Biometrika (2017), **xx**, x, *pp.* 1–16
Printed in Great Britain

Optimal Bayes Classifiers for Functional Data and Density Ratios

By Xiongtao Dai, Hans-Georg Müller
Department of Statistics, University of California, Davis, California 95616, U.S.A.
dai@ucdavis.edu hgmueller@ucdavis.edu

and Fang Yao
Department of Statistical Sciences, University of Toronto, 100 St. George Street, Toronto, Ontario M5S 3G3, Canada
fyao@utstat.toronto.edu

Summary

Bayes classifiers for functional data pose a challenge. One difficulty is that probability density functions do not exist for functional data, so the classical Bayes classifier using density quotients needs to be modified. We propose to use density ratios of projections on a sequence of eigenfunctions that are common to the groups to be classified. The density ratios are then factored into density ratios of individual projection scores, reducing the classification problem to obtaining a series of one-dimensional nonparametric density estimates. The proposed classifiers can be viewed as an extension to functional data of some of the very earliest nonparametric Bayes classifiers that were based on simple density ratios in the one-dimensional case. By means of the factorization of the density quotients the curse of dimensionality that would otherwise severely affect Bayes classifiers for functional data can be avoided. We demonstrate that in the case of Gaussian functional data, the proposed functional Bayes classifier reduces to a functional version of the classical quadratic discriminant. A study of the asymptotic behaviour of the proposed classifiers in the large sample limit shows that under certain conditions the misclassification rate converges to zero, a phenomenon that has been referred to as perfect classification. The proposed classifiers also perform favourably in finite sample applications, as we demonstrate through comparisons with other functional classifiers in simulations and various data applications, including spectral data, functional magnetic resonance imaging data for attention deficit hyperactivity disorder patients, and yeast gene expression data.

Some key words: Common functional principal component; Density estimation; Functional classification; Gaussian process; Quadratic discriminant analysis.

1. Introduction

For the classification of functional data, predictors may be viewed as random trajectories and responses are indicators for two or more categories. The goal of functional classification is to assign a group label to each predictor function, i.e., to predict the group label for each of the observed random curves. Functional classification is a rich topic with applications in many areas of commerce, medicine, the sciences, chemometrics,

and genetics (Leng & Müller, 2006; Song et al., 2008; Zhu et al., 2010, 2012; Francisco-Fernández et al., 2012; Coffey et al., 2014). Within the functional data analysis framework (Wang et al., 2016), each observation is viewed as a smooth random curve on a compact domain. Functional classification also has recently been extended to the related task of classifying longitudinal data (Wu & Liu, 2013; Wang & Qu, 2014; Yao et al., 2016) and also has close connections with functional clustering (Chiou & Li, 2008). The vast literature on functional classification includes distance-based classifiers (Ferraty & Vieu, 2003; Alonso et al., 2012), k-nearest neighbour classifiers (Biau et al., 2005; Cérou & Guyader, 2006; Biau et al., 2010), Bayesian methods (Wang et al., 2007), logistic regression (Araki et al., 2009), or partial least squares (Preda & Saporta, 2005; Preda et al., 2007),

Bayes classifiers based on density quotients are optimal in the sense of minimizing misclassification rates, and this provides one of the major motivations for developing nonparametric density estimation (technical report by Fix & Hodges Jr, 1951 commented by Sillverman & Jones, 1989; Rosenblatt, 1956; Parzen, 1962; Wegman, 1972). However, for multiple predictors unrestricted nonparametric approaches are subject to the curse of dimensionality (Scott, 2015). This leads to very slow rates of convergence for estimating the nonparametric densities for dimensions larger than three or four and renders the resulting classifiers practically worthless. The situation is exacerbated in the case of functional predictors, which are infinite-dimensional and hence afflicted by a particularly bad curse of dimensionality, as small ball probabilities in function space imply that the expected number of functions falling into balls with small radius is so small that densities do not even exist in most cases (Li & Linde, 1999; Delaigle & Hall, 2010).

Hence, in order to define a Bayes classifier through density quotients with reasonably good estimation properties, one needs to invoke sensible restrictions, for example on the class of predictor processes. This approach was adopted in Delaigle & Hall (2012), who consider two Gaussian populations with equal covariance using a functional linear discriminant, in analogy to the linear discriminant, corresponding to the Bayes classifier in the analogous multivariate Gaussian case. Galeano et al. (2015) proposed a closely related functional quadratic method for discriminating two general Gaussian populations, making use of a suitably defined Mahalanobis distance for functional data. In contrast to these previous approaches, we aim here at the construction of a nonparametric Bayes classifier for functional data. The idea is to project the observations onto an orthonormal basis that is common to the two populations, then to construct density ratios through products of the density ratios of the projection scores. This corresponds to the Bayes classifier if scores are independent. The densities themselves are nonparametrically estimated, which is feasible as they are only one-dimensional. We establish the asymptotic equivalence of the proposed functional nonparametric Bayes classifiers and their estimated versions as well as asymptotic perfect classification for the proposed classifiers.

The term perfect classification was introduced in Delaigle & Hall (2012) to denote conditions where the misclassification rate converges to zero as an increasing number of projection scores is used, and we use it in the same sense here. Perfect classification in the Gaussian case requires that there are certain differences between the mean or covariance functions, while such differences are not a prerequisite for the proposed nonparametric approach to succeed. In the special case of Gaussian functional predictors, the proposed classifiers simplify to those considered in Delaigle & Hall (2013). Additionally, we extend our theoretical results to cover the practically important situation where the functional data are not fully observed, but rather are observed as noisy measurements that are made

on a dense grid, while previous approaches were based on the less realistic assumption of fully observed trajectories without noise.

2. Functional Bayes Classifiers

We consider the situation where the observed data come from a common distribution (X, Y), where X is a fully observed square integrable random function in $L^2(\mathcal{T})$, $\mathcal{T}$ is a compact interval, and $Y \in \{0, 1\}$ is a group label. Assuming that X shares the same distribution with $X^{(k)}$ if X is from population Π_k $(k = 0, 1)$, that is, $X^{(k)}$ has the same distribution as X given $Y = k$, and that $\pi_k = \mathrm{pr}(Y = k)$ is the prior probability that an observation falls into Π_k, our goal is to infer the group label Y of a new observation X. The optimal Bayes classification rule that minimizes misclassification error classifies an observation $X = x$ to Π_1 if

$$Q(x) = \frac{\mathrm{pr}(Y = 1 \mid X = x)}{\mathrm{pr}(Y = 0 \mid X = x)} > 1,$$

where we denote realized functional observations by x and random predictor functions by X. We denote the conditional densities of the functional observations X when conditioning on the group label 0 or 1 by g_0 and g_1, assuming that these conditional densities exist. Then Bayes' theorem implies that

$$Q(x) = \frac{\pi_1 g_1(x)}{\pi_0 g_0(x)}. \tag{1}$$

Since translation-invariant densities for functional data do not usually exist (Delaigle & Hall, 2010) and the density quotients are known only for certain classes of Gaussian processes (Baíllo et al., 2011; Berrendero et al., 2015, arXiv:1507.04398), we consider a sequence of approximations with increasing number of components and then use the density ratios (1).

