Statistica Sinica

COPULA-BASED FUNCTIONAL BAYES CLASSIFICATION WITH PRINCIPAL COMPONENTS AND PARTIAL LEAST SQUARES

Wentian Huang, David Ruppert

Cornell University

Abstract: We present a new functional Bayes classifier that uses principal component (PC) or partial least squares (PLS) scores from the common (i.e. pooled) covariance function, that is, the covariance function marginalized over groups. When the groups have different covariance functions, the PC or PLS scores need not be independent or even uncorrelated. We use copulas to model the dependence. Our method is semiparametric; the marginal densities are estimated nonparametrically using kernel smoothing, and the copula is modeled parametrically. We focus on Gaussian and t-copulas, but other copulas can be used. The strong performance of our methodology is demonstrated through simulation, real-data examples, and asymptotic properties.

Key words and phrases: Asymptotic theory, Bayes classifier, functional data, perfect classification, rank correlation, semiparametric model

1. Introduction

Functional classification, where the features are continuous functions on a compact interval, is receiving increasing interest in fields such as chemometrics, medicine, economics, and environmental science. James and Hastie (2001) extended the linear discriminant analysis (LDA) to functional data (FLDA), including the case where the curves are partially observed. James (2002) proposed a functional version of the generalized linear model (FGLM), including functional logistic regression. Thereafter, the FGLM was further researched by, among others, Müller et al. (2005), Li et al. (2010), Zhu et al. (2010), McLean et al. (2014), and Shang et al. (2015). Aside from the FGLM, other classifiers have also been studied. Rossi and Villa (2006) applied support vector machines (SVM) to classify infinite-dimensional data. Cuevas et al. (2007) explored the classification of functional data based on data depth. Li and Yu (2008) suggested a functional segmented discriminant analysis combining an LDA and an SVM, and Cholaquidis et al. (2016) proposed a nonlinear aggregation classifier.

However, certain issues remain. Current methods, such as the FLDA, SVM, and functional centroid classifier (Delaigle and Hall (2012)), distinguish groups by the differences between their functional means. They achieve satisfactory results when the location difference is the dominant feature distinguishing classes, but functional data provide more information than just group means. For example, Fig. 1 from the example in Section 4.1 compares the mean and standard deviation functions of raw and smoothed fractional anisotropy (FA) measured along the corpus callosum (cca) of 141 subjects, 99 with multiple sclerosis (MS) and 42 without. The disparity between the group standard deviations in panel (c) provides additional information that can identify MS patients. As shown in Section 4.1, the LDA and centroid classifiers fail to capture this information, and have higher misclassification rates
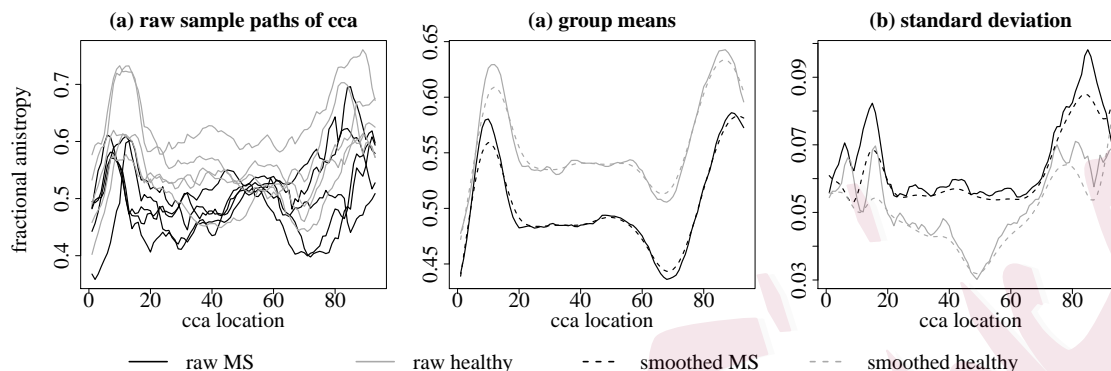
than the classifiers we propose.

Figure 1: Panel (a) shows profiles of FA, five each of cases and controls, and panels (b) and (c) show the group means and standard deviations. Compared to the controls, the MS group has a lower mean and a higher standard deviation.

Both parametric and nonparametric methods have drawbacks in classifying functional data. Parametric models, such as linear and quadratic discriminant analysis, are popular in functional classification, especially because nonparametric methods are likely to encounter the curse of dimensionality. However, parametric methods can cast rigid assumptions on the class boundaries (Li and Yu (2008)). Our interest is in methods that avoid stringent assumptions on the data. Dai et al. (2017) proposed a nonparametric Bayes classifier, assuming that the subgroups share the same sets of eigenfunctions, and that the scores projected on them are independent. With these assumptions and the definition of the density of random functions proposed by Delaigle and Hall (2010), the joint densities of the truncated functional data can be estimated using a *univariate* kernel density estimation (KDE). The Bayes rules estimated this way avoid the curse of dimensionality, but require that the groups have equal sets of eigenfunctions and independent scores.

We propose new semiparametric Bayes classifiers. We project the functions onto the eigenfunctions of the pooled covariance function, that is, the covariance function marginalized over groups. These eigenfunctions can be estimated by applying a functional principal

components analysis (fPCA) to the combined groups. The projections will not be independent or even uncorrelated, unless these common eigenfunctions are also the eigenfunctions of the group-specific covariance functions, an assumption not likely to hold in many situations. For instance, in Section 4 we discuss two real-data examples, and include a comparison of their group eigenfunctions in the Supplementary Material (Fig. S4 and Fig. S8). Both cases appear to violate the equal eigenfunction assumption. We estimate the marginal density of the projected scores using a univariate KDE, as in Dai et al. (2017), and model the association between the scores using a parametric copula. Our semiparametric methodology avoids the restricted range of applications imposed by the assumption of equal group-specific eigenfunctions. It also avoids the curse of dimensionality that a multivariate nonparametric density estimation would entail.

In addition to the principal components (PC) basis, we also consider a partial least squares (PLS) projection basis. PLS has attracted recent attention owing to its effectiveness in prediction and classification problems with high-dimensional and functional data. Preda et al. (2007) discuss a functional LDA combined with PLS. Delaigle and Hall (2012) mention the potential advantage of PLS scores in their functional centroid classifier, when the difference between the group means does not lie primarily in the space spanned by the first few eigenfunctions. We find that PLS scores can be more efficient than PC scores in capturing group mean differences.

This study contributes to the literature in two ways. In our numerical results, the new method shows improved prediction accuracy and strength in dimension reduction, and extends the functional Bayes classification to multiclass classification. In the theoretical analysis, several new conditions are added for the functional data to achieve asymptotic optimality.

These conditions are required because of the unequal group-specific eigenfunctions. More-over, we propose asymptotic sparsity assumptions on the inverse of the copula correlations in our new method, following the design of Yuan (2010) and Liu et al. (2012) for high-dimensional data. We also build a new theorem that uses the special copula structure to achieve asymptotic perfect classification.

In Section 2, we introduce our model and the copula-based functional Bayes classifiers. Section 3 contains a comprehensive simulation study comparing our methods with existing classifiers on both binary and multiclass problems. Section 4 uses two real-data examples to show the strength of our classifiers in terms of accuracy and dimension reduction with respect to data size. In Section 5, we discuss the asymptotic properties of our classifiers. We also establish conditions for our classifiers to achieve perfect classification on data generated by Gaussian and non-Gaussian processes. Finally, in Section 6, we discuss future work, including extending the classification to the case where there are multiple functional predictors. Additional results and detailed proofs are provided in the Supplementary Material.

2. Model Setup & Functional Bayes Classifiers with Copulas

2.1 Methodology

Suppose $(X_{i\cdot\cdot}, Y_i)$, $i = 1, \ldots, n$ are independent and identically distributed (i.i.d.) from the joint distribution of (X, Y), where X is a square integrable function over some compact interval $\mathcal{T}$, that is, $X \in \mathcal{L}^2(\mathcal{T})$. Here $Y = 0, 1$ is an indicator of groups Π_0 and Π_1, respectively, and $\pi_k = P(Y = k)$. In addition, $X_{i\cdot k}$, for $i = 1, \ldots, n_k$ and $k = 0, 1$, denotes the ith sample curve of $X_{\cdot\cdot k} = (X|Y = k)$, and $n = \sum_{k=0,1} n_k$. Our goal is to classify a new observation, x.

Note that throughout the paper, we order the index of X by observation counts (i), joint

basis (j), and group labels (k): for curves, $X_{i..}$ denotes the ith observation of the random function X, and $X_{..k}$ is the random function $X|Y = k$. Therefore, $X_{i \cdot k}$ is the ith sample curve of $X_{..k}$. Furthermore, $X_{\cdot j \cdot}$ and $X_{\cdot jk}$ are random variables from projecting X and $X_{..k}$, respectively, onto the jth joint basis function ψ_j, with X_{ijk} the ith observation of $X_{\cdot jk}$.

Dai et al. (2017) extended the Bayes classification from multivariate to functional data: a new curve x is classified into Π_1 if

$$Q(x) = \frac{P(Y = 1|X = x)}{P(Y = 0|X = x)} = \frac{\overline{f}_1(x)\pi_1}{\overline{f}_0(x)\pi_0} \approx \frac{f_1(x_1, \ldots, x_J)\pi_1}{f_0(x_1, \ldots, x_J)\pi_0} > 1, \tag{2.1}$$

where $\overline{f}_k$ is the density of $X_{..k}$ and f_k is the joint density of the scores $X_{\cdot jk}$ on the basis ψ_j, for $1 \le j \le J$.

A key feature of the Bayes classification on functional data is that the classifiers vary with the choice of basis functions ψ_j and with the estimation of f_0, f_1. Dai et al. (2017) built the original functional Bayes classifier (BC), upon two important assumptions. First, the sets of the first J eigenfunctions, $\{\phi_1, \ldots, \phi_J\}$, of the covariance operators G_1 and G_0 of the two groups are equal. Here, $G_k(\phi_j)(t) = \int_{\mathcal{T}} G_k(s,t)\phi_j(s)ds = \lambda_{jk}\phi_j(t)$, $G_k(s,t) = \text{cov}\{X_{..k}(s), X_{..k}(t)\} = \sum_{j=1}^{\infty} \lambda_{jk}\phi_j(s)\phi_j(t)$, and λ_{jk} is the jth eigenvalue in group k. Second, letting $\psi_j = \phi_j$, for $1 \le j \le J$, the J projected scores $X_{\cdot jk} = \langle X_{..k}, \phi_j \rangle$ are independent. Then, with f_{jk} as the marginal density of $X_{\cdot jk}$, the log ratio of $Q(x)$ in Eq.(2.1) becomes

$$\log Q(x) \approx \log Q_J(x) = \log\left(\frac{\pi_1}{\pi_0}\right) + \sum_{j=1}^{J} \log\left\{\frac{f_{j1}(x_j)}{f_{j0}(x_j)}\right\}. \tag{2.2}$$

A classifier that uses Eq.(2.2) avoids the curse of dimensionality and only needs to estimate the marginal densities, f_{jk}. However, as later simulations and examples show, its

performance can degrade if the two aforementioned assumptions are not met. We propose

new semiparametric Bayes classifiers based on copulas that do not require these two assump-

tions, and yet are free from the curse of dimensionality. The theoretical work in Section 5

proves that these classifiers maintain the advantages of BC over a wider range of data dis-

tributions, and are capable of perfect classification when $n \to \infty$ and $J \to \infty$.

