Advances in Applied Mathematics 应用数学进展, 2016, 5(2), 291-297
Published Online May 2016 in Hans. http://www.hanspub.org/journal/aam
http://dx.doi.org/10.12677/aam.2016.52037

Hans汉斯

Prediction for the Shanghai Stock Index Based on the Functional Data

Lijuan Cheng

School of Mathematics and Computation Science, Lingnan Normal University, Zhanjiang Guangdong
Email: cheng_lijuan@163.com

Received: May 6th, 2016; accepted: May 27th, 2016; published: May 30th, 2016

Abstract

In the research of financial data, the functional data are often encountered. In this paper, the prediction model of functional principal components analysis is established to forecast the Shanghai Stock Index. Based on the principal component analysis theory and calculation method, the Shanghai Composite Index is forecasted by Matlab.

Keywords

Functional Data, Principal Component Analysis, Forecast

基于函数型数据的上证指数预测

程丽娟

岭南师范学院数学与计算科学学院，广东 湛江
Email: cheng_lijuan@163.com

收稿日期：2016年5月6日；录用日期：2016年5月27日；发布日期：2016年5月30日

摘 要

在金融数据的研究中，经常遇到函数型数据。主要建立函数型主成分分析的预测模型，分析函数型数据在上证指数预测中的应用，根据函数型数据分析的原理及其求解主成分分析的方法，使用Matlab对上证

文章引用: 程丽娟. 基于函数型数据的上证指数预测[J]. 应用数学进展, 2016, 5(2): 291-297.
http://dx.doi.org/10.12677/aam.2016.52037

指数进行预测。

关键词

函数型数据，主成分分析，预测

1. 引言

传统的数据分析中，获得的数据包括截面数据、时间序列数据以及面板数据，但是对这三种数据分析时要依赖许多的假设条件，适用数据的类型具有一定的局限性，加拿大学者 J.O. Ramsay 在 1982 年首次给出将泛函分析、拓扑学和统计学相结合的设想，提出“函数型数据”的概念以及函数型数据的分析方法[1]。近年来，函数型数据分析广泛应用在医学、气象学、生物学等领域。金融数据间隔较短，可以视为连续数据，即函数型数据，因而可以进行函数型数据分析[2]。

本文主要建立函数型主成分分析的预测模型，分析函数型数据在上证指数预测中的应用，根据函数型数据分析的原理及其求解主成分分析的方法，使用 Matlab 软件对上证指数进行预测。

2. 函数型数据分析

对于原始数据序列 $y=\left(y_1,y_2,\cdots,y_n\right)$，建立回归模型

$$y_j=x\left(t_j\right)+\varepsilon_j,\ j=1,2,\cdots,n \tag{1}$$

其中，$x\left(t_j\right)$ 为原始数据序列的函数在 t_j 时的取值，ε_j 为噪声，表示观测数据中的扰动因素、误差或其它外生因素[3]。在标准的统计模型中，通常假定 ε_j 服从零均值有限方差的独立分布，且通常假定方差相等。但在处理函数型数据时，可以放松这些假设条件，使模型更加符合现实中的问题。在函数型数据分析中，首先要考虑到原始数据的选取问题，而原始数据的选取要通过抽样率来进行。原始数据在时间轴上的曲率不同，通常曲率值大的地方，相应的数据分布比较密，因此在曲率大的地方应该选取较多的数据点，曲率可以用 $\left|D^2x\left(t\right)\right|$ 表示[4]。

