Biometrika (2012), **xx**, x, *pp*. 1–27

Printed in Great Britain

Componentwise classification and clustering of functional data

BY A. DELAIGLE, P. HALL

Department of Mathematics and Statistics, University of Melbourne, Parkville, Victoria 3010, Australia

A.Delaigle@ms.unimelb.edu.au halpstat@ms.unimelb.edu.au

AND N. BATHIA

Jump Trading, 600 West Chicago Avenue, Chicago, Illinois 60654, United States of America

nbathia@jumptrading.com

SUMMARY

The infinite dimension of functional data can challenge conventional methods for classification and clustering. A variety of techniques have been introduced to address this problem, particularly in the case of prediction, but the structural models that they involve can be too inaccurate, or too abstract, or too difficult to interpret, for practitioners. In this paper we develop approaches to adaptively choose components, enabling classification and clustering to be reduced to finite-dimensional problems. We explore and discuss properties of these methodologies. Our techniques involve methods for estimating classifier error rate and cluster tightness, and for choosing both the number of components, and their locations, to optimize these quantities. A major attraction of this approach is that it allows identification of parts of the function domain that convey

important information for classification and clustering. It also permits us to determine regions that are relevant to one of these analyses but not the other.

Some key words: Bandwidth; classification error rate; kernel method; statistical smoothing; tightness of clusters.

1. INTRODUCTION

Problems of classification for functional data are vexed by difficulties caused by the intrinsic infinite dimension of functions. For simpler methods, such as linear or quadratic discriminant analysis, the difficulty is caused by the necessity to estimate and invert covariance operators. For nonparametric classifiers, which are attractive because of the awkwardness of modelling random functions, parameter-free approaches to infinite-dimensional problems can produce noisy and slowly convergent techniques since they attempt to respond to too many different sources of information. Similar difficulties can arise in problems of clustering, since algorithms can become trapped in local minima if they are calculated using too many dimensions. These difficulties motivate methods for dimension reduction.

In the functional data context, classifiers are constructed from independent data pairs distributed as (X, Y), where X is a random function defined on a compact interval $\mathcal{I}$ and Y is a class label taking the values 0 to $K-1$, with K denoting the number of classes. Clusterers are constructed from data on X alone. In the literature, dimension reduction is often performed by projecting functional curves onto a finite number, p, of functions $\psi_1, \ldots, \psi_p$. Then, standard multivariate classifiers or clusterers are applied to the p-variate projections $(\int_{\mathcal{I}} X\psi_1, \ldots, \int_{\mathcal{I}} X\psi_p)^{\mathrm{T}}$. In this context, the functions ψ_j are often taken to be the first p elements of a basis, where the functional basis is either arbitrary, for example a spline basis, or chosen from the data, for example the principal component basis. See for example Hall et al. (2001), Glendinning and Herbert (2003), Huang and Zheng (2006) and Song et al. (2008).

These approaches to dimension reduction are hindered by the fact that, in general, there is no particular reason why these functions ψ_j would be particularly suitable for classification. In particular, since the ψ_js are not chosen to optimize classification performance then, by projecting the data on a low number, p, of them, we may lose a great deal of information relevant to classification. This is true even for the principal component basis which is constructed from the data, but only in a way that guarantees that the variability of the X functions is well represented by projecting on the first few basis functions. In particular, this basis is chosen regardless of the class labels of the data. To overcome this issue, in an unpublished manuscript, Tian and James suggested an iterative approach combining prediction-based variable selection and control of model complexity. Their solution is one of the first attempts to choose basis functions in a way that takes into account classification error. Their method is interesting, but it is a little complex and not fully data-driven; for example, p is not chosen from the data. Moreover, parts of their algorithm are based on prediction rather than classification.

In this paper, we develop a simple technique that can be employed for virtually any classification or clustering method, and which provides useful practical insight and interpretability. Our approach is very simple; it consists of determining a relatively small number of points $t_1, \ldots, t_p \in \mathcal{I}$ that are chosen so that $X(t_1), \ldots, X(t_p)$ convey particular information for classification or clustering, respectively. Those points then become of special interest to the practitioner, who might wish to consider aspects of the data generating process that influence the function X at $t_1, \ldots, t_p$. Even the fact that the points might be different in different problems, for example problems of classification and clustering, is of interest. It is well known that, while some data features are particularly helpful for characterising the type or nature of the data, they can be unhelpful for prediction. As we shall see, this is also true in the context of classification, where landmark points for classification can often be located at inflection points of the curves,

rather than, for example, at turning points. We shall also see that, when classifying a sample of random functions $X_i(t)$ into groups, the important points t depend on the nature of the groups. In other words, the points that are important for a grouping based on a variable Y are not necessarily important for a grouping based on another variable Z. Therefore, being able to identify the features that are important for the problem of interest, in a readily interpretable way, can be advantageous. A point selection approach was introduced by Ferraty et al. (2010) in the prediction context. The methodology and results in that paper are quite different from those here.

We suggest empirical approaches to choosing both p and the points $t_1, \ldots, t_p \in \mathcal{I}$. Specifically, we construct classifiers and clustering methods when data on X are restricted to $t_1, \ldots, t_r$, for successive values of r, and for each r we estimate the performance of the methodology, stopping when the amount of error incurred by the classifier, or absence of tightness of clusters, drops below a threshold. The value of r at which this occurs represents our empirical approximation, $\hat{p}$, to p. In the context of clustering it is sometimes possible, depending on the clustering method, to use the random functions directly, but even here there is a great deal of insight to be gained by determining a small number of components that have substantial leverage for constructing tight clusters.

Methods for classifying functional data have been discussed by a number of authors; see §2·4. For additional references on classification and clustering, see James and Sugar (2003), Vilar and Pertega (2004), Biau et al. (2005), Fromont and Tuleau (2006), Leng and Müller (2006), López-Pintado and Romo (2006), Rossi and Villa (2006), Cuevas et al. (2007), Wang et al. (2007), Berlinet et al. (2008), Epifanio (2008), Peng and Müller (2008), Araki et al. (2009) and Chamroukhi et al. (2010). For a general introduction to functional data analysis, see Ramsay and Silverman (2005).

2. MODEL AND METHODOLOGY FOR CLASSIFICATION

2·1. *Model*

In problems of classification we assume that independent and identically distributed data pairs $(X_1, I_1), \ldots, (X_n, I_n)$ are observed, where each I_j is a class label taking values in the set $\{0, \ldots, K-1\}$, with K denoting the number of classes. The goal of classification methods is to assign, to one of the K classes, a value x of X that is missing its class label. For brevity here we treat only the case of two populations, numbered 0 and 1 respectively, noting that other settings are similar.

To overcome the difficulties encountered by classifiers applied to infinite dimensional objects, rather than using the whole functions X directly, we identify a small number of points $t_1, \ldots, t_p \in \mathcal{I}$ that have important leverage for classification, and apply a conventional finite dimensional classifier based on the p-dimensional vectors $\bigl(X(t_1), \ldots, X(t_p)\bigr)^{\mathrm{T}}$. We select p and $t_1, \ldots, t_p$ adaptively, in a way that depends on both the data and the particular classifier employed, as described below in §§2·2 and 2·3. Theoretical properties of the method will be studied in the appendix, and proofs are available in the Supplementary Material.

