Principal Component Analysis of Functional Data based on Constant Numerical Characteristics

Meiling Chen*,** Huiwen Wang*,**

*School of Economics and Management, Beihang University, Beijing 100191, China
**Research Center of Complex Data Analysis, Beihang University, Beijing 100191, China
mlchen@foxmail.com,

Abstract. A new approach to principal component analysis (PCA) is proposed for functional data. In prevailing methods of functional principal component analysis (FPCA), the definition of a mean is in the form of a function. However, data centralisation based on this kind of mean actually obtains a residual function. The result of FPCA, given its matrix of residual functions, may thus fail to present the essential variation of the original data. Besides, applications in FPCA are mainly for types of one sample problems. Numerical characteristics of functional data are defined as real constants. Centralisation in terms of constant numerical characteristics implies the relocation of the entire matrix of functional variances in order to obtain original curves whose centres of gravity are settled on the origin. Furthermore, based on the covariance matrix obtained from constant numerical characteristics, functional principal components for multivariate sample problems are proposed. Conclusions are validated by simulation in a real situation.

1 Introduction

We propose an extension of principal component analysis (PCA) for classical data to functional data. Functional data analysis (FDA) was first introduced by Ramsay (1982). Later, Ramsay and Silverman (2005) provided a thorough treatment of FDA. In contrast to classical statistical methodology, FDA treats an entire sequence of measurements for an individual as a single functional entity rather than a set of discrete values. Treating the data as a function retains all of the information contained in the data.

We will focus on PCA of functional data. PCA for functional data (FPCA) is described in full by Ramsay and Silverman (2005). Each functional principal component (FPC) represents a particular movement pattern over the time interval considered and represents variation of all the observations around the mean. More recently, FPCA based on the Karhunen-Loeve decomposition has been successfully applied in many areas. Locantore et al. (1999) explored abnormalities in the curvature of the cornea in the human eye. Viviani et al. (2005) compared its functional and multivariate versions and discovered that the functional approach offers a rather better image of experimental manipulations underlying the data. These investigations have time as the covariate. Kneip and Utikal (2001) applied FPCA to describe a set of density curves where the covariate is income. These applications are for types of problems in one

variable. In real life, we often wish to study the variation of more than one function, such as the variation of hip and knee angles described by Ramsay and Silverman (2005). They discussed FPCA about bivariate and multivariate data and summed the inner products of two components obtained by a bivariate function. Another method for two sample problems, called Common functional principal component analysis (CFPCA), was considered by Benko et al. (2009). Then, Coffey et al. (2011) applied CFPCA to the problem of analysing human movements about angle-time series on several groups of individuals. However, the literature on formal treatments of multivariate sample problems is scarce. Much of the discussion on FPCA uses univariate functions of a single variable as the standard process, where the variables of the PCA under discussion refer to the values of the covariate of functions rather than to the functional data set itself. Therefore further research about multivariate PCA needs to be considered.

We attempt to explore a PCA method for multivariate functional variables. PCA aims to reduce the dimensions of a large data set by reconstructing the covariance matrix. Therefore, the key step in PCA is to find the correct covariance matrix. Moreover, the fundamental elements of the covariance matrix, i.e., variances and covariances and their definitions are important in PCA. In first instance, the constant numerical characteristics about functional data are defined, based on the theory of inner products in the vector function space and numerical characteristics by integral calculation on continuous random variables in probability. From this, we obtain a constant type of mean, variance, covariance and correlation coefficient. We then implement these for dimension reduction modelling of a multivariate functional data set through PCA.

Our paper is structured as follows: section 2 states the existing definitions for numerical characteristics of functional data and FPCA proposed in the literature. Section 3 introduces several basic definitions about numerical characteristics of functional data and investigates PCA for multivariate functional data. Simulation is conducted in section 4 to show how the definitions of numerical characteristics affect the results of PCA for univariate functional variables. Then, a real case about Chinese income data is applied in section 5 to validate the effectiveness of the proposed method. The last section provides for a summary.

2 Preliminaries

Generally speaking, PCA is used to reduce dimensionality and collinearity of multivariate data by transforming a set of correlated variables into a new set of uncorrelated variables called principal components (PCs). Given p random variables $\mathbf{X}_1, \mathbf{X}_2, \cdots, \mathbf{X}_p$, the first $m(< p)$ PCs are linear combinations of variables, which retain most of the variation presented in the original variables. Coefficients of the PCs of $\mathbf{X}_1, \mathbf{X}_2, \cdots, \mathbf{X}_p$ are the eigenvectors of the covariance matrix $\mathbf{\Sigma}$. That is, let $\lambda_1 > \lambda_2 > \cdots > \lambda_p$ be the eigenvalues of $\mathbf{\Sigma}$ and let $\mathbf{u}_1, \mathbf{u}_2, \cdots, \mathbf{u}_p$ be the corresponding eigenvectors where $\|\mathbf{u}_k\| = 1 (k = 1, 2, \cdots, p)$. The k^{th} PC is $\mathbf{Y}_k = \mathbf{u}_k^{'}\mathbf{X}$ whose variance is λ_k. Consequently, PCA is the eigen decomposition for the covariance matrix.

First we recall the existing expression of the covariance matrix of FDA and investigate its standard process of PCA. Then PCA for the univariate functional variable is introduced, since we will conduct a simulation in section 4 to show how the definitions of numerical characteristics affect the PCA results for the univariate functional variable. Our approach used the constant covariance matrix defined in section 3.