Specifically, we represent x and the random X by projecting onto an orthogonal basis $\{\psi_j\}_{j=1}^{\infty}$, yielding the projection scores $\{x_j\}_{j=1}^{\infty}$ and $\{\xi_j\}_{j=1}^{\infty}$, where $x_j = \int_{\mathcal{T}} x(t)\psi_j(t)\,\mathrm{d}t$ and $\xi_j = \int_{\mathcal{T}} X(t)\psi_j(t)\,\mathrm{d}t$ $(j = 1, 2, \dots)$. As noted in Hall et al. (2001), when comparing the conditional probabilities, it is sensible to project the data from both groups onto the same basis. Our goal is to approximate the conditional probabilities $\mathrm{pr}(Y = k \mid X = x)$ by $\mathrm{pr}(Y = k \mid \text{the first } J \text{ scores of } x)$, where $J \to \infty$. Then by Bayes' theorem,

$$Q(x) \approx \frac{\mathrm{pr}(Y = 1 \mid \text{the first } J \text{ scores of } x)}{\mathrm{pr}(Y = 0 \mid \text{the first } J \text{ scores of } x)} = \frac{\pi_1 f_1(x_1, \dots, x_J)}{\pi_0 f_0(x_1, \dots, x_J)}, \tag{2}$$

where f_1 and f_0 are the conditional densities for the first J random projection scores $\xi_1, \dots, \xi_J$.

Since estimating the joint density of $(\xi_1, \dots, \xi_J)$ is impractical and subject to the curse of dimensionality when J is large, it is sensible to introduce conditions that simplify (2). A first simplification is to assume the auto-covariances of the stochastic processes that generate the observed data have the same ordered eigenfunctions for both populations. Denote the mean functions as $\mu_k(t) = E\{X^{(k)}(t)\}$, and the covariance functions as $G_k(s, t) = \mathrm{cov}\{X^{(k)}(s), X^{(k)}(t)\}$ with associated covariance operators

$$G_k : L^2(\mathcal{T}) \to L^2(\mathcal{T}), \quad G_k(f) = \int_{\mathcal{T}} G_k(s, t) f(s)\,\mathrm{d}s.$$

Assuming $G_k(s,t)$ is continuous, by Mercer's theorem (see e.g. Bosq, 2000)

$$G_k(s,t) = \sum_{j=1}^{\infty} \lambda_{jk}\phi_{jk}(s)\phi_{jk}(t),$$

where $\lambda_{1k} \geq \lambda_{2k} \geq \cdots \geq 0$ are the eigenvalues of G_k, ϕ_{jk} are the corresponding orthonormal eigenfunctions ($j = 1, 2, \ldots$), and $\sum_{j=1}^{\infty} \lambda_{jk} < \infty$ ($k = 0, 1$). The common eigenfunction condition is then $\phi_{j0} = \phi_{j1} = \phi_j$, where ϕ_j is the jth common eigenfunction (Flury, 1984; Benko et al., 2009; Boente et al., 2010; Coffey et al., 2011). This assumption can be weakened to the requirement of equality of the set of eigenfunctions, which ignores their order, in which case one can reorder the eigenfunctions and eigenvalues such that $\phi_{j0} = \phi_{j1} = \phi_j$. Choosing the projection directions $\{\psi_j\}_{j=1}^{\infty}$ as the shared eigenfunctions $\{\phi_j\}_{j=1}^{\infty}$, one has $\operatorname{cov}(\xi_j, \xi_l) = 0$ if $j \neq l$, where the scores ξ_j correspond to the functional principal components $\int_{\mathcal{T}}\{X(t) - \mu_k(t)\}\phi_j(t)\,\mathrm{d}t$ only if $\mu_k \equiv 0$.

A second simplification is that we assume that the projection scores are independent under both populations, whence the densities in (2) factor and the criterion function can be rewritten by taking logarithms as

$$Q_J(x) = \log\left(\frac{\pi_1}{\pi_0}\right) + \sum_{j=1}^{J} \log\left\{\frac{f_{j1}(x_j)}{f_{j0}(x_j)}\right\}, \tag{3}$$

where f_{jk} is the density of the jth score under Π_k. We classify into Π_1 if and only if $Q_J(x) > 0$. Due to the zero divided by zero problem, (3) is defined only on a set $\mathcal{X}$ with $\operatorname{pr}(X \in \mathcal{X}) = 1$, and our theoretical arguments in the following are restricted to this set. For the asymptotic analysis we will consider the case where $J = J(n) \to \infty$ as $n \to \infty$. The independent projections assumption is commonly made in functional data analysis, and is satisfied by a large class of processes, including Gaussian processes. For processes with dependent projection scores, the performance of our method will depend on how well the process can be approximated through independent projection scores. Our proposed classifiers demonstrated good performance relative to other classifiers even under violations of the independence assumption; see Section 5·2.

When predictor processes X are Gaussian for group $k = 0, 1$, the projection scores ξ_j are independent and one may substitute Gaussian densities for the densities f_{jk} in (3). Writing the jth projection of the mean function $\mu_k(t)$ of Π_k as $\mu_{jk} = E(\xi_j \mid Y = k) = \int_{\mathcal{T}} \mu_k(t)\phi_j(t)\,\mathrm{d}t$, in this special case of our general nonparametric approach, one obtains the simplified version

$$Q_J^G(x) = \log\left(\frac{\pi_1}{\pi_0}\right) + \frac{1}{2}\sum_{j=1}^{J}\left[(\log\lambda_{j0} - \log\lambda_{j1}) - \left\{\frac{1}{\lambda_{j1}}(x_j - \mu_{j1})^2 - \frac{1}{\lambda_{j0}}(x_j - \mu_{j0})^2\right\}\right]. \tag{4}$$

Here $Q_J^G(X)$ either converges to a random variable almost surely if $\sum_{j\geq 1}(\mu_{j1} - \mu_{j0})^2/\lambda_{j0} < \infty$ and $\sum_{j\geq 1}(\lambda_{j0}/\lambda_{j1} - 1)^2 < \infty$, or otherwise diverges to ∞ or $-\infty$ almost surely, as $J \to \infty$. More details about the properties of $Q_J^G(X)$ can be found in the Supplementary Material. It is apparent that (4) is the quadratic discriminant rule using the first J projection scores, which is the Bayes classifier for multivariate Gaussian data with different covariance structures. If further $\lambda_{j0} = \lambda_{j1}$ ($j = 1, 2, \ldots$) then one has equal covariances and (4) reduces to the functional linear discriminant (Delaigle & Hall, 2012).

As the proposed method does not assume Gaussianity and allows for densities f_{jk} of general form in (3), one may expect better performance than Gaussian-based functional classifiers when the distributions are non-Gaussian. This is borne out by the simulation results in Section 5·2. The densities of the projection scores can be estimated nonparametrically by kernel density estimation (Silverman, 1986) as described in Section 3.