2·2. *Choosing the points in a given dimension*

We start by describing, for a general classifier, the procedure that selects the most important r-dimensional point when r is fixed. Next, in §2·3, we show how to choose the dimension p. Given the dataset $\mathcal{D} = \{(X_1, I_1), \ldots, (X_n, I_n)\}$, let $J(x, \mathcal{D} \,|\, t_{(r)})$ denote the population index, either 0 or 1, to which our classifier assigns the individual with explanatory variable x after dimension has been reduced to $t_{(r)} = (t_1, \ldots, t_r)^{\mathrm{T}}$. In particular, the classifier that produces the result $J(x, \mathcal{D} \,|\, t_{(r)})$ is based on the data vectors $\bigl(X_i(t_1), \ldots, X_i(t_r), I_i\bigr)^{\mathrm{T}}$ for $i = 1, \ldots, n$. The

cross-validation estimator of error rate is

$$\hat{\mathrm{err}}_r(t_{(r)}) = \frac{1}{n}\sum_{i=1}^{n} I\{J(X_i, \mathcal{D}_{-i} \,|\, t_{(r)}) \neq I_i\}\,, \tag{1}$$

where $\mathcal{D}_{-i} = \mathcal{D} \setminus \{(X_i, I_i)\}$ denotes the dataset with the ith data pair removed. We set the most important r-dimensional point $t_{(r)} = \hat{t}_{(r)}$ to be the one that minimizes $\hat{\mathrm{err}}_r(t_{(r)})$.

2·3. *Choosing p*

To describe how to choose p, let $\mathcal{I}_r$ denote the set of all r-vectors $t_{(r)} = (t_1, \ldots, t_r)^{\mathrm{T}}$ with $t_1 < \ldots < t_r$ and $t_1, \ldots, t_r \in \mathcal{I}$, and define

$$T_r = \inf_{t_{(r)} \in \mathcal{I}_r} \hat{\mathrm{err}}_r(t_{(r)}) = \hat{\mathrm{err}}_r\big(\hat{t}_{(r)}\big)\,.$$

We suggest increasing r until the incremental change in the minimum error T_r for r dimensions is a small fraction of the minimum error for the previous value of r, or of T_1. These two approaches can be formalized by respectively defining $\hat{p}$ by $\hat{p} = \inf\{r : T_r - T_{r+1} \leq \rho\, T_{r-1}\}$, the latter being equivalent to

$$\hat{p} = \inf\{r : (1-\rho)\, T_r \leq T_{r+1}\}\,, \tag{2}$$

or by defining $\hat{p}$ by

$$\hat{p} = \inf\{r : T_r - T_{r+1} \leq \rho\, T_1\}. \tag{3}$$

Here, ρ denotes a pre-determined small proportion, for example $\rho = 0{\cdot}05$, $0{\cdot}1$ or $0{\cdot}2$. In our numerical work we used the approach based on (2) with $\rho = 0{\cdot}1$. This gave good results in all cases, but the value of ρ is not very important and we obtained similar results with other values of ρ ranging from 0 to $0{\cdot}2$; see the Supplementary Material for an illustration on some simulated examples. It is inappropriate here to try to drive the error down to zero. Even the Bayes classifier, in finite-dimensional problems where the supports of the distributions representing the two populations have nondegenerate intersection, has strictly positive classification error.

Therefore, in classification problems it does not make sense to continue to increase r until the error falls to a small proportion ρ.

Let $n_k = \sum_{i=1}^n I(I_i = k)$ denote the number of observations drawn from population k. The expected error rate of empirical classifiers is

$$\mathrm{err}_r(t_{(r)}) = \frac{n_0}{n}\,\mathrm{pr}_0\big\{J(X, \mathcal{D} \,|\, t_{(r)}) = 1\big\} + \frac{n_1}{n}\,\mathrm{pr}_1\big\{J(X, \mathcal{D} \,|\, t_{(r)}) = 0\big\}\,, \tag{4}$$

where X is random function that is independent of the dataset $\mathcal{D}$, and pr_k denotes probability measure under the hypothesis that X is from population k. The quantity $\mathrm{err}_r(t_{(r)})$ is estimated by $\hat{\mathrm{err}}_r(t_{(r)})$, at (1). The expected error rate of the classifier when $t_{(r)}$ is replaced by $\hat{t}_{(\hat{p})}$ is

$$\mathrm{err}^{\mathrm{emp}} = \frac{n_0}{n}\,\mathrm{pr}_0\big\{J\big(X, \mathcal{D} \,\big|\, \hat{t}_{(\hat{p})}\big) = 1\big\} + \frac{n_1}{n}\,\mathrm{pr}_1\big\{J\big(X, \mathcal{D} \,\big|\, \hat{t}_{(\hat{p})}\big) = 0\big\}\,. \tag{5}$$

2·4. *Details for specific classifiers*

Next we describe the application of our methodology to five popular classifiers: Fisher's linear and quadratic discriminants (James and Hastie, 2001; Preda et al., 2007; Shin, 2008), a nonparametric Bayes rule, a nonparametric regression-based classifier (Ferraty and Vieu, 2003, 2006) and a classifier based on logistic regression. Let x denote a new function, without a class label, which we wish to classify, and put $\mathrm{x}(t_{(r)}) = \big(x(t_1), \ldots, x(t_r)\big)^{\mathrm{T}}$. Let π_0 and π_1 denote the prior probabilities of the two populations. Often in practice, π_k is taken to be equal to either n_k/n, if we believe that the sample proportions reflect the population ones, or $1/2$ otherwise.

To define Fisher's linear discriminant method combined with our point selection approach, put $\mathrm{X}_i(t_{(r)}) = \big(X_i(t_1), \ldots, X_i(t_r)\big)^{\mathrm{T}}$, let $\hat{\Sigma}(t_{(r)})$ denote the empirical $r \times r$ covariance matrix computed from the data vectors $\mathrm{X}_i(t_{(r)})$ for $i = 1, \ldots, n$, and write $\bar{\mathrm{X}}_0(t_{(r)})$ and $\bar{\mathrm{X}}_1(t_{(r)})$ for the average of $\mathrm{X}_i(t_{(r)})$ over i such that $I_i = 0$ and $I_i = 1$, respectively. Fisher's linear discriminant, for the particular choice $t_{(r)}$ of components, assigns x to population 0 if

$$\big\{\mathrm{x}(t_{(r)}) - \bar{\mathrm{X}}_0(t_{(r)})\big\}^{\mathrm{T}}\,\hat{\Sigma}(t_{(r)})^{-1}\,\big\{\mathrm{x}(t_{(r)}) - \bar{\mathrm{X}}_0(t_{(r)})\big\}$$

$$\leq \left\{\mathrm{x}(t_{(r)}) - \bar{\mathrm{X}}_1(t_{(r)})\right\}^{\mathrm{T}} \hat{\Sigma}(t_{(r)})^{-1} \left\{\mathrm{x}(t_{(r)}) - \bar{\mathrm{X}}_1(t_{(r)})\right\} + C_{01}\,, \tag{6}$$

where $C_{01} = \log(\pi_0/\pi_1)$, or equivalently if

$$\begin{aligned} 2\left\{\bar{\mathrm{X}}_0(t_{(r)}) - \bar{\mathrm{X}}_1(t_{(r)})\right\}^{\mathrm{T}} \hat{\Sigma}(t_{(r)})^{-1}\, \mathrm{x}(t_{(r)}) \geq & \bar{\mathrm{X}}_0(t_{(r)})^{\mathrm{T}}\, \hat{\Sigma}(t_{(r)})^{-1}\, \bar{\mathrm{X}}_0(t_{(r)}) \\ & - \bar{\mathrm{X}}_1(t_{(r)})^{\mathrm{T}}\, \hat{\Sigma}(t_{(r)})^{-1}\, \bar{\mathrm{X}}_1(t_{(r)}) - C_{01}\,, \end{aligned}$$

and to population 1 otherwise.

