.86	81.81	81.73	-
Avg. acc. std.(%)	3.95	3.86	4.02	3.20	3.66	3.62	-
Avg. dim. red.	5.68	6.52	5.89	2.99	6.14	6.14	-
$n = 100$	FCQ	MID	T	PLS	MHR	MHV	Base
Avg. acc.(%)	79.62	81.85	78.93	82.71	82.99	82.81	-
Avg. acc. std.(%)	3.56	3.33	3.62	2.91	3.23	3.24	-
Avg. dim. red.	7.08	7.93	7.44	3.48	6.99	7.04	-
$n = 200$	FCQ	MID	T	PLS	MHR	MHV	Base
Avg. acc.(%)	80.47	82.96	79.83	83.39	83.83	83.53	-
Avg. acc. std.(%)	3.31	2.98	3.35	2.71	3.03	3.03	-
Avg. dim. red.	8.33	9.03	8.88	4.17	7.48	7.46	-
All simulations	FCQ	MID	T	PLS	MHR	MHV	Base
Avg. acc.(%)	79.05	80.95	78.33	82.25	82.32	82.20	-
Avg. acc. std.(%)	3.80	3.66	3.87	3.10	3.50	3.49	-
Avg. dim. red.	6.45	7.27	6.78	3.34	6.53	6.52	-

Table S4: Description of the real datasets: n is the number of observations of dimension *dim*. "Base acc." represents the proportion of classification success obtained with the complete curves (with a k-NN functional classifier).

Dataset	n	*dim.*	Base acc.	References
Growth	93	31	96.77%	Ramsay and Silverman (2005)
Tecator (2nd derivative)	215	100	98.60%	Ferraty and Vieu (2006)
Phoneme (binary)	1717	256	78.97%	Hastie, Tibshirani and Friedman (2009)

heights of 54 girls and 39 boys measured at 31 non-equally distant time points. It has been used in many classification studies, see e.g. Mosler and Mozharovskyi (2014) for a recent summary.

Tecator. This is another well-known data set used many times as a benchmark for comparisons in FDA studies. It is available via the *fda.usc* R package. It consists of 215 near-infrared absorbance spectra of finely chopped meat, obtained using a Tecator Infratec Food & Feed Analyzer. Thus the final data set is made of 215 curves, observed at 100 equispaced points, ranging from 850 to 1050 nm with associated values of moisture, fats and protein contents. Following Ferraty and Vieu (2006), the sample is separated in two classes according to the fat content (smaller or larger than 20%). A particularity of Tecator dataset is the high homogeneity among the raw data, which makes the classification problem harder. For this reason, these data are often used in a differentiated version, that is, they are smoothed (e.g., via splines) and then the first or the second derivative of the smoothed curves is used (e.g. the cited monograph Ferraty and Vieu (2006)). We show here the results corresponding to the second derivatives (which turn out to provide a higher discrimination power than the raw data or the first derivative). A recent review of classification performances for different methods is given in Galeano, Joseph and Lillo (2014).

Phoneme. These are data of speech recognition, discussed in Hastie, Buja and Tibshirani (1995). They can be downloaded from `www-stat.stanford.edu/ElemStatLearn` and are analyzed in Hastie, Tibshirani and Friedman (2009) and Ferraty and Vieu (2006). The original sample has 4509 curves, corresponding to log-periodograms constructed from 32 ms long recordings of males pronouncing five phonemes. Each curve was observed at 256 equispaced points. This five-classes discrimination problem is adapted to our binary setup by taking just (as in Delaigle, Hall and Bathia (2012)) the curves corresponding to the phonemes "aa" and "ao". The sample size is $n = 1717$ (695 from "aa" and 1022 from "ao").

In this real data setting we use the same methods and follow the same methodology as in the simulation study with the only exception of the generation of the training, validation and test samples. Here we consider the usual cross-validation procedure which avoids splitting the sample (sometimes small) into three different sets. Each output is obtained by standard leave-one-out cross-validation. The only exception is the phoneme data set for which this procedure

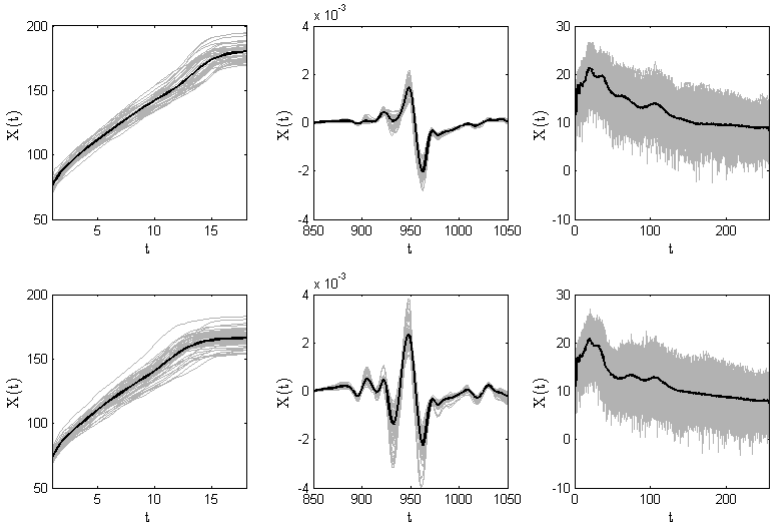

Figure S3: Data trajectories and mean functions from class 0 (first row) and class 1 (second row). Columns correspond to growth, Tecator and phoneme data from left to right.

is extremely time-consuming (due to the large sample size); so we use instead ten-fold cross-validation (10CV). The respective validation steps are done with the same resampling schemes within the training samples. This is an usual way to proceed when working with real data; see e.g. Hastie, Tibshirani and Friedman (2009, Subsection 7.10) for further details. Several outputs obtained with the three considered data sets are given in Tables S5 (accuracy) and S6 (number of variables) below. The complete results (incorporating some additional dimension reduction methods) can be found in the last sheet of the excel file `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 as 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. Note also that maxima-hunting methods (particularly MHR) outperform or are very close to the Base outputs (which uses the entire curves). PLS is overcome by our methods in two of the three problems but it is the clear winner in 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 two first considered examples but relatively fails in the phoneme case. There is maybe some room for improvement in the stopping criterion (recall that we have used the same parameters as in Delaigle, Hall and Bathia (2012)). Recall also that, 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).

