Curve Data Classification via Functional Principal Component Analysis

doi:10.6148/IJITAS.2010.0304.02

International Journal of Intelligent Technologies and Applied Statistics, 3(4), 2010

IJITAS, 3(4), 2010

作者/Author： Pai-Ling Li;Che-Chiu Wang

頁數/Page： 383-399

出版日期/Publication Date：2010/12

International Journal of Intelligent Technologies and Applied Statistics
Vol.3, No.4 (2010) pp.383–399

Curve Data Classification via Functional Principal Component Analysis

Pai-Ling Li* and Che-Chiu Wang

Department of Statistics, Tamkang University, Taipei, Taiwan

ABSTRACT

We propose a best predicted curve classification (BPCC) criterion for classifying the curve data. The data are viewed as realizations of a mixture of stochastic processes and each subprocess corresponds to a known class. Under the assumption that all the groups have different mean functions and eigenspaces, an observed curve is classified into the best predicted class by minimizing the distance between the observed and predicted curves via subspace projection among all classes based on the functional principal component analysis (FPCA) model. The BPCC approach accounts for both the means and the modes of variation differentials among classes while other classical functional classification methods consider the differences in mean functions only. Practical performance of the proposed method is demonstrated through simulation studies and a real data example of matrix assisted laser desorption (MALDI) mass spectrometry (MS) data. The proposed method is also compared with other multivariate and functional classification approaches. Overall, the BPCC method outperforms the others when the mean functions and the eigenspaces among classes are significantly distinct. For classifying the MALDI MS data, we found that functional classification methods perform better than multivariate data approaches, and the dimension reduction via FPCA is advantageous to improving the accuracy of classification.

Keywords: Classification; Functional data analysis; Functional principal component analysis; Mass spectrometry; Proteomics

1. Introduction

With the technological advances in high-throughput, extensive repeated measurements are increasingly collected in many scientific fields. For example, the daily precipitation sampled from different weather stations for many years, the time-course gene expression levels of thousands of genes simultaneously measured at sequential time points in a microarray experiment, and the proteomic spectrum densely collected at sequential mass-per-charge (m/z) ratios in a matrix-assisted laser

* Corresponding author: plli@stat.tku.edu.tw

desorption and ionization time-of-flight (MALDI-TOF) experiment. Data of this type can be viewed as realizations of curves or longitudinally collected functional data. Recently, statistical methods for analyzing curve data based on the framework of functional data analysis (FDA) have been developed actively. A comprehensive introduction of functional data analysis is provided by Ramsay and Silverman [15].

Supervised classification of curve data is a major and interesting topic of functional data analysis. Numerous functional classification methods have been proposed for identifying functional data into correct classes according to pattern or functional features of the curves. James and Hastie [7] proposed the functional linear discrimination analysis extended from classical multivariate data approach via smoothing techniques. Using the generalized linear model or functional logistic regression for functional data classification is another popular approach (James [8]; Müller and Stadtmüller [13]; Leng and Müller [9]; Aguilera et al. [1]). Ferraty and Vieu [5] proposed a nonparametric kernel method for discriminating functional data. Extension of support vector machine (SVM) to functional data classification has been discussed (Rossi and Villa [16]; Park et al. [14]). Moreover, functional classification methods based on Baysian analysis are proposed with applications to gene expression data (Mallick et al. [10]) and MALDI data (Morris et al. [11]). Functional classification of curve data or functional data has been greatly developed in the last decade.

In this study, classification of curve data based functional principal component analysis (FPCA) is discussed. A typical approach of applying FPCA to functional data classification is classifying the FPC scores directly through logistic regression (Müller [12]; Leng and Müller [9]). However, the difference of within-curve covariance or correlation structures between distinct groups may not be shown by the distribution of FPC scores. For this reason, we use the FPCA model to propose the best predicted curve classification (BPCC) criterion for classifying the curve data, which can simultaneously take into account the means and the modes of variation differentials between classes. Chiou and Li [2] proposed the k-centers functional clustering (kCFC), where the cluster centers hinge on the cluster functional principal component subspaces and individual cluster membership is determined by the minimum L^2 distance between the observed curve and the fitted function obtained by cluster subspace projection. The classification criterion of the unsupervised clustering method, kCFC, can be easily extended to the proposed BPCC method. We assume that each observed curve can be viewed as a realization of a random function and is sampled from a mixture of K stochastic processes, where each subprocess represents a class and the class center is defined by the structure of a FPC subspace that corresponds to the Karhunen-Loève expansion. An observed curve is then classified into the best predicted class by minimizing the distance between the observed and predicted curves via subspace projection among all classes based on the FPCA model. It is shown that the proposed BPCC algorithm performs reasonably well for data under various cluster structures in our numerical studies.

The rest of this paper is organized as follows. Section 2 introduces the functional random-effects model of random curves and the relevant FPCA-based functional classification methods, including the BPCC and functional logistic regression (FLR). Simulation studies for investigating the numerical performance of the proposed method are presented in Section 3. Section 4 illustrates a practical application to identifying lung cancer cases by a MALDI mass spectrometry (MS) data set. Concluding remarks are summarized in Section 5.

2. Classification based on FPC subspace projection

2.1 Functional random-effects model of random curves

Let $L^2(dv)$ represents a Hilbert space of square integrable functions with respect to the measure $dv(t) = v(t)dt$ on a real interval $\mathcal{T} = [a, b]$, for $a < b$, where dt is a Lebesgue measure and $v(t)$ is a nonnegative weight function such that $v(t) > 0$ for $t \in T$ and $v(t) = 0$ otherwise. The inner product of two functions f and g in $L^2(dv)$ is defined as $\langle f, g\rangle = \int f(t)g(t)dv(t)$ and the L^2 norm is defined as $\|\cdot\| = \langle\cdot,\cdot\rangle^{1/2}$. Here, we use a constant weight function $v(t) = (b - a)^{-1}I_{[t\in T]}$ in this study. Suppose that n independent random curves $X_1, X_2,..., X_n$ are sampled from a stochastic process X in $L^2(dv)$. Assume that the process X has a smooth mean function $\mu(t) = E(X(t))$ and a smooth covariance function $\Gamma(s, t) = Cov(X(s), X(t))$, where Γ is twice continuously differentiable. Based on the Karhunen-Loève expansion, the random function $X_i(t)$ can expressed as the following functional random-effects model,

