Biometrika (2017), **104**, 3, *pp.* 545–560
Printed in Great Britain
doi: 10.1093/biomet/asx024
Advance Access publication 23 May 2017

Optimal Bayes classifiers for functional data and density ratios

BY XIONGTAO DAI, HANS-GEORG M LLER

Department of Statistics, University of California, One Shields Avenue, Davis, California 95616, U.S.A.

dai@ucdavis.edu hgmueller@ucdavis.edu

AND FANG YAO

Department of Probability & Statistics, School of Mathematical Sciences and Center for Statistical Sciences, Peking University, Beijing 100871, China

fyao@math.pku.edu.cn

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 onto a sequence of eigenfunctions that are common to the groups to be classified. The density ratios are then factorized into density ratios of individual projection scores, reducing the classification problem to obtain-

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 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; Cerou & Guyader, 2006; Biau et al., 2010), Bayesian methods (Wang et al., 2007), logistic regression (Araki et al., 2009), and 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and this provides one of the major motivations for developing methods of nonparametric density estimation; see an unpublished 1951 technical report by E. Fix and J. L. Hodges Jr, commented on by Rosenblatt (1956), Parzen (1962), Wegman (1972) and Silverman & Jones (1989). However, for multiple predictors, unrestricted nonparametric approaches are subject to the curse of dimensionality (Scott, 2015). This leads to very slow rates of convergence in estimating the nonparametric densities for dimensions higher 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 a small radius is so low that densities do not even exist in most cases (Li & Linde, 1999; Delaigle & Hall, 2010).

Therefore, 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 by Delaigle & Hall (2012), who considered two Gaussian populations with equal covariance using a functional linear discriminant, in analogy to the linear discriminant, that corresponds 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, here we aim to construct 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, and then construct density ratios through products of the density ratios of the projection scores. The result 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 by Delaigle & Hall (2012) to refer to conditions where the misclassification rate converges to zero as an increasing number of projection scores is used, and we use the term in the same sense here. Perfect classification in the Gaussian case requires there to be 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 made on a dense grid, whereas previous approaches were based on the less realistic assumption of fully observed trajectories without noise.

2. Functional Bayes classifiers

We consider the situation in which the observed data come from a common distribution (X, Y), where X is a fully observed square-integrable random function in $L^2(\mathcal{T})$, with $\mathcal{T}$ being 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 $(k = 0, 1)$, i.e., $X^{(k)}$ 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. Let g_0 and g_1 denote the conditional densities of the functional observations X when conditioning on the group labels 0 and 1, respectively, 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 (Ba llo et al., 2011; Berrendero et al., 2016) we consider a sequence of approximations with an 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 $\xi_j = \int_{\mathcal{T}} X(t)\psi_j(t)\,\mathrm{d}t$ $(j = 1, 2, \ldots)$. 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,

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

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

Since estimating the joint density of $(\xi_1, \ldots, \xi_J)$ 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 that the auto-covariances of the stochastic processes which generate the observed data have the same ordered eigenfunctions for both populations. Denote the mean functions by $\mu_k(t) = E\{X^{(k)}(t)\}$ and the covariance functions by $G_k(s, t) = \mathrm{cov}\{X^{(k)}(s), X^{(k)}(t)\}$, 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 that $G_k(s,t)$ is continuous, by Mercer's theorem (see, e.g., Bosq, 2000) we have

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

where $\lambda_{1k} \geqslant \lambda_{2k} \geqslant \cdots \geqslant 0$ 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 between the sets of eigenfunctions, ignoring their order, in which case one can reorder the eigenfunctions and eigenvalues so 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 $\mathrm{cov}(\xi_j, \xi_l) = 0$ 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 to assume that the projection scores are independent under both populations, whence the densities in (2) factorize 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 $Q_J(x) > 0$. Because of the zero-divided-by-zero problem, (3) is defined only on a set $\mathcal{X}$ with $\mathrm{pr}(X \in \mathcal{X}) = 1$, 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 § 5·2.