Fisher's quadratic discriminant method is almost identical to the linear discriminant method. For the particular choice $t_{(r)}$ of components, it assigns x to population 0 if (6) is satisfied, except that $\hat{\Sigma}(t_{(r)})$ on the left- and right-hand sides of the inequality is replaced by its variants $\hat{\Sigma}_0(t_{(r)})$ and $\hat{\Sigma}_1(t_{(r)})$ computed solely from the data vectors $\mathrm{X}_i(t_{(r)})$ drawn from populations 0 and 1, respectively. In practice, for the linear and quadratic discriminant classifiers, the error rate can be estimated directly by $n^{-1} \sum_{i=1}^{n} I\{J(X_i, \mathcal{D} \,|\, t_{(r)}) \neq I_i\}$ instead of by the leave-one-out approach at (1).

The third classifier, a nonparametric version of Bayes rule, can be implemented in our context as follows. For $k = 0, 1$, let $f_k(x \,|\, t_{(r)})$ denote the density of $\big(X(t_1), \ldots, X(t_r)\big)^{\mathrm{T}}$ evaluated at $\big(x(t_1), \ldots, x(t_r)\big)^{\mathrm{T}}$, given that X is drawn from population k. For $k = 0, 1$ and $j = 1, \ldots, r$, let $h_{k,j} > 0$ be smoothing parameters called bandwidths, and let K be a smooth, symmetric probability density called the kernel. A multivariate kernel density estimator of $f_k(x \,|\, t_{(r)})$ can be defined by

$$\hat{f}_k\big(x \,|\, t_{(r)}\big) = \frac{c_r}{n_k \prod_{j=1}^{r} h_{k,j}} \sum_{i=1}^{n} I(I_i = k)\, K\Big[\Big\{ \sum_{j=1}^{r} |x(t_j) - X_i(t_j)|^2 / h_{k,j}^2 \Big\}^{1/2}\Big], \tag{7}$$

where $c_r^{-1} = \int K\big\{ \big(\sum_{j=1}^{r} u_j^2 \big)^{1/2} \big\}\, du_1 \cdots du_r$. See for example Wand and Jones (1995). The nonparametric Bayes rule assigns x to population 0 if

$$\pi_0\, \hat{f}_0(x \,|\, t_{(r)}) > \pi_1\, \hat{f}_1(x \,|\, t_{(r)}), \tag{8}$$

and to population 1 otherwise. Choice of the bandwidths will be discussed in §2·5.

The fourth classifier for which we discuss our point selection approach is based on a nonparametric estimator of the regression function $g(x \,|\, t_{(r)}) = E\{I_i \mid \mathrm{X}_i(t_{(r)}) = \mathrm{x}(t_{(r)})\}$. Let K be a kernel and, for $j = 1, \ldots, r$, let $h_j > 0$ be a bandwidth. A multivariate kernel regression estimator of $g(x \,|\, t_{(r)})$ can be defined by

$$\hat{g}\big(x \,|\, t_{(r)}\big) = \frac{\sum_{i=1}^n I_i \, K\Big[\Big\{\sum_{j=1}^r |x(t_j) - X_i(t_j)|^2 / h_j^2\Big\}^{1/2}\Big]}{\sum_{i=1}^n K\Big(\Big\{\sum_{j=1}^r |x(t_j) - X_i(t_j)|^2 / h_j^2\Big\}^{1/2}\Big]}; \tag{9}$$

see Wand and Jones (1995). Motivated by the fact that $g(x \,|\, t_{(r)}) = \mathrm{pr}\{I_i = 1 \mid \mathrm{X}_i(t_{(r)}) = \mathrm{x}(t_{(r)})\}$, the classifier based on $\hat{g}$ assigns x to population 0 if $\hat{g}(x \,|\, t_{(r)}) < 0.5$, and to population 1 otherwise.

Finally, the fifth classifier is based on a parametric estimator of the logistic regression model $g(x \,|\, t_{(r)}) = E\{I_i \mid \mathrm{X}_i(t_{(r)}) = \mathrm{x}(t_{(r)})\} = \exp\{\beta_0 + \mathrm{x}(t_{(r)})^{\mathrm{T}}\beta\}/[1 + \exp\{\beta_0 + \mathrm{x}(t_{(r)})^{\mathrm{T}}\beta\}]$, where $\beta_0 \in \mathbb{R}$ and $\beta \in \mathbb{R}^r$ are unknown parameters. The regression curve $g(x \,|\, t_{(r)})$ is estimated by $\hat{g}(x \,|\, t_{(r)})$, obtained by replacing β_0 and β by their least-squares estimators $\hat{\beta}_0$ and $\hat{\beta}$. The classifier assigns x to population 0 if $\hat{g}(x \,|\, t_{(r)}) < 0.5$ and to population 1 otherwise. In practice, for this classifier too, error rate can be estimated directly by $n^{-1} \sum_{i=1}^n I\{J(X_i, \mathcal{D} \,|\, t_{(r)}) \neq I_i\}$ instead of by the leave-one-out approach at (1).

Since $\hat{\mathrm{err}}_r$ can take at most $n + 1$ different values, its minimum is not always unique. In the Supplementary Material, we describe, for each classifier, procedures that can be used to break ties.

2·5. *Bandwidth choice*

When calculating the nonparametric regression estimator at (9), we define $h_j = \hat{\sigma}_j h$ where $\hat{\sigma}_j^2$ is the empirical variance of the $X_i(t_j)$s calculated from the entire training sample. As in Ferraty and Vieu (2006), we choose h by a nearest neighbour method. More precisely, we take $h = (d_k +$

$d_{k+1})/2$ where d_k^2 is the kth order statistic of $\sum_{j=1}^r |x(t_j) - X_1(t_j)|^2/\hat{\sigma}_j^2, \ldots, \sum_{j=1}^r |x(t_j) - X_n(t_j)|^2/\hat{\sigma}_j^2$, and $k = k(r, t_{(r)})$ is chosen by minimising $\hat{\text{err}}_r$ with respect to k, for r and $t_{(r)}$ fixed.

To calculate the kernel density estimators at (7), for $k = 0, 1$ and $j = 1, \ldots, r$, we take bandwidths $h_{k,j}$ of the form $h_{k,j} = \hat{\sigma}_{k,j} h_k$, where $\hat{\sigma}_{k,j}^2$ is the empirical variance of the $X_i(t_j)$s coming from population k, and h_0 and h_1 are chosen using the nearest neighbour method. More specifically, $h_0 = (d_{0,k} + d_{0,k+1})/2$, where $d_{0,k}^2$ is the kth order statistic of $\sum_{j=1}^r |x(t_j) - X_i(t_j)|^2/\hat{\sigma}_{0,j}^2$, for X_i in group 0, and h_1 is defined similarly.

In both cases, and as in Ferraty and Vieu (2006), we restrict our search of the number of neighbours k to a grid. We use the grid $[5, n^*/2]$, where $n^* = n$ in the regression case, and $n^* = \min(n_0, n_1)$ in the density case. In the latter setting, we use the same value of k for both density estimators. We break ties in the same way as in §2·4.

3. CLUSTERING

In clustering problems, we observe only the functional data $X_1, \ldots, X_n$, and the goal is to cluster them in a certain number, k say, of groups. Unlike the classification case, there are opportunities for clustering functional data without any dimension reduction. For example, the L_2 metric for functions can sometimes be used to good effect for k-means clustering (Chiou and Li, 2007). Nevertheless, in clustering problems there is a great deal of superfluous information in functional data. To appreciate why, note that since the functions are generally continuous then, if t is close to u, $X(t)$ is usually close to $X(u)$, and so clustering on the variables $X_i(t)$ for $i = 1, \ldots, n$ will typically give very similar results to clustering on the $X_i(u)$s.

This viewpoint motivates the problem of determining the places in the interval $\mathcal{I}$ that have particularly good leverage for clustering. Which parts of the interval are especially useful for

discriminating between two clusters, and which parts are largely unnecessary because the information they convey is present in other, nearby places? Answering this question can provide important practical insight. Moreover, in cases where the experimenter knows what parts of the curves they consider as being important, knowing what parts of the curves the clustering algorithm focuses on helps identify if the clustering method is appropriate for their problem or not.