$$X_i(t) = \mu(t) + \sum_{j=1}^{\infty} \xi_{ij}\varphi_j(t), \tag{1}$$

where the random-effects ξ_{ij} are uncorrelated with zero mean and finite variance λ_j. The set of functions $\{\varphi_j\}$ forms an orthornormal basis in L^2 associated with the covariance function Γ. In practice, the random function X_i can possibly be contaminated with measurement errors. Let the trajectory $Y_i(t)$ be a realization of the random function $X_i(t)$ at time t. We consider the additive measurement error model

$$Y_i(t) \quad = \quad X_i(t) + \epsilon_i(t),$$

where $\epsilon_i(t)$ are random measurement errors that are iid with $E(\epsilon_i(t)) = 0$ and $Cov(\epsilon_i(s), \epsilon_i(t)) = \sigma^2\delta_{st}$, where δ_{st} is 1 for $s = t$ and 0 otherwise. The random errors $\epsilon_i(t)$ are assumed to be independent of ξ_{ij}. In practical applications, a truncated expansion of model (2) with finite number of components such as

$$\tilde{X}_i(t) = \mu(t) + \sum_{j=1}^{M} \xi_{ij}\varphi_j(t) \quad (2)$$

is often required for approximating the random process X_i. A common practice is to choose the number of components M according to the proportion of total variations explained by the first few leading principal components $\{\varphi_j,\ j = 1,\ldots, M\}$. We note that when $\mu \in Span\ \{\varphi_1, \varphi_2,\ldots, \varphi M\}$, the mean function $\mu(t)$ can be expressed as $\sum_{j=1}^{M} \langle \mu, \varphi_j \rangle\ \varphi_j(t)$ and thus the multiplicative random-effects ξ_{ij} in model (2) are not identifiable. Therefore, we assume that the mean function μ does not belong to the space spanned by the principal component functions $\{\varphi_j,\ j = 1, 2,\ldots, M\}$ for identifiability concerns.

Let $y_{ij} = y_i(t_{ij})$ be the jth observation of the ith random curve Y_i collected at time point t_{ij}. In estimation of the model components in (1), the mean function $\mu(t)$ is estimated by applying local linear regression to the scatterplot data $\{(t_{ij}, y_{ij})_{i=1,\ldots, n, j=1,\ldots, mi}\}$, and the covariance function Γ is obtained by applying the scatterplot smoothing into the raw covariances $\{r_{i,j\ell}, 1 \le i \le n, 1 \le j \ne \ell \le m_i\}$ to fit a local linear plane, where $r_{i,j\ell} = (y_{ij} - \hat{\mu}(t_{ij}))(y_{i\ell} - \hat{\mu}(t_{i\ell}))$. The estimated eigenvalues $\{\hat{\lambda}_j\}$ and eigenfucntions $\{\hat{\varphi}_j\}$ are obtained by solving an eigensystem

$$\hat{\Gamma}(s,t) = \sum_{j=1}^{\infty} \lambda_j \varphi_j(s)\varphi_j(t).$$

The random-effects ξ_{ij} can be obtained by the shrinkage estimates of Yao et al. [18] to adjust for measurement errors such that $\hat{\xi}_{ij} = \frac{\hat{\lambda}_j}{\hat{\lambda}_j + \hat{\tau}/m_i}\tilde{\xi}_{ij}$, where $\tilde{\xi}_{ij}$ is a discrete approximation to $\langle y_i - \hat{\mu}, \hat{\varphi}_j \rangle$, and $\hat{\tau}$ is the estimate of shrinkage parameter τ , which can be obtained by leave-one-curve-out cross-validation.

2.2 The best predicted curve classification criterion

In this section, we extend the basic principles of k-centers FC proposed by Chiou and Li [2] to functional classification. Suppose the data curves considered comprise K clusters, then X can be viewed as a mixture of K subprocesses in L^2. Let the random variable C is the cluster membership of X, $C \in \{1, 2,\ldots, K\}$. We assume that the mean and covariance funcitons associated with cluster c are conditionally defined $\mu^{(c)}(t) = E(X(t) \mid C = c)$ and $\Gamma^{(c)}(s, t) = Cov(X(s), X(t) \mid C = c)$. Let $\{\varphi_j^{(c)},\ j = 1, 2,\ldots\}$ be the set of orthornormal bases associated with the covariance function $\Gamma^{(c)}(s, t)$ as defined in link with the functional random-effects model (1). Similar to the model (1), the realization $Y_i^{(c)}$ of the random function X_i corresponding to cluster c is given by the conditional model

$$\begin{aligned} Y_i^{(c)}(t) &= X_i^{(c)}(t) + \epsilon_i^{(c)}(t) \\ &= \mu^{(c)}(t) + \sum_{j=1}^{\infty} \xi_{ij}^{(c)} \varphi_j^{(c)}(t) + \epsilon_i^{(c)}(t), \end{aligned} \tag{3}$$

where $\xi_{ij}^{(c)} = \langle X_i - \mu^{(c)}, \varphi_j^{(c)} \rangle$ are uncorrelated random-effects with mean zero and variance $\lambda_j^{(c)}$. The uncorrelated random errors $\epsilon_i^{(c)}(t)$ have zero mean and constant variance $\sigma_{(c)}^2$ and are assumed to be independent of $\xi_{ij}^{(c)}$. When a large proportion of total variances is explained by the first few leading principal components as is often the case in practice, it is appropriate to consider the projection of $X_i^{(c)}$ onto the FPC subspace of cluster c, a truncated nonparametric random-effect model $\tilde{X}_i^{(c)}$, to approximate $X_i^{(c)}$ such that

$$\tilde{X}_i^{(c)}(t) = \mu^{(c)}(t) + \sum_{j=1}^{M_c} \xi_{ij}^{(c)} \varphi_j^{(c)}(t), \tag{4}$$

where the number of components M_c has to be chosen to expand the random process effectively.

