Supplementary material for the paper "On the use of reproducing kernel Hilbert spaces in functional classification", by Berrendero, Cuevas and Torrecilla

S1 An additional result. The proofs of all theorems

S1.1 A consistency result

We first establish the consistency result mentioned in comment (c) of the Subsection 5.1 in the paper.

Theorem S1. *Consider the classification problem (with $p = 1/2$) according to the model (3), for $t \in [0,T]$. Denote $\hat{m}(t) = \hat{m}_1(t) - \hat{m}_0(t)$, where $\hat{m}_j(t) := n_j^{-1}\sum_{i=1}^{n_j} X_{j,i}(t) = \bar{X}_j(t)$ for $j = 0, 1$, and let $\hat{K}_{t_1,\ldots,t_d}$ be the pooled sample covariance matrix, whose (i,j) entry is*

$$\hat{K}_{t_1,\ldots,t_d}(i,j) = \frac{1}{n_1+n_2}\sum_{r\in\{0,1\}}\left(\sum_{\ell=1}^{n_r}(X_{r,\ell}(t_i) - \bar{X}_r(t_i))(X_{r,\ell}(t_j) - \bar{X}_r(t_j))\right).$$

Assume,

(i) $\mathbb{E}\|\epsilon_j^2\|_\infty < \infty$, *for* $j = 0, 1$, *where* $\|\cdot\|_\infty$ *stands for the supremum norm.*

(ii) The variable selection method is performed on a compact set $\Theta \subset [0,T]^d$.

(iii) $K_{t_1,\ldots,t_d}$ *is invertible for all* $(t_1,\ldots,t_d) \in \Theta$ *and their entries are continuous on* Θ.

Then, $L_n \to L^*$ *a.s., as* $n \to \infty$.

Proof of Theorem S1. For the sake of conciseness, denote $\tau := (t_1,\ldots,t_d)$, a generic element of Θ, $\hat{\tau} := (\hat{t}_1,\ldots,\hat{t}_d)$, and $\tau^* := (t_1^*,\ldots,t_d^*)$. We will also use the following notation: for $j = 0, 1$,

$$\tilde{\psi}_j(\tau) := \frac{\left(2(m_{j,\tau} - \hat{\mu}_\tau)^\top \hat{K}_\tau^{-1}\hat{m}_\tau\right)^2}{\hat{m}_\tau^\top \hat{K}_\tau^{-1} K_\tau \hat{K}_\tau^{-1}\hat{m}_\tau},$$

where $m_{j,\tau} := (m_j(t_1),\ldots,m_j(t_d))^\top$ and $\hat{\mu}_\tau = (\hat{m}_{0,\tau} + \hat{m}_{1,\tau})/2$. Then, if μ_τ denotes the population counterpart of $\hat{\mu}_\tau$ and $\psi(\tau) := \frac{\left(2(m_\tau - \mu_\tau)^\top K_\tau^{-1} m_\tau\right)^2}{m_\tau^\top K_\tau^{-1} m_\tau}$ it is not difficult to show that $L^* = 1 - \Phi(\psi(\tau^*)^{1/2}/2)$, and

$$L_n = 1 - \frac{1}{2}\Phi\left(\frac{\tilde{\psi}_0(\hat{\tau})^{1/2}}{2}\right) - \frac{1}{2}\Phi\left(\frac{\tilde{\psi}_1(\hat{\tau})^{1/2}}{2}\right),$$

where Φ is the cumulative distribution function of the standard Gaussian distribution (to obtain these formulas we have used the arguments in Mardia et al. (1980) p. 321, for L^*, and Fan and Fan (2008), p. 2609, for L_n). Since Φ is continuous, the desired conclusion will readily follow if we prove $\tilde{\psi}_j(\hat{\tau}) \to \psi(\tau^*)$ as $n \to \infty$, a.s., for $j = 0, 1$.

Since $\mathbb{E}\|\epsilon_j\|_\infty < \infty$, for $j = 0, 1$, Mourier's Strong Law of Large Numbers (SLLN) for random elements taking values in Banach spaces (see e.g. Laha and Rohatgi (1979), p. 452) implies

$$\sup_{\tau\in\Theta} \|\hat{m}_\tau - m_\tau\| \to 0, \quad \text{as } n \to \infty, \quad \text{a.s.} \tag{S1}$$

Since $\mathbb{E}\|\epsilon_j^2\|_\infty < \infty$ for $j = 0, 1$, Mourier's SLLN also implies that the entries of $\hat{K}_\tau$ converge uniformly to those of K_τ, that is for $i, j = 1, \ldots, d$,

$$\sup_{\tau\in\Theta} |\hat{K}_\tau(i,j) - K_\tau(i,j)| \to 0, \quad \text{as } n \to \infty, \quad \text{a.s.} \tag{S2}$$

Observe that

$$\hat{K}_\tau^{-1} = \frac{\text{adj}(\hat{K}_\tau)}{\det(\hat{K}_\tau)},$$

where $\text{adj}(K)$ and $\det(K)$ denote the adjugate and the determinant of a matrix K, respectively. By (S2), the entries of $\text{adj}(\hat{K}_\tau)$ converge uniformly to those of $\text{adj}(K_\tau)$, and $\det(\hat{K}_\tau)$ converges uniformly to $\det(K_\tau)$. Moreover, $\inf_{\tau\in\Theta} \det(K_\tau) > 0$ because $\det(K_\tau)$ is continuous in τ and, by assumption, $\det(K_\tau) > 0$, for all $\tau \in \Theta$, where Θ is a compact set. As a consequence of all these observations,

$$\sup_{\tau\in\Theta} |\hat{K}_\tau^{-1}(i,j) - K_\tau^{-1}(i,j)| \to 0, \quad \text{as } n \to \infty, \quad \text{a.s.} \tag{S3}$$

By (S1) and (S3), it also holds

$$\sup_{\tau\in\Theta} \|\hat{K}_\tau^{-1}\hat{m}_\tau - K_\tau^{-1}m_\tau\| \to 0, \quad \text{as } n \to \infty, \quad \text{a.s.}$$

From this convergence, together with (S1), we deduce

$$\sup_{\tau\in\Theta} |\hat{\psi}(\tau) - \psi(\tau)| \to 0, \quad \text{as} \quad n \to \infty, \quad \text{a.s.} \tag{S4}$$

and

$$\sup_{\tau\in\Theta} |\tilde{\psi}_j(\tau) - \psi(\tau)| \to 0, \quad \text{as} \quad n \to \infty, \quad \text{a.s.} \quad j = 0, 1. \tag{S5}$$

Due to (S4), with probability one, given $\epsilon > 0$ there exists N such that for $n \geq N$ it holds $\hat{\psi}(\tau) - \epsilon \leq \psi(\tau) \leq \hat{\psi}(\tau) + \epsilon$, for all $\tau \in \Theta$. Taking the maximum in these inequalities we get $\hat{\psi}(\hat{\tau}) - \epsilon \leq \psi(\tau^*) \leq \hat{\psi}(\hat{\tau}) + \epsilon$. That is, we have

$$\hat{\psi}(\hat{\tau}) \to \psi(\tau^*), \quad \text{as } n \to \infty \quad \text{a.s.} \tag{S6}$$

Finally, note that for $j = 0, 1$,

$$|\tilde{\psi}_j(\hat{\tau}) - \psi(\tau^*)| \leq |\tilde{\psi}_j(\hat{\tau}) - \psi(\hat{\tau})| + |\psi(\hat{\tau}) - \hat{\psi}(\hat{\tau})| + |\hat{\psi}(\hat{\tau}) - \psi(\tau^*)|.$$

Then, from (S4), (S5) and (S6) we get $\tilde{\psi}_j(\hat{\tau}) \to \psi(\tau^*)$ as $n \to \infty$, a.s. for $j = 0, 1$, as desired. □

S1.2 Proofs of the theorems and corollaries stated in the paper

Proof of Theorem 2. Equation (4) follows straightforwardly from the combination of (1) and (2). To prove the expression for the Bayes error notice that $\langle X - m_0, m \rangle_K$ lies in $\bar{\mathcal{L}}(X - m_0)$ and therefore the random variable $\eta^*(X)$ is Gaussian both under $Y = 1$ and $Y = 0$. Furthermore, Equations (6.19) and (6.20) in Parzen (1961) yield

$$\begin{aligned}
\mathbb{E}(\eta^*(X)|Y = 0) &= - \parallel m \parallel_K^2 /2 - \log\left(\frac{1-p}{p}\right), \\
\mathbb{E}(\eta^*(X)|Y = 1) &= \parallel m \parallel_K^2 /2 - \log\left(\frac{1-p}{p}\right), \\
\mathrm{Var}(\eta^*(X)|Y = 0) &= \mathrm{Var}(\eta^*(X)|Y = 1) = \parallel m \parallel_K^2 .
\end{aligned}$$

