Achieving near-perfect classification for functional data

Aurore Delaigle and Peter Hall

Department of Mathematics and Statistics, University of Melbourne, Australia.

Abstract: We show that, in functional data classification problems, perfect asymptotic classification is often possible, making use of the intrinsic very high dimensional nature of functional data. This performance is often achieved by linear methods, which are optimal in important cases. These results point to a marked contrast between classification for functional data and its counterpart in conventional multivariate analysis, where dimension is kept fixed as sample size diverges. In the latter setting, linear methods can sometimes be quite inefficient, and there are no prospects for asymptotically perfect classification, except in pathological cases where, for example, a variance vanishes. By way of contrast, in finite samples of functional data, good performance can be achieved by truncated versions of linear methods. Truncation can be implemented by partial least-squares or projection onto a finite number of principal components, using, in both cases, cross-validation to determine the truncation point. We establish consistency of the cross-validation procedure.

Keywords: Bayes classifier, centroid-method classifier, cross-validation, dimension reduction, error rate, Gaussian process, linear model, projection.

1 Introduction

We present a new viewpoint for interpreting and solving classification problems involving functional data, and we argue that those problems have unusual, and fascinating, properties that set them apart from their finite-dimensional counterparts. In particular we show that, in many quite standard settings, the performance of simple classifiers constructed from training samples becomes perfect as the sizes of those samples diverge. That is, the classifiers can be constructed so that the probability of correctly classifying a new data function converges to 1 as the training sample sizes increase. That property never holds for finite dimensional data, except in pathological cases where, for example, one or more of the variances vanish. In important cases

we demonstrate that, unlike their low-dimensional counterparts, basic linear methods (e.g. the centroid-based classifier and linear discriminant analysis) are optimal in a range of functional data problems. We discuss the relationship between our findings and the very good practical performance of dimension reduced methods based on partial least-squares or on carefully chosen principal component projections.

The perfect asymptotic classification property holds in a variety of settings, including those where the standard deviations of high-order principal component scores are of the same order as, or smaller than, the sizes of the corresponding mean components. That context is far from pathological; for example, it holds for mathematical models that are routinely used in simulation studies for functional data. The theoretical foundation for these findings is an intriguing dichotomy of properties, and is as interesting as the findings themselves. It can be summarised as follows.

Consider the simple, common centroid-based classifier and apply it, in the context of infinite training samples, to classify a new data function X into one of two groups. We do this by projecting X onto the real line as the point $\int_{\mathcal{I}} \psi X$, where $\mathcal{I}$ denotes the support of X and the function ψ determines the projection. We choose ψ to minimise classification error in the resulting one-dimensional classification problem. If the ψ that minimises error is well defined then it is not possible to achieve perfect classification, although in Gaussian cases the univariate classifier has minimum classification error among all possible approaches. On the other hand, if the problem of choosing the best ψ does not have a well defined solution, it is because much of the important information for classification lies arbitrarily far out in the tails of the principal component expansion of X. This might seem to be a theoretical blind alley, leading nowhere of practical value. However, it can actually be exploited in practice by choosing a projection ψ which, for as many finite-order terms as empirically feasible, accurately captures properties of the infinite expansion.

Reflecting this dichotomy, in cases where perfect classification is impossible it is often feasible to capture virtually all of the classification information in the data using relatively low-order terms in the principal component expansion. In the context of perfect classification, however, we must venture further into the expansion in order to find the information that is necessary for good classification decisions. This issue calls to mind a remark made by Cox (1968), in the context of prediction rather than classification. Arguing that a conventional principal component approach might give

poor results, Cox noted that "A difficulty seems to be that there is no logical reason why the dependent variable should not be closely tied to the least important principal component." (See Cook, 2007, for discussion.) This difficulty arises in the second part of the dichotomy referred to above, but it can be addressed successfully and results in practical, adaptive classifiers with asymptotically perfect classification performance, as training sample sizes diverge.

There is a sizeable literature on methods for classifying functional data. It includes methodology suggested by James and Hastie (2001), Preda et al. (2007) and Shin (2008), who suggested linear discriminant analysis; Ferraty and Vieu (2003), who constructed classifiers based on kernel estimators of prior probabilities; James and Sugar (2003), who introduced clustering methods for functional data; Hall et al. (2001), Glendinning and Herbert (2003), Huang and Zheng (2006) and Song et al. (2008), who used classifiers based on principal component analysis; Vilar and Pertega (2004), who developed a classifier based on a distance measure; Biau et al. (2005) and Fromont and Tuleau (2006), who proposed methodology founded on a functional data version of the nearest neighbour classification rule; Leng and Müller (2006) and Chamroukhi et al. (2010), who introduced model-based classifiers; López-Pintado and Romo (2006) and Cuevas et al. (2007), who suggested classifiers based on the notion of data depth; Rossi and Villa (2006), who developed methodology founded on the support vector machine; Wang et al. (2007) and Berlinet et al. (2008), who employed wavelet methods; Epifanio (2008), who developed classifiers based on shape descriptors; Tian and James (2010), who suggested a classifier founded on projections; and Araki (2009), who proposed a Bayesian approach. See also Ramsay and Silverman's (2005) general introduction to functional data analysis.

2 Model, theory and methodology

2.1 Model

Suppose we observe independent and identically distributed data pairs (X_i, I_i), where X_i is a random function defined on a compact interval $\mathcal{I}$, and I_i is a class label, or indicator function, taking only the values 0 or 1. In effect the population from which we sample is a mixture of sub-populations Π_0 and Π_1, say, corresponding to $I_i = 0$

and 1, respectively. To reflect that fact we write X_{ki}, where $1 \le i \le n_k$, for the ith function among $X_1, \ldots, X_n$ for which the corresponding class label equals k, with $k = 0$ or 1 and $n_0 + n_1 = n$. Finally, let π be the probability that a new data curve X comes from sub-population Π_0.

Throughout section 2 we study properties of classifiers in cases where the functions drawn from populations Π_0 and Π_1 have uniformly bounded covariance, and differ only in location. (More general cases are discussed in section 4.2.) In particular, no matter whether X is drawn from Π_0 or Π_1 its covariance is

$$\mathrm{cov}\{X(u), X(v)\} = K(u,v) = \sum_{j=1}^{\infty} \theta_j \, \phi_j(u) \, \phi_j(v) \,, \tag{2.1}$$

say, where $\theta_1 \ge \theta_2 \ge \ldots$. The far right-hand side of (2.1) represents the standard principal component expansion of the covariance in terms of nonzero eigenvalues θ_j, and the respective orthonormal eigenfunctions ϕ_j, of the linear transformation K defined by $K(\psi)(v) = \int_{\mathcal{I}} \psi(v) \, K(u,v) \, du$. This dual use of the notation K is conventional, and serves to connect the transformation K to its "kernel," the function K.

Let E_k denote the expectation operator, which we shall apply to a general functional of X under the constraint that X comes from Π_k. We assume that:

$$\begin{aligned} &E_0(X) = 0, \; E_1(X) = \mu, \text{ and the covariance } K \text{ is strictly positive definite} \\ &\text{and uniformly bounded.} \end{aligned} \tag{2.2}$$

Strict positive definiteness of K is equivalent to asserting that each θ_j in (2.1) is strictly positive. Uniform boundedness implies that $\sum_j \theta_j < \infty$. We write

$$\mu = \sum_{j=1}^{\infty} \mu_j \, \phi_j \tag{2.3}$$

for the generalised Fourier decomposition of μ with respect to the basis $\phi_1, \phi_2, \ldots$. Note that our assumption that only one of the two means, rather than both, is nonzero does not materially influence the results in the following section. Indeed, looking at the case of two different means is operationally equivalent to taking one to be zero and the other to equal the difference of means.

2.2 Centroid classifier

For $k = 0$ and 1, let $\bar{X}_k(t) = n_k^{-1} \sum_{j=1}^{n_k} X_{kj}(t)$, and let X be a new data function that we wish to assign to one of the two populations. The centroid classifier assigns X to

Π_0 or Π_1 according as the statistic $T(X) = D^2(X, \bar{X}_1) - D^2(X, \bar{X}_0)$, is positive or negative, respectively, where D denotes a distance measure between two functions. When $\pi = P(X \in \Pi_0)$ is known, $T(X)$ is often replaced by $T(X) + c$, where c is an adjustment for unequal prior probabilities; see Remark 1 below. In this paper we assume that π is unknown and take $c = 0$.

In Theorems 2.1 and 2.2 below we show that, when the distance is given by $D(X, \bar{X}_k) = \big|\langle X, \psi\rangle - \langle \bar{X}_k, \psi\rangle\big|$, where $\langle X, \psi\rangle = \int_{\mathcal{I}} X\psi$ and ψ is a function defined on $\mathcal{I}$, the centroid classifier enjoys optimality properties. Note that this approach amounts to projecting the data onto a space of dimension 1 determined by a function ψ defined on $\mathcal{I}$. Equipped with this particular distance, if $E_1(X) = \mu$ and $E_0(X) = 0$, then, as n_0 and n_1 diverge,

$$T(X) = \big(\langle X, \psi\rangle - \langle \bar{X}_1, \psi\rangle\big)^2 - \big(\langle X, \psi\rangle - \langle \bar{X}_0, \psi\rangle\big)^2 \tag{2.4}$$

converges to

$$T^0(X) = \big(\langle X, \psi\rangle - \langle \mu, \psi\rangle\big)^2 - \langle X, \psi\rangle^2 \,. \tag{2.5}$$

In Theorem 2.1 we study the error of this asymptotic classifier T^0, and derive the function ψ that guarantees optimal classification in the Gaussian case. In Theorem 2.2 we show that the same function can give perfect classification in important non-Gaussian settings. In particular, these two theorems explain theoretically why, in practice, despite their lack of sophistication, linear classifiers can perform very well in the case of functional data.