Table S5: Classification accuracy (in percentages) for the real data with both classifiers (k-NN above and LDA below).

				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 S6: Average number of variables (or components) selected for the real data sets using both classifiers (k-NN above and LDA below).

				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	-

Also note that the use of the LDA classifier didn't lead to any significant gain; in fact, the results are globally worse than those of k-NN except for a few particular cases.

Although in principle our methodology is not primarily targeted to the best classification rate, but to the choice of the most representative variables, we can conclude that maxima-hunting procedures combined with the simple k-NN are competitive when compared with PLS and other successful and sophisticated methods in 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.

S7 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. We have considered the **classification accuracy**, that is, the proportion of correct classification obtained with the different procedures. Thus, the global scores of the different methods with respect to this criterion along the 400 experiments are computed in three alternative ways. According to the **relative ranking** assessment, 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: $10(u-w)/(W-w)$. Second, the **positional ranking** scoring criterion just gives 10 points to the winner in every experiment, 9 points to the second one, etc. Finally, the **F1 ranking** rewards strongly the winner. For each experiment, points are divided as in a F1 Grand Prix, that is, the winner receives 25 points and the rest 18, 15, 10, 8, 6 and 4 successively. The final scores are given in Table S7 below. For each sample size, the results are the average over the 100 experiments corresponding to that sample size. The winner and the second best classifier in each category appear marked in the table. Also, the relative ranking scores of the full 400 models are shown in Figure S4.

S8 Some results and proofs

To prove Theorem 2 we need two lemmas dealing with the uniform strong consistency of one-sample and two-sample functional U-statistics, respectively.

Lemma 1. *Let $X : T \to \mathbb{R}$ be a process with continuous trajectories a.s. defined on the compact rectangle $T = \prod_{i=1}^{d}[a_i, b_i] \subset \mathbb{R}^d$. Let $X_1, \ldots, X_n$ be a sample of n independent trajectories of X. Define the functional U-statistic*

$$U_n(t) = \frac{2}{n(n-1)} \sum_{i<j} k[X_i(t), X_j(t)],$$

where the kernel k is a real continuous, permutation symmetric function. Assume that

$$\mathbb{E}\big(\sup_{t\in T} |k[X(t), X'(t)]|\big) < \infty,$$

Table S7: Final scores of the considered methods for the simulation experiments. The rankings correspond to the observed performances in classification accuracy. The individual scores are in turn combined according to three different ranking criteria (proportional, positional and F1).

				k-NN rankings			
Ranking criterion	FCQ	MID	T	PLS	MHR	MHV	Base
Relative ($n = 30$)	4.66	4.79	3.61	6.94	8.64	7.64	2.68
Relative ($n = 50$)	4.62	5.45	3.25	6.94	8.50	7.48	3.25
Relative ($n = 100$)	4.37	6.23	2.71	7.06	8.35	7.21	3.97
Relative ($n = 200$)	4.04	6.72	2.15	7.02	8.19	7.06	4.64
Relative (whole)	4.42	5.80	2.93	6.99	8.42	7.35	3.64
Positional ($n = 30$)	6.52	6.22	5.59	7.93	9.09	8.06	5.59
Positional ($n = 50$)	6.55	6.50	5.64	7.90	8.72	7.95	5.74
Positional ($n = 100$)	6.42	6.83	5.48	8.03	8.58	7.72	5.98
Positional ($n = 200$)	6.26	7.30	5.27	7.96	8.34	7.62	6.25
Positional (whole)	6.44	6.71	5.50	7.96	8.68	7.84	5.89
F1 ($n = 30$)	11.64	10.58	9.54	17.37	19.55	15.93	9.39
F1 ($n = 50$)	12.01	11.27	9.77	17.29	18.12	15.80	9.74
F1 ($n = 100$)	11.58	12.39	9.51	17.71	17.46	15.03	10.41
F1 ($n = 200$)	11.24	13.90	9.01	17.19	16.71	14.89	11.06
F1 (whole)	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 ($n = 30$)	3.57	3.46	1.79	7.60	8.15	8.11	-
Relative ($n = 50$)	3.74	4.61	1.89	7.20	8.60	8.16	-
Relative ($n = 100$)	3.83	5.95	1.90	6.70	8.96	8.18	-
Relative ($n = 200$)	3.89	6.74	2.27	6.09	8.78	7.83	-
Relative (whole)	3.76	5.19	1.96	6.90	8.62	8.07	-
Positional ($n = 30$)	6.75	6.51	5.71	8.54	8.75	8.74	-
Positional ($n = 50$)	6.72	6.71	5.87	8.39	8.80	8.52	-
Positional ($n = 100$)	6.72	7.15	5.92	7.95	8.79	8.47	-
Positional ($n = 200$)	6.62	7.58	6.18	7.63	8.81	8.23	-
Positional (whole)	6.70	6.99	5.92	8.13	8.79	8.49	-
F1 ($n = 30$)	11.96	11.12	9.58	19.08	17.95	18.31	-
F1 ($n = 50$)	11.91	11.68	10.12	18.57	18.33	17.41	-
F1 ($n = 100$)	12.20	12.92	10.24	16.64	18.58	17.42	-
F1 ($n = 200$)	11.74	14.35	10.92	15.66	18.76	16.74	-
F1 (whole)	11.95	12.52	10.22	17.49	18.41	17.47	-