2.2 Copula-Based Bayes Classifier with PC

Allowing for possibly unequal group eigenfunctions, the covariance function of group k is

$$G_k(s, t) = \text{cov}\left(X_{..k}(s), X_{..k}(t)\right) = \sum_{j=1}^{\infty} \lambda_{jk} \phi_{jk}(s) \phi_{jk}(t), \ \ k = 0, 1,$$

with $\phi_{1k}, \ldots, \phi_{Jk}$ as the eigenfunctions. For simplicity, we assume the group means are

$E(X|Y = 0) = 0$ and $E(X|Y = 1) = \mu_d$. The joint covariance operator G then has the

kernel $G(s, t) = \pi_1 G_1(s, t) + \pi_0 G_0(s, t) + \pi_1 \pi_0 \mu_d(s) \mu_d(t)$.

As later examples suggest, the unequal group eigenfunction case is common. To ac-

commodate this case, we can project data from both groups onto the same basis functions.

Therefore, we use the eigenfunctions $\phi_1, \ldots, \phi_J$ of G as the basis $\psi_1, \ldots, \psi_J$.

The joint density f_k, for $k = 0, 1$, in Eq.(2.1) allows for potential score correlation and

tail dependency, which we use copulas to model. A copula is a multivariate cumulative

distribution function (CDF) with univariate marginal distributions that are all uniform, and

it characterizes only the dependency between the components; see, for example, Ruppert

and Matteson (2015). Here, we extend its use to truncated scores of functional data.

Let $x_j = \langle x, \phi_j \rangle = \int_{\mathcal{T}} x(t) \phi_j(t) dt$ be the jth projected score of x. The copula function

C_k describes the distribution of the first J scores in Π_k by

$$F_k\left(x_1, \ldots, x_J\right) = C_k\left\{F_{1k}(x_1), \ldots, F_{Jk}(x_J)\right\}, \tag{2.3}$$

$$f_k\left(x_1, \ldots, x_J\right) = c_k\left\{F_{1k}(x_1), \ldots, F_{Jk}(x_J)\right\} f_{1k}(x_1) \cdots f_{Jk}(x_J). \tag{2.4}$$

F_k in Eq.(2.3) is the joint CDF of $X_{\cdot 1k}, \ldots, X_{\cdot Jk}$, and C_k is the CDF of the uniformly distributed variables $F_{1k}(X_{\cdot 1k}), \ldots, F_{Jk}(X_{\cdot Jk})$, where F_{jk} is the univariate CDF of $X_{\cdot jk}$. In Eq.(2.4), the joint density f_k is decomposed into score marginal densities f_{jk} and the copula density c_k for the dependency between the projected scores. Our revised classifier is $\mathbb{1}\left\{\log Q_J^*(x) > 0\right\}$; that is, the new curve x belongs to Π_1 if

$$\log Q_J^*(x) = \log\left(\frac{\pi_1}{\pi_0}\right) + \sum_{j=1}^{J} \log\left\{\frac{f_{j1}(x_j)}{f_{j0}(x_j)}\right\} + \log\left\{\frac{c_1\{F_{11}(x_1), \ldots, F_{J1}(x_J)\}}{c_0\{F_{10}(x_1), \ldots, F_{J0}(x_J)\}}\right\} > 0. \tag{2.5}$$

We also consider situations in which Y has more than two classes. A more general procedure for multiclass classification is described in the Supplementary Material Section S2.

2.3 Choice of Copula and Correlation Estimator

There are a number of approaches to copula estimation. Genest et al. (1995) studied the asymptotic properties of semiparametric estimation in copula models. Chen and Fan (2006) discussed semiparametric copula estimation to characterize the temporal dependence in time series data. Kauermann et al. (2013) estimated the copula density nonparametrically using penalized splines, and Gijbels et al. (2012) applied multivariate kernel density estimation to copulas.

To address the high dimensionality of functional data, we model the copula densities c_1 and c_0 parametrically, and use a kernel estimation for the univariate densities $f_{1k}, \ldots, f_{Jk}$, for $k = 0, 1$. We study the properties of Bayes classification using both Gaussian copulas and t-copulas, denoted by BCG and BCt, respectively. When c_k is modeled by a Gaussian copula in Eq.(2.4), $c_k(\cdot) = c_{G,k}(\cdot | \Omega_{G,k})$, where $c_{G,k}$ is the Gaussian copula density with $J \times J$ correlation matrix $\Omega_{G,k}$. When there is tail dependency between the scores, a t-copula is used: $c_k(\cdot) = c_{t,k}(\cdot | \Omega_{t,k}, \nu_k)$, with $c_{t,k}$ the t-copula density, $\Omega_{t,k}$ the correlation matrix, and ν_k the tail index.

There are several ways to estimate the correlation matrices $\Omega_{G,k}$ or $\Omega_{t,k}$. We use rank correlations, and specifically, Kendall's τ. Kendall's τ between the projected scores of $X_{\cdot \cdot k}$ on the jth and j'th basis is $\rho_\tau (X_{\cdot jk}, X_{\cdot j'k}) = E \left[\operatorname{sign} \left\{ \left(X_{\cdot jk}^{(1)} - X_{\cdot jk}^{(2)} \right) \left(X_{\cdot j'k}^{(1)} - X_{\cdot j'k}^{(2)} \right) \right\} \right]$, $\operatorname{sign}(x) = \mathbb{1}\{x > 0\} - \mathbb{1}\{x < 0\}$, and $X_{\cdot \cdot k}^{(1)}, X_{\cdot \cdot k}^{(2)}$ are i.i.d. samples of $X_{\cdot \cdot k}$. The robustness of the rank correlation and its optimal asymptotic error rate are studied by Liu et al. (2012).

A relationship exists between the (j, j')th entry of the copula correlation Ω_k and Kendall's τ: $\Omega_k^{jj'} = \sin \left(\frac{\pi}{2} \rho_\tau (X_{\cdot jk}, X_{\cdot j'k}) \right)$ for both Gaussian copulas and t-copulas (Kendall (1948); Kruskal (1958); Ruppert and Matteson (2015)). Then, $\Omega_k^{jj'}$ is estimated by Kendall's τ as $\hat{\Omega}_k^{jj'} = \sin \left(\frac{\pi}{2} \hat{\rho}_{\tau,k}^{jj'} \right)$, where

$$\hat{\rho}_{\tau,k}^{jj'} = \frac{2}{n_k (n_k - 1)} \sum_{1 \le i \le i' \le n_k} \operatorname{sign} \left\{ \langle X_{i\cdot k} - X_{i'\cdot k}, \hat{\phi}_j \rangle \langle X_{i\cdot k} - X_{i'\cdot k}, \hat{\phi}_{j'} \rangle \right\}.$$

It is possible that $\hat{\Omega}_k$ is not positive definite, but this problem is easily remedied (Ruppert and Matteson (2015)). Another rank correlation, Spearman's ρ, is similar and is omitted here. In the Supplementary Material S5.4, we show that for Gaussian copulas, the difference

between the log determinant of $\hat{\boldsymbol{\Omega}}_k$, as estimated, and that of $\boldsymbol{\Omega}_k$ is $Op\left(J\sqrt{(\log J)/n}\right)$.

Additionally for t-copulas with $\hat{\boldsymbol{\Omega}}_{t,k}$, we apply a pseudo-maximum likelihood to estimate

the tail parameter $\nu_k > 0$ by maximizing the log copula density

$$\sum_{i=1}^{n_k} \log\left[c_{t,k}\left\{\hat{F}_{1k}\left(X_{i1k}\right),\ldots,\hat{F}_{Jk}\left(X_{iJk}\right)|\hat{\boldsymbol{\Omega}}_{t,k},\nu_k\right\}\right],\text{ with }\hat{F}_{jk}\left(x\right)=\sum_{i=1}^{n_k}\mathbb{1}\left\{X_{ijk}\leq x\right\}/\left(n_k+1\right).$$

Mashal and Zeevi (2002) discuss the maximum pseudo-likelihood estimation of t-copulas, and

apply it to model extreme co-movements of financial assets.

2.4 Marginal Density f_{jk} Estimation

We estimate the marginal density f_{jk} of the projected scores $X_{\cdot jk}$ using a kernel density

estimation: $\hat{f}_{jk}\left(\hat{x}_j\right) = \dfrac{1}{n_k h_{jk}}\sum_{i=1}^{n_k} K\left(\dfrac{\langle x - X_{i\cdot k}, \hat{\phi}_j\rangle}{h_{jk}}\right)$, with K the standard Gaussian kernel,

$\hat{\phi}_j$ the estimated jth joint eigenfunction, $h_{jk} = \hat{\sigma}_{jk}h$ the bandwidth for scores projected on $\hat{\phi}_j$

in group k, $\hat{\sigma}_{jk}$ as the estimated standard deviation of $\sigma_{jk} = \sqrt{\text{Var}\left(X_{\cdot jk}\right)}$, and $\hat{x}_j = \langle x, \hat{\phi}_j\rangle$.

Then, $\log Q_J^*\left(x\right)$ in Eq.(2.5) is estimated by

$$\log \hat{Q}_J^*\left(x\right) = \log\left(\frac{\hat{\pi}_1}{\hat{\pi}_0}\right) + \sum_{j=1}^{J}\log\left\{\frac{\hat{f}_{j1}(\hat{x}_j)}{\hat{f}_{j0}(\hat{x}_j)}\right\} + \log\left\{\frac{\hat{c}_1\{\hat{F}_{11}(\hat{x}_1),\ldots,\hat{F}_{J1}(\hat{x}_J)\}}{\hat{c}_0\{\hat{F}_{10}(\hat{x}_1),\ldots,\hat{F}_{J0}(\hat{x}_J)\}}\right\},$$

where $\hat{c}_k$ is the Gaussian copula or t-copula density with the estimated parameters, and

$\hat{\pi}_k = n_k/n$. Proposition 1 in Section 5 shows that with an additional mild assumption,

when the group eigenfunctions are unequal, $|\hat{f}_{jk}(\hat{x}_j) - f_{jk}(x_j)|$ is asymptotically bounded

at the same rate as when the eigenfunctions are equal. Detailed proofs are included in

Supplementary Material.

2.5 Copula-Based Bayes Classifier with Partial Least Squares

An interesting alternative to using PCs is to use functional partial least squares (FPLS).

FPLS finds directions that maximize the covariance between the projected X and Y scores,

rather than focusing on the variation in X alone, as with PCA. As the algorithm in the Supplementary Material S1 describes, FPLS iteratively generates a weight function w_j at each step j, for $1 \leq j \leq J$, which solves $\max_{w_j \in \mathcal{L}^2(\mathcal{T})} \operatorname{cov}^2 \left\{ Y^{j-1}, \langle X^{j-1}, w_j \rangle \right\}$, such that $\|w_j\| = 1$ and $\langle w_j, G(w'_j) \rangle = 0$, for all $1 \leq j' \leq j - 1$. Recall that G is the joint covariance operator of the random function X. Here, Y^{j-1} and X^{j-1} are the updated function X and the indicator Y at step $j - 1$ (see S1), respectively, and their corresponding sample values are denoted as Y_i^{j-1} and $X_{i\cdot\cdot}^{j-1}$, for $i = 1, \ldots, n$.

