Hindawi Publishing Corporation
Discrete Dynamics in Nature and Society
Volume 2014, Article ID 365204, 7 pages
http://dx.doi.org/10.1155/2014/365204

Research Article

Functional Principal Components Analysis of Shanghai Stock Exchange 50 Index

Zhiliang Wang, Yalin Sun, and Peng Li

College of Mathematics and Informatics, North China University of Water Conservancy and Hydroelectric Power, Zhengzhou 450000, China

Correspondence should be addressed to Zhiliang Wang; wzl@ncwu.edu.cn

Received 28 February 2014; Revised 23 May 2014; Accepted 4 July 2014; Published 22 July 2014

Academic Editor: Seenith Sivasundaram

The main purpose of this paper is to explore the principle components of Shanghai stock exchange 50 index by means of functional principal component analysis (FPCA). Functional data analysis (FDA) deals with random variables (or process) with realizations in the smooth functional space. One of the most popular FDA techniques is functional principal component analysis, which was introduced for the statistical analysis of a set of financial time series from an explorative point of view. FPCA is the functional analogue of the well-known dimension reduction technique in the multivariate statistical analysis, searching for linear transformations of the random vector with the maximal variance. In this paper, we studied the monthly return volatility of Shanghai stock exchange 50 index (SSE50). Using FPCA to reduce dimension to a finite level, we extracted the most significant components of the data and some relevant statistical features of such related datasets. The calculated results show that regarding the samples as random functions is rational. Compared with the ordinary principle component analysis, FPCA can solve the problem of different dimensions in the samples. And FPCA is a convenient approach to extract the main variance factors.

1. Introduction

In the present study of data analysis we have learned, the data we research is either cross-sectional data or panel data. In the practical research, however, we often meet with such data which has functional characteristics. Functional data is multivariate data with an ordering on the dimensions [1]. The data seem to deserve the label "functional" since they so clearly reflect the smooth curves that we assume generated them. The typical dataset of this sort consists of time series and cross-sectional data, such as the time series of stock price, and some datasets even may take on curves or images. Advances in data collection and storage have tremendously increased the presence of such functional data, whose graphical representations are curves, images, or shapes. The theoretical and practical developments in functional data analysis are mainly from the last four decades, due to the rapid development of computer recording and storing facilities. As a new area of statistics, functional data analysis extends existing methodologies and theories from the fields of data analysis, generalized linear models, multivariate data analysis, nonparametric statistics, and many others. Recently, there were several impressive attempts to analyze functional dataset such as Ramsay et al. [2–5], who proposed some new concepts and methods in the field of FDA.

FPCA is the functional analogue of the well-known dimension reduction technique in the multivariate statistical analysis and is useful in determining the common factors or trends that are present in the dynamics of the underlying recovered functions.

The advance of FPCA can be seen when Karhunen [7] and Loève [8] independently developed a theory on the optimal series expansion of a continuous stochastic process. Motivated by a dataset of growth curve, Rao [9] developed some preliminary ideas on FPCA and proposed statistical tests for the equality of average growth curves over a period of time. Much later, Dauxois et al. [10] introduced a functional exposition of PCA with applications to statistical inference. Several other notable developments have arisen out of the

Table 1: The differences in notation between PCA and FPCA [6].

	PCA	FPCA
Variables	$X = [x_1, x_2, \dots, x_n], x_i = \left[x_{1i}, \dots, x_{pi}\right]', \ i = 1, \dots, n$	$f(t) = \left[f_1(t), f_2(t), \dots, f_{\ n}(t)\right], t \in \left[x_1, x_p\right]$
Data	Vectors $\in R^P$	Curves $\in L_2\left[x_1, x_p\right]$
Covariance	Matrix $V = \mathrm{Cov}(X) \in R^P \times R^P$	Operator V bounded between x_1 and x_p, $\phi_k(t) \in L_2\left[x_1, x_p\right], \int_{x_1}^{x_p} V\xi_k(t)\,dt = \lambda_k \xi_k(t)$ $V: L_2\left[x_1, x_p\right] \rightarrow L_2\left[x_1, x_p\right]$
Eigen structure	Vector $\Phi_k \in R$, $V\Phi_k = \lambda_k \Phi_k$, for $1 \le k \le \min(n, p)$	Function $\phi_k(t) \in L_2\left[x_1, x_p\right]$, $\int_{x_1}^{x_p} V\phi_k(t)\,dt = \lambda_k \phi_k(t)$, for $1 \le k \le n$
Components	Random variables in R^P	Random variables in $L_2\left[x_1, x_p\right]$

systematic research of the functional data analysis group named the Toulouse School of Functional Data Analysis [11].