2.1. 数据平滑

最简单的平滑方法为线性平滑，线性平滑法是指用离散观测值的线性组合去估计函数 $\hat{x}\left(t\right)$，线性平滑可以衍生出平滑算子的许多性质，并且其运算速度较快，但是有时非线性平滑方法在不同的数据和同一个数据的不同部分，结果会更优良，因而本文考虑基函数平滑法。基函数平滑法是将离散的观测数据转化为相应函数的平滑方法，基函数能够保留待估函数的性质。它是用 K 个已知的基函数 $\phi_k\left(t\right)\left(k=1,2,\cdots,K\right)$ 的线性组合得到函数 $x\left(t\right)$ 的估计 $\hat{x}\left(t\right)$，即

$$\hat{x}\left(t\right)=\sum_{k=1}^{K}c_k\phi_k\left(t\right) \tag{2}$$

对函数 $x\left(t\right)$ 的估计包括两部分，第一步是选取用来拟合的基函数，第二步是估计每个基函数前面的系数。基函数的个数 K 决定了观测数据 y_j 的平滑度，K 越小，拟合的函数越平滑，但拟合程度就越差。

采用基函数平滑方法时，常选用的基函数包括傅立叶基，B-Spline 基，多项式基，常数基等。对于周期性的数据常采用傅里叶基，对于非周期性的数据通常用 B-Spline 基进行拟合[5]。在基函数 $\phi_k\left(t\right)$ 确定的情况下，系数向量 $c=\left(c_1,c_2,\cdots,c_K\right)'$ 的一组取值唯一地确定一个函数。估计系数 c_k 的最简单方法是最

小二乘法，即最小化残差平方和：

$$SMSSE\left(y|c\right)=\sum_{j=1}^{n}\left[y_j-\sum_{k=1}^{K}c_k\phi_k\left(t_j\right)\right]^2 \tag{3}$$

解得 $c=\left(\Phi'\Phi\right)^{-1}\Phi'y$。

从观测值 $y_j=x\left(t_j\right)+\varepsilon_j$ 来估计曲线 x，则曲线估计的两个目标：一方面，估计的曲线和数据较好地拟合，即使残差平方和最小。另一方面，如果得到的曲线波动很大或局部变化过大的话不希望拟合的太好，即放弃一定程度的拟合度，达到较好的平滑度。因此，我们定义带惩罚的残差平方和为：

$$PENSSR_\lambda\left(x|y\right)=\sum_j\left\{y_j-x\left(t_j\right)\right\}^2+\lambda\times PEN_2\left(x\right) \tag{4}$$

其中，$PEN_2\left(x\right)=\int\left\{D^2x\left(s\right)\right\}^2ds=\left\|D^2x\right\|^2$，即为惩罚项，$\lambda$ 称为平滑参数。通过最小化 $PENSSR_\lambda\left(x|y\right)$，以获得估计函数。

上式中的第一项度量了曲线的拟合效果，第二项度量了曲线的平滑度。正的常数 λ 决定了拟合度与平滑度之间的平衡，当 λ 较小时，惩罚项对 $PENSSR_\lambda\left(x|y\right)$ 影响较小；当 λ 较大时，在 $PENSSR_\lambda\left(x|y\right)$ 上产生较大的粗糙惩罚。λ 的值由广义交叉验证法确定。

2.2. 函数型主成分分析

函数型数据主成分分析是将变量看作函数的形式，其样本的协方差矩阵也变为函数的形式，因此可以避免出现高维的协方差矩阵。在函数数据下，观测矩阵为 $\left\{x\left(t\right)=\left(x_1\left(t\right),x_2\left(t\right),\cdots,x_N\left(t\right)\right)^T\middle|t\in\mathrm{T}=\left\{1,\cdots,T\right\}\right\}$，它表示对同一个变量进行了 N 次重复观测，每次所得到的观测数据构成一个函数 $x_i\left(t\right)$，此时，函数型数据的主成分的线性组合为：

$$f_i=\int\xi(t)x_i\left(t\right)\mathrm{d}t=\left\langle\xi,x_i\right\rangle,i=1,\cdots,N \tag{5}$$

其中下标 i 表示第 i 次重复观测，$\xi\left(t\right)$ 为权重函数，即函数型主成分的特征函数。

$x\left(s\right)$ 和 $x\left(t\right)$ 的协方差为 $v\left(s,t\right)$，则有 $v\left(s,t\right)=\frac{1}{N-1}\sum_{i=1}^{N}x_i\left(s\right)x_i\left(t\right)$。

