Knowledge-Based Systems 97 (2016) 224–236

Contents lists available at ScienceDirect

Knowledge-Based Systems

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

Comparison study of orthonormal representations of functional data in classification

Yinfeng Meng[a,b], Jiye Liang[a,c,*], Yuhua Qian[a,c]

[a] School of Computer and Information Technology, Shanxi University, Taiyuan 030006, Shanxi, China
[b] School of Mathematical Sciences, Shanxi University, Taiyuan 030006, Shanxi, China
[c] Key Laboratory of Computational Intelligence and Chinese Information Processing of Ministry of Education, Taiyuan 030006, Shanxi, China

ARTICLE INFO

Article history:
Received 23 June 2015
Revised 20 November 2015
Accepted 25 December 2015
Available online 11 January 2016

Keywords:
Functional data
Orthonormal representation
Orthonormal basis
Classification

ABSTRACT

Functional data type, which is an important data type, is widely prevalent in many fields such as economics, biology, finance, and meteorology. Its underlying process is often seen as a continuous curve. The classification process for functional data is a basic data mining task. The common method is a two-stage learning process: first, by means of basis functions, the functional data series is converted into multivariate data; second, a machine learning algorithm is employed for performing the classification task based on the new representation. The problem is that a majority of learning algorithms are based on Euclidean distance, whereas the distance between functional samples is L_2 distance. In this context, there are three very interesting problems. (1) Is seeing a functional sample as a point in the corresponding Euclidean space feasible? (2) How to select an orthonormal basis for a given functional data type? (3) Which one is better, orthogonal representation or non-orthogonal representation, under finite basis functions for the same number of basis? These issues are the main motivation of this study. For the first problem, theoretical studies show that seeing a functional sample as a point in the corresponding Euclidean space is feasible under the orthonormal representation. For the second problem, through experimental analysis, we find that Fourier basis is suitable for representing stable functions(especially, periodic functions), wavelet basis is good at differentiating functions with local differences, and data driven functional principal component basis could be the first preference especially when one does not have any prior knowledge on functional data types. For the third problem, experimental results show that orthogonal representation is better than non-orthogonal representation from the viewpoint of classification performance. These results have important significance for studying functional data classification.

1. Introduction

Recent years have witnessed considerable improvements in data acquisition technology and data storage abilities. As a result, it has become imperative to classify individual systems in various research fields based on one or more data series. The underlying process of every data series is an unknown function (continuous curve), called functional data. The classification process for functional data is typically the same as that for their underlying generation functions.

At present, for the classification of functional data, there are two types of commonly used methods. One involves constructing functional classifiers, such as a functional support vector machine (SVM) by means of kernel techniques [3,33,48] and functional logistic regression [5,18,24,43,47,49], and the other is a two-stage classification method [28]. For the second method, in the first stage, usually, functional samples are represented in a finite dimensional functional subspace by means of basis functions; thus, functional data with infinite dimension becomes multivariate data, which consists of coefficients before the basis functions. In the second stage, a classical learning algorithm for finite dimensional data is used. The reason is that the high dimensionality of data series renders many data mining methods ineffective and fragile [8]. This obstacle is sometimes referred to as the "curse of dimensionality" [14]. In most data series mining problems, there is a need for dimensionality reduction and forming new data series representations [27]. It is required that the new representation preserves sufficient information for solving data series mining problems correctly. Once the basis is chosen, the optimal value for the number of basis functions can be derived from the data [48].

* Corresponding author at: School of Computer and Information Technology, Shanxi University, Taiyuan 030006, Shanxi, China. Tel./fax: +86 0351 7018176.
E-mail addresses: mengyf@sxu.edu.cn (Y. Meng), ljy@sxu.edu.cn (J. Liang), jinchengqyh@sxu.edu.cn (Y. Qian).

http://dx.doi.org/10.1016/j.knosys.2015.12.016
0950-7051/

Representing data series in the transformed domain is a common dimensionality reduction approach. Some of the popular transformation techniques are Fourier transform [15,33,53] and wavelet transform [11,16,32,37]. Functional principal component analysis(FPCA) [10,21,29,39,43,46,54–56] is a popular technique that uses statistical methods. Other methods include B-spline functions [1,3,35,59], Mercer kernel transforms [36,38], radial basis functions [4,5,26], etc.

In fact, the representation of functional data is essentially a kind of approximation of itself. In the process of machine learning of functional data, a kind of structured representation using basis functions is used to transform functional data into multivariate data, and then, the distances between functional samples are converted into the Euclidean distances between the corresponding multivariate data. However, the representability of using the corresponding multivariate data to represent functional data, and the rationality of using the distance between the corresponding two multivariate data to replace the distance between two functional samples have not been studied in detail. Therefore, the relationship of different spaces is first introduced, and then the orthonormal representation theory is employed to explain the representability and rationality.

Theoretically, under orthonormal basis, for any two different functional samples, the distance between them can be approximated based on the distance between their low-dimensional representations, which is isomorphic to the corresponding Euclidean distance. At this time, choosing an appropriate orthonormal basis is still a problem. Therefore, three kinds of common orthonormal basis and their differences are considered. The three kinds of orthonormal basis are normal Fourier basis, wavelet basis, and functional principal component basis, the eigenequation of FPCA is derived by means of variational theory.

It is well known that non-orthogonal representation can also represent a functional data series as certain multivariate data. Therefore, it is important to verify if orthogonal basis has a stronger representation ability than non-orthogonal basis for functional data under the same number of basis functions from the viewpoint of classification performance.

In order to verify the representation ability of the above orthonormal basis in classification, the extracted features(the coefficient vector, which consists of coefficients before the basis functions) of the functional data will be used in classification model construction. It has been pointed out in the literature [17] that support vector machine(SVM) and random forest are two preferred classification methods, and thus, LibSVM [12] and RandomForest [9,44] are first used to classify the functional data for three kinds of orthonormal representations. As other choices, logistic regression [29,40], K-nearest neighbor [30,31], and artificial neuron network [34,41] will also be used as classifiers for discriminating functional samples. Based on these classifiers, we shall also compare the classification performance of orthogonal representation with that of non-orthogonal representation.

The main objective of this paper is to explain the rationality behind converting functional samples into corresponding multivariate data that are to be used for training a classifier. At the same time, from the point of view of experiments, we shall explain that among the three basis candidates, Fourier basis is suitable for representing stable signals(especially, periodic functions), wavelet representation can yield better results than Fourier representation for non-stationary signals, and orthonormal basis obtained through functional principal components offers good representation ability for some functional data with complex trend characteristics. Functional principal component analysis (FPCA), in particular, can be the first choice when people do not have any prior knowledge. Furthermore, we also demonstrate that orthogonal basis is indeed better than non-orthogonal basis from the viewpoint of classification performance.

Table 1
The observation form of functional data.

Sample	t_1	t_2	$\cdots$	t_p
X_1	$X_1(t_1)$	$X_1(t_2)$	$\cdots$	$X_1(t_p)$
X_2	$X_2(t_1)$	$X_2(t_2)$	$\cdots$	$X_2(t_p)$
$\vdots$	$\vdots$	$\ddots$	$\vdots$	
X_N	$X_N(t_1)$	$X_N(t_2)$	$\cdots$	$X_N(t_p)$

The remainder of this paper is organized as follows. Some basic concepts of functional data and some approximation theory under orthonormal representation are presented in Section 2. Section 3 describes three kinds of common orthonormal representations for functional data, and in particular, the eigenequation for functional principal component is derived using the variational principle. Section 4 introduces several classification methods including LibSVM, RandomForest, logistic regression, K-nearest neighbor, and artificial neuron network. Furthermore, four classification performance indexes such as the precision, the recall, $F1$ score, and the accuracy are introduced in detail. Section 5 provides numerical studies for feature extraction and classification methods for functional data. In this section, we analyze the classification performance of three different kinds orthonormal basis, point out which kind of orthonormal basis is appropriate to represent what type of functional data, and answer whether orthogonal representation is better than non-orthogonal representation for classifying functional data for the same number of finite basis functions. Section 6 concludes the paper with some remarks and discussions.

2. Orthonormal representation for functional data

2.1. The basic concepts of functional data

Advances in data collection and storage have led to an increased presence of functional data, whose graphical representations are curves, images, or shapes [51]. The observation form of the functional data is also a two-dimensional table, which is shown in Table 1, in which $X_i(t)$ (abbreviated as X_i), $t \in I$, $i = 1, 2, \ldots, N$ is an underlying continuous and smooth function, and $X_i \in L^2(I)$, where $L^2(I)$ is the space of the square-integrable functions defined on the compact set I, $X : I \to \mathcal{R}$, $(\int_I X^2(t)dt)^{1/2} < \infty$, $\mathcal{R}$ is the real number space. At the same time, $L^2(I)$ is a separable Hilbert space with the inner product $< X, Y >= \int_I X(t)Y(t)dt$ and the norm $\|X\|_2 = (\int_I X^2(t)dt)^{1/2}$. $X_i(t_j)$ denotes the observed value for $X_i(t)$ at a discrete point t_j for the ith functional sample.

To understand the $L^2(I)$ space, the relationship among different spaces is first introduced. It is well known that the introduction of the distance is for the purpose of studying the convergence. People, therefore, defined the metric space. In the metric space, the distance between any two elements can be computed. If the concept of completion (any Cauchy sequence is a convergent sequence [58]) is introduced in the metric space, the space will become a complete metric space.

However, the metric space only has a topological structure, which restricts its application area. If a linear operation is introduced to the metric space, a linear normal space [58] can be obtained and the algebraic operation between elements can be carried out. In this case, the distance is transformed into the norm, which combines the metric and the linear operations perfectly. In other words, the linear normal space not only keeps its topological structure but also maintains its algebraic structure. The

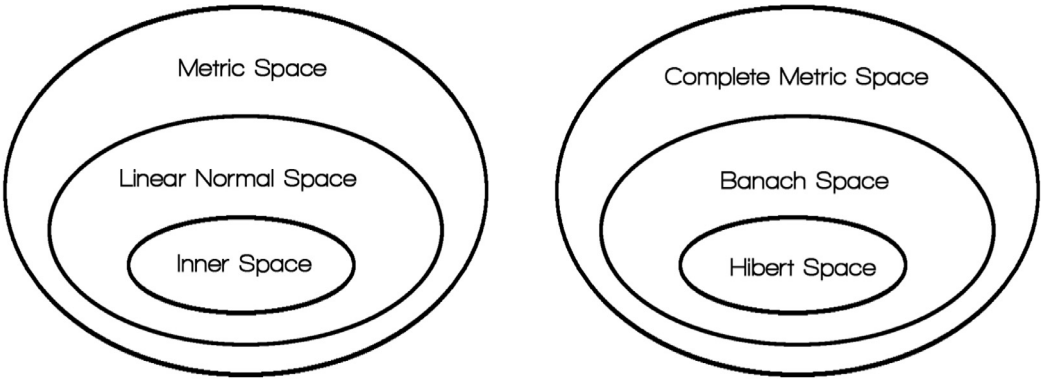

Fig. 1. The relationship of different spaces.

Banach space [2,52], especially, is a complete linear normal space, in which any element can be approached by a linear combination of basis vectors.