3. Estimation

Under the common eigenfunction assumption, we may write $G_k(s,t) = \mathrm{cov}\{X^{(k)}(s), X^{(k)}(t)\} = \sum_{j=1}^{\infty} \lambda_{jk}\phi_j(s)\phi_j(t)$ where the ϕ_j are the common eigenfunctions. We then estimate the ϕ_j, which serve as the projection directions, by pooling data from the two groups in the training data to obtain a joint covariance estimate for the joint covariance operator $G = \pi_0 G_0 + \pi_1 G_1$. Then ϕ_j is also the jth eigenfunction of G with eigenvalue $\lambda_j = \pi_0\lambda_{j0} + \pi_1\lambda_{j1}$. Assume we have $n = n_0 + n_1$ functional predictors $X_1^{(0)}, \ldots, X_{n_0}^{(0)}$ and $X_1^{(1)}, \ldots, X_{n_1}^{(1)}$ sampled from Π_0 and Π_1. We estimate the mean and covariance functions by $\hat{\mu}_k(t)$ and $\hat{G}_k(s,t)$, the sample mean and sample covariance function for group k, and estimate π_k by $\hat{\pi}_k = n_k/n$. Setting $\hat{G}(s,t) = \hat{\pi}_0\hat{G}_0(s,t) + \hat{\pi}_1\hat{G}_1(s,t)$ and denoting the jth eigenvalue-eigenfunction pair of $\hat{G}$ by $(\hat{\lambda}_j, \hat{\phi}_j)$, we obtain the projections for a generic functional observation X as $\hat{\xi}_j = \int_{\mathcal{T}} X(t)\hat{\phi}_j(t)\,\mathrm{d}t$ $(j = 1, \ldots, J)$, denoting the projection scores of $X_i^{(k)}$ by $\hat{\xi}_{ijk}$, where we assume fully observed noise-free predictor trajectories. The eigenvalues λ_{jk} are estimated by $\hat{\lambda}_{jk} = \int_{\mathcal{T}}\int_{\mathcal{T}} \hat{G}_k(s,t)\hat{\phi}_j(s)\hat{\phi}_j(t)\,\mathrm{d}s\,\mathrm{d}t$, which is motivated by $\lambda_{jk} = \int_{\mathcal{T}}\int_{\mathcal{T}} G_k(s,t)\phi_j(s)\phi_j(t)\,\mathrm{d}s\,\mathrm{d}t$, the pooled eigenvalues by $\hat{\lambda}_j = \hat{\pi}_0\hat{\lambda}_{j0} + \hat{\pi}_1\hat{\lambda}_{j1}$, and the jth projection scores μ_{jk} of $\mu_k(t)$ by $\hat{\mu}_{jk} = \int_{\mathcal{T}} \hat{\mu}_k(t)\hat{\phi}_j(t)\,\mathrm{d}t$. The resulting estimators for μ_k, G_k, ϕ_j, and λ_{jk} are consistent; see the Appendix.

We then proceed to obtain nonparametric estimates of the densities for each of the projection scores by applying kernel density estimates (Silverman, 1986) to the sample projection scores from group k. The kernel density estimate for the jth projection in group k is

$$\hat{f}_{jk}(u) = \frac{1}{n_k h_{jk}} \sum_{i=1}^{n_k} K\Big(\frac{u - \hat{\xi}_{ijk}}{h_{jk}}\Big), \tag{5}$$

where $u \in \mathbb{R}$ and $h_{jk} = h\hat{\lambda}_{jk}^{1/2}$ are bandwidths adapted to the variance of the j-th projection score, see Sections 4 and 5·1, leading to corresponding estimates of the density ratios $\hat{f}_{j1}(u)/\hat{f}_{j0}(u)$ that are used to obtain an estimated version of (3). An alternative estimate for the density ratios based on nonparametric kernel regression (Nadaraya, 1964; Watson, 1964) is discussed in the Supplementary Material. Writing $\hat{x}_j = \int_{\mathcal{T}} x(t)\hat{\phi}_j(t)\,\mathrm{d}t$, the estimated criterion function based on kernel density estimate is thus

$$\hat{Q}_J(x) = \log\frac{\hat{\pi}_1}{\hat{\pi}_0} + \sum_{j\le J} \log\frac{\hat{f}_{j1}(\hat{x}_j)}{\hat{f}_{j0}(\hat{x}_j)}. \tag{6}$$

In practice, the assumption that functional data are fully observed trajectories is often unrealistic. Rather, one encounters observations of the functions that have been taken on a regular or irregular design, possibly with some missing observations, where the measurements are contaminated with measurement errors that one may assume are in-

dependent with zero mean and finite variance. In this situation, one can smooth the discrete observations using local linear kernel smoothers, and then regard the smoothed trajectory as a fully observed functional predictor. We provide theoretical justification for this approach by showing that one may obtain the same asymptotic classification results as for fully observed functional data, with details given before Theorem 1. Specifically, for each curve we pre-smooth the noisy measurements

$$W_{ikl} = X_i^{(k)}(t_{ikl}) + \varepsilon_{ikl} \quad (i = 1, \ldots, n_k;\ k = 0, 1;\ l = 1, \ldots, m_{ik}),$$

by local linear smoothers, where m_{ik} is the number of measurements per curve. For each $t \in \mathcal{T}$ we set $\tilde{X}_i^{(k)}(t) = \hat{\beta}_0$, where

$$(\hat{\beta}_0, \hat{\beta}_1) = \underset{\beta_0, \beta_1}{\arg\min} \sum_{l=1}^{m_{ik}} K_0\left(\frac{t - t_{ikl}}{w_{ik}}\right) \{W_{ikl} - \beta_0 - \beta_1(t - t_{ikl})\}^2,$$

K_0 is a kernel function, and w_{ik} is the bandwidth used for pre-smoothing.

Denoting the sample covariance function of the smoothed predictors in group k by $\tilde{G}_k$, the estimated pooled covariance by $\tilde{G}(s,t) = \hat{\pi}_0\tilde{G}_0(s,t) + \hat{\pi}_1\tilde{G}_1(s,t)$, the estimated jth eigenfunction of $\tilde{G}$ by $\tilde{\phi}_j(t)$, and the estimated projection scores by $\tilde{\xi}_{ijk} = \int_{\mathcal{T}} \tilde{X}_i^{(k)}(t)\tilde{\phi}_j(t)\,\mathrm{d}t$ and $\tilde{x}_j = \int_{\mathcal{T}} x(t)\tilde{\phi}_j(t)\,\mathrm{d}t$ for a random function $\tilde{X}_i^{(k)}$ or fixed function x, the densities of the projection scores are obtained by kernel density estimates

$$\tilde{f}_{jk}(u) = \frac{1}{n_k h_{jk}} \sum_{i=1}^{n_k} K\left(\frac{u - \tilde{\xi}_{ijk}}{h_{jk}}\right) \tag{7}$$

analogous to (5). Finally, $\tilde{Q}_J$ is the criterion function using J components analogous to $\hat{Q}_J$ in (6), but with kernel density estimates $\tilde{f}_{jk}$ as in (7).

4. Asymptotic Properties

We present the asymptotic equivalence of the estimated and the true Bayes classifiers in Theorem 1 and give conditions for the proposed nonparametric Bayes classifiers to achieve perfect classification in Theorem 2, with proofs in the Supplementary Material. We assume here the following simplifications, which can easily be weakened. Without loss of generality we denote the mean functions of Π_0 and Π_1 by 0 and $\mu(t)$, respectively, since we can subtract the mean function of Π_0 from all samples, whereupon $\mu(t)$ becomes the difference in the mean functions. We also assume that $\pi_0 = \pi_1$ and $n_0 = n_1$ and use a common multiplier h for all bandwidths $h_{jk} = h\lambda_{jk}^{1/2}$ in the kernel density estimates. Write $I(\cdot)$ for the indicator function that has value 1 if the condition inside the brackets holds and 0 otherwise. In order for $I\{Q_J(x) \geq 0\}$ to be the Bayes classifier based on the first J projection scores, we need the following assumptions:

Condition 1. the covariance operators $G_k(s,t)$ under Π_0 and Π_1 have common eigenfunctions;

Condition 2. for all $j \geq 1$, the projection scores ξ_j onto the common eigenfunctions ϕ_j are independent under Π_0 and Π_1, and their densities exist.

Condition 1 means that the covariance functions $G_k(s,t)$ under Π_0 and Π_1 can be decomposed as $G_k(s,t) = \mathrm{cov}\{X^{(k)}(s), X^{(k)}(t)\} = \sum_j \lambda_{jk}\phi_j(s)\phi_j(t)$, where the ϕ_j are

the common eigenfunctions and λ_{jk} are the associated eigenvalues. For our analysis, the common eigenfunctions serve as projection directions and are assumed to be such that the projection scores become independent, as is for example the case if predictor processes satisfy the more restrictive Gaussian assumption; see Section 6 for further discussion. Additional assumptions are provided in the Appendix.