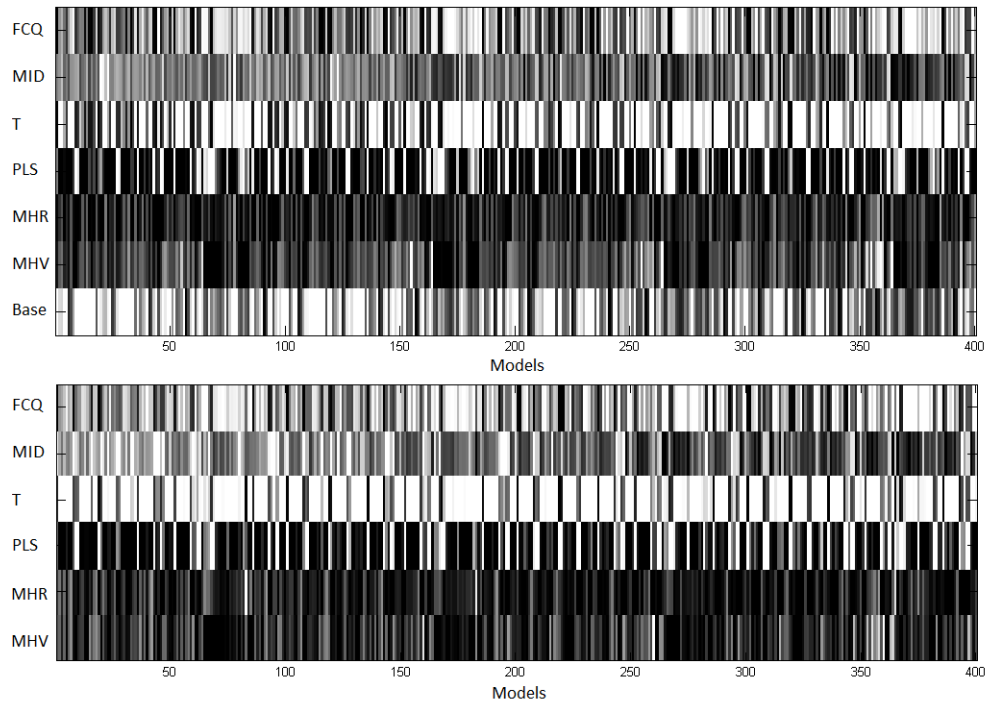

Figure S4: Display of relative ranking scores, the darker the better (black corresponds to 10 and white 0). Each column represents a simulation model and each file corresponds to a dimension reduction method. The ranking outputs are obtained with both k-NN (first display) and LDA (second display) classifiers. Maxima-hunting with $\mathcal{R}$ is often the best and never the worst.

where X and X' denote two independent copies of the process. Then, as $n \to \infty$, $\|U_n - U\|_\infty \to 0$, a.s., where $U(t) = \mathbb{E}(k[X(t), X'(t)])$.

Proof. First, we show that $U(t)$ is continuous. Let $t_n \subset T$ such that $t_n \to t$. Then, due to the continuity assumptions on the process and the kernel, $k[X(t_n), X'(t_n)] \to k[X(t), X'(t)]$, a.s. Using the assumption $\mathbb{E}\big(\sup_{t\in T} |k[X(t), X'(t)]|\big) < \infty$, Dominated Convergence Theorem (DCT) allows us to deduce $U(t_n) \to U(t)$.

Let $M_\delta(t) = \sup_{s:|s-t|_d \leq \delta} |h(s) - h(t)|$ where, for the sake of simplicity, we denote $h(t) = k[X(t), X'(t)]$. The next step is to prove that, as $\delta \downarrow 0$,

$$\sup_{t\in T} \mathbb{E}(M_\delta(t)) \to 0. \tag{S8.1}$$

Both $M_\delta(t)$ and $\lambda_\delta(t) = \mathbb{E}(M_\delta(t))$ are continuous functions. Since $h(t)$ is uniformly continuous on $\{s : |s - t|_d \leq \delta\}$, $M_\delta(t)$ is also continuous. The fact that $\lambda_\delta(t)$ is continuous follows directly from DCT since $|M_\delta(t)| \leq 2 \sup_{t\in T} |h(t)|$ and, by assumption, $\mathbb{E}(\sup_{t\in T} |h(t)|) < \infty$. By continuity, $M_\delta(t) \to 0$ and $\lambda_\delta(t) \to 0$, as $\delta \downarrow 0$. Now, since $\delta > \delta'$ implies $\lambda_\delta(t) \geq \lambda_{\delta'}(t)$, for

all $t \in T$, we can apply Dini's Theorem to deduce that $\lambda_\delta(t)$ converges uniformly to 0, that is, $\sup_{t\in T} \lambda_\delta(t) \to 0$, as $\delta \downarrow 0$.