On the other hand, the inner product space [50,58] is also a linear normal space, in which the norm is induced by the inner product (i.e.,$\|X\|_2 =< X, X >^{1/2}$). Different from an ordinary linear normal space, in the inner product space, people can define the angle so as to further discuss the orthogonality. In particular, the Hilbert space[2] is a complete inner space, in which people can discuss both approximation and angle. In other words, any element in the Hilbert space can be infinitely approached by a linear combination of orthogonal basis.

The relationship of different spaces is clearly shown in Fig. 1.

Through Fig. 1, for a functional object, the complete metric space can be used to judge whether the object can be approached or not; the Banach space answers the problem of how to approach it (i.e., what is used to approximate it); and the Hilbert space indicates that it can be approached by a linear combination of a family of orthogonal basis. Functional data belong to the space of the square-integrable functions defined on the compact set I, just $L^2(I)$.

2.2. Some approximation theory for functional data

Owing to $L^2(I)$ being a Hilbert space, in this subsection, we discuss several important properties of the system of normalized orthogonal functions in $L^2(I)$.

Lemma 1. *[58] Let $\{\varphi_i\}$ be a system of normalized orthogonal functions in $L^2(I)$, $X \in L^2(I)$. For a given k, suppose*

$$X^{(k)} = \sum_{i=1}^{k} a_i\varphi_i, \tag{1}$$

where a_i $(i = 1, 2, \ldots, k)$ is a real number, then $\|X - X^{(k)}\|_2$ achieves its minimum value if $a_i =< X, \varphi_i >$ $(i = 1, 2, \ldots, k)$.

Lemma 2. *[58] Let $\{\varphi_i\}$ be a complete system of orthogonal functions in $L^2(I)$, $X \in L^2(I)$. Given $a_i =< X, \varphi_i >$ $(i = 1, 2, \ldots)$, one has that*

$$\lim_{k\to\infty} \|X^{(k)} - X\|_2 = 0. \tag{2}$$

Remark 1. When $\{\varphi_i\}$ is a complete system of normalized orthogonal functions, as per Lemma 1, one knows that $X^{(k)}$ is the optimal approximation of X in the k-dimensional subspace H_0 of $L^2(I)$, where H_0 is spanned by $\{\varphi_1, \varphi_2, \ldots, \varphi_k\}$. Moreover, Lemma 2 shows that the approximation performance will improve as k increases.

Lemma 3. *[57] Let $\mathscr{X}$ be a Hilbert space. If $\{\varphi_i\}$ is a complete system of normalized orthogonal functions of $\mathscr{X}$, then $\mathscr{X}$ has the corresponding Parseval equivalent formulation, i.e., $\forall X \in \mathscr{X}$,*

$$\|X\|_2^2 = \sum_{a\in A} | < X, \varphi_i > |^2, \tag{3}$$

where A is an index set.

Remark 2. Since $L^2(I)$ is a Hilbert space, if $\{\varphi_i\}$ is the complete system of normalized orthogonal functions in $L^2(I)$, for $\forall X \in L^2(I)$, take $a_i =< X, \varphi_i >$ $(i = 1, 2, \ldots)$, we have

$$\|X\|_2^2 = \sum_{i=1}^{\infty} a_i^2. \tag{4}$$

Through the above lemmas, one can draw the following conclusion.

Theorem 1. *Let $\{\varphi_i\}$ be a complete system of normalized orthogonal functions in $L^2(I)$, $\forall X,\ Y \in L^2(I)$, and $a_i =< X, \varphi_i >$ $(i = 1, 2, \ldots)$, $b_i =< Y, \varphi_i >$ $(i = 1, 2, \ldots)$. Given $X^{(k)} = \sum_{i=1}^{k} a_i\varphi_i$, $Y^{(k)} = \sum_{i=1}^{k} b_i\varphi_i$, we have that*

(1) $\| X^{(k)} - Y^{(k)} \|_2^2 = \sum_{i=1}^{k} (a_i - b_i)^2$,

(2) $\| X - Y \|_2^2 = \lim_{k\to\infty} \| X^{(k)} - Y^{(k)} \|_2^2$.

Proof.

(1) It is evident that

$$\begin{aligned} \| X^{(k)} - Y^{(k)} \|_2^2 &= \| \sum_{i=1}^{k} a_i\varphi_i - \sum_{i=1}^{k} b_i\varphi_i \|_2^2 \\ &= \| \sum_{i=1}^{k} (a_i - b_i)\varphi_i \|_2^2 = \sum_{i=1}^{k} (a_i - b_i)^2. \end{aligned}$$

(2) By Lemma 3 and argument (1) of this theorem,

$$\begin{aligned} \|X - Y\|_2^2 &= \sum_{i=1}^{\infty} | < X - Y, \varphi_i > |^2 \\ &= \sum_{i=1}^{\infty} | < X, \varphi_i > - < Y, \varphi_i > |^2 \\ &= \sum_{i=1}^{\infty} (a_i - b_i)^2 = \lim_{k\to\infty} \sum_{i=1}^{k} (a_i - b_i)^2 \\ &= \lim_{k\to\infty} \| X^{(k)} - Y^{(k)} \|_2^2. \end{aligned}$$

This completes the proof. □

Remark 3. Theorem 1 shows that the distance between two elements in $L^2(I)$ can be approximated by the corresponding distance between their low-dimensional representations in a subspace of $L^2(I)$. In fact, in a complete system of normalized orthogonal functions, the distance between two elements in the low-dimensional subspace equals the Euclidean distance between their coefficient vectors.

For a classification problem, based on the observation $X_i(t_j), i = 1, 2, \ldots, N, j = 1, 2, \ldots, p$, one first finds an approximation $X_i^{(k)}(t)$ of $X_i(t)$, $i = 1, 2, \ldots, N$ in a given subspace of $L^2(I)$. Based on $X_i^{(k)}(t) = \sum_{j=1}^{k} a_{ij}\varphi_j(t)$, we know that $(a_{i1}, a_{i2}, \ldots, a_{ik})$ can be used to represent X_i, $i = 1, 2, \ldots, N$. In this case, many classification algorithms can be directly applied to the objects characterized by the new features $(a_{i1}, a_{i2}, \ldots, a_{ik})$, $i = 1, 2, \ldots, N$.

3. Several orthonormal representations

3.1. Fourier basis

It is well known that the Fourier series can provide a basis expansion. Let $X \in L^2(I)$ and T be the measure of I, then X can be represented by the following orthonormal basis,

$$\sqrt{\frac{1}{T}}, \sqrt{\frac{2}{T}}\sin\left(\frac{2\pi}{T}t\right), \sqrt{\frac{2}{T}}\cos\left(\frac{2\pi}{T}t\right), \ldots, \sqrt{\frac{2}{T}}\sin\left(\frac{2k\pi}{T}t\right), \sqrt{\frac{2}{T}}\cos\left(\frac{2k\pi}{T}t\right), \cdots$$

Moreover, the fast Fourier transform(FFT) provides a strategy to determine the coefficients extremely efficiently if p (the observation number of X) is a power of 2, and the arguments are equally spaced [46]. A point worth noting is that: a Fourier series is well suited for representing stable functions (especially, periodic functional instances), while it is inappropriate for those functions with strong local features or discontinuous features.

3.2. Functional principal component analysis (FPCA)

Functional principal component basis is also a kind of orthonormal basis. However, it differs from Fourier basis in that people cannot write out its explicit expression formula, which is often depicted by some trend characteristics. Its basic idea originated from Ramsay's work [45]. In order to clearly express the process of acquiring functional principal components, in this subsection, multivariate PCA is introduced and the eigenequation for functional PCA is derived.

3.2.1. Multivariate PCA

We first introduce and discuss the method of multivariate PCA. For p-dimensional multivariate data $\mathbf{X}_1, \mathbf{X}_2, \cdots, \mathbf{X}_N$, let

$$\mathbf{X} = \begin{pmatrix} \mathbf{X}_1' \\ \mathbf{X}_2' \\ \vdots \\ \mathbf{X}_N' \end{pmatrix} = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{N1} & x_{N2} & \cdots & x_{Np} \end{pmatrix}$$

be a standardized matrix. Multivariate PCA can be used to find a linear transformation matrix $\mathbf{A}_{k\times p}(k \le p)$ with a low-dimensional subspace, in which sample variance can be maximized on each dimension. Let ξ_j be the jth column of $\mathbf{A}'$, $f_{ij} = \xi_j'\mathbf{X}_i$ represents the score of the ith sample on the jth dimension, and ξ_j is the jth principal component vector. The overall information mean (the mean squares of the scores) of all samples on the jth dimension is represented as $\frac{1}{N}\sum_{i=1}^{N} f_{ij}^2$, $j = 1, \ldots, k$. In fact,

$$\begin{aligned} &\frac{1}{N}\sum_{i=1}^{N} f_{ij}^2 \\ &= \frac{1}{N}(f_{1j}\ f_{2j}\ \cdots\ f_{Nj})\begin{pmatrix} f_{1j} \\ f_{2j} \\ \vdots \\ f_{Nj} \end{pmatrix} \\ &= \frac{1}{N}\xi_j'\begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{N1} & x_{N2} & \cdots & x_{Np} \end{pmatrix}'\begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{N1} & x_{N2} & \cdots & x_{Np} \end{pmatrix}\xi_j \\ &= \frac{1}{N}\xi_j\mathbf{X}'\mathbf{X}\xi_j. \end{aligned} \tag{5}$$

Let $\mathbf{V} = \frac{1}{N}\mathbf{X}'\mathbf{X}$, where $\mathbf{V}$ is a sample covariance matrix. Hence, the jth principal component vector ξ_j has

$$\max \xi_j'\mathbf{V}\xi_j. \tag{6}$$

Furthermore, in order to guarantee the uniqueness of solutions, the constraint condition $\xi_j'\xi_j = 1$ needs to be considered. Based on this consideration, the above optimization problem becomes the following conditional extreme value problem:

$$F(\xi) = \xi'\mathbf{V}\xi - \lambda(\xi'\xi - 1). \tag{7}$$

Taking the derivative with respect to ξ, the following equation is obtained:

$$F'(\xi) = 2\mathbf{V}\xi - 2\lambda\xi = 0, \tag{8}$$

i.e., each principal component vector should satisfy the following eigen equation:

$$\mathbf{V}\xi = \lambda\xi. \tag{9}$$

In practice, we only select the k eigenvectors $\xi_1, \xi_2, \ldots, \xi_k$ with the top k eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_k$, which constitute an orthonormal basis of a k-dimensional subspace. In this case, $A' = (\xi_1, \xi_2, \ldots, \xi_k)$. A popular method for choosing the parameter k is the scree plot, which is a graphical method [23]. To apply it, one plots the successive eigenvalues λ_j against j. The method recommends determining the j for which the decrease of the eigenvalues appears to level off. This point is used as the selected value of k.