We shall answer the question in the case of the popular k-means clustering algorithm. There, if we reduce each function X_i to the vector $\mathrm{X}_i(t_{(r)}) = \big(X_i(t_1), \ldots, X_i(t_r)\big)^{\mathrm{T}}$, the following iterative algorithm is used to determine clusters based on that choice of components. (a) Given an assignment of data to k clusters, determine the mean or centroid, $\bar{\mathrm{X}}_\ell(t_{(r)})$ say, of the data $\mathrm{X}_i(t_{(r)})$ in the ℓth cluster, for $\ell = 1, \ldots, k$. (b) Recompute the clusters by assigning each $\mathrm{X}_i(t_{(r)})$ to the cluster corresponding to the value of $\bar{\mathrm{X}}_\ell(t_{(r)})$ that is nearest to $\mathrm{X}_i(t_{(r)})$. Steps (a) and (b) are iterated until convergence is achieved. At that point we consider the ℓth cluster, $\mathcal{C}_\ell(t_{(r)})$ say, to consist of functions X_i, not just the vectors $\mathrm{X}_i(t_{(r)})$, and we write $\bar{X}_\ell$ for the mean, or centroid, of functions $X_i \in \mathcal{C}_\ell(t_{(r)})$. A measure of the tightness of the clusters is given by

$$S_r(t_{(r)}) = \sum_{\ell=1}^{k} \sum_{X_i \in \mathcal{C}_\ell(t_{(r)})} \|X_i - \bar{X}_\ell\| \,,$$

where on this occasion $\|\cdot\|$ denotes the L_2 metric on functions. Then $S_r(t_{(r)})$ is our measure of the tightness of the clusters when the components of the data functions are determined by $t_{(r)}$. In this notation, and making the assumption that tighter clusters are better, we use

$$T_r = \inf_{t_{(r)} \in \mathcal{I}_r} S_r(t_{(r)}) \tag{10}$$

as our benchmark for performance, and take the most important r-dimensional point $t_{(r)} = \hat{t}_{(r)}$ to be the one that minimizes $S_r(t_{(r)})$.

Empirical algorithms for choosing a particular value, $\hat{p}$, of r are similar to those suggested earlier. For example, we can define $\hat{p}$ as at (2) or (3), using the definition of T_r at (10). We used (2) with $\rho = 0{\cdot}1$.

4. NUMERICAL PROPERTIES

4·1. *Full versus sequential approaches*

It is tempting to proceed sequentially using a greedy algorithm, and to define first an estimator $\hat{t}_1$ of the value $t_1 \in \mathcal{I}$ that produces the smallest value of T_1. Then, given $\hat{t}_1$, estimate t_2 as the value $\hat{t}_2$ which, when adjoined to $\hat{t}_1$, leads to the smallest value of T_2; and so on. This is the approach taken by Ferraty et al. (2010) in a related problem of functional prediction. However, it usually does not lead to consistent estimation of the optimal values of t_j. That is perhaps best seen by considering the case $p = 2$, where it can be shown that, although the pair $(\hat{t}_1, \hat{t}_2)$ generally converges in probability to a limit (t_1', t_2'), the set $\{t_1', t_2'\}$ is usually different from the pair $\{t_1, t_2\}$ that gives optimal prediction of Y from $\big(X(t_1), X(t_2)\big)$. The problem is that t_1' was, in a sense, a compromise between t_1 and t_2, and so by adding a new point t_2' without also revising the value of t_1' we are incurring performance losses because of the initial compromise. For similar reasons the sequential algorithm may not even converge.

On the other hand, a full search taking into account, for successively higher values of r, all possible sequences $t_{(r)} = (t_1, \ldots, t_r)^{\mathrm{T}}$, can be feasible for $r = 1$, 2 or 3, but becomes computationally too costly for higher values of r. We suggest using an approach that makes a compromise between the full and the sequential search, as follows. For each $r \geq 1$, at step $r + 1$, i.e. on going from r points to $r + 1$ points, first use a sequential approach, adjoining $\hat{t}_{r+1}$ to the points $\hat{t}_1, \ldots, \hat{t}_r$ selected at the rth step. Then refine this choice by constructing a neighbour-

hood around each point $\hat{t}_1, \ldots, \hat{t}_{r+1}$, and performing a full search over $(t_1, \ldots, t_{r+1})^{\mathrm{T}}$ in that neighbourhood. Then continue to step $r + 2$, proceeding similarly.

Another computational saving can easily be made by noticing that neighbouring points t and u usually have very similar values of $X(t)$ and $X(u)$, therefore rendering quite inefficient a method that would consider all possible sequences $t_{(r)}$. Motivated by this, the next paragraph describes two time-saving simplifications. These are based purely on empirical and computational considerations, and can of course be modified if a visual inspection of the curves suggests that finer grids should be employed in all or parts of $\mathcal{I}$, for example in areas where $X(t)$ changes rapidly. However, we believe that our prescription can be used as a default in most cases.

For the sequential part of the algorithm we suggest performing the search for each t_i on a grid of approximately 150 equispaced points over the interval $\mathcal{I}$, and never letting any two points t_i and t_j, for $i \neq j$, be closer than $2\Delta t$, where Δt denotes the space between two adjacent points of the grid. If the curves $X_i(t)$ are observed only for a number $L < 150$ of t values, then we replace 150 by L. For the refining part of the algorithm described two paragraphs above, as r increases we suggest taking shorter and shorter grids, our default being to use, for each t_j, 20 neighbouring points equispaced by $2\Delta t$ for $r = 2$ and $r = 3$, ten points equispaced by $2\Delta t$ for $r = 4$, and to perform only a sequential approach for $r \geq 5$. Further simplifications can be made to reduce computational time for $r = 4$, for example, by performing the multidimensional refinement on only three of the components. In general we do not expect more than just a few points to be selected by the procedure. In all the examples on which we tested our method, we rarely selected more than three or four points. In our experience, such algorithms run reasonably fast, for example they rarely take more than two minutes of CPU time for $n = 100$ on a computer equipped with an Intel Xeon W3520@2.67GHz processor.

4·2. *Real data illustrations*

We applied the five classification methods described in §2·4 on three real datasets. As we shall see below, overall the methods that performed the best were the nonparametric regression-based and the logistic regression-based classifiers. The nonparametric Bayes classifier gave results similar to the nonparametric regression-based one. However, in small sample sizes the two empirical bandwidths required by the former often implied that it was beaten marginally by the latter. Therefore, for brevity, we do not discuss the Bayes classifier below. More detailed results are available in the Supplementary Material.

For comparison, we also considered classifiers based on functional approaches that project the data via partial least-squares or principal components; such functional approaches were used by, for example, Ferraty and Vieu (2006), Leng and Müller (2006), Escabias et al. (2007), Preda et al. (2007) and Delaigle and Hall (2012). In the partial least-squares case we applied the classifiers of §2·4 to the setting where, instead of the projecting on $\hat{t}_1, \ldots, \hat{t}_{\hat{p}}$, we used the univariate projection $\int_{\mathcal{I}} X_i \hat{\beta}$, where $\hat{\beta}$ was the partial least-squares approximation to the slope function of the linear regression of I_i on X_i. Such classifiers are defined in the same way as in §2·4, except that we replace the dimension r by 1 and each occurrence of $x(t_j)$ and $X_i(t_j)$ by $\int_{\mathcal{I}} x\hat{\beta}$ and $\int_{\mathcal{I}} X_i\hat{\beta}$, respectively. As detailed in Delaigle and Hall (2012), the partial least-squares slope estimator $\hat{\beta}$ is defined by a linear combination of q basis functions, and we chose q by minimising the cross-validation estimator of classification error defined in §2. For the linear discriminant, we know from Delaigle and Hall (2012) that, in a variety of settings, the partial least-squares projection is optimal. Hence in this case we do not expect our point selection method to improve often on the performance of the one based on partial least-squares, but the attraction of our approach lies in the insight brought by the points it selects.