We need the following notation (here, X is drawn from Π_0 or Π_1):

$$Q = \int_{\mathcal{I}} \psi\,(X - EX)\,, \quad \nu = \langle \mu, \psi\rangle = \int_{\mathcal{I}} \psi\,\mu\,, \tag{2.6}$$

$$\sigma^2 \equiv \mathrm{var}(Q)\,, \quad \psi^{(r)} = \sum_{j=1}^{r} \theta_j^{-1} \mu_j\, \phi_j \;. \tag{2.7}$$

The proof of Theorem 2.1 is given in section A.1; Theorem 2.2 can be proved similarly.

Theorem 2.1. *Assume that the data are Gaussian and* (2.2) *holds. Then, no matter what the value of the prior probability* π*:*
(a) The probability of misclassification for the classifier T^0 *equals* $\mathrm{err} = 1 - \Phi(\nu/2\sigma)$.
(b) The minimum value, err_0*, of this error is given by* $\mathrm{err}_0 = 1 - \Phi\{\frac{1}{2}\,(\sum_{j\ge1}\,\theta_j^{-1}\,\mu_j^2)^{1/2}\}$.
If $\sum_{j\ge1}\,\theta_j^{-2}\,\mu_j^2 < \infty$, err_0 *is the classification error for the classifier* T^0 *computed with*

$\psi = \psi^{(\infty)}$. *Otherwise,* err_0 *is achieved as the limit of classification errors for the classifier* T^0 *computed at the sequence* $(\psi^{(r)})_{r \geq 1}$ *of functions* ψ.
(c) If $\sum_{j \geq 1} \theta_j^{-1} \mu_j^2 = \infty$ *then* $\mathrm{err}_0 = 0$ *and there is perfect classification.*
(d) In this problem, no classifier can give better results than those described above, in the following sense: if $\sum_{j \geq 1} \theta_j^{-1} \mu_j^2 = \infty$, *then, in the case of infinite training samples the error rate of an optimal classifier, based on a likelihood ratio test for functional data, also equals* err_0. *If* $\sum_{j \geq 1} \theta_j^{-2} \mu_j^2 < \infty$, *and if* T^0, *at* (2.5), *is replaced by its version for known* π *(see Remark 1), then the optimal error rate also equals* err_0.

Theorem 2.2. *If the data are not Gaussian, then if the populations* Π_0 *and* Π_1 *have prior probabilities* π *and* $1-\pi$, *respectively, and if* (2.2) *holds, then: (a) The probability of misclassification for the classifier* T^0 *equals* $\mathrm{err} = \pi\, P(R > \nu/2\sigma) + (1-\pi)\, P(R < -\nu/2\sigma)$, *where* $R = Q/\sigma$ *has zero mean and unit variance. (b) If* $\sum_{j \geq 1} \theta_j^{-1} \mu_j^2 = \infty$ *then, choosing* $\psi = \psi^{(r)}$ *and letting* r *diverge, the probability of misclassification for the classifier* T^0 *tends to* $\mathrm{err}_0 = 0$, *and there is perfect classification.*

Note that in Theorem 2.2 we discuss only the perfect classification property, since this theorem treats general distributions. Indeed, since (in non-Gaussian cases) properties of the classifier depend intimately on those distributions via Q, then the general case seems not to admit an elementary, insightful derivation of the optimal ψ.

Theorem 2.1 provides detailed information about the dichotomy discussed in section 1. It implies that when $\sum_{j \geq 1} \theta_j^{-2} \mu_j^2 < \infty$, perfect classification is not possible but the centroid method is optimal in Gaussian cases. Here, the function ψ that guarantees optimality is well defined, and is given by $\psi^{(\infty)} = \sum_{j=1}^{\infty} \theta_j^{-1} \mu_j \phi_j$. Taking $\psi = \psi^{(\infty)}$, the classifier assigns X to Π_0 or Π_1 according as the statistic $-2 \sum_{j=1}^{\infty} \theta_j^{-1} \mu_j X_j + \sum_{j=1}^{\infty} \theta_j^{-1} \mu_j^2$ is positive or negative, respectively. This rule is analogous to that for the Bayes classifier employed in finite dimensional classification problems, and our results are a natural extension of the optimality property of the multivariate Bayes classifier. In this case, this rule is also equivalent to the linear discriminant method studied by James and Hastie (2001) and Preda et al. (2007).

The most interesting result is that when $\sum_{j \geq 1} \theta_j^{-2} \mu_j^2 = \infty$, since this case exploits the functional nature of the data and can result in perfect classification. This contrasts with earlier work on related functional linear regression problems, where this case is usually considered to be degenerate; see section 2.3 for discussion. When

$\sum_{j\ge1}\theta_j^{-1}\mu_j^2=\infty$, no rescaled version of $\psi^{(r)}$ converges to a nondegenerate limit in L_2, but perfect classification can be obtained as n diverges.

We should stress that perfect classification holds if and only if $\sum_{j\ge1}\theta_j^{-1}\mu_j^2=\infty$, and that optimal classification is achieved for a particular $\psi=\psi^{(\infty)}$ if and only if $\sum_{j\ge1}\theta_j^{-2}\mu_j^2<\infty$. The relationship between these two conditions divides the classification problem into three parts — one, where $\sum_{j\ge1}\theta_j^{-2}\mu_j^2<\infty$ and optimal classification is imperfect but is achieved using a single ψ; another, where $\sum_{j\ge1}\theta_j^{-2}\mu_j^2=\infty$ but $\sum_{j\ge1}\theta_j^{-1}\mu_j^2<\infty$, and optimal classification is imperfect but is achieved only along the sequence $\psi^{(r)}$, which does not converge; and a third, where $\sum_{j\ge1}\theta_j^{-1}\mu_j^2=\infty$ and perfect classification is achieved, but only along the non-convergent series $\psi^{(r)}$.

In view of the preceding discussion one might be tempted to construct a test that sheds light on which of the three different regimes discussed above obtains. However, it is perhaps more appropriate to construct several classifiers, one or more of which is sensitive to which regime pertains, while the others are more conventional. The error rates of these classifiers can be estimated, and the results used to guide the final approach.

Remark 1. When π is known, the centroid method can be adjusted to take the probabilities of each population into account. In this case, $T^0(X)$ at (2.5) is replaced by $T^0(X)=\big(\langle X,\psi\rangle-\langle\mu,\psi\rangle\big)^2-\langle X,\psi\rangle^2-2\log\{(1-\pi)/\pi\}$. However, in most situations π is unknown and is either taken equal to $\frac12$, leading to formula (2.5) for T^0, or is estimated by the proportion n_0/n of observations in the sample coming from Π_0.

2.3 Empirical choice of ψ: truncation and partial least squares

First approach. In practice we can calculate only finite sums, and therefore, even if $\psi^{(\infty)}$ is well defined, we can use only truncated versions of it. One obvious possibility is to apply the centroid classifier with $\psi^{(r)}$, where r is finite. Of course, in practice θ_j, ϕ_j and μ_j are unknown, and are estimated from the data. We take $\widehat\theta_j$ and $\widehat\phi_j$ to be the estimators obtained by empirical principal component analysis of the entire dataset (taking there each subpopulation to be centered at its empirical mean), and $\widehat\mu_j=\int_{\mathcal I}\widehat\mu\widehat\phi_j$, with $\widehat\mu(t)=n_1^{-1}\sum_{i=1}^{n_1}X_{1i}(t)-n_0^{-1}\sum_{i=1}^{n_0}X_{0i}(t)$. Then, for $r=1,\dots,n$ we

deduce empirical versions $\widehat{\psi}^{(r)}(t) = \sum_{j=1}^{r} \widehat{\theta}_j^{-1} \widehat{\mu}_j \widehat{\phi}_j(t)$ of $\psi^{(r)}(t)$. We suggest choosing r by minimising a cross-validation (CV) estimator of classification error.

To appreciate how this is done, let $\mathcal{C}(X \,|\, \psi)$ denote the index of the population to which a new function X is allocated by the classifier using the projection based on a function ψ, and let P_k denote probability measure under the assumption that X comes from Π_k. The error rate of $\mathcal{C}$ is defined by

$$\mathrm{err}(\psi) = \pi\, P_0\{\mathcal{C}(X \,|\, \psi) = 1\} + (1 - \pi)\, P_1\{\mathcal{C}(X \,|\, \psi) = 0\}\,. \tag{2.8}$$

To estimate error rate we randomly create B partitions of the training sample, as in, for example, Hastie et al. (2009). For $b = 1, \ldots, B$, each partition splits the sample into two subsamples, $\{X_{1b}^*, \ldots, X_{mb}^*\}$, where $m = [n/K]$ and $\{X_{m+1,b}^*, \ldots, X_{nb}^*\}$, with $\{X_{1b}^*, \ldots, X_{nb}^*\}$ denoting a random permutation of the $n = n_0 + n_1$ observations from training sample. Then we estimate the classification error by taking an average of B leave-m-out CV estimators. More precisely, we take

$$r = \mathrm{argmin}_{1 \le j \le n} \widehat{\mathrm{err}}(\widehat{\psi}^{(j)}), \tag{2.9}$$

where $\widehat{\mathrm{err}}$ is the empirical error,

$$\widehat{\mathrm{err}}(\psi) = \frac{1}{B(n-m)} \sum_{b=1}^{B} \sum_{i=m+1}^{n} \sum_{k=0}^{1} I\{\mathcal{C}_{k,-b}(X_{ib}^* \,|\, \psi) = 1 - k, X_{ib}^* \in \Pi_k\}\,, \tag{2.10}$$

with $\mathcal{C}_{k,-b}$ denoting the version of $\mathcal{C}$ constructed by omitting $\{X_{m+1,b}^*, \ldots, X_{n,b}^*\}$ from the training sample; and the indicator function I is defined by $I(\mathcal{E}) = 1$ if the event $\mathcal{E}$ holds, and $I(\mathcal{E}) = 0$ otherwise.