3.2.2. Functional PCA

Many statistical applications today involve data that do not fit into classical univariate or multivariate frameworks; for example, growth curves, spectral curves, and time-dependent gene expression profiles [19]. These functional objects can be regarded as the samples in the space of square-integrable functions $L^2(I)$, where I is a compact set. In order to improve the performance of machine learning and speed up machine learning algorithms, functional PCA can be used to extract the important discriminant features of functional data. In essence, for any $X(t) \in L^2(I)$, the aim of functional PCA is to find an optimal approximation $X^{(k)}(t)$ in a low-dimensional functional subspace of $L^2(I)$, where $X^{(k)}(t) = \sum_{i=1}^{k} a_i\xi_i(t)$. The low-dimensional functional subspace H_0 is spanned by $\{\xi_1(t), \xi_2(t), \ldots, \xi_k(t)\}$. In addition, different from Fourier basis, wavelet basis, spline basis, etc., the functional principal component basis is driven by data.

In this part, we mainly focus on the problem of acquiring functional principal components. Given some centralized functional objects $X_1(t), X_2(t), \ldots, X_N(t)$, the objective of functional PCA is to find a functional subspace H_0 so that the information of the

functional data is maximized on each eigen dimension, where $\xi_1(t), \xi_2(t), \ldots, \xi_k(t)$ are eigen functions. $f_{ij} = \int_T \xi_j(t) X_i(t) dt$ represents the score of the ith sample on the jth dimension and $\xi_j(t)$ is the jth principal component function. Similar to multivariate PCA, the overall information mean of the samples on the jth eigen dimension can be also represented as

$$\frac{1}{N}\sum_{i=1}^{N} f_{ij}^2 = \frac{1}{N}\sum_{i=1}^{N}\left[\int_T \xi_j(t) X_i(t) dt\right]^2, \ j = 1, \ldots, k. \tag{10}$$

In order to guarantee the uniqueness of the solutions, the constraint condition $\int_T \xi_j(t)^2 dt = 1$ needs to be considered. The process of mining the principal components basically becomes the process of finding $\xi's$ that maximize:

$$F(\xi) = \frac{1}{N}\sum_{i=1}^{N}\left[\int_T \xi(t) X_i(t) dt\right]^2 - \lambda\left[\int_T \xi(t)^2 dt - 1\right]. \tag{11}$$

Given $\forall \epsilon \in \mathcal{R}$, $\forall \eta \in L^2(I)$, $\mathcal{R}$ is the real number space, based on the formula (11), one can obtain the following variational equation:

$$F(\xi + \epsilon\eta) = \frac{1}{N}\sum_{i=1}^{N}\left[\int_T (\xi(t) + \epsilon\eta(t)) X_i(t) dt\right]^2 - \lambda\left[\int_T (\xi(t) + \epsilon\eta(t))^2 dt - 1\right]. \tag{12}$$

Especially, when $\epsilon = 0$, variational Eq. (12) becomes formula (11).

Let $\xi = \xi(t)$ maximize formula (11). For formula (12), taking the derivative with respect to ϵ, one has that

$$\left.\frac{dF}{d\epsilon}\right|_{\epsilon=0} = 0, \tag{13}$$

in detail,

$$\left.\frac{dF}{d\epsilon}\right|_{\epsilon=0} = \frac{1}{N}\sum_{i=1}^{N} 2\int_T \xi(t) X_i(t) dt \int_T \eta(s) X_i(s) ds - 2\lambda \int_T \xi(s)\eta(s) ds = 2\int_T \eta(s)\left[\int_T \upsilon(s,t)\xi(t) dt - \lambda\xi(s)\right] ds, \tag{14}$$

where $\upsilon(s,t) = \frac{1}{N}\sum_{i=1}^{N} X_i(s) X_i(t)$.

Combining (13), (14) with the arbitrariness of $\eta(s)$, we have the following eigen equation

$$\int_T \upsilon(s,t)\xi(t) dt = \lambda\xi(s). \tag{15}$$

Thus, $\xi's$ maximizing formula (11) are solutions of Eq. (15). The left side of (15) is an integral transform $\mathbf{V}$ of the eigen function ξ defined by

$$\mathbf{V}\xi = \int_T \upsilon(s,t)\xi(t) dt. \tag{16}$$

The integral transform is named as the covariance operator $\mathbf{V}$. Therefore, we may also express the eigen Eq. (15) directly as (9). For the computational methods of functional principal components, see [46].

In practice, we only select the corresponding k eigenfunctions $\xi_1, \xi_2, \ldots, \xi_k$ of the top k eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_k$. In this case, $\xi_1, \xi_2, \ldots, \xi_k$ constitute an orthonormal basis in a k-dimensional functional subspace. $(f_{i1}, f_{i2}, \ldots, f_{ik})$ can be used to represent $X_i(t)$, $i = 1, 2, \ldots, N$, $t \in I$.

Remark 4. To avoid the influence of different variable units, multivariate data usually need to be standardized. However, univariate functional data are not influenced by the units. For FPCA, functional data only need to be centralized. Of course, in order to better capture critical features, longitudinal transformation for functional data can be first carried out before principal component analysis. The longitudinal transformations include logarithmic transform, first order difference transform, second order difference transform, and so on.

3.3. Wavelet basis

In the literature [20], it is mentioned that one common approach of functional data analysis is to project the functional samples onto a finite dimensional subspace of $L^2(I)$ and to use the basis coefficients in a learning algorithm. It is well known that functional data with dramatically local changes appear in many fields including economics, medical science, engineering, etc. Unlike Fourier basis and splines, wavelet transform can easily capture local properties of a functional signal. In general, wavelet basis is constructed using multiresolution analysis. For any primary resolution level $j_0 \geq 0$, the collection $\{\phi_{j_0k}, k = 0, 1, \ldots, 2^{j_0} - 1; \psi_{jk}, j \geq j_0, k = 0, 1, \ldots, 2^j - 1\}$ constitutes an orthonormal basis of $L^2(I)$, and $\phi_{jk}(\text{resp.}\psi_{jk})$ is obtained by translations [13] and dilatations of a compactly supported function $\phi(\text{resp. } \psi)$, which is called as a 'father' wavelet(resp. a 'mother' wavelet), where

$$\phi_{jk}(t) = \sum_{l\in Z} 2^{j/2}\phi(2^j(t-l) - k) \tag{17}$$

and

$$\psi_{jk}(t) = \sum_{l\in Z} 2^{j/2}\psi(2^j(t-l) - k). \tag{18}$$

The idea underlying the wavelet approach is that a broad class of functions can be arbitrarily well approximated by a wavelet series [7]; i.e., for any function $X(t) \in L^2(I)$,

$$X(t) = \sum_{k=0}^{2^{j_0}-1} < X, \phi_{j_0k} > \phi_{j_0k}(t) + \sum_{j=j_0}^{\infty}\sum_{k=0}^{2^j-1} < X, \psi_{j,k} > \psi_{j,k}(t). \tag{19}$$

The coefficient $< X, \phi_{j_0k} >$ and $< X, \psi_{j,k} >$ are called the scaling and wavelet coefficients of $X(t)$, respectively.

The first term in Eq. (19) is the smooth approximation of $X(t)$ at level j_0, and the second term is the detail part of the wavelet representation. We assume that each functional curve X is observed on a fine sampling grid $t_1, \ldots, t_p$. Note that a wavelet decomposition of X can also be given in a form similar to that in (19). For $j_0 = 0$, we have

$$X(t_l) = < X, \phi_{00} > \phi_{00}(t_l) + \sum_{j=0}^{J-1}\sum_{k=0}^{2^j-1} < X, \psi_{j,k} > \psi_{j,k}(t), \tag{20}$$

where $J := log_2(N)$ is the maximal number of wavelet levels and $< X, \phi_{00} >$ and $< X, \psi_{j,k} >$ are, respectively, the scale and wavelet coefficients of the discretized curve X at position k for resolution level j. These empirical coefficients can be efficiently computed using the discrete wavelet transform described in the literature [42].

There are many types of wavelet transforms in the literature. In this paper, we shall adopt the Daubechies wavelet, which is an orthogonal basis with a compact support.

3.4. Time complexity of different representation methods

In this subsection, we analyze the computational complexity of each of the above three orthonormal representations. In fact, for the representation of functional data, it is critical to find the basis coefficients. Suppose that a functional dataset has N functional samples and each functional sample has p observation points. For Fourier representation, the FFT makes it possible to find all coefficients extremely efficiently, and its time complexity is $\mathcal{O}(Np \log p)$.

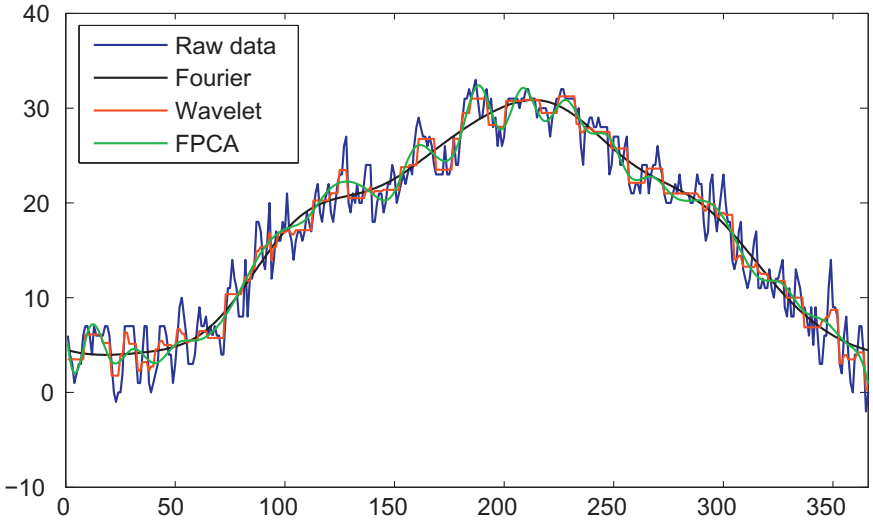

Fig. 2. Raw curve and its three fitted curves of the average temperature of Shanghai in 2012.

For wavelet representation, discrete wavelet transform(DWT) provides p coefficients closely related to the wavelet coefficients of each functional curve in $\mathcal{O}(p)$ operations [46]. As a consequence, given a functional dataset, the time complexity based on wavelet representation is $\mathcal{O}(Np)$. For a functional PCA representation, it is important to compute the covariance function in Eq. (15), therefore, its time complexity is $\mathcal{O}(Np^2)$. According to the above analysis, one can see that the time complexity of wavelet representation is the lowest, that of functional PCA is the highest, and that of Fourier representation is in the middle. Of course, in practice, the characteristic of the data itself should be first considered for the choice of basis functions.

3.5. Intuitionistic description of different orthonormal representations

In this subsection, we employ the average temperature data of Shanghai in 2012 to show its intuitionistic characteristic after orthonormal representation by using the above three orthonormal basis (see Fig. 2). In Fig. 2, the blue curve represents the raw average temperature curve, the black curve is based on Fourier representation, the red curve is based on wavelet representation, and the green curve is based on FPCA representation. From Fig. 2, one can see that wavelet basis can cope well with rapid changes in behavior, functional PCA can better fit the raw curve, and Fourier basis is smooth and stable.

4. Classification method and performance index

4.1. Classification methods

After the orthonormal representation of functional data, each functional sample becomes a point in Euclidean space. Using the Weka platform, we choose five kinds of classification methods including LibSVM, RandomForest, logistic regression, K-nearest neighbor(KNN), and artificial neuron network(ANN) to classify the functional data.