The result follows using these values to standardize the variable $\eta^*(X)$ in $L^* = (1-p)\mathbb{P}(\eta^*(X) > 0|Y = 0) + p\mathbb{P}(\eta^*(X) < 0|Y = 1)$. □

Proof of Theorem 3. We will use the following result

Theorem S2. *[Th. 1 in Shepp (1966)]. Let P_B, P_i be the distributions corresponding to the standard Brownian Motion $\{B(t),\ t \in [0,T]\}$ and to a Gaussian process $\{X(t),\ t \in [0,T]\}$ with mean function m_i in the Dirichlet space $\mathcal{D}[0,T]$ and covariance function K_i. Then $P_i \sim P_0$ if and only if there exists a function $\tilde{K}_i \in L^2([0,T] \times [0,T])$ such that (5) holds, with $1 \notin \sigma(\tilde{K}_i)$, the spectrum of $\tilde{K}_i$.*

We will also need Lemmas 1 and 2 in Shepp (1966), p. 334-335 which give the expression of the Radon-Nikodym derivative dP_i/dP_B in the case $P_i << P_B$ under the conditions of Theorem S2. According to these lemmas, we have for $i = 0, 1$,

$$\frac{dP_i}{dP_B}(X) = \exp\left\{-\frac{1}{2}\sum_{j=1}^{\infty}\left(\log(1-\lambda_{i,j}) + \frac{(x_{i,j} - \xi_{i,j})^2}{1-\lambda_{i,j}} - x_{i,j}^2\right)\right\}.$$

Then, using the chain rule for Radon-Nikodym derivatives:

$$\log \frac{dP_1}{dP_0}(X) = \log \frac{dP_1}{dP_B}(X) - \log \frac{dP_0}{dP_B}(X) = 2\sum_{j=1}^{\infty} \eta_j(X),$$

where $\eta_j(X)$ is defined in Equation (6) in the paper. The result follows from the last expression and (1). □

Proof of Corollary 1. For the Brownian bridge we have $K(s,t) = \min\{s,t\} - st$, so that (5) amounts to

$$\int_0^s \int_0^t \tilde{K}(u,v)dudv = st.$$

As a consequence, $\tilde{K} \equiv 1$. It is not difficult to show that $\lambda = T$ is the only non-zero eigenvalue for $\tilde{K}$ and $\varphi(t) \equiv 1/\sqrt{T}$ is its corresponding unit eigenfunction. From Theorem S2, $P_0 \sim P_1$ if and only if $T < 1$. Moreover, since $m_i(0) = 0$ for $i = 0, 1$ we have $\xi_{i,j} = 0$ for $i = 0, 1$ and $j = 1, 2, \ldots$. Also, the only value $x_{i,j}$ which does not vanish is $x_{1,1} = X(T)/\sqrt{T}$. The corollary is obtained by plugging these values in the expression of the optimal rule provided by Theorem 3. □

Proof of Theorem 5. Observe that, if $\theta_j > 0$ for all $j \geq 1$,

$$m_1 = \sum_{j=1}^{\infty} \mu_j \phi_j = \sum_{j=1}^{\infty} \frac{\mu_j}{\sqrt{\theta_j}} \sqrt{\theta_j}\phi_j,$$

where $\{\sqrt{\theta_j}\phi_j : \theta_j > 0\}$ is an orthonormal basis of $\mathcal{H}(K)$ [see, e.g., Theorem 4.12, p. 61 in Cucker and Zhou (2007)]. Then, by Parseval's formula, $m_1 \in \mathcal{H}(K)$ if and only if $\| m_1 \|_K^2 = \sum_{j=1}^{\infty} \theta_j^{-1}\mu_j^2 < \infty$. As a consequence, we have the desired equivalence:

$$P_1 \sim P_0 \Leftrightarrow m_1 \in \mathcal{H}(K) \Leftrightarrow \| m_1 \|_K < \infty \Leftrightarrow \sum_{j=1}^{\infty} \theta_j^{-1}\mu_j^2 < \infty.$$

Moreover,

$$\text{err}_0 = 1 - \Phi\left(\frac{1}{2}(\sum_{j=1}^{\infty} \theta_j^{-1}\mu_j^2)^{1/2}\right) = 1 - \Phi\left(\frac{1}{2}\| m_1 \|_K\right),$$

what gives the coordinate-free expression of the Bayes error.

Now, if we further assume (as in Delaigle and Hall (2012a)) that $\psi \in L^2$, the optimal classifier proposed by these authors (8) is equivalent to $T^0(X) = 1$ if and only if

$$\langle m_1, \psi\rangle^2_{L^2} - 2\langle m_1, \psi\rangle_{L^2}\langle X, \psi\rangle_{L^2} < 0. \tag{S7}$$

Since $m_1 = \sum_{j=1}^{\infty} \mu_j\phi_j$, with $m_1 \neq 0$, and $\psi = \sum_{j=1}^{\infty} \theta_j^{-1}\mu_j\phi_j$, we have $\langle m_1, \psi\rangle_{L^2} = \sum_{j=1}^{\infty} \theta_j^{-1}\mu_j^2 =\| m_1 \|^2_K \neq 0$. Therefore, (S7) holds if and only if

$$\langle X, \psi\rangle_{L^2} - \frac{\| m_1 \|^2_K}{2} > 0.$$

To end the proof it is enough to show $\langle X, m_1\rangle_K = \langle X, \psi\rangle_{L^2}$. The linearity of $\langle X, \cdot\rangle_K$ and the fact that θ_j and ϕ_j are respectively eigenvalues and eigenfunctions of the integral operator with kernel K imply

$$\langle X, m_1\rangle_K = \sum_{j=1}^{\infty} \theta_j^{-1}\mu_j\langle X, \theta_j\phi_j\rangle_K = \sum_{j=1}^{\infty} \theta_j^{-1}\mu_j \int_0^T \langle X, K(\cdot, u)\rangle_K\phi_j(u)du.$$

Now, from Equation (6.18) in Parzen (1961),

$$\int_0^T \langle X, K(\cdot, u)\rangle_K\phi_j(u)du = \int_0^T X(u)\phi_j(u)du = \langle X, \phi_j\rangle_{L^2}.$$

Finally, combining the two last displayed equations,

$$\langle X, m_1\rangle_K = \sum_{j=1}^{\infty} \theta_j^{-1}\mu_j\langle X, \phi_j\rangle_{L^2} = \langle X, \sum_{j=1}^{\infty} \theta_j^{-1}\mu_j\phi_j\rangle_{L^2} = \langle X, \psi\rangle_{L^2}.$$

□

Proof of Theorem 6. Let $X = \sum_{i=1}^{\infty} Z_i\phi_i$, the Karhunen-Loève expansion of X, with the Z_i uncorrelated. For a given trajectory $x = \sum_{i=1}^{\infty} z_i\phi_i$. Define $x^n = \sum_{i=1}^{n} z_i\phi_i$, This is a trajectory drawn from the process $X^n = \sum_{i=1}^{n} Z_i\phi_i$, whose distribution under P_i is denoted by P_{in} (for $i = 0, 1$, the covariance function is $K_n(s, t) = \sum_{i=1}^{n} \theta_i\phi_i(s)\phi_i(t)$, where $\theta_i = \mathbb{E}(Z_i^2)$, and the mean function under P_{1n} is

$$m_n(t) = \sum_{i=1}^{n} \mathbb{E}(Z_i)\phi_i(t),$$

Note that, under P_0, $\mathbb{E}(Z_j) = 0$, so that the mean function is 0. From Karhunen-Loève Theorem (see Ash and Gardner (1975), p. 38) $m_n(t) \to m(t)$ for all t .

Note also that $m_n \in \mathcal{H}(K)$. Again this follows from the fact that $\{\sqrt{\theta_i}\phi_i : \theta_i > 0\}$ is an orthonormal basis of $\mathcal{H}(K)$ [see, e.g., Theorem 4.12, p. 61 in Cucker and Zhou (2007)].

We now prove that we must necessarily have $\lim_n \|m_n\|_K = \infty$. Indeed, if we had $\lim_n \|m_n\|_K < \infty$ for some subsequence of $\{m_n\}$ (denoted again $\{m_n\}$) we would have that such $\{m_n\}$ would be a Cauchy sequence in $\mathcal{H}(K)$, since for $q > p$, $\|m_p - m_q\|_K^2 = \sum_{i=p+1}^{q} \frac{\mathbb{E}(Z_i)^2}{\theta_i}$. This, together with the pointwise convergence $m_n(t) \to m(t)$ leads, from Corollary 1 (see Berlinet and Thomas-Agnan (2004), p. 10) to $m \in \mathcal{H}(K)$. But, from Parzen's Theorem 1, this would entail $P_1 << P_0$, in contradiction with $P_1 \perp P_0$. We thus conclude $\|m_n\|_K \to \infty$.

