Computational Statistics and Data Analysis 52 (2008) 4790–4800

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

journal homepage: www.elsevier.com/locate/csda

Classification of functional data: A segmentation approach

Bin Li*, Qingzhao Yu

Louisiana State University, 70803 Baton Rouge, LA, United States

ARTICLE INFO

Article history:
Received 27 September 2007
Received in revised form 20 March 2008
Accepted 21 March 2008
Available online 30 March 2008

ABSTRACT

We suggest a classification and feature extraction method on functional data where the predictor variables are curves. The method, called *functional segment discriminant analysis* (FSDA), combines the classical linear discriminant analysis and support vector machine. FSDA is particularly useful for irregular functional data, characterized by spatial heterogeneity and local patterns like spikes. FSDA not only reduces the computation and storage burden by using a fraction of the spectrum, but also identifies important predictors and extracts features. FSDA is highly flexible, easy to incorporate information from other data sources and/or prior knowledge from the investigators. We apply FSDA to two public domain data sets and discuss the understanding developed from the study.

1. Introduction

Data arising in a wide range of scientific investigations are often obtained in the form of curves and the observed data consist of sets of curves sampled on a fine grid. The book by Ramsay and Silverman (2005) provides a clear overview of the foundations and applications of functional data analysis (FDA). Although functional data shares many common principles with multivariate data, they are different in the "atom" of a statistical analysis. From the FDA's perspective, each observed instance is represented as a function (with some random noise) in an infinite-dimensional space, while each observation is only a realization of the function at a given point. Most existing approaches for projecting the functional data on to a finite-dimensional space can be categorized as *regularization* or *filtering* methods. The regularization approach works on a discretization of the input space, such as time interval. Since the resulting data vectors are often correlated and high-dimensional, some form of regularization constraint needs to be imposed (e.g. DiPillo (1976), Friedman (1989) and Hastie et al. (1995)). On the other hand, the filtering approach approximates each function by a linear combination of a finite number of basis functions, and represents the data by the resulting basis coefficients.

One of the main tasks in FDA is classification: one wants to predict the physical or chemical properties of a sample via its functional form. However, due to the complexity and high dimensionality of functional data, it is challenging to build a discriminant model that is reasonably flexible to accommodate the important features, and yet feasible to fit. For example, *irregular* functional data, characterized by spatial heterogeneity and local patterns like spikes, are commonly encountered in the biomedical field such as spectrometric data.

Linear discriminant analysis (LDA) and the support vector machine (SVM) (Boser et al., 1992; Vapnik, 1995) are two popular classification methods. LDA is a time-honoured tool for classification and data reduction. Although LDA enjoys a number of nice properties, e.g. it is reasonably robust to non-normality and different class covariances, it has the following two deficiencies. (1) It is too rigid when the class boundaries (in the predictor space) are highly nonlinear. (2) It is too flexible when there are many highly-correlated predictors, which is often the case in FDA. The support vector machine has demonstrated superior performances and can be easily extended to nonlinear feature space via a technique called the *kernel*

[*] Corresponding author. Tel.: +1 225 578 1343.
E-mail address: bli@lsu.edu (B. Li).

0167-9473/$ – see front matter © 2008 Elsevier B.V. All rights reserved.
doi:10.1016/j.csda.2008.03.024

trick. However, since its decision rule utilizes all the variables without discrimination, the standard SVM suffers from the presence of redundant variables (Hastie et al., 2001; Guyon et al., 2002).

In this paper, we suggest a classification method, called *functional segment discriminant analysis* (FSDA), for functional data that combines the LDA as a data reduction tool and the support vector machine as the classifier. Originally, we aimed to propose this method specifically for irregular functional data, however, the method is also highly competitive on data without obvious local patterns.

The outline of this paper is as follows. After describing the classification problems and notations, Section 2 presents the details of the proposed method. In Section 3 we present (1) the prediction performance of FSDA together with other competitors on two real datasets; (2) the simulation studies that provide indications on the expected performance of our method under different scenarios. Related issues are discussed in Section 4.

2. Methodology

In this paper, we focus on the functional data which were sampled on a fine grid. If the data are not sampled on a regular grid such as sparse functional data, then we can pre-smooth the curves and discretize its estimated function to a fine grid. Consider a classification problem with K classes and N training samples. One is given N observations of the form

$$\{\mathbf{x}_i\}_1^N = \{x_{i1}, \ldots, x_{ip}\}_1^N \tag{1}$$

considered to be a random sample of functions to a fine grid of p equally spaced values, indexed by $T = \{1, \ldots, p\}$. Hence, there are p predictors (independent variables) in total. In engineering and spectrometry, these predictors are often called *channels*. To reach out to all readers, we refer to the predictor as *channel* from now on. Let $y_i \in \{1, \ldots, K\}$ be the class status for the ith observation and C_k be the indices of the n_k samples in class k. The mean of the jth channel in class k is $\bar{x}_{jk} = \sum_{i \in C_k} x_{ij}/n_k$, the overall mean for jth channel is $\bar{x}_{j\cdot} = \sum_i x_{ij}/N$.

As a classification problem, a classifier $f(\mathbf{x})$ is constructed based on the training data $\{\mathbf{x}_i, y_i\}_1^N$. A common learning task is to achieve accurate prediction, i.e. given the value of an input vector $\mathbf{x}$, the classifier $f(\mathbf{x})$ makes a good prediction of the class status, denoted by $\hat{y}$.

Sections 2.1 and 2.2 present the two-stage feature extraction procedure in FSDA, in which linear discriminant variables are extracted on short curve segments. The linear discriminant variables are further input into the support vector machine which generates a nonlinear classification boundary on low dimensional space. This is presented in Section 2.3. The model selection issue is discussed in Section 2.4.