LibSVM uses the library LibSVM and calls from Weka for classification with Gaussian kernel, based on gamma=1 and tolerance=0.001.

RandomForest implements a forest of RandomTree base classifiers with 100 trees, using $\lfloor \log(\#inputs + 1) \rfloor$ inputs and unlimited depth trees.

Logistic learns a multinomial logistic regression model with a ridge estimator, using ridge in the log-likelihood $R = 10^{-8}$.

ANN is a multilayer perceptron network with sigmoid hidden neurons, learning rate 0.3, momentum 0.2, 500 training epochs, and #hidden neurons equal (#inputs and #classes)/2.

KNN is a K-nearest neighbor classifier, which tunes K by using cross-validation with linear neighbor search and Euclidean distance.

4.2. Some classification performance indexes

To test the representation ability of different basis in the classification of functional data, four classification performance measures are employed including recall(Rec), precision(Pre), $F1$ score, and accuracy (Acc).

Joachims [25] proposed the precision and the recall of a decision rule ψ in binary classification problems. Now, we generalize them to multi-class classification problems. For multi-class classification problems with K classes, given a decision rule ψ, the recall $Rec(\psi)_i$ of the ith class is defined to be the probability that an example X with label $y = i$ is classified correctly, i.e., $\psi(X) = i$:

$$\begin{aligned} Rec(\psi)_i &= P(\psi(X) = i | y = i) \\ &= \frac{P(\psi(X) = i, y = i)}{\sum_{j=1}^{K} P(\psi(X) = j, y = i)}. \end{aligned} \tag{21}$$

Similarly, the precision $Pre(\psi)_i$ of the ith class is defined to be the probability that an example X classified as $\psi(X) = i$ does indeed have the same label, i.e., $y = i$:

$$\begin{aligned} Pre(\psi)_i &= P(y = i | \psi(X) = i) \\ &= \frac{P(\psi(X) = i, y = i)}{\sum_{j=1}^{K} P(\psi(X) = i, y = j)}. \end{aligned} \tag{22}$$

The $F1$ score of the ith class is used to reconcile the precision with the recall, which is formally defined as follows:

$$F1(\psi)_i = \frac{2Pre(\psi)_i Rec(\psi)_i}{Pre(\psi)_i + Rec(\psi)_i}. \tag{23}$$

Based on the recall, the precision, and $F1$ score of the ith class, we can obtain the weighted recall (Rec), the weighted precision(Pre), and the weighted $F1$ score(F1), which are called the recall, the precision, and F1 score of the decision rule ψ, respectively.

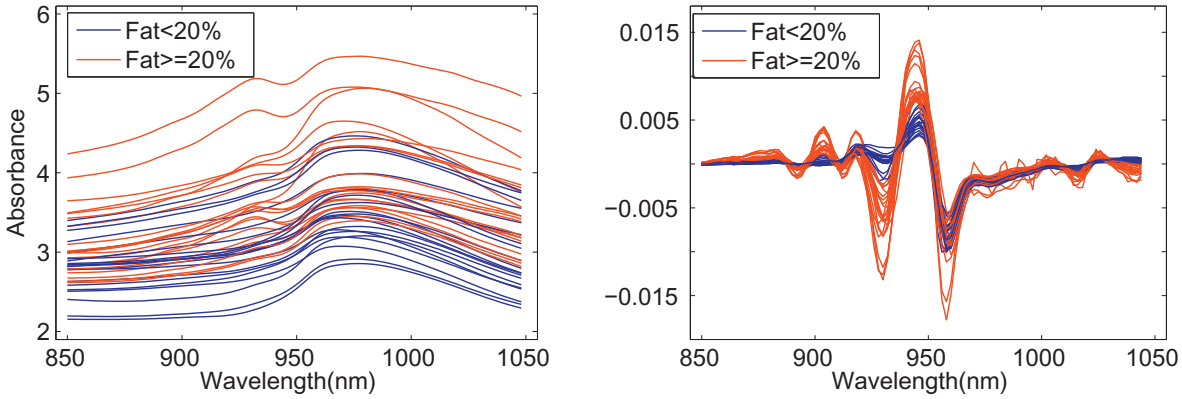

Fig. 3. Left: sample curves for Tecator data. Each class has 20 sample curves. Right: corresponding sample curves after second order difference.

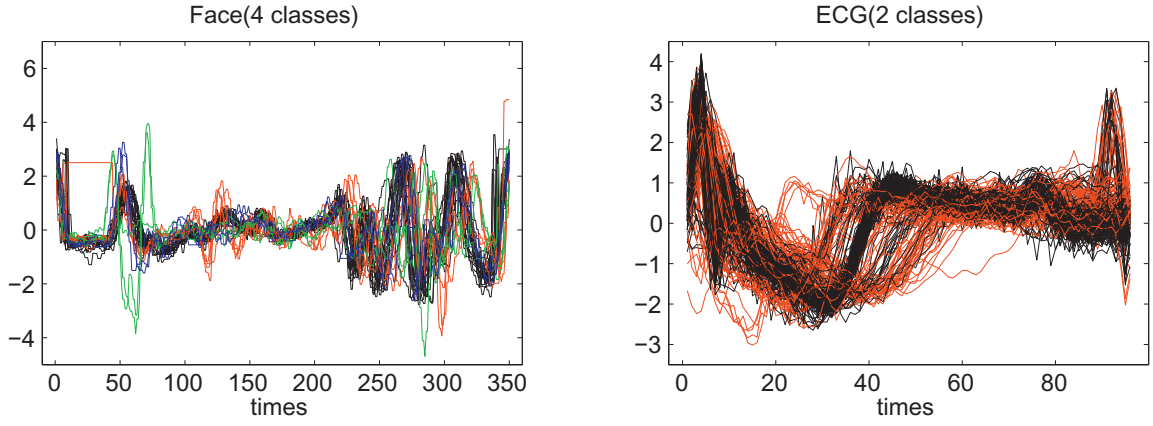

Fig. 4. Left: Face data. Right: ECG data.

Their definitions are as follows:

$$Rec(\psi) = \sum_{i=1}^{K} P(y=i) Rec(\psi)_i, \tag{24}$$

$$Pre(\psi) = \sum_{i=1}^{K} P(y=i) Pre(\psi)_i, \tag{25}$$

and

$$F1(\psi) = \sum_{i=1}^{K} P(y=i) F1(\psi)_i, \tag{26}$$

respectively.

Furthermore, the accuracy $Acc(\psi)$ of a decision rule ψ is defined as follows:

$$Acc(\psi) = \frac{1}{N} \sum_{l=1}^{N} I(\psi(X_l) = y_l), \tag{27}$$

where

$$I(\psi(X_l) = y_l) = \begin{cases} 1, & \psi(X_l) = y_l, \\ 0, & \psi(X_l) \neq y_l. \end{cases}$$

5. Experimental analysis

5.1. Data description

Tecator data is available at http://lib.stat.cmu.edu/datasets/tecator. The dataset(see Fig. 3) consists of 215 nearinfrared absorbance spectra of meat samples, recorded on a Tecator Infratec Food Analyzer. Each observation consists of a 100-channel absorbance spectrum in the wavelength range of 850–1050 nm. The goal here is to predict whether the fat percentage is greater than 20% from the spectra. Among the 215 samples, 138 have fat percentage less than 20%.

Face data, ECG data, and **ItalyPower data** are taken from the UCR Time Series Classification and Clustering website.[1] The Face dataset (see the figure on the left in Fig. 4) consists of 112 curves sampled from 4 groups at 350 instants of time. The ECG dataset(see the figure on the right in Fig. 4) consists of 200 electrocardiogram from 2 groups of patients sampled at 96 time instants. The ItalyPower dataset (see Fig. 5) consists of 1029 curves sampled from 2 groups at 24 time instants.

Sdata dataset is a simulated dataset. In this dataset, we make a simulated classification example with three known underlying generation mechanism: $g_2(t) = \cos(1.5\pi t)$, $g_2(t) = \sin(1.5\pi t)$, $g_3(t) = \sin(\pi t)$ $t \in [0, 1]$. Fig. 6 presents underlying generation curves and three sample curves of each class, where each class consists of 200 curves.

Phoneme data was formed by selecting five phonemes for classification based on digitized speech from the 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

[1] http://www.cs.ucr.edu/~eamonn/time_series_data/

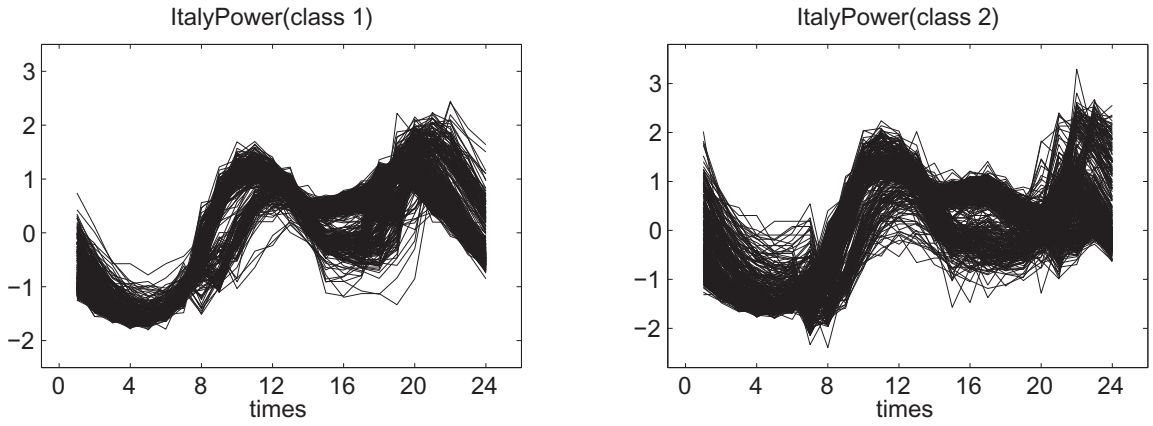

Fig. 5. ItalyPower data.