Based on the FPC subspace projection framework, the best cluster membership c^* of an observed curve Y^* is determined by a metric that properly measures the distance between Y^* and its projection onto the FPC subspaces of different classes. When the L^2 distance is chosen as the dissimilarity measure, we may define the distance measure by $d_L(f, g) = \|f - g\|^2$ for any functions f and g in L^2. The best predicted cluster membership c^* of Y^* is determined by the classification criterion

$$c^* = \underset{c \in \{1,\ldots,K\}}{\arg\min}\ d_{L2}(Y^*, \hat{X}^{(c)}), \tag{5}$$

where $\hat{X}^{(c)}$ is the projected function of Y^* onto FPC subspace of cluster c with $\hat{X}^{(c)}(t) = \hat{\mu}^{(c)}(t) + \sum_{j=1}^{Mc} \hat{\xi}_j^{(c)*} \hat{\varphi}_j^{(c)}(t)$, $\hat{\xi}_j^{(c)*}$ is the shrinkage estimator of $\langle Y^* - \mu^{(c)}, \varphi_j^{(c)} \rangle$, $j = 1,\ldots, M_c$, based on model (4).

Let $\{y_i^{(c)}, i = 1,\ldots, n_c, c = 1,\ldots, K\}$ be a training data set of K known classes, and y^* be a newly observed curve to be classified. The BPCC algorithm comprises two basic steps: (a) to estimate the mean functions $\mu^{(c)}$ and eigenfunctions $\{\varphi_j^{(c)}\}$ of each class of the training data, and (b) to classify y^* based on the criterion (5). We note that the BPCC method discovers homogeneous subgroups of curve data according to the structure of the means as well as the modes of variation differentials through the projections of the FPC subspaces. Moreover, the L^2 distance measure d_{L2} can be replaced by other distance measures according to different classification purposes.

For instance, the functional correlation $d_{FC}(f, g) = \langle \frac{\tilde{f}}{\|\tilde{f}\|}, \frac{\tilde{g}}{\|\tilde{g}\|} \rangle$ (Chiou and Li [3]), where $\tilde{f} = f - \langle f, 1 \rangle$ is a centered version of f and $\tilde{g}$ is defined analogously, for $f, g \in L^2$; the nonparametric rank correlation $d_{RC}(f, g) = \frac{\langle R^f, R^g \rangle}{\|R^f\| \|R^g\|}$ of Heckman and Zamar [6], where $R^f(t) = r^f(t) - \langle r^f, 1 \rangle$, $r^f(t) = Pr\{f(S) < f(t)\} + \frac{1}{2} Pr\{f(S) = f(t)\}$ is the random function of f, and $S \sim Uniform(a, b)$. Both the distances measures d_{FC} and d_{RC} can be used for classification based on the shape similarity between curves. For this case, the classification criterion (5) is replaced by

$$c^* = \arg\max_{c \in \{1,\ldots,K\}} d_{FC}(Y^*, \hat{X}^{(c)})$$

or

$$c^* = \arg\max_{c \in \{1,\ldots,K\}} d_{RC}(Y^*, \hat{X}^{(c)}).$$

2.3 Functional logistic regression based on FPCA

Functional logistic regression is a special case of generalized functional linear model and is widely used to the classification problem of two groups. Functional logistic regression based on FPCA has been also discussed in other previous literatures (Escabias et al. [4]; Müller [12]). Leng and Müller [9] applied the functional discrimination through logistic regression based on functional principal components to analysis of yeast cell-cycle temporal data. We introduce the basic idea of functional logistic regression based on FPCA model in this section.

Let the response Z denote membership in one of two groups, and define $Z = 1$ if the observation comes from the group G_1 and $Z = 0$ if it comes from the other group G_0. Given the data $(\{X_i(t)\}, Z_i)$, $i = 1,\ldots, n$, where $X_i(t)$ is the predictor function of ith subject. The functional logistic regression is given by $Z_i = \pi_i + e_i$, where $\pi_i = P(Z_i = 1 \mid \{X_i(t), t \in \mathcal{T}\})$, and e_i are i.i.d random errors with zero mean and variance $\pi_i(1 - \pi_i)$. The probability of response $Z = 1$ for the predictor function $X_i(t)$ is given by

$$\pi_i = g^{-1}\left(\alpha^* + \int_{\mathcal{T}} X_i(t)\beta(t)dt\right), \tag{6}$$

where g is the logit function, i.e., $g(x) = \log\{x/(1 - x)\}$. The parameters α^* and $\beta(t)$ are a constant and a smooth function, respectively. Suppose the coefficient function $\beta(\cdot)$ can be expanded by $\beta(t) = \sum_{j=1}^{\infty} \beta_j \varphi_j(t)$ by the orthnormal eigenfunctions defined in the model (1). It is easy to shown that the logistic model (6) can be expressed as

$$\pi_i = g^{-1}\left(\alpha + \sum_{j=1}^{\infty} \beta_j \xi_{ij}\right) \tag{7}$$

based on the FPCA model (1) of the predictor function $X_i(t)$. Intuitively, the classification rule of the logistic regression model (7) is determined by the marginal FPC scores ξ_{ij} of the two groups. We note that the model (7) is usually approximated by the truncated model based on the M leading FPC scores in practice, and the FPC scores ξ_{ij} can be estimated by the method described in Section 2.1. The coefficients $(\alpha, \beta_1, \ldots, \beta_M)$ are obtained by the usual estimating equation approach. We estimate the π_i by $\hat{\pi}_i = g^{-1}\left(\hat{\alpha} + \sum_{j=1}^{M} \hat{\beta}_j \hat{\xi}_{ij}\right)$ and classify the ith subject into the group G_1 by $\hat{\pi} \le p_1$, otherwise into group G_0. In this study, we use $p_1 = 0.5$ in simulations and data example.

3. Simulations

We examine the performance of the proposed BPCC method and compare it with the functional logistic regression based on FPCA model (denoted by FLR) presented in Section 2.3 and the multivariate logistic regression (denoted by LR) approaches for two classes classification. The synthetic curves of two groups are generated based on the model

$$y_i^{(k)}(t_{il}) = \mu^{(k)}(t_{il}) + \sum_{j=1}^{2} \xi_{ij}^{(k)} \varphi_j^{(k)}(t_{il}) + \epsilon_{il}, \tag{8}$$

where $l = 1, \ldots, m$, $i = 1, \ldots, n_k$, and $k = 1, 2$. The recording times are generated from an equally spaced design on [0, 1] such that $t_{il} = (l - 1)/(m - 1)$. The random-effects $\xi_{ij}^{(k)}$ are independently generated from $N(0, \lambda_j)$, and the random errors $\epsilon_{il}^{(k)}$ are independently generated from $N(0, \sigma^2)$, for all k. We consider the following three settings for mean functions of two classes in the simulations:

Setting 1: Two mean functions differ in vertical shift. The mean functions of two clusters are set as $\mu^{(1)}(t) = -2(t - 0.5)^2 + t$ and $\mu^{(2)}(t) = \mu^{(1)}(t) - 2$.