Ramsay and Silverman (2005) proposed how to summarise statistics for functional data. Existing research for FDA are based on the following definitions. For a univariate functional variable set $\mathbf{X} = \{x_i(t) \in L^2[a,b], i = 1,2,\ldots,n\}$, the mean function with values

$$\overline{x}(t) = n^{-1}\sum_{i=1}^{n} x_i(t) \tag{1}$$

is the average of the functions point-wise across replications. Similarly the variance function

$$\text{var}_X(t) = (n-1)^{-1}\sum_{i=1}^{n}[x_i(t) - \overline{x}(t)]^2 \tag{2}$$

is a measure of how far a set of curves lies from the mean function and the standard deviation function is the square root of the variance function. The covariance function summarises the dependence of records across different values of the argument and is computed for all t_1, t_2

$$\text{cov}_X(t_1,t_2) = (n-1)^{-1}\sum_{i=1}^{n}[x_i(t_1) - \overline{x}(t_1)][x_i(t_2) - \overline{x}(t_2)] \tag{3}$$

and the associated correlation function is

$$\text{corr}_X(t_1,t_2) = \frac{\text{cov}_X(t_1,t_2)}{\sqrt{\text{var}_X(t_1)\text{var}_X(t_2)}} \tag{4}$$

For pairs of observed functions (x_i, y_i), measured over time, the cross-covariance can be used to quantify the dependence between the two functional variables, as follows:

$$\text{cov}_{X,Y}(t_1,t_2) = (n-1)^{-1}\sum_{i=1}^{n}[x_i(t_1) - \overline{x}(t_1)][y_i(t_2) - \overline{y}(t_2)] \tag{5}$$

and the corresponding cross-correlation function:

$$\text{corr}_{X,Y}(t_1,t_2) = \frac{\text{cov}_{X,Y}(t_1,t_2)}{\sqrt{\text{var}_X(t_1)\text{var}_Y(t_2)}} \tag{6}$$

In general

$$\text{corr}_{X,Y}(t_1,t_2) = \text{corr}_{Y,X}(t_2,t_1) \tag{7}$$

In order to make a distinction between the two kinds of numerical characteristics for functional data, the mean function, covariance function, associated correlation function, cross-covariance and cross-correlation function defined above are called functional numerical characteristics (FNC). These indicators have the following two properties: (i) they are functions with time t as the covariate; (ii) for the purpose of calculating the correlation coefficient and covariance, t_1, t_2 are regarded as different variables. Therefore, it would be difficult to measure the relationship between the two functional variables $\mathbf{X}, \mathbf{Y}$.

According to the formula(1), centralizing of $x_i(t), (i = 1,2,\ldots,n)$ is given by

$$\tilde{x}_i(t) = x_i(t) - \overline{x}(t) \tag{8}$$

With the centralisation in terms of the mean function shown in equation (8), the mean of the functions point-wise across replications is zero.

Next, PCA for univariate functional variables referred to as FPCA (Ramsay and Silverman, 2005) will be introduced. After obtaining a set of functional data, we often need to study the characteristics of these functions further, trying to find their internal rules of change. The object of FPCA is to find out that group of indicators which can express these internal rules of a given series sample most efficiently.

For univariate functional data, $\mathbf{X} = \{x_i(t) \in L^2[a,b], i = 1, 2, \ldots, n\}$. The way of consolidating every $x_i(t)$ in the interval $[a,b]$ to a comprehensive variable is as follows:

$$f_i = \int_a^b \beta(t) x_i(t) dt \tag{9}$$

where $\beta(t)$ is the weighting coefficient function, corresponding to the factor axis in the multivariate situation. We require that the comprehensive variable accounts for the total variation as much as possible, i.e., its variance should be the largest. For each comprehensive variable, its variance is expressed as follows:

$$\text{var}(\mathbf{f}_k) = n^{-1} \sum_{i=1}^{n} [\int_a^b \beta_k(t) x_i(t) dt]^2 \tag{10}$$

where $\mathbf{f}_k = (f_{1k}, f_{2k}, \cdots, f_{nk}), k = 1, 2, \cdots, m(m < p)$ is the k^{th} principal component of functional observations and $\text{var}(\mathbf{f}_1) \geq \text{var}(\mathbf{f}_2) \geq \cdots \geq \text{var}(\mathbf{f}_m)$.

The methodology of FPCA is similar to that of the multivariate PCA for ordinary data. For the univariate functional data set $\mathbf{X} = \{x_i(t) \in L^2[a,b], i = 1, 2, \ldots, n\}$, centralisation for every sample function is performed by subtracting the mean of all sample functions. To simplify the expression, we can also use $x_i(t)$ for the function after centralisation. Then, the solution of the weighting coefficient function $\beta_1(t)$ corresponding to the first principal component has become the following maximisation problem:

$$\begin{aligned} \max \quad & n^{-1} \sum_{i=1}^{n} (\int_a^b \beta_1(t) x_i(t) dt)^2 \\ s.t. \quad & \int_a^b [\beta_1(t)]^2 dt = \|\beta_1(t)\|^2 = 1 \end{aligned} \tag{11}$$

Similarly, the solution of the corresponding weighting coefficient function $\beta_k(t)$ of the k^{th} principal component becomes the following maximisation problem:

$$\begin{aligned} \max \quad & n^{-1} \sum_{i=1}^{n} (\int_a^b \beta_k(t) x_i(t) dt)^2 \\ s.t. \quad & \int_a^b [\beta_k(t)]^2 dt = \|\beta_k(t)\|^2 = 1 \\ & \int_a^b \beta_k(t) \beta_{k-m}(t) dt = 0, m = 1, \cdots, k-1 \end{aligned} \tag{12}$$

After obtaining the weighting coefficient function corresponding to the k^{th} principal component, the principal component score of the i^{th} sample curve on the k^{th} principal component is calculated as follows:

$$f_{ik} = \int_a^b \beta_k(t) x_i(t) dt \tag{13}$$

More details about FPCA can be found in Ramsay and Silverman (2005). Comparing FPCA with classical PCA, we can see clearly that the variables discussed on FPCA refer to different values of the independent variable, such as time t. It is natural to ask the question, how to conduct FPCA for multivariable functional data. Ramsay and Silverman (2005) provide a treatment of FPCA for bivariate and multivariate functional data based on the univariate one. They argue that one of the most important features of the functional data analysis approach to PCA is that, once the inner product has been defined appropriately, PCA looks formally the same. How to explore an appropriate definition of this inner product to simplify the process of PCA for variables of functional data, motivates our study.