In the case of principal components we applied the nonparametric regression-based classifier defined in §2·4 in the setting where, instead of projecting on $\hat{t}_1, \ldots, \hat{t}_{\hat{p}}$, we used the p-dimensional projection $(\int_{\mathcal{I}} X_i \hat{\phi}_1, \ldots, \int_{\mathcal{I}} X_i \hat{\phi}_p)^{\mathrm{T}}$, where $\hat{\phi}_1, \ldots, \hat{\phi}_p$ were the first p eigenfunctions obtained by empirical principal component analysis (Hall and Hosseini-Nassab, 2006), and where p was chosen to minimize the cross-validation estimator of classification error defined in §2. This classifier is defined by the formula of the fourth classifier described on page 9, if we replace r by p and each occurrence of $x(t_j)$ and $X_i(t_j)$ by $\int_{\mathcal{I}} x\hat{\phi}_j$ and $\int_{\mathcal{I}} X_i \hat{\phi}_j$, respectively.

Finally, using ideas similar to those used in the prediction context by Ferraty and Vieu (2009), we implemented a boosting version of our nonparametric and logistic regression-based procedures, by adding to the fitted curve $\hat{g}$ a nonparametric estimator of the regression of the fitted residuals on X_i. More precisely, we calculated $\hat{m}(x) = \sum_{i=1}^n \hat{\epsilon}_i K\big(\|x - X_i\|/h\big) / \sum_{i=1}^n K\big(\|x - X_i\|/h\big)$, where $\hat{\epsilon}_i = Y_i - \hat{g}(X_i \,|\, t_{(r)})$ and $\|x\|^2 = \int_{\mathcal{I}} x^2$. We took h to be the kth smallest value of $\|x - X_1\|, \ldots, \|x - X_n\|$, where k minimized this cross-validation estimate of classification error of the classifier that assigns a new data function x to population 0 if $\hat{g}(x) + \hat{m}(x) < 0.5$, and to population 1 otherwise. Our boosted classifier assigns a new data function x to population 0 if $\hat{\gamma}(x) < 0.5$, and to population 1 otherwise, where $\hat{\gamma}$ is, among the two fitted curves $\hat{g}$ and $\hat{g} + \hat{m}$, the one that leads to the smallest cross-validation estimate of classification error.

For each of the three datasets, we let N denote the total number of observations, of which N_k are in group k, for $k = 0$ and 1. To assess the performance of the classification methods on a given dataset, we randomly divided the dataset into a training sample of size n and a test sample of size $N - n$, for each of $n = 30$, 50 and 100. Each training sample was obtained by drawing uniformly n observations, without replacement, from the main dataset. In each case we generated 200 pairs of training and test samples; for each pair we constructed the classifier from

the training sample, applied it to classify the observations from the test sample, and calculated the resulting classification error rate. Each boxplot shown in the figures below was constructed from 200 such error rates, and so were the tables with additional numerical results, provided in the Supplementary Material.

All our codes were written in Matlab. Parts of our codes that calculate nonparametric regression and density estimators reflect the freely available R codes of Ferraty and Vieu (2006). For the nonparametric estimator we used the bandwidth described in §2·5, and the Epanechnikov kernel $K(u) = (1 - u^2)\, 1\{|u| \le 1\}$. In each case we took the prior probability equal to $1/2$.

Next we describe our datasets. In the rainfall data, which are available at `http://dss.ucar.edu/datasets/ds482.1`, we considered $N = 190$ rainfall curves from $N_0 = 43$ northern and $N_1 = 147$ southern Australian weather stations, used by Delaigle and Hall (2010). Each $X_i(t)$ denotes rainfall at time t for the ith weather station, where $t \in [0, 365]$ represents the period that has passed, in a given year, at the time of measurement, and, as in Delaigle and Hall (2010), rainfall is averaged, by local linear smoothing, over the years for which the station has been operating. Fig. 1 shows for each group the curves and their means $\bar{X}_0 = N_0^{-1} \sum_{i=1}^{N_0} X_i$ and $\bar{X}_1 = N_1^{-1} \sum_{i=N_0+1}^{N} X_i$.

The Tecator data, available at `http://lib.stat.cmu.edu/datasets/tecator`, consist of $N = 240$ observations of near infrared absorbance spectra of finely chopped meat, recorded, using a Tecator Infratec Food & Feed Analyzer, at 100 equispaced values of t ranging from 850 nanometres to 1050 nanometres, and numbered 1 to 100 in the graphs. As usual with chemometrics data, for $i = 1, \ldots, 240$ we took the curves $X_i(t)$ to be smooth versions of the first derivative of the spectra; see Remark 1 below. The fat content, Y, of each meat sample was also available. Since these data had no natural grouping, we artificially split them into two groups. First, as in Ferraty and Vieu (2006), §8.4.2, we put the $N_0 = 85$ curves for which $Y > 20$ in

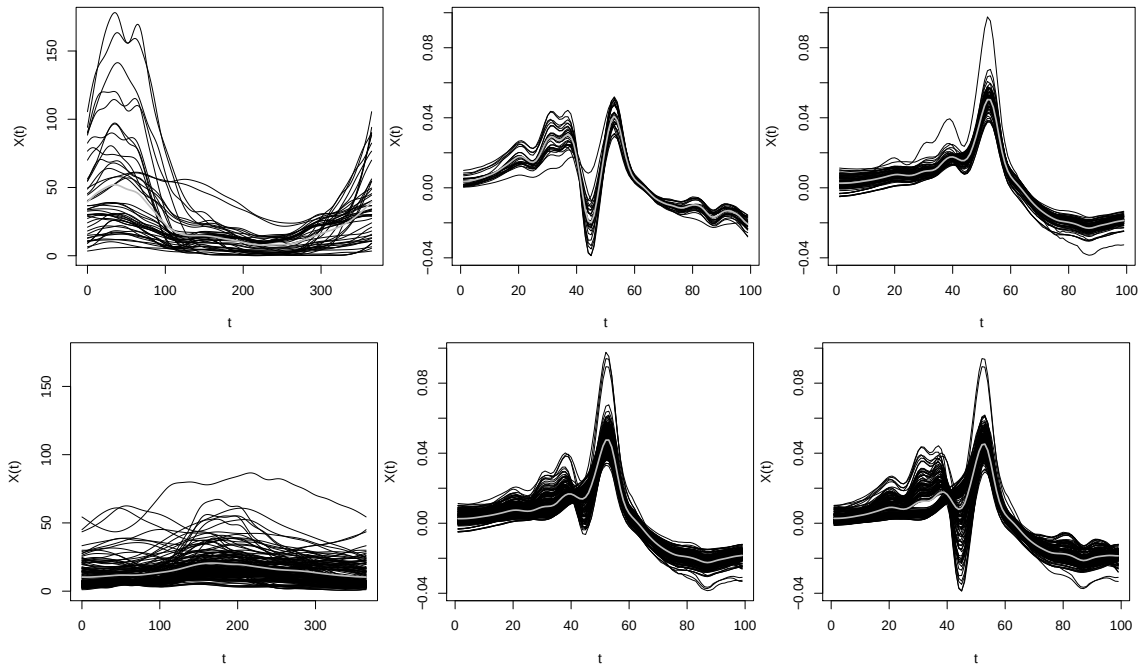

Fig. 1. Rainfall and Tecator data. First column: rain curves; second column: derivative spectra of the Tecator data, case I; third column: derivative spectra of the Tecator data, case II. First row: data from group 0; second row: data from group 1. The mean curves of each group are shown in grey.

group 0, and the remaining $N_1 = 155$ curves in group 1. We refer to this as case I. Then we considered a more complex case, which we refer to as case II, where we grouped the data so that the mean curves of the two groups were almost identical. There we put the $N_0 = 75$ curves for which $Y \in [10, 25]$ in group 0, and the remaining $N_1 = 165$ curves in group 1. Since linear and quadratic discriminant methods are based on mean differences, these classifiers are clearly inadequate here and cannot give an average classification error rate much lower than 0.5, but will be included in our discussion for illustrative purposes. Fig. 1 shows the curves $X_i(t)$ and the mean curves for each group.