Second approach. An alternative approach can be constructed if we note the connection between our problem and linear regression. Let $Y = I(X \in \Pi_1)$, and let Φ_r be the space generated by $\phi_1, \ldots, \phi_r$. It is not difficult to prove that

$$\mathrm{argmin}_{\beta \in \Phi_r} E\Big\{Y - EY - \int_{\mathcal{I}} \beta(X - EX)\Big\}^2 = \alpha_r \psi^{(r)}, \tag{2.11}$$

where α_r is a constant. In other words, $\psi^{(r)}$ is, up to a constant multiple, equal to the slope of the best approximation (in the mean squared sense) to Y by a linear function of X, where the slope function is restricted to lie in Φ_r. Since the centroid classifier is invariant to changes of scale of ψ, it gives the same results whether ψ is equal to $\psi^{(r)}$

or to the slope $\beta \in \Phi_r$. Note that, when $\sum_{j\geq 1} \theta_j^{-2}\, \mu_j^2 < \infty$, the minimisation problem at (2.11) is well defined for $r \to \infty$, and the well defined solution, in $L_2(\mathcal{I})$, is that of a standard linear regression problem. On the other hand, when $\sum_{j\geq 1} \theta_j^{-2}\, \mu_j^2 = \infty$ the limit of $\psi^{(r)}$ as $r \to \infty$ is not well defined. This implies that we cannot define a standard unrestricted linear regression model for the relation between Y and X, and in the literature it has been noted that, in this case, the least squares criterion provides inconsistent estimators of the regression function; see for example Preda and Saporta (2002). While this is true, it does not cause difficulty for classification in either theory or practice; see sections 2.2 and 3.

The connection with linear regression suggests that we could choose ψ in practice using iterative partial least squares (PLS); see Preda and Saporta (2003) and Preda et al. (2007). Here we compare the asymptotic form of PLS, which we refer to as APLS, with the asymptotic approach based on $\psi^{(r)}$. At step r, APLS approximates Y by $Y = EY + \int_{\mathcal{I}} \beta_r\,(X - EX) + f_r$, where $\beta_r(t) = \sum_{j=1}^{r} c_j W_j(t)$, with the W_js orthogonal and each of norm 1, and with f_r denoting the residual of the approximation. The function W_1 is chosen to maximise $\mathrm{cov}(Y, \int_{\mathcal{I}} W_1 X)$, and, given the jth subspace generated by $W_1, \ldots, W_{j-1}$, W_j is chosen to maximise the covariance of $\int_{\mathcal{I}} W_j X$ and the part of Y that is left to explain, that is f_{j-1}. See section A.2 for details.

Like $\psi^{(r)}$, the function β_r provides an approximation to $Y - EY$ of the form $\int_{\mathcal{I}} \beta\,(X - EX)$, but at each step r, β_r is chosen within the r dimensional space generated by $W_1, \ldots, W_r$, which explains the majority of the covariance of X and Y. By comparison, the space Φ_r is the r dimensional space that explains the largest part of the empirical covariance of X. Clearly, since PLS tries to capture as much of the linear relation between X and Y as possible, then in general the r-dimensional space generated by PLS is better suited for classification than the space $\Phi^{(r)}$. In practice, if the main differences between the means of Π_0 and Π_1 come from μ_j with j relatively small, then the two approaches will give very similar results. The importance of PLS becomes clearer in cases where the most important differences between the means of the two populations come mostly from components μ_j for large j, see our numerical investigation in section 3. We give details of the iterative PLS algorithm in section A.2, which provides a finite sample version $\widehat{\beta}_r$ of β_r. We suggest choosing r for PLS by minimising the empirical error at (2.10), with ψ there replaced by $\widehat{\beta}_r$.

3 Numerical properties

3.1 Classifiers considered in our numerical work

Based on the asymptotic arguments developed in the previous sections, we suggest applying the centroid classifier using either $\psi = \widehat{\psi}^{(r)}$ or $\psi = \widehat{\beta}_r$, with r chosen by CV as described in section 2.3. We implemented both methods, and in each case took $B = 200$ and $K = 5$. To reduce the amount of computation, and rule out the most noisy eigenfunctions, we restricted our search to $r \leq n/2$. As usual with smoothing parameter selection via CV, the empirical criterion $\widehat{\text{err}}$ typically has several local minima, and the global minimum is not necessarily the best one. In these cases we kept only the two smallest values of r corresponding to a local minimum of $\widehat{\text{err}}$, and chose a global minimum among those two. For general considerations about taking the second left-most minimisers of CV criteria, see Hall and Marron (1991).

We compared these two procedures with an approach also used in practice, where the centroid-based method is applied to a multivariate projection $(\int_{\mathcal{I}} X\widehat{\phi}_1, \ldots, \int_{\mathcal{I}} X\widehat{\phi}_p)$, using p empirical principal component scores. We chose p by minimising the p-variate projection version of (2.10). We also considered a more sophisticated nonparametric classifier suggested by Ferraty and Vieu (2006), based on the model $Y = g(\int_{\mathcal{I}} \beta X) + \epsilon$, where g is an unknown regression curve and $Y = I\{X \in \Pi_1\}$. The classifier assigns X to Π_0 if $\widehat{g}(\int_{\mathcal{I}} \widehat{\beta}_r X) < \frac{1}{2}$, and to Π_1 otherwise, where $\widehat{g}$ is a nonparametric estimator of g and $\widehat{\beta}_r$ is the curve obtained by an rth stage PLS. We implemented this classifier using the function `funopare.knn.gcv` of Ferraty and Vieu (2006), where we chose r by a CV estimator of the error rate for this classifier. We denote the fully empirical centroid classifiers based on $\widehat{\psi}^{(r)}$, $\widehat{\beta}_r$ and $(\widehat{\phi}_1, \ldots, \widehat{\phi}_p)$ by, respectively, CENT_{PC1}, CENT_{PLS} and $\text{CENT}_{\text{PC}p}$. We denote the nonparametric classifier by NP.

3.2 Simulated examples

In this section we illustrate several properties of the centroid classifier. In all our examples we generated n curves from two populations, Π_0 and Π_1, of respective sizes $n_0 = n/2$ and $n_1 = n/2$. For $i = 1, \ldots, n_k$, where $k = 0, 1$, we took $X_{ki}(t) = \sum_{j=1}^{40} (\theta_j^{1/2} Z_{jk} + \mu_{jk})\, \phi_j(t)$. Here the Z_{jk}s were independent standard normal random variables, or independent exponential exp(1) variables centered to zero. In each case

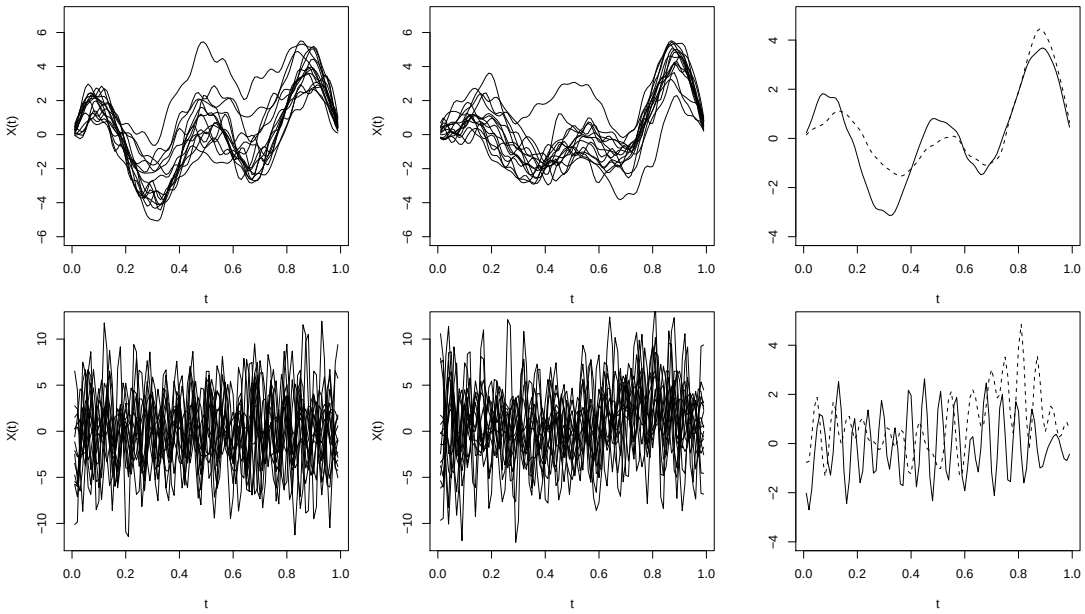

Figure 1: Sample of size $n_k = 15$ from population Π_k, for $k = 0$ (left) and $k = 1$ (middle), and empirical mean curves from the two samples (right). First row: Example 2, second row: Example 3.

we took $\phi_j(t) = \sqrt{2}\sin(\pi j t)$, let $t \in \mathcal{I} = [0, 1]$ and generated the data on a discrete grid of 100 equispaced points in $\mathcal{I}$. We chose the μ_{jk}s and θ_js depending on the property of the classifier that we wanted to illustrate (see below). In each case we generated $M = 200$ samples. For each sample we constructed the data-driven version of each classifier, and then tested the practical performance of the resulting classifier by assigning to Π_0 or Π_1 each of $B = 200$ randomly generated test data (100 from Π_0 and 100 from Π_1). For each $m = 1, \ldots, M$, and each classifier, we calculated the percentage P_m of the B test curves that were misclassified. Below, to assess the performance of the classifiers, we show means and standard deviations of these M values of P_m.