The algorithm gives the decomposition $X_{i\cdot\cdot}(t) = \sum_{j=1}^{J} s_{ij} P_j(t) + E_i(t)$, for $t \in \mathcal{T}$, where $\mathbf{s}_i = (s_{i1}, \ldots, s_{iJ})^T$ is the length J score vector, $P_j \in \mathcal{L}^2(\mathcal{T})$, for $1 \leq j \leq J$, are loading functions, and E_i is the residual. Preda et al. (2007) investigated PLS in linear discriminant analysis (LDA), and defined score vectors $\mathbf{S}_j$ as eigenvectors of the product of the Escoufier's operators of X and Y (Escoufier (1970)). For our case, the classifiers BCG and BCt now act on the PLS scores $\mathbf{s}_i = (s_{i1}, \ldots, s_{iJ})^T$ of each observation $X_{i\cdot\cdot}$. We refer to these classifiers as BCG-PLS and BCt-PLS, respectively.

The dominant PCA directions might only have large within-group variances and small between-group differences in means. Such directions will have little power to discriminate between groups. This problem can be fixed by FPLS. The advantages of FPLS have been discussed, for example, by Preda et al. (2007) and Delaigle and Hall (2012). The latter found that when the difference between the group means projected on the jth PC direction is large only for large j, their functional centroid classifier with PLS scores has lower misclassification rates than when using PCA scores. As later examples show, FPLS is especially effective in such situations.

3. Comparison of Classifiers using Simulated Data

3.1 Data Design

To set up the simulation, for simplicity, we use $\pi_1 = \pi_0 = 0.5$. By Karhunen–Loève expansions, the functions $X_{i \cdot k}$, for $i = 1, \ldots, n_k$, of group $k = 0, 1$ can be decomposed as $X_{i \cdot k} = \mu_k + \sum_{j=1}^{J} \sqrt{\lambda_{jk}} \xi_{ijk} \phi_{jk}$, where μ_k is the group mean, λ_{jk} is the jth eigenvalue in group k corresponding to eigenfunction ϕ_{jk}, and $\lambda_{1k} > \cdots > \lambda_{Jk}$. The variables ξ_{ijk} are distributed with $E(\xi_{ijk}) = 0$, $\mathrm{var}(\xi_{ijk}) = 1$, and $\mathrm{cov}(\xi_{ijk}, \xi_{ij'k}) = 0$, for $\forall j \neq j'$. The compact interval $\mathcal{T}$ is $[0, 1]$, and the functions $X_{i \cdot k}$ are observed at the equally spaced grid $t_1 = 0, t_2 = 1/50, \ldots, t_{51} = 1$, with i.i.d. Gaussian noise $\epsilon_{ik}(t)$ centered at zero and with standard deviation 0.5. The classifiers are implemented both with and without pre-smoothing the data. Because they have similar performance, we report only the results using pre-smoothing. The total sample size is $n = 250$, with 100 training and 150 test cases. The number of eigenfunctions for curve generation is $J = 201$, double the size of the training data set, to imitate the infinite dimensions of the functional data. For each j, the bandwidth h_{jk} for KDE is selected by the direct plug-in method (Sheather and Jones (1991)). Simulations are repeated $N = 1000$ times. The Supplementary Material S3.1 includes additional results with increased training size.

The distribution of (X, Y) is determined by four factors: the eigenfunctions (whether common or group-specific), difference between group means, eigenvalues, and score distributions. The factors are varied according to a $2 \times 2 \times 2 \times 3$ full factorial design, described below. We adopt a four-letter system to label the 24 factor-level combinations, which we call "scenarios."

Factor 1: Eigenfunctions $\phi_{1k}, \ldots, \phi_{Jk}$ **of group** k**:** The first factor specifies the eigenfunctions of the covariance operators G_1 and G_0. When the two sets $\phi_{1k}, \ldots, \phi_{Jk}$, for $k = 0, 1$, are the same, let the common eigenfunctions be the Fourier basis on $\mathcal{T} = [0, 1]$, where $\phi_{1k}(t) = 1, \phi_{jk}(t) = \sqrt{2}\cos(j\pi t)$ or $\sqrt{2}\sin((j-1)\pi t)$, for $1 < j \leq 201$ even or odd.

When the two groups have unequal eigenfunctions, the group $k = 0$ uses the Fourier basis $\phi_{10}, \ldots, \phi_{J0}$ as above, but the group $k = 1$ has a Fourier basis rotated by iterative updating:

i) let the starting value of $\phi_{11}, \ldots, \phi_{J1}$ be the original Fourier basis functions, as above;

ii) at step (j, j'), where $1 \leq j \leq J - 1$, $j' = j + 1, \ldots, J$, the pair of functions $(\phi_{j1}^*, \phi_{j'1}^*)$ is generated by a Givens rotation of angle $\theta_{jj'}$ of the current pair $(\phi_{j1}, \phi_{j'1})$ such that

$$\phi_{j1}^*(t) = \cos(\theta_{jj'})\phi_{j1}(t) - \sin(\theta_{jj'})\phi_{j'1}(t), \quad \phi_{j'1}^*(t) = \sin(\theta_{jj'})\phi_{j1}(t) + \cos(\theta_{jj'})\phi_{j'1}(t).$$

iii) the rotation angle for each pair of (j, j') is $\theta_{jj'} = \dfrac{\pi}{3}(\lambda_{j0} + \lambda_{j'0})$, with $\lambda_{j0}, \lambda_{j'0}$ the jth and j'th eigenvalues, respectively, of group $k = 0$. Hence, the major eigenfunctions receive greater rotations, with the angles proportional to their eigenvalues;

iv) then, we update $\phi_{j1}, \phi_{j'1}$ with the new $\phi_{j1}^*, \phi_{j'1}^*$ and continue the rotations until each pair of (j, j'), with $1 \leq j \leq J - 1$, $j' = j + 1, \ldots, J$, is rotated.

The rotated Fourier basis of group $k = 1$ guarantees that both groups Π_1 and Π_0 span the same eigenspace and satisfy the null hypothesis of the test of equal eigenspaces developed by Benko et al. (2009). This test was used by Dai et al. (2017) to check whether the two groups have the same eigenfunctions. However, having equal eigenspaces is a necessary, but not sufficient condition for having equal sets of eigenfunctions, as proved by the rotated basis. Because of the unequal eigenfunctions of the operators G_1 and G_0, the scores X_{ijk} are correlated, which can be modeled by the new copula-based classifiers.

We also tested other choices of the second set of eigenfunctions, including the Haar wavelet system on $\mathcal{L}^2([0,1])$. However, the results are similar, and so are omitted. We denote the scenario where Π_1 and Π_0 have equal eigenfunctions as S (same), and otherwise as R (rotated).

Factor 2: Difference, μ_d, Between the Group Means: The second factor, which is at two levels, S (same) and D (different), is the difference between the group means, $\mu_d = \mu_1 - \mu_0$. For simplicity, we let $\mu_0 = 0$, $\mu_1 = \mu_d$. Here, $\mu_d(t) = t$.

Factor 3: Eigenvalues $\lambda_{1k}, \ldots, \lambda_{Jk}$ of Group k: The third factor, at two levels labeled S and D, is whether the eigenvalues $\lambda_{1k}, \ldots, \lambda_{Jk}$ depend on k. We label the level where $\lambda_{j1} = \lambda_{j0} = 1/j^2$ as S, and that when $\lambda_{j1} = 1/j^3$ and $\lambda_{j0} = 1/j^2$ as D, for $1 \le j \le J$.

Factor 4: Distribution of the standardized scores ξ_{ijk}: The fourth factor, at three levels N (normal), T (tail dependence and skewness), and V (varied), is the distribution of ξ_{ijk}.

N: $\xi_{i1k}, \ldots, \xi_{iJk}$ have a Gaussian distribution $N(0,1)$ for both $k = 0$ and 1.

T: This level includes tail dependency by setting $\xi_{ijk} = (\delta_{ijk} - b)/\eta_{ik}$, where $\delta_{ijk} \sim \text{Exp}(\lambda^*)$, $\lambda^* = 5\sqrt{3}/3$, $b = 1/\lambda^*$, and $\eta_{ik} \sim \chi^2(5)/5$, for all $j = 1, \ldots, J$. All δ_{ijk} and η_{ik} are mutually independent, whereas the scores ξ_{ijk} on each basis j are uncorrelated, but dependent, because they share the same denominator, η_{ik}. The scores are skewed in both groups.

V: In this level, the scores in the two groups have different types of distributions, with $\xi_{ij1} \sim N(0,1)$, and $\xi_{ij0} \sim \text{Exp}(1) - 1$. Simulation results of a different choice of the varied distributions of ξ_{ij1} and ξ_{ij0} are included in Supplementary Material Section S3.1 Table S1.

Table 1 lists all 24 scenarios used in the simulations:

	$\xi_{ijk} \sim$ N	$\xi_{ijk} \sim$ T	$\xi_{ijk} \sim$ V
$\mu_d = 0,\ \lambda_{j1} = \lambda_{j0}$	(R/S)SSN	(R/S)SST	(R/S)SSV
$\mu_d = 0,\ \lambda_{j1} \neq \lambda_{j0}$	(R/S)SDN	(R/S)SDT	(R/S)SDV
$\mu_d \neq 0,\ \lambda_{j1} = \lambda_{j0}$	(R/S)DSN	(R/S)DST	(R/S)DSV
$\mu_d \neq 0,\ \lambda_{j1} \neq \lambda_{j0}$	(R/S)DDN	(R/S)DDT	(R/S)DDV

Table 1: Simulation scenarios. The labels are ordered: eigenfunctions (R/S), group mean (S, D), eigenvalues (S, D), and ξ_{ijk} distributions (N, T, V). Note that in SSSN and SSST, functions from both groups have the same distribution. We simply include them to have a full factorial design.

3.2   Functional Classifiers

The classifiers used in this study are listed below. Five of them are Bayes classifiers, and the last three are non-Bayes. The methods proposed in this paper are described in (ii) - (iii).

(i) BC: the original Bayes classifier of Dai et al. (2017), with the log density ratio given by Eq.(2.2). The scores are by projection onto PCs;

(ii) BCG, BCG-PLS: Bayes classifiers with a Gaussian copula to model correlation, using PC and PLS scores, respectively. The rank correlation used is Kendall's τ. Both the Gaussian copula and the t-copula densities can be implemented using the R package `copula` (Hofert et al. (2018));

(iii) BCt, BCt-PLS: Bayes classifiers similar to (ii), but using a t-copula instead;

(iv) CEN: functional centroid classifier in Delaigle and Hall (2012), where observation x is classified to group $k = 1$ if $T(x) = (\langle x, \psi \rangle - \langle \mu_1, \psi \rangle)^2 - (\langle x, \psi \rangle - \langle \mu_0, \psi \rangle)^2 \leq 0$, with μ_1 and μ_0 the group means. Here, $\psi = \sum_{j=1}^{J^*} \lambda_j^{-1} \mu_j \phi_j$ is a function of the first J^* joint eigenfunctions ϕ_j, the corresponding eigenvalues λ_j, and $\mu_j = \langle \mu_1 - \mu_0, \phi_j \rangle$;

(v) PLSDA (PLS discriminant analysis): binary classifier using Fisher's linear discriminant

rule, with FPLS as a dimension-reduction method. It is implemented in the R package
`pls` (Mevik et al. (2011));

(vi) Logistic regression: logistic regression on functional PCs, implemented by the R func-
tion `glm`. It is one of the functional generalized regressions discussed in Müller et al.
(2005).

In each simulation, J^* is selected using 10-fold cross validation on the training data. The
candidate J values range from 1 to 30 (2 to 30 for classifiers using copulas). The estimation
of the joint eigenfunctions ϕ_j follows the discretization approach of the fPCA, as described
in Chapter 8.4 of Ramsay and Silverman (2005). A similar discretization strategy is used
for the PLS basis.