Setting 2: Two mean functions differ only in two local features. The mean functions of two clusters are set as

$$\mu^{(1)}(t) = 1/(1 + e^{5-t})$$

and

$$\mu^{(2)}(t) = \mu^{(1)}(t) + 0.5I(t = t_2) - 0.75I(t = t_m),$$

where $I(\cdot)$ is the indicator function.

Setting 3: Two mean functions differ in shape. The mean functions of two clusters are set as

$$\mu^{(1)}(t) = 4(t - 0.5)^2 + 1$$

and

$$\mu^{(2)}(t) = 2.5 \exp\{-25(t-0.25)^2\} + 2 \exp\{-50(t-0.75)^2\}.$$

Define $E_1 = span\{\rho_{11}, \rho_{12}\}$ and $E_2 = span\{\rho_{21}, \rho_{22}\}$ as two distinct eigensapces spanned by different sets of orthonormal eigenfunctions, where

$$\rho_{11}(t) = \sqrt{2}\sin(\pi t), \rho_{12}(t) = \sqrt{2}\cos(\pi t), \rho_{21}(t) = \sqrt{2}\sin(2\pi t),$$

and $\rho_{22}(t) = \sqrt{2}\cos(2\pi t)$

are orthonormal basis functions. We consider the following six cases in the simulations:

Case A1: The mean functions of two clusters are set as Setting 1 and the eigenspaces of two clusters are set as the same space E_1. The eigenvalues $(\lambda_1^{(k)}, \lambda_2^{(k)})$ are set as (.2, .1), for k = 1, 2, and the variance of measurement errors are set as σ^2 = .25.

Case A2: The mean functions of two clusters are set as Setting 2, and the variance of measurement errors are set as σ^2 = .00064. The eigenspaces of two clusters and eigenvalues $(\lambda_1^{(k)}, \lambda_2^{(k)})$ are the same as Case A1.

Case A3: The mean functions of two clusters are set as Setting 3, the eigenvalues $(\lambda_1^{(k)}, \lambda_2^{(k)})$ are set as (.4, .3), for k = 1, 2, and the variance of measurement errors are set as σ^2 = .25. The eigenspaces of two clusters are the same as in Case A1.

Case B1: The mean functions of two clusters are set as Setting 1, and the eigenspaces of the first and second clusters are set E_1 and E_2, respectively. The eigenvalues $(\lambda_1^{(k)}, \lambda_2^{(k)})$ are set as (2, 1), for k = 1, 2, and the variance of measurement errors are set as σ^2 = 1.

Case B2: The mean functions of two clusters are set as Setting 2, and the variance of mea-surement errors are set as σ^2 = .81. The eigenspaces of two clusters and eigenvalues $(\lambda_1^{(k)}, \lambda_2^{(k)})$ are the same as in Case B1.

Case B3: The mean functions of two clusters are set as Setting 3, the eigenvalues $(\lambda_1^{(k)}, \lambda_2^{(k)})$ are set as (4, 3), for k = 1, 2, and the variance of measurement errors are set as σ^2 = .36. The eigenspaces of two clusters are the same as in Case B1.

The first three Cases A1-A3 are designed for two clusters with different mean functions but the same eigenspace, while the other three Cases B1-B3 are set for clusters with different mean functions and eigenspaces. We generate $2n$ synthetic curves for a sample, where $n = n_1 + n_2$, and the first and second clusters have $2n_1$ and $2n_2$ curves, respectively. We randomly select n_k curves of cluster k as the training data for estimating the model components and building the classification rule, for k = 1, 2. The remaining n curves are taken as test data for calculating the accuracy of classification. Various sample sizes n and number of time point m are considered in simulations. We also consider the balanced design ($n_1 = n_2$) and the unbalanced design ($n_1 \neq n_2$) for the sample size of each cluster.

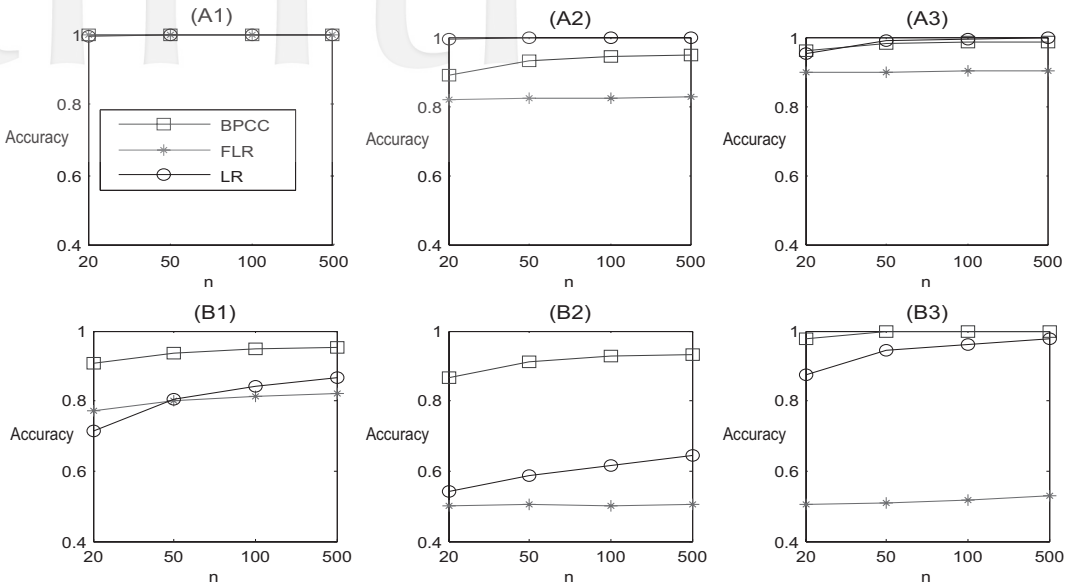

Figure 1. Average accuracy of test data resulted from the BPCC, FLR and LR methods under $m = 11$ and different sample sizes n based on 1,000 synthetic samples.