Perfect classification when the mean differences arise early in the sequence $(\mu_j)_{j\geq 1}$. Our first illustration, Example 1, is constructed to show that near perfect classification is possible in practice and in realistic settings. For $j = 1, \ldots, 40$, we took $\theta_j = j^{-2}$; for $k = 0, 1$ and $j > 6$ we let $\mu_{jk} = 0$, and we set the other components equal to $(\mu_{10}, \mu_{20}, \mu_{30}, \mu_{40}, \mu_{50}, \mu_{60}) = (0, -0.5, 1, -0.5, 1, -0.5)$, $(\mu_{11}, \mu_{21}, \mu_{31}, \mu_{41}, \mu_{51}, \mu_{61}) = (0, -0.75, 0.75, -0.15, 1.4, 0.1)$. The Z_{jk}s were independent centered exp(1) variables. Next, in Example 2, to show that the method also works when the variances of the two populations differ, we took the same setting as in Example 1 but, for the data of

Table 1: Percentage of misclassified observations in the simulated examples: Mean of P_m (stdev of P_m) calculated from $M = 200$ Monte Carlo simulations.

Data	n	CENT_{PC1}	CENT_{PLS}	$\text{CENT}_{\text{PC}p}$
Example 1	30	4.45 (4.31)	3.13 (1.84)	13.6 (8.62)
	50	3.13 (2.26)	2.73 (1.41)	10.7 (7.43)
Example 2	30	6.51 (5.45)	4.98 (2.76)	17.4 (9.73)
	50	4.84 (4.32)	3.56 (2.00)	14.6 (8.33)
Example 3	30	50.0 (3.64)	3.64 (3.85)	50.0 (3.41)
	50	49.6 (3.70)	0.14 (0.30)	49.7 (3.72)

population Π_0, we replaced θ_j by $\theta_{j0} = 1.5\,\theta_j$. In Figure 1, we show 15 typical curves from Π_0 and 15 curves Π_1, as well as the two empirical mean curves $\bar{X}_0(t)$ and $\bar{X}_1(t)$ calculated from these data. As can be seen from this graph, the curves are quite variable, and the empirical means are not dramatically different. However, in both cases (Examples 1 and 2), $\sum_j \mu_j^2 \theta_j^{-1}$ is large, where θ_j is the jth largest eigenvalue of the pooled centered data, and μ_j is the projection, on the jth eigenfunction, of the difference between the mean curves from Π_0 and Π_1. As predicted by the theory, CENT_{PC1} and CENT_{PLS} work nearly perfectly in these cases; see Table 1, where we show means and standard deviations of the percentage P_m of misclassified data. Clearly, $\text{CENT}_{\text{PC}p}$ does not compete with these approaches.

Superiority of PLS. Our Example 3 illustrates the superiority of PLS in cases where the most important mean differences do not come from the first few components μ_j. For $j = 1, \ldots, 40$, we took $\theta_j = \exp\{-\{2.1 - (j-1)/20\}^2\}$, $\mu_{j0} = 0$ and $\mu_{j1} = 0.75 \cdot (-1)^{j+1} \cdot I\{j \leq 3\}$. The Z_{jk}s were independent standard normal variables. In this case the largest θ_js are for j close to 40, whereas the μ_js are all zero, except for the first three (note that, to avoid having to redefine the functions ϕ_j, we did not index the θ_js in decreasing order). Even though the θ_js do not decay very fast to zero, it is clear that CENT_{PC1} will encounter difficulties in practice, as in order to work well this method needs to estimate reasonably well the principal components corresponding to the smallest 38th to 40th eigenvalues. That is possible only if the sample size is much larger than 40. This difficulty experienced by the PC approach is partly a consequence of the eigenvalues being spaced rather closely together, as well as of the common challenge that PC methods face when one is estimating relatively high order components. On the other hand, PLS is able to quickly focus on the important components for classification, and as a result, CENT_{PLS} gives near

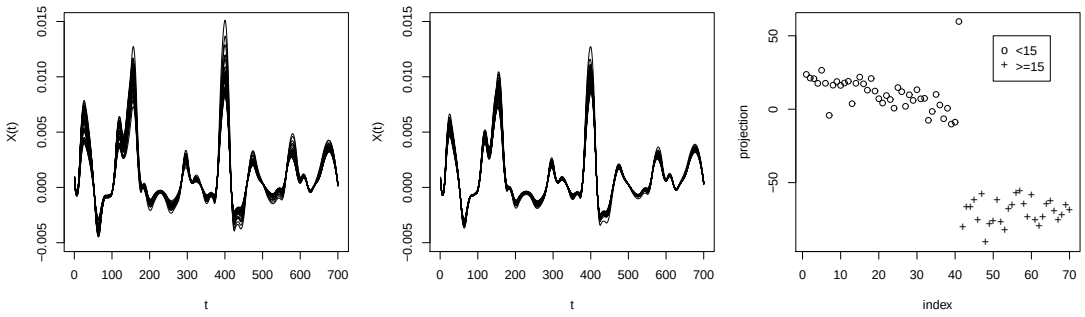

Figure 2: Wheat derivative curves when the protein level is less (left) or greater (middle) than 15. Right: a typical test sample projected on $\widehat{\psi}^{(r)}$ when $n = 30$.

perfect classification. As predicted by our theory, the classification problem is much easier when the mean differences come from components that correspond to small eigenvalues θ_j. However, only the PLS approach is able to pick up those differences in practice, by focusing on basis functions which explain at most the linear relation between Y and X.

3.3 Real-data examples

We applied our method to three real-data examples. The first two datasets show that very good performance can be achieved in practice by the centroid classifier. The last one is an example where the differences between the two populations are difficult to capture. In each case we had a sample of $N = N_0 + N_1$ observations (N_0 from Π_0 and N_1 from Π_1), which we split randomly, $M = 200$ times, into a training sample of size n, with $n = 30$, 50 and 100, and a test sample of size $N - n$. In each case we constructed M times the four empirical classifiers, described in section 3.1, from the training data, and used them to classify the test data. We then calculated, for each case and each classifier, the corresponding M values of the percentage P_m of test observations that were misclassified.

Our first example illustrates the perfect classification property. The data consist of near infrared spectra of 100 wheat samples with known protein content, and measured from 1100nm to 2500nm in 2nm intervals. See Kalivas (1997) for a description. We used the protein content to separate the data into two populations Π_0 (protein content less than 15) and Π_1 (protein content greater than 15) of sizes $N_0 = 41$ and $N_1 = 59$. As usual with chemometrics data, we worked with the derivative curves of the spectra, which we estimated with splines as in Ferraty and Vieu (2006). The

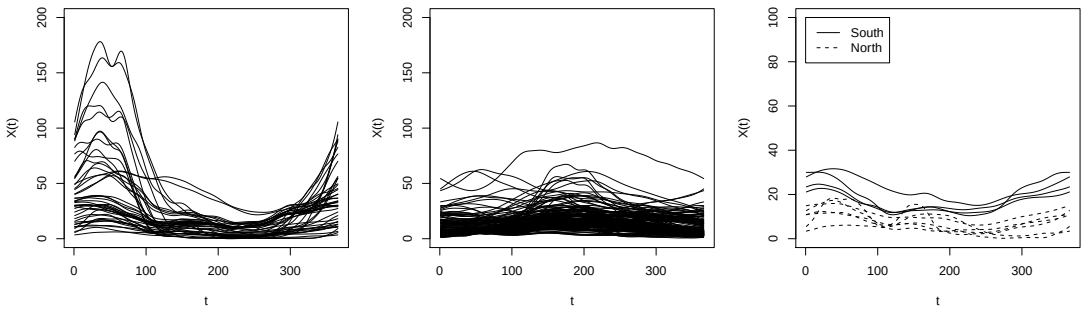

Figure 3: Rainfall curves for Australian northern (left) and southern (middle) weather stations, and misclassified curves for a typical test sample when $n = 100$ (right).

Table 2: Percentage of misclassified observations: Mean of P_m (stdev of P_m) calculated from $M = 200$ randomly chosen test samples.

Data	n	CENT_{PC1}	CENT_{PLS}	NP	$\text{CENT}_{\text{PC}p}$
Wheat	30	0.89 (2.49)	0.46 (1.24)	0.49 (1.29)	15.0 (5.24)
	50	0.22 (1.09)	0.06 (0.63)	0.01 (0.14)	14.4 (5.52)
Rain	30	10.6 (3.09)	10.4 (3.24)	10.6 (3.62)	12.9 (3.04)
	50	10.3 (2.90)	9.80 (3.06)	9.32 (3.60)	13.3 (2.63)
	100	9.45 (3.02)	8.98 (2.96)	8.34 (3.33)	13.3 (3.32)
Phoneme	30	22.5 (3.59)	24.2 (5.37)	24.4 (5.31)	23.7 (2.37)
	50	20.8 (2.08)	21.5 (3.02)	21.9 (2.91)	23.4 (1.80)
	100	20.0 (1.09)	20.1 (1.12)	20.1 (1.37)	23.4 (1.36)

data for the two groups are plotted in Figure 2. In this example we can get near perfect classification using the centroid classifier, when the data are projected onto the right function. In Figure 2 we show, for a typical training sample of size $n = 30$, the test data projected onto the empirical function $\widehat{\psi}^{(r)}$. In this sample the projected data from the two groups are perfectly separated. Means and standard deviations of the percentage P_m of misclassified observations are shown in Table 2. On average for these data, CENT_{PC1} and CENT_{PLS} gave near perfect classification since the mean percentage of misclassification was less than 1%, even for training samples of size $n = 30$. Unsurprisingly in this case, the nonparametric classifier gave essentially the same results as these two methods (although slightly worse for $n = 30$ and slightly better for $n = 50$). The p-dimensional centroid classifier $\text{CENT}_{\text{PC}p}$ misclassified a much higher proportion of the data.