3.3   Classifier Performance

Table 2 contains the average misclassification rates over 1000 simulations by each method
on each scenario. In addition, for each simulation, we use 10-fold cross-validation to select
the classifier with the best performance on the training data among the eight classifiers in
Section 3.2. The average misclassification rates of the CV-selected classifier are listed in the
CV column. The column Ratio(CV) contains the percentage difference between the CV-
selected (CV) and the best (opt) classifier: $\text{Ratio(CV)} = \{\text{err(CV)} - \text{err(opt)}\}/\text{err(opt)} \times$
100%. For each scenario, the lowest error rates of the eight classifiers are in bold. We label
those within the optimal case's margin of error (MOE) for each data scenario γ in italics:
$\text{MOE}_\gamma = 1.96 \times \sigma_\gamma^*/\sqrt{1000}$, where σ_γ^* is the sample standard deviation of the best classifier's
(at scenario γ) error rates from 1000 simulations. The simulations enable a comprehensive
understanding of the classifiers' behaviors, which we now discuss.

	BC	BCG	BCGPLS	BCt	BCtPLS	CEN	PLSDA	logistic	CV	Ratio (CV)
SSSN	0.502	0.502	0.500	0.500	0.501	0.502	0.501	0.500	0.501	0.23%
SSDN	**0.227**	0.244	0.345	0.258	0.443	0.464	0.495	0.466	0.232	2.43%
SDSN	0.347	0.351	0.361	0.351	0.363	**0.275**	0.304	0.279	0.291	5.88%
SDDN	**0.169**	0.173	0.303	0.175	0.327	0.231	0.262	0.234	0.173	2.64%
SSST	0.507	0.502	0.500	0.505	0.499	0.499	0.499	0.499	0.502	0.69%
SSDT	**0.438**	*0.441*	0.454	0.456	0.471	0.488	0.497	0.490	0.452	3.19%
SDST	0.188	0.183	0.270	0.184	0.311	**0.167**	0.234	*0.169*	0.170	1.96%
SDDT	0.166	0.161	0.237	0.160	0.296	**0.148**	0.233	*0.150*	0.152	2.59%
SSSV	**0.355**	0.361	0.484	0.363	0.493	0.476	0.481	0.489	0.363	2.20%
SSDV	**0.253**	0.270	0.373	0.276	0.430	0.455	0.477	0.462	0.257	1.78%
SDSV	**0.264**	0.275	0.401	0.276	0.408	0.279	0.315	0.283	0.273	3.27%
SDDV	**0.202**	0.209	0.309	0.207	0.313	0.236	0.280	0.238	0.210	3.95%
RSSN	0.327	**0.147**	0.183	**0.147**	0.180	0.494	0.497	0.485	0.151	2.67%
RSDN	0.252	**0.090**	0.140	0.093	0.164	0.489	0.500	0.482	0.093	2.93%
RDSN	0.287	**0.128**	0.154	**0.128**	0.152	0.327	0.333	0.329	0.131	2.71%
RDDN	0.208	**0.077**	0.112	*0.079*	0.128	0.287	0.300	0.288	0.080	3.44%
RSST	0.435	**0.354**	0.373	*0.357*	0.372	0.486	0.490	0.489	0.361	1.95%
RSDT	0.400	**0.326**	0.348	0.336	0.365	0.486	0.491	0.485	0.339	3.87%
RDST	0.178	**0.148**	0.248	0.154	0.261	0.174	0.252	*0.175*	0.156	5.80%
RDDT	0.166	**0.137**	0.217	0.142	0.255	0.159	0.249	*0.158*	0.147	7.68%
RSSV	0.266	**0.147**	0.202	*0.149*	0.204	0.472	0.481	0.475	0.150	1.71%
RSDV	0.233	**0.100**	0.143	0.105	0.157	0.465	0.475	0.469	0.104	3.85%
RDSV	0.241	**0.145**	0.183	*0.146*	0.191	0.332	0.349	0.337	0.148	2.28%
RDDV	0.238	**0.116**	0.157	0.120	0.167	0.299	0.325	0.300	0.121	3.97%

Table 2: Misclassification rates of eight classifiers on 24 scenarios, each an average from 1000 simulations. Lowest rates of each data case are in bold, and cases within margin of error (see text) of the lowest are in italics. The column labeled CV contains error rates of the classifier selected by cross-validation. Ratio(CV) is the percent difference from the best of the eight classifiers for that scenario. CV error rates are not included in the rankings that determine coloring. SSSN and SSST are in gray because there is actually no difference between groups in these scenarios, and, because $\pi_0 = \pi_1 = 1/2$, the true misclassification rate is 0.5.

– *Equal versus Unequal Eigenfunctions.* A comparison between the top and bottom half of Table 2 demonstrates the strength of our copula-based classifiers, especially on unequal eigenfunctions (bottom half). By its nature, BC has strong performance when the two groups have the same set of eigenfunctions and the scores ξ_{ijk} are mutually independent, for example, in SSDN and SSDV. However, when the data have a more complicated structure, such as score tail dependency and location difference, CEN and logistic obtain better results (SDST, SDDT). Note that in every case with equal eigenfunctions, BCG/BCt are always the ones with rates closest to those of BC.
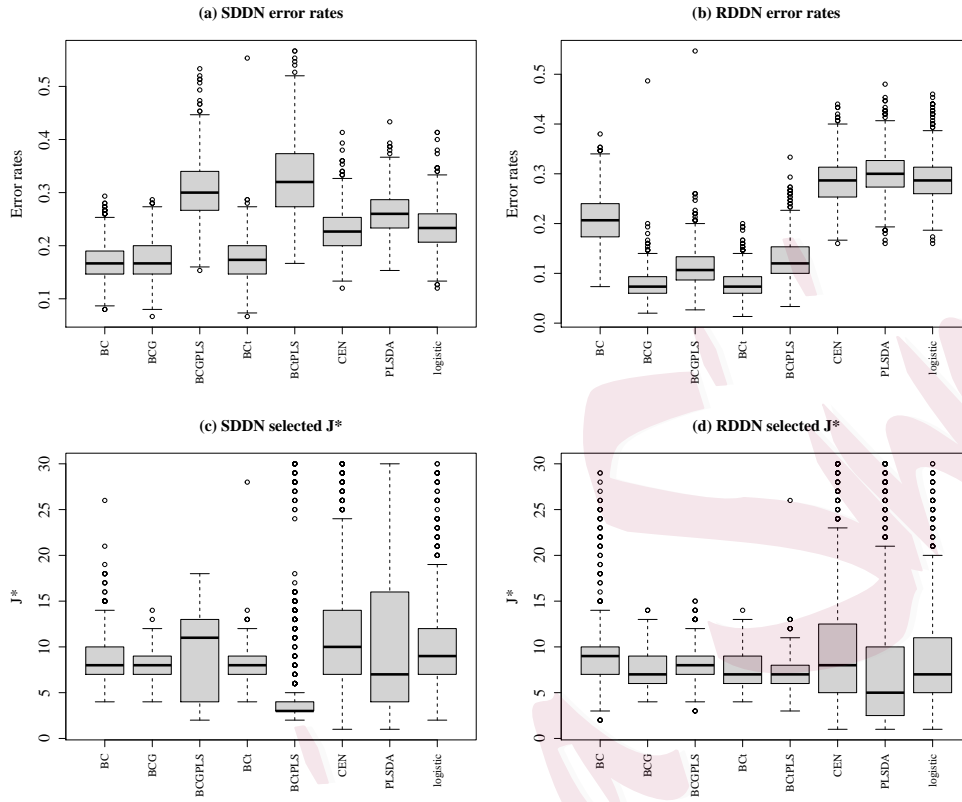
Figure 2: Part (a) and (b) are box plots of the error rates by the eight classifiers in scenarios SDDN and RDDN. The bottom two plots (c) and (d) are box plots of cross-validated J^* in each simulation.

On the other hand, when the group eigenfunctions are different, BC and the three non-Bayes classifiers fail to outperform BCG/BCt in any scenario, even though the group eigenspaces remain equal. BCG maintains its robust performance of lowest error rates throughout all cases. BCt is not far behind, and falls into BCG's MOE 50% of the time as labeled.

Fig. 2 compares the misclassification rates and the corresponding J^* selected in each of the 1000 simulations at two scenarios, SDDN and RDDN. These two scenarios differ only in their eigenfunction setting. In Plot (a), where the groups have equal eigenfunctions, BC, BCG, and BCt show similar behaviors in classification. In Plot (b), where the group eigenfunctions differ, BCG and BCt have the lowest error rates

and variation, followed by BCG-PLS and BCt-PLS. In Plots (c) and (d), we find that BCG and BCt are the only classifiers that have a stable choice of optimal J^*: both methods choose $J^* < 10$ more than 75% of the time with few outliers, regardless of whether the group eigenfunctions are equal or not.

– *Difference between the group means.* Under the equal eigenfunction setting, non-Bayes classifiers such as CEN and the logistic regression are naturally sensitive to a location difference, especially when other factors are kept the same; see for example, SDSN, SDST. However, in the bottom half of Table 2, where the group eigenfunctions differ, BCG shows the strongest performance in all cases, with BCt a close second.

In this table, the PC-based methods BCG and BCt show an advantage over their PLS counterparts in scenarios with a location difference. That is because μ_d is effectively captured by PCs. In Section 3.4, when the new μ_d has nonzero projections only on the last several bases, PLS-based classifiers can do a better job than other methods in distinguishing such a difference, as mentioned in Delaigle and Hall (2012). This phenomenon is also discussed in Section 4.

– *Difference in group eigenvalues and score distributions.* In general, we find that the marginal densities of the scores and their eigenvalues have similar effects on the classifiers' performance. They contribute to the difference of the functional distributions in each group, which the three non-Bayes methods (CEN, PLSDA, logistic) fail to detect. For all scenarios in Table 2 without a location difference, CEN, PLSDA, and the logistic regression all show very poor performance, with error rates close to 50%.

The two right-most columns in Table 2 show that the CV-selected method achieves

comparable performance to the optimal result of each scenario. This demonstrates the stability and strength of our copula-based Bayes classifiers, especially under the unequal eigenfunction setting. Sections S3.2 and S3.3 in the Supplementary Material report the correlations between the first 10 scores in the scenarios RSDN and RSDT, respectively. These high correlations are consistent with the strong performance of the copula-based classifiers in the scenarios where the two groups have different eigenfunctions.

3.4 Multiclass Classification Performance

We also investigate the performance of the aforementioned methods in terms of classifying data into more than two labels, because the group eigenfunctions from multiple different classes are more likely to be unequal, making it increasingly necessary to consider the dependency of the scores on the joint basis.

We now denote the group labels as $Y = k$, for $k = 0, 1, 2$, and set up the multiclass scenarios following the design in Section 3.1. The first column in Table 3 lists the 12 scenarios considered. The first letter M labels unequal group eigenfunctions: when $Y = 0$ and 1, the group eigenfunctions are the Fourier basis and its rotated counterpart, respectively, as described in type R of Factor 1 for binary data; when $Y = 2$, the group basis is again the rotated Fourier functions on $\mathcal{T} = [0, 1]$, but the rotation angle factor used in iii) of Factor 1 in Section 3.1 is now $\pi/4$ instead of $\pi/3$. We omit cases of equal group eigenfunctions, because similar results can be found in the binary setup, and the likelihood of an unequal basis increases as the levels of Y increase.

The second letter S or D again denotes equal group means or not, respectively. When the group means μ_k are unequal (labeled D), we set $\mu_0 = 0$, μ_1 is the identity function used previously, and $\mu_2 = \sum_{j=192}^{201} \phi_{j0}$. The function μ_2 follows a similar design to that of Delaigle

and Hall (2012), where the group mean only has nonzero weights on the last three of 40 eigenfunctions. We assign the nonzero weights to the last 10 of the 201 bases.