When predictor processes X are Gaussian for group $k = 0$ or 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 $\sum_{j\geqslant 1}(\mu_{j1} - \mu_{j0})^2/\lambda_{j0} < \infty$ and $\sum_{j\geqslant 1}(\lambda_{j0}/\lambda_{j1} - 1)^2 < \infty$, or else 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 moreover $\lambda_{j0} = \lambda_{j1}$ ($j = 1, 2, \ldots$), 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 might expect better performance than Gaussian-based functional classifiers when the distributions are non-Gaussian. This is borne out by the simulation results in § 5·2. The densities of the projection scores can be estimated nonparametrically by kernel density estimation (Silverman, 1986) as described in § 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 that 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 functions 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

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

where $u \in \mathbb{R}$ and $h_{jk} = h\hat{\lambda}_{jk}^{1/2}$ are bandwidths adapted to the variance of the jth projection score (see § § 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 the kernel density estimate is therefore

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

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 independent 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 can obtain the same

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 ϕ_j are the common eigenfunctions and λ_{jk} 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 the case if predictor processes satisfy the more restrictive Gaussian assumption, for example; see § 6 for further discussion. Additional assumptions are provided in the Appendix.

THEOREM 1. *Under Conditions 1, 2 and A1–A9 in the Appendix, 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* $\mathrm{pr}(S \cap [I\{\tilde{Q}_J(X) \geqslant 0\} \neq I\{Q_J(X) \geqslant 0\}]) \to 0$ *as* $n \to \infty$.

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

The next result states 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, A10 and A11, the Bayes classifier* $I\{Q_J(x) \geqslant 0\}$ *achieves perfect classification as* $J \to \infty$ *if* $\sum_{j\geqslant 1}(r_j - 1)^2 = \infty$ *or* $\sum_{j\geqslant 1} m_j^2 = \infty$.

This theorem extends previous results on perfect classification, such as those in Delaigle & Hall (2012, 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 methods involve 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 scatterplots of the pooled estimated scores and group labels. For the kernel estimates we use a Gaussian kernel for which the bandwidth multiplier h is chosen by ten-fold crossvalidation, 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 in Delaigle & Hall (2013) for example, 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 J of projections used is selected by ten-fold crossvalidation, jointly with the selection of h for the nonparametric methods.

5·2. *Simulation results*

We illustrate the proposed Bayes classifiers in three simulation settings that involve varying distributions and dependency assumptions for the projection scores. In 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 zero and variance λ_{jk}, which are generated under two distribution scenarios: Scenario A, where the A_{ijk} are normally distributed; and Scenario B, where the A_{ijk} are centred exponentially distributed. In 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 $\chi^2_{30}/30$.

In each setting, we generate n training samples, each having $1/2$ chance of being from Π_0 or Π_1, and we let ϕ_j be the jth function in the Fourier basis, where $\phi_1(t) = 1$, $\phi_2(t) = \sqrt{2}\cos(2\pi t)$, $\phi_3(t) = \sqrt{2}\sin(2\pi t)$, and so on, for $t \in [0, 1]$. We set $\mu_0(t) = 0$ and set $\mu_1(t) = 0$ or t for the same- or different-mean scenarios, respectively. The variances of A_{ijk} under Π_0 are $\lambda_{j0} = \exp(-j/3)$, and those under Π_1 are $\lambda_{j1} = \exp(-j/3)$ or $\exp(-j/2)$ $(j = 1, \ldots, 50)$ for the same- or different-variance scenarios, respectively. The random functions are sampled at 51 equally spaced time-points from 0 to 1, with small measurement errors in the form of independent Gaussian noise with mean zero 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 of comparing the predictive performance of the following functional classification methods: the centroid method (Delaigle & Hall, 2012); the proposed nonparametric Bayes classifier in three versions, where estimation is based on Gaussian densities, nonparametric densities or nonparametric regression, as discussed in § 5·1; logistic regression; the functional quadratic discriminant as in Galeano et al. (2015); and Gaussian process logistic regression (Rasmussen & Williams, 2006) with squared exponential function and automatic relevance determination. The functional quadratic discriminant was never the winner in any scenario of our simulation study, so we omit it from the tables. We show in Table 1 the results corresponding to presmoothing the predictors by local linear smoothers with crossvalidation bandwidth choice. Since all classifiers show improved performance when presmoothing of the predictor functions takes place, the results obtained without presmoothing are relegated to the Supplementary Material.