函数型数据的第一主成分 f_1 的求解为求下面有约束条件的最大化问题：

$$\begin{cases}\max\frac{1}{N-1}\sum_{i=1}^{N}\left(\int\xi\left(t\right)x_i\left(t\right)\mathrm{d}t\right)^2\\ s.t.\int\left(\xi\left(t\right)\right)^2\mathrm{d}t=\left\|\xi\right\|^2=1\end{cases} \tag{6}$$

同样的，第 k 个主成分 f_k 的求解就是在上式的最大化问题中添加约束条件 $\int\xi_k\xi_{k-m}=0\left(\text{其中}m=1,\cdots,k-1\right)$ 求得[6]。

函数型主成分的权重函数 $\xi\left(s\right)$ 满足特征方程 $\int v\left(s,t\right)\xi\left(t\right)\mathrm{d}t=\lambda\xi\left(s\right)$，其中 λ 为特征值。

记 $V\xi\left(s\right)=\int v\left(s,t\right)\xi\left(t\right)\mathrm{d}t$，则 $V\xi\left(s\right)=\lambda\xi\left(s\right)$。

函数型主成分分析中数据 $x\left(t\right)$ 是函数的形式，样本的个数 N 决定了其协方差算子 V 的秩为 $N-1$。因此其非零特征值的最大个数为 $N-1$，进而满足约束条件的主成分的最大个数为 $N-1$。在实际观测中，我们通常选取样本个数 N 与观测点的个数 T，至少应该满足 $N<T/2$。

函数型主成分的选取思想与多元主成分的选取相同，根据所研究问题的需要确定累积贡献率，选择合适的 K 使得 $\sum_{k=1}^{K}\lambda_k\Big/\sum_{k=1}^{N-1}\lambda_k$ 达到所确定的累积贡献率，一般要求累积贡献率不小于 85%。

2.3. 函数型主成分预测模型

设随机过程$\{X(t):t\in[T_1,T_2]\}$均方连续，且二次可积。设随机变量$\tilde{X}(t)$是$X(t)$的线性均方估计，则有$E\left[\left|\tilde{X}(t)-X(t)\right|^2\right]=\inf\left\{E\left[\left|Z-X(t)\right|^2\right]:Z\in L_X^2\right\}$，其中$L_X^2$是希尔伯特空间$L^2(\Omega)$的一个希尔伯特子空间。$\tilde{X}(t)$是$X(t)$在子空间$L_X^2$上的正交投影，即$\left(\tilde{X}(t)-X(t)\right)\perp L_X^2$，从而有

$$E\left[\left(X(s)-\bar{X}(s)\right)\left(X(t)-\bar{X}(t)\right)\right]=E\left[\left(\tilde{X}(s)-\bar{X}(s)\right)\left(\tilde{X}(t)-\bar{X}(t)\right)\right],\forall t\in[T_1,T_2] \tag{7}$$

设$\varepsilon^2(s)$为$X(s)$在$s\in[T_3,T_4]$的线性预测的均方误差，表示为$\varepsilon^2(s)=E\left[\left|\tilde{X}(t)-X(t)\right|^2\right]$。

分别提取两个区间的主成分，随机过程$\{X(t)\}$在$[T_1,T_2]$中第i个主成分定义为：

$$\xi_i=\int_{T_1}^{T_2}\left(X(t)-\mu(t)\right)f_i(t)\mathrm{d}t \tag{8}$$

其中，f_i称为主因子，它是对应于协方差函数$C(t,s)$的第i大特征值λ的正则化的特征函数，μ是$\{X(t)\}$的均值函数[7]。

第i个主成分解释的方差(即方差的累积贡献率)为$V_i=\lambda_i/V$，其中V是$\{X(t)\}$在区间$[T_1,T_2]$中的总方差，表示的样本数据波动的程度，即$V=E\left[\int_{T_1}^{T_2}\left(X(t)-\mu(t)\right)^2\mathrm{d}t\right]=\sum_i\lambda_i$。