In the second example we consider the $N = 191$ rainfall curves from $N_0 = 43$ northern (Π_0) and $N_1 = 148$ southern (Π_1) Australian weather stations, used by Delaigle and Hall (2010) and available at `http://dss.ucar.edu/datasets/ds482.1`. Here each curve $X_{ik}(t)$ represents rainfall at time t, where $t \in [0, 365]$ denotes the

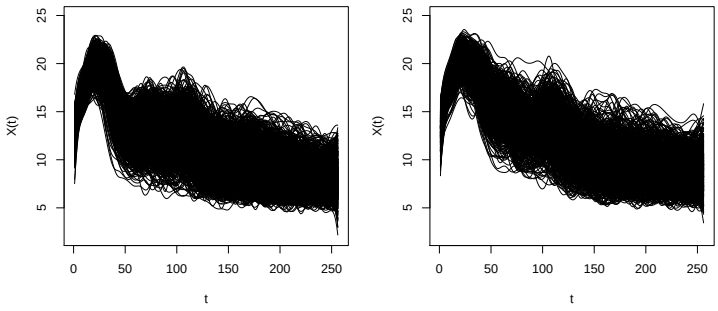

Figure 4: Phoneme curves for the sound "ao" (left) and found the sound "aa" (right).

period that had passed, in a given year, at the time of measurement. The raw data were observed daily and, as in Delaigle and Hall (2010), for each station we averaged the rainfall over the years for which the station had been operating, using a local linear smoother. The data curves are shown in Figure 3. In this dataset perfect classification is not possible because some of the stations geographically located in the North have a rainfall pattern typical of the South, and vice-versa; and others have an almost flat rainfall curve. This is illustrated in Figure 3 as well, where we show, for a training sample of size $n = 50$, the misclassified northern and southern curves out of the 141 test curves. Table 2 depicts means and standard deviations of the percentage P_m of misclassified observations for each method and various training sample sizes n. The numbers reveal a low misclassification rate of 10% or less when using CENT_{PC1} or CENT_{PLS}, but a higher classification error for the $\text{CENT}_{\text{PC}p}$ approach. In this example too, as suggested by the theory, the improvement obtained by the nonparametric classifier NP was minor: the mean of P_m decreased slightly, but its standard deviation increased.

In the third example we analysed the digitised speech phoneme data described in Hastie et al. (2009) and available at `www-stat.stanford.edu /ElemStatLearn`. The dataset contains log-periodograms constructed from 32 msec long recordings of males pronouncing five different phonemes. We took Π_0 and Π_1 to be the population of, respectively, the phonemes "aa" as in "dark" and "ao" as in "water", whose log-periodograms are hard to distinguish from one another, see Figure 4. Here, $N_0 = 695$ and $N_1 = 1022$, and each $X_{ik}(t)$ was observed at 256 equispaced frequencies t. We made the curves continuous by passing a local linear smoother through the points, and kept only the least noisy part by truncating them to the interval $\mathcal{I} = [1, 50]$. The means and standard deviations of the percentages of misclassified curves are shown in Table 2. Here, for CENT_{PC1} and CENT_{PLS}, the percentage of misclassification is

higher (about 23% for $n = 30$ and 20% for $n = 100$), and further from zero than in the previous examples. However, despite its much higher level of sophistication, the nonparametric classifier does not do any better than these simple classifiers.

4 Further theoretical properties

4.1 Theoretical properties of empirically chosen ψ

In this section, for simplicity we establish theoretical properties of the leave-one-out CV estimator of classification error. Similar methods and results apply in the case of multi-fold CV. Likewise, to keep the proofs relatively short and transparent we prove the results when CV is used to select r in $\psi^{(r)}$, and when the search for r is restricted to integers less than a certain number $\widehat{R}$ (see below), which prevents us from having to impose stringent restrictions on the eigenvalue spacings $\theta_j - \theta_{j+1}$. However, if we were able to rely on the differences $\theta_j - \theta_{j+1}$ being no smaller than a specified function of j then we could define $\widehat{R}$ differently, in particular taking it larger. In practice, as in more conventional problems where CV is used, we found that better results were obtained when restricting to the first two local minima of the CV criterion. Justification can be given as in Hall and Marron (1990).

Let T_{0r} be the version of the classifier T^0, defined at (2.5), when $\psi = \psi^{(r)}$. We assume that:

(a) for some $\eta \in (0, 1]$, $(n_0 + n_1)^\eta = O\{\min(n_0, n_1)\}$; (b) for $k = 0$ and 1, $\sup_{t \in \mathcal{T}} E_k\{|X(t)|^C\} < \infty$, where $C \geq 4$ depends on η; (c) no ties exist among the eigenvalues θ_j; and (d) for $k = 0, 1$, $P_k\{\sup_{r \geq 1} |T_{0r}(X)| \leq c\}$ converges to zero as $c \to 0$. (4.1)

Let $\widehat{R} + 1 = \inf\{j \geq 1 : \widehat{\theta}_j - \widehat{\theta}_{j+1} < \eta_n\}$, where η_n denotes a sequence of constants decreasing to zero but such that $n^{1/5}\,\eta_n \to \infty$ as $n \to \infty$. In the theorem below, where we write $\widehat{\mathrm{err}}(\widehat{\psi}^{(r)})$ for the leave-one-out CV estimator of $\mathrm{err}(\psi^{(r)})$, we show that this approach is consistent if r is constrained to lie in the range $1 \leq r \leq \widehat{R}$. The proof is given in section A.1.

Theorem 4.1. *If* (2.1), (2.2) *and* (4.1) *hold, and if* $n_0/n = \pi$, *then, as* n *increases,* $\widehat{R} \to \infty$ *and*

$$\max_{1 \leq r \leq \widehat{R}} \left| \widehat{\mathrm{err}}\big(\widehat{\psi}^{(r)}\big) - \mathrm{err}\big(\psi^{(r)}\big) \right| \to 0\,, \tag{4.2}$$

where in both cases the convergence is in probability.

4.2 Classification when the population covariances differ

The results of section 2 can be extended to more general situations, where the populations may not differ only in terms of means. In this section we adapt our theory to such cases. The method is implemented as discussed in section 2, and the results of section 2.3 remain valid.

Assume that a random function X drawn from population Π_k (resp., from Π, the mixture of the two centered populations) has bounded covariance function K_k (resp. K), for $k = 0, 1$. In analogy with (2.1), let

$$K_k(u,v) = \sum_{j=1}^{\infty} \theta_{kj}\, \phi_{kj}(u)\, \phi_{kj}(v)\ , \quad K(u,v) = \sum_{j=1}^{\infty} \theta_j\, \phi_j(u)\, \phi_j(v), \tag{4.3}$$

respectively, denote the standard principal component expansion of K_k (resp. K) where θ_{kj} and ϕ_{kj} (resp. θ_j and ϕ_j) are nonzero eigenvalues, and the respective eigenfunctions, of the linear operator K_k (resp. K) with kernel equal to the function K_k (resp. K), and $k = 0$ or 1. Reflecting (2.3), express the mean μ in terms of the basis $\phi_1, \phi_2, \ldots$: $\mu = \sum_{j=1}^{\infty} \mu_j\, \phi_j$.

Let ψ denote a function on $\mathcal{I}$; define $T(X)$, and its large-sample limit $T^0(X)$, as at (2.4) and (2.5), respectively; and determine that X comes from Π_0 if $T^0(X) > 0$, and from Π_1 otherwise. Recall that E_k and P_k denote expectation and probability measure under the hypothesis that X comes from Π_k. We impose the following condition:

(i) $E_0(X) = 0$ and $E_1(X) = \mu$, (ii) the covariance functions K, K_0 and K_1 are strictly positive definite and bounded, (iii) $\|\mu\| < \infty$, (iv) $(\sum_{1 \le j \le r} \theta_j^{-1}\, \mu_j^2)^2 / (\sum_{j=1}^{\infty} \theta_{\ell j} \left\{ \sum_{1 \le m \le r} \theta_m^{-1}\, \mu_m \int \phi_m(u)\, \phi_{\ell j}(u)\, du \right\}^2)$ $\to \infty$ as $r \to \infty$, for $\ell = 0$ and 1. (4.4)

Assumption (4.4)(ii) implies that the eigenvalues θ and θ_{kj} in (4.3) and (4.3) are all strictly positive. Parts (i)–(iii) of (4.4) involve only standard assumptions in functional data analysis, and part (iv) holds in many cases too. The following result is derived in section A.1.

Theorem 4.2. *Assume* (4.4)*(i)–(iii), and suppose too that the populations* Π_0 *and* Π_1 *have prior probabilities* π *and* $1-\pi$*, respectively. Then: (a) The probability of misclassification, when using the centroid classifier* T^0 *based on* $\psi = \psi_r = \sum_{1 \le j \le r} \theta_j^{-1}\, \mu_j \phi_j$*,*

equals

$$\pi P\Big\{R_0 \geq \sum_{1\leq j\leq r} \theta_j^{-1}\,\mu_j^2/(2\alpha_0)\Big\} + (1-\pi)\,P\Big\{\,R_1 > \sum_{1\leq j\leq r} \theta_j^{-1}\,\mu_j^2/(2\,\alpha_1)\Big\}$$

where, for $\ell = 0$ *and* 1, *the random variables* R_ℓ *have zero mean and unit variance, and* $\alpha_\ell^2 = \sum_{j=1}^\infty \theta_{\ell j}\,\{\int \psi(u)\,\phi_{\ell j}(u)\,du\}^2$. *(b) If* (4.4)*(iv) is satisfied then we obtain perfect classification, in the sense that, as we proceed along the sequence* $(\psi_r)_r$, *we have* $\pi P_0\{T^0(X) < 0\} + (1-\pi)P_1\{T^0(X) > 0\} \to 0$.