THEOREM 1. *Under Conditions* 1–2 *and A*1–*A*9*, for any* $\epsilon > 0$ *there exist a set* S *with* $\mathrm{pr}(S) > 1 - \epsilon$ *and a sequence* $J = J(n, \epsilon) \to \infty$ *such that* $\mathrm{pr}(S \cap [I\{\tilde{Q}_J(X) \geq 0\} \neq I\{Q_J(X) \geq 0\}]) \to 0$ *as* $n \to \infty$.

Theorem 1 provides the asymptotic equivalence of the estimated classifier based on the kernel density estimates (7) of pre-smoothed observations and the Bayes classifier $I\{Q_J(x) \geq 0\}$ based on the first J projections. This implies that it is sufficient to investigate the asymptotics of the Bayes classifier based on Q_J to establish asymptotic perfect classification.

The next result shows that the proposed nonparametric Bayes classifiers achieve perfect classification under certain conditions. Let $m_j = \mu_j/\lambda_{j0}^{1/2}$ and $r_j = \lambda_{j0}/\lambda_{j1}$.

THEOREM 2. *Under Conditions* 1–2 *and A*10–*A*11*, the Bayes classifier* $I\{Q_J(x) \geq 0\}$ *achieves perfect classification if* $\sum_{j\geq 1}(r_j - 1)^2 = \infty$ *or* $\sum_{j\geq 1} m_j^2 = \infty$*, as* $J \to \infty$.

This theorem extends previous results on perfect classification, as in Delaigle & Hall (2012) and Delaigle & Hall (2013), to classifiers of a more general nonparametric form. The conditions for perfect classification in Theorem 2 are sufficient but not necessary. The general case that we study here has the interesting feature that when Π_1 and Π_0 are non-Gaussian, perfect classification may occur even if the mean and covariance functions under the two groups are the same. This may happen for instance when the distributions of the infinitely many independent projection scores have different shapes, which provides information for discrimination. For example, the projection scores ξ_j may be independent random variables with the same mean and variance for both populations, but may follow normal distributions under Π_1 and Laplace distributions under Π_0; see the Supplementary Material. In such cases, attempts at classification under Gaussian assumptions are doomed, as mean and covariance functions are the same between the groups, while the proposed nonparametric Bayes classifiers can reflect these differences.

5. NUMERICAL PROPERTIES

5·1. *Practical Considerations*

We propose three practical implementations for estimating the projection score densities $f_{jk}(\cdot)$ that will be compared in our data illustrations, along with other previously proposed functional classification methods. All of these involve the choice of tuning parameters, namely bandwidths and the number of components included. We describe below how these are specified. Our first implementation is the nonparametric density classifier as in (6), where one estimates the density of each projection by applying kernel density estimators to the observed sample scores as in (5). The second implementation is the nonparametric regression approach detailed in the Supplementary Material, where we apply kernel smoothing (Nadaraya, 1964; Watson, 1964) to the scatter plots of the pooled estimated scores and group labels. For the kernel estimates we use a Gaussian ker-

nel where the bandwidth multiplier h is chosen by ten-fold cross-validation, minimizing the misclassification rate.

The third implementation is referred to as the Gaussian method. Each of the projections is assumed to be normally distributed with mean and variance estimated by the sample mean $\hat{\mu}_{jk} = \sum_{i=1}^{n_k} \hat{\xi}_{ijk}/n_k$ and sample variance $\hat{\lambda}_{jk} = \sum_{i=1}^{n_k} (\hat{\xi}_{ijk} - \hat{\mu}_{jk})^2/(n_k - 1)$ of $\hat{\xi}_{ijk}$ $(i = 1, \ldots, n)$. We then use the density of $N(\hat{\mu}_{jk}, \hat{\lambda}_{jk})$ as $\hat{f}_{jk}(\cdot)$. This Gaussian implementation differs from the quadratic discriminant implementation discussed for example in Delaigle & Hall (2013), as in our approach we always force the projection directions for the two populations to be the same. This has the practical advantage of providing more stable estimates for the eigenfunctions and is a prerequisite for constructing nonparametric Bayes classifiers for functional predictors. For all classifiers included in our comparisons, the number of projections J used is selected by ten-fold cross-validation, jointly with the selection of h for the nonparametric methods.

5·2. *Simulation Results*

We illustrate the proposed Bayes classifiers in three simulation settings for varying distributions and dependency assumptions for the projection scores. For the first two scenarios, the samples are generated by $X_i^{(k)}(t) = \mu_k(t) + \sum_{j=1}^{50} A_{ijk}\phi_j(t)$ $(i = 1, \ldots, n_k;$ $k = 0, 1)$, where n_k is the number of samples in Π_k. The A_{ijk} are independent random variables with mean 0 and variance λ_{jk}, which are generated under two distribution scenarios: Scenario A, the A_{ijk} are normally distributed; Scenario B, the A_{ijk} are centred exponentially distributed. For Scenario C, we generate samples with uncorrelated but dependent scores by $X_i^{(k)}(t) = \mu_k(t) + \sum_{j=1}^{50}(A_{ijk}/B_{ik})\phi_j(t)$, where the A_{ijk} are the same as in Scenario B, and the B_{ik} are independent and follow the same distribution as $\chi^2_{30}/30$, or Gamma$(30, 30)$.

In each setting, we generate n training samples, each having $1/2$ chance to be from Π_0 or Π_1, and let ϕ_j be the jth function in the Fourier basis, where $\phi_1(t) = 1$, $\phi_2(t) = \sqrt{2}\cos(2\pi t)$, $\phi_3(t) = \sqrt{2}\sin(2\pi t)$, etc., $t \in [0, 1]$. We set $\mu_0(t) = 0$, and $\mu_1(t) = 0$ or t for the same or the different mean scenarios, respectively. The variances of A_{ijk} under Π_0 are $\lambda_{j0} = e^{-j/3}$, and those under Π_1 are $\lambda_{j1} = e^{-j/3}$ or $e^{-j/2}$ $(j = 1, \ldots, 50)$ for the same or the different variance scenarios, respectively. The random functions are sampled at 51 equally spaced time points from 0 to 1, with additional small measurement errors in the form of independent Gaussian noise with mean 0 and variance 0.01 added to each observation for all scenarios. We use modest sample sizes of $n = 50$ and $n = 100$ for training the classifiers, and 500 samples for evaluating the predictive performance.

Each simulation experiment is repeated 500 times with the goal to compare the predictive performance of the following functional classification methods: the centroid method (Delaigle & Hall, 2012); the proposed nonparametric Bayes classifier in three versions: basing estimation on Gaussian densities, nonparametric densities, or nonparametric regression, as discussed in Section 5·1; logistic regression; the functional quadratic discriminant as in Galeano et al. (2015); and the Gaussian process logistic regression (Rasmussen & Williams, 2006) with squared exponential function and automatic relevance determination. The functional quadratic discriminant was never the winner for any scenario in our simulation study so we omitted it from the tables. We show the results corresponding to pre-smoothing the predictors by local linear smoothers with cross-validation bandwidth choice in Table 1. Since all classifiers improve performance when pre-smoothing the pre-

dictor functions, the results obtained without pre-smoothing are only presented in the Supplementary Material.