随机过程$\{X(t)\}$可以 Karhunen-Loève 正交展开，则有$X(t)=\mu(t)+\sum_{i=1}^{\infty}\xi_i f_i(t),t\in[T_1,T_2]$，同样，设$g_j$和$\eta_j$分别表示在区间$[T_3,T_4]$上的特征函数和主成分，则随机过程$\{X(s):s\in[T_3,T_4]\}$用 Karhunen-Loève 正交展开表示为$X(s)=\mu(s)+\sum_{j=1}^{\infty}\eta_j g_j(s),s\in[T_3,T_4]$，若区间$[T_3,T_4]$表示的是未来的一个时间区间，则$X(s)$的线性最小二乘估计$\tilde{X}(s)$为$\tilde{X}(s)=\mu(s)+\sum_{j=1}^{\infty}\tilde{\eta}_j g_j(s),s\in[T_3,T_4]$[8]，其中，$\tilde{\eta}_j$为随机变量$\{X(t):t\in[T_1,T_2]\}$的主成分$\xi_i(i=1,2,\cdots)$的线性最小二乘估计，表示为

$$\tilde{\eta}_j=\sum_{i=1}^{\infty}\frac{E\left[\eta_j\xi_i\right]}{\lambda_i}\xi_i=\sum_{i=1}^{\infty}\beta_i^j\xi_i,j=1,2,\cdots \tag{9}$$

对上式中的$\tilde{\eta}_j$序列，每个序列我们取前p_j个主成分拟合，得到最小二乘估计量$\tilde{\eta}_j^{p_j}$为$\tilde{\eta}_j^{p_j}=\sum_{i=1}^{p_j}\beta_i^j\xi_i$，因此，主成分预测模型$PCP\left(q;p_1,p_2,\cdots,p_q\right)$为

$$\tilde{X}^q(s)=\mu(s)+\sum_{j=1}^{q}\tilde{\eta}_j^{p_j}g_j(s)=\bar{X}(s)+\sum_{j=1}^{q}\sum_{i=1}^{p_j}\beta_i^j\xi_i g_j(s),s\in[T_3,T_4] \tag{10}$$

3. 实证分析

本文选取我国上证指数作为研究对象，选取 2015 年 4 月 7 日至 6 月 8 日的 5 分钟收益率数据。为了避免“隔夜效应”的影响，在进行数据分析时，舍弃每天的第一个数据(即 9:35 时刻)，把每天的 5 分钟收益率数据视为其对应函数产生的一组样本观测值，使用 B 样条插值法对这些样本观测值进行拟合，并进行平滑处理，然后根据得到的光滑曲线求出其变化速度曲线，即一阶导数曲线，如下图所示。图 1 表示的是收益率曲线及其平滑曲线，其中绿线表示的原数据，蓝线表示的是 B 样条插值法得到的曲线，红