In recent years, Hall and Hosseini-Nasab [12, 13] showed how the properties of functional principal component analysis can be elucidated through stochastic expansions and related results. Yao et al. [14] proposed a FPCA procedure via a conditional expectation method, which is aimed at estimating functional principal component scores for sparse longitudinal data. Hall and Vial [15] have investigated the properties of FPCA and have given some insights into methodology and convergence rates. Di et al. [16] introduced multilevel FPCA, which is designed to extract the intra- and intersubject geometric components of multilevel functional data. Based on FPCA, Hyndman and Shang [17] proposed graphical tools for visualizing functional data and detecting functional outliers.

Due to the theoretical and practical developments, FPCA has been successfully applied to many practical problems, such as the analysis of cornea curvature in the human eye [18], the analysis of electronic commerce [19], the analysis of growth curve [20], the analysis of income density [21], the analysis of implied volatility surface in finance [22], the analysis of longitudinal primary biliary liver cirrhosis [23], and the analysis of spectroscopy data [24]. Furthermore, Hyndman and Shahid Ullah [25] proposed a smoothed and robust FPCA and used it to forecast age-specific mortality and fertility rates.

The objective of this paper is to study the monthly volatility of return of Shanghai 50 index which consists of 50 stocks. Treating stock price series as random function in a space spanned by finite dimensional functional bases, we intensively explore methods of functional data analysis, especially functional principal component analysis.

In the area of finance, some impressive papers with the functional data analysis are found such as Ramsay and Ramsey [26], Muller and Ulrich [27], and Miao [28]. But, few republications are found with research on the increasingly flourishing Chinese financial market. This paper will fill the blank both in theory and in application.

Our study can be described as an exploratory data approach:

$$\text{Data collection} \longrightarrow \text{Data Analysis} \longrightarrow \text{Conclusions.}$$

This paper is organized as follows. In Section 2, we describe the functional principal component analysis (FPCA), which plays a significant role in the development of functional data analysis. It is also an essential ingredient of functional principal component regression (FPCR). Section 3 will illustrate the empirical study with the application of the theory in Section 2. Some further discussion and a conclusion are presented in Section 4.

2. Methodology

As mentioned before, an important tool in the functional data analysis toolbox is FPCA, that is, functional principal component analysis. The main idea of FPCA is just like multivariate principal component analysis (PCA) but its principal component weights or harmonics are functions of time. They carry the main features of the functional data object and can be interpreted separately.

The differences in notation between PCA and FPCA are summarized in Table 1.

The basic assumption of FDA is that data generating process can be described as a smooth function. FPCA finds the set of orthogonal principal component function by maximizing the variance along each component.

The first functional principal component $\phi_1(t)$ is defined by

$$\phi_1(t) = \arg \frac{\max}{\|\Phi\|^2 = 1} \frac{1}{N} \sum_{i=1}^{N} \left(\int_L \phi(t) f_i(t)\,dt \right)^2 \tag{1}$$

subject to

$$\|\phi_1\| = 1. \tag{2}$$

The kth functional principal component $\phi_k(t)$ can be found analogously, subject to the additional constraint

$$\int_L \phi_j(t)\,\phi_k(t)\,dt = 0, \quad \forall j < k. \tag{3}$$

The sample covariance function of $f(x) = [f_1(x), f_2(x), \dots, f_n(x)]$, $x \in [x_1, x_p]$ is given by

$$V(s,t) = \frac{1}{N} \sum_{i=1}^{N} f_i(s)\, f_i(t), \tag{4}$$

where function $f_i(t)$ has usually been first centered.

Covariance operator V extends the concept of a sample covariance matrix to functional data; it is easy to show that V is a positive compact symmetric linear operator. It is obvious that

$$(V\phi_K)(t) = \lambda_k \phi_K(t), \quad \lambda_1 \geq \lambda_2 \geq \cdots \geq 0. \tag{5}$$

Detailed calculation procedure is provided below.

Step 1. The data we need in this paper is collected through some public resources such as WIND database.

Step 2. The data we get may be dirty, so data preprocessing is necessary. Then, the raw data are collected, cleaned, and organized.

Step 3. The data are next converted to functional form. Through this step, the raw data for observation i are used to define a function f_i that can be evaluated at all values of t over interval $[x_1, x_p]$. In order to do this, a basis must be specified. A nonparametric method is used to estimate $f_i(t)$ for $t \in [x_1, x_p]$, $i = 1, \ldots, n$.