Then, given $\epsilon > 0$, choose n such that

$$(1-p)\Phi\left(-\frac{\| m_n \|_K}{2} - \frac{1}{\| m_n \|_K}\log\left(\frac{1-p}{p}\right)\right) + p\Phi\left(-\frac{\| m_n \|_K}{2} + \frac{1}{\| m_n \|_K}\log\left(\frac{1-p}{p}\right)\right) < \epsilon, \tag{S8}$$

Now, consider the problem $X^n \sim P_{1n}$ vs $X^n \sim P_{0n}$ Note that $X^n \sim P_{in}$ if and only if $X \sim P_i$, for $i = 0, 1$. Since $m_n \in \mathcal{H}(K_n)$, we have $P_{0n} \sim P_{1n}$ (using again Parzen's Theorem 1).

Now, according to Theorem 2 (on the expression of the optimal rules in the absolutely continuous case under homoskedasticity), the optimal rule is $g_n(X) = \mathbb{I}_{\{\eta_n(X)>0\}}$, where

$$\eta_n(x) = \langle x, m_n\rangle_K - \frac{1}{2}\| m_n \|_K^2 - \log\left(\frac{1-p}{p}\right), \tag{S9}$$

whose probability of error, is exactly the expression on the left-hand side of (S8). So this probability can be made arbitrarily small. □

Proof of Theorem 7.

(a) Suppose $m(\cdot) = m_1(\cdot) - m_0(\cdot) = \sum_{i=1}^{d} \alpha_i K(\cdot, t_i)$. Then, $m \in \mathcal{H}(K)$ which implies $P_0 \sim P_1$, according to Theorem 1. From Theorem 2, the optimal rule to classify a trajectory x between P_0 and P_1 is $g^*(x) = \mathbb{I}_{\{\eta^*(x)>0\}}$, where $\eta^*(x)$ is given in Equation (4):

$$\begin{aligned}\eta^*(x) =& \langle x - m_0, \sum_{i=1}^{d}\alpha_i K(\cdot, t_i)\rangle_K - \frac{1}{2}\| \sum_{i=1}^{d}\alpha_i K(\cdot, t_i) \|_K^2 - \log\left(\frac{1-p}{p}\right)\\ =& \sum_{i=1}^{d}\alpha_i(x(t_j) - m_0(t_j)) - \frac{1}{2}\sum_{i=1}^{d}\sum_{j=1}^{d}\alpha_i\alpha_j K(t_i, t_j) - \log\left(\frac{1-p}{p}\right),\end{aligned}$$

where we have used the reproducing property to obtain the last equality.

Observe that $m(\cdot) = \sum_{i=1}^{d} \alpha_i K(\cdot, t_i)$ implies the following relationship between $\alpha_1, \ldots, \alpha_d$ and $t_1, \ldots, t_d$:

$$m_{t_1,\ldots,t_d} = K_{t_1,\ldots,t_d} \cdot (\alpha_1, \ldots, \alpha_d)^\top. \tag{S10}$$

Then, we can rewrite the previous expression as

$$\eta^*(x) = \sum_{i=1}^{d} \alpha_i \left(x(t_i) - \frac{m_0(t_i) + m_1(t_i)}{2} \right) - \log\left(\frac{1-p}{p}\right), \tag{S11}$$

which exactly coincides with the discriminant score of the optimal (Bayes) rule for the finite dimensional discrimination problem based on the d-dimensional marginals $(X(t_1), \ldots, X(t_d))$.

(b) If m is given by $m(\cdot) = \sum_{i=1}^{d} \alpha_i K(\cdot, t_i)$, then

$$\|m\|_K^2 = \sum_{i=1}^{d} \sum_{j=1}^{d} \alpha_i \alpha_j K(t_i, t_j) = m_{t_1,\ldots,t_d}^\top K_{t_1,\ldots,t_d}^{-1} m_{t_1,\ldots,t_d}.$$

(c) In Theorem 2 (b) we established that the minimal probability of misclassification was given by

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

But we have just established that $\| m \|_K$ coincides with the Mahalanobis distance between m_0 and m_1. For simplicity, let us denote by Δ such distance. Then, the statement amounts to prove that

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

is a decreasing function of Δ. Indeed, a direct calculation provides

$$\frac{\partial}{\partial \Delta} L^*(\Delta) = -\frac{\sqrt{\frac{1}{p} - 1} p e^{-\frac{4\log^2(\frac{1}{p}-1) + \Delta^4}{8\Delta^2}}}{\sqrt{2\pi}}.$$

Since $\frac{\partial}{\partial \Delta} L^*(\Delta) < 0$ for all Δ, the result follows.

□

S2 Simulations

S2.1 Some practical issues

Recall that, as explained in Section 5.1, the variable selection method considered here (denoted RK-VS) boils down to maximizing the function

$$\hat{\psi}(t_1, \ldots, t_d) := \hat{m}_{t_1,\ldots,t_d}^{\top} \hat{K}_{t_1,\ldots,t_d}^{-1} \hat{m}_{t_1,\ldots,t_d},$$

where $\hat{m}_{t_1,\ldots,t_d}$ are empirical estimators of the difference between the mean vector of $X(t_1), \ldots, X(t_d)|Y = i$, for $i = 0, 1$ and $\hat{K}_{t_1,\ldots,t_d}$ is an estimator of the common covariance matrix.

The algorithm. In general, $\hat{\psi}(t_1, \ldots, t_d)$ is a non-concave function with potentially many local maxima so that the maximization process could be hard to implement even for moderately large values of d. Hence, in practice, we can use the following "greedy" algorithm.

1. Initial step: consider a large enough grid of points in $[0, T]$ and find $\hat{t}_1$ such that $\hat{\psi}(\hat{t}_1) \geq \hat{\psi}(t)$ when t ranges over the grid. Observe that this initial step amounts to find the point maximizing the signal-to-noise ratio since

$$\hat{\psi}(t) = \frac{\hat{m}(t)^2}{\hat{\sigma}_t^2} = \frac{(\bar{X}_1(t) - \bar{X}_0(t))^2}{\hat{\sigma}_t^2},$$

   where $\hat{\sigma}_t^2$ is the considered estimator of the variance at t.

2. Repeat until convergence: once we have computed $\hat{t}_1, \ldots, \hat{t}_{d-1}$, find $\hat{t}_d$ such that
   $\hat{\psi}(\hat{t}_1, \ldots, \hat{t}_{d-1}, \hat{t}_d) \geq \hat{\psi}(\hat{t}_1, \ldots, \hat{t}_{d-1}, t)$ for all t in the grid.

Whereas we have no guarantee that this algorithm converges to the global maximum of $\hat{\psi}(t_1, \ldots, t_d)$, it is computationally affordable and shows good performance in practice.

On the estimation of m and K. In principle (unless some strong parametric assumptions are made), the estimation of $m = m_1 - m_0$ will be done in the simplest way, using the sample means, i.e., $\hat{m} = \hat{m}_1(t) - \hat{m}_0(t)$, where $\hat{m}_j(t) := n_j^{-1} \sum_{i=1}^{n_j} X_{j,i}(t) = \bar{X}_j(t)$, for $j = 0, 1$.

The estimation of $\hat{K}$ might look as a more delicate issue. It is well-known that in some functional data analysis techniques (including functional linear regression and principal components analysis) there is a need to use smooth estimators of the covariance operator K; see, for example, Cuevas (2014, Secs. 5.2 and 7.1). Of course, such smoothed estimators could be also applied here but the underlying

(functional) reasons to use them are not present in this case since in fact we are only concerned with the covariance matrices $K_{t_1,\ldots,t_d}$ of finite dimensional projections $(X(t_1),\ldots,X(t_d))$. Thus, unless otherwise stated, we will estimate $K_{t_1,\ldots,t_d}$ by the natural empirical counterpart $\hat{K}_{t_1,\ldots,t_d}$ constructed from the sample covariances. This has been the method we have used (with overall good results) in our empirical studies.