线表示对 B 样条插值法得到的曲线平滑处理后得到的新曲线。图 2 表示的是 B 样条插值法得到曲线的一阶导数曲线。

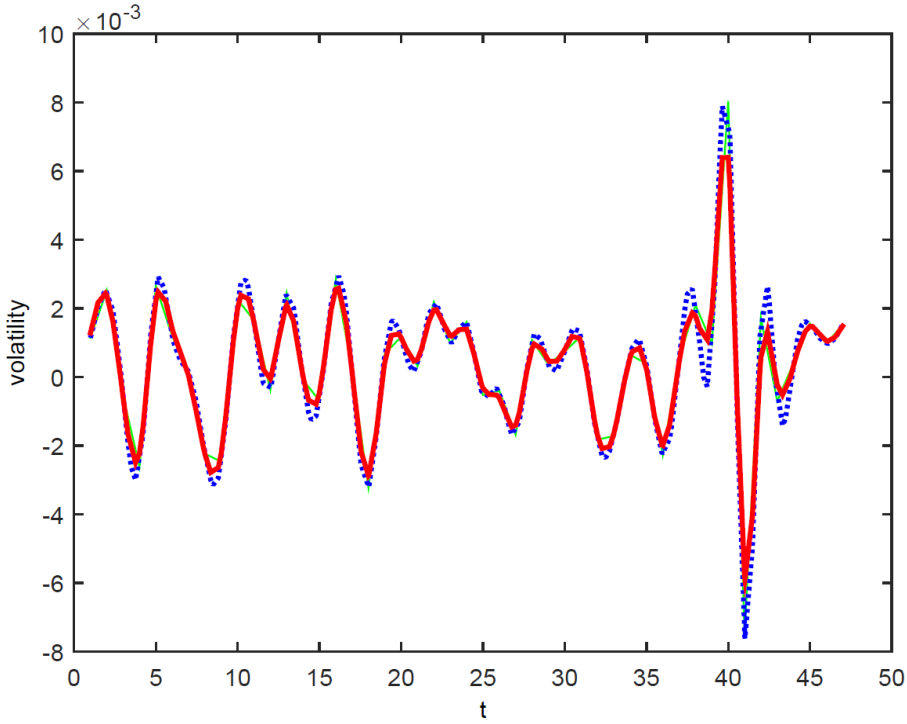

Figure 1. Yield curve and smooth curve
图 1. 收益率曲线及其平滑曲线

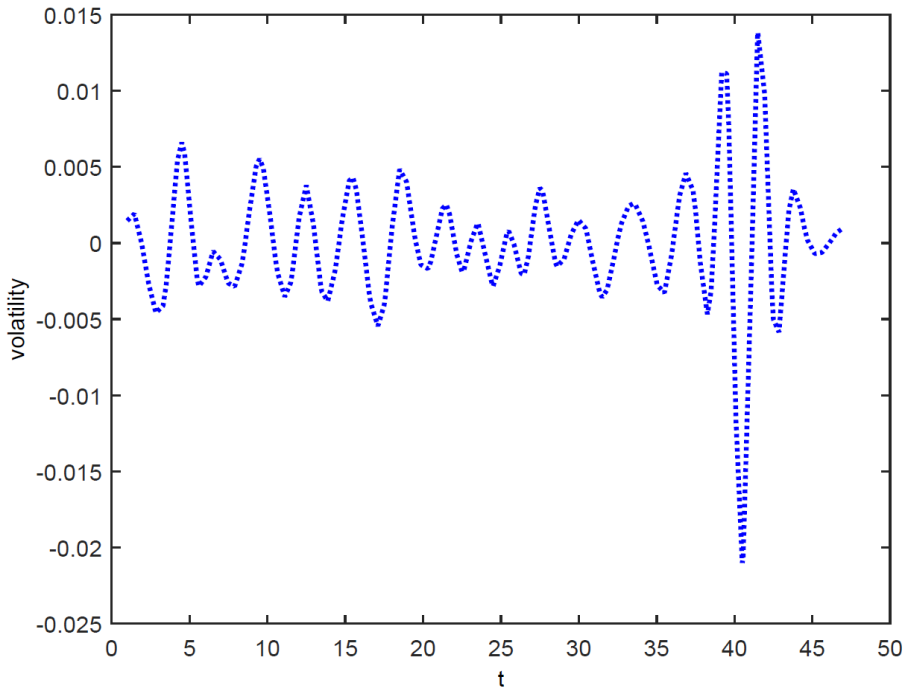

Figure 2. First derivative curve
图 2. 一阶导数曲线

得到平滑函数后，为分析收益率在时间上的差异，将收益率分为两个区间，其中前 30 天为第一区间，后 14 天为第二区间，分别对两个区间的收益率函数进行函数型主成分分析，如下图所示。图 3 表示第一区间的前两个主成分，其中 PC1 对应的方差贡献率是 52.9%，PC2 对应的方差贡献率是 47.1%，图 4 表示第二区间的前两个主成分，其中 PC1 对应的方差贡献率是 58.2%，PC2 对应的方差贡献率是 27.3%。