Then, we express each function as a linear combination of basic functions and approximate each function by a finite number of basis functions φ_k. Consider

$$f_i(t) \approx \sum_{k=1}^{K} \beta_{i,k}\varphi_k, \quad i = 1, \ldots, n. \tag{6}$$

Some popular basis functions, such as polynomial basis functions, Bernstein polynomial basis functions, Fourier basis functions, and wavelet basis function and B-spline, are used to estimate the functions. B-spline is our first choice because of its goodness of fitting nonperiodic data in our study.

Step 4. The function may also need to be registered or aligned in order to show some important features. Vertical amplitude variation and horizontal variation can be separated by this step. In our study, this step is not used due to our data characteristics.

Step 5. Next, a variety of preliminary displays and summary statistics are developed. For example, first and second derivative curves estimated from these data using techniques discussed before are displayed and we can elude that some curves have larger variation, while other curves are with less impressed variation.

Step 6. Then exploratory analyses such as FPCA can be carried out.

The first principal component can be found by solving

$$V\phi_1(t) = \lambda_1\phi_1(t), \quad \|\phi_1\| = 1. \tag{7}$$

Step 7. The kth functional principal component is a solution of

$$V\phi_k(t) = \lambda_k\phi_k(t), \quad \|\phi_k\| = 1 \tag{8}$$

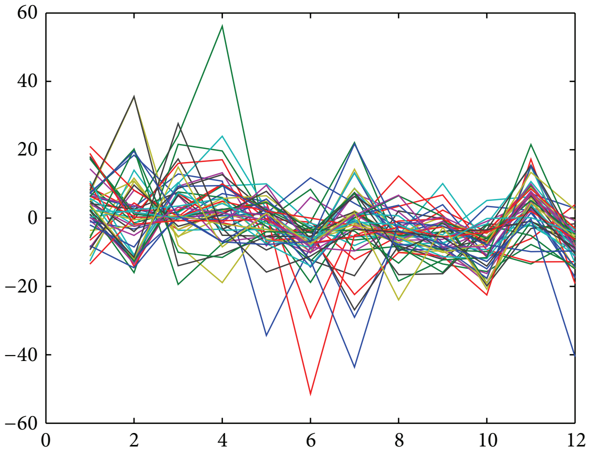


Figure 1: The monthly rate of return of 50 stocks.

subject to

$$\int_L \phi_j(t)\,\phi_k(t) = 0, \quad \forall j < k. \tag{9}$$

Step 8. Accumulative percentage of explained variance is calculated, and some discussion and economic explanation about the functional principal component are provided finally.

3. Application

We now represent the monthly rate of return of 50 stocks in Figure 1, which constitute the SSE50 index.

As we can see, almost nothing can be seen in this form of plot. So, some work must be taken to study the data.

Then, the datasets are converted to functional form, which means functions that can be evaluated at all values of T over some interval. The 50 functions are displayed in Figure 2, with the estimated mean function in bold. There are features in this data too subtle to see in this type of plot.

An impression is that some curves are high (with good investment return) and that some curves are low (with not so good investment return).

We therefore conclude that some of the variation from curve to curve can be explained at the level of certain derivatives. The fact that derivatives are of interest is further reason to think of the records as functions, rather than vectors of observations in discrete time.

Next, we will give the fitted curve of the 50 curves we have got. With the stock code 600019, we can see that the fitted result is pretty good as illustrated in Figure 3.

Figures 4 and 5 display the first and second derivative curves estimated from these data using techniques discussed before. We can elude that some curves have larger variation, while other curves are with less impressed variation.

Now, in Figure 6, we illustrate the variance-covariance structure of return rate. The peak point at the middle of the diagonal represents the largest variance in October.

At last, the principal component functions are represented in Figures 7–11 as perturbations of the mean function

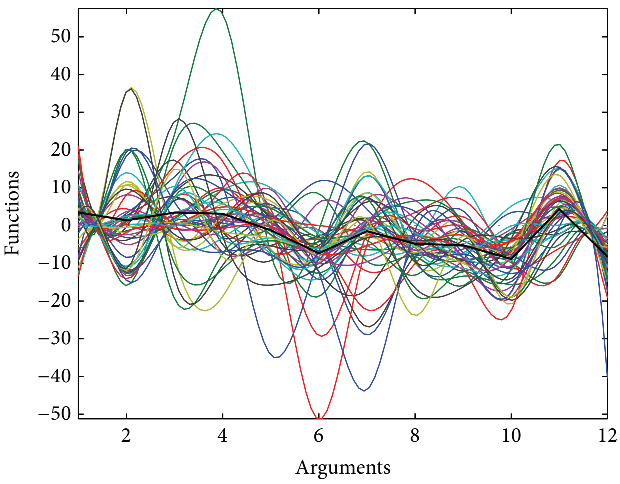


Figure 2: The functional form of 50 stocks. Note: the black bold line is the mean function.