3 PCA of Functional Data based on Constant Numerical Characteristics

3.1 The Constant Covariance Matrix for Functional Data

In this section, we first define the inner product operator for functional data. We then propose several basic concepts about constant numerical characteristics (CNC), used in the investigation of PCA of multivariate functional data, such as the constant-style mean, variance, covariance and correlation coefficient of functional data. Included is the method of centralising and standardising functional data.

Definition 1: In the space of functional data, the inner product of two functions $x(t) \in L^2[a,b]$ and $y(t) \in L^2[a,b]$ is defined as

$$\langle x(t), y(t) \rangle = \int_a^b x(t) y(t) dt \tag{14}$$

Accordingly, the inner product of two n dimensional functional variables $\mathbf{X} = (x_1(t), x_2(t), \cdots, x_n(t))'$, $\mathbf{Y} = (y_1(t), y_2(t), \cdots, y_n(t))'$ is defined as

$$\langle \mathbf{X}, \mathbf{Y} \rangle = \sum_{i=1}^{n} (\int_a^b x_i(t) y_i(t) dt) \tag{15}$$

The inner product for vectors of functional data satisfies the property of inner product space as follows: (its proof is found in mathematical analysis).

Proposition 1. For any functional data vectors $\mathbf{X}, \mathbf{Y}, \mathbf{Z}$, the inner product satisfies

(1) Positive definiteness, i.e., $\langle \mathbf{X}, \mathbf{X} \rangle \geq 0$; the equality holds iff $\mathbf{X} = \mathbf{0}$;

(2) Symmetry, i.e., $\langle \mathbf{X}, \mathbf{Y} \rangle = \langle \mathbf{Y}, \mathbf{X} \rangle$;

(3) Linearity, i.e., $\langle \mathbf{X} + \mathbf{Y}, \mathbf{Z} \rangle = \langle \mathbf{X}, \mathbf{Z} \rangle + \langle \mathbf{Y}, \mathbf{Z} \rangle$ and $\langle \alpha \mathbf{X}, \mathbf{Y} \rangle = \alpha \langle \mathbf{X}, \mathbf{Y} \rangle, \alpha \in \mathbb{R}$.

Furthermore, the inner product also can induce a squared norm and a distance in functional vector space in the standard way.

$$\|\mathbf{X}\|^2 = \langle \mathbf{X}, \mathbf{X} \rangle = \sum_{i=1}^{n} (\int_a^b x_i(t)^2 dt) \tag{16}$$

$$d^2(\mathbf{X}, \mathbf{Y}) = \|\mathbf{X} - \mathbf{Y}\|^2 \tag{17}$$

Then, CNC of functional data, such as the constant-style mean of functional data, variance, covariance and correlation coefficient of functional data can be proposed as follows:
Definition 2: For a functional variable $\mathbf{X} = (x_1(t), x_2(t), \cdots, x_n(t))'$, $x_i(t) \in L^2[a, b], i = 1, 2, \cdots, n$, the constant-style mean is given by

$$E(\mathbf{X}) = n^{-1} \sum_{i=1}^{n} \int_a^b x_i(t) dt \tag{18}$$

Accordingly, the centralisation of $x_i(t)$ is given by:

$$\tilde{x}_i(t) = x_i(t) - E(\mathbf{X}), i = 1, 2, \cdots, n \tag{19}$$

Definition 3: For the functional variable $\mathbf{X} = (x_1(t), x_2(t), \cdots, x_n(t))'$, $x_i(t) \in L^2[a, b], i = 1, 2, \cdots, n$, the constant-style variance is given as follows:

$$\text{var}(\mathbf{X}) = n^{-1} \sum_{i=1}^{n} \int_a^b [x_i(t) - E(\mathbf{X})]^2 dt \tag{20}$$

based on Definition 1, it can also be denoted as

$$\text{var}(\mathbf{X}) = n^{-1} \langle \mathbf{X} - E(\mathbf{X}), \mathbf{X} - E(\mathbf{X}) \rangle \tag{21}$$

according to equations (18) and (20), the functional variable can be standardised as follows:

$$x_i^*(t) = \frac{x_i(t) - E(\mathbf{X})}{\sqrt{\text{var}(\mathbf{X})}}, i = 1, 2, \cdots, n \tag{22}$$

It can be easily deduced that the standardised functional variable $\mathbf{X}^* = (x_1^*(t), x_2^*(t), \cdots, x_n^*(t))$ satisfies the following properties:

Proposition 2. $E(\mathbf{X}^*) = n^{-1} \sum_{i=1}^{n} \int_a^b x_i^*(t) dt = 0$.

Proposition 3. $\text{var}(\mathbf{X}^*) = n^{-1} \sum_{i=1}^{n} \int_a^b [x_{ij}^*(t) - E(\mathbf{X}^*)]^2 dt$=1.

Definition 4: For any two functional variables $\mathbf{X} = (x_1(t), x_2(t), \cdots, x_n(t))'$, and $\mathbf{Y} = (y_1(t), y_2(t), \cdots, y_n(t))'$, the constant-style covariance is given by

$$\text{cov}(\mathbf{X}, \mathbf{Y}) = n^{-1} \sum_{i=1}^{n} \int_a^b [x_i(t) - E(\mathbf{X})][y_i(t) - E(\mathbf{Y})] dt \tag{23}$$

Following the convention of equation (15), it can also be denoted as

$$\text{cov}(\mathbf{X},\mathbf{Y}) = n^{-1}\langle \mathbf{X} - E(\mathbf{X}), \mathbf{Y} - E(\mathbf{Y})\rangle \tag{24}$$

Combining equation (20) and (23), the correlation coefficient of two functional variables can be defined as:

$$corr(\mathbf{X},\mathbf{Y}) = \frac{\text{cov}(\mathbf{X},\mathbf{Y})}{\sqrt{\text{var}(\mathbf{X})}\sqrt{\text{var}(\mathbf{Y})}} \tag{25}$$