A natural alternative to such estimators would arise in those cases in which we are assuming a precise parametric model, such as for example a Brownian motion for which $K(s,t) = K(\theta,s,t) = \theta\min(s,t)$ depending on an unknown parameter θ. In such models one could naturally consider parametric estimations of type $K(\hat{\theta},s,t)$. From this point of view, the RK methods are completely flexible allowing for including additional knowledge about the covariance structure. Hence, the appropriate estimator $\hat{K}_{t_1,\ldots,t_d}$ depends on the assumptions we are willing to make about the processes involved in the classification problem; see next subsection for more details on this.

S2.2 An illustrative example. The price of estimating the covariance function

The purpose of this subsection is to gain some practical insight on the meaning and performance of our RK methods. In particular, we will take into account that the RK methods can incorporate information on the assumed underlying model, via a known (or partially known) covariance function. In what follows we will assume that the data trajectories come from a Brownian Motion with different (unknown) mean functions. So we would incorporate this information in our "variable selection + classification" task by just using the, supposedly true, $K(s,t)$, instead of its estimator in (12). We will denote by RK_B-VS and RK_B-C the resulting "oracle" methods for variable selection and classification, respectively, implemented with $K(s,t) = \min\{s,t\}$.

While the assumption that K is known might seem too strong, it is still useful to compare the performance of the oracle RK_B-VS and RK_B-C methods with the standard RK-VS and RK-C versions in which $K(s,t)$ is estimated from the sample; in addition note that, under the model of densely observed functional data, we might even consider a parametric model with $K(s,t) = \theta\min\{s,t\}$, since the scale parameter θ can be estimated with arbitrary precision using just one densely observed trajectory.

In any case, we want to assess the loss of efficiency involved in the estimation of $K(s,t)$. To this end, consider a simulated example under the general model (3) in which P_0 and P_1 are Brownian motions whose mean functions fulfill $m(t) = m_1(t) - m_0(t) = \sum_{i=1}^{r} a_i\Phi_{m,k}(t)$, where $t \in [0,1]$, the a_i are constants and the $\{\Phi_{m,k}\}$ are continuous piecewise linear functions as those considered in Mörters and Peres (2010, p. 28); they are obtained by integrating the piecewise constant functions of

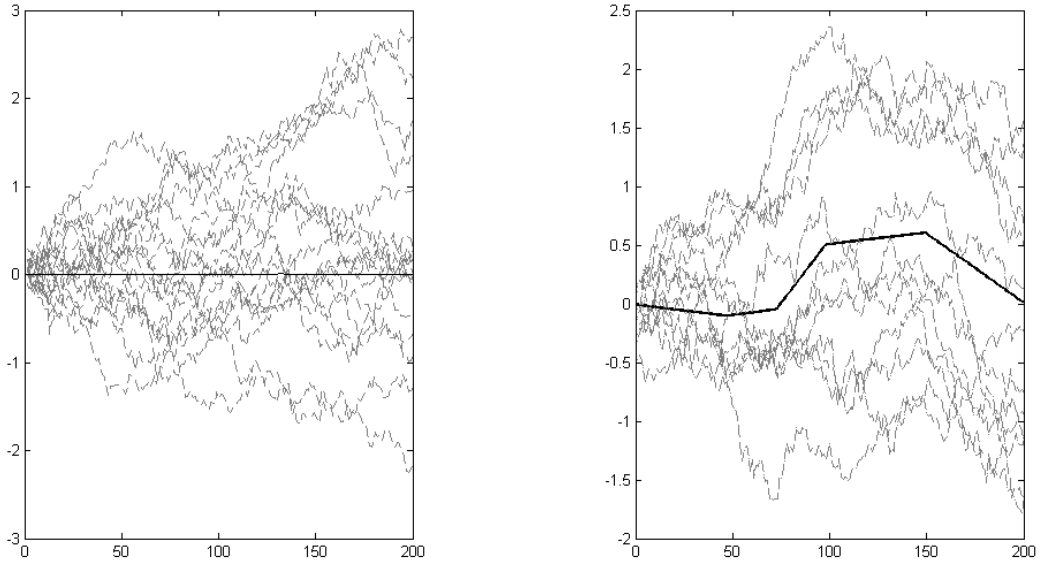

Figure S1: Some trajectories from the toy example $B(t)$ (left) vs $B(t) + \Phi_{1,1}(t) - \Phi_{2,1}(t) + \Phi_{2,2}(t) - \Phi_{3,2}(t)$ (right). Thick solid lines correspond to the mean functions.

the Haar basis. Explicit expressions can be found in Appendix A. In fact, it can be proved there that the $\{\Phi_{m,k}\}$ form a orthonormal basis of the Dirichlet space $\mathcal{D}[0,1]$ which is the RKHS space corresponding to this model. As a consequence, the equivalence condition in Theorem 2 is automatically fulfilled. In addition, given the simple structure of the "peak" functions $\Phi_{m,k}$, it is easy to see that the "sparsity condition" $m(\cdot) = \sum_{i=1}^{d} \alpha_i K(\cdot, t_i)$ also holds in this case. To be more specific, in our simulation experiments we have taken $m_0(t) = 0$, $m_1(t) = \Phi_{1,1}(t) - \Phi_{2,1}(t) + \Phi_{2,2}(t) - \Phi_{3,2}(t)$, and $p = \mathbb{P}(Y = 1) = 1/2$, so that the Bayes rule given by Theorem 2 depends only on the values $x(t)$ at $t = 0$, $1/4$, $3/8$, $1/2$, $3/4$ and 1 and the Bayes error is 0.1587. Note that in this particular example $t = 0$ is irrelevant in practice since all trajectories start at 0 and $K(\cdot, 0) = 0$. Some typical trajectories are shown in Figure S1. Using $K(s,t) = \min(s,t) = \sum_{m=0}^{\infty} \sum_{k=o}^{\infty} \Phi_{kj} \Phi_{jk}$ it is not difficult to show that $m_1(t)$ has the alternative representation

$$m_1(t) \approx -0.83K(t, 1/4) - 4K(t, 3/8) + 4K(t, 1/2) + 2.83K(t, 3/4) - 2.41K(t, 1).$$

Now, we analyze the performance of RK and RK_B in this example. The left panel of Figure S2 shows the evolution of the classification error as the sample size increases for RK-C (blue line with circles) and RK_B-C (red line with diamonds). The dashed black line indicates the Bayes error. Each output is obtained by averaging 100 independent runs with test samples of size 200; for each sample

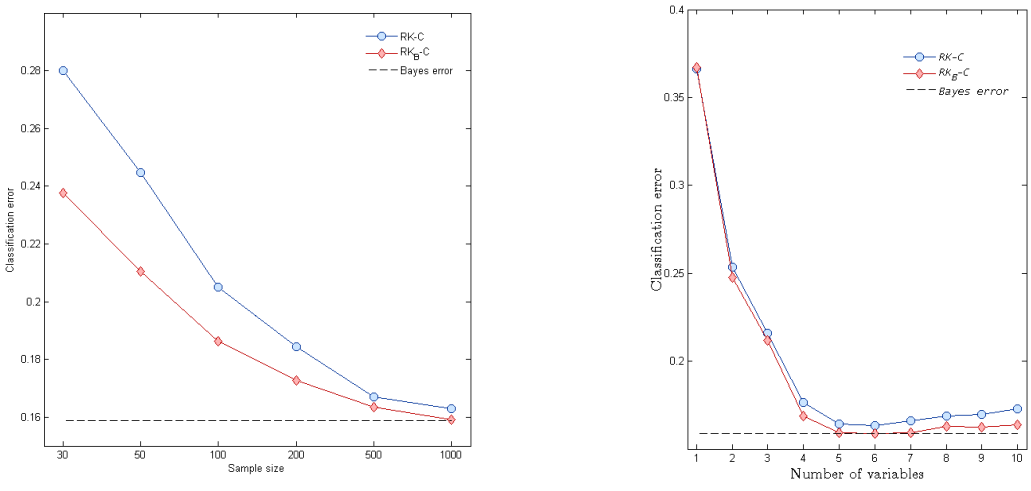

Figure S2: Evolution of the classification error of RK-C and RK$_B$-C in terms of the sample size (left panel) and the number of selected variables (right panel).

size, the number of selected variables is set through a validation sample. The right panel of Figure S2 shows the classification error in terms of the number of variables for RK-C and RK$_B$-C for $n = 500$. Finally, Figure S3 shows the frequency of selection of each variable among the first six (by construction, we know there are just six relevant points) corresponding to 100 independent runs of RK-VS for three different sample sizes. The theoretical relevant points are marked by vertical dashed lines. So, to sum up, whereas Figure S2 summarizes the results in terms of classification performance, Figure S3 is more concerned with capacity of identifying the right relevant variables.