Table 1. *Misclassification rates (%), with standard errors in brackets for pre-smoothed predictors for the simulation scenarios*

n	μ	λ	Centroid	Gaussian	NPD	NPR	Logistic	GP Logistic
Scenario A (Gaussian case)								
50	same	diff	48.9 (0.14)	22.7 (0.17)	23.1 (0.20)	25.7 (0.21)	48.9 (0.13)	30.3 (0.30)
	diff	same	36.5 (0.24)	38.3 (0.22)	40.7 (0.22)	39.3 (0.23)	32.2 (0.26)	42.5 (0.26)
	diff	diff	33.4 (0.25)	18.0 (0.16)	18.4 (0.18)	20.3 (0.20)	28.1 (0.26)	24.9 (0.27)
100	same	diff	48.9 (0.14)	17.1 (0.11)	18.1 (0.12)	19.4 (0.13)	49.1 (0.14)	20.3 (0.15)
	diff	same	29.8 (0.23)	31.6 (0.23)	33.6 (0.25)	31.9 (0.25)	25.4 (0.15)	34.7 (0.35)
	diff	diff	27.0 (0.24)	13.0 (0.11)	14.0 (0.12)	14.8 (0.13)	21.1 (0.14)	15.3 (0.15)
Scenario B (exponential case)								
50	same	diff	48.5 (0.15)	28.3 (0.18)	29.1 (0.21)	31.4 (0.24)	48.6 (0.14)	33.0 (0.29)
	diff	same	35.0 (0.24)	38.4 (0.22)	38.0 (0.22)	36.5 (0.23)	30.9 (0.23)	36.6 (0.25)
	diff	diff	30.3 (0.24)	20.2 (0.18)	20.9 (0.22)	21.4 (0.22)	27.0 (0.23)	23.3 (0.25)
100	same	diff	48.5 (0.15)	25.1 (0.13)	24.0 (0.14)	25.0 (0.14)	48.4 (0.15)	24.3 (0.18)
	diff	same	29.2 (0.23)	33.3 (0.23)	32.3 (0.20)	31.1 (0.21)	25.4 (0.17)	30.0 (0.25)
	diff	diff	26.1 (0.22)	16.5 (0.14)	14.6 (0.13)	14.7 (0.13)	21.6 (0.16)	14.6 (0.16)
Scenario C (dependent case)								
50	same	diff	48.5 (0.15)	32.2 (0.20)	34.1 (0.23)	36.0 (0.24)	48.4 (0.15)	38.1 (0.28)
	diff	same	36.1 (0.25)	39.8 (0.24)	40.1 (0.22)	38.6 (0.23)	31.4 (0.24)	38.3 (0.24)
	diff	diff	31.6 (0.24)	24.6 (0.20)	25.6 (0.22)	26.3 (0.22)	27.5 (0.23)	26.7 (0.25)
100	same	diff	48.6 (0.15)	29.5 (0.13)	29.7 (0.14)	30.8 (0.14)	48.7 (0.15)	31.2 (0.20)
	diff	same	31.0 (0.22)	35.1 (0.23)	34.6 (0.19)	32.8 (0.21)	25.8 (0.18)	31.6 (0.26)
	diff	diff	27.6 (0.23)	21.8 (0.16)	20.3 (0.14)	20.4 (0.16)	21.9 (0.16)	17.8 (0.24)

Centroid, the method of Delaigle & Hall (2012); Gaussian, NPD, and NPR correspond to the Gaussian, nonparametric density, and nonparametric regression implementations of the proposed Bayes classifiers, respectively; Logistic, functional logistic regression; GP logistic, Gaussian process logistic regression.

For Scenario A, the proposed nonparametric Bayes classifiers have superior performance for those scenarios where covariance differences in the populations are present, while the logistic methods work best for those cases where the differences are exclusively in the mean. This is because the proposed nonparametric Bayes classifiers take into account both mean and covariance differences between the populations. For Scenario B, the proposed Bayes classifiers continue to outperform all other methods when covariance differences occur, especially when the sample size is small. When there are differences between the covariances, the Gaussian implementation performs the best when the sample size is small, while the nonparametric density implementation and the Gaussian process logistic regression perform the best when the sample size is large. This is likely due to the nonparametric classifiers having larger variance than the parametric classifiers so that they require more training data to perform well.

Scenario C is more challenging compared to the other scenarios due to the dependency in the projection scores, which violates Condition 2. Nevertheless the proposed classifiers outperform the other classifiers in the presence of covariance differences, especially if the sample size is small. Gaussian process logistic regression performs best when differences exist in both the mean functions and the covariance functions and the sample size is large, due to its capacity to tackle dependent predictors.

5·3. *Data Illustrations*

We present four data examples to illustrate the performance of the proposed Bayes classifiers for functional data. We pre-smoothed the yeast data by a local linear smoother with cross-validation bandwidth choice since the original observations are quite noisy as seen in Fig. 1, while for the wine and the attention deficit hyperactivity disorder datasets we used the curves as provided in the data, which were already preprocessed and smooth. Following the procedure described in Benko et al. (2009), we tested whether the eigenspaces generated by the first $J = 20$ eigenfunctions are common, as almost always the number of included components selected by cross-validation is less than 20; see Section 6 for further discussion. All p-values obtained from 2000 bootstrap samples are larger than 0.1, so the common eigenspace assumption appears to be reasonable.

We used repeated ten-fold cross-validation misclassification error rates to evaluate the performance of the classifiers. In order to obtain the correct cross-validation misclassification error rate, the selection of the number of components and bandwidth was carried out using only the training data in each cross-validation partition. We repeated the process 500 times and report the mean misclassification rates and the standard errors in Table 2. The proposed Bayes classifiers had the best performance for three of the four data sets, while functional quadratic discriminant also performed very well overall, indicating that covariance operator differences contain crucial information for the classification task.

Table 2. *500-repeat ten-fold cross-validation misclassification rates (%) with standard errors in brackets for real data*

Data	Centroid	Gaussian	NPD	NPR	Logistic	Quadratic
ADHD	41.7 (0.2)	34.1 (0.13)	36.7 (0.2)	36.8 (0.21)	47.4 (0.21)	34.6 (0.19)
Wheat	0.03 (0.01)	0.126 (0.016)	0.006 (0.003)	0.088 (0.014)	0 (0)	0.098 (0.014)
Yeast	20 (0.084)	12.5 (0.085)	15.1 (0.11)	14.4 (0.11)	20.8 (0.12)	14.5 (0.089)
Wine	6.84 (0.064)	5.08 (0.067)	5.09 (0.063)	4.67 (0.057)	7.56 (0.083)	5.93 (0.075)

ADHD, attention deficit hyperactivity disorder.

Our first data example concerns classifying attention deficit hyperactivity disorder patients from brain imaging data. The data were obtained in the Attention Deficit Hyperactivity Disorder-200 Sample. Attention deficit hyperactivity disorder is the most commonly diagnosed behavioural disorder in childhood and it is of interest to what extent it can be diagnosed from brain signals alone, where we use filtered preprocessed resting state functional magnetic resonance imaging data as predictors. The data were collected and preprocessed by the New York University Child Study Center (Tzourio-Mazoyer et al., 2002), with signals from 116 brain regions of interest. We consider only subjects for which the attention deficit hyperactivity disorder index is below the first quartile, which constitute group Π_0 with $n_0 = 36$, or above the third quartile, defining group Π_1 with $n_1 = 34$, aiming to predict group membership. The functional predictors are taken to be the average of the mean blood-oxygen-level dependent signals of the 91st to 108th regions, corresponding to the cerebellum that is known to have significant impact on the attention deficit hyperactivity disorder index (Berquin et al., 1998). These average signals are shown in the right panel of Fig. 1 for 172 time points.