For a $n \times p$ functional sample data matrix $\mathbf{X}_{n\times p}$, we have

$$\mathbf{X}_{n\times p} = (\mathbf{X}_1, \mathbf{X}_2, \cdots, \mathbf{X}_p) = \begin{pmatrix} \mathbf{O}_1 \\ \mathbf{O}_2 \\ \vdots \\ \mathbf{O}_n \end{pmatrix} = \begin{pmatrix} x_{11}(t) & x_{12}(t) & \cdots & x_{1p}(t) \\ x_{21}(t) & x_{22}(t) & \cdots & x_{2p}(t) \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1}(t) & x_{n2}(t) & \cdots & x_{np}(t) \end{pmatrix} \tag{26}$$

where $x_{ij}(t) \in L^2[a,b]$; accordingly, the covariance matrix $\mathbf{\Sigma}$ of $\mathbf{X}_{n\times p}$ can be given by

$$\mathbf{\Sigma} = n^{-1} \begin{pmatrix} \langle \mathbf{X}_1 - E(\mathbf{X}_1), \mathbf{X}_1 - E(\mathbf{X}_1)\rangle & \cdots & \langle \mathbf{X}_1 - E(\mathbf{X}_1), \mathbf{X}_p - E(\mathbf{X}_p)\rangle \\ \vdots & \ddots & \vdots \\ \langle \mathbf{X}_1 - E(\mathbf{X}_1), \mathbf{X}_p - E(\mathbf{X}_p)\rangle & \cdots & \langle \mathbf{X}_p - E(\mathbf{X}_p), \mathbf{X}_p - E(\mathbf{X}_p)\rangle \end{pmatrix} \tag{27}$$

Next, if all data units $x_{ij}(t)$ have been centralised by equation (19), we have

$$\text{var}(\mathbf{X}_j) = n^{-1}\langle \mathbf{X}_j, \mathbf{X}_j\rangle \tag{28}$$

$$\text{cov}(\mathbf{X}_j, \mathbf{X}_k) = n^{-1}\langle \mathbf{X}_j, \mathbf{X}_k\rangle \tag{29}$$

then the covariance matrix $\mathbf{\Sigma}$ can be simply defined as

$$\mathbf{\Sigma} = n^{-1} \begin{pmatrix} \langle \mathbf{X}_1, \mathbf{X}_1\rangle & \langle \mathbf{X}_1, \mathbf{X}_2\rangle & \cdots & \langle \mathbf{X}_1, \mathbf{X}_p\rangle \\ \langle \mathbf{X}_1, \mathbf{X}_2\rangle & \langle \mathbf{X}_2, \mathbf{X}_2\rangle & \cdots & \langle \mathbf{X}_2, \mathbf{X}_p\rangle \\ \vdots & \vdots & \ddots & \vdots \\ \langle \mathbf{X}_1, \mathbf{X}_p\rangle & \langle \mathbf{X}_2, \mathbf{X}_p\rangle & \cdots & \langle \mathbf{X}_p, \mathbf{X}_p\rangle \end{pmatrix} \tag{30}$$

Similarly, the correlation matrix $\mathbf{R}$ can be given by

$$\mathbf{R} = \begin{pmatrix} 1 & \frac{\text{cov}(\mathbf{X}_1,\mathbf{X}_2)}{\sqrt{\text{var}(\mathbf{X}_1)}\sqrt{\text{var}(\mathbf{X}_2)}} & \cdots & \frac{\text{cov}(\mathbf{X}_1,\mathbf{X}_p)}{\sqrt{\text{var}(\mathbf{X}_1)}\sqrt{\text{var}(\mathbf{X}_p)}} \\ \frac{\text{cov}(\mathbf{X}_1,\mathbf{X}_2)}{\sqrt{\text{var}(\mathbf{X}_1)}\sqrt{\text{var}(\mathbf{X}_2)}} & 1 & \cdots & \frac{\text{cov}(\mathbf{X}_2,\mathbf{X}_p)}{\sqrt{\text{var}(\mathbf{X}_2)}\sqrt{\text{var}(\mathbf{X}_p)}} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\text{cov}(\mathbf{X}_1,\mathbf{X}_p)}{\sqrt{\text{var}(\mathbf{X}_1)}\sqrt{\text{var}(\mathbf{X}_p)}} & \frac{\text{cov}(\mathbf{X}_2,\mathbf{X}_p)}{\sqrt{\text{var}(\mathbf{X}_2)}\sqrt{\text{var}(\mathbf{X}_p)}} & \cdots & 1 \end{pmatrix} \tag{31}$$

Note that the covariance matrix $\mathbf{\Sigma}$ will be equal to the correlation matrix $\mathbf{R}$, if all the functional units $x_{ij}(t)$ have been standardised. From here on, the covariance matrix $\mathbf{\Sigma}$ whose elements are the constant-style variances and covariances, will be referred to as the constant-style covariance matrix of multivariable functional data.

3.2 Multivariate functional PCA based on Constant Covariance Matrix

Based on the constant-style covariance matrix of multivariate functional data, we begin by deriving the functional PCs. For simplicity of notation, we assumed that all the functional data units have been centralised. Similar to the numeric case, the k^{th} functional PC (FPC) $\mathbf{Y}_k (k = 1, 2, \cdots, p)$ is a linear combination of $\mathbf{X}_1, \mathbf{X}_2, \cdots, \mathbf{X}_p$, i.e., $\mathbf{Y}_k = \mathbf{X}\mathbf{u}_k = \sum_{j=1}^{p} u_{kj}\mathbf{X}_j$, where $u_{kj} \in \mathbb{R}$ $(j = 1, 2, \cdots, p)$, with the constraints $\|\mathbf{u}_k\| = 1$ and $\mathbf{u}_k' u_l = 0$ $(l = 1, 2, \cdots, p,\ l \neq k)$. Also, the first m FPC $\mathbf{Y}_1, \mathbf{Y}_2, \cdots, \mathbf{Y}_m$ must maximise total variance as much as possible to represent the original information carried by $\mathbf{X}_1, \mathbf{X}_2, \cdots, \mathbf{X}_p$. According to the definitions proposed earlier, we have