Similarly, S or D in the third position represents the same or different group eigenvalues, respectively. When the group eigenvalues are equal, $\lambda_{jk} = 10/j^2$ for all k; otherwise, $\lambda_{jk} = 10/j^2, 10/j^3, 10/j$, respectively, for $k = 0, 1, 2$, for $j \geq 1$. Finally, the last letter inherits the design from Factor 4 of Section 3.1 to describe the standardized score distribution patterns: similarly to the binary case, N and T denote the Gaussian and skewed distributions, respectively, for all three levels, while for V, we define the scores ϵ_{ijk} to follow a standard Gaussian, centered exponential with rate one, or skewed distribution in T, for $k = 0, 1, 2$ respectively.

The other setup details of the noise, data pre-smoothing, and bandwidth selection are all similar to Section 3.1 for binary data. For each simulation, we have 100 training and 150 test cases. The optimal cut-off J^* is selected using cross-validation from $J \leq 10$. Table 3 presents the misclassification rates from 1000 Monte Carlo repetitions by seven of the eight classifiers in Section 3.2. Note that functional centroid classifier is not applicable to multiclass data, and thus is excluded here. As in the binary case, the Supplementary Material Table S2 includes additional results with an increased training size and a different set of score distributions (V).

Table 3 indicates that for data of multiple labels, the behaviors of the seven classifiers follow a similar pattern to that of the binary case when the group eigenfunctions are unequal. In particular, BCt shows strength under increased data complexity, followed closely by BCG. BCG-PLS/BCt-PLS also prove their advantage in detecting location differences on minor basis functions in MDSN. Although they fail to outperform their PC-based counterparts

	BC	BCG	BCGPLS	BCt	BCtPLS	PLSDA	logistic	CV	Ratio(CV)
MSSN	0.520	**0.325**	0.392	*0.327*	0.392	0.641	0.637	0.328	0.89%
MDSN	0.356	0.247	*0.237*	0.245	**0.235**	0.446	0.427	0.226	-3.88%
MSDN	0.213	*0.169*	0.281	**0.168**	0.310	0.636	0.618	0.173	3.00%
MDDN	0.194	*0.156*	0.272	**0.156**	0.295	0.540	0.509	0.157	1.11%
MSST	0.560	**0.450**	0.503	**0.450**	0.492	0.635	0.638	0.456	1.25%
MDST	0.343	**0.286**	0.303	**0.286**	0.333	0.424	0.364	0.284	-0.72%
MSDT	0.449	*0.399*	0.444	**0.397**	0.467	0.624	0.616	0.401	0.95%
MDDT	0.342	0.297	0.355	**0.287**	0.403	0.483	0.401	0.293	2.38%
MSSV	0.325	**0.259**	0.394	*0.261*	0.475	0.633	0.615	0.264	2.23%
MDSV	0.288	*0.237*	0.356	**0.234**	0.433	0.436	0.399	0.241	2.93%
MSDV	0.385	0.314	0.427	**0.302**	0.435	0.631	0.627	0.311	3.00%
MDDV	0.272	*0.223*	0.322	**0.219**	0.340	0.475	0.434	0.224	2.18%

Table 3: Misclassification rates averaged over 1000 simulations of the seven classifiers on 12 multiclass data scenarios. Best case in each scenario is in bold, and cases within margin of error of the lowest are in italic. $P(Y = k) = 1/3$, for $k = 0, 1, 2$, so the true misclassification rate of any method is approximately 0.667.

under more complicated scenarios such as MDST and MDSV, we believe this is because the group means are not the only dominant difference in these two data cases.

Tables 2 and 3 give us clear guidelines that deciding whether or not to use copulas in a classification makes a more significant impact on the outcome than the type of copulas, because both BCG and BCt present competitive performance. The tables also reveal the strength of copula-based methods in dimension reduction. Classifiers using copulas are able to achieve high accuracy with small cut-off J^*, which indicates their advantage in small samples. In addition, in general, PCs are preferable to PLS, owing to their robustness and simplicity of implementation. BCG-PLS and BCt-PLS should be considered when the group mean difference is significant and located at minor eigenfunctions, which we discuss further in the real-data examples.

4.   Real-Data Examples

In this section, we use two real-data examples to illustrate the strength of our new method in terms of classification and dimension reduction with respect to the data size n.

4.1 Classification of Multiple Sclerosis Patients

Our first example explores the classification of multiple sclerosis (MS) cases based on FA profiles of the cca tract. FA is the degree of anisotropy of water diffusion along a tract, and is measured by diffusion tensor imaging (DTI). Outside the brain, water diffusion is isotropic (Goldsmith et al. (2012)). MS is an autoimmune disease leading to lesions in white matter tracts such as the cca. These lesions decrease FA.

The DTI data set in the R package `refund` (Goldsmith et al. (2018)) contains FA profiles at 93 locations on the cca of 142 subjects. The data were collected at Johns Hopkins University and the Kennedy–Krieger Institute. The numbers of visits per subject range from one to eight, but we used the 142 FA curves from the first visits only. One subject with partially missing FA data was removed. Among the 141 subjects, 42 are healthy ($k = 0$) and 99 were diagnosed with MS ($k = 1$). We use local linear regression for data pre-smoothing. To determine the optimal number of dimensions J^* for each method, we use cross-validation with maximal $J = 30$. The misclassification rates from using 10-fold cross-validation were recorded for 1000 repetitions.

As discussed in Section 1, Panel (a) in Fig. 1 plots 5 FA profiles from each group, and panels (b) and (c) display the group means and standard deviations of the cases and controls, using raw and pre-smoothed data. Compared with the controls, MS patients have lower mean FA values and greater variability. We see that smoothing removes some noise.

Method	BC	BCG	BCGPLS	BCt	BCtPLS	CEN	PLSDA	logistic
Error Rate	0.228	0.199	0.211	**0.192**	0.211	0.264	0.219	0.216

Table 4: Average misclassification rates of eight functional classifiers by 1000 repetitions of 10-fold CV. BCt has the best performance. The best case is in bold.

As shown in Table 4 and Part (a) of Fig. 3, BCt achieves the lowest error rate at 0.192,
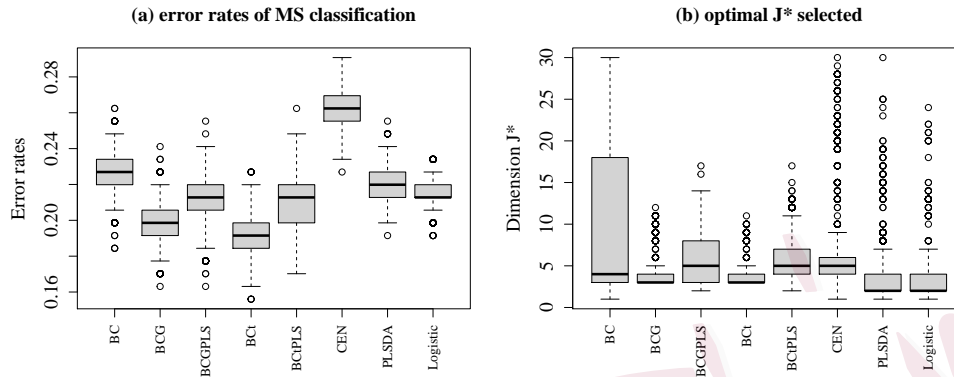
Figure 3: Box plots of misclassification rates and optimal number of components J^* in the MS study over 1000 repetitions of 10-fold cross-validation. BCt achieves the lowest average error rate, while requiring a very small number of components ($J^* < 5$) with lowest variation.

with a margin of error 0.0007. The rates of the other methods fail to fall into this range, and are all significantly higher than that of BCt. In fact, the third quartile for BCt is below the first quartile of all other methods, except BCG. Part (b) is a box plot of cross-validated J^* during each simulation for all classifiers. Here, BCt and BCG achieve the lowest error rates, with a minimal number of dimensions. In addition, compared with methods such as CEN, PLSDA, or logistic regression, their choice of optimal J^* is very stable, with the smallest variation and few outliers. In contrast, BC is prone to employing a large number of components in classification. This tendency can be found in other examples too.

In the Supplementary Material, we compare the loadings (S3), score distributions (S5, and group eigenfunctions (S4) between using PC and PLS. The difference explains why PC is a better choice for this example. Note that it is not our intent to develop DTI as a technique for diagnosing MS. DTI is too expensive and time-consuming for that purpose. Instead, we are looking for differences in FA between cases and controls, because these could inform researchers about the nature of the disease. We have found clear differences between cases and controls in the mean and variance of FA. The strong positive correlation between the

second and the third PC scores in the healthy cases (Spearman's ρ at 0.525 and an adjusted p-value 2×10^{-2}) is diminished in the MS group. BCt and BCG are best able to use a compact model to capture subtle differences, such as correlations.

4.2 Particulate Matter (PM) Emission of Heavy-Duty Trucks

As a second example, we investigate the relationship between the movement patterns of heavy-duty trucks and particulate matter (PM) emissions. We use the data in McLean et al. (2015), originally extracted from the Coordinating Research Council E55/59 emissions inventory program documentary (Clark et al. (2007)). The data set contains 108 records of truck speed in miles/hour over 90-second intervals, and the logarithms of their PM emission in grams (log PM), captured by 70 mm filters.

We dichotomize log PM. The 41 of 108 cases with log PM above average are called high emission ($k = 1$), and the other cases are low emission ($k = 0$). We classify log PM level using the 90-second velocity profiles. The misclassification rates are estimated using 10-fold cross-validation, repeated 1000 times.
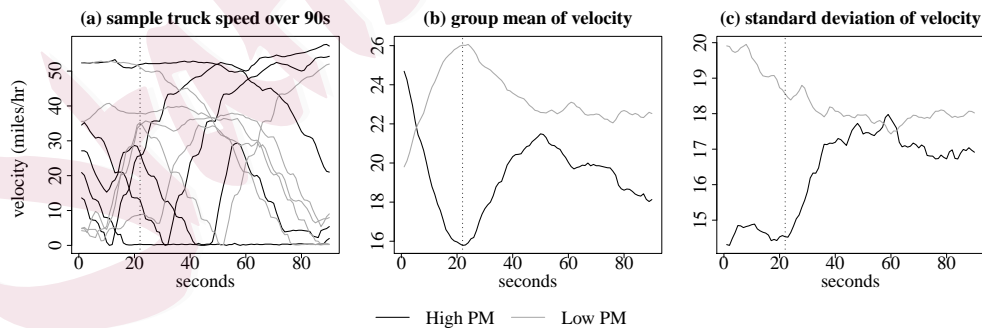

Figure 4: Plots of five sample paths in each PM group, as well as group mean and standard deviation of truck velocity data. On average, trucks in high PM group have lowest speed at 22 seconds, marked with a dashed line on each plot.

As Fig. 4 shows, during the first 20 seconds, vehicles in the high PM group, on average, decelerate to a minimum speed, whereas the low PM group tends to speed up. The high PM

group also has much lower variation than the low PM group.