For Scenario A, the proposed nonparametric Bayes classifiers show superior performance in those settings where covariance differences in the populations are present, whereas the logistic methods work best in those cases where the differences lie exclusively in the mean. This is because the proposed nonparametric Bayes classifiers take into account both mean and covariance

Table 1. *Misclassification rates (%), with standard errors in parentheses, for presmoothed predictors in the three 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)

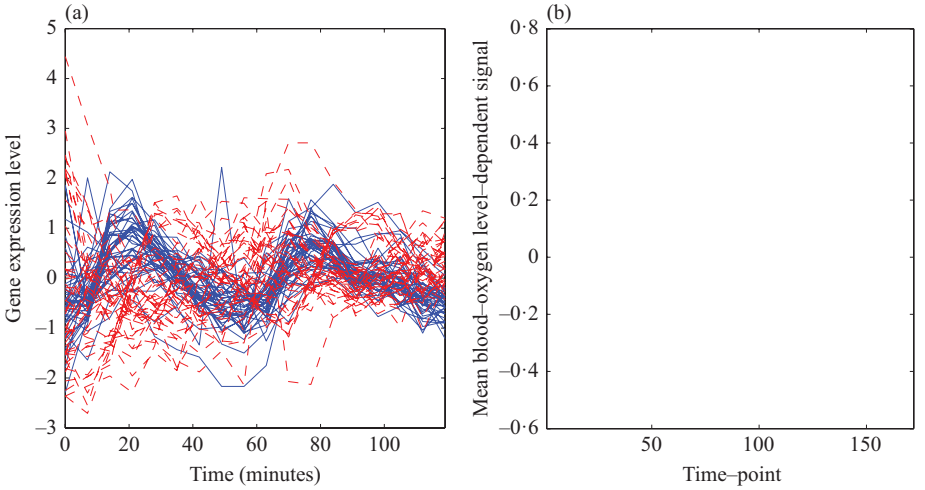
Centroid, the method of Delaigle & Hall (2012); Gaussian, NPD and NPR, the Gaussian, nonparametric density and nonparametric regression implementations of the proposed Bayes classifiers, respectively; Logistic, functional logistic regression; GP logistic, Gaussian process logistic regression.

differences between the populations. In 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 than the other scenarios because of the dependency in the projection scores, which violates Condition 2. Nevertheless, the proposed classifiers outperform the others 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 when the sample size is large, owing to its capacity to handle dependent predictors.

5·3. *Data illustrations*

We present four real-data examples to illustrate the performance of the proposed Bayes classifiers for functional data. We presmoothed the yeast data by a local linear smoother with crossvalidation bandwidth choice, since the original observations are quite noisy, as can be seen from Fig. 1; for the wine dataset and the attention deficit hyperactivity disorder dataset we used the curves as provided in the data, which were already pre-processed and smooth. Following the procedure described in Benko et al. (2009), we tested whether the eigenspaces generated by the



to bimodal while the other density is not. The nonparametric implementations of the proposed Bayes estimators based on nonparametric regression and on nonparametric density estimation are capable of reflecting such shape differences and therefore outperform the classifiers based on Gaussian assumptions.

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 lies 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 datasets. 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 set.

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 processes is closed with respect to componentwise transformation. Processes generated by a nonlinear 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 restricted 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 $k \neq k^*$.

Acknowledgement

This work was supported by the U.S. National Science Foundation and the Natural Sciences and Engineering Research Council of Canada. We thank the reviewers for helpful comments. Yao is also affiliated with the Department of Statistical Sciences, University of Toronto.

Supplementary material

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

Appendix

Assumptions and additional results

For simplicity of presentation, throughout all proofs we adopt the simplifying assumptions mentioned in the first paragraph of § 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 $\hat{Q}_J(x)$ in (6) is asymptotically equivalent to $Q_J(x)$ 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 of $(\xi_j - \mu_j)/\lambda_{j1}^{1/2}$ when $k = 1$, let $\hat{g}_{jk}$ be the kernel density estimates of g_{jk} using the estimated functional principal components, and let $\bar{g}_{jk}$ 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 a detailed proof given in Delaigle & Hall (2011); our derivations utilize their result.

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

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 $r \geqslant 1$, $\lambda_{jk}^{-r} E_{\Pi_k}[\int_{\mathcal{T}}\{X(t) - E_{\Pi_k}X(t)\}\phi_j(t)\,\mathrm{d}t]^{2r}$ 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 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 $n \to \infty$.

Condition A5. The densities g_{jk} are bounded away from zero on any compact interval within their respective supports; that is, 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 $j \geqslant 1$.