The phoneme data are available at `www-stat.stanford.edu/ElemStatLearn`. Here, the $N = 1717$ curves $X_i(t)$, for $i = 1, \ldots, N$, are log-periodograms constructed from 32 milliseconds long recordings of males pronouncing two phonemes: $N_0 = 695$ curves are observations of the phoneme aa as in dark, and $N_1 = 1022$ curves concern the phoneme ao as in water.

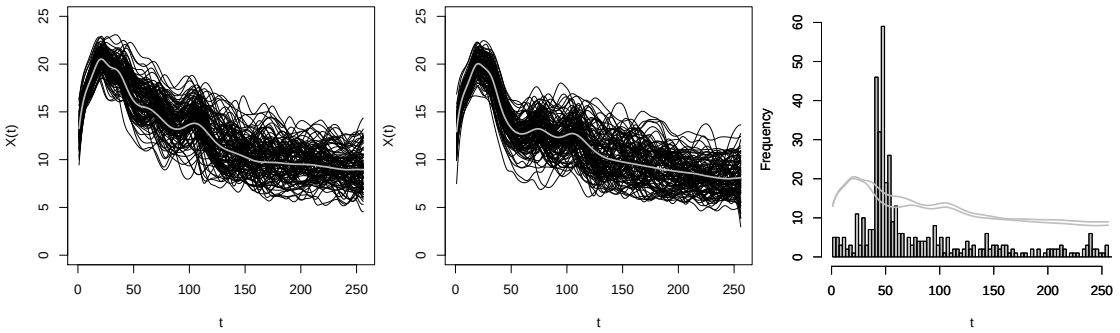

Fig. 2. Phoneme data. Left: 100 curves from the phoneme aa; middle: 100 curves from the phoneme ao; right: histogram of points selected by the nonparametric regression-based classifier, calculated for 200 samples when $n = 100$. The grey curves are the group means.

Each curve was observed at 256 equispaced frequencies t, denoted on the horizontal axes of the graphs by 1 to 256. A sample of 100 curves and the means from the two groups are shown in Fig. 2.

Remark 1. Spectrometric curves are generally very smooth, and to first order they generally differ from one another mostly by a vertical shift. Taking the derivatives of these curves removes this shift and permits us to focus on more subtle differences, which can significantly improve the performance of nonlinear regression methods, as illustrated in Ferraty and Vieu (2006). We found the same to be true for nonlinear classifiers, which performed poorly with the non differentiated curves, compared to classifiers based on the first or second derivatives. In such cases the cross-validation estimate of the classification error, based on the spectra, was usually much larger than that based on their first or second derivatives. This indicates that practitioners who do not have sufficient knowledge about properties of their data can be guided by cross-validation to choose which derivative to work with.

Our numerical investigation revealed some interesting facts. (i) Overall the method that worked the best was nonparametric regression combined with our point selection approach. In cases where the two groups were divided in a rather simple way, the three logistic-based tech-

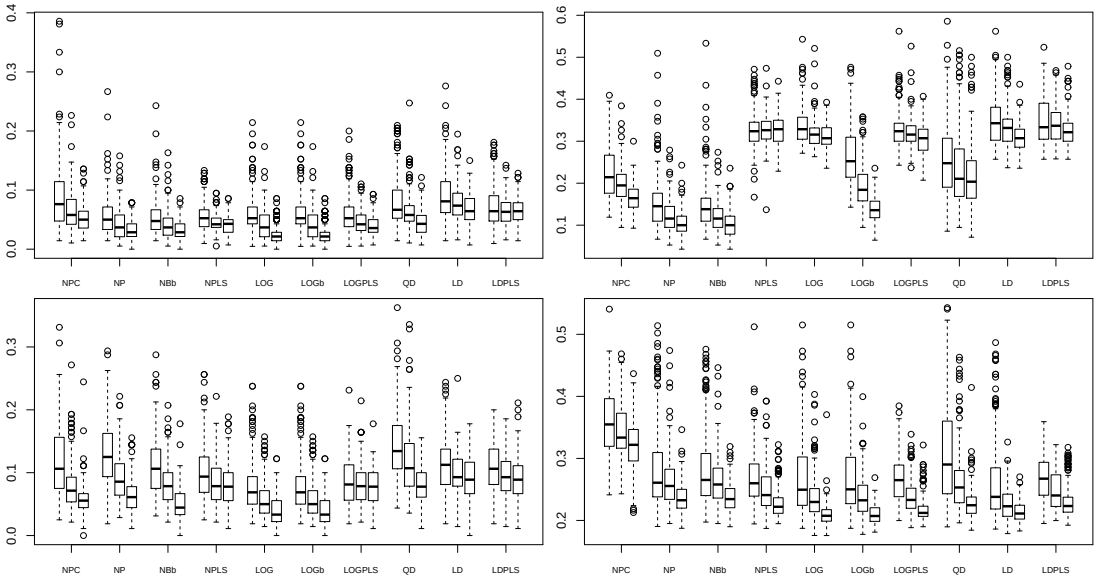

Fig. 3. Boxplots of classification error rates based on 200 samples. Top left: Tecator data, case I; top right: Tecator data, case II; bottom left: rainfall data; bottom right: phoneme data. We show boxplots for the nonparametric regression-based methods combined with our approach (NP), with principal components (NPC) or with partial least-squares (NPLS), the boosting version of NP (NPb), the logistic regression methods combined with our approach (LOG), with partial least-squares (LOGPLS) and with boosting (LOGb), the linear discriminant method combined with our approach (LD) and with partial least-squares (LDPLS), and the quadratic discriminant method (QD). In each group of three boxes, the first is for training samples of size $n = 30$, the second for $n = 50$, the third for $n = 100$.

niques and the nonparametric method based on partial least-squares performed very well, often slightly better than the nonparametric procedure based on our point selection method. See the results for the Tecator case I and phoneme data in Fig. 3. In these cases, the main advantage of our approach is the additional insight brought by the identification of those points that are most important for classification. When the groups were created in a more complex way, the nonparametric method combined with our point selection approach performed best, and sometimes considerably better than the other approaches. See the results for case II of the Tecator data in Fig. 3. (ii) The three logistic methods often gave results similar to each other, but the best ones were those based on our approach, which also has the advantage discussed at (i). (iii) Linear