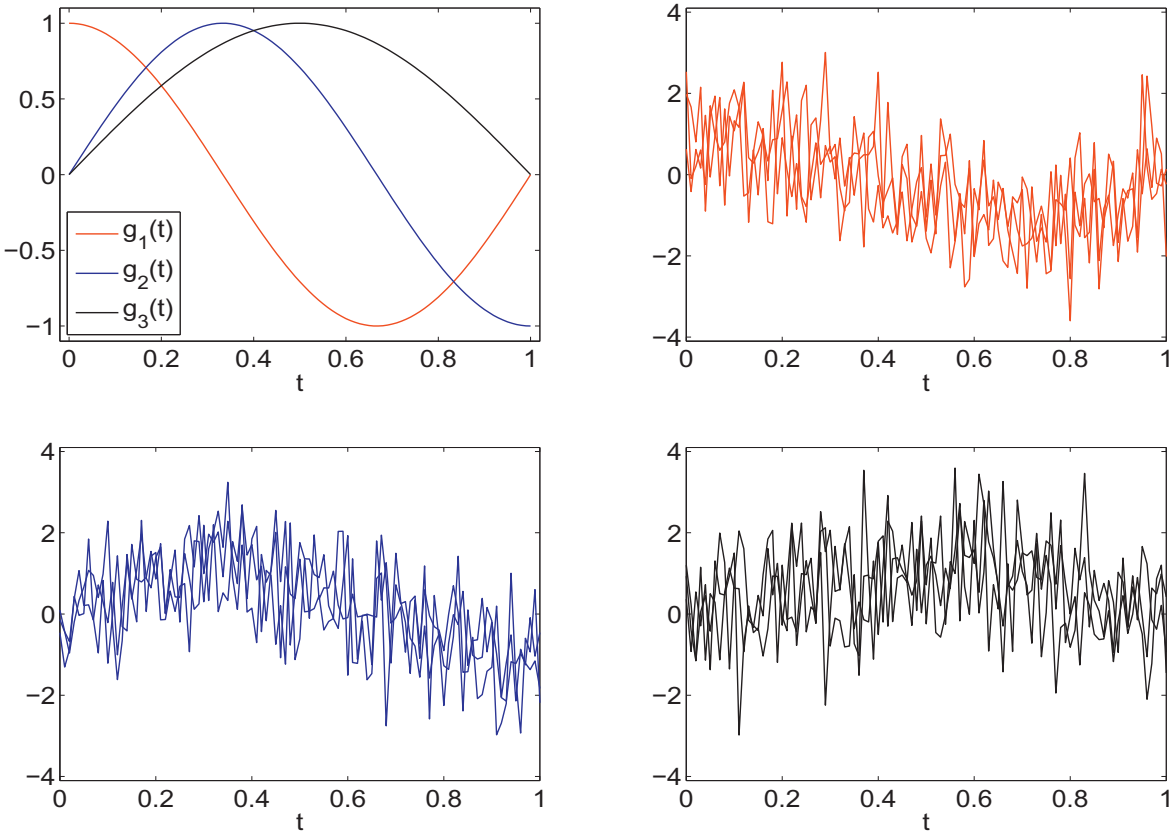

Fig. 6. Illustration of the generation mechanism (left upper) and sample curves (right upper and under) of Sdata data.

paper on penalized discriminant analysis (PDA) by Hastie et al. [22]. Fig. 7 shows five examples for each of the five classes.

From Figs. 3 and 4, Tecator dataset, Face dataset, and ECG dataset have local significant feature differences. From Figs. 5–7, ItalyPower dataset, Sdata dataset, and Phoneme dataset have global feature differences.

5.2. Classification performance comparison of different orthonormal basis

For machine learning of functional data, one needs to first find the critical features of functional data based on their low-dimensional orthonormal representations. That is, let $\varphi_1, \varphi_2, \ldots, \varphi_k$ be the orthonormal basis in certain k-dimensional subspace, then X can be approximated by $X^{(k)} = < X, \varphi_1 > \varphi_1 + < X, \varphi_2 > \varphi_2 + \cdots + < X, \varphi_k > \varphi_k$. In addition, k-dimensional coefficient vector $(< X, \varphi_1 >, , < X, \varphi_2 >, \ldots, < X, \varphi_k >)$ can be used to represent functional sample X. Secondly, a learning algorithm like SVM is performed with the 'reduced', low-dimensional data [6]. All of the following classification experiments are carried out according to the 10-fold cross-validation criterion.

In most applications, it is important to determine a value of k such that the actual data can be replaced by the approximation $\sum_{i=1}^{k} < X, \varphi_i > \varphi_i$. For example, in the case of functional principal component basis, a subjective decision for the choice of k can be made from a scree plot [28], which shows percentages of variation of functional samples. As for the Fourier basis and wavelet basis, k can be selected based on the total mean-squared error [45], which makes a trade-off between bias and sampling variance. Ramsay [45] pointed out that people find it difficult to attempt to fix model

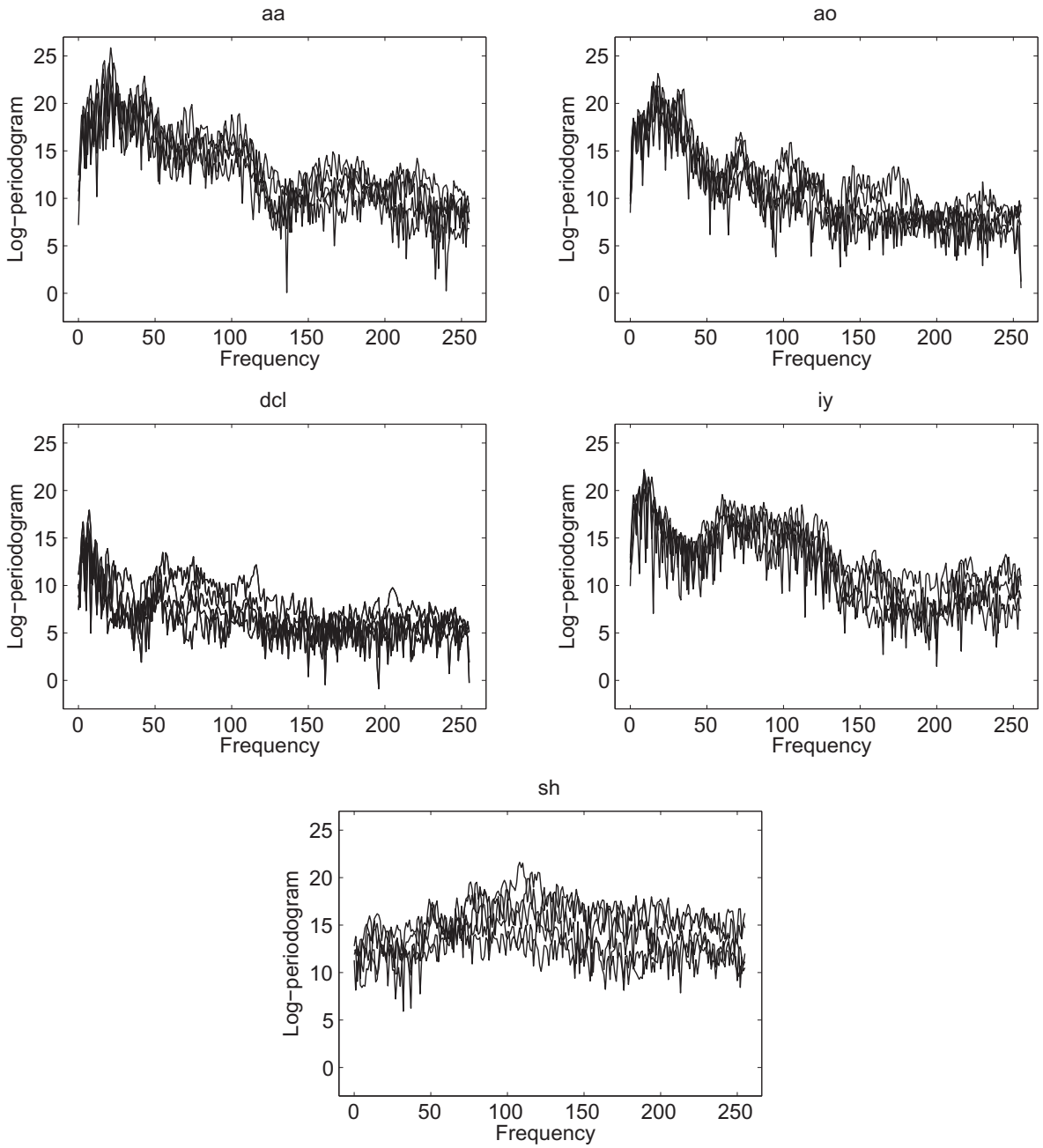

Fig. 7. Sample curves for Phoneme data. Each class has five sample curves.

dimensionality since there is no one gold standard method for the variable selection problem. In this paper, we mainly focus on the classification performance difference of three orthonormal representations with the same number of basis and the same classifier. Therefore, according to the characteristic of functional data set, different number of basis is utilized to represent the functional data set. The comparisons of their classification performance on the six data sets are shown in Tables 2–7. In every table, the value of k indicates the number of basis functions.

From Tables 2–4, it can be seen that the classification performance of wavelet representation is statistically better than that of Fourier representation. It shows that wavelet basis may be better at capturing local characteristics. This results from the fact that the wavelet expansion of a function X yields a multiresolution analysis and thus wavelets provide a systematic sequence of degrees of locality (see formula (18)). From Tables 5–7, it is evident that the classification performance of Fourier representation is statistically better than that of wavelet representation, and is more robust with respect to number of basis. It shows that Fourier basis is appropriate for representing periodic functions and stable signals. In fact, Fourier expansion is a linear combination of sine functions and cosine functions, and is generally uniformly smooth. From these tables, we argue that functional PCA may be a better representation method especially when one does not have any prior knowledge for the characteristics of functional data. For complex functional data such as Face dataset, ECG dataset, and Phoneme dataset, functional PCA exhibits much better performance than others since it is data-driven. Besides, we can also see that the classification performance is very relevant to the number of orthonormal basis. However, it should be noted that too many basis functions may cause an over-fitting problem for the classifier. Through the experiments, it can be also seen that LibSVM, RandomForest, and

Table 2
Classification performance induced by different representations on Tecator.

Basis	Classifier	Fourier				FPCA				Wavelet				Non-or			
		Acc	Pre	Rec	F1	Acc	Pre	Rec	F1	Acc	Pre	Rec	F1	Acc	Pre	Rec	F1
	LibSVM	0.967	0.968	0.967	0.967	**0.986**	**0.986**	**0.986**	**0.986**	0.991	0.991	0.991	0.991	0.828	0.826	0.828	0.825
	RandomForest	0.972	0.972	0.972	0.972	0.977	0.977	0.977	0.977	0.991	0.991	0.991	0.991	0.814	0.812	0.814	0.812
$k = 4$	Logistic	0.972	0.972	0.972	0.972	0.977	0.977	0.977	0.977	0.986	0.986	0.986	0.986	0.809	0.807	0.809	0.805
	KNN	0.981	0.981	0.981	0.981	0.967	0.967	0.967	0.967	0.991	0.991	0.991	0.991	0.772	0.776	0.772	0.774
	ANN	0.972	0.972	0.972	0.972	0.981	0.981	0.981	0.981	0.995	0.995	0.995	0.995	0.800	0.798	0.800	0.798
	LibSVM	**0.991**	**0.991**	**0.991**	**0.991**	0.977	0.977	0.977	0.977	0.972	0.972	0.972	0.972	0.828	0.826	0.828	0.825
	RandomForest	0.967	0.968	0.967	0.967	0.981	0.981	0.981	0.981	0.977	0.977	0.977	0.977	0.828	0.826	0.828	0.825
$k = 7$	Logistic	0.986	0.986	0.986	0.986	0.981	0.981	0.981	0.981	0.986	0.987	0.986	0.986	0.781	0.777	0.781	0.777
	KNN	0.986	0.986	0.986	0.986	0.977	0.978	0.977	0.977	0.972	0.972	0.972	0.972	0.828	0.826	0.828	0.827
	ANN	0.986	0.986	0.986	0.986	0.986	0.986	0.986	0.986	0.986	0.986	0.986	0.986	0.805	0.804	0.805	0.804
	LibSVM	0.991	0.991	0.991	0.991	0.972	0.972	0.972	0.972	**1.000**	**1.000**	**1.000**	**1.000**	0.953	0.953	0.953	0.953
	RandomForest	0.977	0.977	0.977	0.977	0.981	0.981	0.981	0.981	0.977	0.977	0.977	0.977	0.963	0.963	0.963	0.963
$k = 13$	Logistic	0.972	0.972	0.972	0.972	0.963	0.964	0.963	0.963	0.991	0.991	0.991	0.991	0.944	0.944	0.944	0.944
	KNN	0.981	0.981	0.981	0.981	0.953	0.954	0.953	0.953	0.967	0.967	0.967	0.967	0.972	0.972	0.972	0.972
	ANN	0.981	0.981	0.981	0.981	0.972	0.972	0.972	0.972	0.972	0.972	0.972	0.972	0.940	0.940	0.940	0.940
	LibSVM	0.977	0.977	0.977	0.977	0.963	0.963	0.963	0.963	0.972	0.972	0.972	0.972	**0.977**	**0.977**	**0.977**	**0.977**
	RandomForest	0.972	0.972	0.972	0.972	0.972	0.972	0.972	0.972	0.972	0.972	0.972	0.972	0.967	0.968	0.967	0.967
$k = 25$	Logistic	0.953	0.953	0.953	0.953	0.949	0.950	0.949	0.949	0.981	0.981	0.981	0.981	0.953	0.953	0.953	0.953
	KNN	0.986	0.986	0.986	0.986	0.935	0.941	0.935	0.933	0.972	0.972	0.972	0.972	0.972	0.972	0.972	0.972
	ANN	0.967	0.968	0.967	0.967	0.962	0.963	0.963	0.963	0.986	0.986	0.986	0.986	0.958	0.958	0.958	0.958