The theorem implies, for example, that if two populations share the same eigenfunctions and differ only through their eigenvalues, then the classification error is small if $\sum_{1\leq j\leq r} \theta_j^{-1}\mu_j^2$ is large, and converges to zero as $r \to \infty$ if the series diverges. This result is similar to that in the case where the two populations have the same covariance function.

The centroid-method classifier is designed to detect differences between the populations Π_0 and Π_1 that occur in location rather than scale. In order to focus them more sharply on location the definitions of $T(X)$ and $T^0(X)$, in section 2.2, could incorporate adjustments for scale. However, in principle, differences in scale can be valuable, and by not adjusting for them the classifier may gain a little extra power when scale differences are significant.

5 Conclusion

Our theoretical study has revealed an intriguing dichotomy between two cases arising in functional classification problems. In the first case, asymptotically perfect classification is not possible, and properties of linear classifiers are simple extensions of their finite dimensional counterparts. In particular, the centroid classifier is equivalent to classification based on prediction by linear regression. In the second case, where asymptotically perfect classification is possible, the centroid classifier can give perfect classification when the data are projected onto a diverging sequence of functions $\psi^{(r)}$, with r increasing. Here, the standard linear regression model is not well defined, but asymptotically perfect classification can be obtained by considering a sequence of linear models in subspaces of dimension r, where we let r diverge. We have shown that these findings explain why, in the functional data context, simple linear classifiers

often perform very well. The two main contenders have been shown to be classifiers based on principal components (PC) and PLS, with PLS having advantages in cases where PC methods would require a large number of terms.

A Appendix: Theoretical details

A.1 Proofs

Proof of Theorem 2.1. (a). If X is drawn from Π_0 then $T^0(X) = \bigl(Q - \int_{\mathcal{I}} \psi\,\mu\bigr)^2 - Q^2 = \nu^2 - 2\,\nu\,Q$, where Q, μ and ν were defined at (2.6) and (2.7). Writing N for a random variable with the normal $\mathrm{N}(0,1)$ distribution, and Φ for the distribution function of N, we see that the probability that the classifier mistakenly assigns X to Π_1, when X is actually drawn from Π_0, is given by: $P_0\bigl(T^0 \le 0\bigr) = P\bigl(\nu^2 - 2\,\nu\,\sigma\,N < 0\bigr) = 1 - \Phi(\nu/2\sigma)$, provided that $\nu \neq 0$ (P_k is defined at Page 8). This is also the probability of incorrect classification if X is drawn from Π_1. Result (a) follows from the fact that $P_0\bigl(T^0 \le 0\bigr) = P_1\bigl(T^0 > 0\bigr)$.

(b)–(c). No matter whether X comes from Π_0 or Π_1, $X - E_k X$, with $k = 0$ or 1, has the distribution of Z, where Z is a zero-mean Gaussian process with covariance function K. Using the spectral decomposition of K at (2.1) we can express Z in its Karhunen-Loève expansion:

$$Z = \sum_{j=1}^{\infty} \theta_j^{1/2}\,Z_j\,\phi_j\,, \tag{A.1}$$

where the variables $Z_1, Z_2, \ldots$ are independent and identically distributed as normal $\mathrm{N}(0,1)$. Assume first that

$$\sum_{j=1}^{\infty} \theta_j^{-2}\,\mu_j^2 < \infty\,. \tag{A.2}$$

We seek to find the function $\psi(t) = \sum_{j=1}^{\infty} \lambda_j\,\phi_j(t)$ that minimises classification error. From (A.1) and the definition of Q, μ and ν, we can write $Q = \sum_{j=1}^{\infty} \lambda_j\,\theta_j^{1/2}\,Z_j$, $\nu = \sum_{j=1}^{\infty} \lambda_j\,\mu_j$ and $\sigma^2 = \sum_{j=1}^{\infty} \lambda_j^2\,\theta_j$. To minimise the error it suffices to maximise $\nu^2/\sigma^2 = (\sum_{j=1}^{r} \lambda_j\,\mu_j)^2 / \sum_{j=1}^{r} \lambda_j^2\,\theta_j$ with respect to the λ_js, that is, to take $\lambda_k = c \cdot \theta_k^{-1}\mu_k$, for any finite constant c, (the error rate does not depend on c). Hence the minimum error is achieved by taking $\psi = \psi^{(\infty)}$. By (A.2) we have that $\psi = \psi^{(\infty)} \in L_2(\mathcal{I})$. With this ψ, the probability of misclassification is at its minimum and equals $1 - \Phi\{\frac{1}{2}\,(\sum_{j\ge1}\,\theta_j^{-1}\,\mu_j^2)^{1/2}\}$.

If $\sum_{j=1}^\infty \theta_j^{-2}\,\mu_j^2 = \infty$ then $\psi^{(\infty)}$ is not in $L_2(\mathcal{I})$, but for each $0 < r < \infty$, $\psi^{(r)}$ defined at (2.7) is in $L_2(\mathcal{I})$. Taking $\psi = \psi^{(r)}$, the probability of misclassification is given by $1 - \Phi\{\frac{1}{2}\,(\sum_{j=1}^r \theta_j^{-1}\,\mu_j^2)^{1/2}\}$, which converges to err_0 as $r \to \infty$. In particular, when $\sum_{j=1}^\infty \theta_j^{-1}\,\mu_j^2 = \infty$, $\mathrm{err}_0 = 0$.

(d). For simplicity, assume that $\pi = \frac{1}{2}$. Under the Gaussian process model in section 2.2, with $n_0 = n_1 = \infty$, we wish to classify a random function X that can have either the distribution of Z defined at (A.1), or that of

$$Z + \mu = \sum_{j=1}^\infty \theta_j^{1/2}\,(Z_j + \theta_j^{-1/2}\,\mu_j)\,\phi_j\,. \tag{A.3}$$

Distinguishing between the hypotheses represented by the models at (A.1) and (A.3) is equivalent to testing the null hypothesis H_0 that the members of a sequence $\zeta_1, \zeta_2, \ldots$ of independent normal random variables all have the $\mathrm{N}(0,1)$ distribution, versus the alternative H_1 that they have respective $\mathrm{N}(\theta_j^{-1/2}\,\mu_j, 1)$ distributions. The optimal test is based on the likelihood ratio, and is equivalent to deciding in favour of H_1 if and only if

$$\lim_{r\to\infty} \left\{ \sum_{j=1}^r \left(\zeta_j - \theta_j^{-1/2}\,\mu_j\right)^2 - \sum_{j=1}^r \zeta_j^2 \right\} < 0\,,$$

or equivalently, if and only if $\nu^2 - 2\,\nu\,Q \le 0$, where ν and Q are as at (2.6) when $\psi = \psi^{(r)}$ is as defined at (2.7), and in particular Q is normal $\mathrm{N}(0, \sigma^2)$. Therefore, applying the likelihood ratio test is equivalent to applying the centroid-based classifier. □

Proof of Theorem 4.1. It is straightforward to show that $\widehat{R} \to \infty$, and so we derive only (4.2). The classifier $\mathcal{C}$, applied to an observation x, asserts that $\mathcal{C}(x\,|\,\psi) = 0$ if and only if $T(x) > 0$, where T is defined at (2.4). The analogue $\mathcal{C}_0$ of $\mathcal{C}$, employing infinite training samples, asserts that $\mathcal{C}_0(x\,|\,\psi) = 0$ if and only if $T^0(x) > 0$, where T^0 is defined at (2.5). Using $\widehat{\psi}^{(r)}$ and $\psi^{(r)}$ to construct $T(x)$ and $T^0(x)$, respectively, and denoting the results by T_r and T_{0r} respectively, we have:

$$T_r(x) \;= \left\langle \bar{X}_1 - \bar{X}_0, \widehat{\psi}^{(r)} \right\rangle \left(\left\langle \bar{X}_1 + \bar{X}_0, \widehat{\psi}^{(r)} \right\rangle - 2\,\left\langle x, \widehat{\psi}^{(r)} \right\rangle \right), \tag{A.4}$$

$$T_{0r}(x) \;= \left\langle \mu_1 - \mu_0, \psi^{(r)} \right\rangle \left(\left\langle \mu_1 + \mu_0, \psi^{(r)} \right\rangle - 2\,\left\langle x, \psi^{(r)} \right\rangle \right), \tag{A.5}$$

and moreover,

$$\left| \left\langle \bar{X}_1 \pm \bar{X}_0, \widehat{\psi}^{(r)} \right\rangle - \left\langle \mu_1 \pm \mu_0, \psi^{(r)} \right\rangle \right| = \left| \langle \mu_1 \pm \mu_0, D_r \rangle + \left\langle D_\pm, \psi^{(r)} \right\rangle + \langle D_\pm, D_r \rangle \right|, \tag{A.6}$$

where $D_r = \widehat{\psi}^{(r)} - \psi^{(r)}$ and $D_\pm = \bar{X}_1 - \mu_1 \pm (\bar{X}_0 - \mu_0)$.