2.1. Marker selection by F-statistic

First, for each channel $j \in \{1, \ldots, p\}$, an F-statistic F_j is calculated by

$$F_j = \frac{\sum_{k=1}^{K} n_k(\bar{x}_{jk} - \bar{x}_{j\cdot})^2/(K-1)}{\sum_{k=1}^{K} \sum_{i \in C_k} (x_{ij} - \bar{x}_{jk})^2/(N-K)}. \tag{2}$$

As the F-statistic measures the overall separation among all the class clusters, the channels with a large F-statistic can be used to judge the class status. We call these channels *markers* from now on, since they indicate the class status (just as the term *biomarker* used in biological science). However, since the functional data are usually autocorrelated, the markers often form clusters. In order to identify these clusters without choosing too many markers, the first m *h-separated* channels with the largest F-statistic are selected as markers. *H-separated* refers to the index of all m selected markers which should be separated from each other by at least h. For example, $S = (1, 3, 20)$ is not a valid index because the gap between 1 and 3 is less than $h = 4$, whereas $S = (1, 5, 20)$ is valid. Note that m and h are two tuning parameters in FSDA. The issue of how to find the optimal m and h is discussed in Section 2.4.

2.2. Feature extraction by LDA

One purpose of LDA is to utilize the class information in finding a sequence of orthogonal projections $\mathbf{w}$, in which the total K class centroids are most separated. To that purpose LDA considers maximizing the following objective:

$$R(\mathbf{w}) = \frac{\mathbf{w}^{\mathrm{T}} V_B \mathbf{w}}{\mathbf{w}^{\mathrm{T}} V_W \mathbf{w}} \tag{3}$$

where V_B is the between-class covariance matrix and V_W is the within-class covariance matrix. They are defined as

$$V_B = \sum_{k=1}^{K} n_k(\bar{\mathbf{x}}_k - \bar{\mathbf{x}}_{\cdot})(\bar{\mathbf{x}}_k - \bar{\mathbf{x}}_{\cdot})^{\mathrm{T}} \tag{4}$$

$$V_W = \sum_{k=1}^{K} \sum_{i \in C_k} (\mathbf{x}_i - \bar{\mathbf{x}}_k)(\mathbf{x}_i - \bar{\mathbf{x}}_k)^{\mathrm{T}} \tag{5}$$

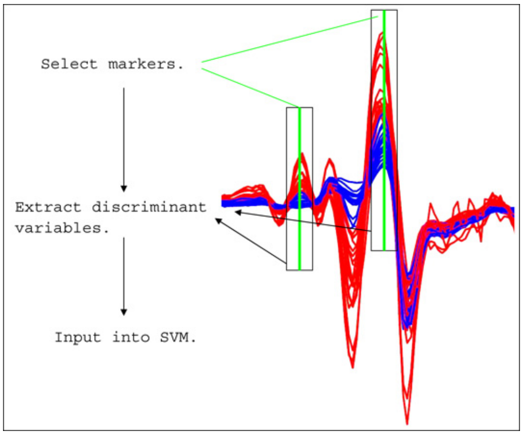

Fig. 1. Illustration of FSDA.

where $\bar{\mathbf{x}}_k = (\bar{x}_{1k}, \ldots, \bar{x}_{pk})$ is the centroid for group k and $\bar{\mathbf{x}}_. = (\bar{x}_{1.}, \ldots, \bar{x}_{p.})$ is the overall centroid. It is known that the projections $\mathbf{w}$ in Eq. (3) are the leading eigenvectors of $V_W^{-1}V_B$ with the largest eigenvalues. For data reduction, LDA entails a sequence of linear discriminant variables $\mathbf{w}^{\mathrm{T}}\mathbf{x}$ representing a subspace for which the class centroids are spread out as much as possible.

For each selected marker with index $s \in S$, LDA is applied on the curve segment which consists of the marker itself and the h nearest neighbours on each side (under the boundary constrain). For example, if $s = 12$ and $h = 6$, then the channel indices for the selected curve segment are $(6, \ldots, 12, \ldots, 18)$. The leading l $(l \leq K - 1)$ linear discriminant variables are then used as the extracted features.

2.3. Classification by SVM

Consider a binary classification problem with training set $\{\mathbf{x}_i, y_i\}_1^N$, where $\mathbf{x}_i \in \mathcal{R}^p$ is the input vector and $y_i \in \{-1, +1\}$ is its class label. SVM finds $f(\mathbf{x}) = b + \mathbf{c} \cdot h(\mathbf{x})$ which minimizes

$$\frac{1}{N}\sum_{i=1}^{N}[1 - y_i(b + \mathbf{c} \cdot h(\mathbf{x}_i))]_+ + \lambda\|\mathbf{c}\|^2 \tag{6}$$

where b is a constant, $\mathbf{c}$ is the directional vector and $\mathcal{D} = \{h_1(\mathbf{x}), \ldots\}$ is a set of basis functions. The tuning parameter λ controls the trade-off between maximizing the margin and minimizing the hinge loss $[1 - yf(\mathbf{x})]_+$, where the subscript $+$ is defined as the positive part of the argument. From the geometric perspective, SVM is a large margin classifier. Specifically, for separable data, SVM separates two classes by maximizing the margin between them, while for non-separable data, the soft-margin SVM chooses a hyperplane that splits two classes as cleanly as possible, while still maximizing the distance to the nearest cleanly split examples.

For classifying the data with complex structures where a linear separation is not plausible, a technique called *kernel trick* can be applied to convert linear SVM to nonlinear SVM. A desirable property of SVM is that its solution only depends on a subset of the training examples called *support vectors*. However, since all the input variables are used for constructing the classifier, SVM cannot select important variables and its performance will degrade when many irrelevant variables exist (Hastie et al., 2001; Guyon et al., 2002).

In FSDA, the SVM is applied to the leading discriminant variables from LDA. The SVM is run by using the **e1071** package, an **R** interface to **LIBSVM** (Chang and Lin, 2001). In this paper, the nonlinear SVM with Gaussian kernel is used. The tuning parameter λ is chosen by minimizing the cross-validation measure through grid search (for details see (Chang and Lin, 2001)). Note that for a multi-class classification problem (i.e. $K > 2$), **LIBSVM** uses the 'one-against-one' approach, in which $K(K-1)/2$ binary classifiers are trained; the appropriate class is found by a voting scheme.