Table 3
Classification performance induced by different representations on face.

Basis	Classifier	Fourier				FPCA				Wavelet				Non-or			
		Acc	Pre	Rec	F1	Acc	Pre	Rec	F1	Acc	Pre	Rec	F1	Acc	Pre	Rec	F1
	LibSVM	0.696	0.715	0.696	0.692	0.848	0.853	0.848	0.847	0.723	0.729	0.723	0.715	0.241	0.150	0.241	0.185
	RandomForest	0.750	0.754	0.750	0.744	0.821	0.826	0.821	0.823	0.732	0.732	0.732	0.729	0.313	0.305	0.313	0.304
$k = 6$	Logistic	0.714	0.722	0.714	0.714	0.839	0.838	0.839	0.838	0.670	0.664	0.670	0.666	0.304	0.259	0.304	0.279
	KNN	0.705	0.704	0.705	0.703	0.830	0.837	0.830	0.832	0.670	0.672	0.670	0.669	0.268	0.256	0.268	0.260
	ANN	0.732	0.731	0.732	0.728	0.804	0.802	0.804	0.803	0.696	0.699	0.696	0.695	0.205	0.211	0.205	0.207
	LibSVM	0.768	0.770	0.768	0.761	0.929	0.933	0.929	0.930	0.857	0.865	0.857	0.859	0.304	0.232	0.304	0.247
	RandomForest	0.813	0.820	0.813	0.814	0.893	0.896	0.893	0.894	0.839	0.842	0.839	0.837	0.313	0.314	0.313	0.308
$k = 11$	Logistic	0.813	0.817	0.813	0.814	0.884	0.886	0.884	0.885	0.911	0.916	0.911	0.912	0.295	0.289	0.295	0.290
	KNN	0.786	0.796	0.786	0.789	0.920	0.923	0.920	0.920	0.804	0.807	0.804	0.801	0.279	0.283	0.277	0.275
	ANN	0.839	0.842	0.839	0.840	0.911	0.912	0.911	0.911	0.920	0.924	0.920	0.920	0.277	0.278	0.277	0.277
	LibSVM	0.875	0.878	0.875	0.875	0.920	0.927	0.920	0.920	0.938	0.940	0.938	0.937	0.402	0.396	0.402	0.372
	RandomForest	0.866	0.871	0.866	0.866	0.902	0.902	0.902	0.902	0.911	0.916	0.911	0.912	0.509	0.505	0.509	0.505
$k = 22$	Logistic	0.920	0.925	0.920	0.921	0.938	0.939	0.938	0.937	0.929	0.934	0.929	0.927	0.375	0.382	0.375	0.378
	KNN	0.866	0.869	0.866	0.864	0.920	0.920	0.920	0.919	0.902	0.905	0.902	0.900	0.384	0.396	0.384	0.386
	ANN	**0.946**	**0.955**	**0.946**	**0.948**	**0.964**	**0.967**	**0.964**	**0.965**	0.955	0.958	0.955	0.955	0.393	0.403	0.393	0.397
	LibSVM	0.902	0.903	0.902	0.900	0.884	0.904	0.884	0.886	0.884	0.890	0.884	0.885	0.598	0.610	0.598	0.594
	RandomForest	0.884	0.888	0.884	0.885	0.911	0.916	0.911	0.911	0.938	0.941	0.938	0.938	**0.652**	**0.664**	**0.652**	**0.656**
$k = 44$	Logistic	0.848	0.860	0.848	0.850	0.893	0.897	0.893	0.892	0.946	0.947	0.946	0.946	0.491	0.496	0.491	0.491
	KNN	0.830	0.843	0.830	0.830	0.884	0.885	0.884	0.883	0.902	0.903	0.902	0.902	0.491	0.539	0.491	0.500
	ANN	0.911	0.917	0.911	0.911	0.929	0.933	0.929	0.928	**0.964**	**0.966**	**0.964**	**0.964**	0.625	0.625	0.625	0.622

Table 4
Classification performance induced by different representations on ECG.

Basis	Classifier	Fourier				FPCA				Wavelet				Non-or			
		Acc	Pre	Rec	F1	Acc	Pre	Rec	F1	Acc	Pre	Rec	F1	Acc	Pre	Rec	F1
	LibSVM	0.765	0.760	0.765	0.762	0.760	0.757	0.760	0.758	0.750	0.744	0.750	0.729	0.790	0.787	0.790	0.788
	RandomForest	0.775	0.771	0.775	0.773	0.790	0.789	0.790	0.789	0.660	0.660	0.660	0.660	0.770	0.763	0.770	0.762
$k = 3$	Logistic	0.790	0.786	0.790	0.786	0.785	0.784	0.785	0.785	0.740	0.730	0.740	0.724	0.770	0.772	0.770	0.771
	KNN	0.775	0.775	0.775	0.757	0.795	0.793	0.795	0.794	0.740	0.733	0.740	0.716	0.800	0.796	0.800	0.793
	ANN	0.755	0.750	0.755	0.751	0.790	0.800	0.790	0.793	0.740	0.732	0.740	0.718	0.805	0.801	0.805	0.798
	LibSVM	0.775	0.771	0.775	0.773	0.855	0.853	0.855	0.853	0.780	0.775	0.780	0.770	0.790	0.789	0.790	0.789
	RandomForest	0.790	0.786	0.790	0.786	0.830	0.829	0.830	0.829	0.815	0.812	0.815	0.813	0.795	0.790	0.795	0.789
$k = 6$	Logistic	0.765	0.760	0.765	0.762	0.835	0.833	0.835	0.833	0.765	0.758	0.765	0.755	0.775	0.778	0.775	0.776
	KNN	0.775	0.775	0.775	0.757	**0.905**	**0.905**	**0.905**	**0.905**	0.850	0.849	0.850	0.849	0.795	0.790	0.795	0.789
	ANN	0.755	0.750	0.755	0.751	0.865	0.863	0.865	0.864	0.815	0.812	0.815	0.813	0.785	0.783	0.785	0.784
	LibSVM	0.850	0.848	0.850	0.848	0.885	0.884	0.885	0.884	0.865	0.864	0.865	0.862	0.805	0.802	0.805	0.803
	RandomForest	0.860	0.859	0.860	0.859	0.845	0.843	0.845	0.843	0.875	0.876	0.875	0.875	0.815	0.816	0.815	0.815
$k = 12$	Logistic	0.795	0.791	0.795	0.792	0.800	0.797	0.800	0.798	0.805	0.804	0.805	0.805	0.790	0.787	0.790	0.788
	KNN	0.875	0.874	0.875	0.874	0.895	0.894	0.895	0.893	0.875	0.874	0.875	0.874	0.805	0.801	0.805	0.800
	ANN	0.825	0.826	0.825	0.825	0.860	0.859	0.860	0.859	0.870	0.871	0.870	0.870	0.770	0.770	0.770	0.770
	LibSVM	0.850	0.849	0.850	0.849	0.890	0.889	0.890	0.889	0.855	0.853	0.855	0.853	0.815	0.812	0.815	0.813
	RandomForest	0.830	0.838	0.830	0.818	0.870	0.869	0.870	0.868	0.845	0.843	0.845	0.843	0.795	0.796	0.795	0.795
$k = 24$	Logistic	0.835	0.833	0.835	0.834	0.850	0.849	0.850	0.849	0.810	0.810	0.810	0.810	0.835	0.833	0.835	0.835
	KNN	**0.880**	**0.881**	**0.880**	**0.880**	0.875	0.874	0.875	0.873	**0.885**	**0.885**	**0.885**	**0.885**	**0.845**	**0.843**	**0.845**	**0.843**
	ANN	0.850	0.849	0.850	0.849	0.865	0.865	0.865	0.865	0.860	0.860	0.860	0.860	0.815	0.811	0.815	0.811

Table 5
Classification performance induced by different representations on ItalyPower.