	BC	BCG	BCGPLS	BCt	BCtPLS	CEN	PLSDA	logistic
Error rate	0.285	0.280	**0.207**	0.280	**0.207**	0.278	0.256	0.228

Table 5: Average misclassification rates of eight functional classifiers by 1000 repetitions of 10-fold cross-validation. BCt-PLS and BCG-PLS have the best performance. The best cases are in bold.
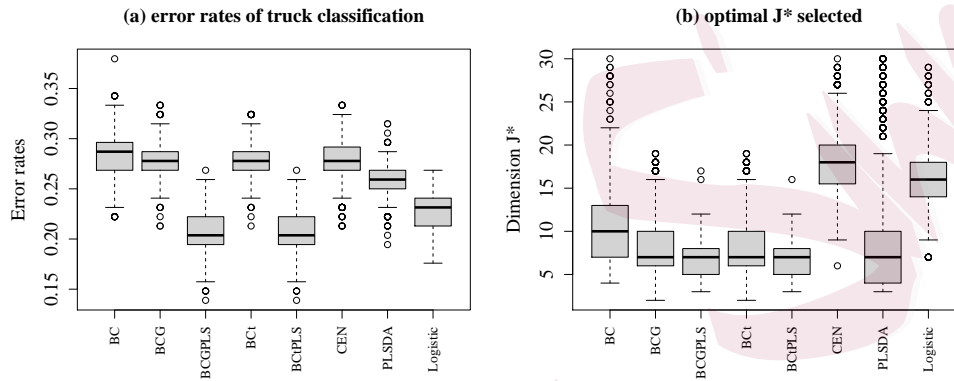

Figure 5: Box plots of misclassification rates and optimal number of components J^* in the truck emission case over 1000 repetitions of 10-fold cross-validation. BCt-PLS and BCG-PLS achieve the lowest average error rate with J^* concentrated around 7.

From Fig. 5 and Table 5, BCG-PLS and BCt-PLS have the lowest misclassification rates. The third quartiles of their error rates are below the first quartiles of the other classifiers, except for the logistic regression. In addition, both methods keep the classification model compact by requiring small J^* with low variation. BC and the three methods on the right of plot (b) of Fig. 5 again demand more components with bigger variation in classification. In Section S4 of the Supplementary Material, we include additional results for both data examples to validate their different choices of PC- and PLS-based classifiers.

5. Theoretical Asymptotic Properties

An interesting feature of functional classifiers is asymptotic perfect classification. That is, under certain conditions, the error rate goes to zero as $J \to \infty$, owing to the infinite-dimensional nature of functional data (Delaigle and Hall (2012)). Dai et al. (2017) discussed

the perfect classification by BC under equal group eigenfunctions. In this section, we prove
that when the group eigenfunctions differ, perfect classification is retained by our classifier
$\mathbb{1}\{\log Q_J^*(X) > 0\}$ for both Gaussian and non-Gaussian processes. The scores $X_{\cdot jk}$, for
$1 \leq j \leq J$, in this section are all projected onto joint eigenfunctions $\phi_1, \ldots, \phi_J$.

We first show that $\log Q_J^*(X)$ and the estimated $\log \hat{Q}_J^*(X)$ are asymptotically equivalent
under mild conditions. Then, the behavior of the Bayes classifier $\mathbb{1}\{\log Q_J^*(X) > 0\}$ is studied
in two settings: first, when the random function $X_{\cdot \cdot k}$ is a Gaussian process for both $k = 0, 1$;
and second, the more general case, when X is non-Gaussian, but its projected scores are
meta-Gaussian distributed in each group. For simplicity, we assume here that $\pi_1 = \pi_0$.

5.1 Asymptotic equivalence of $\log \hat{Q}_J^*(X)$ and $\log Q_J^*(X)$

We first list several assumptions, which help establish the asymptotic equivalence of both
the marginal and the copula density components of $\log \hat{Q}_J^*(X)$ and $\log Q_J^*(X)$.

Assumption A1. *For all $C > 0$ and some $\delta > 0$: $\sup_{t \in \mathcal{T}} E\{|X(t)|^C\} < \infty$,*
$\sup_{s,t \in \mathcal{T}: s \neq t} E[\{|s - t|^{-\delta}|X(s) - X(t)|\}^C] < \infty$.

Assumption A2. *For integers $r \geq 1$, $\lambda_j^{-r} E[\int_{\mathcal{T}}\{X - E(X)\}\phi_j]^{2r}$ is bounded uniformly in j.*

Assumption A3. *There are no ties among the eigenvalues $\{\lambda_j\}_{j=1}^\infty$.*

Assumption A4. *The density g_j of the jth standardized score $\langle X - E(X), \phi_j \rangle / \sqrt{\lambda_j}$ is
bounded and has a bounded derivative; 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 ratio $f_{j1}(X_{\cdot j\cdot})/f_{j0}(X_{\cdot j\cdot})$ is atomless for all $j \geq 1$.*

For all $c > 0$, let $\mathcal{S}(c) = \{x \in \mathcal{L}^2(\mathcal{T}) : \|x\| \leq c\}$. Assumptions A1–A4 are from De-
laigle and Hall (2010), adapted here to bound the difference $D_{jk}(x_j) = \hat{g}_{jk}(\hat{x}_j) - \bar{g}_{jk}(x_j)$ s.t.
$\sup_{x \in \mathcal{S}(c)} |D_{jk}(x_j)| = o_p\{(nh)^{-1/2}\}$. We let $\hat{g}_{jk}(\hat{x}_j) = 1/(n_k h) \sum_{i=1}^{n_k} K\left\{\langle X_{i\cdot k} - x, \hat{\phi}_j\rangle / (\hat{\sigma}_{jk}h)\right\}$

be the estimated density of the standardized scores of group k on basis $\hat{\phi}_j$, with $\bar{g}_{jk}(x_j)$ using

ϕ_j and σ_{jk}. In addition, the following assumption is added for $D_{jk}(x_j)$, for both $k = 0, 1$:

Assumption A5. $\sup_{x \in \mathcal{S}(c)} |\hat{\pi}_k D_{jk}(x_j) / (\hat{\pi}_0 D_{j0}(x_j) + \hat{\pi}_1 D_{j1}(x_j))| = O_p\left(1 + \sqrt{\dfrac{\log n}{nh^3}}\right).$

We use A5 to give a mild bound simply to avoid the case where the magnitudes of both

$D_{jk}(x_j)$, for $k = 0, 1$, are too large and close, but with opposite signs. A5 guarantees that

the difference between the estimated marginal density $\hat{f}_{jk}(\hat{x}_j)$ and $f_{jk}(x_j)$ is able to be

bounded by the same rate as when the group eigenfunctions are equal. However, this is not

a necessary condition for the asymptotic equivalence of $\log \hat{Q}_J^*(X)$ and $\log Q_J^*(X)$, and we

can certainly relax its bound for Theorem 1 below.

Then, $\hat{f}_{jk}(\hat{x}_j) = (1/\hat{\sigma}_{jk})\,\hat{g}_{jk}(\hat{x}_j)$, and we have Proposition 1 (see the Supplementary

Material for the proof):

Proposition 1. *Under Assumptions A1–A5, when the group eigenfunctions are unequal, the*

estimated marginal density $\hat{f}_{jk}$ using scores $\langle X_{i \cdot k}, \hat{\phi}_j \rangle$ achieves an asymptotic error bound:

$\sup_{x \in \mathcal{S}(c)} |\hat{f}_{jk}(\hat{x}_j) - f_{jk}(x_j)| = O_p\left\{ h + \sqrt{\dfrac{\log n}{nh}} \right\}$, *where the rate is the same as in Dai et al.*

(2017), where the group eigenfunctions are equal.

Assumption A6. *The CDFs F_{jk} of scores $X_{\cdot jk}$ are continuous and strictly increasing, with*

correspondent marginal densities f_{jk} continuous and bounded. In addition, f_{jk} are bounded

away from zero on any compact interval within their supports.

A6 ensures that the scores $X_{\cdot jk}$ and their monotonic transformations are atomless; this

also follows Condition 5 in Dai et al. (2017).

Then, in addition to the marginal densities, we establish the equivalence of Ω_k^{-1} and $\hat{\Omega}_k^{-1}$

in $\log Q_J^*(X)$ and $\log \hat{Q}_J^*(X)$, respectively, as $n \to \infty$. As mentioned in Section 2.3, we

calculate $\hat{\mathbf{\Omega}}_k$ using rank correlations. In addition, when J is large, the inverse of $\hat{\mathbf{\Omega}}_k$ can be estimated using the graphical Dantzig selector (Yuan (2010)), which solves the matrix inverse by connecting the entries of the inverse correlation matrix to a multivariate linear regression, and exploits the sparsity of the inverse matrices (Yuan (2010)). Liu et al. (2012) provided a q-norm Op bound of the difference between the inverse Gaussian copula matrix and its estimation by the Dantzig estimator for high-dimensional problems, and is extended here for the difference between $\mathbf{\Omega}_k^{-1}$ and $\hat{\mathbf{\Omega}}_k^{-1}$.

Our sparsity assumptions on the inverse correlation matrices follow the design of Yuan (2010) and Liu et al. (2012): let $\mathbf{\Omega}_k$ belong to the class of matrices $\mathcal{C}\left(\kappa, \tau, M, J\right) := \{\mathbf{\Omega}^{J \times J} : \mathbf{\Omega} \succ \mathbf{0}, \operatorname{diag}(\mathbf{\Omega}) = \mathbf{1}, \|\mathbf{\Omega}^{-1}\|_1 \leq \kappa, \frac{1}{\tau} \leq \lambda_{\min}(\mathbf{\Omega}) \leq \lambda_{\max}(\mathbf{\Omega}) \leq \tau, \operatorname{deg}(\mathbf{\Omega}^{-1}) \leq M\}$, where $\kappa, \tau \geq 1$ are constants determining the tuning parameter in the graphical Dantzig selector, and the parameter M bounding $\operatorname{deg}(\mathbf{\Omega}^{-1}) = \max_{1 \leq j \leq J} \sum_{j'=1}^{J} I(\mathbf{\Omega}_{jj'}^{-1} \neq 0)$ is dependent on J. Assuming these sparsity conditions, we have the following theorem.

Theorem 1. *Under A1–A6, $\forall \epsilon > 0$, as $n \to \infty$, there exists a sequence $J\left(n, \epsilon, M\right) \to \infty$, and a set S dependent on $J\left(n, \epsilon, M\right)$, $P\left(S\right) \geq 1 - \epsilon$, such that*

$$P\left(S \cap \left\{\mathbb{1}\left\{\log \hat{Q}_J^*\left(X\right) \geq 0\right\} \neq \mathbb{1}\left\{\log Q_J^*\left(X\right) \geq 0\right\}\right\}\right) \to 0,$$

provided that $MJ\sqrt{\log J} = o\left(\sqrt{n}\right)$.

Theorem 1 proves that under unequal group eigenfunctions, $\log \hat{Q}_J^*\left(X\right)$ using copulas retains the property in Theorem A1 of Dai et al. (2017) for the estimated Bayes classifiers with equal group eigenfunctions and independent scores: as $n \to \infty$, $\log \hat{Q}_J^*\left(X\right)$ gets arbitrarily close to the true Bayes classifier $\log Q_J^*\left(X\right)$, which enables us to discuss the performance of

our method using the properties of the true Bayes classifier.