2.4. Model selection

Fig. 1 illustrates the FSDA procedure described in the last three sections. Note that FSDA has three parameters: m (the number of selected curve segments), h (related to the size of the curve segment), and $l \leq K - 1$ (the number of leading discriminant variables used as features). In binary classification, l can only be one. In multi-class classification, our experience is that using the first few ($l \leq 3$) leading discriminant variables is usually enough to achieve good performance. For fixed l, optimal m and h are searched for systematically by monitoring, for instance, the cross-validation prediction error. In practice, we applied a grid search for the optimal m and h by minimizing the five-fold cross-validation errors on the training set.

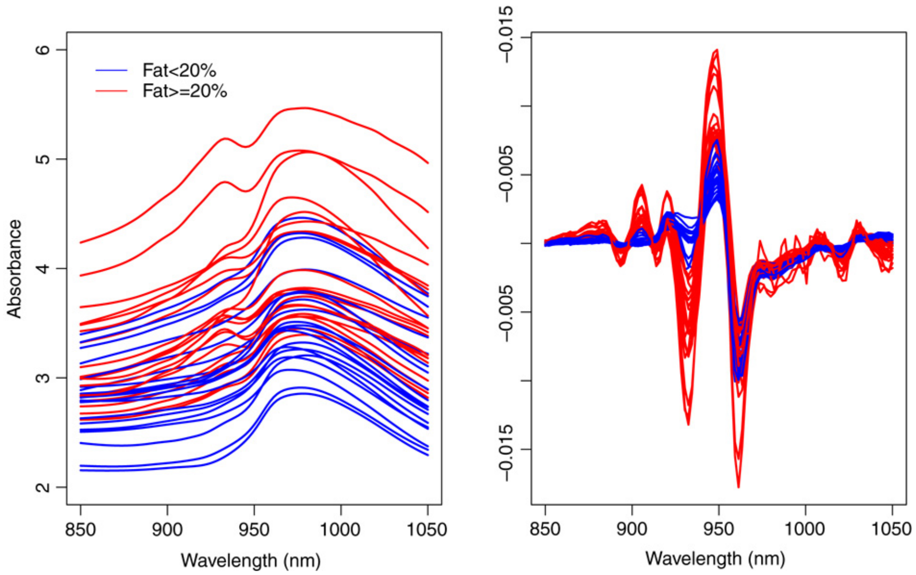

Fig. 2. Left: Sample curves for tecator data. Each class has 20 sample curves. Right: Corresponding sample curves after second order differencing.

3. Numerical studies

To evaluate the predictive performance, we compared FSDA with other competitors on two public domain datasets. Then the sensitivity analysis on the tuning parameters and identification of important channels were explored. Simulation studies are presented in order to shed light on the expected performance of FSDA under different scenarios.

3.1. Data description

Tecator data consists of 215 near-infrared absorbance spectra of meat samples, recorded on a Tecator Infratec Food Analyzer. Each observation consists of a 100-channel absorbance spectrum in the wave length range 850–1050 nm, contents of moisture (water), fat and protein. The goal here is to predict whether the fat percentage is greater than 20% from the spectra. Among 215 samples, 138 have fat percentage less than 20%. The data set is available at http://lib.stat.cmu.edu/datasets/tecator Fig. 2 shows twenty examples in each of two classes.

Phoneme data was formed by selecting five phonemes for classification based on digitized speech from TIMIT database. The dataset consists of 4509 speech frames with “aa” (695), “ao” (1022), “dcl” (757), “iy” (1163) and “sh” (872). The phonemes are transcribed as follows: “sh” as in “she”, “dcl” as in “dark”, “iy” as the vowel in “she”, “aa” as the vowel in “dark”, and “ao” as the first vowel in “water”. From each speech frame, a log-periodogram of length 256 was computed. The data, which is available at http://www-stat.stanford.edu/~tibs/ElemStatLearn/, was used in the paper on *penalized discriminant analysis* (PDA) by Hastie et al. (1995). Fig. 3 shows five examples in each of five classes.

3.2. Competitors

We compared our method FSDA with three competitors denoted as PCA, Wavelet and PDA.

- Principal component analysis (PCA) is a key dimension reduction technique used in functional data analysis. In the PCA approach, the first few leading principle components from the input matrix are extracted as features. For various issues and references of using PCA in FDA, we refer the readers to Chapter 8 of Ramsay and Silverman (2005). More recently, Leng and Müller (2006) used functional principal components to classify temporal gene expression data. Hall et al. (2006) showed the asymptotic estimation properties in functional PCA. Yao (2007) proposed a nonparametric approach for jointly modelling longitudinal and survival data using functional principal components. In order to have a fair comparison, we chose the number of leading principal components adaptively by the same scheme as for FSDA, i.e. five-fold cross-validation, and the extracted leading principal components are input to the SVM for classification.
- Wavelet bases are comparatively recent, and they have considerable promise in many FDA contexts. We refer the readers to Section 3.8 of Ramsay and Silverman (2005) for references in this area. More recently, Morris and Carroll (2006) developed wavelet-based functional mixed models. In this paper, the discrete wavelet decomposition algorithm (Mallat, 1989) was applied to the input matrix and the wavelet coefficients, after hard thresholding, are used as the extracted