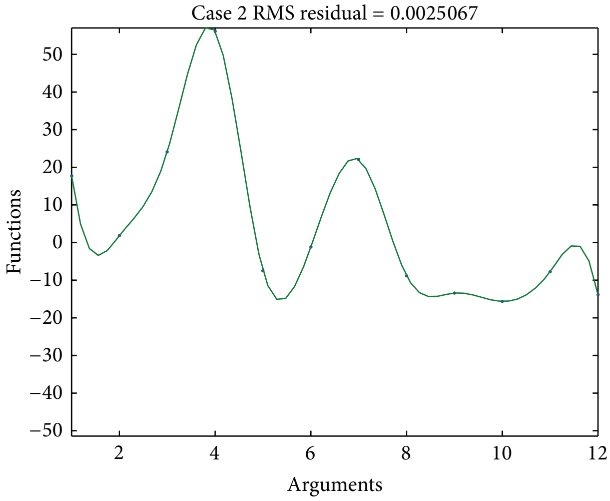


Figure 3: The functional form of the stock code 600019.

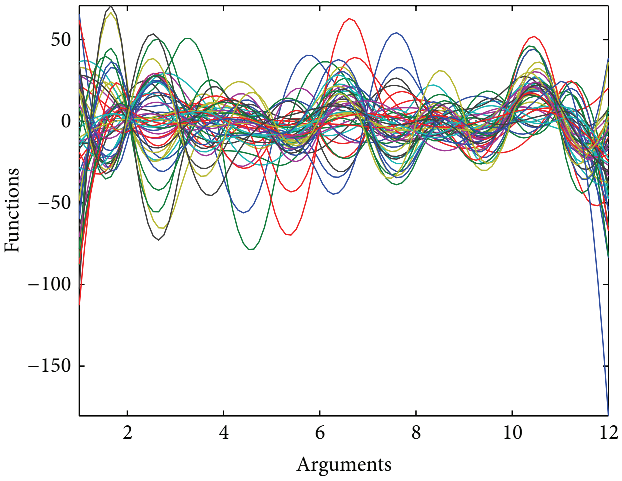


Figure 4: The first derivative curves.

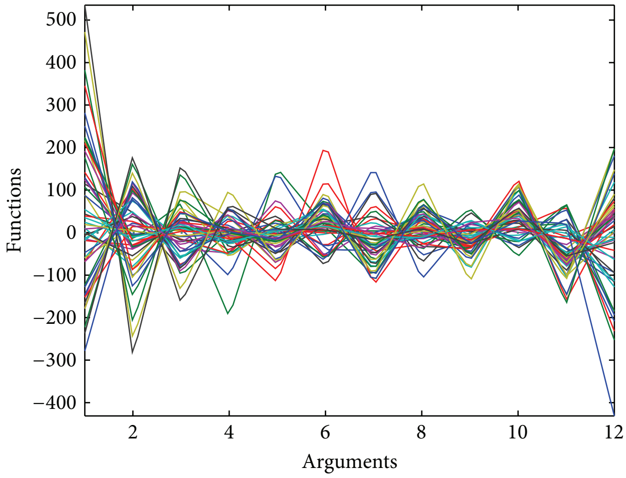


Figure 5: The second derivative curves.

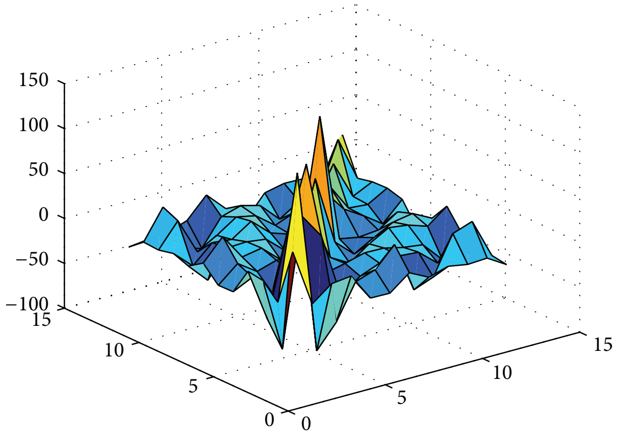


Figure 6: The variance-covariance structure of return rate.

Table 2: Accumulative percentage of explained variance.