The second data example concerns the classification of spectrometric data of wheat samples. This classification problem was originally described in Kalivas (1997). The goal is to use the near infrared spectra measured from 1100 to 2500 nm in 2 nm intervals to predict groups defined by moisture content, Π_0 with $n_0 = 41$ if the moisture content is

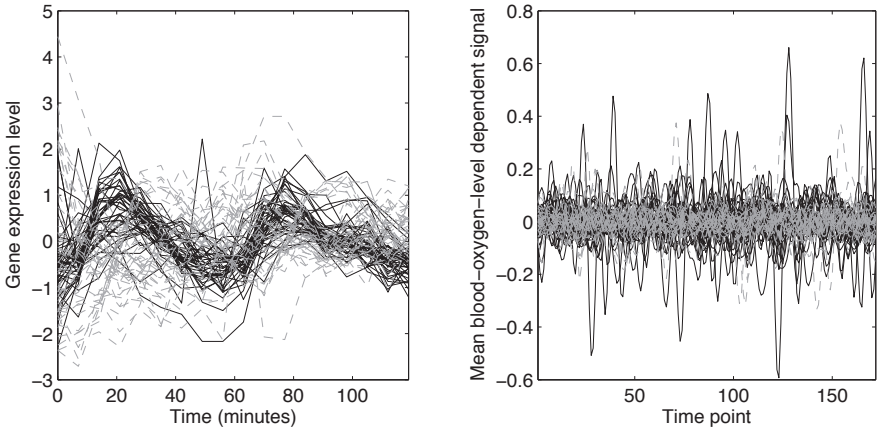

Fig. 1. Gene expression trajectories for the yeast data, left panel; the attention deficit hyperactivity disorder data, right panel, where for both panels Π_0, dashed line; Π_1 solid line.

less than 15, and Π_1 with $n_1 = 59$ otherwise. Following Delaigle & Hall (2012), we use B-splines to smooth the curves, and then consider the first derivatives of the functional observations. It turns out that for this example, functional logistic regression is the best classifier, while the nonparametric density estimation version of the Bayes classifier has the best performance among the other methods.

Our third data example focuses on yeast gene expression time courses during the cell cycle as predictors (Spellman et al., 1998). The predictors are gene expression level time courses for $n = 89$ genes, observed at 18 equally spaced time points from 0 minute to 119 minutes. The expression trajectories for genes related to G1 phase regulation of the cell cycle were regarded as group Π_1 with $n_1 = 44$ and all non-G1 related trajectories as group Π_0 with $n_0 = 45$. The Gaussian implementation of the proposed Bayes classifiers outperformed the other methods by a margin of at least 2%, while the functional quadratic discriminant was also competitive for this classification problem.

In our fourth example we analyse wine spectra data. These data were made available by Professor Marc Meurens, Université Catholique de Louvain and contain a training set of 93 samples and a testing set of 30 samples, which we combined into a dataset of size $n = 123$. For each sample the mean infrared spectrum on 256 points and the alcohol content are observed. Samples with alcohol contents greater than 12 were regarded as Π_1 with $n_1 = 78$, and the remaining samples as Π_0 with $n_0 = 45$.

The kernel density estimates of the first four projection scores for the wine example are displayed in Fig. 2, with Π_0 in dashed lines and Π_1 in solid lines. Clearly the densities are not normal, and some of them appear to be bimodal. The differences between each pair of densities are not limited to location and scale, but also manifest themselves in the shapes of the densities; in the second and the fourth plots the density estimate from one group is close to bimodal and the other density is not. The nonparametric implementations of the proposed Bayes estimators based on nonparametric regression or nonparametric density estimation are capable of reflecting such shape differences and therefore outperform the classifiers based on Gaussian assumptions.

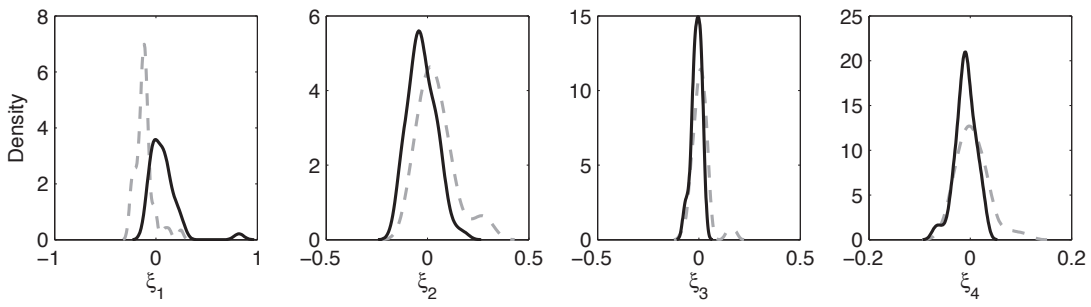

Fig. 2. Kernel density estimates for the first four projection scores for the wine spectra. Π_0, dashed line; Π_1 solid line.

In all examples, the quadratic discriminant outperforms the centroid method, suggesting that in these examples there is information contained in the differences between the covariance operators of the two groups to be classified. In the presence of such more subtle differences and additional shape differences in the distributions of projection scores the proposed nonparametric Bayes methods are expected to work particularly well.

6. Discussion

As the two groups to be classified often share common characteristics, the working assumption that the covariance functions of the predictor functions share some common structure is not unreasonable. We assume that the commonality between the covariances lie in the principal modes of variation, and the projection scores reflect latent factors that differ between groups. The common eigenfunction assumption is more general than the common or proportional covariance assumption and leads to sensible directions of projection for constructing the proposed Bayes classifiers, permitting meaningful between-group comparisons of variation (Benko et al., 2009; Coffey et al., 2011).

To justify the common eigenfunction assumption in practice, we tested whether the sets of eigenfunctions are common to two groups in real data applications. Since our method allows the eigenfunctions to have different orders, we implemented the test by following Benko et al. (2009), i.e., testing whether the eigenspaces generated by the first $J = 20$ components are the same; the null hypothesis was not rejected for any of the dataset. The common eigenspace assumption is weaker than the common eigenfunction assumption because the former allows one set of eigenfunctions to be a rotation of the other.

Processes with independent projections are generated by a fixed set of orthonormal directions of variation and a set of independent random variables representing the independent variation in each of the directions. The independent projection assumption seems restrictive but is satisfied by a reasonably large class of processes which includes Gaussian processes. This class of process is closed with respect to componentwise transformation. Processes generated by a non-linear transformation of a finite set of independent random variables are excluded from this class, however, because the functional principal components in the infinite dimensional space are then bound to lie on a certain manifold

with dependent projections. Independent component analysis (Hyvärinen & Oja, 2000) also assumes independence among components. For processes with dependent projection scores, Bayes classifiers can be constructed through estimating (2), but the joint densities may be practically estimated only for a small number of projections, due to the curse of dimensionality. Even in cases with dependent projection scores one may be able to approximate the multivariate joint density through product densities, as corroborated by our simulation results.

The proposed Bayes classifiers can be naturally extended to K-class classification by projecting observations onto a set of eigenfunctions common to all groups, estimating projection densities f_{jk} $(j = 1, \ldots, J; k = 1, \ldots, K)$ from the projections ξ_{jk} for group k, and then classifying into the group with highest posterior probability $\mathrm{pr}(Y = k \mid \xi_1, \ldots, \xi_J)$. This is equivalent to classifying into group k^* if the product density quotient of group k^* over k is greater than 1 for all $k \neq k^*$.

Acknowledgements

This work was supported by National Science Foundation and the Natural Sciences and Engineering Research Council of Canada. We thank the reviewers for helpful comments.

Supplementary Material

Supplementary material available at *Biometrika* online includes the description of our nonparametric regression estimate, an example for perfect classification when the mean and the covariance functions are the same, additional simulation results, and proofs.