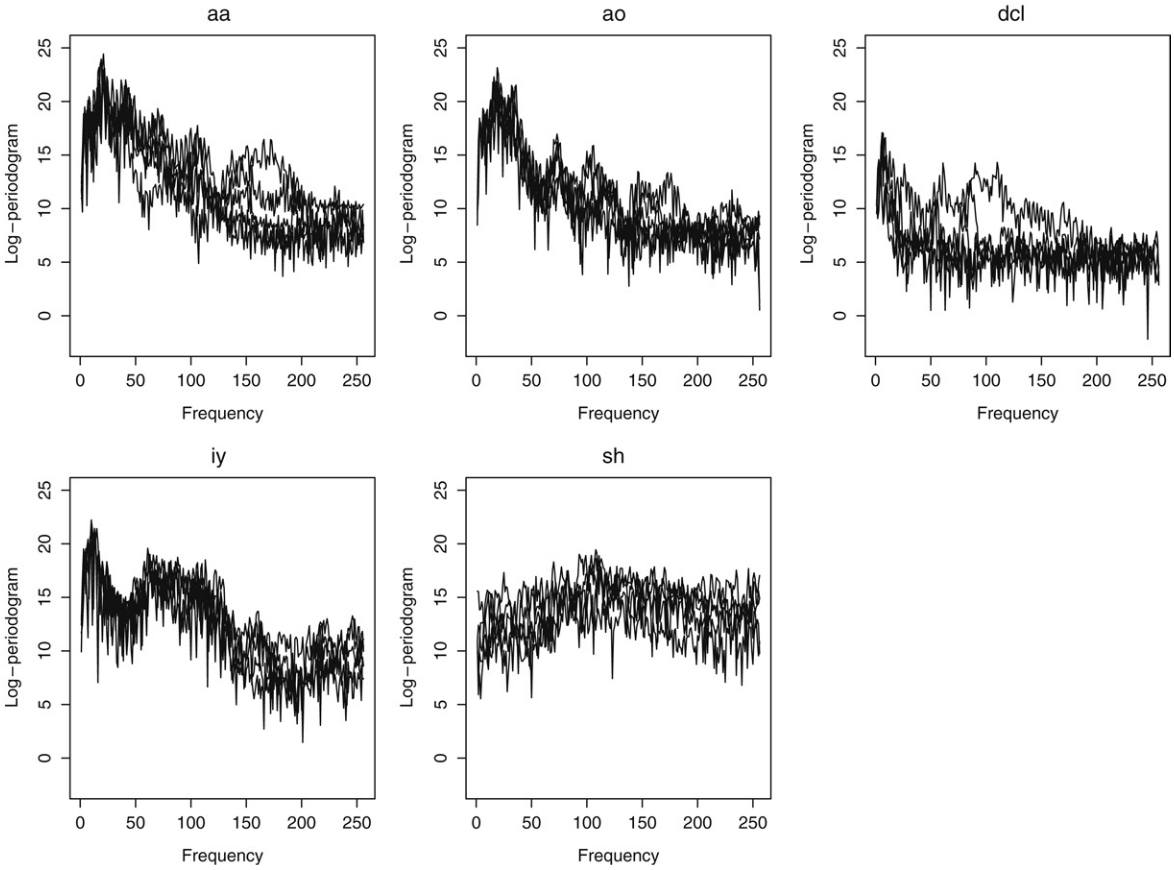

Fig. 3. Sample curves for phoneme data. Each class has five sample curves.

features. As in PCA, we chose the threshold values adaptively through five-fold cross-validation and the extracted features are imported to the SVM for classification.

- Penalized discriminant analysis, as developed by Hastie et al. (1995), produces linear combinations of the input variables that contribute to the discrimination rule as LDA does. However, unlike LDA, which is known to fail when faced with the high dimension and high correlation of adjacent spectral bands, PDA often performs well with hyperspectral data by regularizing the solutions with an additional penalty term. Note that, the penalty parameter is tuned through five-fold cross-validation.

Recently, Rossi and Villa (2006) investigated the use of a support vector machine for function data classification. In order to take into account the functional nature of the data, they suggest applying the standard filtering approach of FDA combined with SVM on the resulting basis coefficients (FSVM). We additionally compared our method with their approach in Section 3.4.

3.3. Prediction performance

In the tecator example, we first applied the second-order differencing to the raw data. Then 120 samples were randomly selected as the training set and the remaining 95 samples were used as the testing set. In the phoneme example, we randomly selected 1000 samples as the training set and the remaining 3509 samples were the testing set. As a multi-class classification problem, the first three linear discriminant variables were used as the extracted features in this example, i.e. $l = 3$, since it achieves the best prediction performance. The classification results shown in Fig. 4 are based on 50 replications for both examples. Within each replication, the randomly assigned training set was used to fit the model, where the performance is evaluated on the test set. In the tecator example, we see that FSDA outperformed the three competitors, and has the smallest classification errors for 41 (out of 50) of the trials. Furthermore, the difference between FSDA and PDA, which outperformed PCA and Wavelet, is statistically significant as indicated by the Wilcoxon signed-rank test (p-value is 0.0001). In the phoneme example, FSDA is the best on 45 trials and on average it is only 0.22% higher than the best for each trial. The difference between FSDA and other three competitors is also statistically significant as indicated by the Wilcoxon signed-rank test (all three p-values are less than 0.0001).

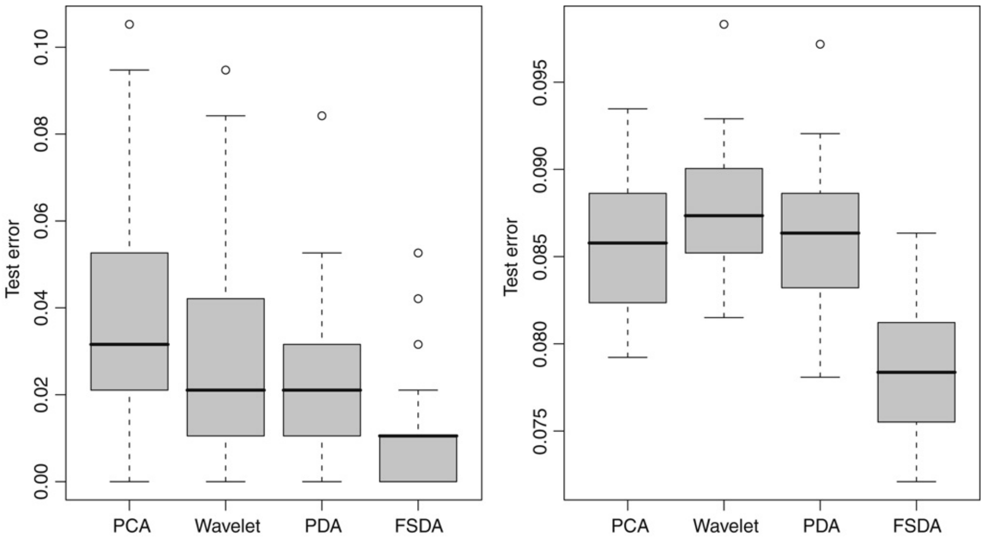

Fig. 4. Test errors in tecator (left) and phoneme (right) examples based on 50 replications.