Condition A1 requires Hölder continuity for processes X, and is a slightly modified version of a condition in Hall & Hosseini-Nasab (2006, 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 of the density quotients. For the case of completely observed predictors, the following results state the equivalence of the estimated classifiers $I\{\hat{Q}_J(X) \geqslant 0\}$ and $I\{\hat{Q}_J^R(X) \geqslant 0\}$ based on the completely observed predictor functions, see the Supplementary Material, and the Bayes classifier using J components, $I\{Q_J(X) \geqslant 0\}$.

THEOREM A1. *Under Conditions 1, 2 and A1–A5, for any $\epsilon > 0$ there exists a set S with $\mathrm{pr}(S) > 1 - \epsilon$ and a sequence $J = J(n, \epsilon) \to \infty$ such that $\mathrm{pr}(S \cap [I\{\hat{Q}_J(X) \geqslant 0\} \neq I\{Q_J(X) \geqslant 0\}]) \to 0$ as $n \to \infty$.*

THEOREM A2. *Under Conditions 1, 2 and A1–A5, for any $\epsilon > 0$ there exists a set S with $\mathrm{pr}(S) > 1 - \epsilon$ and a sequence $J = J(n, \epsilon) \to \infty$ such that $\mathrm{pr}(S \cap [I\{\hat{Q}_J^R(X) \geqslant 0\} \neq I\{Q_J(X) \geqslant 0\}]) \to 0$ as $n \to \infty$.*

To obtain theoretical results under presmoothing, we require Conditions A6–A9 below, 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 $\int_{\mathcal{T}} E\{\mathrm{d}^2 X^{(k)}(t)/\mathrm{d}t^2\}\,\mathrm{d}t < \infty$.

Condition A7. For $i = 1, \ldots, n$ and $k = 0, 1$, $\{t_{ikl} : l = 1, \ldots, m_{ik}\}$ are considered deterministic and arranged in increasing order. There exist design densities $u_{ik}(t)$, uniformly smooth over i and satisfying $\int_{\mathcal{T}} u_{ik}(t)\,\mathrm{d}t = 1$ and $0 < c_1 < \inf_i\{\inf_{t \in T} u_{ik}(t)\} < \sup_i\{\sup_{t \in T} u_{ik}(t)\} < c_2 < \infty$, which 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 exists a common sequence of bandwidths 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 define $m = m(n) = \inf_{i=1,\ldots,n;\, k=0,1} m_{ik}$. 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 $j \geqslant 1$.

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 $j \geqslant 1$.

REFERENCES

ALONSO, A. M., CASADO, D. & ROMO, J. (2012). Supervised classification for functional data: A weighted distance approach. *Comp. Statist. Data Anal.* **56**, 2334–46.

ARAKI, Y., KONISHI, S., KAWANO, S. & MATSUI, H. (2009). Functional logistic discrimination via regularized basis expansions. *Commun. Statist.* A **38**, 2944–57.

Ba´llo, A., Cuevas, A. & Cuesta-Albertos, J. A. (2011). Supervised classification for a family of Gaussian functional models. *Scand. J. Statist.* **38**, 480–98.

Benko, M., H rdle, W. & Kneip, A. (2009). Common functional principal components. *Ann. Statist.* **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**, 1087–93.

Berrendero, J. R., Cuevas, A. & Torrecilla, J. L. (2016). On the use of reproducing kernel Hilbert spaces in functional classification. *arXiv:* 1507.04398v3.

Biau, G., Bunea, F. & Wegkamp, M. (2005). Functional classification in Hilbert spaces. *IEEE Trans. Info. Theory* **51**, 2163–72.

Biau, G., C rou, F. & Guyader, A. (2010). Rates of convergence of the functional *k*-nearest neighbor estimate. *IEEE Trans. Info. Theory* **56**, 2034–40.

Boente, G., Rodriguez, D. & Sued, M. (2010). Inference under functional proportional and common principal component models. *J. Mult. Anal.* **101**, 464–75.

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

C rou, F. & Guyader, A. (2006). Nearest neighbor classification in infinite dimension. *ESAIM Prob. Statist.* **10**, 340–55.

Chiou, J.-M. & Li, P.-L. (2008). Correlation-based functional clustering via subspace projection. *J. Am. Statist. Assoc.* **103**, 1684–92.

Coffey, N., Harrison, A., Donoghue, O. & Hayes, K. (2011). Common functional principal components analysis: A new approach to analyzing human movement data. *Hum. Movement Sci.* **30**, 1144–66.

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

Delaigle, A. & Hall, P. (2010). Defining probability density for a distribution of random functions. *Ann. Statist.* **38**, 1171–93.