$$\begin{aligned}
\text{var}(\mathbf{Y}_k) &= n^{-1}\langle \mathbf{Y}_k, \mathbf{Y}_k \rangle \\
&= n^{-1}\langle u_{k1}\mathbf{X}_1 + u_{k2}\mathbf{X}_2 + \cdots + u_{kp}\mathbf{X}_p, u_{k1}\mathbf{X}_1 + u_{k2}\mathbf{X}_2 + \cdots + u_{kp}\mathbf{X}_p \rangle \\
&= n^{-1}(u_{k1}, u_{k2}, \cdots, u_{kp}) \begin{pmatrix} \langle \mathbf{X}_1, \mathbf{X}_1 \rangle & \langle \mathbf{X}_1, \mathbf{X}_2 \rangle & \cdots & \langle \mathbf{X}_1, \mathbf{X}_p \rangle \\ \langle \mathbf{X}_1, \mathbf{X}_2 \rangle & \langle \mathbf{X}_2, \mathbf{X}_2 \rangle & \cdots & \langle \mathbf{X}_2, \mathbf{X}_p \rangle \\ \vdots & \vdots & \ddots & \vdots \\ \langle \mathbf{X}_1, \mathbf{X}_p \rangle & \langle \mathbf{X}_2, \mathbf{X}_p \rangle & \cdots & \langle \mathbf{X}_p, \mathbf{X}_p \rangle \end{pmatrix} \begin{pmatrix} u_{k1} \\ u_{k2} \\ \vdots \\ u_{kp} \end{pmatrix} \\
&= \mathbf{u}_k' \mathbf{\Sigma} \mathbf{u}_k \qquad (32)
\end{aligned}$$

where $\mathbf{\Sigma}$ represents the covariance matrix of $\mathbf{X}_1, \mathbf{X}_2, \cdots, \mathbf{X}_p$.

The following derivation is the same as classical PCA for numeric data, i.e., looking for m orthonormalised vectors $\mathbf{u}_1, \mathbf{u}_2, \cdots, \mathbf{u}_m$ that maximise $\sum_{k=1}^{m} \mathbf{var}(\mathbf{Y}_k)$, where $\mathbf{var}(\mathbf{Y}_1) \geq \mathbf{var}(\mathbf{Y}_2) \geq \cdots \geq \mathbf{var}(\mathbf{Y}_m)$ in the following optimisation problem.

$$\begin{aligned}
\max \quad & \sum_{k=1}^{m} \mathbf{u}_k' \mathbf{\Sigma} \mathbf{u}_k \\
s.t. \quad & \|\mathbf{u}_k\| = 1 \\
& \mathbf{u}_k' \mathbf{u}_l = 0 \\
& \mathbf{u}_1' \mathbf{\Sigma} \mathbf{u}_1 \geq \mathbf{u}_2' \mathbf{\Sigma} \mathbf{u}_2 \geq \cdots \geq \mathbf{u}_m' \mathbf{\Sigma} \mathbf{u}_m \\
& k = 1, 2, \cdots, p, l \neq k \qquad (33)
\end{aligned}$$

Thus, $\mathbf{u}_1, \mathbf{u}_2, \cdots, \mathbf{u}_m$ are the eigenvectors of $\mathbf{\Sigma}$, corresponding to the eigenvalues $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_m$. The derivation of the FPC coefficients is also the eigen decomposition of the constant-style covariance matrix. Finally, we obtain the k^{th} FPC $\mathbf{Y}_k = \mathbf{X}\mathbf{u}_k$.

Therefore, functional PCA based on a constant covariance matrix can be summarised as the following algorithm:

Step 1: Compute the constant-style covariance matrix $\mathbf{\Sigma}$ (the correlation matrix $\mathbf{R}$) of $\mathbf{X}_{n\times p}$, using equation (27) (the correlation matrix $\mathbf{R}$ obtained by (31)).

Step 2: Solve $\mathbf{\Sigma}\mathbf{u}_p = \lambda_k \mathbf{u}_p$ $(1 \leq k \leq p)$ for for the eigenvalues $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p$ and the corresponding orthonormalised eigenvectors $\mathbf{u}_1, \mathbf{u}_2, \cdots, \mathbf{u}_p$.

Step 3: Compute the k^{th} FPC $\mathbf{Y}_k = \mathbf{X}\mathbf{u}_k$ $(1 \leq k \leq p)$, where $\mathbf{Y}_k$ is an n-dimensional vector function.

We call this method the Multivariate Functional PCA (abbreviated as MFPCA).

3.3 Properties and interpretation

Similar to those from classical PCA, the FPCs $\mathbf{Y}_1, \mathbf{Y}_2, \cdots, \mathbf{Y}_m (1 \leq m \leq p)$, derived by MFPCA, satisfy the following properties.

If $\mathbf{X}_1, \mathbf{X}_2, \cdots, \mathbf{X}_p$ have been standardised, we have

Proposition 4.

$$E(\mathbf{Y}_h) = 0, h = 1, 2, \cdots, m \tag{34}$$

Proof. $E(\mathbf{Y}_h) = E(\sum_{j=1}^{p} u_{hj}\mathbf{X}_j) = \sum_{j=1}^{p} u_{hj}E(\mathbf{X}_j) = 0.$

Proposition 5.

$$\text{var}(\mathbf{Y}_h) = \lambda_h,\ h = 1, 2, \cdots, m \tag{35}$$

Proof. $\text{var}(\mathbf{Y}_h) = \mathbf{u}_h'\mathbf{\Sigma}\mathbf{u}_h = \mathbf{u}_h'\lambda_h\mathbf{u}_h = \lambda_h\mathbf{u}_h'\mathbf{u}_h = \lambda_h.$

Proposition 6.