These results are quite positive; RK-C seems to be a good estimator of the optimal classifier as the error rate converges swiftly to the Bayes error even when the number of variables is unknown and fixed by validation. The right panel in Figure S2 shows that for the true number of variables (five-six) the algorithm achieves the best performance. By contrast, a wrong choice of the number of variables can entail an important increase of the misclassification rate, so this is a sensitive issue. In addition, the selected variables (represented in Figure S3) are mostly in coincidence with the theoretical ones. Even for small sample sizes, RK$_B$-VS and RK-VS variables are grouped around the relevant variables. Only the variable $X(0)$ is omitted since it is in fact nearly irrelevant. This good performance in detecting the important variables is in principle better than one might expect for a greedy algorithm (that, therefore, might not provide the true global optimum). Note also that the inclusion of some additional information seems specially beneficial for smaller sample sizes. Finally, it is worth mentioning that the RK-based

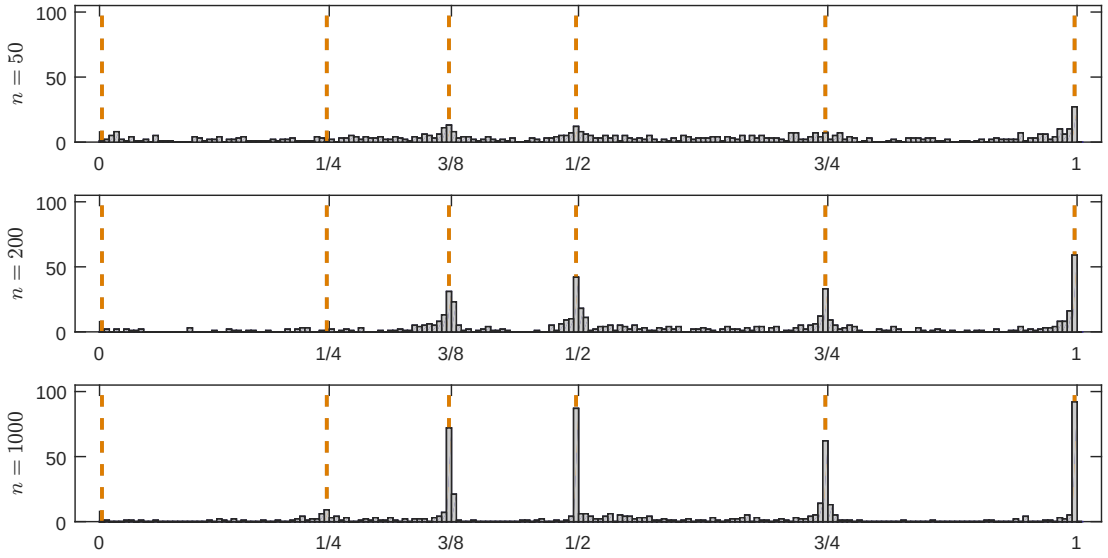

Figure S3: Histograms of the six first selected variables by RK-VS over 100 runs for sample sizes 50 (top panel), 200 (middle panel) and 1000 (bottom panel).

methods seem to be relatively inexpensive from the computational point of view. Our experience suggests that the increase in the computation time as the sample size increases is much slower than that of other competing dimension reduction methods.

S2.3 Simulation study

The simulation experiments include 94 models, previously considered in the studies by Berrendero *et al.* (2016a,b). These models can be grouped into three classes.

(i) *Gaussian models*: they are defined via the marginal Gaussian distributions (Brownian-like, Ornstein Uhlenbeck,...) P_i of $X(t)|Y = i$ for $i = 0, 1$. In all cases $p = \mathbb{P}(Y = 1) = 1/2$.

(ii) *Logistic-type models*: they are defined through the function $\eta(X) = \mathbb{P}(Y = 1|X(t))$ and the marginal of X. It is assumed that $\eta(x) = (1 + e^{-\Psi(x(t_1),\cdots,x(t_d))})^{-1}$, with different choices for the link function Ψ.

(iii) *Finite mixtures* of different types of Gaussian models.

Detailed descriptions of the 94 considered models can be found in Appendix A. We should emphasize that only 7 among these 94 models fulfill all the conditions imposed in our theoretical results. They are grouped under the label RKHS in the extended output tables of Appendix B. The remaining "unorthodox" models aim at checking the behavior of our proposal when some departures from the assumptions are present.

Training samples of sizes $n = 30, 50, 100, 200$ are considered for each model. Sample trajectories are discretized in 100 equispaced points in the interval [0,1]. The criterion of comparison is the classification accuracy for an independent test sample of size 200. The number of selected variables as well as the classification parameters (if needed) are fixed in a validation step, using, for each test sample, another independent validation sample of size 200. The final output is the average classification accuracy over 200 runs of this experiment.

Comparison of variable selection methods

The primary aim of the study is to check the performance of our RK variable selection method against other dimension reduction procedures, chosen among the winners in Berrendero *et al.* (2016a,b). To be specific, these are the methods considered in the experiments:

- RK-VS, as defined in Subsection S2.1.
- RK_B-VS, the "oracle" version RK-VS defined in Subsection S2.2 by assuming that the common covariance structure coincides with that of the Brownian motion. As mentioned above, RK_B is included only for illustration purposes, just to check the price of the estimation in $K(s,t)$ and the (sometimes surprising) resistance against departures from the assumptions on the covariance structure.
- mRMR-RD: this is a modified version of the popular minimum redundancy maximum relevance algorithm (mRMR) for variable selection proposed by Ding and Peng (2005). The aim of mRMR is to select the subset S of variables that maximizes the difference $\text{rel}(S) - \text{red}(S)$, where $\text{rel}(\cdot)$ and $\text{red}(\cdot)$ are appropriate measures of relevance and redundancy which are defined in terms of an association measure between random variables. The improved version of mRMR considered here (denoted mRMR-RD) has been recently proposed in Berrendero *et al.* (2016a). It relies on the use of the increasingly popular *distance correlation* (Székely *et al.*, 2007) association measure to define relevance and redundancy in the mRMR algorithm.
- MHR: the maxima hunting method (Berrendero *et al.*, 2016b) also uses the distance correlation $R^2(t) = \mathcal{R}^2(X(t), Y)$, between $X(t)$ and the binary response Y to select the points $t_1, \ldots, t_k$ corresponding to the local maxima of $R^2(t)$. This automatically takes into account the relevance-redundancy trade-off (though in a qualitative way, quite different to that of the mRMR methodology).
- PLS: partial least squares, a well-known dimension reduction technique; see e.g. Delaigle and Hall (2012c) and references therein.

Table S1: Percentage of correct classification with the three considered classifiers

Classifier	Sample size	Dimension reduction methods				
		mRMR-RD	PLS	MHR	RK-VS	RK_B-VS
LDA	$n = 30$	81.04	**82.87**	82.44	81.50	80.89
	$n = 50$	82.37	**83.78**	83.68	83.44	82.54
	$n = 100$	83.79	84.70	84.97	**85.30**	84.46
	$n = 200$	84.88	85.46	85.90	**86.51**	85.90
kNN	$n = 30$	81.88	82.45	**82.46**	82.28	81.92
	$n = 50$	82.95	83.49	83.43	**83.75**	83.25
	$n = 100$	84.31	84.77	84.73	**85.59**	84.95
	$n = 200$	85.38	85.79	85.91	**87.16**	86.50
SVM	$n = 30$	83.22	84.12	**84.62**	84.28	84.12
	$n = 50$	84.21	85.04	85.44	**85.60**	85.20
	$n = 100$	85.27	86.03	86.29	**86.96**	86.48
	$n = 200$	86.10	86.79	86.86	**87.90**	87.50

All these methods for variable selection (or, in the case of PLS, for projection-based dimension reduction) are data-driven, i.e., independent on the classifier, so we can combine them with different classifiers. For illustrative purposes we show the results we have obtained with the Fisher linear classifier (LDA), k nearest neighbors (kNN) and support vector machine with a linear kernel (SVM).

Some aggregated results are in Table S1. Variable selection methods and PLS are in columns and each row corresponds to a sample size and a classifier. Each output is the average classification accuracy of the 94 models over 200 runs. Boxed outputs denote the best result for each sample size and classifier. For readability, additional, more detailed, summary tables are included in Appendix B. The full results of the 1128 experiments (94 models $\times$ 4 samples sizes $\times$ 3 classifiers) are available in the supplementary file *outputs*.