The last step is to show $\|U_n - U\|_\infty \to 0$ a.s., as $n \to \infty$. For $i \neq j$, denote $M_{ij,\delta}(t) = \sup_{s:|s-t|_d<\delta} |h_{ij}(s) - h_{ij}(t)|$, where $h_{ij}(t) = k[X_i(t), X_j(t)]$, and $\lambda_\delta(t) = \mathbb{E}(M_{ij,\delta}(t))$. Fix $\epsilon > 0$. By (S8.1), there exists $\delta > 0$ such that $\lambda_\delta(t) < \epsilon$, for all $t \in T$. Now, since T is compact, there exist $t_1, \ldots, t_m$ in T such that $T = \cup_{k=1}^m B_k$, where $B_k = \{t : |t - t_k|_d \leq \delta\} \cap T$. Then,

$$\begin{aligned}
\|U_n - U\|_\infty &= \max_{1\leq k\leq m} \sup_{t\in B_k} |U_n(t) - U(t)| \\
&\leq \max_{1\leq k\leq m} \sup_{t\in B_k} [|U_n(t) - U_n(t_k)| + |U_n(t_k) - U(t_k)| + |U(t_k) - U(t)|] \\
&\leq \max_{1\leq k\leq m} \sup_{t\in B_k} |U_n(t) - U_n(t_k)| + \max_{k=1,\ldots,m} |U_n(t_k) - U(t_k)| + \epsilon,
\end{aligned}$$

since $|s - t|_d \leq \delta$ implies $|U(s) - U(t)| = |\mathbb{E}[h(s) - h(t)]| \leq \mathbb{E}|h(s) - h(t)| \leq \lambda_\delta(t) < \epsilon$.

For the second term, we have $\max_{k=1,\ldots,m} |U_n(t_k) - U(t_k)| \to 0$ a.s., as $n \to \infty$, applying SLLN for U-statistics (see e.g. DasGupta (2008), Theorem 15.3(b), p. 230). As for the first term, observe that using again SLLN for U-statistics,

$$\begin{aligned}
\sup_{t\in B_k} |U_n(t) - U_n(t_k)| &\leq \frac{2}{n(n-1)} \sum_{i<j} \sup_{t\in B_k} |h_{ij}(t_k) - h_{ij}(t)| \\
&= \frac{2}{n(n-1)} \sum_{i<j} M_{ij,\delta}(t_k) \to \lambda_\delta(t_k), \quad \text{a.s.},
\end{aligned}$$

where $\lambda_\delta(t_k) < \epsilon$. Therefore,

$$\begin{aligned}
\limsup_n \|U_n - U\|_\infty &\leq \limsup_n \max_{k=1,\ldots,m} \sup_{t\in B_k} |U_n(t) - U_n(t_k)| \\
&\quad + \limsup_n \max_{k=1,\ldots,m} |U_n(t_k) - U(t_k)| + \epsilon \leq 2\epsilon.
\end{aligned}$$

□

Lemma 2. *Let $X^{(0)} : T \to \mathbb{R}$ and $X^{(1)} : T \to \mathbb{R}$ be a pair of independent processes with continuous trajectories a.s. defined on the compact rectangle $T = \prod_{i=1}^d [a_i, b_i] \subset \mathbb{R}^d$. Let $X_1^{(0)}, \ldots, X_{n_0}^{(0)}$ and $X_1^{(1)}, \ldots, X_{n_1}^{(1)}$ be samples of n_0 and n_1 independent trajectories of $X^{(0)}$ and $X^{(1)}$, respectively. Define the functional two-sample U-statistic*

$$U_{n_0,n_1}(t) = \frac{1}{n_0 n_1} \sum_{i=1}^{n_0} \sum_{j=1}^{n_1} k[X_i^{(0)}(t), X_j^{(1)}(t)],$$

where the kernel k is a continuous, permutation symmetric function. Assume that

$$\mathbb{E}\big(\sup_{t\in T} |h(t)| \log^+ |h(t)|\big) < \infty,$$

with $h(t) = k[X^{(0)}(t), X^{(1)}(t)]$. *Then, as* $\min(n_0, n_1) \to \infty$,

$$\|U_{n_0,n_1} - U\|_\infty \to 0, \quad a.s.,$$

where $U(t) = \mathbb{E}(k[X^{(0)}(t), X^{(1)}(t)])$.

Proof. It is analogous to the proof of Lemma 1 so it is omitted. We need to apply a strong law of large numbers for two-sample U-statistics. This result can be guaranteed under slightly stronger conditions on the moments of the kernel; see Sen (1977, Th.1). Hence the condition $\mathbb{E}\big(\sup_{t\in T} |h(t)| \log^+ |h(t)|\big) < \infty$ in the statement of the lemma. □

Proofs of the results of the main document

In what follows all the equation labels [(2.1), (3.1), etc.] not starting with S refer to equations in the main body of the paper.

Theorem 1. (a) From (2.1), as X_t is d-dimensional and Y is one-dimensional, taking into account $c_1 = \pi$, we have

$$\begin{aligned}\mathcal{V}^2(X_t, Y) =& \| \varphi_{X_t,Y}(u,v) - \varphi_{X_t}(u)\varphi_Y(v) \|_w^2 \\ &= \tfrac{1}{\pi c_d} \int_{\mathbb{R}} \int_{\mathbb{R}^d} |\varphi_{X_t,Y}(u,v) - \varphi_{X_t}(u)\varphi_Y(v)|^2 \tfrac{1}{|u|_d^{d+1} v^2} du dv.\end{aligned}$$