PC	The percentage of explained variance
PC1	30.22%
PC2	22.07%
PC3	19.91%
PC4	11.02%
PC5	7.54%
Total	90.76%

by adding and subtracting a multiple of each principal component function.

Table 2 includes accumulative percentage of explained variance.

We can see that the first principal component function (Figure 7), which accounts for 30.22 percent of the variation, has always had an obvious positive effect on the mean function between February and April 2011. In fact, the concept of high-speed rail provoked the Chinese financial market vigorously during that period. Therefore, we can reasonably

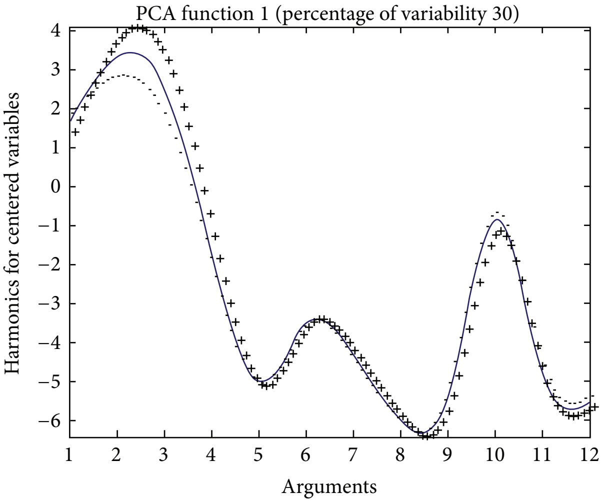


Figure 7: The first principal component function.

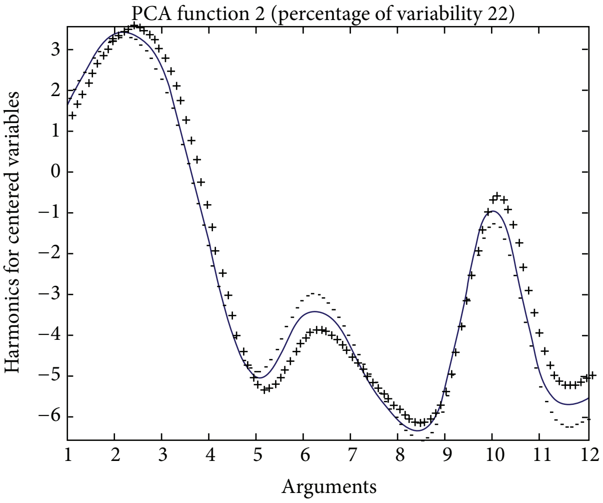


Figure 8: The second principal component function.

believe that the first principal component represents the speculation boom.

The second principal component function (Figure 8), which accounts for 22.07 percent of the variance, is seen to pick up the influence of tighten monetary policy to control the excessive price rises, especially the price of real estate.

With the stock price getting lower and lower, more and more investors believe the current price is worth taking risk, which forms some power of buying driving price briefly rebounded. The third principal component (Figure 9), which accounts for 19.91 percent of the variation, is believed to be the representative of the influence.

With the excess drop in price, a growing number of blue chips are underestimated and the investment value will promote a price return. The effect is summarized to the fourth principal component (Figure 10), which accounts for 11.02 percent of the variation.

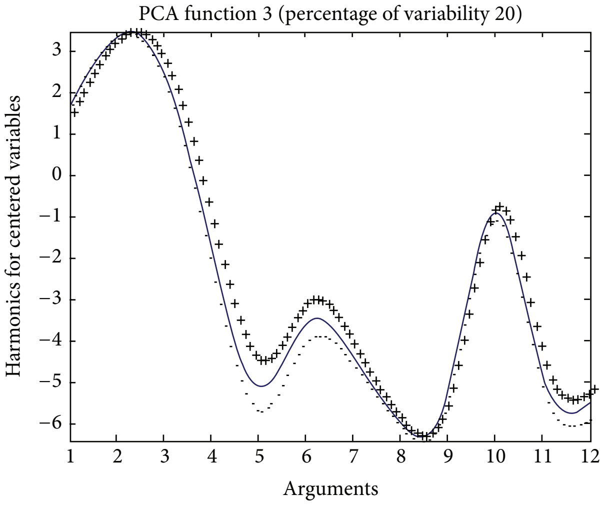


Figure 9: The third principal component function.