Table 1
Mean test errors for tecator example

Derivatives	FSVM (%)	FSDA (%)
Raw data	7.5	3.98 (0.28)
First derivative	–	2.91 (0.29)
Second derivative	2.6	1.09 (0.16)

Table 2
Mean test errors for phoneme example

FSVM (%)	FSDA (%)
22	18.5 (0.2)

3.4. FSDA versus FSVM

Rossi and Villa (2006) applied their method to both tecator and phoneme data sets. For the tecator data, we use the same data splitting scheme as theirs. Table 1 compares the test results for FSDA with FSVM in the tecator example under different derivatives. The number in parentheses is the standard error for the average test error based on 50 random partitions of the training and test sets. Based on Table 1, we have the following remarks: (1) FSDA achieves better prediction performance than FSVM in both cases and the differences are large compared to standard errors. (2) Using the second order derivative of the spectrum as the input substantially reduces the test error rate in both FSDA and FSVM.

In the phoneme example, Rossi and Villa (2006) focused on the binary classification problem for “aa” against “ao”, since it is the most difficult sub-problem. Table 2 compares the results for FSVM and FSDA methods in phoneme example. Like the tecator example, the number in parentheses is the standard error for the mean test error based on 50 random partitions of training and test sets, and FSDA achieves a lower test error rate than FSVM.

In both tecator and phoneme examples, FSDA achieves better prediction performance than FSVM. A natural explanation is that the standard SVM suffers from the presence of irrelevant input variables (i.e. basis coefficients), which means, unlike FSDA, FSVM does not consider the class labels in the feature extraction process (i.e. reduce the dimension via the filtering/projection process).

3.5. Sensitivity analysis on tuning parameters

Fig. 5 shows the average test errors, in the 50 replications described above, from using different values for the tuning parameters m and h. Note that since cross-validation provides an unbiased estimate of the expected value of the prediction error, Fig. 5 can be viewed as an approximated distribution of the cross-validation errors. In the tecator example, we see the optimal h is 1 and using a different m doesn't affect the prediction performance much. The small optimal value for h implies the data entails local differential patterns, e.g. spikes, between two classes. In the phoneme example, we see that as h increases, the test errors decrease and stabilize after h reaches 6. The optimal m is around 7. The relatively large optimal

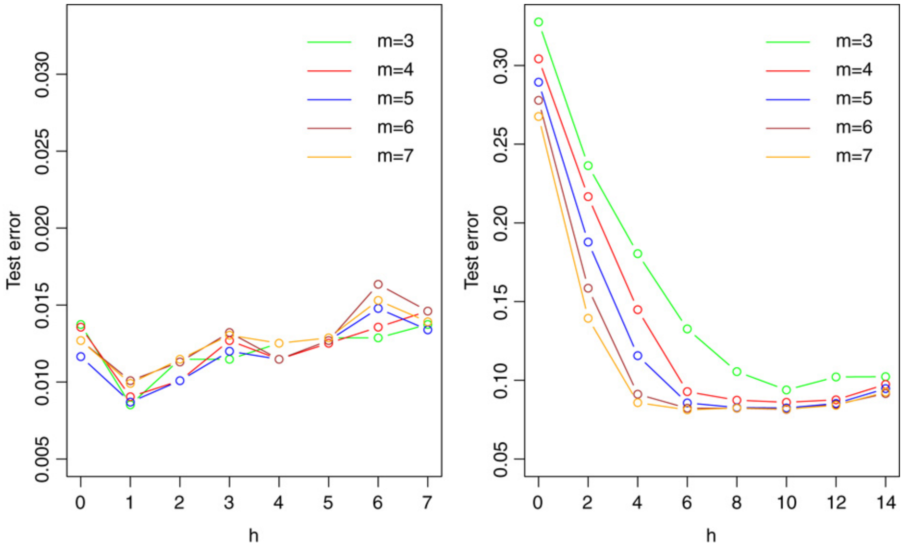

Fig. 5. Average test errors in tecator (left) and phoneme (right) examples with different h and m values.

values for both h and m imply that the data is characterized by some sort of global differential patterns, e.g. 'bumps' instead of 'spikes'.

3.6. *Mapping selected channels*

Identification of markers, the channels indicating class status, is one of the key problems in functional data analysis. FSDA automatically selects markers and their 'neighbourhood'. In practice, we can apply the method on resampled data several times. Those channels that have been included in the selected curve segments most frequently can be viewed as strong candidates for the 'true' markers. Fig. 6 shows the relative frequencies (in 50 replications) of being the marker (black) and included in curve segments (gray) for each channel in tecator (left) and phoneme (right) examples. In addition, since the markers are separated by only at least h channels, it is possible that two selected curve segments overlap. In the tecator example, we see the channels around 905, 935 and 1045 nm have the highest frequencies of being selected as markers and included in the curve segments. Shown in the right panel of Fig. 2, we see these regions do have a good separation of two classes. In the phoneme example, we see that the channels selected most are in the low frequency region. This agrees with the fact that the signals tend to stabilize in the high frequency region.