$$\text{cov}(\mathbf{Y}_h, \mathbf{Y}_l) = 0,\ \ for\ \ h \neq l \tag{36}$$

Proof. $\text{cov}(\mathbf{Y}_h, \mathbf{Y}_l) = \mathbf{u}_h'\mathbf{\Sigma}\mathbf{u}_l = \mathbf{u}_h'\lambda_l\mathbf{u}_l = \lambda_l\mathbf{u}_h'\mathbf{u}_l = 0.$

The cumulative contribution rate (CCR) of the first m FPCs to the total information can be measured by:

$$\text{CCR} = \frac{\sum_{k=1}^{m} \text{var}(\mathbf{Y}_k)}{\sum_{j=1}^{p} \text{var}(\mathbf{X}_j)} = \frac{\sum_{k=1}^{m} \lambda_k}{\sum_{j=1}^{p} \lambda_j} \tag{37}$$

CCR reports the percentage of the original information contained in m-dimensional space, spanned by the reserved factor axes $\mathbf{u}_1, \mathbf{u}_2, \cdots, \mathbf{u}_m$ and works as an assistant in determining m, the number of reserved principal components.

As well, a loading plot based on $corr(\mathbf{Y}_h, \mathbf{X}_j)$ can display the interpretation of principal components by the original variables.

4 Experimental Results of a Synthetic Data Set

In this section, we carry out a simple experiment to test the results of the FPCA (PCA for univariate functional data) according to section 2, based on different numerical characteristics. We pretreated the same generated synthetic data by two kinds of numerical characteristics of functional data and compared the FPCA results.

4.1 The Synthetic Data

We employ the following method to generate 15 random functional observations.

$$x(t) = U + \sin(t) + \varepsilon(t), t \in [0.2, 5.2] \tag{38}$$

where, U follows the uniform distribution in [2,3], and the random error $\varepsilon(t)$ follows an uniform distribution in [-0.25,0.25]. As a result, an univariate functional dataset $\mathbf{X} = (x_1(t), x_2(t), \cdots, x_{15}(t))'$, as show in Fig.1, is obtained.

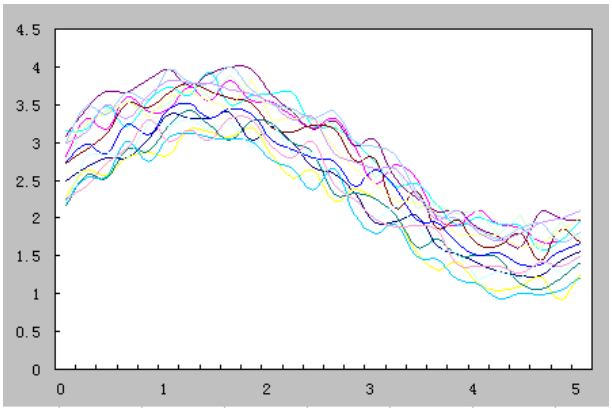

FIG. 1 – *Generated simulation data* $x(t)$

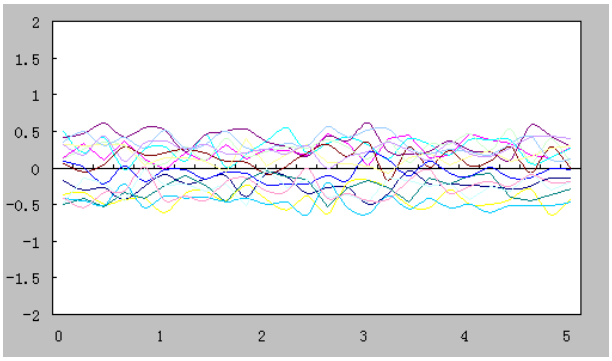

FIG. 2 – *Centralised curves based on FNC*

4.2 FPCA Experimental Result based on FNC

The experiment can be implemented by the two steps as follows: the first step is centralisation of the simulation data and the second step is treating the data by FPCA.

By using equations (1) and (8), the mean function and the centralised functions are obtained. Fig.2 presents the result of the centralised functions. It should be noted that only residual curves remained after subtracting the mean function, whose gravity is settled on the horizontal axis across the argument interval.

Next, by using FPCA, the extracted PC variances and CCRs are obtained. The first and second component, respectively, account for 21.3% and 15.8% of the original information. Fig.3 displays the weighting coefficient functions corresponding to the first two principal components. As we can see from Fig.3, the values of the corresponding weighting coefficient functions of these two principal components fluctuate randomly throughout the entire interval.

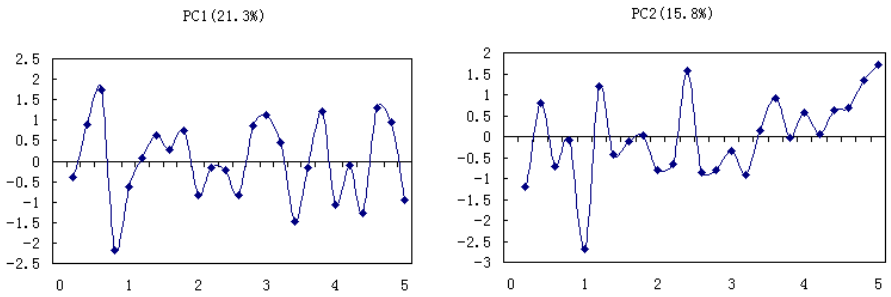

FIG. 3 – *Weighting coefficient functions for the first two FPCs*

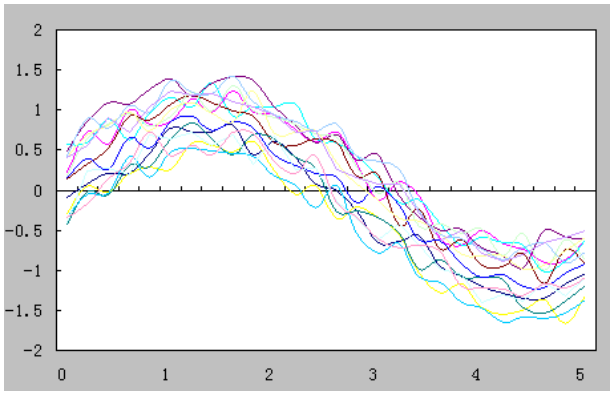

FIG. 4 – *Centralised curves based on CNC*

4.3 FPCA Experimental Result based on CNC

In the following, we will treat the same synthetic data set to FPCA with CNC. Thus we can compare the FPCA results based on FNC and CNC. Similarly, our experiment consists of the following steps.