Basis	Classifier	Fourier				FPCA				Wavelet				Non-or			
		Acc	Pre	Rec	F1	Acc	Pre	Rec	F1	Acc	Pre	Rec	F1	Acc	Pre	Rec	F1
	LibSVM	0.961	0.961	0.961	0.961	0.967	0.967	0.967	0.967	0.824	0.829	0.824	0.823	0.716	0.726	0.716	0.713
	RandomForest	0.950	0.951	0.950	0.950	0.964	0.964	0.964	0.964	0.764	0.764	0.764	0.764	0.721	0.724	0.721	0.720
$k = 3$	Logistic	0.958	0.958	0.958	0.958	0.968	0.968	0.968	0.968	0.824	0.826	0.824	0.824	0.704	0.705	0.704	0.703
	KNN	0.960	0.960	0.960	0.960	0.962	0.962	0.962	0.962	0.836	0.839	0.836	0.835	0.725	0.753	0.725	0.717
	ANN	0.960	0.960	0.960	0.960	0.964	0.964	0.964	0.964	0.825	0.839	0.825	0.823	0.713	0.734	0.713	0.707
	LibSVM	0.961	0.961	0.961	0.961	0.971	0.971	0.971	0.971	0.958	0.958	0.958	0.958	0.948	0.950	0.948	0.948
	RandomForest	0.957	0.957	0.957	0.957	0.965	0.965	0.965	0.965	0.954	0.954	0.954	0.954	0.942	0.943	0.942	0.942
$k = 6$	Logistic	0.951	0.952	0.951	0.951	0.969	0.969	0.969	0.969	0.949	0.949	0.949	0.949	0.946	0.947	0.946	0.946
	KNN	0.955	0.955	0.955	0.955	0.954	0.956	0.954	0.954	0.944	0.944	0.944	0.944	0.939	0.942	0.939	0.939
	ANN	0.964	0.964	0.964	0.964	0.966	0.966	0.966	0.966	0.954	0.954	0.954	0.954	0.942	0.942	0.942	0.942
	LibSVM	0.966	0.966	0.966	0.966	**0.975**	**0.975**	**0.975**	**0.975**	0.968	0.968	0.968	0.968	0.953	0.954	0.953	0.953
	RandomForest	0.960	0.960	0.960	0.960	0.968	0.968	0.968	0.968	0.966	0.966	0.966	0.966	0.945	0.945	0.945	0.945
$k = 12$	Logistic	0.960	0.960	0.960	0.960	0.967	0.967	0.967	0.967	0.970	0.970	0.970	0.970	0.942	0.942	0.942	0.942
	KNN	0.954	0.954	0.954	0.954	0.971	0.971	0.971	0.971	0.961	0.961	0.961	0.961	0.938	0.946	0.938	0.938
	ANN	0.961	0.961	0.961	0.961	0.960	0.960	0.960	0.960	0.968	0.969	0.969	0.969	0.938	0.938	0.938	0.938
	LibSVM	**0.974**	**0.974**	**0.974**	**0.974**	0.974	0.974	0.974	0.974	**0.972**	**0.972**	**0.972**	**0.972**	**0.955**	**0.956**	**0.955**	**0.955**
	RandomForest	0.972	0.972	0.972	0.972	0.966	0.966	0.966	0.966	0.972	0.972	0.972	0.972	0.946	0.946	0.946	0.946
$k = 24$	Logistic	0.972	0.972	0.972	0.972	0.969	0.969	0.969	0.969	0.973	0.973	0.973	0.973	0.942	0.942	0.942	0.942
	KNN	0.972	0.972	0.972	0.972	0.964	0.964	0.964	0.964	0.969	0.969	0.969	0.969	0.937	0.940	0.937	0.937
	ANN	0.961	0.961	0.961	0.961	0.969	0.969	0.969	0.969	0.967	0.967	0.967	0.967	0.944	0.944	0.944	0.944

Table 6
Classification performance induced by different representations on Sdata.

Basis	Classifier	Fourier				FPCA				Wavelet				Non-or			
		Acc	Pre	Rec	F1	Acc	Pre	Rec	F1	Acc	Pre	Rec	F1	Acc	Pre	Rec	F1
	LibSVM	0.995	0.995	0.995	0.995	**0.997**	**0.997**	**0.997**	**0.997**	0.975	0.975	0.975	0.975	0.978	0.978	0.978	0.978
	RandomForest	0.993	0.993	0.993	0.993	0.988	0.988	0.988	0.988	0.968	0.968	0.968	0.968	0.980	0.980	0.980	0.980
$k = 4$	Logistic	0.987	0.987	0.987	0.987	0.987	0.987	0.987	0.987	0.970	0.970	0.970	0.970	0.977	0.977	0.977	0.977
	KNN	0.993	0.993	0.993	0.993	0.993	0.993	0.993	0.993	0.963	0.963	0.963	0.963	0.972	0.972	0.972	0.972
	ANN	0.992	0.992	0.992	0.992	0.992	0.992	0.992	0.992	0.977	0.977	0.977	0.977	0.975	0.975	0.975	0.975
	LibSVM	0.993	0.993	0.993	0.993	0.995	0.995	0.995	0.995	0.990	0.990	0.990	0.990	0.975	0.975	0.975	0.975
	RandomForest	0.993	0.993	0.993	0.993	0.992	0.992	0.992	0.992	0.982	0.982	0.982	0.982	0.977	0.977	0.977	0.977
$k = 7$	Logistic	0.988	0.988	0.988	0.988	0.985	0.985	0.985	0.985	0.982	0.982	0.982	0.982	0.975	0.975	0.975	0.975
	KNN	0.993	0.993	0.993	0.993	0.990	0.990	0.990	0.990	0.965	0.965	0.965	0.965	0.942	0.942	0.942	0.942
	ANN	0.990	0.990	0.990	0.990	0.988	0.988	0.988	0.988	0.977	0.977	0.977	0.977	0.967	0.967	0.967	0.967
	LibSVM	0.992	0.992	0.992	0.992	0.995	0.995	0.995	0.995	0.990	0.990	0.990	0.990	0.968	0.968	0.968	0.968
	RandomForest	0.993	0.993	0.993	0.993	0.990	0.990	0.990	0.990	0.987	0.987	0.987	0.987	**0.980**	**0.980**	**0.980**	**0.980**
$k = 13$	Logistic	0.995	0.995	0.995	0.995	0.987	0.987	0.987	0.987	0.977	0.977	0.977	0.977	0.968	0.968	0.968	0.968
	KNN	0.983	0.983	0.983	0.983	0.963	0.966	0.963	0.963	0.973	0.973	0.973	0.973	0.910	0.911	0.910	0.910
	ANN	0.993	0.993	0.993	0.993	0.995	0.995	0.995	0.995	0.987	0.987	0.987	0.987	0.980	0.980	0.980	0.980
	LibSVM	**0.997**	**0.997**	**0.997**	**0.997**	0.997	0.997	0.997	0.997	**0.995**	**0.995**	**0.995**	**0.995**	0.963	0.963	0.963	0.963
	RandomForest	0.993	0.993	0.993	0.993	0.987	0.987	0.987	0.987	0.982	0.982	0.982	0.982	0.977	0.977	0.977	0.977
$k = 26$	Logistic	0.980	0.980	0.980	0.980	0.985	0.985	0.985	0.985	0.982	0.982	0.982	0.982	0.938	0.940	0.938	0.939
	KNN	0.950	0.950	0.950	0.950	0.907	0.907	0.907	0.907	0.967	0.970	0.967	0.967	0.893	0.895	0.893	0.893
	ANN	0.988	0.988	0.988	0.988	0.992	0.992	0.992	0.992	0.980	0.980	0.980	0.980	0.953	0.953	0.953	0.953

Table 7
Classification performance induced by different representations on Phoneme.

Basis	Classifier	Fourier				FPCA				Wavelet				Non-or			
		Acc	Pre	Rec	F1	Acc	Pre	Rec	F1	Acc	Pre	Rec	F1	Acc	Pre	Rec	F1
	LibSVM	0.846	0.846	0.846	0.844	0.836	0.833	0.836	0.832	0.690	0.676	0.690	0.653	0.577	0.548	0.577	0.538
	RandomForest	0.833	0.833	0.833	0.832	0.818	0.816	0.818	0.817	0.653	0.649	0.653	0.650	0.543	0.531	0.543	0.535
$k = 4$	Logistic	0.843	0.843	0.843	0.842	0.836	0.834	0.836	0.834	0.682	0.658	0.682	0.637	0.576	0.532	0.576	0.538
	KNN	0.840	0.843	0.840	0.838	0.831	0.829	0.831	0.826	0.690	0.685	0.690	0.674	0.575	0.561	0.575	0.560
	ANN	0.840	0.843	0.840	0.838	0.836	0.835	0.836	0.835	0.682	0.671	0.682	0.656	0.569	0.541	0.569	0.535
	LibSVM	0.891	0.890	0.891	0.889	0.896	0.895	0.896	0.895	0.815	0.815	0.815	0.813	0.584	0.568	0.583	0.552
	RandomForest	0.879	0.878	0.879	0.877	0.882	0.881	0.882	0.881	0.813	0.814	0.813	0.812	0.570	0.553	0.570	0.556
$k = 8$	Logistic	0.890	0.890	0.890	0.889	0.893	0.892	0.893	0.892	0.801	0.801	0.801	0.800	0.580	0.562	0.580	0.549
	KNN	0.839	0.839	0.839	0.839	0.888	0.889	0.888	0.886	0.802	0.811	0.802	0.800	0.570	0.565	0.570	0.539
	ANN	0.886	0.886	0.886	0.886	0.889	0.889	0.889	0.888	0.812	0.815	0.812	0.811	0.574	0.558	0.574	0.549
	LibSVM	0.912	0.912	0.912	0.911	0.913	0.912	0.913	0.912	0.907	0.907	0.907	0.907	0.633	0.628	0.633	0.628
	RandomForest	0.895	0.895	0.895	0.894	0.896	0.895	0.896	0.894	0.901	0.901	0.901	0.900	0.630	0.622	0.630	0.622
$k = 16$	Logistic	0.907	0.906	0.907	0.906	0.908	0.907	0.908	0.907	0.902	0.901	0.902	0.901	0.625	0.618	0.625	0.619
	KNN	0.900	0.900	0.900	0.900	0.892	0.897	0.892	0.886	0.887	0.890	0.887	0.884	0.618	0.623	0.618	0.617
	ANN	0.890	0.889	0.890	0.889	0.896	0.895	0.896	0.895	0.890	0.889	0.890	0.890	0.614	0.613	0.614	0.611
	LibSVM	**0.914**	**0.914**	**0.914**	**0.913**	**0.918**	**0.917**	**0.918**	**0.917**	**0.914**	**0.913**	**0.914**	**0.913**	**0.703**	**0.699**	**0.703**	**0.700**
	RandomForest	0.899	0.899	0.899	0.897	0.893	0.893	0.893	0.892	0.912	0.912	0.912	0.911	0.696	0.691	0.696	0.685
$k = 32$	Logistic	0.903	0.903	0.903	0.903	0.909	0.909	0.909	0.909	0.902	0.902	0.902	0.902	0.694	0.689	0.694	0.690
	KNN	0.903	0.907	0.903	0.901	0.872	0.886	0.872	0.862	0.885	0.889	0.885	0.881	0.691	0.701	0.691	0.687
	ANN	0.891	0.890	0.891	0.890	0.891	0.891	0.891	0.891	0.898	0.897	0.898	0.897	0.651	0.650	0.651	0.649

artificial neuron network(ANN) show much better classification performance.

5.3. Classification performance comparison between orthogonal basis and non-orthogonal basis

In this subsection, we compare the classification performance of classifiers induced by orthogonal representation and non-orthogonal representation. Now, Fourier basis $\{1, \sin(t), \cos(t), \ldots, \sin(kt), \cos(kt), \ldots\}$ is used to represent functional samples. If the measure T of I is not 2π, then the above basis is not orthogonal basis. The classification performance for non-orthogonal representation (just Non-or) can be seen from Tables 2–7.

First, we can see that the classification performance of classifiers induced by orthogonal representation is statistically better than that induced by non-orthogonal representation on every data set. Second, since Fourier basis is appropriate to represent functional data with periodic characteristic, the difference between two kinds of representations is not too large for the Sdata data set. However, the classification performance of classifiers induced by orthogonal representation is still better than that of non-orthogonal representation. Third, for functional data with local difference characteristics including Tecator dataset and Face dataset, there are obvious difference between the classifiers induced by orthogonal representation and those induced by non-orthogonal representation for classification performance.

6. Conclusions

The main motivations of this study were to answer three important problems: (1)Why can a functional sample be seen as a point in the corresponding Euclidean space after certain orthonormal representation? (2)How to select orthonormal basis for a given functional data type? (3) For orthogonal representation and non-orthogonal representation, which one is better under finite basis functions with the same number of basis?