3.7. *Simulation studies*

In order to gain insight into the nature of our method, we applied it to two simulated binary classification examples with known underlying generation mechanism. Fig. 7 illustrates the underlying generation mechanism and three sample curves in each class. Panel (a) shows the underlying target functions for each class (blue and red) in Example 1. Panel (b) shows three sample curves in each class in Example 1. Panels (c) and (d) show the generation mechanism and sample curves in Example 2, respectively. Note that in both examples the simulated curves are generated by adding some correlated Gaussian noise to their corresponding target function. We see that the difference between two classes in Example 1 lies only in the middle spikes, where in Example 2 the difference lies in the whole spectrum. In other words, Example 1 (2) has local (global) differential pattern.

Fig. 8 shows the test errors in Example 1 and 2 based on 50 replications. For each replication, the model is fitted to 300 randomly generated curves (about 150 in each class), while the performance is evaluated on an independent test set containing 3000 observations. We see that in Example 1, our method is the best in 47 runs. This agrees with our original motivation that FSDA is particularly useful in detecting local patterns. In Example 2, due to the (global) nature of the generation mechanism, it is no wonder that PCA performs the best. PDA and FSDA perform slightly worse than PCA but better than Wavelet method.

Fig. 9 shows the average test errors (on 50 replications) by using different values for the tuning parameters m and h in Example 1 and 2. In Example 1, we see the optimal h is 5 and using a different m doesn't seem to have much effect on the predictive performance. In Example 2, we see that as h increases, the test errors decrease sharply at the beginning and stabilize after h reaches 40. The small (large) optimal value for h in Example 1 (2) agrees with its underlying generating mechanism. Interestingly, we see that Fig. 9 is similar to Fig. 5. Namely, in both tecator and Example 1, we see a sharp elbow at the optimal h, where the ratio of the optimal value of h to the width of the whole spectrum is small (i.e in both cases the

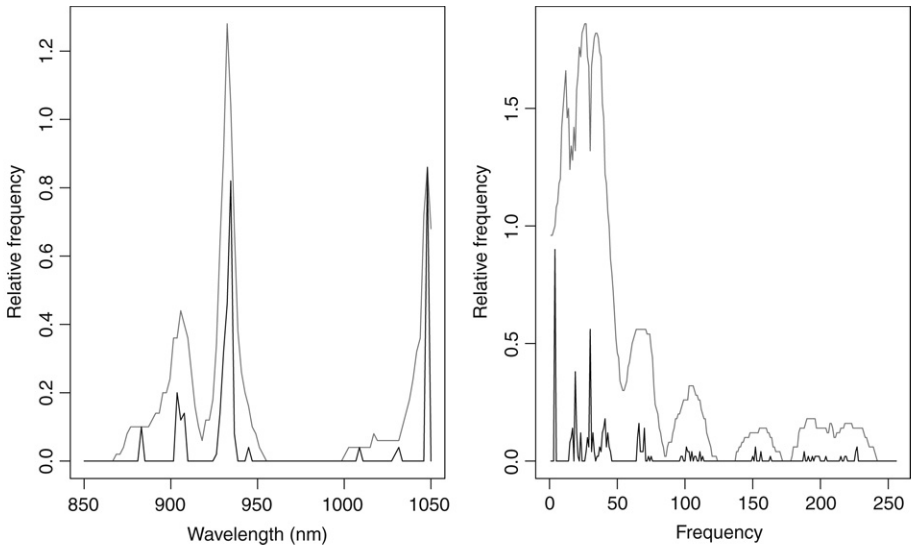

Fig. 6. The relative frequencies of being the marker (black) and included in curve segments (gray) for each channel in tecator (left) and phoneme (right) examples.

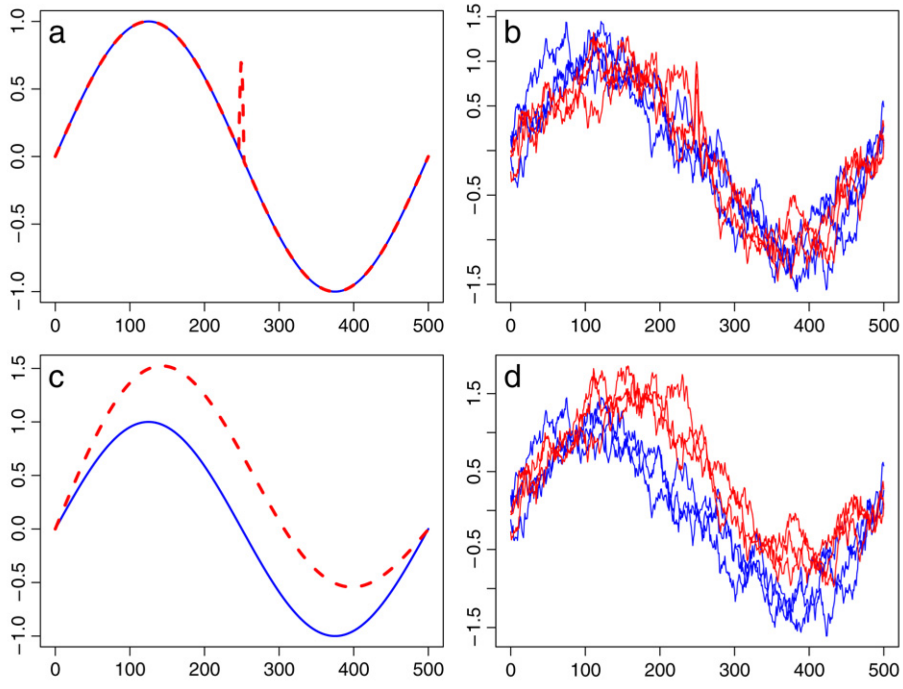

Fig. 7. Illustration of the generation mechanism and samples curves in Example 1 and 2. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

ratio is about 1%). On the other hand, in both phoneme and Example 2, we see a smooth elbow and the optimal *h* is relatively large. This implies that tecator data is characterized by local patterns, while phoneme data features global patterns.

Fig. 10 shows the relative frequencies (in 50 replications) of being the marker (black) and included in curve segments (gray) for each channel in Example 1 and 2. In Example 1, we see FSDA selects the middle spike in every replication (the peak height of being included in curve segment is 1). In Example 2, we see that the channels being selected as markers are scattered between 150 and 350, and the highest frequency region of the gray curve (being included in selected curve segments) is located between 200 and 300. This not only confirms the global differential patterns in Example 2 but also indicates the region where the target functions differ most between two classes.