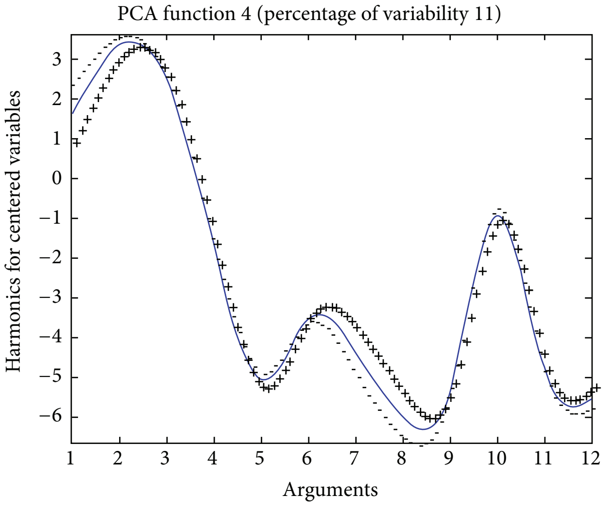


Figure 10: The fourth principal component function.

The fifth principal component (Figure 11), which accounts for only 7.5 percent of the variation, having little effect on the mean function, will not be discussed in this paper.

4. Conclusion and Further Discussion

FPCA attempts to find the dominant modes of variation around an overall trend function and is thus a key technique in functional data analysis. As we described before, modern data analysis has benefited and will continue to benefit greatly from the development of functional data analysis. In this paper, we mainly illustrate the functional principal components analysis by the research on monthly return rate of stocks constituting Shanghai 50 index. We extracted the main variance factors over time by extracting principal component regarding the samples as random functions, which has strong theoretical and practical value. A functional feature of the proposed approach that distinguishes it from

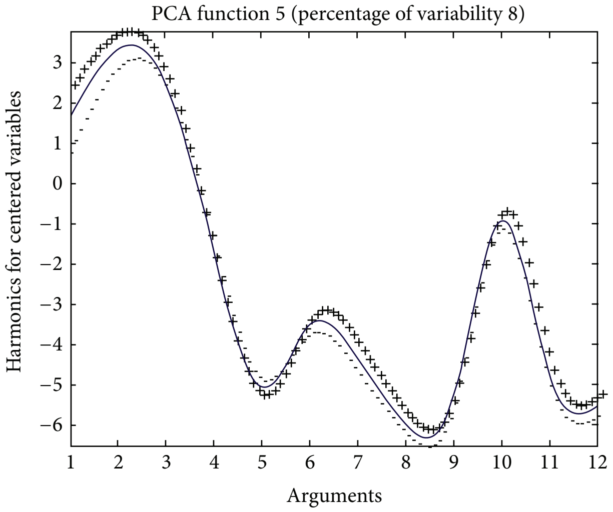


Figure 11: The fifth principal component function.

established methods for spot volatility analysis is that it is geared towards the analysis of observations drawn from all realizations of the volatility process, rather than observations from a single realization. As we described before, the first principal component function has always had an obvious positive effect on the mean function and thus could be summarized as the representation of the speculation boom. The second principal component function, on the other hand, having outstanding negative effect on the mean function between May to July 2011, is seen to pick up the influence of fiscal austerity to curb the fast rising prices.

In addition, the proposed FPCA method is easy to program and implement. By smoothing the underlying functions or curves, the principal components we need are extracted easily. The fast computational speed of our method makes it feasible to be applied in empirical studies with a large number of observations.

Besides the proposed method in this paper, several other methods, such as curves classification, nonparametric analysis, and functional depth analysis, can be utilized to analyze functional data. These methods will be considered in the next study. Moreover, in order to emphasize the interest of doing the functional approach and compare the corresponding results, we will treat the curves as high dimensional standard vectors in the future work.

Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

References

[1] P. Hall, H. Müller, and J. Wang, "Properties of principal component methods for functional and longitudinal data analysis," *Annals of Statistics*, vol. 34, no. 3, pp. 1493–1517, 2006.

[2] J. O. Ramsay and C. J. Dalzell, "Some tools for functional data analysis (with discussion)," *Journal of the Royal Statistical Society B*, vol. 53, no. 3, pp. 539–572, 1991.

[3] J. O. Ramsay and B. W. Silverman, *Applied Functional Data Analysis: Methods and Case Studies*, Springer, New York, NY, USA, 2002.

[4] J. O. Ramsay and B. W. Silverman, *Functional Data Analysis*, Springer, New York, NY, USA, 2nd edition, 2005.

[5] J. O. Ramsay, G. Hooker, and S. Graves, *Functional Data Analysis with R and MATLAB*, Springer, New York, NY, USA, 2009.

[6] H. L. Shang, *A Survey of Functional Principal Component Analysis (Working Paper06/11)*, Department of Econometrics and Business Statistics, Monash University, 2011.