For the first problem, in this paper, we have given a theorem for illustrating the reasonability of representing a functional sample as a point in the corresponding Euclidean space, which is isomorphic to the low-dimensional representation space for the functional sample. In this case, the distance between two functional samples becomes the Euclidean distance between two points in the classification process, and thus, based on this representation, some machine learning algorithms can be utilized to classify a functional data set.

For the second problem, we have performed a series of comparison analysis in-between Fourier basis, functional principal component basis and wavelet basis. Experiment results show that the selection of orthonormal basis may depend on the characteristics of functional data themselves, which is helpful for obtaining a classifier of functional data with better generalization ability. Fourier basis may be suitable for stable functional data (especially periodic data), wavelet basis may be appropriate for functional data with local characteristic, and data driven functional principal component basis could be the preferred choice when prior information about the characteristics of functional data is not known. In particular, the eigenequation of FPCA is obtained by means of variational theory.

For the third problem, experimental results have also shown that orthogonal representation may be much better than non-orthogonal representation from the viewpoint of classification performance, because orthogonal representation may include more information under finite basis functions with same number of basis.

To summarize, the research results would be very helpful for guiding researchers to reasonably use orthonormal representation methods for machine learning of functional data.

Acknowledgment

This work was supported by the National Natural Science Fund of China (Nos. 61432011, U1435212), National Key Basic Research and Development Program of China (973) (Nos. 2013CB329404).

References

[1] C. Abraham, P.A. Cornillon, E. Matzner-Lober, Unsupervised curve clustering using B-splines, Scand. J. Stat. 30 (2003) 581–595.
[2] R.A. Adams, J.J.F. Fournier, Sobolev Spaces, 2nd ed., Elsevier(Singapore) Pte Ltd, 2009.
[3] T. Ando, Penalized optimal scoring for the classification of multi-dimensional functional data, Stat. Methodol. 6 (2009) 565–576.
[4] T. Ando, S. Imoto, S. Konishi, Nonlinear regression modeling via regularized radial basis function networks, J. Stat. Plan. Inference 138 (2008) 3616–3633.
[5] Y. Araki, S. Konishi, S. Imoto, Functional discriminant analysis for time-series gene expression data via radial basis function expansion, in: Proceeding of the COMPSTAT, 2004, pp. 613–620.
[6] J.R. Berrendero, A. Cuevas, J.L. Torrecilla, The mRMR variable selection method: a comparative study for functional data, J. Stat. Comput. Simul. (2015), doi:10.1080/00949655.2015.1042378.
[7] P. Besbeas, I.D. Feis, T. Sapatinas, A comparative simulation study of wavelet shrinkage estimators for poisson counts, Int. Stat. Rev. 72 (2) (2004) 209–237.
[8] K. Beyer, J. Goldstein, R. Ramakrishnan, U. Shaft, When is "nearest neighbor" meaningful? in: Proceedings of the 7th International Conference on Database Theory-ICDT'99, 1999, pp. 217–235.
[9] L. Breiman, Random forests, Mach. Learn. 45 (2001) 5–32.
[10] P.E. Castro, W.H. Lawton, E.A. Sylvestre, Principal modes of variation for processes with continuous sample curves, Technometrics 28 (1986) 329–337.
[11] K.P. Chan, A.C. Fu, Efficient time series matching by wavelets, in: Proceedings of the 15th IEEE International Conference on Data Engineering, 1999, pp. 126–133.
[12] C.C. Chang, C.J. Lin, LIBSVM: A library for support vector machines, ACM Trans. Intell. Syst. Technol. 2 (3) (2011) 1–27.
[13] R.E. Edwards, Fourier Series: A Modern Introduction, vol. 1, 2nd ed., Springer-Verlag, 2003.
[14] J. Elder IV, D. Pregibon, A statistical perspective on knowledge discovery in databases, in: U.M. Fayyad, G. Shapiro, P. Smyth, R. Uthurusamy (Eds.), Advances in Knowledge Discovery and Data Mining, AAAI Press, Menlo Park, CA, 1996, pp. 83–113.
[15] C. Faloutsos, M. Ranganathan, Y. Manolopulos, Fast subsequence matching in time-series databases, SIGMOD Rec. 23 (1994) 519–529.
[16] R. Favero, R. King, Wavelet parameterization for speech recognition: variations in translation and scale parameters[c]//speech, image processing and neural networks, 1994, in: Proceedings of the International Symposium on ISSIPNN' 94., IEEE, 1994, pp. 694–697.
[17] M. Fernández-Delgado, E. Cernadas, S. Barro, Do we need hundreds of classifiers to solve real world classification problems? J. Mach. Learn. Res. 15 (2014) 3133–3181.
[18] F. Ferraty, P. Vieu, Curves discrimination: a nonparametric functional approach, Comput. Stat. Data Anal. 44 (2003) 161–173.
[19] D. Gervini, Outlier detection and trimmed estimation for general functional data, Stat. Sin. 22 (4) (2012) 1639–1660.
[20] B. Gregorutti, B. Michel, P. Saint-Pierre, Grouped variable importance with random forests and application to multiple functional data analysis, Comput. Stat. Data Anal. 90 (2015) 15–35.
[21] P. Hall, H.G. Müller, J.L. Wang, Properties of principal component methods for functional and longitudinal data analysis, Ann. Stat. 34 (2006) 1493–1517.
[22] T. Hastie, A. Buja, R. Tibshirani, Penalized discriminant analysis, Ann. Stat. 23 (1995) 73–102.
[23] L. Horváth, P. Kokoszka, Inference for Functional Data with Applications, Springer, New York, 2012.
[24] G.M. James, Generalized linear models with functional predictors, J. R. Stat. Soc. Ser. B 64 (2002) 411–432.
[25] T. Joachims, Estimating the generalization performance of a SVM efficiently, in: P. Langley (Ed.), Proceedings of ICML-00, 17th International Conference on Machine Learning, Morgan Kaufmann Publishers, San Francisco, CA, 2000, pp. 431–438.
[26] S. Konishi, T. Ando, S. Imoto, Bayesian information criteria and smoothing parameter selection in radial basis function networks, Biometrika 91 (2004) 27–43.
[27] M. Krawczak, G. Szkatula, An approach to dimensionality reduction in time series, Inf. Sci. 260 (2014) 15–36.
[28] H.J. Lee, Functional Data Analysis: Classification and Regression [d], Texas A&M University, 2004.
[29] X.Y. Leng, H.G. Müller, Classification using functional data analysis for temporal gene expression data, Bioinformatics 22 (2006) 68–76.

[30] B. Li, Y.W. Chen, Y.Q. Chen, The nearest neighbor algorithm of local probability centers, IEEE Trans. Systems, Man, Cybern. B 38 (1) (2008) 141–154.
[31] Z.G. Liu, Q. Pan, J. Dezert, A new belief-based k-nearest neighbor classification method, Pattern Recognit. 46 (3) (2013) 834–844.
[32] Y.H. Liu, U. Aickelin, J. Feyereisi, L.G. Durrant, Wavelet feature extraction and genetic algorithm for biomarker detection in colorectal cancer data, Knowl.-Based Syst. 37 (2013) 502–514.
[33] M. López, J. Martínez, J.M. Matías, J. Taboadaa, J.A. Vilán, Functional classification of ornamental stone using machine learning techniques, J. Comput. Appl. Math. 234 (2010) 1338–1345.
[34] J. Lu, V. Behbood, P. Hao, H. Zuo, S. Xue, G. Zhang, Transfer learning using computational intelligence: A survey, Knowl.-Based Syst. 80 (2015) 14–23.
[35] Y.H. Luan, H.Z. Li, Clustering of temporal gene expression data using a mixed-effects model with b-splines, Bioinformatics 19 (2003) 474–482.
[36] Y.F. Meng, J.Y. Liang, Regression analysis for functional data based on least squares support vector machine, Pattern Recognit. Artif. Intell. 27 (12) (2014) 1124–1130.(in Chinese)
[37] J.S. Morris, R.J. Carroll, Wavelet-based functional mixed models, J. R. Stat. Soc., Ser. B 68 (2006) 179–199.
[38] A. Munõz, J. González, Representing functional data using support vector machines, Pattern Recognit. Lett. 31 (2010) 511–516.
[39] H.G. Müller, Functional modelling and classification of longitudinal data, Scand. J. Stat. 32 (2005) 223–240.
[40] H.G. Müller, U. StadtMüller, Generalized functional linear models, Ann. Stat. 33 (2005) 774–805.
[41] D.W. Patterson, Artificial Neural Networks: Theory and Applications, Prentice Hall PTR Upper Saddle River, NJ, USA, 1998.
[42] D.B. Percival, A.T. Walden, Wavelet Methods for Time Series Analysis, Cambridge University Press, 2000.
[43] C. Preda, G. Saporta, C. Lévéder, PLS classification of functional data, Comput. Stat. 22 (2007) 223–235.
[44] N. Quadrianto, Z. Ghahramani, A very simple safe-bayesian random forest, IEEE Trans. Pattern Anal. Mach. Intell. 37 (6) (2015) 1297–1303.
[45] J.O. Ramsay, When the data are functions, Psychometrika 47 (4) (1982) 379–396.
[46] J.O. Ramsay, B.W. Silverman, Functional Data Analysis, 2nd ed., Springer-Verlag, New York, 2005.
[47] S.J. Ratcliffe, G.Z. Heller, L.R. Leader, Functional data analysis with application to periodically stimulated foetal heart rate data. II: Functional logistic regression, Stat. Med. 21 (2002) 1115–1127.
[48] F. Rossi, N. Villa, Support vector machine for functional data classification, Neurocomputing 69 (2006) 730–742.
[49] N. Rossi, X. Wang, J.O. Ramsay, Nonparametric item response function estimates with the EM algorithm, J. Educ. Behav. Stat. 27 (2002) 291–317.
[50] W. Rudin, Functional Analysis, 2nd ed., China Machine Press, 2004.
[51] H.L. Shang, A survey of functional principal component analysis, Adv. Stat. Anal. 98 (2014) 121–142.
[52] L. Vasak, On the generalization of weakly compactly generated banach spaces, Studia Math 70 (1979) 11–19.
[53] J. Wang, K.F. Huang, H.W. Wang, A cluster method of functional data analysis, Appl. Stat. Manag. 28 (5) (2009) 839–844.
[54] F. Yao, Functional principal component analysis for longitudinal and survival data, Stat. Sin. 17 (2007) 965–983.
[55] F. Yao, H.G. Müller, A.J. Clifford, et al., Shrinkage estimation for functional principal component scores with application to the population kinetics of plasma folate, Biometrics 59 (2003) 676–685.
[56] F. Yao, H.G. Müller, J.L. Wang, Functional data analysis for sparse longitudinal data, J. Am. Stat. Assoc. 100 (2005) 577–590.
[57] G.Q. Zhang, Y.Q. Lin, Functional Analysis, Beijing University Press, 1999.(in Chinese)
[58] M.Q. Zhou, Real Function Theory, Beijing University Press, 1999 (in Chinese).
[59] J.J. Zhou, M. Chen, Spline estimators for semi-functional linear model, Stat. Probab. Lett. 82 (2012) 505–513.