The results are quite similar for all considered classifiers: RK-VS methodology outperforms the other competitors on average with a better performance for bigger sample sizes. Although RK-VS could have more difficulties to estimate the covariance matrix for small sample sizes, it is very close to MHR, which seems to be the winner in that case. Besides, the number of variables selected by RK-VS (not reported here for the sake of clarity; see Table S4) is comparable to that of mRMR-RD and MHR for kNN and SVM but it is about half of the number selected by mRMR-RD and MHR for LDA (the number of PLS components is often smaller but they lack interpretability). Note that, according with the available ex-

Table S2: Average classification accuracy (%) over all considered models.

n	kNN	SVM	RK-C	RK_B-C	LDA-Oracle
30	79.61	83.86	81.50	80.89	84.97
50	80.96	85.01	83.44	82.54	86.23
100	82.60	86.20	85.30	84.46	87.18
200	83.99	87.07	86.51	85.90	87.69

perimental evidence (Berrendero *et al.*, 2016b,a), the competing selected methods (mRMR-RD, MHR and PLS) have themselves a good general performance. So, these outputs are remarkable and encouraging especially taking into account that only 7 out of 94 models under study fulfill all the regularity conditions required for the best performance of RK-VS. Note that, somewhat surprisingly, the failure of the "Brownian assumption" implicit in the RK_B-VS method does not entail a big loss of accuracy with respect to the "non-parametric" RK-VS version.

Comparison of classifiers

We also assess the performance of the classifiers RK-C and RK_B-C; see the definitions in Subsections 5.1 and S2.2, respectively. The competitors are kNN and SVM (with linear kernel), two standard all-purpose classification methods.

Table S2 provides again average percentages of correct classification over 200 runs of the previously considered 94 functional models. The results are grouped by sample size (in rows). Classification methods are in columns. The full detailed outputs are given in the supplementary file *outputs*.

The difference with Table S1 is that, in this case, the classifiers kNN and SVM are used with no previous variable selection. So, the original whole functional data are used. This is why we have replaced the standard linear classifier LDA (which cannot be used in high-dimensional or functional settings) with the LDA-Oracle method which is just the Fisher linear classifier based on the "true" relevant variables (which are known beforehand since we consider models for which the Bayes rule depends only on a finite set of variables). Of course this classifier is not feasible in practice; it is included here only for comparison purposes.

As before, RK-C results are better for higher sample sizes and the distances between SVM or LDA-Oracle and RK-C are swiftly shortened with n; and again, RK_B-C is less accurate than RK-C but not too much. While the global winner is SVM, the slight loss of accuracy associated with the use of RK-C and RK_B-C can be seen as a reasonable price for the simplicity and ease of interpretability of these methods. Note also that the associated procedure of variable selection can be seen as a plus of RK-C. In fact, the combination of RK-VS with SVM outperforms

Table S3: Average classification accuracy (%) for the models satisfying the assumptions of Th. S1

n	kNN	SVM	RK-C	RK_B-C	LDA-Oracle
30	83.20	87.29	88.30	89.95	90.91
50	84.90	88.81	89.81	90.69	91.41
100	86.61	89.88	90.81	91.18	91.64
200	87.94	90.48	91.13	91.30	91.71

SVM based on the entire curves (see Table S1).

Table S3 shows average percentages of correct classification over 200 runs of the subset of models among all seven models that satisfy the assumptions in Theorem S1, which establishes the consistency of the procedure proposed in Section 5. It is not surprising that for these models RK-C and RK_B-C have a better performance than kNN and SVM. In fact the RK percentages of correct classification are very close to those of LDA-Oracle, which means that there is not much room for improvement under these assumptions

S3 Computational details

All considered methodologies have been implemented in MATLAB. The code is available upon request. Some details:

- We have followed the implementation of the the minimum Redundancy Maximum Relevance algorithm given in Berrendero *et al.* (2016a). This version allows us to introduce different association measures.
- We have implemented the original iterative PLS algorithm that can be found, e.g. in Delaigle and Hall (2012c).
- Maxima-hunting and the distance correlation measure have been computed as described in Berrendero *et al.* (2016b).
- Our k-NN implementation is built around the MATLAB function *pdist2* and allows for the use of different distances; we have employed the usual Euclidean distance. Also, the computation for different numbers of neighbors can be simultaneously made with no additional cost.
- Our LDA is a faster implementation of the MATLAB function *classify*.
- The linear SVM has been performed with the MATLAB version of the LIBLINEAR library (see Fan *et al.* (2008)) using the parameters *bias* and *solver*

type 2. It obtains (with our data) very similar results to those of the default *solver type* 1, but faster. LIBLINEAR is much faster than the more popular LIBSVM library when using linear kernels.

- The cost parameter C of the linear SVM classifier, the number k of nearest neighbors in the k-NN rule, the smoothing parameter h in MHR and the number of selected variables are chosen by standard validation procedures.

- The DHB algorithm has been implemented according to the instructions given in Delaigle, Hall, and Bathia (2012b), including leave-one-out cross-validation and computational savings. We have also used the same parameters and the first stopping criterion proposed by these authors.

References

Ash, R.B. and Gardner, M.F. (1975) *Topics in Stochastic Processes*. Academic Press, New York.

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

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

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

Cucker, F. and Zhou, D.X. (2007). *Learning theory: an approximation theory viewpoint*. Cambridge University Press.

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

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

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

Delaigle, A. and P. Hall (2012c) Methodology and theory for partial least squares applied to functional data. *The Annals of Statistics*. **40**, 322–352.

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 Fan, Y. (2008). High-Dimensional Classification Using Features Annealed Independence Rules. *Annals of Statistics*, 2605–2637.

Fan R-E, Chang K-W, Hsieh C-J,Wang X-R, Lin C-J. (2008). LIBLINEAR: a library for large linear classification. *Journal of Machine Learning Research*, **9** 1871–1874.

Laha, R.G. and Rohatgi, V.K. (1979). *Probability Theory*. Wiley.

Mardia, K. V., Kent, J. T., and Bibby, J. M. (1980). *Multivariate analysis*. Academic Press.

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

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

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

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

APPENDIX A. Models used in the simulation study

The general structure is similar to that of the simulation studies in Berrendero *et al.* (2016b) and Berrendero *et al.* (2016a) which are devoted to the assessment of variable selection methods in the functional classification setting. Here we consider the 94 models for which the mean functions m_0 and m_1 are different. The optimal classification rule in each case depends only on a finite number of variables. Models differ in complexity and number of relevant variables. They are defined giving either:

(E1) A pair of distributions for $X|Y = 0$ and $X|Y = 1$ (corresponding to P_0 and P_1, respectively) as well as the prior probability $p = \mathbb{P}(Y = 1)$; in all cases, we take $p = \mathbb{P}(Y = 1) = 1/2$.

(E2) The marginal distribution of X plus the conditional distribution $\eta(x) = \mathbb{P}(Y = 1|X = x)$.

All the 94 considered models belong to one of the following classes:

Gaussian models: they are denoted by G. Gaussian models are generated according to the general pattern (E1). In all cases the distributions of $X(t)|Y = i$ are chosen among one of the Gaussian distributions described below.

Logistic models: they are defined through the general pattern (E2). The process $X = X(t)$ follows one of the above mentioned distributions and $Y \sim \text{Binom}(1, \eta(X))$ with

$$\eta(x) = \frac{1}{1 + e^{-\Psi(x(t_1),\cdots,x(t_d))}},$$

a function of the relevant variables $x(t_1), \cdots, x(t_d)$. The 15 versions and the few variants of this model considered are identified with the general label L. They correspond to different choices for the link function Ψ (both linear and nonlinear) and for the distribution of X.

Mixtures: they are obtained by combining (via mixtures) the above mentioned Gaussian distributions assumed for $X|Y = 0$ and $X|Y = 1$ in several ways. These models are denoted by M in the output tables.

The processes involved are chosen among the following: first, the **standard Brownian Motion**, B. Second, BT denotes a **Brownian Motion 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 **Ornstein–Uhlenbeck process**, with zero mean (OU) or different mean functions $m(t)$ (OUt). Finally some "smooth" processes have been also included. They are obtained by convolving Brownian trajectories with Gaussian kernels. We have considered two levels of smoothing denoted by sB and ssB; in the list of models below those labeled ssB are smoother than those with label sB.

In the following list of models, P_i denotes the distribution of $X|Y = i$ and *variables* is the set of relevant variables in each Gaussian or Mixture case. We call them "relevant" in the sense that the optimal classification rule depends only on these variables. In the list below the variables written in boldface are "especially relevant" in terms of their relative discriminating capacity.

All considered sample data are discretized in 100 equispaced points $X_1, \ldots, X_100$ in the interval [0,1]. To avoid degeneracies we have excluded the point 0 and the point 1 in the Brownian Bridge type models.

1. Gaussian models considered:

1. **G2** : $\begin{cases} P_0 : & B(t) + t \\ P_1 : & B(t) \end{cases}$

   $variables = \{X_{100}\}$.