First, we calculated the constant-style mean by using equation (18). Then the simulation data were centralised and the results are shown in Fig.4. The results of the centralisation show extreme differences between Figs. 2 and 4. It is to be noted in Fig.4, that the curves have moved but did not change shape. Centralisation in this method only translates the curves as a whole to a new position whose barycentre is settled on the horizontal axis. This definition is more in accord with the nature of traditional centralisation than the existing definition.

The implemented dimension reduction through FPCA to the centralised curves, shown in Fig.4 provided the extracted PC variances and CCRs. Fig.5 denotes the weighting coefficient functions corresponding to the first principal component. The results, shown in Fig.5, indicate that the first principal component explains 96.8% of the overall variation in the data, and can be interpreted as the main development trend over all the observations.

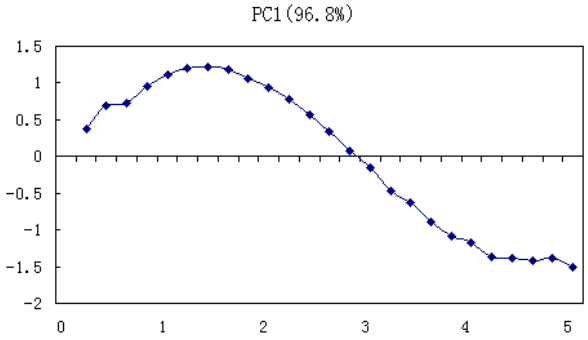

FIG. 5 – *Weighting coefficient functions for PC1*

4.4 Discussion

Experimental results of our synthetic data set show that the fundamental elements of the covariance matrix such as the variances and the covariances, as well as their mode of definition, are important in PCA.

For FDA, the mean function is the average of the functions point-wise across replications and reflects the average level in every point of the sample function, while the constant-style mean is the gravity of the entire functional variable set, when treating all the curves as an entity. Moreover, FPCA based on FNC analyses the error functions, while FPCA based on CNC explores the main variation patterns.

5 Analysis of Real-life Applications

For illustration, we applied MFPCA to the data of Chinese national income. Theoretically, the pattern of national income refers to its distribution among government, enterprises and residents. We attempted to adopt MFPCA to investigate the structure of these three types of income and, as well, validated the effectiveness of the proposed approach.

5.1 Data

The set includes data from 31 provinces and areas in China, identified by three variables, i.e., compensation of employees, government revenue and income of enterprises, for the period 1993 to 2007, adjusted for inflation. We transformed the discrete data of these three variables to smoothing functions, with the b-spline functions as the basic functions, using the Roughness Penalty Smoothing method. Then the three groups (93 smoothing functions) represent the process of dynamic changes in the distribution pattern of our national income in terms of these three variables. We defined compensation of employees, government revenue and income of enterprises as independent functional variables, denoted as $\mathbf{X}_1, \mathbf{X}_2, \mathbf{X}_3$ respectively.

	$\mathbf{X}_1$	$\mathbf{X}_2$	$\mathbf{X}_3$
$\mathbf{X}_1$	1	0.96	0.96
$\mathbf{X}_2$	0.96	1	0.98
$\mathbf{X}_3$	0.96	0.98	1

TAB. 1 – *Correlation matrix of independent variables*

FPCs	FPC Variance	CCR(%)
1	2.981	99.4
2	0.013	99.8
3	0.006	100

TAB. 2 – *FPC Variances and CCRs*

5.2 The Results of MFPC

The correlation matrix is obtained from equation (31), as shown in Tab.1, showing correlation coefficients of the independent variables as high as 0.95 or higher.

By using the proposed analytical approach, we obtained the extracted variances of FPCs and the CCRs in Tab.2.

The first component accounts for 97.63% of the variation. As a consequence, more than 95% of the total information can be summarised for the original variables. In other words, the first component is sufficient to represent the original variable data set.

$$\mathbf{Y}_1 = 0.989\mathbf{X}_1 + 0.995\mathbf{X}_2 + 0.994\mathbf{X}_3 \tag{39}$$

The correlation between the PCs and the original variables is shown in Fig.6. As can be seen, the first FPC denoted FPC1 is positively associated with all three variables. Therefore, we can refer to FPC1 as the "distribution volume". Furthermore, we synthetically evaluated the speed of income development of the 31 provinces and areas according to FPC1 curves of each province or area. The FPC1 curves can be visualized. Each curve in Fig.7 is a linear combination of the three independent functional variables.

5.3 Analysis of FPCA of Chinese Income Data

On the basis of the results of sections 5.2 we applied FPCA, mentioned in section 2, to the set of curves $\mathbf{Y}_1$ shown in Fig.7. The results indicate that the first principal component (PC1) explains 99.4% of the overall variation in the data. Fig.8 displays the weighting coefficient function corresponding to the first principal component. The curve in Fig.8 can be interpreted as the main development trend over all 31 provinces and areas. Furthermore, scores of the 31 FPC1 curves on PC1 and the second principal component (PC2) can be calculated from equation (13). Fig.9 presents the observations along the PC1 and PC2. The PC2 scores of the 31 provinces and areas are not significance, since PC2 explains only an additional 0.4%. Therefore, according to the location on the PC1 axis of each area, we synthetically evaluated the income situation of the 31 provinces and areas. As well, we can rank the 31 provinces and