[7] K. Karhunen, "Zur Spektraltheorie stochastischer Prozesse," vol. 1946, no. 34, 7 pages, 1946.

[8] M. Loève, "Fonctions aleatoires a decomposition orthogonal exponentielle," *La Revue Scientique*, vol. 84, pp. 159–162, 1946.

[9] C. R. Rao, "Some statistical methods for comparison of growth curves," *Biometrics*, vol. 14, no. 1, pp. 1–17, 1958.

[10] J. Dauxois and A. Pousse, *Les analyses factorielles en calcul des probabilities et en statistique: essai d'etude synthetique [Ph.D. thesis]*, l'Universite Paul-Sabatier de Toulouse, Toulouse, France, 1976.

[11] J. Dauxois, A. Pousse, and Y. Romain, "Asymptotic theory for the principal component analysis of a vector random function: some applications to statistical inference," *Journal of Multivariate Analysis*, vol. 12, no. 1, pp. 136–154, 1982.

[12] P. Hall and M. Hosseini-Nasab, "On properties of functional principal components analysis," *Journal of the Royal Statistical Society B: Statistical Methodology*, vol. 68, no. 1, pp. 109–126, 2006.

[13] P. Hall and M. Hosseini-Nasab, "Theory for high-order bounds in functional principal components analysis," *Mathematical Proceedings of the Cambridge Philosophical Society*, vol. 146, no. 1, pp. 225–256, 2009.

[14] F. Yao, H. Müller, and J. Wang, "Functional data analysis for sparse longitudinal data," *Journal of the American Statistical Association*, vol. 100, no. 470, pp. 577–590, 2005.

[15] P. Hall and C. Vial, "Assessing the finite dimensionality of functional data," *Journal of the Royal Statistical Society B*, vol. 68, no. 4, pp. 689–705, 2006.

[16] C. Di, C. M. Crainiceanu, B. S. Caffo, and N. M. Punjabi, "Multilevel functional principal component analysis," *The Annals of Applied Statistics*, vol. 3, no. 1, pp. 458–488, 2009.

[17] R. J. Hyndman and H. L. Shang, "Rainbow plots, bagplots, and boxplots for functional data," *Journal of Computational and Graphical Statistics*, vol. 19, no. 1, pp. 29–45, 2010.

[18] N. Locantore, J. S. Marron, D. G. Simpson, N. Tripoli, J. T. Zhang, and K. L. Cohen, "Robust principal component analysis for functional data," *Test*, vol. 8, no. 1, pp. 1–73, 1999.

[19] S. Wang, W. Jank, and G. Shmueli, "Explaining and forecasting online auction prices and their dynamics using functional data analysis," *Journal of Business & Economic Statistics*, vol. 26, no. 2, pp. 144–160, 2008.

[20] J. M. Chiou and P. L. Li, "Functional clustering and identifying substructures of longitudinal data," *Journal of the Royal Statistical Society B: Statistical Methodology*, vol. 69, no. 4, pp. 679–699, 2007.

[21] A. Kneip and K. J. Utikal, "Inference for density families using functional principal component analysis," *Journal of the*

American Statistical Association, vol. 96, no. 454, pp. 519–532, 2001.

[22] F. Cont and J. Fonseca, "The dynamics of implied volatility surfaces," *Quantitative Finance*, vol. 2, no. 1, pp. 45–60, 2002.

[23] F. Yao, H. G. Muller, and J. L. Wang, "Functional linear regression analysis for longitudinal data," *Annals of Statistics*, vol. 33, no. 6, pp. 2873–2903, 2005.

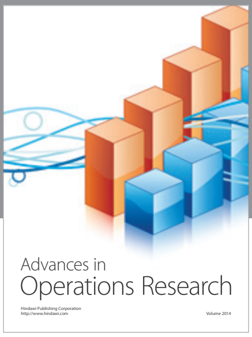
[24] F. Yao and H. Müller, "Functional quadratic regression," *Biometrika*, vol. 97, no. 1, pp. 49–64, 2010.

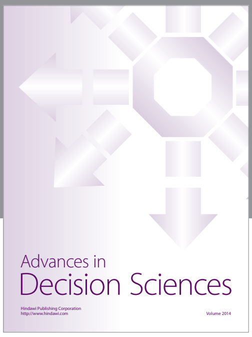
[25] R. J. Hyndman and M. Shahid Ullah, "Robust forecasting of mortality and fertility rates: a functional data approach," *Computational Statistics and Data Analysis*, vol. 51, no. 10, pp. 4942–4956, 2007.