Let's analyze the integrand,

$$\begin{aligned}\varphi_{X_t,Y}(u,v) - \varphi_{X_t}(u)\varphi_Y(v) &= \mathbb{E}\left[e^{iu^\top X_t} e^{ivY}\right] - \mathbb{E}\left[e^{iu^\top X_t}\right] \mathbb{E}\left[e^{ivY}\right] \\ &= \mathbb{E}\left[(e^{iu^\top X_t} - \varphi_{X_t}(u))(e^{ivY} - \varphi_Y(v))\right] \\ &= \mathbb{E}\left[\mathbb{E}\left[(e^{iu^\top X_t} - \varphi_{X_t}(u))(e^{ivY} - \varphi_Y(v))|X\right]\right] \\ &= \mathbb{E}\left[(e^{iu^\top X_t} - \varphi_{X_t}(u))\mathbb{E}\left[(e^{ivY} - \varphi_Y(v))|X\right]\right] \\ &\overset{(*)}{=} \mathbb{E}\left[(e^{iu^\top X_t} - \varphi_{X_t}(u))(e^{iv} - 1)(\eta(X) - p)\right] \\ &= (e^{iv} - 1)\mathbb{E}\left[(e^{iu^\top X_t} - \varphi_{X_t}(u))(\eta(X) - p)\right] \\ &= (e^{iv} - 1)\mathbb{E}\left[e^{iu^\top X_t}(\eta(X) - p)\right] = (e^{iv} - 1)\zeta(u,t).\end{aligned}$$

Step (*) in the above chain of equalities is motivated as follows:

$$\begin{aligned}\mathbb{E}\left[(e^{ivY} - \varphi_Y(v))|X\right] &= \mathbb{E}\left[e^{ivY}|X\right] - \varphi_Y(v) = (e^{iv} - 1)\eta(X) - (e^{iv} - 1)p \\ &= (e^{iv} - 1)((\eta(X) - p)).\end{aligned}$$

Therefore, since $\int_{\mathbb{R}} \frac{|e^{iv}-1|^2}{\pi v^2} dv = 2$,

$$\mathcal{V}^2(X_t, Y) = \int_{\mathbb{R}} \frac{|e^{iv} - 1|^2}{\pi v^2} dv \int_{\mathbb{R}^d} \frac{|\zeta(u,t)|^2}{c_d |u|_d^{d+1}} du = \frac{2}{c_d} \int_{\mathbb{R}^d} \frac{|\zeta(u,t)|^2}{|u|_d^{d+1}} du.$$

(b) Since $\zeta(u,t) = \mathbb{E}\left[(\eta(X)-p)\, e^{iu^\top X_t}\right]$,

$$\begin{aligned}
|\zeta(u,t)|^2 &= \mathbb{E}\left[(\eta(X)-p)e^{iu^\top X_t}\right]\mathbb{E}\left[(\eta(X')-p)e^{-iu^\top X_t'}\right] \\
&= \mathbb{E}\left[(\eta(X)-p)(\eta(X')-p)e^{iu^\top (X_t-X_t')}\right] \\
&= \mathbb{E}\left[(\eta(X)-p)(\eta(X')-p)\cos(u^\top(X_t-X_t'))\right] \\
&= -\mathbb{E}\left[(\eta(X)-p)(\eta(X')-p)(1-\cos(u^\top(X_t-X_t')))\right],
\end{aligned}$$

where we have used $|\zeta(u,t)|^2 \in \mathbb{R}$ and $\mathbb{E}\left[(\eta(X)-p)(\eta(X')-p)\right] = 0$. Now, using expression (3.1),

$$\begin{aligned}
\mathcal{V}^2(X_t,Y) &= -2\mathbb{E}\left[(\eta(X)-p)(\eta(X')-p)\int_{\mathbb{R}^d}\frac{1-\cos(u^\top(X_t-X_t'))}{c_d|u|_d^{d+1}}du\right] \\
&= -2\mathbb{E}\left[(\eta(X)-p)(\eta(X')-p)|X_t-X_t'|_d\right] \\
&= -2\mathbb{E}\left[(Y-p)(Y'-p)|X_t-X_t'|_d\right],
\end{aligned}$$

since [see e.g. Lemma 1 in Székely, Rizzo and Bakirov (2007)],

$$\int_{\mathbb{R}^d}\frac{1-\cos(u^\top x)}{c_d|u|_d^{d+1}}du = |x|_d, \quad \text{for all } x\in\mathbb{R}^d.$$

(c) By conditioning on Y and Y' we have

$$\begin{aligned}
\mathbb{E}[(Y-p)(Y'-p)|X_t-X_t'|_d] &= p^2 I_{00}(t)(1-p)^2 - p(1-p)I_{01}(t)2p(1-p) \\
&\quad + (1-p)^2 I_{11}(t)p^2 = p^2(1-p)^2(I_{00}(t)+I_{11}(t)-2I_{01}(t)).
\end{aligned}$$

Now, using (3.2), $\mathcal{V}^2(X_t,Y) = 4p^2(1-p)^2\left[I_{01}(t) - \frac{I_{00}(t)+I_{11}(t)}{2}\right]$. □

Theorem 2. Continuity of $\mathcal{V}_n^2(X_t,Y)$ is straightforward from DCT. It suffices to prove the result for sequences of samples $X_1^{(0)},\ldots,X_{n_0}^{(0)}$, and $X_1^{(1)},\ldots,X_{n_1}^{(1)}$, drawn from $X|Y=0$ and $X|Y=1$, respectively, such that $n_1/(n_0+n_1)\to p = \mathbb{P}(Y=1)$.