Since, by (4.1)(b), $E(\|X\|^2) < \infty$, then

$$\int_{\mathcal{I}} D_\pm^2 = O_p(n^{-1})\,, \tag{A.7}$$

and we shall prove shortly that

$$\max_{1 \le r \le \widehat{R}} \int_{\mathcal{I}} D_r^2 = o_p(1)\,, \quad n^{-1/2} \max_{1 \le r \le \widehat{R}} \left\|\psi^{(r)}\right\| = o_p(1)\,. \tag{A.8}$$

Results (A.4)–(A.8), and again the property $E(\|X\|^2) < \infty$, imply that for each $C_1 > 0$,

$$\sup_{x\,:\,\|x\| \le C_1} \max_{1 \le r \le \widehat{R}} |T_r(x) - T_{0r}(x)| = o_p(1)\,. \tag{A.9}$$

Let X be a new function from Π_k, independent of the data. It follows from (A.9), and the assumption in (4.1)(d) that $P_k\{\sup_{r \ge 1} |T_{0r}(X)| \le c\} \to 0$ as $c \to 0$, that

$$P\big\{T_r(X) \text{ and } T_{0r}(X) \text{ have the same sign for all } 1 \le r \le \widehat{R}\big\} \to 1\,.$$

Equivalently, the probability that the classifiers $\mathcal{C}(X \,|\, \widehat{\psi}^{(r)})$ and $\mathcal{C}_0(X \,|\, \psi^{(r)})$ give the same results for each r in the range $1 \le r \le \widehat{R}$ converges to 1.

A similar argument can be used to prove that, if we define $\mathcal{C}^{(i)}(X \,|\, \widehat{\psi}^{(r,i)})$ to be the version of $\mathcal{C}(X \,|\, \widehat{\psi}^{(r)})$ defined not in terms of T_r but its analogue $T_r^{(i)}$, based on the training data from which X_i (in either of the training samples) has been excluded, then the probability that $\mathcal{C}^{(i)}(X_i \,|\, \widehat{\psi}^{(r,i)})$ and $\mathcal{C}_0(X_i \,|\, \psi^{(r)})$ give the same results for each r in the range $1 \le r \le \widehat{R}$, uniformly in data functions X_i from the training data, converges to 1. (In particular, (4.1)(b) can be used to prove that removing a single observation from the training data alters the value of $|T_r(x) - T_{0r}(x)|$ by $o_p(1)$, uniformly over all indices of the removed observation, as well as uniformly over x for which $\|x\| \le C_1$, and uniformly over $1 \le r \le \widehat{R}$.)

Finally we derive (A.8). Observe that

$$D_r = \sum_{j=1}^{r} \big(\frac{\widehat{\mu}_{1j}}{\widehat{\theta}_j}\,\widehat{\phi}_j - \theta_j^{-1}\,\mu_{1j}\,\phi_j\big) = \sum_{j=1}^{r} \big(\widehat{\theta}_j^{-1}\,\widehat{\mu}_{1j} - \theta_j^{-1}\,\mu_{1j}\big)\,\widehat{\phi}_j + \sum_{j=1}^{r} \theta_j^{-1}\,\mu_{1j}\,\big(\widehat{\phi}_j - \phi_j\big)\,,$$

$$\tfrac{1}{2} \int_{\mathcal{I}} D_r^2 \le \sum_{j=1}^{r} \Big(\widehat{\theta}_j^{-1}\widehat{\mu}_{1j} - \theta_j^{-1}\mu_{1j}\Big)^2 + \|\mu_1\|^2 \sum_{j=1}^{r} \theta_j^{-2} \|\widehat{\phi}_j - \phi_j\|^2\,,$$

$$\Big|\widehat{\theta}_j^{-1}\widehat{\mu}_{1j} - \theta_j^{-1}\mu_{1j}\Big| \le \widehat{\theta}_j^{-1}|\widehat{\mu}_{1j} - \mu_{1j}| + \theta_j^{-1}\,\widehat{\theta}_j^{-1}|\mu_{1j}|\;|\widehat{\theta}_j - \theta_j|$$

$$\leq \widehat{\theta}_j^{-1} \left| \int_{\mathcal{I}} (\bar{X}_1 - \mu_1)\, \widehat{\phi}_j + \int_{\mathcal{I}} \mu_1\, (\widehat{\phi}_j - \phi_j) \right| + \theta_j^{-1}\, \widehat{\theta}_j^{-1} |\mu_{1j}|\, |\widehat{\theta}_j - \theta_j|\,,$$

and hence,

$$A_1 \sum_{j=1}^{r} \left(\widehat{\theta}_j^{-1} \widehat{\mu}_{1j} - \theta_j^{-1} \mu_{1j} \right)^2 \leq \|\bar{X}_1 - \mu_1\|^2 \sum_{j=1}^{r} \widehat{\theta}_j^{-2} + \|\mu_1\|^2 \sum_{j=1}^{r} \widehat{\theta}_j^{-2}\, \|\widehat{\phi}_j - \phi_j\|^2 + \|\mu_1\|^2 \sum_{j=1}^{r} (\theta_j\, \widehat{\theta}_j)^{-2} (\widehat{\theta}_j - \theta_j)^2\,,$$

where, here and below, A_1, A_2 and A_3 will denote absolute constants. Therefore,

$$A_2 \int_{\mathcal{I}} D_r^2 \leq \|\bar{X}_1 - \mu_1\|^2 \sum_{j=1}^{r} \widehat{\theta}_j^{-2} + \|\mu_1\|^2 \sum_{j=1}^{r} \frac{\|\widehat{\phi}_j - \phi_j\|^2}{\min(\theta_j^2, \widehat{\theta}_j^2)} + \|\mu_1\|^2 \sum_{j=1}^{r} \frac{(\widehat{\theta}_j - \theta_j)^2}{(\theta_j\, \widehat{\theta}_j)^2}\,. \tag{A.10}$$

Hall and Hosseini-Nasab (2006) noted that

$$|\widehat{\theta}_j - \theta_j| \leq \widehat{\Delta}\,, \quad \|\widehat{\phi}_j - \phi_j\| \leq 8^{1/2}\, \widehat{\Delta}\, \delta_j^{-1}\,,$$

where $\widehat{\Delta}^2 = \int_{\mathcal{I}} \int_{\mathcal{I}} (\widehat{K} - K)^2$ and $\delta_j = \min_{k \leq j} (\theta_k - \theta_{k+1})$. Therefore, by (A.10).

$$A_3 \int_{\mathcal{I}} D_r^2 \leq \left(\|\bar{X}_1 - \mu_1\|^2 + \|\mu_1\|^2\, \widehat{\Delta}^2 \right) \sum_{j=1}^{r} (\theta_j\, \widehat{\theta}_j)^{-2} + \|\mu_1\|^2\, \widehat{\Delta}^2 \sum_{j=1}^{r} \left\{ \delta_j\, \min(\theta_j, \widehat{\theta}_j) \right\}^{-2}\,. \tag{A.11}$$

Now,

$$\widehat{R} \leq \sum_{j=1}^{\infty} I\big(\widehat{\theta}_j - \widehat{\theta}_{j+1} \geq \eta_n\big) \leq \sum_{j=1}^{\infty} \eta_n^{-1} \left(\widehat{\theta}_j - \widehat{\theta}_{j+1}\right) I\big(\widehat{\theta}_j - \widehat{\theta}_{j+1} \geq \eta_n\big) \leq \eta_n^{-1}\, \widehat{\theta}_1\,,$$

and if $j \leq \widehat{R}$ and $\widehat{\Delta} \leq \frac{1}{2}\, \eta_n$ then $\widehat{\theta}_j \geq \eta_n$ and $\theta_j \geq \widehat{\theta}_j - \Delta \geq \frac{1}{2}\, \eta_n$. Therefore, if $\widehat{\Delta} \leq \frac{1}{2}\, \eta_n$ then

$$\sum_{j=1}^{\widehat{R}} \left(\theta_j\, \widehat{\theta}_j\right)^{-2} \leq \eta_n^{-1}\, \widehat{\theta}_1 \cdot \left(\eta_n \cdot \tfrac{1}{2}\, \eta_n\right)^{-2} = 4\, \eta_n^{-5}\, \widehat{\theta}_1\,, \tag{A.12}$$

$$\sum_{j=1}^{\widehat{R}} \left\{\theta_j\, \min\left(\theta_j, \widehat{\theta}_j\right)\right\}^{-2} \leq \eta_n^{-1}\, \widehat{\theta}_1 \cdot \left(\tfrac{1}{2}\, \eta_n^2\right)^{-2} = 4\, \eta_n^{-5}\, \widehat{\theta}_1\,. \tag{A.13}$$

Furthermore, by assumption on η_n, $n^{-1/5} = o(\eta_n)$; and since $\sup_{t \in \mathcal{T}} E_k\{|X(t)|\}^4 < \infty$ for $k = 1, 2$, by (4.1)(b); then $\widehat{\Delta} = O_p(n^{-1/2})$ and $\|\bar{X}_1 - \mu_1\| = O_p(n^{-1/2})$. It follows

that $P(\widehat{\Delta} \le \frac{1}{2}\,\eta_n) \to 1$. These properties and (A.11)–(A.13) imply the first part of (A.8). The second part of (A.8) follows more simply:

$$n^{-1} \max_{1\le r\le \widehat{R}} \bigl\|\psi^{(r)}\bigr\|^2 = n^{-1} \max_{1\le r\le \widehat{R}} \sum_{j=1}^{r} \theta_j^{-2}\,\mu_{1j}^2 \le n^{-1}\,\bigl(\tfrac{1}{2}\,\eta_n\bigr)^{-2}\,\|\mu_1\|^2 \to 0\,,$$

since $n^{1/5}\,\eta_n \to \infty$. (The inequality holds when $\widehat{\Delta} \le \frac{1}{2}\,\eta_n$.) □