根据表 1 中的方差贡献率，在第一区间选取全部 2 个主成分。在第二个区间里，第一、二主成分的方差贡献率已经达到了 85.5%，因此选取前两个主成分表示第二区间内收益率的变化特征。

观察表 2 中的相关系数，$\hat{\eta}_1$ 与 $\hat{\xi}_1$ 的相关系数为 1.0000，$\hat{\eta}_2$ 与 $\hat{\xi}_2$ 的相关系数为 1.0000，远远大于其它的相关，因此，选取 $\hat{\xi}_1$ 对 $\hat{\eta}_1$ 做最小二乘估计，选取 $\hat{\xi}_2$ 对 $\hat{\eta}_2$ 做最小二乘估计，由此得到 $PCP\left(q;p_1,p_2,\cdots,p_q\right)$ 模型为

$$PCP\left(2;2\right):\quad \tilde{X}^2\left(s\right)=\bar{X}\left(s\right)+\tilde{\eta}_1^2 g_1\left(s\right)+\tilde{\eta}_2^2 g_2\left(s\right)$$

$$\tilde{\eta}_1^2=-0.0259\xi_1+0.0949$$

$$\tilde{\eta}_2^2=-0.4318\xi_2-0.0069$$

将观测到的真实值、模型估计出的预测值及由两者得到的误差百分比如下表所示。

由表 3 可以看出，除 6 月 12 日误差百分比较大以外，绝大部分的误差百分比都较小。因此说明建立的 $PCP\left(2;2\right)$ 模型具有很好的准确性。综上知，建立函数型主成分预测模型对我国上证指数收益率进行预测，预测的准确程度较高。

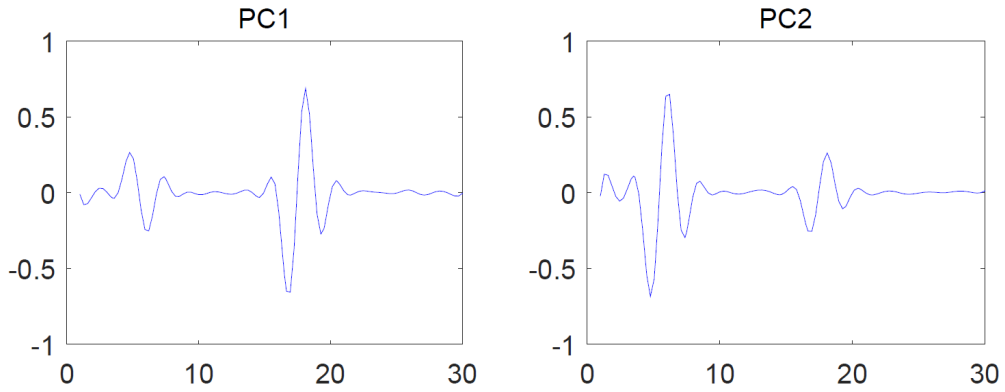

Figure 3. Principal component weighting function (the first interval)
图 3. 主成分权重函数(第一区间)