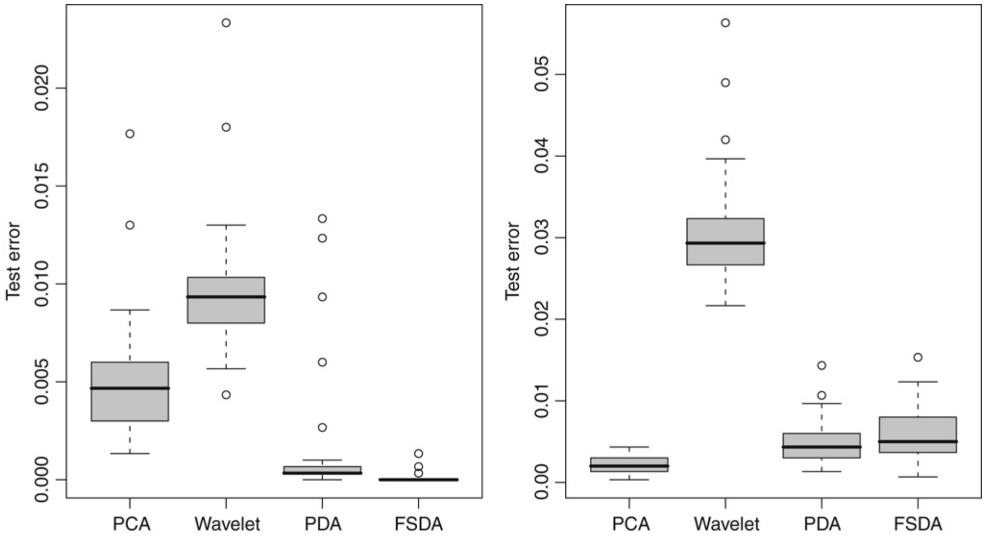

Fig. 8. Test errors in Example 1 (left) and Example 2 (right) based on 50 replications.

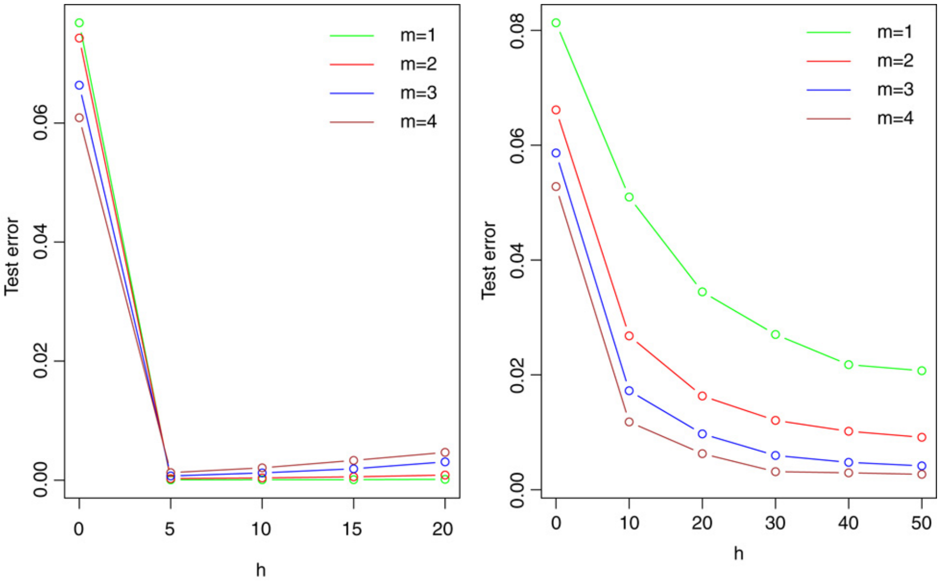

Fig. 9. Average test errors in Example 1 (left) and Example 2 (right) with different h and m values.

4. Discussion

The intrinsic limitation of classical LDA is that its objective function in Eq. (3) requires the within-class covariance matrix V_W to be nonsingular. However, due to the high dimensionality and autocorrelated structure, V_W is often (or close to) singular in functional data. Chen et al. (2000) even suggested searching for the discriminant information from the null space of V_W in image data analysis.

In FSDA, the linear discriminant variables are extracted from the short curve segments (low dimensional space), which potentially contain most discriminant information. The discriminant variables are then input into a nonlinear support vector machine, which enables FSDA to have a nonlinear classification boundary on a low-dimensional feature space. Hence, FSDA is well suited for small n (sample size) and large p (dimensionality) problems, which is not uncommon for functional data.

The comparison of PCA and LDA is an old topic. It is generally believed that classification methods based on LDA achieve better performance than those based on PCA. However, recently Martínez and Kak (2001) pointed out that when the training dataset is small in n (relative to p), PCA can outperform LDA. In FSDA, the *small n large p problem* is circumvented by applying LDA on short curve segments instead of the whole spectrum. A natural explanation of the superiority of FSDA over the Wavelet method is that the latter does not consider the class labels in the feature extraction process. The reason that