Appendix

Assumptions and Additional Results

For simplicity of presentation we adopt throughout all proofs the simplifying assumptions mentioned in the first paragraph of Section 4. We remark that $\hat{\mu}_k$, $\hat{G}_k$, $\hat{\phi}_j$, and $\hat{\lambda}_{jk}$ constructed from the sample mean, covariance, eigenfunctions and eigenvalues of the completely observed functions are consistent estimates for their corresponding targets, as per Hall & Hosseini-Nasab (2006). Theorem A1 below states that $\hat{Q}_J(x)$ as in (6) is asymptotically equivalent to $Q_J(x)$ as in (3), for all J. We define the kernel density estimator using the true projection scores $\xi_{ijk} = \int_{\mathcal{T}} X_i^{(k)}(t)\phi_j(t)\,\mathrm{d}t$ as

$$\bar{f}_{jk}(u) = \frac{1}{n_k h_{jk}} \sum_{i=1}^{n_k} K\left(\frac{u - \xi_{ijk}}{h_{jk}}\right).$$

Let g_{jk} be the density functions of the standardized functional principal components $\xi_j/\lambda_{j0}^{1/2}$ when $k = 0$ and that of $(\xi_j - \mu_j)/\lambda_{j1}^{1/2}$ when $k = 1$, $\hat{g}_{jk}$ be the kernel density estimates of g_{jk} using the estimated functional principal components, and $\bar{g}_{jk}$ be the kernel density estimates using the true functional principal components, analogous to $\hat{f}_{jk}$ and $\bar{f}_{jk}$. Delaigle & Hall (2010) provide the uniform convergence rate of $\hat{g}_{jk}$ to $\bar{g}_{jk}$ on a compact domain, with detailed proof available in Delaigle & Hall (2011); our derivations utilize this result.

We make the following assumptions for $k = 0, 1$; here Conditions A1–A4 parallel assumptions (3.6)–(3.9) in Delaigle & Hall (2010), namely:

Condition A1. for all large $C > 0$ and some $\delta > 0$, $\sup_{t\in\mathcal{T}} E_{\Pi_k}\{|X(t)|^C\} < \infty$ and $\sup_{s,t\in\mathcal{T}:s\neq t} E_{\Pi_k}[\{|s-t|^{-\delta}|X(s) - X(t)|\}^C] < \infty$;

Condition A2. for each integer $r \geq 1$, $\lambda_{jk}^{-r} E_{\Pi_k}[\int_{\mathcal{T}}\{X(t) - E_{\Pi_k}X(t)\}\phi_j(t)\,\mathrm{d}t]^{2r}$ is bounded uniformly in j;

Condition A3. the eigenvalues $\{\lambda_j\}_{j=1}^{\infty}$ are all different, and so are the eigenvalues in each of the sequences $\{\lambda_{jk}\}_{j=1}^{\infty}$, for $k = 0, 1$;

Condition A4. the densities g_{jk} are bounded and have a bounded derivative; the kernel K is a symmetric, compactly supported density function with two bounded derivatives; for some $\delta > 0, h = h(n) = O(n^{-\delta})$ and $n^{1-\delta}h^3$ is bounded away from zero as $n \to \infty$; the distributions of $f_{j1}(\xi_j)/f_{j0}(\xi_j)$ are atomless;

Condition A5. the densities g_{jk} are bounded away from zero on any compact interval within their respective support, i.e. for all compact intervals $\mathcal{I} \subset \mathrm{supp}(g_{jk})$, $\inf_{x_j \in \mathcal{I}} g_{jk}(x_j) > 0$ for $k = 0, 1$ and $j \geq 1$.

Condition A1 requires Hölder continuity for processes X, and is a slightly modified version of a condition in Hall & Hosseini-Nasab (2006) and Hall & Hosseini-Nasab (2009). Condition A2 is satisfied if the standardized functional principal components have moments of all orders that are uniformly bounded. In particular, Gaussian processes satisfy Condition A2 since the standardized functional principal components identically follow the standard normal distribution. Condition A3 is standard (Bosq, 2000); here the λ_j are the eigenvalues of the pooled covariance operator. Conditions A4 and A5 are needed for constructing consistent estimators for the density quotients and the classifiers. For the case of completely observed predictors, the following results state the equivalence of the estimated classifiers $I\{\hat{Q}_J(X) \geq 0\}$ and $I\{\hat{Q}_J^R(X) \geq 0\}$ based on the completely observed predictor functions, see the Supplementary Material, and the Bayes classifier using J components $I\{Q_J(X) \geq 0\}$.

Theorem A1. *Under Conditions* 1–2 *and A*1–*A*5*, for any* $\epsilon > 0$ *there exist a set* S *with* $\mathrm{pr}(S) > 1 - \epsilon$ *and a sequence* $J = J(n, \epsilon) \to \infty$ *such that* $\mathrm{pr}(S \cap [I\{\hat{Q}_J(X) \geq 0\} \neq I\{Q_J(X) \geq 0\}]) \to 0$ *as* $n \to \infty$.

Theorem A2. *Under Conditions* 1–2 *and A*1–*A*5*, for any* $\epsilon > 0$ *there exist a set* S *with* $\mathrm{pr}(S) > 1 - \epsilon$ *and a sequence* $J = J(n, \epsilon) \to \infty$ *such that* $\mathrm{pr}(S \cap [I\{\hat{Q}_J^R(X) \geq 0\} \neq I\{Q_J(X) \geq 0\}]) \to 0$ *as* $n \to \infty$.

To obtain theoretical results under pre-smoothing, we require Conditions A6–A9, which parallel assumptions (B2)–(B4) in the Supplementary Material of Kong et al. (2016):

Condition A6. for $k = 0, 1$, $X^{(k)}$ is twice continuously differentiable on $\mathcal{T}$ with probability 1, and $\int_{\mathcal{T}} E\{d^2X^{(k)}(t)/dt^2\}\,\mathrm{d}t < \infty$;

Condition A7. for $i = 1, \ldots, n$ and $k = 0, 1$, $\{t_{ikl} : l = 1, \ldots, m_{ik}\}$ are considered deterministic and ordered increasingly. There exist design densities $u_{ik}(t)$ which are uniformly smooth over i satisfying $\int_{\mathcal{T}} u_{ik}(t)\,\mathrm{d}t = 1$ and $0 < c_1 < \inf_i\{\inf_{t\in T} u_{ik}(t)\} < \sup_i\{\sup_{t\in T} u_{ik}(t)\} < c_2 < \infty$ that generate t_{ikl} according to $t_{ikl} = U_{ik}^{-1}\{l/(m_{ik}+1)\}$, where U_{ik}^{-1} is the inverse of $G_{ik}(t) = \int_{-\infty}^{t} u_{ik}(s)\,\mathrm{d}s$;

Condition A8. for each $k = 0, 1$, there exist a common sequence of bandwidth w such that $0 < c_1 < \inf_i w_{ik}/w < \sup_i w_{ik}/w < c_2 < \infty$, where w_{ik} is the bandwidth for smoothing $\tilde{X}_i^{(k)}$. The kernel function K_0 for local linear smoothing is twice continuously differentiable and compactly supported;

Condition A9. let $\delta_{ik} = \sup\{t_{ik,l+1} - t_{ikl} : l = 0, \ldots, m_{ik}\}$ and $m = m(n) = \inf_{i=1,\ldots,n;\, k=0,1} m_{ik}$. Then $\sup_{i,k} \delta_{ik} = O(m^{-1})$, w is of order $m^{-1/5}$, and $mh^5 \to \infty$, where h is the common bandwidth multiplier in the kernel density estimator.

To obtain asymptotic perfect classification properties, we impose the following conditions on the standardized functional principal components:

Condition A10. the densities $g_{j0}(\cdot)$ and $g_{j1}(\cdot)$ are uniformly bounded for all $j \geq 1$;

Condition A11. the first four moments of $\xi_j/\lambda_{j0}^{1/2}$ under Π_0 and those of $(\xi_j - \mu_j)/\lambda_{j1}^{1/2}$ under Π_1 are uniformly bounded for all $j \geq 1$.

References

Alonso, A. M., Casado, D. & Romo, J. (2012). Supervised classification for functional data: A weighted distance approach. *Computational Statistics and Data Analysis* **56**, 2334–2346.