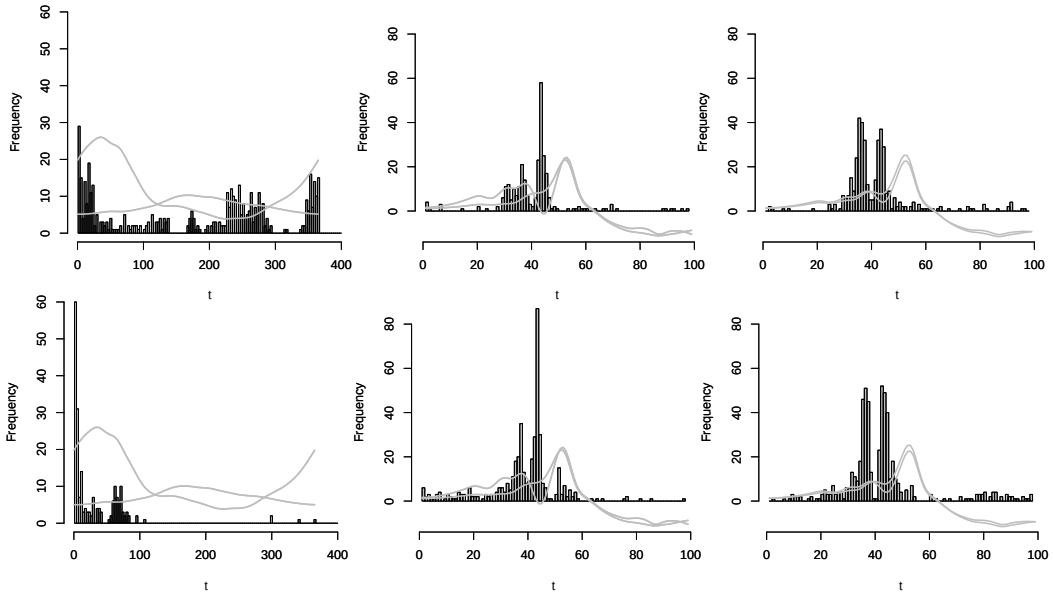

Fig. 4. Rainfall and Tecator data. Left: histograms of points selected by the nonparametric regression-based classifier (top) or the centroid clusterer (bottom) for the rainfall data, calculated from 200 samples, when $n = 100$. Middle: histograms of points selected by the nonparametric regression-based classifier for the Tecator data, case I, calculated for 200 samples, when $n = 30$ (top) or $n = 100$ (bottom). Right: same as middle, for Tecator, case II. The grey curves show a constant multiple of the mean curves of each group.

discrimination combined with our approach or with partial least-squares, performed very similarly. This shows that our point selection method works particularly well since, for the linear discriminant method, it is often virtually impossible to beat partial least-squares projection; see Delaigle and Hall (2012). Again, our approach has the attractiveness discussed in (i).

In Fig. 2. of the Supplementary Material we show graphs indicating the number of points selected by the nonparametric regression-based classifier. We learn from those figures that our procedure rarely chooses more than three points. Overall, the number of points selected tended to increase with sample size. This is connected to the fact that nonparametric methods work well in higher dimensions only when the sample size is large enough, and cross-validation is able to detect this.

For a given method, the number and location of points selected by our procedure varied among the 200 pairs of samples, but points that had high leverage for classification were selected in many of these 200 samples. To illustrate this we constructed histograms showing the frequency at which each point was selected over the 200 test samples, and to visualize the features of the curves on which our method focused we superimposed a rescaled version of the group mean curves. Such histograms, for the nonparametric regression-based classifier, are shown in the third column of Fig. 2 and in Fig. 4. We can see that, for a given dataset, the selected points depend on the way the groups were created; compare cases I and II of the Tecator data. For rapidly changing curves, such as with the Tecator dataset, the points frequently selected generally correspond to a mode or an inflection point of the curves $X_i(t)$. Moreover, the location of the points is quite sharply determined. For curves that vary more slowly, such as the phoneme or rainfall data, neighbouring points carry similar information and, as a result, the location of the points is more widespread. Interestingly, the points selected by the nonparametric regression-based classifier are different from those selected by the clustering method, which we applied to the same 200 subsamples of sizes $n = 30$, 50 and 100 for these rainfall data, using the k-means clustering algorithm described in §3. Remember that when data are clustered, there is no test sample for which the group is known, and grouping is based only on the X values. The histograms of the points selected by this method are shown for $n = 100$ in Fig. 4; similar points were selected for $n = 30$ and $n = 50$.

ACKNOWLEDGEMENT

Research supported by grants and fellowships from the Australian Research Council. The Australian weather data were assembled by the Australian Bureau of Meteorology. We thank Bob Dattore for providing the data, which are available from the Research Data Archive, maintained

by the Computational and Information Systems Laboratory at the National Center for Atmospheric Research sponsored by the National Science Foundation. We are also grateful to two referees, an associate editor and the editor for helpful suggestions.

SUPPLEMENTARY MATERIAL

Supplementary material available at *Biometrika* online includes a description of procedures for breaking ties, additional simulation results and all proofs.

APPENDIX: THEORETICAL PROPERTIES

Results in the case of classification

Recall the definitions of err_r and $\hat{\mathrm{err}}_r$ at (4) and (1). Let $t^0_{(r)}$ denote the vector that minimizes $\mathrm{err}_r(t_{(r)})$ over $t_{(r)} \in \mathcal{I}_r$. Technical conditions for the theorems are given at page 24. We assume that the density estimators used are kernel estimators, in the case of the empirical nonparametric Bayes rule, or, under the assumption that the data are Gaussian and when Fisher's linear or quadratic discriminator is employed, are constructed by maximum likelihood. Of course, many alternative assumptions are possible in the latter setting, but we make the simplifying Gaussian assumption because Fisher's discriminators are optimal in that case.

In Theorems 1 and 2, below, we state properties of the first three classifiers introduced in §2·4. The properties of the regression-based classifier are identical and can be derived with essentially the same proofs. Fisher's linear and quadratic discriminators become unreliable if the covariance matrices $\hat{\Sigma}(t_{(r)})$ and $\hat{\Sigma}_k(t_{(r)})$ used in their construction are close to being singular, so we restrict attention to the set $\mathcal{J}_r(c)$ of r-vectors $t_{(r)} \subseteq \mathcal{I}_r$ for which the determinants of the corresponding true covariance matrices exceed a given, small positive constant c. Analogously, in the case of the empirical nonparametric Bayes rule we confine ourselves to $t_{(r)}$ in the class $\mathcal{J}_r(c)$ for which the true densities $f_k(\,\cdot \,|\, t_{(r)})$, for $k = 0, 1$, are bounded above by c^{-1}, and in either case we suppose that, for some $\eta > 0$, a sphere centred at $t^0_{(r)}$ and of

radius η is contained in $\mathcal{J}_r(c)$. In practical terms these restrictions amount to asking that, for the vectors $t_{(r)}$ that we consider, none of the components are too close to one another.

Our first result shows that $\hat{\text{err}}(t_{(r)})$ and $\hat{t}_{(r)}$ are consistent for $\text{err}_r(t_{(r)})$ and $t^0_{(r)}$, respectively. Condition A is given at page 24 and the proof is given in the Supplementary Material, for the nonparametric Bayes method. The arguments are similar for the linear and quadratic discriminants.

THEOREM 1. *Fix $r \geq 1$ and assume that either Condition A holds, in the case of empirical nonparametric Bayes rule, or the process X is Gaussian and satisfies $E\{\sup_{t\in\mathcal{I}} |X'(t)|^C\} < \infty$ for some $C > 0$, in the context of Fisher's linear and quadratic discriminators. Then, as $n \to \infty$, (i) $\hat{\text{err}}(t_{(r)}) = \text{err}_r(t_{(r)}) + o_p(1)$ uniformly in $t_{(r)} \in \mathcal{J}_r(c)$, and (ii) $\hat{t}_{(r)} = t^0_{(r)} + o_p(1)$.*

Our next result, a corollary of Theorem 1, shows that error rates of the empirical classifiers, defined at (5), converge in probability to the minimum error rate suggested by the respective algorithm. In the theorem, we select $\hat{p}$ as in (2) and (3), except that we restrict our search to $r \leq r_0$, where $r_0 \geq 1$ is a finite upper bound. That is, we use

$$\hat{p} = \inf\{r \leq r_0 : (1-\rho)\, T_r \leq T_{r+1}\} \tag{A1}$$

or

$$\hat{p} = \inf\{r \leq r_0 : T_r - T_{r+1} \leq \rho\, T_1\}. \tag{A2}$$

This does not change anything in practice, but it makes the proofs considerably simpler. The proof of the theorem is given in the Supplementary Material, for the nonparametric Bayes method. The arguments are similar for the linear and quadratic discriminants.