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. *J. R. Statist. Soc.* B **74**, 267–86.

Delaigle, A. & Hall, P. (2013). Classification using censored functional data. *J. Am. Statist. Assoc.* **108**, 1269–83.

Ferraty, F. & Vieu, P. (2003). Curves discrimination: A nonparametric functional approach. *Comp. Statist. Data Anal.* **44**, 161–73.

Flury, B. N. (1984). Common principal components in *k* groups. *J. Am. Statist. Assoc.* **79**, 892–8.

Francisco-Fern ndez, M., Tarr´o-Saavedra, J., Mallik, A. & Naya, S. (2012). A comprehensive classification of wood from thermogravimetric curves. *Chemomet. Intel. Lab. Syst.* **118**, 159–72.

Galeano, P., Joseph, E. & Lillo, R. E. (2015). The Mahalanobis distance for functional data with applications to classification. *Technometrics* **57**, 281–91.

Hall, P. & Hosseini-Nasab, M. (2006). On properties of functional principal components analysis. *J. R. Statist. Soc.* B **68**, 109–26.

Hall, P. & Hosseini-Nasab, M. (2009). Theory for high-order bounds in functional principal components analysis. *Math. Proc. Camb. Phil. Soc.* **146**, 225–56.

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

Hyv rinen, A. & Oja, E. (2000). Independent component analysis: Algorithms and applications. *Neural Networks* **13**, 411–30.

Kalivas, J. H. (1997). Two data sets of near infrared spectra. *Chemomet. Intel. Lab. Syst.* **37**, 255–9.

Kong, D., Xue, K., Yao, F. & Zhang, H. H. (2016). Partially functional linear regression in high dimensions. *Biometrika* **103**, 147–59.

Leng, X. & M ller, 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. *Ann. Prob.* **27**, 1556–78.

Nadaraya, E. (1964). On estimating regression. *Theory Prob. Appl.* **9**, 141–2.

Parzen, E. (1962). On estimation of a probability density function and mode. *Ann. Math. Statist.* **33**, 1065–76.

Preda, C. & Saporta, G. (2005). PLS regression on a stochastic process. *Comp. Statist. Data Anal.* **48**, 149–58.

Preda, C., Saporta, G. & L v der, C. (2007). PLS classification of functional data. *Comp. Statist.* **22**, 223–35.

Rasmussen, C. E. & Williams, C. K. I. (2006). *Gaussian Processes for Machine Learning*. Cambridge, Massachusetts: MIT Press.

Rosenblatt, M. (1956). Remarks on some nonparametric estimates of a density function. *Ann. Math. Statist.* **27**, 832–7.

SCOTT, D. W. (2015). *Multivariate Density Estimation: Theory, Practice, and Visualization*. Hoboken, New Jersey: Wiley.

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

SILVERMAN, B. W. & JONES, M. C. (1989). An important contribution to nonparametric discriminant analysis and density estimation: Commentary on Fix and Hodges (1951). *Int. Statist. Rev.* **57**, 233–47.

SONG, J., DENG, W., LEE, H. & KWON, D. (2008). Optimal classification for time-course gene expression data using functional data analysis. *Comp. Biol. Chem.* **32**, 426–32.

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. *Molec. Biol. Cell* **9**, 3273–97.

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**, 273–89.

WANG, J.-L., CHIOU, J.-M. & MÜLLER, H.-G. (2016). Functional data analysis. *Ann. Rev. Statist. Applic.* **3**, 257–95.

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. *Hum. Brain Map.* **28**, 967–78.

WANG, X. & QU, A. (2014). Efficient classification for longitudinal data. *Comp. Statist. Data Anal.* **78**, 119–34.

WATSON, G. S. (1964). Smooth regression analysis. *Sankhyā* A **26**, 359–72.

WEGMAN, E. J. (1972). Nonparametric probability density estimation: I. A summary of available methods. *Technometrics* **14**, 533–46.

WU, Y. & LIU, Y. (2013). Functional robust support vector machines for sparse and irregular longitudinal data. *J. Comp. Graph. Statist.* **22**, 379–95.

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**, 1260–8.

ZHU, H., VANNUCCI, M. & COX, D. D. (2010). A Bayesian hierarchical model for classification with selection of functional predictors. *Biometrics* **66**, 463–73.

[*Received on* 29 *November* 2016. *Editorial decision on* 6 *March* 2017]