Proof of Theorem 4.2. Define $\psi_r = \sum_{j=1}^{r} \theta_j^{-1}\,\mu_j\,\phi_j$. If X is from Π_0 then $T^0(X) = \bigl(\langle X,\psi_r\rangle - \langle \mu,\psi_r\rangle\bigr)^2 - \langle X,\psi_r\rangle^2 = \langle \mu,\psi_r\rangle^2 - 2\,\alpha_0\,\langle \mu,\psi_r\rangle\,R_0$, where the random variable R_0 satisfies $E(R_0) = 0$ and $E(R_0^2) = 1$, and $\alpha_0^2 = \sum_{j=1}^{\infty} \theta_{0j}\,\bigl\{\int \psi_r\,\phi_{0j}\bigr\}^2 \ge 0$. Similarly, if X is from Π_1 then $T^0(X) = \bigl(\langle X,\psi_r\rangle - \langle \mu,\psi_r\rangle\bigr)^2 - \langle X,\psi_r\rangle^2 = -\langle \mu,\psi_r\rangle^2 - 2\,\alpha_1\,\langle \mu,\psi_r\rangle\,R_1$, where $0 \le \alpha_1^2 = \sum_{j=1}^{\infty} \theta_{1j}\,\{\int \psi_r\,\phi_{1j}\}^2$ and the random variable R_1 satisfies $E(R_1) = 0$ and $E(R_1^2) = 1$. Therefore,

$$\begin{aligned}
\mathrm{err}(\psi_r) &= \pi\,P_0\bigl(T^0 \le 0\bigr) + (1-\pi)\,P_1\bigl(T^0 > 0\bigr) \\
&= \pi\,P\bigl\{R_0 \ge \langle \mu,\psi_r\rangle/(2\alpha_0)\bigr\} + (1-\pi)\,P\bigl\{\,-R_1 > \langle \mu,\psi_r\rangle/(2\,\alpha_1)\bigr\} \\
&\le 4\{\pi\alpha_0^2 + (1-\pi)\alpha_1^2\}/\langle \mu,\psi_r\rangle^2,
\end{aligned}$$

where we used Markov's inequality. In the particular case where (4.4)(iv) holds we deduce that $\mathrm{err}(\psi_r) \to 0$ as $r \to \infty$. □

A.2 Details of PLS algorithm

The iterative PLS procedures constructs successive approximations (for $\ell = 1,\ldots,r$ with $r > 0$ some finite number) to $X(t)$ and Y in the form

$$X_i(t) = \bar{X}(t) + \sum_{1\le j\le \ell} T_{ij} P_j(t) + E_{i\ell}(t) \text{ and } Y_i = \bar{Y} + \sum_{1\le j\le \ell} T_{ij}\,c_j + f_{i\ell}\,, \qquad \text{(A.14)}$$

$$E_{i\ell}(t) = X_i(t) - \bar{X}(t) - \sum_{1\le j\le \ell} T_{ij} P_j(t) \text{ and } f_{i\ell} = Y_i - \bar{Y} - \sum_{1\le j\le \ell} T_{ij}\,c_j\,, \qquad \text{(A.15)}$$

where $T_{ij} = \int_{\mathcal{I}} W_j(t)\{X_i(t) - \bar{X}(t)\}\,dt$ for some function $W_j(t)$, and with $E_{i\ell}$ and $f_{i\ell}$ the residuals. At step r, PLS provides a linear approximation to Y_i through $Y_i = \widehat{\alpha}_r + \int_{\mathcal{I}} \widehat{\beta}_r(t) X_i(t) + f_{iq}$, with $\widehat{\alpha}_r = \bar{Y} - \sum_{j=1}^{r} \int_{\mathcal{I}} c_j W_j(t) \bar{X}(t)$ and $\widehat{\beta}_r(t) = \sum_{j=1}^{r} c_j W_j(t)$. W_j, P_j and c_j are constructed iteratively as follows. First, choose the function W_1 that satisfies $\|W_1\| = 1$ and maximises the covariance of Y_i and $T_{i1} = \int_{\mathcal{I}} W_1\,X_i$. Write (A.14) for $\ell = 1$, where $P_1(t)$ (respectively c_1) is the LS estimator of the

slope of the linear regression of X_i (respectively Y_i), on T_{i1}. Then, more generally, given an orthonormal sequence $W_1, \ldots, W_{j-1}$, choose W_j, subject to $\|W_j\| = 1$ and $\int_{\mathcal{I}} W_j\, W_k = 0$ for $1 \le k \le j-1$, to maximise the covariance of $f_{i,j-1}$ and $\int_{\mathcal{I}} W_j\, E_{i,j-1} = \int_{\mathcal{I}} W_j\, X_i = T_{ij}$. Then $P_j(t)$ (respectively c_j) is the LS estimator of the slope of the intercept-free linear regression of E_{ij} (respectively f_{ij}), on T_{ij}.

Acknowledgements

Research supported by grants and fellowships from the Australian Research Council.

References

Araki, Y., Konishi, S., Kawano, S. and Matsui, H. (2009). Functional logistic discrimination via regularized basis expansions. *Commun. Statistics–Theory and Methods* **38**, 2944–2957.

Biau, G., Bunea, F. and Wegkamp, M.H. (2005). Functional classification in Hilbert spaces. *IEEE Trans. Inform. Theory* **51**, 2163–2172.

Berlinet, A., Biau, G. and Rouvière, L. (2008) Functional classification with wavelets. *Annales de l'Institut de statistique de l'université de Paris*, **52**, 61–80.

Chamroukhi, F., Same, A., Govaert, G. and Aknin, P. (2010). A hidden process regression model for functional data description. Application to curve discrimination. *Neurocomputing* **73**, 1210–1221.

Cook, R.D. (2007). Fisher lecture: Dimension reduction in regression. *Statist. Sci.* **22**, 1–26.

Cox, D.R. (1968). Notes on some aspects of regression analysis. *J. R. Statist. Soc.* Ser. A **131**, 265–279.

Cuevas, A., Febrero, M. and Fraiman, R. (2007). Robust estimation and classification for functional data via projection-based depth notions. *Comput. Statist.* **22**, 481–496.

Delaigle, A. and Hall, P. (2010). Defining probability density for a distribution of random functions. *Ann. Statist.* **38**, 1171–1193.

Epifanio, I. (2008). Shape descriptors for classification of functional data. *Technometrics* **50**, 284–294.

Ferraty, F. and Vieu, P. (2003). Curves discrimination: a nonparametric functional approach. *Comput. Statist. Data Anal.* **4**, 161–173.

Fromont, M. and Tuleau, C. (2006). Functional classification with margin conditions. In *Learning Theory–Proceedings of the 19th Annual Conference on Learning Theory, Pittsburgh, PA, USA, June 22-25, 2006*, Eds J.G. Carbonell and J. Siekmann. Springer, New York.

Glendinning, R.H. and Herbert, R.A. (2003). Shape classification using smooth principal components. *Patt. Recognition Lett.* **24**, 2021–2030.

Hall, P. and Hosseini-Nasab, M. (2006). On properties of functional principal components analysis. *J. R. Stat. Soc.* Ser. B **68**, 109–126.

Hall, P. and Marron, S.J. (1991). Local minima in cross-validation functions. *J. R. Stat. Soc.* Ser. B **53**, 245–252.

Hall, P., Poskitt, D. and Presnell, B. (2001). A functional data-analytic approach to signal discrimination. *Technometrics* 43, 1–9.

Hastie, T. Tibshirani, R. and Friedman, J. (2009). The Elements of. Statistical Learning: Data Mining, Inference, and Prediction. 2nd Ed. *Springer-Verlag*.

Huang, D.-S. and Zheng, C.-H. (2006). Independent component analysis-based penalized discriminant method for tumor classification using gene expression data. *Bioinformatics* **22**, 1855–1862.

James, G. and Hastie, T. (2001). Functional linear discriminant analysis for irregularly sampled curves. *J. Roy. Statist. Soc. Ser. B* **63**, 533–550.

James, G. and Sugar, C. (2003). Clustering for Sparsely Sampled Functional Data. *J. Amer. Statist. Assoc.* **98**, 397–408.

Kalivas, J. H. (1997). Two data sets of near infrared spectra. *Chemometrics and Intelligent Laboratory Systems*, **37**, 255–259.

Leng, X.Y. and Müller, H.-G. (2006). Classification using functional data analysis for temporal gene expression data. *Bioinformatics* **22**, 68–76.

López-Pintado, S. and Romo, J. (2006). Depth-based classification for functional data. In *DIMACS Series in Discrete Mathematics and Theoretical Computer Science* **72**, 103–120.

Müller, H.-G. and Stadmüller, U. (2005). Generalized functional linear models. *Ann. Statist.* **33**, 774–805.

Preda, C., and Saporta, G. (2005). PLS regression on a stochastic process *Comput. Statist. Data Anal.* **48**, 149–158.

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

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

Rossi, F. and Villa, N. (2006). Support vector machine for functional data classification. *Neurocomputing* **69**, 730–742.

Shin, H. (2008). An extension of Fisher's discriminant analysis for stochastic processes. *J. Multivar. Anal.* **99**, 1191–1216.

Song, J.J., Deng, W., Lee, H.-J. and Kwon, D. (2008). Optimal classification for time-course gene expression data using functional data analysis. *Comp. Biology and Chemistry* **32**, 426–432.

Tian, S.T. and James, G. (2010). Interpretable dimensionality reduction for classification with functional data. Manuscript.

Vilar, J.A. and Pertega, S. (2004). Discriminant and cluster analysis for Gaussian

stationary processes: Local linear fitting approach. *J. Nonparam. Statist.* **16**, 443–462.

Wang, X.H., Ray, S. and Mallick, B.K. (2007). Bayesian curve classification using wavelets. *J. Amer. Statist. Assoc.* **102**, 962–973.