From the triangle inequality it is enough to prove the uniform convergence of $\hat{I}_{00}(t)$, $\hat{I}_{11}(t)$ and $\hat{I}_{01}(t)$ to $I_{00}(t)$, $I_{11}(t)$ and $I_{01}(t)$, respectively. For the first two quantities we apply Lemma 1 to the kernel $k(x,x') = |x-x'|$. For the last one we apply Lemma 2 to the same kernel. Observe that $\mathbb{E}\|X\|_\infty < \infty$ implies the moment condition of Lemma 1 whereas $\mathbb{E}(\|X\|_\infty \log^+\|X\|_\infty) < \infty$ implies the moment condition of Lemma 2. The last statement readily follows from the uniform convergence and the compactness of $[0,1]^d$. □

Proposition 1. We know $g^*(x) = \mathbb{I}_{\{\eta(x)>1/2\}}$. Then, we use equation (4.1), which provides $\eta(x)$ in terms of the Radon-Nikodym derivative $d\mu_0/d\mu_1$, and the expression for $d\mu_0/d\mu_1$ given in Liptser and Shiryayev (1977), p. 239. This gives

$$\eta(x) = \left[\frac{1-p}{p}\sqrt{2}e^{-x_1^2/4}+1\right]^{-1}.$$

Now, from $g^*(x) = \mathbb{I}_{\{\eta(x)>1/2\}}$, we get $g^*(x) = 1$ if and only if $x_1^2 > 4\log\left(\frac{\sqrt{2}(1-p)}{p}\right)$. □

Proposition 2. Again, we use expression (4.1) to derive the expression of the optimal rule $g^*(x) = \mathbb{I}_{\{\eta(x)>1/2\}}$. In this case the calculation is made possible using the expression of the Radon-Nikodym derivative for the distribution of a Brownian process with trend, $F(t) + B(t)$, with respect to that of a standard Brownian:

$$\frac{d\mu_1}{d\mu_0}(B) = \exp\left\{-\frac{1}{2}\int_0^1 F'(s)^2 ds + \int_0^1 F' dB\right\}, \tag{S8.2}$$

for μ_0-almost all $B \in \mathcal{C}[0,1]$; see, Mörters and Peres (2010), Th. 1.38 and Remark 1.43, for further details. Observe that in this case we have $F(t) = ct$. Thus, from (4.1), we finally get $\eta(x) = \left[\frac{1-p}{p}\exp\left(\frac{c^2}{2} - cx_1\right) + 1\right]^{-1}$, which again only depends on x through $x(1) = x_1$. The result follows easily from this expression. □

Proposition 3. In this case, the trend function is $F(t) = \Phi_{m,k}(t)$. So $F^{'}(t) = \varphi_{m,k}$ and $F^{''}(t) = 0$. From equations (4.1) and (S8.2), we readily get (4.3) and (4.4). □

Proposition 4. Let us first consider the model in Proposition 1 (i.e., Brownian vs. Brownian with a stochastic trend). Such model entails that $X_t|Y = 0 \sim N(0, \sqrt{t})$ and $X_t|Y = 1 \sim N(0, \sqrt{t^2+t})$. Now, recall that if $\xi \sim N(m, \sigma)$, then,

$$\mathbb{E}|\xi| = \sigma\sqrt{\frac{2}{\pi}}e^{-\frac{m^2}{\sigma^2}} + m\left(2\Phi\left(\frac{m}{\sigma}\right) - 1\right), \tag{S8.3}$$

where $\Phi(z)$ denotes the distribution function of the standard normal.

Now, using (3.3) and (S8.3) we have the following expressions,

$$I_{01}(t) = \mathbb{E}|\sqrt{t}Z - \sqrt{t^2+t}Z'| = \sqrt{\frac{2(t^2+2t)}{\pi}},$$

$$I_{00}(t) = \mathbb{E}|\sqrt{t}Z - \sqrt{t}Z'| = \sqrt{\frac{4t}{\pi}},$$

$$I_{11}(t) = \mathbb{E}|\sqrt{t^2+t}Z - \sqrt{t^2+t}Z'| = \sqrt{\frac{4(t^2+t)}{\pi}},$$

where Z and Z' are independent $N(0,1)$ random variables.

Then, the function $\mathcal{V}^2(X_t, Y) = 4p^2(1-p)^2\left(I_{01}(t) - \frac{I_{00}(t)+I_{11}(t)}{2}\right)$ grows with t so it is maximized at $t^* = 1$, which is the only point that has an influence on the Bayes rule.

Let us now consider the model in Proposition 2 (i.e., Brownian vs. Brownian with a linear trend). Again, from (S8.3) we have in this case,

$$I_{01}(t) = \mathbb{E}|ct + \sqrt{t}Z - \sqrt{t}Z'| = 2\sqrt{\frac{t}{\pi}}e^{-\frac{c^2 t}{2}} + ct\left(2\Phi\left(c\sqrt{\frac{t}{2}}\right) - 1\right),$$

$$I_{00}(t) = I_{11}(t) = \mathbb{E}|\sqrt{t}Z - \sqrt{t}Z'| = \sqrt{\frac{4t}{\pi}},$$

where Z and Z' are iid standard Gaussian variables. Therefore using (3.3),

$$\mathcal{V}^2(X_t, Y) = C\left[2\sqrt{\frac{t}{\pi}}\left(e^{-\frac{c^2 t}{2}} - 1\right) + ct\left(2\Phi\left(c\sqrt{\frac{t}{2}}\right) - 1\right)\right],$$

where $C = 4p^2(1-p)^2$. We can check numerically that this an increasing function which reaches its only maximum at $t^* = 1$. According to Proposition 1 this is the only relevant point for the Bayes rule. □

Appendix: models

We now list all the models included in the simulation study. The relevant variables are indicated in brackets (for Gaussian and mixture models) or in the expression of $\psi(X)$ (for the logistic-type models). Variables in bold face had found to be specially relevant.