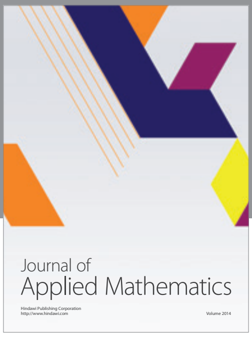
[26] J. O. Ramsay and J. B. Ramsey, "Functional data analysis of the dynamics of the monthly index of nondurable goods production," *Journal of Econometrics*, vol. 107, no. 1-2, pp. 327–344, 2002.

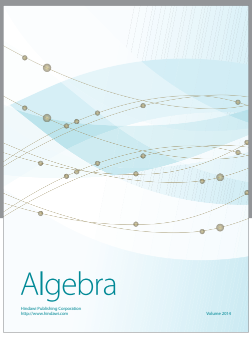
[27] H. G. Muller and M. S. Ulrich, "Functional data analysis for volatility," Journal of Economics Literature Classification Codes: C14, C51, C52, G12, G17, 2011.

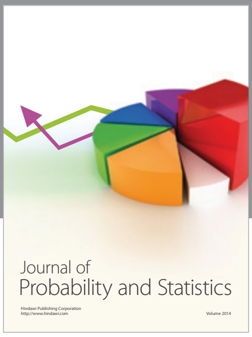
[28] H. Miao, "Potential applications of function data analysis in high-frequency financial research," *Journal of Business & Financial Affairs*, vol. 2, no. 1, article e125, 2013.


Advances in
Operations Research

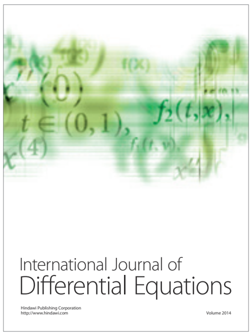

Advances in
Decision Sciences


Journal of
Applied Mathematics

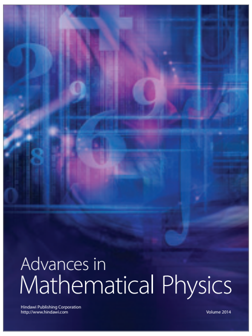

Algebra

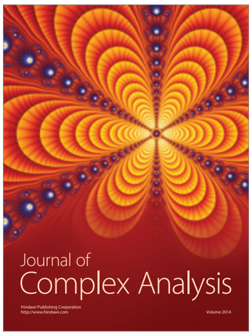

Journal of
Probability and Statistics

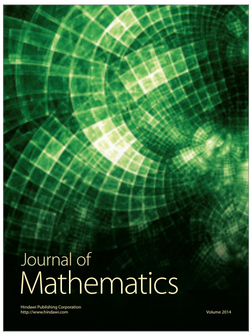
The Scientific
World Journal

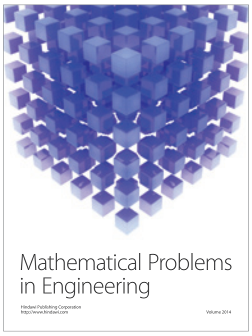

International Journal of
Differential Equations

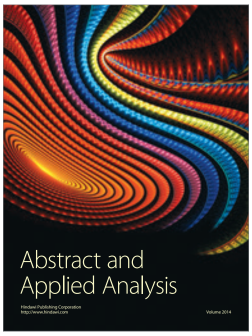
International Journal of
Combinatorics

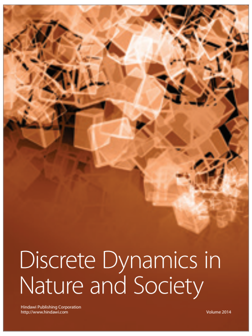

Advances in
Mathematical Physics

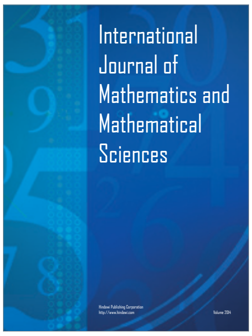

Journal of
Complex Analysis


Journal of
Mathematics


Mathematical Problems
in Engineering

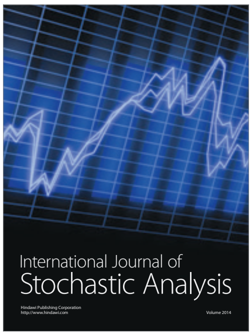

Abstract and
Applied Analysis


Discrete Dynamics in
Nature and Society


International
Journal of
Mathematics and
Mathematical
Sciences

Journal of
Discrete Mathematics

Journal of
Function Spaces


International Journal of
Stochastic Analysis

Journal of
Optimization