The methods BPCC, FLR and LR are compared by the accuracy of correct classification based on 1,000 synthetics samples. For the FLR and BPCC methods, the local polynomial smoothing techniques are implemented for estimation and the bandwidths are automatically chosen by the cross-validation method. The numbers of principal components are determined by the 90% criterion for the proportion of total variance explained. We compare the three classification methods across various combinations of sample sizes (n = 20, 50, 100, 500) and numbers of observation points (m = 11, 21, 41) under the balanced sample size design. Figure 1 shows the average accuracy of the three compared methods under different number of curves for m = 11. Obviously, the accuracy of classification increases in sample size for all methods. The BPCC and LR methods performs better than FLR, and the BPCC outperforms the others for Cases B1-B3. Similar results can be obtained for the larger number of observation points. Figure 2 displays the average accuracy under different number of time points m. Except for Case A2, the accuracy obtained from BPCC increases in the observation numbers while the effect of m is not consistent among all the cases for LR and FLR. The accuracy decreases in m for Case A2 could be because the local features of the second cluster are not detected for larger number of time points. For most of the cases, the BPCC method still performs better than LR and FLR under all considered number of time points, especially for Cases B1-B3. Overall, the BPCC method has the outstanding advantage over LR and FLR when distinct clusters differs in mean functions and eigenspaces.

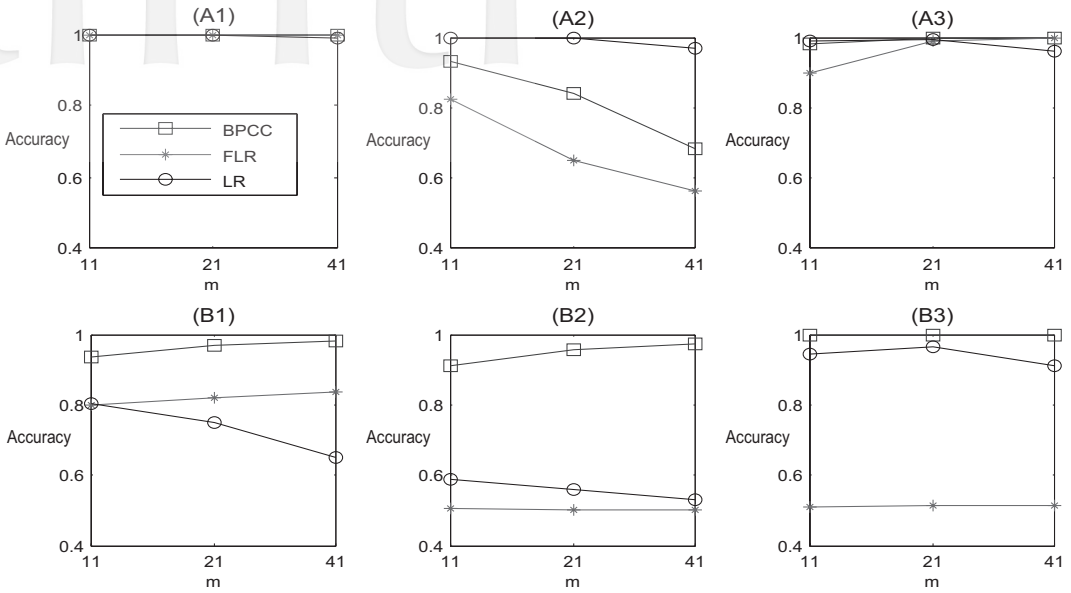

Figure 2. Average accuracy of test data resulted from the BPCC, FLR and LR methods under $n = 100$ and different number of time points m based on 1,000 synthetic samples.

Furthermore, we compare the performances of all the methods under an unbalanced design of sample sizes. The number of curves for the first cluster is set as a small sample size $n_1 = 10$, and the size of second cluster is set as a larger number $n_2 = 40$. Table 1 presents the averages and standard deviations of accuracy for all cases. In addition to the accuracy of all curves, the accuracy of each class is also respectively provided in Table 1. In general, the BPCC outperforms the others for Cases B1-B3 while there is no best approach for Cases A1-A3. In addition, the accuracy of cluster with smaller sample size ($n_1 = 10$) obtained from LR and FLR is lower than the cluster with larger sample size ($n_2 = 40$), whereas the BPCC method performs well for both clusters. Therefore, compared with the LR and FLR approaches, the BPCC criterion has more stable performance of accuracy, specificity, and sensitively for clusters with very different sample sizes.

4. An application to the classification of mass spectrometry proteomic data

We introduce an application of the FPCA-based functional classification methods of a MALDI MS data set of Yildiz et al. [19]. The data consist of 288 serum proteomic profiles collected from a case-control study which aims to distinguish lung cancer cases from matched controls through MALDI MS analysis of the the most abundant peptides in the serum. The cases and controls were matched to avoid confounding

Table 1. Averages and standard deviations (SD) of accuracy for the test data obtained by LR, FRL and BPCC methods based on 1,000 simulation replicates for Cases A1–B3 under unbalanced design of cluster sizes. The sample sizes of two groups are set as $(n_1, n_2) = (10, 40)$.