1. **Gaussian** models:

1. **G1** : $\begin{cases}\mu_0 : & B(t) \\ \mu_1 : & B(t) + \theta t \quad , \theta \sim N(0,3)\end{cases}$
   $variables = \{X_{100}\}$.
2. **G1b** : $\begin{cases}\mu_0 : & B(t) \\ \mu_1 : & B(t) + \theta t \quad , \theta \sim N(0,5)\end{cases}$
   $variables = \{X_{100}\}$.
3. **G2** : $\begin{cases}\mu_0 : & B(t) + t \\ \mu_1 : & B(t)\end{cases}$
   $variables = \{X_{100}\}$.
4. **G2b** : $\begin{cases}\mu_0 : & B(t) + 3t \\ \mu_1 : & B(t)\end{cases}$
   $variables = \{X_{100}\}$.
5. **G3** : $\begin{cases}\mu_0 : & BB(t) \\ \mu_1 : & B(t)\end{cases}$
   $variables = \{X_{100}\}$.
6. **G4** : $\begin{cases}\mu_0 : & B(t) + hillside_{0.5,4}(t) \\ \mu_1 : & B(t)\end{cases}$
   $variables = \{X_{47}, \boldsymbol{X}_{100}\}$.
7. **G5** : $\begin{cases}\mu_0 : & B(t) + 3\Phi_{1,1}(t) \\ \mu_1 : & B(t)\end{cases}$
   $variables = \{X_1, \boldsymbol{X}_{48}, X_{100}\}$.
8. **G6** : $\begin{cases}\mu_0 : & B(t) + 5\Phi_{2,2}(t) \\ \mu_1 : & B(t)\end{cases}$
   $variables = \{X_{48}, \boldsymbol{X}_{75}, X_{100}\}$.
9. **G7** : $\begin{cases}\mu_0 : & B(t) + 5\Phi_{3,2}(t) + 5\Phi_{3,4}(t) \\ \mu_1 : & B(t)\end{cases}$
   $variables = \{X_{22}, \boldsymbol{X}_{35}, X_{49}, X_{74}, \boldsymbol{X}_{88}, X_{100}\}$.
10. **G8** : $\begin{cases}\mu_0 : & B(t) + 3\Phi_{2,1.25}(t) + 3\Phi_{2,2}(t) \\ \mu_1 : & B(t)\end{cases}$
    $variables = \{X_9, \boldsymbol{X}_{35}, X_{48}, X_{62}, \boldsymbol{X}_{75}, X_{100}\}$.

2. These are the ψ functions used to define the different **logistic** models,

L1: $\psi(X) = 10X_{65}$.

L2: $\psi(X) = 10X_{30} + 10X_{70}$.

L3: $\psi(X) = 10X_{30} - 10X_{70}$.

L4: $\psi(X) = 20X_{30} + 50X_{50}20X_{80}$.

L5: $\psi(X) = 20X_{30} - 50X_{50} + 20X_{80}$.

L6: $\psi(X) = 10X_{10} + 30X_{40} + 10X_{72} + 10X_{80} + 20X_{95}$.

L7: $\psi(X) = \sum_{i=1}^{10} 10X_{10i}$.

L8: $\psi(X) = 20X_{30}^2 + 10X_{50}^4 + 50X_{80}^3$.

L9: $\psi(X) = 10X_{10} + 10|X_{50}| + 0X_{30}^2X_{85}$.

L10: $\psi(X) = 20X_{33} + 20|X_{68}|$.

L11: $\psi(X) = \frac{20}{X_{35}} + \frac{30}{X_{77}}$.

L12: $\psi(X) = \log X_{35} + \log X_{77}$.

L13: $\psi(X) = 40X_{20} + 30X_{28} + 20X_{62} + 10X_{67}$.

L14: $\psi(X) = 40X_{20} + 30X_{28} - 20X_{62} - 10X_{67}$.

L15: $\psi(X) = 40X_{20} - 30X_{28} + 20X_{62} - 10X_{67}$.

The variations included are,

L3b: $\psi(X) = 30X_{30} - 20X_{70}$.

L4b: $\psi(X) = 30X_{30} + 20X_{50} + 10X_{80}$.

L5b: $\psi(X) = 10X_{30} - 10X_{50} + 10X_{80}$.

L6b: $\psi(X) = 20X_{10} + 20X_{40} + 20X_{72} + 20X_{80} + 20X_{95}$.

L8b: $\psi(X) = 10X_{30}^2 + 10X_{50}^4 + 10X_{80}^3$.

3. **Mixture** models:

1. **M1** : $\begin{cases} \mu_0: & \begin{cases} B(t)+3t & ,1/2 \\ B(t)-2t & ,1/2 \end{cases} \\ \mu_1: & B(t) \end{cases}$
   $variables = \{X_{100}\}$.

2. **M2** : $\begin{cases} \mu_0: & \begin{cases} B(t)+3\Phi_{2,2}(t) & ,1/2 \\ B(t)+5\Phi_{3,2}(t) & ,1/2 \end{cases} \\ \mu_1: & B(t) \end{cases}$
   $variables = \{X_{22},\boldsymbol{X}_{35}, X_{48},\boldsymbol{X}_{75}, X_{100}\}$.

3. **M3** : $\begin{cases} \mu_0: & \begin{cases} B(t)+3\Phi_{2,2}(t) & ,1/10 \\ B(t)+5\Phi_{3,2}(t) & ,9/10 \end{cases} \\ \mu_1: & B(t) \end{cases}$
   $variables = \{X_{22},\boldsymbol{X}_{35}, X_{48},\boldsymbol{X}_{75}, X_{100}\}$.

4. **M4**: $\begin{cases} \mu_0: & \begin{cases} B(t)+3\Phi_{2,2}(t) & ,1/2 \\ B(t)+5\Phi_{3,3}(t) & ,1/2 \end{cases} \\ \mu_1: & B(t) \end{cases}$
   $variables = \{X_{48},\boldsymbol{X}_{62},\boldsymbol{X}_{75}, X_{100}\}$.