THEOREM 2. *Assume that the conditions of Theorem 1 hold for $r = 1, \ldots, r_0 + 1$. (i) Define*

$$p = \inf\left\{r \leq r_0 : (1-\rho)\,\text{err}_r\big(t^0_{(r)}\big) \leq \text{err}_{r+1}\big(t^0_{(r+1)}\big)\right\},$$

where the set on the right-hand side is assumed to be non-empty, and suppose that there exists $\eta > 0$ such that

$$\inf_{r \le r_0} \{\mathrm{err}(t^0_{(r+1)}) - (1-\rho)\,\mathrm{err}(t^0_{(r)})\} > \eta. \tag{A3}$$

Then, if $\hat{p}$ is selected as in (A1)*, we have, as $n \to \infty$:*

$$\mathrm{pr}(\hat{p} = p) \to 1, \quad \mathrm{err}^{\mathrm{emp}} \to \mathrm{err}(t^0_{(p)})\,. \tag{A4}$$

(ii) Define

$$p = \inf\big\{r \le r_0 : \mathrm{err}_r\big(t^0_{(r)}\big) - \mathrm{err}_{r+1}\big(t^0_{(r+1)}\big) \le \rho\,\mathrm{err}_1\big(t^0_{(1)}\big)\big\},$$

where the set on the right-hand side is assumed to be non-empty, and suppose that there exists $\eta > 0$ such that

$$\inf_{r \le r_0} \{\rho\,\mathrm{err}_1\big(t^0_{(1)}\big) - \mathrm{err}_r\big(t^0_{(r)}\big) + \mathrm{err}_{r+1}\big(t^0_{(r+1)}\big)\} > \eta. \tag{A5}$$

Then, if $\hat{p}$ is selected as in (A2)*,* (A4) *holds as $n \to \infty$.*

Condition A

Let E_k denote expectation for data from population k, and recall that n_k is the number of data pairs (X_i, I_i) for which $I_i = k$, where $k = 0$ or 1, and that c is the small positive constant in the definition of $\mathcal{J}_r(c)$, introduced prior to Theorem 1. Define $n = n_1 + n_2$. For simplicity we take the bandwidths $h_{k1}, \ldots, h_{kr}$ to be identical and to equal $h = h(n)$, say, for each r.

Condition A:

(a) The kernel K is a symmetric, compactly supported, univariate probability density satisfying the Hölder continuity condition $|K(u) - K(v)| \le C_1\,|u - v|^{C_2}$ for constants $C_1 > 0$ and $0 < C_2 \le 1$, and for all real u and v;

(b) the bandwidth h used when computing $\hat{f}_k^{-i}(\cdot \,|\, t_{(r)})$ and $\hat{f}_k(\cdot \,|\, t_{(r)})$, for $k = 0, 1$, satisfies $h = O(n^{-C_3})$ and $(nh^r)^{-1} = O(n^{-C_3})$ for some $C_3 > 0$;

(c) for $k = 0$ and 1 the ratio n_k/n is bounded away from zero as $n \to \infty$;

(d) X is differentiable on $\mathcal{I}$, $E_k[\sup_{t\in\mathcal{I}} \{|X(t)|^C + |X'(t)|^C\}] < \infty$ for $k = 0, 1$ and for sufficiently large $C > 0$;

(e) the joint densities $f_0(\cdot \,|\, t_{(r)})$ and $f_1(\cdot \,|\, t_{(r)})$ of $\big(X(t_1), \ldots, X(t_r)\big)^{\mathrm{T}}$, in populations 0 and 1 respectively, satisfy $\sup_{x\in\mathbb{R}^r} \sup_{\mathrm{u}_{(r)},\mathrm{v}_{(r)}\in\mathcal{J}_r(c):\|\mathrm{u}_{(r)}-\mathrm{v}_{(r)}\|\le\epsilon} |f_k(x \,|\, \mathrm{u}_{(r)}) - f_k(x \,|\, \mathrm{v}_{(r)})| \to 0$ as $\epsilon \to 0$;

(f) the multivariate distributions of X have the property that, for each $\epsilon > 0$, there exist $\delta > 0$ and $n_0 \ge 1$ such that, for all $n \ge n_0$, $|\mathrm{err}(t_{(r)}) - \mathrm{err}(t^0_{(r)})| > \delta$ whenever $\|t_{(r)} - t^0_{(r)}\| > \epsilon$ and $t_{(r)} \in \mathcal{J}_r(c)$;

(g) for $k = 0$ or 1,

$$\lim_{\epsilon\downarrow 0} \sup_{t_{(r)}\in\mathcal{J}(c)} \mathrm{pr}_k\Big\{|\pi_k f_k(X \,|\, t_{(r)}) - \pi_{1-k} f_{1-k}(X \,|\, t_{(r)})| \le \epsilon\Big\} = 0.$$

Condition A(b) is satisfied by the majority of kernels used in practice. The conditions on h in A(b), or stronger ones, are conventionally imposed when deriving consistency of nonparametric estimators of smooth functions of r variables. The other parts of Condition A are self evident.

REFERENCES

ARAKI, Y., KONISHI, S., KAWANO, S. AND MATSUI, H. (2009). Functional logistic discrimination via regularized basis expansions. *Commun. Statist. Theory 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.

CHIOU, J.-M. AND LI, P.-L. (2007). Functional clustering and identifying substructures of longitudinal data. *J. R. Statist. Soc.* B **69**, 679–699.

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.

DELAIGLE, A. AND HALL, P. (2012). Achieving near-perfect classification for functional data. *J. R. Statist. Soc.* B, doi: 10.1111/j.1467-9868.2011.01003.x, to appear.

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

ESCABIAS, M., AGUILERA, A.M. AND VALDERRAMA,M.J. (2007). Functional PLS logit regression model. *Comput. Statist. Data Anal.* **51**, 4891–4902.

FERRATY, F. AND VIEU, P. (2003). Curves discrimination: a nonparametric functional approach. *Comput. Statist. Data Anal.* **4**, 161–173.

FERRATY, F. AND VIEU, P. (2006). *Nonparametric Functional Data Analysis*. New York: Springer.

FERRATY, F. AND VIEU, P. (2009). Additive prediction and boosting for functional data. *Comput. Statist. Data Anal.* **53**, 1400–1413.

FERRATY, F., HALL, P. AND VIEU, P. (2010). Most-predictive design points for functional data predictors. *Biometrika* **97**, 807–824.

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. New York: Springer.

GLENDINNING, R.H. AND HERBERT, R.A. (2003). Shape classification using smooth principal components. *Patt. Recognition Lett.* **24**, 2021–2030.

HALL, P., POSKITT, D. AND PRESNELL, B. (2001). A functional data-analytic approach to signal discrimination. *Technometrics* **43**, 1–9.

HALL, P. AND HOSSEINI-NASSAB , M. (2006). On properties of functional principal components analysis. *J. R. Statist. Soc.* B **68**, 109–126.

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. R. Statist. Soc.* B **63**, 533–550.

JAMES, G. AND SUGAR, C. (2003). Clustering for Sparsely Sampled Functional Data. *J. Am. Statist. Assoc.* **98**, 397–408.

LENG, X. 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.

PENG, J. AND MÜLLER, H.-G. (2008). Distance-based clustering of sparsely observed stochastic processes, with applications to online auctions. *Ann. Appl. Statist.* **2**, 1056–1077.

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. New York: Springer.

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. Mult. 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. Biol. Chem.* **32**, 426–432.

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.

WAND, M. P AND JONES, M.C. (1995). *Kernel Smoothing*. London: Chapman & Hall.

WANG, X.H., RAY, S. AND MALLICK, B.K. (2007). Bayesian curve classification using wavelets. *J. Am. Statist. Assoc.* **102**, 962–973.