5.2   Perfect classification when X is a Gaussian process in both groups

Let $X_{\cdot \cdot k}$ be a centered Gaussian process such that $X_{\cdot \cdot k} = \sum_{q=1}^{\infty} \sqrt{\lambda_{qk}} \xi_{qk} \phi_{qk}$, with $\xi_{qk} \sim N(0,1)$, for $k = 0, 1$. We denote the $J \times J$ covariance matrix of scores $X_{\cdot jk}$, for $1 \le j \le J$, as $\mathbf{R}_k$, where its (j, j')th entry is equal to $\mathrm{cov}\,(X_{\cdot jk}, X_{\cdot j'k}) = \sum_{q=1}^{\infty} \lambda_{qk} \langle \phi_{qk}, \phi_j \rangle \langle \phi_{qk}, \phi_{j'} \rangle$, and its eigenvalues are $d_{1k}, \ldots, d_{Jk}$. Let $\vec{\mu}_J$ be a length-J vector $(\mu_1, \ldots, \mu_J)^T$ by projecting μ_d on first J bases, $\mu_j = \langle \mu_d, \phi_j \rangle$. By the law of total covariance and the result that the trace of a matrix is equal to the sum of its eigenvalues, we derive the following relationship between the two sets of eigenvalues (i.e. λ_j, λ_{jk}, and d_{jk}): $\sum_{j=1}^{J} \lambda_j = \pi_1 \sum_{j=1}^{J} d_{j1} + \pi_0 \sum_{j=1}^{J} d_{j0} + \pi_1 \pi_0 \sum_{j=1}^{J} \mu_j^2$, and $\sum_{j=1}^{J} d_{jk} = \sum_{j=1}^{J} \sum_{q=1}^{\infty} \lambda_{qk} \langle \phi_{qk}, \phi_j \rangle^2$. The following assumption is standard in functional data for the distribution of X, and ensures that $d_{jk} > 0$, for $1 \le j \le J$, $k = 0, 1$:

Assumption A7. *Both the group covariance operators, G_1, G_0, and the covariance matrices $\mathbf{R}_0$, $\mathbf{R}_1$ are bounded and positive definite, and $\mu_d \in \mathcal{L}^2(\mathcal{T})$.*

When X is Gaussian in both groups, $\log Q_J^*(X)$ is a quadratic form in $\mathbf{X}_J$ ($\mathbf{X}_J$ is a length-J vector with jth entry $\langle X, \phi_j \rangle$):

$$\log Q_J^*(X) = -\frac{1}{2} \left(\mathbf{X}_J - \vec{\mu}_J\right)^T \mathbf{R}_1^{-1} \left(\mathbf{X}_J - \vec{\mu}_J\right) + \frac{1}{2} \mathbf{X}_J^T \mathbf{R}_0^{-1} \mathbf{X}_J + \log \sqrt{\frac{|\mathbf{R}_0|}{|\mathbf{R}_1|}}. \tag{5.1}$$

With potentially unequal group eigenfunctions, entries in $\mathbf{X}_J$ at $Y = k$ can be correlated, which complicates the distribution of $\log Q_J^*(X)$ in each group.

Therefore, we implement a linear transformation of $\mathbf{X}_J$ in Steps i)–iii):

i) The eigendecomposition of the matrix product gives $\mathbf{R}_0^{1/2} \mathbf{R}_1^{-1} \mathbf{R}_0^{1/2} = \mathbf{P}^T \boldsymbol{\Delta} \mathbf{P}$, where

$\mathbf{\Delta} = \mathrm{diag}\{\Delta_1, \ldots, \Delta_J\}$, Δ_j as eigenvalues of $\mathbf{R}_0^{1/2}\mathbf{R}_1^{-1}\mathbf{R}_0^{1/2}$. By the equivalence of the determinants, $\prod_{j=1}^{J} \frac{d_{j0}}{d_{j1}} = \prod_{j=1}^{J} \Delta_j$. In addition, $\Delta_j > 0$, for all j, under A7;

ii) Let $\mathbf{Z} = \mathbf{R}_0^{-1/2}\mathbf{X}_J$, $\mathbf{U} = \mathbf{PZ}$;

iii) When $k = 0$, the jth entry U_j of the vector $\mathbf{U}$ has a standard Gaussian distribution; at $k = 1$, $U_j \sim N(-b_j, 1/\Delta_j)$, with b_j the jth entry of $\mathbf{b} = -\mathbf{PR}_0^{-1/2}\vec{\mu}_J$.

Consequently, the entries of $\mathbf{U}$ are uncorrelated for both $k = 0$ and 1, Eq.(5.1) becomes

$$\log Q_J^*(X) = -\frac{1}{2}\sum_{j=1}^{J} \Delta_j \left(U_j + b_j\right)^2 + \frac{1}{2}\sum_{j=1}^{J} U_j^2 + \frac{1}{2}\sum_{j=1}^{J} \log \Delta_j,$$

and the asymptotic behaviors of the Bayes classifier for Gaussian processes are concluded.

Theorem 2. *With A7, when the random function X is a Gaussian process at both $Y = 0$ and 1 and the group eigenfunctions of G_0, G_1 are unequal, the functional Bayes classifier $\mathbb{1}\{\log Q_J^*(X) > 0\}$ achieves perfect classification when either $\|\mathbf{R}_0^{-1/2}\vec{\mu}_J\|^2 \to \infty$, or $\sum_{j=1}^{J}(\Delta_j - 1)^2 \to \infty$, as $J \to \infty$. Otherwise, its error rate $err(\mathbb{1}\{\log Q_J^*(X) > 0\}) \not\to 0$.*

Theorem 2 is a natural extension of Theorem 2 in Dai et al. (2017). It again reveals that the error rate of the Bayes classifier approaches zero asymptotically when Π_1 and Π_0 are sufficiently different in terms of either the group means or the scores' variances. In addition, recognizing the different correlation patterns between group scores helps improve the classification accuracy. Instead of adopting $\mu_j/\sqrt{\lambda_{j0}}$ and $\lambda_{j0}/\lambda_{j1}$ to build conditions for perfect classification, as in Dai et al. (2017), we use the transformed $\mathbf{R}_0^{-1/2}\vec{\mu}_J$ and Δ_j to accommodate the potentially unequal group eigenfunctions and the dependent scores. For the special case when the eigenfunctions are actually equal, the covariance matrices

$\mathbf{R}_k = \text{diag}\{\lambda_{1k}, \ldots, \lambda_{Jk}\}$ with $\Delta_j = \lambda_{j0}/\lambda_{j1}$, and consequently the two conditions in Theorem 2 become the same as those proposed in Dai et al. (2017). The proof of Theorem 2 is in Section S6.2 of the Supplementary Material.

5.3   When X is a non-Gaussian process

For non-Gaussian processes, when the projected scores $X_{\cdot jk}$, for $1 \leq j \leq J$, fit a Gaussian copula model, that is, they are meta-Gaussian distributed, we derive sufficient conditions in terms of the marginal densities f_{jk} and the score correlations in order to achieve an asymptotically zero misclassification rate.

First, we let $\mathbf{u}_k = (u_{1k}, \ldots, u_{Jk})^T$ be a length-J random vector with $u_{jk} = \Phi^{-1}(F_{jk}(X_{\cdot j\cdot}))$, where $\Phi(\cdot)$ is the CDF of $N(0,1)$. When $Y = k$, $(u_{jk}|Y = k) \sim N(0,1)$, and $\text{var}(\mathbf{u}_k|Y = k) = \boldsymbol{\Omega}_k$, as denoted before. Let the eigendecomposition be $\boldsymbol{\Omega}_k = \mathbf{V}_k \mathbf{D}_k \mathbf{V}_k^T$, with $\mathbf{D}_k$ the diagonal matrix with eigenvalues ω_{jk}, for $j = 1, \ldots, J$. On the other hand, $u_{jk}|Y = k'$ follows a more complicated distribution when $k' \neq k$. We denote $\text{var}(\mathbf{u_k}|Y = k') = \mathbf{M}_k$ with the eigendecomposition $\mathbf{M}_k = \mathbf{U}_k \tilde{\mathbf{D}}_k \mathbf{U}_k^T$, and the eigenvalues of $\mathbf{M}_k$ are v_{jk}, for $j = 1, \ldots, J$.

Therefore, the log density ratio $\log Q_J^*(X)$ in the Bayes classifier with a Gaussian copula can be represented as

$$\log Q_J^*(X) = \sum_{j=1}^J \log \frac{f_{j1}(X_{\cdot j\cdot})}{f_{j0}(X_{\cdot j\cdot})} + \frac{1}{2}\log\frac{|\boldsymbol{\Omega}_0|}{|\boldsymbol{\Omega}_1|} - \frac{1}{2}\mathbf{u}_1^T\left(\boldsymbol{\Omega}_1^{-1} - \mathbf{I}\right)\mathbf{u}_1 + \frac{1}{2}\mathbf{u}_0^T\left(\boldsymbol{\Omega}_0^{-1} - \mathbf{I}\right)\mathbf{u}_0$$

$$= \sum_{j=1}^J \log \frac{f_{j1}(X_{\cdot j\cdot})}{f_{j0}(X_{\cdot j\cdot})} \bigg/ \frac{\sqrt{\omega_{j1}}}{\sqrt{\omega_{j0}}} - \frac{1}{2}\mathbf{u}_1^T\left(\boldsymbol{\Omega}_1^{-1} - \mathbf{I}\right)\mathbf{u}_1 + \frac{1}{2}\mathbf{u}_0^T\left(\boldsymbol{\Omega}_0^{-1} - \mathbf{I}\right)\mathbf{u}_0. \quad (5.2)$$

Similarly to A7, we make an assumption on the covariances of $\mathbf{u}_k$, conditional on Y:

Assumption A8. *The matrices $\boldsymbol{\Omega}_k$ and $\mathbf{M}_k$, for $k = 0, 1$, are bounded and positive definite.*

Next, we define a sequence of ratios g_j, for $j = 1, 2, \ldots$, by $g_j = \dfrac{f_{j1}(X_{\cdot j \cdot})}{f_{j0}(X_{\cdot j \cdot})} \Big/ \dfrac{\sqrt{\omega_{j1}}}{\sqrt{\omega_{j0}}}$, where g_j compares the ratio of the marginal densities to the ratio of the eigenvalues of the correlation matrices. In addition, let

$$s_{jk} = \frac{\mathrm{var}\left(\langle V_{jk}, \mathbf{u}_k \rangle | Y = k\right)}{\mathrm{var}\left(\langle V_{jk}, \mathbf{u}_k \rangle | Y = k'\right)} = \frac{\mathbf{V}_{jk}^T \boldsymbol{\Omega}_k \mathbf{V}_{jk}}{\mathbf{V}_{jk}^T \mathbf{M}_k \mathbf{V}_{jk}} = \frac{\omega_{jk}}{\sum_{q=1}^{J} C_{(j,q)k}^2 \upsilon_{qk}},$$

where $C_{(j,q)k} = \langle \mathbf{U}_{qk}, \mathbf{V}_{jk} \rangle$, $\sum_{q=1}^{J} C_{(j,q)k} = 1$, and $\mathbf{U}_{qk}$ and $\mathbf{V}_{jk}$ are the qth and jth columns, respectively, of the eigenvector matrices $\mathbf{U}_k$ and $\mathbf{V}_k$. As a result, s_{jk} compares the jth eigenvalue of $\boldsymbol{\Omega}_k$ against a convex combination of the eigenvalues of $\mathbf{M}_k$, the individual weights of which are determined by projecting $\mathbf{V}_{jk}$ onto the eigenvalues of $\mathbf{M}_k$, $\mathbf{U}_{qk}$.

In terms of the sequences g_j and s_{jk}, for $j = 1, 2, \ldots$, we derive the following theorem for non-Gaussian processes; the proof is in Section S6.3 of the Supplementary Material.