5. **M5** : $\begin{cases} \mu_0: & \begin{cases} B(t)+3\Phi_{2,1}(t) & ,1/3 \\ B(t)+3\Phi_{2,2}(t) & ,1/3 \\ B(t)+5\Phi_{3,2}(t) & ,1/3 \end{cases} \\ \mu_1: & B(t) \end{cases}$
   $variables = \{X_1,\boldsymbol{X}_{22},\boldsymbol{X}_{35}, X_{48},\boldsymbol{X}_{75}, X_{100}\}$.

6. **M6** : $\begin{cases} \mu_0: & \begin{cases} B(t)+3\Phi_{2,1}(t) & ,1/2 \\ B(t)+3t & ,1/2 \end{cases} \\ \mu_1: & B(t) \end{cases}$
   $variables = \{X_1,\boldsymbol{X}_{22}, X_{49},\boldsymbol{X}_{100}\}$.

7. **M7** : $\begin{cases} \mu_0: & \begin{cases} B(t)+3\Phi_{1,1}(t) & ,1/2 \\ BB(t) & ,1/2 \end{cases} \\ \mu_1: & B(t) \end{cases}$
   $variables = \{X_1,\boldsymbol{X}_{48},\boldsymbol{X}_{100}\}$.

8. **M8** : $\begin{cases} \mu_0: & \begin{cases} B(t)+\theta t, \ \ \theta\sim N(0,5) & ,1/2 \\ B(t)+hillside_{0.5,5}(t) & ,1/2 \end{cases} \\ \mu_1: & B(t) \end{cases}$
   $variables = \{X_{47},\boldsymbol{X}_{100}\}$.

9. **M9** : $\begin{cases} \mu_0: & \begin{cases} B(t)+\theta t, \ \ \theta\sim N(0,5) & ,1/2 \\ BB(t) & ,1/2 \end{cases} \\ \mu_1: & B(t) \end{cases}$
   $variables = X_{100}$.

10. **M10** : $\begin{cases} \mu_0: & \begin{cases} B(t)+3\Phi_{1,1}(t) & ,1/3 \\ B(t)-3t & ,1/3 \\ BB(t) & ,1/3 \end{cases} \\ \mu_1: & B(t) \end{cases}$
    $variables = \{X_1,\boldsymbol{X}_{48},\boldsymbol{X}_{100}\}$.

11. **M11** : $\begin{cases} \mu_0: & \begin{cases} B(t)+3\Phi_{1,1}(t) & ,1/4 \\ B(t)-3t & ,1/4 \\ B(t)+hillside_{0.5,5}(t) & ,1/4 \\ BB(t) & ,1/4 \end{cases} \\ \mu_1: & B(t) \end{cases}$
    $variables = \{X_1,\boldsymbol{X}_{48},\boldsymbol{X}_{100}\}$.

Finally, the full list of models involved is, in summary, as follows:

1. L1 OU
2. L1 OUt
3. L1 B
4. L1 sB
5. L1 ssB
6. L2 OU
7. L2 OUt
8. L2 B
9. L2 sB
10. L2 ssB
11. L3 OU
12. L3b OU
13. L3 OUt
14. L3b OUt
15. L3 B
16. L3b B
17. L3 sB
18. L3 ssB
19. L4 OU
20. L4b OU
21. L4 OUt
22. L4b OUt
23. L4 B
24. L4 sB
25. L4 ssB
26. L5 OU
27. L5b OU
28. L5 OUt

29. L5 B
30. L5 sB
31. L5 ssB
32. L6 OU
33. L6b OU
34. L6 OUt
35. L6b OUt
36. L6 B
37. L6 sB
38. L6 ssB
39. L7 OU
40. L7b OU
41. L7 OUt
42. L7b OUt
43. L7 B
44. L7 sB
45. L7 ssB
46. L8 B
47. L8 sB
48. L8 ssB
49. L8b OU
50. L9 B
51. L9 sB
52. L9 ssB
53. L10 OU
54. L10 B
55. L10 sB
56. L10 ssB
57. L11 OU
58. L11 OUt
59. L11 B
60. L11 sB
61. L11 ssB
62. L12 OU
63. L12 OUt
64. L12 B
65. L12 sB
66. L12 ssB
67. L13 OU
68. L13 OUt
69. L13 B
70. L13 sB
71. L13 ssB
72. L14 OU
73. L14 OUt
74. L14 B
75. L14 sB
76. L15 OU
77. L15 OUt
78. L15 B
79. L15 sB
80. G1
81. G1b
82. G2
83. G2b
84. G3
85. G4
86. G5
87. G6
88. G7
89. G8
90. M1
91. M2
92. M3
93. M4
94. M5
95. M6
96. M7
97. M8
98. M9
99. M10
100. M11

References

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

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

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. *Statistica Sinica* **20**, 101–148.

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

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

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

Hastie, T., Buja, A. and Tibshirani, R. (1995). Penalized discriminant analysis. *Ann. Statist.* **23**, 73–102.

Hastie, T. and Tibshirani, R. and Friedman, J.. (2009). The elements of statistical learning: data mining, inference and prediction. Springer.

Helland, I. S.. (1988). On the structure of partial least squares regression *Communications in statistics-Simulation and Computation.* **17(2)**, 581–607.

Liptser, R. S. and Shiryayev, A. N. (1977). *Statistics of random processes.* Springer-Verlag.

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

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 *Computational Statistics.* **22**, 223–235.

Ramsay, J.O. and Silverman, B.W. (2005). *Functional data analysis.* Springer.

Sen, P.K. (1977). Almost sure convergence of generalized U-statistics, *Ann. Probab.* **5**, 287–290.

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