		Average Accuracy (SD)		
Case	Method	All	Class 1	Class 2
A1	*LR*	0.9912 (0.0211)	0.9894 (0.0373)	0.9917 (0.0218)
	FLR	1.0000 (0.0000)	1.0000 (0.0000)	1.0000 (0.0000)
	BPCC	1.0000 (0.0000)	1.0000 (0.0000)	1.0000 (0.0000)
A2	*LR*	0.9714 (0.0383)	0.9617 (0.0726)	0.9739 (0.0387)
	FLR	0.7912 (0.0242)	0.0313 (0.0721)	0.9812 (0.0348)
	BPCC	0.6659 (0.1321)	0.5438 (0.2415)	0.6964 (0.1781)
A3	*LR*	0.9635 (0.0454)	0.9512 (0.0818)	0.9666 (0.0459)
	FLR	0.9992 (0.0077)	0.9977 (0.0287)	0.9996 (0.0041)
	BPCC	0.9998 (0.0024)	0.9991 (0.0122)	1.0000 (0.0000)
B1	*LR*	0.7236 (0.0928)	0.5906 (0.1824)	0.7569 (0.1064)
	FLR	0.9226 (0.0495)	0.7060 (0.1634)	0.9768 (0.0395)
	BPCC	0.9825 (0.0236)	0.9572 (0.0839)	0.9888 (0.0231)
B2	*LR*	0.5989 (0.0820)	0.4030 (0.1673)	0.6478 (0.1009)
	FLR	0.8089 (0.0245)	0.0672 (0.1144)	0.9943 (0.0180)
	BPCC	0.9739 (0.0321)	0.9424 (0.1120)	0.9818 (0.0322)
B3	*LR*	0.9161 (0.0655)	0.8632 (0.1343)	0.9293 (0.0663)
	FLR	0.8054 (0.0294)	0.0596 (0.1157)	0.9919 (0.0249)
	BPCC	0.9976 (0.0140)	0.9880 (0.0700)	1.0000 (0.0000)

variables such as age, sex, and total pack-year history. Yildiz et al. [19] split the the matched cases and controls into training (n = 182) and test (n = 106) sets. Since the MS data are high-dimensional even after preprocessing, they selected seven MS features based on the preprocessed spectra and applied the selected features to multivariate class-prediction models. Based on the seven discriminant features, the overall classification accuracy, sensitivity, and specificity of their matched blinded test set obtained from the logistic regression (denoted by LR [7 features]), SVM, and the Weighted Flexible Compound Covariate Method (WFCCM) of Shyr and Kim [17] are shown in Table 2. Among these three methods, the SVM approach performs slightly better than the others.

In this study, we are interested in applying the functional classification methods to the densely collected serum proteomic profiles, especially the FPCA-based approaches. We consider the 288 MALDI MS serum spectra after preprocessing, in which each spectrum represents the ion current intensities measured at 184 m/z

Table 2. Accuracy, sensitivity, and specificity of the blinded test set of MALDI MS data selected by Yildiz et al. [19].

Method	(M_1, M_2)	Accuracy	Sensitivity	Specificity
LR (7 features)		0.7547	0.6200	0.8750
SVM (7 features)		0.7925	0.6400	0.9286
WFCCM (7 features)		0.7260	0.5800	0.8570
LR		0.6226	0.5560	0.6786
FLR	(43,43)	0.9057	0.8800	0.9286
BPCC	(26,20)	0.9057	0.9600	0.8571
BPCC ($M_1 = M_2$)	(71,71)	0.9340	0.9600	0.9107

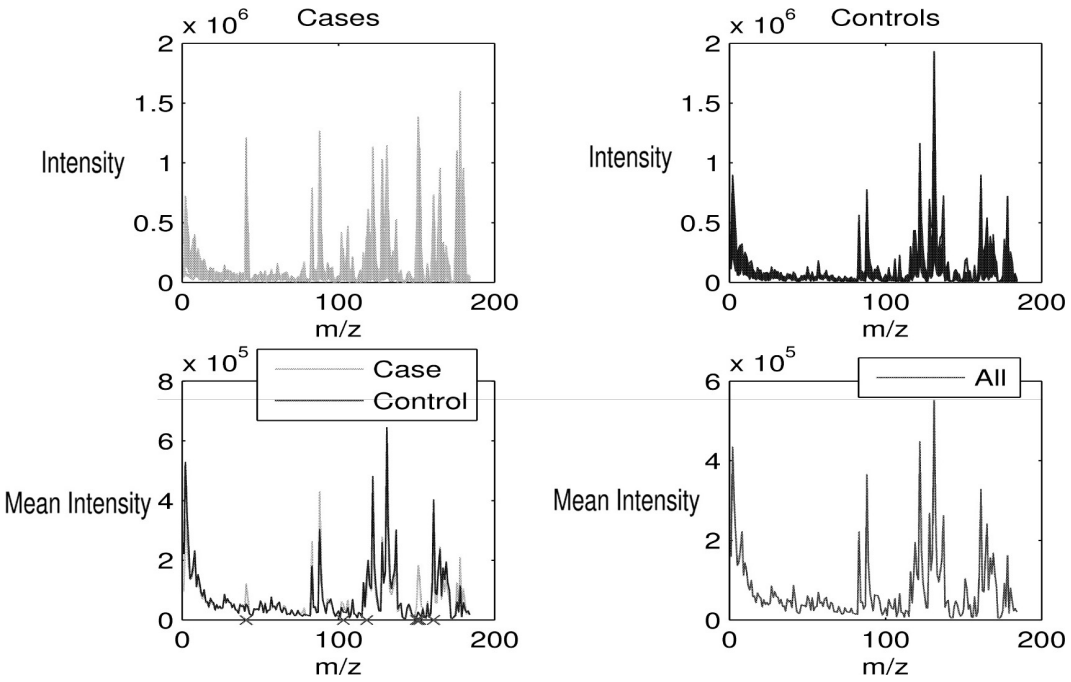

Figure 3. Raw trajectories of MALDI MS data (upper panels) and the marginal mean function (lower right panel) and conditional mean functions (lower left panel) of Case and Control groups. The notation '×' denote the seven discriminant features selected by Yildiz et al. [19].

locations for a subject. For classification purposes, we treat the m/z values as equally spaced with the aim to capture the major patterns of proteomic profiles. This should not affect the classification results since all the spectra are treated the same on the realigned m/z values. Figure 3 shows the raw trajectories of the pre-processed data

of 142 cases and 146 controls, the overall mean function, and the conditional mean function of each group. Obviously, the difference between case and control groups is shown in the mean intensities at some m/z values, especially at the seven features selected by Yildiz et al. [19]. We analyze the MALDI MS spectra by the functional random-effects models in Section 2.1. The local linear smoothing methods are applied in estimation of the model components but the measurement errors are not taken into account. The bandwidths of one- and two-dimensional smoothing methods are chosen by the cross-validation method. Based on the marginal FPCA model (1), the four leading eigenfunctions, the proportions of total variance explained, and the scatter plots of pairwise FPC scores are displayed in Figure 4. It shows that some of the cases and controls can be easily separated through the distributions of marginal FPC scores. This indicates FPCA could be a proper dimension reduction technique for this data set. We further estimate the eigenfucnctions of each group based on