Theorem 3. *With Assumptions A6, A7, and A8, when the projected scores $X_{\cdot j k}$, for $j = 1, \ldots, J$, are meta-Gaussian distributed at each group Π_k, perfect classification by the Bayes classifier $\mathbb{1}\{\log Q_J^*(X) > 0\}$ is achieved asymptotically if a subsequence $g_r^* = g_{j_r}$ of g_j exists, with corresponding $s_{j_r k}$, such that one of the following conditions is satisfied as $r \to \infty$:*

a) $g_{j_r} = o_p(1)$, and $s_{j_r 0} \to 0$;

b) $1/g_{j_r} = o_p(1)$, and $s_{j_r 1} \to 0$;

or when g_{j_r} has distinct behaviors in subgroups:

c) $g_{j_r} = o_p(1)$ at $Y = 1$, $1/g_{j_r} = o_p(1)$ at $Y = 0$, with both $s_{j_r 0}$ and $s_{j_r 1} \to 0$;

d) $1/g_{j_r} = o_p(1)$ at $Y = 1$, and $g_{j_r} = o_p(1)$ at $Y = 0$.

Based on the structure of the log density ratio described in Eq.(5.2), Theorem 3 discusses the occurrence of perfect classification in two aspects: g_j, which mainly depicts the relative magnitude of the score marginal densities at each $k = 0, 1$; and s_{jk}, which compares the correlation between the scores conditioned at each group. Either part showing enough disparity between groups results in perfect classification.

For example, in Theorem 3 a), when there exists a subsequence $g_{j_r} \to 0$ in probability, indicating the dominance of the marginal densities by the group $Y = 0$, the misclassification tends to occur at $Y = 1$. However, as $s_{j_r \cdot 0} \to 0$, the covariance of $\mathbf{u}_0$ conditioned at $Y = 1$ becomes much larger than at $Y = 0$. As a result, the nonnegative $\mathbf{u}_0^T \mathbf{\Omega}_0^{-1} \mathbf{u}_0^T$ in Eq.(5.2) with large variation when $Y = 1$ compensates to eventually avoid misclassifying X to group 0. When g_{j_r} behaves perfectly, as in case d), where the corresponding group marginal densities are dominant in each subgroup $Y = k$, we do not need to impose requirements on the copula correlation to achieve perfect classification.

Remark. Theorem 3 provides sufficient, but not necessary conditions for the Bayes classifier to achieve asymptotic perfect classification under unequal group eigenfunctions. Owing to the optimality of the Bayes classifier in minimizing the zero-one loss, various conditions from other functional classifiers to achieve an asymptotically zero error also work here. For example, Delaigle and Hall (2012) proposed conditions in terms of group eigenvalues and the mean difference for the functional centroid classifier to reach perfect classification. These also work as sufficient conditions for $\mathbb{1}\{\log Q_J^*(X) > 0\}$ in our case. With a copula model, which is not found in previous work, Theorem 3 uses the relation between the scores' marginal densities and correlations to reduce the error rate to zero asymptotically.

6. Discussion

6.1 Remarks

Our copula-based Bayes classifiers remove the assumptions of equal group eigenfunctions and independent scores. As our two examples show, it is not uncommon to have unequal group eigenfunctions (see Fig. S4 and Fig. S8). The new methods also prove to have stronger performance in terms of dimension reduction than that of the original BC. Our simulation results prove the strength of our method in distinguishing groups by the differences in their functional means and their covariance functions. We examined the two choices of projection directions, PC and PLS. PLS can detect location differences on eigenfunctions corresponding to smaller eigenvalues. We discussed new conditions for the estimated classifier to be asymptotically equivalent to the true Bayes classifier, and for perfect classification to occur. These differ from those of previous works, owing to the unequal group eigenfunction setting. We also imposed sparsity conditions on the inverse of the copula correlations.

6.2 Future Work

In future work, we would like to extend the copula-based classification to the problem with multiple functional covariates. Some previous works discuss this situation in the framework of functional generalized models: Crainiceanu et al. (2009) proposed a generalized multi-level regression model where there are repeated curve measurements for each subject; Zhu et al. (2010) discussed an FGLM approach for the classification of multilevel functions with Bayesian variable selection; and Li et al. (2010) present a generalized functional linear model where there are both functional and multivariate covariates, and use a semiparametric single-index function to model the interaction between them. We plan to approach the problem

from a different angle, using functional Bayes classification again, owing to its strong performance in the single functional predictor case. Furthermore, because it is natural to assume that the response depends on the covariates and their interactions, it becomes more important for our method to model the dependency between the projected scores. Another aspect we would like to consider is how to choose a proper functional basis for multiple functional predictors.

Supplementary Materials

The Supplementary Materials for this document contain additional results for the simulations, for the fractional anisotropy (FA) example, and for the example using truck emissions. They also contain proofs of Theorems 1, 2, and 3.

Acknowledgements

References

Aguilera, A. M., Escabias, M., Preda, C., and Saporta, G. (2010). Using basis expansions for estimating functional pls regression: applications with chemometric data. *Chemometrics and Intelligent Laboratory Systems*, 104(2):289–305.

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

Chen, X. and Fan, Y. (2006). Estimation of copula-based semiparametric time series models. *Journal of Econometrics*, 130(2):307–335.

Cholaquidis, A., Fraiman, R., Kalemkerian, J., and Llop, P. (2016). A nonlinear aggregation type classifier. *Journal of Multivariate Analysis*, 146:269–281.

Clark, N. N., Gautam, M., Wayne, W. S., Lyons, D. W., Thompson, G., and Zielinska, B. (2007). Heavy-duty vehicle chassis dynamometer testing for emissions inventory, air quality modeling, source apportionment and air toxics emissions inventory. *Coordinating Research Council, incorporated*.

Crainiceanu, C. M., Staicu, A.-M., and Di, C.-Z. (2009). Generalized multilevel functional regression. *Journal of the American Statistical Association*, 104(488):1550–1561.

Cuevas, A., Febrero, M., and Fraiman, R. (2007). Robust estimation and classification for functional data via projection-based depth notions. *Computational Statistics*, 22(3):481–496.

Dai, X., Müller, H.-G., and Yao, F. (2017). Optimal bayes classifiers for functional data and density ratios. *Biometrika*, 104(3):545–560.

Delaigle, A. and Hall, P. (2010). Defining probability density for a distribution of random functions. *The Annals of Statistics*, 38(2):1171–1193.

Delaigle, A. and Hall, P. (2011). Theoretical properties of principal component score density estimators in functional data analysis. *Bulletin of St. Petersburg University. Maths. Mechanics. Astronomy*, (2):55–69.

Delaigle, A. and Hall, P. (2012). Achieving near perfect classification for functional data. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 74(2):267–286.

Escoufier, Y. (1970). *Echantillonnage dans une population de variables aléatoires réelles*. Department de math.; Univ. des sciences et techniques du Languedoc.

Genest, C., Ghoudi, K., and Rivest, L.-P. (1995). A semiparametric estimation procedure of dependence parameters in multivariate families of distributions. *Biometrika*, 82(3):543–552.

Gijbels, I., Omelka, M., and Veraverbeke, N. (2012). Multivariate and functional covariates and conditional copulas. *Electronic Journal of Statistics*, 6:1273–1306.

Goldsmith, J., Crainiceanu, C. M., Caffo, B., and Reich, D. (2012). Longitudinal penalized functional regression for

cognitive outcomes on neuronal tract measurements. *Journal of the Royal Statistical Society: Series C (Applied*

Statistics), 61(3):453–469.

Goldsmith, J., Scheipl, F., Huang, L., Wrobel, J., Gellar, J., Harezlak, J., McLean, M., Swihart, B., Xiao, L.,

Crainiceanu, C., Reiss, P., Chen, Y., Greven, S., Huo, L., Kundu, M., Park, S., Miller, D. s., and Staicu, A.-M.

(2018). refund: Regression with functional data. *R package version*, 0.1(17).

Hall, P. and Hosseini-Nasab, M. (2009). Theory for high-order bounds in functional principal components analysis.

In *Mathematical Proceedings of the Cambridge Philosophical Society*, volume 146, pages 225–256. Cambridge

University Press.

Hofert, M., Kojadinovic, I., Maechler, M., and Yan, J. (2018). *copula: Multivariate Dependence with Copulas*. R

package version 0.999-19.1.

James, G. M. (2002). Generalized linear models with functional predictors. *Journal of the Royal Statistical Society:*

Series B (Statistical Methodology), 64(3):411–432.

James, G. M. and Hastie, T. J. (2001). Functional linear discriminant analysis for irregularly sampled curves. *Journal*

of the Royal Statistical Society: Series B (Statistical Methodology), 63(3):533–550.

Kauermann, G., Schellhase, C., and Ruppert, D. (2013). Flexible copula density estimation with penalized hierarchical

b-splines. *Scandinavian Journal of Statistics*, 40(4):685–705.

Kendall, M. G. (1948). Rank correlation methods.

Kruskal, W. H. (1958). Ordinal measures of association. *Journal of the American Statistical Association*, 53(284):814–

861.

Li, B. and Yu, Q. (2008). Classification of functional data: A segmentation approach. *Computational Statistics &*

Data Analysis, 52(10):4790–4800.

Li, Y., Wang, N., and Carroll, R. J. (2010). Generalized functional linear models with semiparametric single-index

interactions. *Journal of the American Statistical Association*, 105(490):621–633.

Liu, H., Han, F., Yuan, M., Lafferty, J., and Wasserman, L. (2012). High-dimensional semiparametric gaussian copula

graphical models. *The Annals of Statistics*, 40(4):2293–2326.

Mashal, R. and Zeevi, A. (2002). Beyond correlation: Extreme co-movements between financial assets. *Unpublished,*

Columbia University.

McLean, M. W., Hooker, G., and Ruppert, D. (2015). Restricted likelihood ratio tests for linearity in scalar-on-

function regression. *Statistics and Computing*, 25(5):997–1008.

McLean, M. W., Hooker, G., Staicu, A.-M., Scheipl, F., and Ruppert, D. (2014). Functional generalized additive

models. *Journal of Computational and Graphical Statistics*, 23(1):249–269.

Mevik, B.-H., Wehrens, R., and Liland, K. H. (2011). pls: Partial least squares and principal component regression.

R package version, 2(3).

Müller, H.-G., Stadtmüller, U., et al. (2005). Generalized functional linear models. *Annals of Statistics*, 33(2):774–805.

Preda, C., Saporta, G., and Lévéder, C. (2007). Pls classification of functional data. *Computational Statistics*,

22(2):223–235.

Ramsay, J. O. and Silverman, B. W. (2005). *Functional data analysis*. New York: Springer.

Rossi, F. and Villa, N. (2006). Support vector machine for functional data classification. *Neurocomputing*, 69(7-

9):730–742.

Ruppert, D. and Matteson, D. S. (2015). *Statistics and Data Analysis for Financial Engineering with R examples*.

Springer.

Shang, Z., Cheng, G., et al. (2015). Nonparametric inference in generalized functional linear models. *The Annals of*

Statistics, 43(4):1742–1773.

Sheather, S. J. and Jones, M. C. (1991). A reliable data-based bandwidth selection method for kernel density

estimation. *Journal of the Royal Statistical Society: Series B (Methodological)*, 53(3):683–690.

Singh, S. and Póczos, B. (2017). Nonparanormal information estimation. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 3210–3219. JMLR.org.

Stone, C. J. (1983). Optimal uniform rate of convergence for nonparametric estimators of a density function or its derivatives. In *Recent advances in statistics*, pages 393–406. Elsevier.

Yuan, M. (2010). High dimensional inverse covariance matrix estimation via linear programming. *Journal of Machine Learning Research*, 11(Aug):2261–2286.

Zhu, H., Vannucci, M., and Cox, D. D. (2010). A bayesian hierarchical model for classification with selection of functional predictors. *Biometrics*, 66(2):463–473.

Department of Statistics and Data Science, Cornell University

E-mail: wh365@cornell.edu

School of Operations Research and Information Engineering, and Department of Statistics and Data Science, Cornell University

E-mail: dr24@cornell.edu