2. **G2b** : $\begin{cases} P_0 : & B(t) + 3t \\ P_1 : & B(t) \end{cases}$

   $variables = \{X_{100}\}$.

3. **G4** : $\begin{cases} P_0 : & B(t) + hillside_{0.5,4}(t) \\ P_1 : & B(t) \end{cases}$

   $variables = \{X_{47}, \boldsymbol{X}_{100}\}$.

4. **G5** : $\begin{cases} P_0 : & B(t) + 3\Phi_{1,1}(t) \\ P_1 : & B(t) \end{cases}$

   $variables = \{X_1, \boldsymbol{X}_{48}, X_{100}\}$.

5. **G6** : $\begin{cases} P_0 : & B(t) + 5\Phi_{2,2}(t) \\ P_1 : & B(t) \end{cases}$

   $variables = \{X_{48}, \boldsymbol{X}_{75}, X_{100}\}$.

6. **G7** : $\begin{cases} P_0 : & B(t) + 5\Phi_{3,2}(t) + 5\Phi_{3,4}(t) \\ P_1 : & B(t) \end{cases}$

   $variables = \{X_{22}, \boldsymbol{X}_{35}, X_{49}, X_{74}, \boldsymbol{X}_{88}, X_{100}\}$.

7. **G8** : $\begin{cases} P_0 : & B(t) + 3\Phi_{2,1.25}(t) + 3\Phi_{2,2}(t) \\ P_1 : & B(t) \end{cases}$

   $variables = \{X_9, \boldsymbol{X}_{35}, X_{48}, X_{62}, \boldsymbol{X}_{75}, X_{100}\}$.

2. LOGISTIC-TYPE MODELS UNDER STUDY: they are all defined according method (E2) (see Sec. 6.1 in the main paper). The process $X = X(t)$ follows one of the distributions mentioned above and $Y = \text{Binom}(1, \eta(X))$ with $\eta(x) = (1 + e^{-\psi(x(t_1),\cdots,x(t_k))})^{-1}$, a function of the relevant variables $x(t_1), \cdots, x(t_k)$.

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

Some variations of these models have been also considered:

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-TYPE MODELS: 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, ..., M10 in the output tables.

1. **M2** : $\begin{cases} P_0: & \begin{cases} B(t) + 3\Phi_{2,2}(t), & 1/2 \\ B(t) + 5\Phi_{3,2}(t), & 1/2 \end{cases} \\ P_1: & B(t) \end{cases}$

   $variables = \{X_{22}, \boldsymbol{X}_{35}, X_{48}, \boldsymbol{X}_{75}, X_{100}\}$.

2. **M3** : $\begin{cases} P_0: & \begin{cases} B(t) + 3\Phi_{2,2}(t), & 1/10 \\ B(t) + 5\Phi_{3,2}(t), & 9/10 \end{cases} \\ P_1: & B(t) \end{cases}$

   $variables = \{X_{22}, \boldsymbol{X}_{35}, X_{48}, \boldsymbol{X}_{75}, X_{100}\}$.

3. **M4**: $\begin{cases} P_0: & \begin{cases} B(t) + 3\Phi_{2,2}(t), & 1/2 \\ B(t) + 5\Phi_{3,3}(t), & 1/2 \end{cases} \\ P_1: & B(t) \end{cases}$

   $variables = \{X_{48}, \boldsymbol{X}_{62}, \boldsymbol{X}_{75}, X_{100}\}$.

4. **M5** : $\begin{cases} P_0: & \begin{cases} B(t) + 3\Phi_{2,1}(t) & ,1/3 \\ B(t) + 3\Phi_{2,2}(t), & 1/3 \\ B(t) + 5\Phi_{3,2}(t), & 1/3 \end{cases} \\ P_1: & B(t) \end{cases}$

   $variables = \{X_1, \boldsymbol{X}_{22}, \boldsymbol{X}_{35}, X_{48}, \boldsymbol{X}_{75}, X_{100}\}$.

5. **M6** : $\begin{cases} P_0: & \begin{cases} B(t) + 3\Phi_{2,1}(t) & ,1/2 \\ B(t) + 3t & ,1/2 \end{cases} \\ P_1: & B(t) \end{cases}$

   $variables = \{X_1, \boldsymbol{X}_{22}, X_{49}, \boldsymbol{X}_{100}\}$.

6. **M7** : $\begin{cases} P_0: & \begin{cases} B(t) + 3\Phi_{1,1}(t) & ,1/2 \\ BB(t) & ,1/2 \end{cases} \\ P_1: & B(t) \end{cases}$

   $variables = \{X_1, \boldsymbol{X}_{48}, \boldsymbol{X}_{100}\}$.

7. **M8** : $\begin{cases} P_0: & \begin{cases} B(t) + \theta t, \ \ \theta \sim N(0,5) & ,1/2 \\ B(t) + hillside_{0.5,5}(t) & ,1/2 \end{cases} \\ P_1: & B(t) \end{cases}$

   $variables = \{X_{47}, \boldsymbol{X}_{100}\}$.

8. **M10** : $\begin{cases} P_0: & \begin{cases} B(t) + 3\Phi_{1,1}(t) & ,1/3 \\ B(t) - 3t & ,1/3 \\ BB(t) & ,1/3 \end{cases} \\ P_1: & B(t) \end{cases}$

   $variables = \{X_1, \boldsymbol{X}_{48}, \boldsymbol{X}_{100}\}$.

Finally, we consider here those models for which the mean functions m_0 and m_1 are different (otherwise any linear method is blind to discriminate between P_0 and P_1). The full list of models involved is as follows:

1. L1 OU	6. L2 OU	11. L3 OU	16. L3b B
2. L1 OUt	7. L2 OUt	12. L3b OU	17. L3 sB
3. L1 B	8. L2 B	13. L3 OUt	18. L3 ssB
4. L1 sB	9. L2 sB	14. L3b OUt	19. L4 OU
5. L1 ssB	10. L2 ssB	15. L3 B	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. G2
81. G2b
82. G4
83. G5
84. G6
85. G7
86. G8
87. M2
88. M3
89. M4
90. M5
91. M6
92. M7
93. M8
94. M10

APPENDIX B. Some additional results

- Table S4 is a complement for Table S1 by showing the average number of variables (or components).
- Tables S5, S6 and S7 show the classification accuracy (percentage of correct classification) for different groups of models and methods obtained with LDA, kNN and SVM classifiers respectively. Results from the different considered classifiers are quite similar in relative terms. Let us recall that the full results of the 1128 experiments (94 models $\times$4 samples sizes $\times$3 classifiers) are available in the supplementary file *outputs*. The methods appear in columns; apart from methods in Table 1 we have included *Base* (except for LDA) and *Oracle* versions of each

Table S4: Average number of selected variables (or components) with the three considered classifiers. Remember that the original dimension is 100.

Classifier	Sample size	Dimension reduction methods				
		mRMR-RD	PLS	MHR	RK-VS	RK_B-VS
LDA	$n = 30$	4.9	2.6	5.4	2.7	3.7
	$n = 50$	5.9	2.8	6.1	2.8	4.1
	$n = 100$	7.2	3.3	7.0	3.2	4.8
	$n = 200$	8.1	4.0	7.5	3.9	5.6
kNN	$n = 30$	7.8	4.3	6.2	7.6	8.1
	$n = 50$	8.0	4.8	6.2	7.3	7.9
	$n = 100$	8.4	5.5	6.2	6.7	7.6
	$n = 200$	8.6	6.2	5.9	6.3	7.2
SVM	$n = 30$	9.3	3.3	8.0	9.3	10.0
	$n = 50$	9.4	3.8	7.9	8.7	9.6
	$n = 100$	9.7	4.6	7.9	8.0	9.2
	$n = 200$	9.8	5.6	7.5	7.6	8.9

method. The first is based on the entire trajectories and *Oracle* only uses the true relevant variables. The simulation outputs are grouped in different categories (in rows) by model type and sample size n. The rows are labelled by the general model type, that is, logistic, Gaussian and mixtures. The logistic models are also divided by the type of processes involved according to the notation given above. RKHS denotes the models that fulfil the hypotheses of RK-VS (G2, G2b, G4,...,G8) and "All models" includes the outputs of all the 94 considered models for each n. We have followed the methodology described in the main paper and the outputs are averaged over 200 independent runs. The marked values correspond to the best performance in each row (excluding *Oracle* which is not feasible in practice).