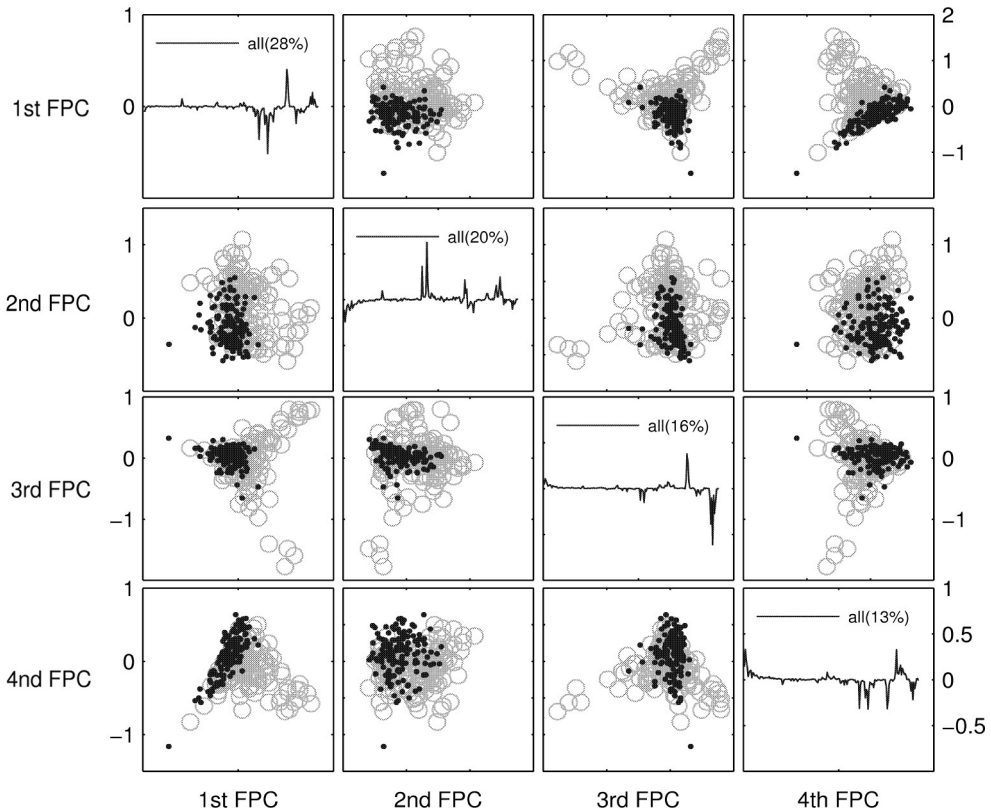

Figure 4. The first four marginal eigenfunctions (diagonal panels) and the scatter plots of pairwise FPC scores based on model (1). The notations '∘' and '•' represent the case and control groups, respectively. The percentages in parentheses indicate the proportions of total variance explained by the principal components.

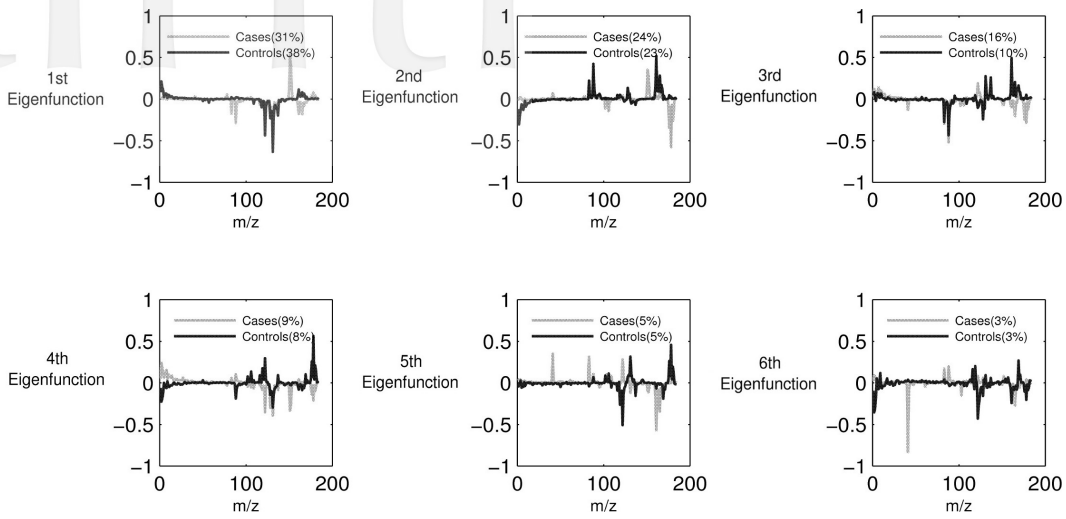

Figure 5. The first five conditional eigenfunctions of two groups based on model (3). The percentages in parentheses indicate the proportions of total variance explained by the principal components.

the conditional model (3). The first six eigenfunctions and corresponding proportions of total variance explained are presented in Figure 5, which shows very distinct covariance structures between case and control groups. Therefore, the two clusters of this data set are different in the structure of the means as well as the modes of variation.

The three compared classification methods in simulation study are implemented to the same training and test sets of Yildiz et al. [19]. For the FLR and BPCC methods, the numbers of principal components are automatically chosen by the cross-validation method based on maximizing the classification accuracy of training data. For the BPCC approach, we also consider setting equal numbers of principal components ($M_1 = M_2$) for two clusters. The results of accuracy, sensitivity, and specificity of the blinded test set obtained by different methods are shown in Table 2. Compared with the classification results based on the seven features, the FPCA-based functional classification approaches, FLR and BPCC, improve the accuracy, sensitivity, and specificity. The BPCC method under $M_1 = M_2$ especially outperforms the others. Moreover, we investigate the performance of all the methods by 4-fold cross-validation. Table 3 shows the averages and standard deviations of the three classification measures for 100 4-fold cross-validation replicates. Overall, the FPCA-based functional classification methods perform better than the multivariate data approache in terms of average and standard deviation. The FLR obtains higher sensitivity than BPCC when the sample size of control group is restricted to four times the size of case group, while the BPCC method performs better than FLR in

Table 3. Averages and standard deviations of accuracy, sensitivity, and specificity based on 100 four-fold cross-validation replicates of MALDI MS data.