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

Baíllo, A., Cuevas, A. & Cuesta-Albertos, J. A. (2011). Supervised classification for a family of Gaussian functional models. *Scandinavian Journal of Statistics* **38**, 480–498.

Benko, M., Härdle, W. & Kneip, A. (2009). Common functional principal components. *The Annals of Statistics* **37**, 1–34.

Berquin, P. C., Giedd, J. N., Jacobsen, L. K., Hamburger, S. D., Krain, A. L., Rapoport, J. L. & Castellanos, F. X. (1998). Cerebellum in attention-deficit hyperactivity disorder: a morphometric MRI study. *Neurology* **50**, 1087–1093.

Biau, G., Bunea, F. & Wegkamp, M. (2005). Functional classification in Hilbert spaces. *IEEE Transactions on Information Theory* **51**, 2163–2172.

Biau, G., Cérou, F. & Guyader, A. (2010). Rates of convergence of the functional k-nearest neighbor estimate. *IEEE Transactions on Information Theory* **56**, 2034–2040.

Boente, G., Rodriguez, D. & Sued, M. (2010). Inference under functional proportional and common principal component models. *Journal of Multivariate Analysis* **101**, 464–475.

Bosq, D. (2000). *Linear Processes in Function Spaces: Theory and Applications*. New York: Springer-Verlag.

Cérou, F. & Guyader, A. (2006). Nearest neighbor classification in infinite dimension. *ESAIM: Probability and Statistics* **10**, 340–355.

Chiou, J.-M. & Li, P.-L. (2008). Correlation-based functional clustering via subspace projection. *Journal of the American Statistical Association* **103**, 1684–1692.

Coffey, N., Harrison, A., Donoghue, O. & Hayes, K. (2011). Common functional principal components analysis: A new approach to analyzing human movement data. *Human Movement Science* **30**, 1144–1166.

Coffey, N., Hinde, J. & Holian, E. (2014). Clustering longitudinal profiles using P-splines and mixed effects models applied to time-course gene expression data. *Computational Statistics and Data Analysis* **71**, 14–29.

Delaigle, A. & Hall, P. (2010). Defining probability density for a distribution of random functions. *The Annals of Statistics* **38**, 1171–1193.

Delaigle, A. & Hall, P. (2011). Theoretical properties of principal component score density estimators in functional data analysis. *Sankt-Peterburgskii Universitet. Vestnik. Seriya 1. Matematika, Mekhanika, Astronomiya* **2011**, 55–69.

Delaigle, A. & Hall, P. (2012). Achieving near perfect classification for functional data. *Journal of the Royal Statistical Society: Series B* **74**, 267–286.

Delaigle, A. & Hall, P. (2013). Classification using censored functional data. *Journal of the American Statistical Association* **108**, 1269–1283.

Ferraty, F. & Vieu, P. (2003). Curves discrimination: a nonparametric functional approach. *Computational Statistics and Data Analysis* **44**, 161–173.

Flury, B. N. (1984). Common principal components in k groups. *Journal of the American Statistical Association* **79**, 892–898.

Francisco-Fernández, M., Tarrío-Saavedra, J., Mallik, A. & Naya, S. (2012). A comprehensive classification of wood from thermogravimetric curves. *Chemometrics and Intelligent Laboratory Systems* **118**, 159–172.

Galeano, P., Joseph, E. & Lillo, R. E. (2015). The Mahalanobis distance for functional data with applications to classification. *Technometrics* **57**, 281–291.

Hall, P. & Hosseini-Nasab, M. (2006). On properties of functional principal components analysis. *Journal of the Royal Statistical Society: Series B* **68**, 109–126.

Hall, P. & Hosseini-Nasab, M. (2009). Theory for high-order bounds in functional principal components analysis. *Mathematical Proceedings of the Cambridge Philosophical Society* **146**, 225–256.

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

Hyvärinen, A. & Oja, E. (2000). Independent component analysis: algorithms and applications. *Neural Networks* **13**, 411–430.

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

Kong, D., Xue, K., Yao, F. & Zhang, H. H. (2016). Partially functional linear regression in high dimensions. *Biometrika* **103**, 147–159.

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

Li, W. V. & Linde, W. (1999). Approximation, metric entropy and small ball estimates for gaussian measures. *The Annals of Probability* **27**, 1556–1578.

Nadaraya, E. (1964). On estimating regression. *Theory of Probability and Its Applications* **9**, 141–142.

Parzen, E. (1962). On estimation of a probability density function and mode. *Annals of Mathematical Statistics* **33**, 1065–1076.

Preda, C. & Saporta, G. (2005). PLS regression on a stochastic process. *Computational Statistics and Data Analysis* **48**, 149–158.

Preda, C., Saporta, G. & Lévéder, C. (2007). PLS classification of functional data. *Computational Statistics* **22**, 223–235.

Rasmussen, C. & Williams, C. (2006). *Gaussian Processes for Machine Learning*. Bognor Regis: University Press Group Limited.

Rosenblatt, M. (1956). Remarks on some nonparametric estimates of a density function. *Annals of Mathematical Statistics* **27**, 832–837.

Scott, D. W. (2015). *Multivariate Density Estimation: Theory, Practice, and Visualization*. Hoboken: John Wiley & Sons.

Sillverman, B. W. & Jones, M. C. (1989). An important contribution to nonparametric discriminant analysis and density estimation: commentary on Fix and Hodges (1951). *International Statistical Review* **57**, 233–247.

Silverman, B. W. (1986). *Density Estimation for Statistics and Data analysis*. London: Chapman & Hall.

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

Spellman, P. T., Sherlock, G., Zhang, M. Q., Iyer, V. R., Anders, K., Eisen, M. B., Brown, P. O., Botstein, D. & Futcher, B. (1998). Comprehensive identification of cell cycle-regulated genes of the yeast saccharomyces cerevisiae by microarray hybridization. *Molecular Biology of the Cell* **9**, 3273–3297.

Tzourio-Mazoyer, N., Landeau, B., Papathanassiou, D., Crivello, F., Etard, O., Delcroix, N., Mazoyer, B. & Joliot, M. (2002). Automated anatomical labeling of activations in SPM using a macroscopic anatomical parcellation of the MNI MRI single-subject brain. *NeuroImage* **15**, 273–289.

Wang, J.-L., Chiou, J.-M. & Müller, H.-G. (2016). Functional data analysis. *Annual Review of Statistics and its Application* **3**, 257–295.

Wang, K., Liang, M., Wang, L., Tian, L., Zhang, X., Li, K. & Jiang, T. (2007). Altered functional connectivity in early Alzheimer's disease: A resting-state fMRI study. *Human Brain Mapping* **28**, 967–978.

Wang, X. & Qu, A. (2014). Efficient classification for longitudinal data. *Computational Statistics and Data Analysis* **78**, 119–134.

Watson, G. S. (1964). Smooth regression analysis. *Sankhyā Series A* **26**, 359–372.

Wegman, E. J. (1972). Nonparametric probability density estimation: I. a summary of available methods. *Technometrics* **14**, 533–546.

Wu, Y. & Liu, Y. (2013). Functional robust support vector machines for sparse and irregular longitudinal data. *Journal of Computational and Graphical Statistics* **22**, 379–395.

Yao, F., Wu, Y. & Zou, J. (2016). Probability-enhanced effective dimension reduction for classifying sparse functional data (with discussion). *Test* **25**, 1–58.

Zhu, H., Brown, P. J. & Morris, J. S. (2012). Robust classification of functional and quantitative image data using functional mixed models. *Biometrics* **68**, 1260–1268.

Zhu, H., Vannucci, M. & Cox, D. D. (2010). A Bayesian hierarchical model for classification with selection of functional predictors. *Biometrics* **66**, 463–473.