Table S5: Percentage of correct classification with LDA

Models	n	mRMR-RD	PLS	MHR	RK-VS	RK_B-VS	LDA-Oracle
All models	30	81.04	82.87	82.44	81.50	80.89	84.97
	50	82.37	83.78	83.68	83.44	82.54	86.23
	100	83.79	84.70	84.97	85.30	84.46	87.18
	200	84.88	85.46	85.90	86.51	85.90	87.69
Logistic OU	30	78.70	80.11	79.36	78.21	76.47	81.92
	50	80.12	80.96	80.75	80.23	78.33	83.24
	100	81.70	81.90	82.30	82.16	80.69	84.27
	200	83.05	82.74	83.65	83.66	82.61	84.84
Logistic OUt	30	80.12	81.30	80.87	79.60	78.56	83.11
	50	81.21	82.05	81.98	81.42	80.20	84.44
	100	82.39	82.91	83.14	83.14	82.15	85.45
	200	83.35	83.51	84.03	84.29	83.66	85.93
Logistic B	30	82.79	84.57	84.19	83.52	82.32	87.54
	50	84.18	85.55	85.59	85.65	84.21	88.83
	100	85.74	86.60	87.16	87.71	86.47	89.90
	200	86.88	87.50	88.33	89.17	88.18	90.51
Logistic sB	30	82.95	84.63	84.26	83.43	82.37	87.10
	50	84.18	85.59	85.59	85.39	84.11	88.46
	100	85.51	86.60	87.02	87.52	86.34	89.55
	200	86.71	87.38	88.20	88.84	87.98	90.18
Logistic ssB	30	84.56	85.73	85.58	84.93	84.51	86.54
	50	85.65	86.49	86.54	86.42	85.93	87.90
	100	86.86	87.25	87.38	87.89	87.39	88.81
	200	87.83	88.01	87.72	88.83	88.59	89.38
Gaussian	30	85.28	88.63	88.70	88.30	89.95	90.91
	50	86.72	89.45	89.38	89.81	90.69	91.41
	100	88.21	89.91	89.86	90.81	91.18	91.64
	200	89.00	90.38	89.96	91.13	91.30	91.71
Mixture	30	71.95	76.19	75.40	73.93	76.65	79.09
	50	73.88	77.66	77.03	76.63	78.30	80.29
	100	75.54	78.91	78.61	79.13	79.89	81.07
	200	76.46	79.66	79.29	80.21	80.61	81.39
RKHS	30	85.28	88.63	88.70	88.30	89.95	90.91
	50	86.72	89.45	89.38	89.81	90.69	91.41
	100	88.21	89.91	89.86	90.81	91.18	91.64
	200	89.00	90.38	89.96	91.13	91.30	91.71

Table S6: Percentage of correct classification with kNN

Models	n	mRMR-RD	PLS	MHR	RK-VS	RK_B-VS	Base	kNN-Oracle
All models	30	81.88	82.45	82.46	82.28	81.92	79.61	84.56
	50	82.95	83.49	83.43	83.75	83.25	80.96	86.16
	100	84.31	84.77	84.73	85.59	84.95	82.60	87.94
	200	85.38	85.79	85.91	87.16	86.50	83.99	89.25
Logistic OU	30	78.71	79.22	79.20	78.58	77.82	75.63	81.15
	50	79.64	80.04	80.02	79.98	79.05	76.87	82.63
	100	80.96	81.13	81.26	81.66	80.68	78.44	84.30
	200	82.10	82.07	82.56	83.21	82.23	79.73	85.49
Logistic OUt	30	81.87	82.71	82.30	81.91	81.37	79.50	84.46
	50	82.83	83.52	83.18	83.13	82.49	80.62	85.89
	100	84.12	84.52	84.33	84.90	84.03	82.02	87.35
	200	85.00	85.31	85.30	86.23	85.31	83.14	88.49
Logistic B	30	83.29	84.01	83.94	83.94	83.04	81.10	86.61
	50	84.38	85.08	84.90	85.47	84.55	82.35	88.24
	100	85.68	86.30	86.31	87.40	86.41	83.92	90.19
	200	86.78	87.39	87.63	89.27	88.25	85.35	91.66
Logistic sB	30	84.00	84.48	84.55	84.40	83.66	81.90	86.59
	50	84.87	85.36	85.31	85.65	84.93	83.02	88.24
	100	86.09	86.61	86.62	87.51	86.62	84.44	90.11
	200	87.07	87.58	87.84	89.17	88.35	85.73	91.59
Logistic ssB	30	85.92	85.97	86.35	86.39	86.09	84.47	88.01
	50	86.86	86.78	87.11	87.49	87.10	85.41	89.44
	100	87.93	87.86	88.05	88.89	88.55	86.71	91.04
	200	88.89	88.81	88.75	90.24	89.88	87.91	92.34
Gaussian	30	83.96	85.35	85.79	86.16	87.13	83.20	87.46
	50	84.80	86.61	86.68	87.62	88.20	84.99	88.55
	100	85.69	87.85	87.58	88.91	89.19	86.61	89.56
	200	86.30	88.74	88.19	89.68	89.84	87.94	90.11
Mixture	30	74.20	74.40	74.40	74.42	75.92	71.05	76.83
	50	76.59	76.92	76.70	77.45	78.43	73.92	79.58
	100	79.46	79.68	79.20	80.76	81.36	77.32	82.70
	200	81.48	81.51	81.42	83.21	83.61	79.98	84.74
RKHS	30	83.96	85.35	85.79	86.16	87.13	83.20	87.46
	50	84.80	86.61	86.68	87.62	88.20	84.99	88.55
	100	85.69	87.85	87.58	88.91	89.19	86.61	89.56
	200	86.30	88.74	88.19	89.68	89.84	87.94	90.11

Table S7: Percentage of correct classification with SVM

Models	n	mRMR-RD	PLS	MHR	RK-VS	RK_B-VS	Base	SVM-Oracle
All models	30	83.22	84.12	84.62	84.28	84.12	83.86	87.53
	50	84.21	85.04	85.44	85.60	85.20	85.01	88.21
	100	85.27	86.03	86.29	86.96	86.48	86.20	88.75
	200	86.10	86.79	86.86	87.90	87.50	87.07	89.03
Logistic OU	30	79.98	80.79	80.81	80.19	79.65	80.18	83.93
	50	81.13	81.64	81.69	81.66	80.95	81.36	84.62
	100	82.39	82.51	82.59	83.15	82.44	82.50	85.17
	200	83.51	83.30	83.50	84.32	83.74	83.42	85.49
Logistic OUt	30	83.38	83.84	84.33	83.70	83.28	83.77	87.24
	50	84.37	84.69	85.14	85.00	84.39	84.82	87.88
	100	85.43	85.67	86.07	86.34	85.75	85.94	88.37
	200	86.15	86.34	86.71	87.26	86.74	86.71	88.64
Logistic B	30	85.24	85.81	87.01	86.56	85.97	86.01	90.58
	50	86.23	86.83	87.92	88.11	87.20	87.17	91.23
	100	87.35	87.92	88.99	89.58	88.69	88.50	91.80
	200	88.16	88.85	89.85	90.71	89.95	89.50	92.09
Logistic sB	30	85.55	85.98	87.06	86.68	86.22	86.22	90.22
	50	86.33	86.96	87.92	87.86	87.32	87.32	90.96
	100	87.13	88.01	88.88	89.41	88.69	88.51	91.53
	200	88.04	88.84	89.55	90.41	89.80	89.40	91.81
Logistic ssB	30	87.16	87.31	87.69	88.26	88.25	87.65	90.08
	50	87.93	88.02	88.28	89.07	88.90	88.47	90.57
	100	88.82	88.96	88.55	89.91	89.77	89.37	91.00
	200	89.47	89.73	88.54	90.60	90.57	90.16	91.25
Gaussian	30	86.42	88.72	88.97	89.00	89.99	87.29	90.54
	50	87.33	89.44	89.27	89.94	90.49	88.81	91.02
	100	88.48	90.03	89.60	90.63	90.93	89.88	91.38
	200	88.98	90.41	89.51	91.03	91.21	90.48	91.45
Mixture	30	73.01	76.52	76.12	75.53	76.93	74.88	78.71
	50	74.39	77.90	77.42	77.50	78.35	76.51	79.89
	100	75.55	79.27	78.65	79.41	79.72	78.20	80.76
	200	76.35	80.10	79.06	80.26	80.50	79.21	81.16
RKHS	30	86.42	88.72	88.97	89.00	89.99	87.29	90.54
	50	87.33	89.44	89.27	89.94	90.49	88.81	91.02
	100	88.48	90.03	89.60	90.63	90.93	89.88	91.38
	200	88.98	90.41	89.51	91.03	91.21	90.48	91.45