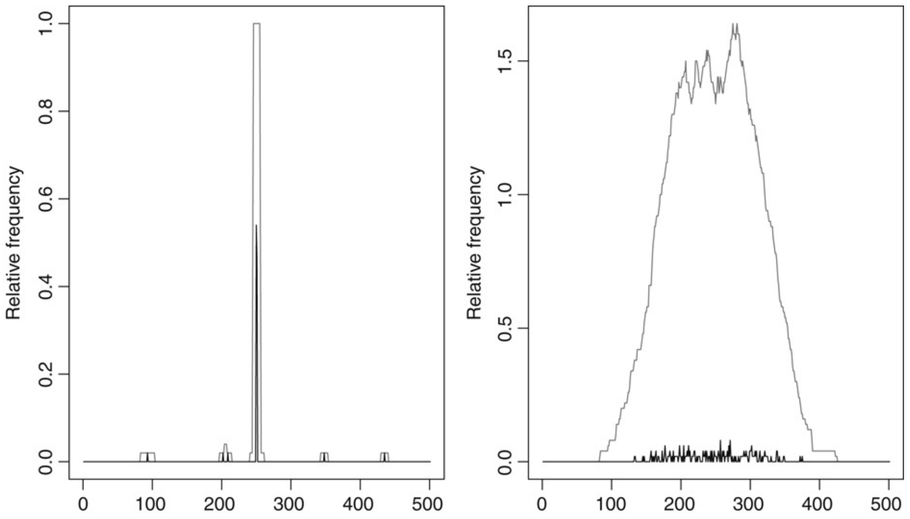

Fig. 10. The relative frequencies of being the marker (black) and included in curve segments (gray) for each channel in Example 1 (left) and 2 (right).

FSDA outperform PDA is two-fold: (1) PDA uses all the input variables for constructing the classifier, thus its performance will degrade when many irrelevant variables exist; (2) unlike PDA, FSDA provides a more flexible nonlinear classification boundary by using the Gaussian kernel in SVM.

Most of the current methods in functional data analysis are based on smoothing (or regularization) methods using global penalties and bandwidth (e.g. Rice and Silverman (1991), Zhang et al. (1998), Wang (1998), Rice and Wu (2001) and Guo (2002)), which are not well suited for modelling irregular data. Unlike these methods, FSDA cuts the curves into short pieces and utilizes those with the most discriminant power. As a result, it relieves the storage and computation burden in handling the commonly high-dimensional functional data. In the tecator and phoneme examples, the percentage of channels being used in FSDA is about 10% and 40%, respectively. In addition, FSDA is highly flexible, i.e. it is easy to incorporate information from other data sources and/or prior knowledge from the investigators. This is because FSDA separates the feature extraction and classification. Thus, in practice, we may input discriminant variables as well as the additional features to the classifier.

In FSDA, since each curve is treated as a vector, it can be applied to any functional data sampled on a fine grid with no need to choose a suitable basis function set for a specific dataset. On the other hand, unlike other regularization methods, FSDA incorporates partial order information by including the 'neighbourhoods' of the selected markers. It is worth noting that a major drawback of FSDA is that it is relatively more sensitive to misalignment than to filtering approach. Thus, we suggest the user aligns the data before applying FSDA if misalignment could be a problem.

In functional data analysis, we may often find it useful to remove trends by using first-order derivatives. Naturally higher order derivative spectra can also be obtained. This not only exploits the intrinsic smoothness in process, but also may get closer to the underlying driving forces at work. Thus, in practice, we may include the order of derivatives as one of the design parameters and choose its optimal value adaptively. For example, in the tecator data, we found using the second order derivative spectra achieves better performance than using the others.

References

Boser, E., Guyon, M., Vapnik, V., 1992. A training algorithm for optimal margin classifiers, in: Proceedings of the Fifth ACM Workshop on Computational Learning Theory, Pittsburgh, PA, pp. 144–152.

Chang, C., Lin, C., 2001. LIBSVM: A library for support vector machines. Available at: http://www.csie.ntu.edu.tw/~cjlin/libsvm.

Chen, L., Liao, H., Ko, M., Lin, J., Yu, G., 2000. A new LDA-based face recognition system which can solve the small sample size problem. Pattern Recognition 33, 1713–1726.

DiPillo, P.J., 1976. The application of bias to discriminant analysis. Communications in Statistics, Part A - Theory and Methods A5, 843–854.

Friedman, J., 1989. Regularized discriminant analysis. Journal of the American Statistical Association 84, 165–175.

Guo, W., 2002. Functional mixed effects models. Biometrics 58, 121–128.

Guyon, I., Weston, J., Barnhill, S., Vapnik, V., 2002. Gene selection for cancer classification using SVM. Machine Learning 46, 389–422.

Hall, P., Müller, H.-G., Wang, J.-L., 2006. Properties of principal component methods for functional and longitudinal data analysis. The Annals of Statistics 34, 1493–1517.

Hastie, T., Buja, A., Tibshirani, R., 1995. Penalized discriminant analysis. Annals of Statistics 23, 73–102.

Hastie, T., Tibshirani, R., Friedman, J., 2001. The Elements of Statistical Learning. Springer, NY.

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

Mallat, S., 1989. A theory for multiresolution signal decomposition: The wavelet representation. IEEE Transaction on Pattern Analysis and Machine Intelligence 11, 674–693.

Martínez, A.M., Kak, A.C., 2001. PCA versus LDA. IEEE Transactions on Pattern Analysis and Machine Intelligence 23, 228–233.

Morris, J.S., Carroll, R.J., 2006. Wavelet-based functional mixed models. Journal of the Royal Statistical Society, Series B 68, 179–199.

Ramsay, J.O., Silverman, B.W., 2005. Functional Data Analysis. Springer, NY.

Rice, J.A., Silverman, B.W., 1991. Estimating the mean and covariance structure nonparametrically when the data are curves. Journal of the Royal Statistical Society, Series B 53, 233–243.

Rice, J.A., Wu, C.O., 2001. Nonparametric mixed effects models for unequally sampled noisy curves. Biometrics 57, 253–259.
Rossi, F., Villa, N., 2006. Support vector machine for functional data classification. Neurocomputing 69, 730–742.
Vapnik, V., 1995. The Nature of Statistical Learning Theory. Springer, NY.
Wang, Y., 1998. Mixed effects smoothing spline analysis of variance. Journal of the Royal Statistical Society, Series B 60, 159–174.
Yao, F., 2007. Functional principal component analysis for longitudinal and survival data. Statistica Sinica 17, 965–983.
Zhang, D., Lin, X., Raz, J., Sowers, M.F., 1998. Semiparametric stochastic mixed models for longitudinal data. Journal of the American Statistical Association 93, 710–719.