$n_1 : n_2$	Method	Accuracy(SD)	Sensitivity(SD)	Specificity(SD)
No Restriction	*LR* (7 features)	0.7640 (0.0430)	0.6693 (0.0754)	0.8587 (0.0601)
	LR	0.7828 (0.0543)	0.7597 (0.0772)	0.8006 (0.0796)
	FLR	0.9182 (0.0301)	0.9084 (0.0486)	0.9329 (0.0495)
	BPCC	0.9296 (0.0292)	0.9362 (0.0464)	0.9238 (0.0460)
	BPCC ($M_1 = M_2$)	0.9295 (0.0283)	0.9149 (0.0502)	0.9452 (0.0382)
Restricted to 1:4	*LR* (7 features)	0.8628 (0.0398)	0.4628 (0.1625)	0.9628 (0.0334)
	LR	0.8962 (0.0512)	0.7519 (0.1573)	0.9322 (0.0497)
	FLR	0.9369 (0.0389)	0.8253 (0.1169)	0.9649 (0.0345)
	BPCC	0.9328 (0.0353)	0.7883 (0.1425)	0.9690 (0.0325)
	BPCC ($M_1 = M_2$)	0.9126 (0.0374)	0.6553 (0.1656)	0.9769 (0.0275)

all three measures under the ratio of sample sizes without restriction. In summary, the results show that the FPCA-based functional classification provides another useful tool for diagnosis of lung caner via MALDI MS data.

5. Concluding remarks

The proposed BPCC method is shown to perform reasonably well when different groups of curve data have distinct means and eigenspaces through numerical studies. When shape patterns of curves are of primary interest in classification, we may consider the correlation-based distance measures such as d_{FC} and d_{RC} defined in Section 2.2. We have demonstrated how to use the FPCA-based functional classification to analyze MALDI MS data. The results show that FPCA is useful to explore the structures of cases and controls and provide additional insight of the spectra data. In addition, the functional data approach avoids reliance on feature detection. Classification of MS data via functional data approach improves the accuracy, sensitivity, and specificity for considering the whole profile of a spectrum. Furthermore, the BPCC method takes the within-spectrum correlation into account, which facilitates classifying the proteomic MS data. In a future work, it is interesting to extend the proposed method by adding informative clinical covariates.

ACKNOWLEDGEMENTS

We thank Dr. Yu Shyr of the Cancer Biostatistics Center, Vanderbilt University, for allowing the use of MALDI serum data set. This research was supported by grant from the National Science Council (NSC 99-2118-M-032-007).

References

[1] A. M. Aguilera, M. Escabias and M. J. Valderrama (2008). Discussion of different logistic models with functional data, application to systemic lupus erythematosus, *Computational Statistics & Data Analysis*, 53, 151-163.

[2] J.-M. Chiou and P.-L. Li (2007). Functional clustering and identifying substructures of longitudinal data, *Journal of the Royal Statistical Society: Series B*, 69, 679-699.

[3] J.-M. Chiou and P.-L. Li (2008). Correlation-based functional clustering via subspace projection, *Journal of the American Statistical Association*, 69, 679-699.

[4] M. Escabias, A. M. Aguilera and M. J. Valderrama (2004). Principal component estimation of functional logistic regression: Discussion of two different approaches, *Journal of Nonparametric Statistic*, 16, 365-384.

[5] F. Ferraty and P. Vieu (2006). *Nonparametric Functional Data Analysis: Theory and Practice*, Springer Verlag, New York.

[6] N. E. Heckman and R. H. Zamar (2000). Comparing the shapes of regression functions, *Biometrika*, 87, 135-144.

[7] G. M. James and T. J. Hastie (2001). Functional linear discriminant analysis for irregularly sampled curves, *Journal of the Royal Statistical Society: Series B*, 63, 533-550.

[8] G. M. James (2002). Generalized linear models with functional predictors, *Journal of the Royal Statistical Society: Series B*, 64, 411-432.

[9] X. Leng and H. G. Müller (2005). Classification using functional data analysis for temporal gene expression data, *Bioinformatics*, 22, 68-76.

[10] B. Mallick, D. Ghosh and M. Ghosh (2005). Bayesian classification of tumours by using gene expression data, *Journal of the Royal Statistical Society: Series B*, 67, 219-234.

[11] J. S. Morris, P. J. Brown, R. C. Herrick, K. A. Baggerly and K. R. Coombes (2008). Bayesian analysis of mass spectrometry proteomic data using wavelet-based functional mixed models, *Biometrics*, 64, 479-489.

[12] H. G. Müller (2005). Functional modelling and classification of longitudinal data, *Scandinavian Journal Statistics*, 32, 223-240.

[13] H. G. Müller and U. Stadtmüller (2005). Generalized functional linear models, *Annals of Statistics*, 33, 774-805.

[14] C. Park, J.-Y. Koo, S. Kim, I. Sohn and J.-W. Lee (2008). Classification of gene functions using support vector machine for time-course gene expression data, *Computational Statistics & Data Analysis*, 52, 2578-2587.

[15] J. O. Ramsay and B. W. Silverman (2005). *Functional Data Analysis* (2nd ed.), Springer Verlag, New York.

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

[17] Y. Shyr and K. Kim (2003). Weighted flexible compound covariate method for classifying microarray data, In D. P. Berrar, W. Dubitzky and M. Granzow (Eds.), 186-200, *A Practical Approach to Microarray Data Analysis*, Kluwer Academic, New York.

[18] F. Yao, H. G. Müller, A. J. Clifford, S. R. Dueker, J. Follett, Y. Lin, et al. (2003). Shrinkage estimation for functional principal component scores with application to the population kinetics of plasma folate, *Biometrics*, 59, 676-685.

[19] P. B. Yildiz, Y. Shyr, J. S. M. Rahman, N. R. Wardwell, L. J. Zimmerman, B. Shakhtour, et al. (2007). Diagnostic accuracy of MALDI mass spectrometric analysis of unfractionated serum in lung cancer, *Journal Thoracic Oncology*, 2, 893-901.