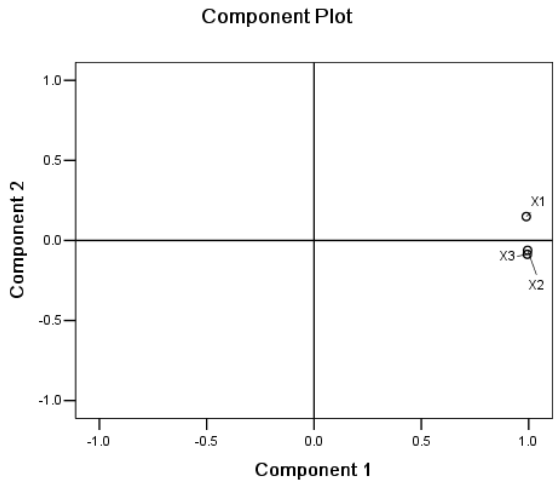

FIG. 6 – *linear correlation between the first two FPCs and original variables*

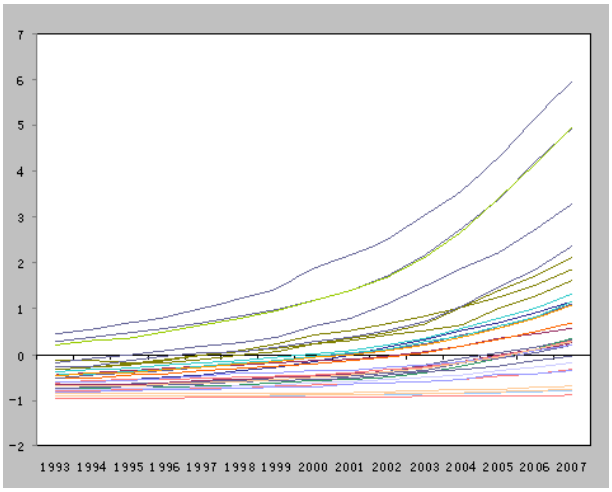

FIG. 7 – *FPC1 plot for Chinese income data*

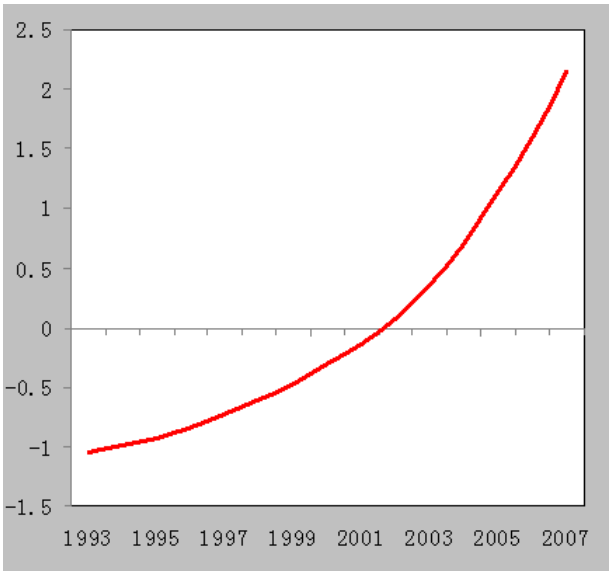

FIG. 8 – *Weighting coefficient functions for FPC1* $\beta_1(t)$

areas by the scores on PC1. It should be noted that the data only shows the total amount of income, but does not reflect the average income of the 31 different areas.

6 Concluding remarks

We proposed and have presented an analytical approach to principal components analysis for multivariate functional data. Our approach featured an analytical variance-covariance structure based on constant numerical characteristics and an easy-to-implement algorithm to obtain the coefficients vectors for the principal components and portions of information on the variation in the original data. We also presented a simple but suitable experiment data set to test the viability of the theoretical results on PCA for univariate functional data, based on different numerical characteristics. The application of our method to the Chinese income data suggests the potential of our method to shed new light on actual problems. This ensures a wide range of possible applications of our method in economics and management.

Acknowledgements

This project was supported by the National Natural Science Foundation of China (Grants No. 71031001, 70771004).

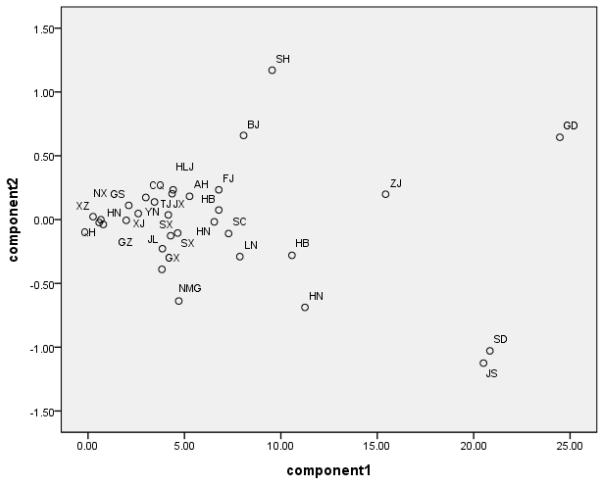

FIG. 9 – *PC1 × PC2 plot for Chinese income data*

References

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

Coffey, N., A. J. Harrison, O. Donoghue, and K. Hayes (2011). Common functional principal components analysis: A new approach to analyzing human movement data. *Human Movement Science 30*, 1144–1166.

Kneip, A. and K. Utikal (2001). Inference for density families using functional principal components analysis. *Journal of the American Statistical Association 96*, 519–542.

Locantore, N., J. S. Marron, D. G. Simpson, N. Tripoli, J. T. Zhang, and K. L. Cohen (1999). Robust principal component analysis for functional data. *Test 8*, 1–34.

Ramsay, J. (1982). When the data are functions. *Psychometrika 47*, 379–396.

Ramsay, J. and B. Silverman (2005). *Functional Data Analysis-Second Edition*. New York: Springer Science+Business Media.

Viviani, R., G. Grön, and M. Spitzer (2005). Functional principal component analysis of fmri data. *Human Brain Mapping 24*, 109–129.