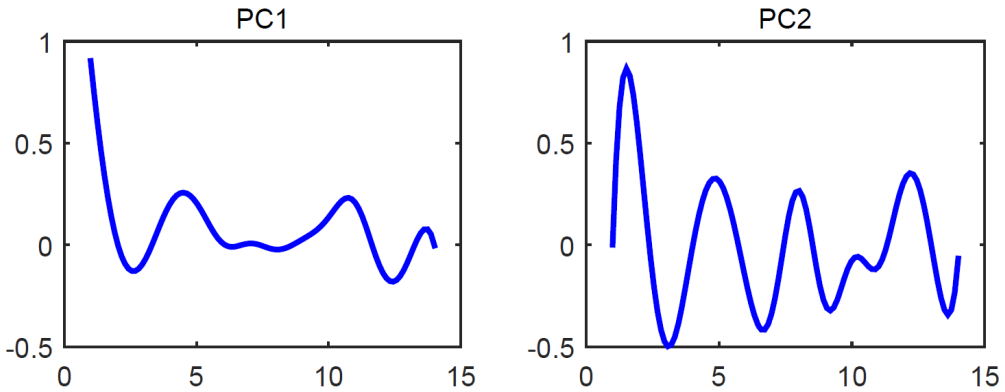

Figure 4. Principal component weighting function (second interval)
图 4. 主成分权重函数(第二区间)

Table 1. Principal component characteristic value and contribution rate
表 1. 主成分特征值与贡献率

主成分	前 30 天			后 14 天		
	特征值	贡献率	累计贡献率	特征值	贡献率	累计贡献率
1	0.0099	52.9%	52.9%	0.0056	58.2%	58.2%
2	0.0088	47.1%	100%	0.0026	27.3%	85.5%

Table 2. Correlation coefficient
表 2. 相关系数

	$\hat{\eta}_1$	$\hat{\eta}_2$
$\hat{\xi}_1$	1.0000	-0.0655
$\hat{\xi}_2$	-0.0655	1.0000

Table 3. Real value and predictive value
表 3. 真实值与预测值

	6 月 9 日	6 月 10 日	6 月 11 日	6 月 12 日
真实值	−0.00357	−0.00146	0.00304	0.00874
预测值	−0.00338	−0.00139	0.00309	0.00793
误差百分比	5.32%	4.79%	−4.93%	9.27%

4. 结论

本文对函数型数据分析方法和函数型数据分析方法在上证指数中的应用进行了研究，介绍了函数型数据分析的研究意义，以及函数型数据的均值函数、方差函数、协方差函数、相关系数函数等描述统计量。研究了函数型数据的主成分分析方法以及如何建立函数型主成分预测模型。通过对上证指数收益率进行预测的实证分析，得到建立函数型主成分预测模型对我国上证指数收益率进行预测，预测的准确程度较高。

基金项目

岭南师范学院自然科学青年项目《基于函数型数据的统计分析及应用》(QL1407)。

参考文献 (References)

[1] Ramsay, J.O. (1982) When the Data Are Functions. *Psychometrika*, **47**, 379-396. http://dx.doi.org/10.1007/BF02293704

[2] Delicado, P. (2011) Dimensionality Reduction When Data Are Density Functions. *Computational Statistics and Data Analysis*, **55**, 401-420. http://dx.doi.org/10.1016/j.csda.2010.05.008

[3] 严明义. 季节数据分析: 一种基于数据的函数性视角的分析方法[J]. 当代经济科学, 2007(1): 108-113.

[4] 严明义, 等. 中国消费价格指数季节被动的函数性数据分析[J]. 统计与信息论坛, 2010(8): 100-106.

[5] Mallor, F., Leon, T. and Gaston, M. (2010) Changes in Power Curve Shapes as an Indicator of Fatigue during Dynamic Contractions. *Journal of Biomechanics*, **43**, 1627-1631. http://dx.doi.org/10.1016/j.jbiomech.2010.01.038

[6] Ramsay, J.O. and Hooker, G. (2009) Functional Data Analysis with R and MATLAB. Springer, New York. http://dx.doi.org/10.1007/978-0-387-98185-7

[7] Shang, H.L. (2010) Nonparametric Modeling and Forecasting Electricity Demand: An Empirical Study. Working Paper.

[8] Berrendero, J.R. (2011) Principal Components for Multivariate Functional Data. *Computational Statistics and Data Analysis*, **55**, 2619-2634. http://dx.doi.org/10.1016/j.csda.2011.03.011