基于经验相关矩阵的区间主成分分析①

郭均鹏，李汶华
（天津大学管理学院，天津 300072）

摘要：给出了针对区间数据样本的主成分分析方法. 为此，首先研究了区间数据样本的经验描述统计量，其中包括单变量的均值与方差、双变量的协方差与相关系数. 然后，基于经验相关矩阵，给出了区间主成分分析的算法，该算法最终得到区间数表达形式的主成分取值. 最后给出算例，分析表明文中方法实施简单，克服了区间主成分分析现有方法的缺点.

关键词：主成分分析；区间数；相关矩阵
中图分类号：O242.29 **文献标识码**：A **文章编号**：1007－9807(2008)03－0049－04

0 概 述

主成分分析(principal component analysis，简称 PCA)是把原来多个变量划为少数几个综合指标的统计分析方法，从数学角度来看是降维处理技术. 传统的 PCA 方法是针对实数样本的，即每一个样本数据用“点”数来表示. 而在实际中由于观测误差或信息不完备等原因，经常会遇到用区间数表示的样本观测值. 针对区间数据的主成分分析方法称为区间主成分分析.

一些文献对区间主成分分析方法进行了研究[1~5]. 其中最重要的两种方法是顶点法(V-PCA)[1,2]和中点法(C-PCA)[1]. 顶点法从几何意义上将每一个区间数据样本看成一个具有 2^p 个(p 为变量个数) 顶点的超矩形(hyper-rectangle)，并将原始的数据矩阵化为一个$(n \cdot 2^p) \times p$ 维(n 为样本个数) 的实数矩阵，然后对其实施传统的 PCA. 显然该方法存在着在随着 p 的变大而计算量指数增长的缺陷[3]. 而中点法则简单地将每一个区间数用其中点来代替，虽然计算量少，但却丢失了数据信息.

本文接下来将首先给出一些文中将要用到的概念和理论. 然后讨论区间样本数据的描述统计量的计算，其中包括单变量的均值与方差，以及双变量的协方差与相关系数. 基于此，讨论针对区间数据的主成分分析方法. 最后给出算例.

1 基本概念与理论

1.1 区间数及其运算

$X = [\underline{x}, \bar{x}] = \{x : \underline{x} \leqslant x \leqslant \bar{x}, x \in \Re\}$ 称为一个区间数，$\underline{x}$ 和 $\bar{x}$ 分别称为区间数的下限和上限. 当 $\underline{x} = \bar{x}$ 时，区间数退化为一个实数. 以区间数为分量的向量称为区间向量.

X、Y 为任两个区间数，$* \in \{+, -, \times, /\}$ 是一个二元运算符，则两个区间数的运算定义为 $X * Y = \{x * y \mid x \in X, y \in Y\}$，运算结果仍为一个区间数(当 $*$ 为除法时，$0 \notin Y$)[6~9].

1.2 主成分分析

设 $\boldsymbol{X} = (x_{ij})_{n \times p}$ 为样本矩阵，考虑对标准化的数据进行主成分分析. 记样本相关矩阵为 $\boldsymbol{R}$，$\lambda_1 \geqslant \lambda_2 \geqslant \cdots \geqslant \lambda_p$ 为 $\boldsymbol{R}$ 的特征根，相应的单位特征向量记作 $\boldsymbol{b}_1, \boldsymbol{b}_2, \cdots, \boldsymbol{b}_p$，则 $\boldsymbol{B} = (\boldsymbol{b}_1, \boldsymbol{b}_2, \cdots, \boldsymbol{b}_p)$ 为列

① 收稿日期：2006－07－21；修订日期：2007－07－23.
基金项目：国家自然科学基金资助项目(70701026).
作者简介：郭均鹏(1973—)，男，山东昌邑人，博士，副教授，Email：guojp@tju.edu.cn

正交矩阵. $\boldsymbol{Y} = \boldsymbol{XB}$ 称为主成分. $a_i = \dfrac{\lambda_i}{\sum_{j=1}^{p}\lambda_j}$ 表示第 i 个主成分的方差贡献率,定义 $\sum_{i=1}^{m} a_i$ 为前 m 个主成分的累计贡献率[10~12].

2 区间样本数据的描述统计量

设 $i = 1,\cdots,n$ 表示 n 个样本点,$X_1,\cdots,X_p$ 为 p 个区间变量,$\xi_{ij} = [\underline{x_{ij}}, \overline{x_{ij}}]$ 表示样本 i 在变量 X_j 上的区间观测值,则样本矩阵可表示为

$$[\boldsymbol{X}] = \begin{bmatrix} [\underline{x_{11}}, \overline{x_{11}}] & \cdots & [\underline{x_{1p}}, \overline{x_{1p}}] \\ \vdots & & \vdots \\ [\underline{x_{n1}}, \overline{x_{n1}}] & \cdots & [\underline{x_{np}}, \overline{x_{np}}] \end{bmatrix} \tag{1}$$

其中,$[\boldsymbol{X}]$ 表示区间样本矩阵,以区别于一般的样本矩阵 $\boldsymbol{X}$.

2.1 均值和方差

考虑某一特定的区间变量 $X_j \equiv Z$,假设第 k 个样本的观测值为 $Z(k) = [\underline{z_k}, \overline{z_k}], k \in E = \{1,\cdots,n\}$,并且假设变量在该区间上服从均匀分布,即

$$P\{Z \leqslant x\} = \begin{cases} 0, & x < \underline{z_k} \\ \dfrac{x - \underline{z_k}}{\overline{z_k} - \underline{z_k}}, & \underline{z_k} \leqslant x < \overline{z_k} \\ 1, & x \geqslant \overline{z_k} \end{cases} \tag{2}$$

进一步,假设每个样本均以 $1/n$ 的等概率被观测,那么变量 Z 的经验分布函数为 n 个服从均匀分布的随机变量 $\{Z(k), k = 1,\cdots,n\}$ 的综合. 因此,由式(2)可得 Z 的经验分布函数为

$$F_Z(x) = \frac{1}{n}\left(\sum_{x \in Z(k)} \frac{x - \underline{z_k}}{\overline{z_k} - \underline{z_k}} + |\{k \mid x \geqslant \overline{z_k}\}|\right) \tag{3}$$

其中,$|A|$ 表示集合 A 的元素个数.

例如,设区间值变量在样本集合 $E = \{1,\cdots,8\}$ 上的观测值为

$$Z(E) = \{[0,2];[1,3];[1.5,2.5];[2,4];[3.5,5];[4.5,5.5];[5,7];[6.5,7.5]\} \tag{4}$$

为了求得 $F_Z(2)$,根据式(3)有

$$F_Z(2) = \frac{1}{8}\left(\frac{2-1}{3-1} + \frac{2-1.5}{2.5-1.5} + \frac{2-2}{4-2} + 1\right) = \frac{1}{4} \tag{5}$$

式(3)对 x 求导,可得 Z 的经验密度函数

$$f_Z(x) = \frac{1}{n}\sum_{k:x \in Z(k)} \frac{1}{\overline{z_k} - \underline{z_k}} = \frac{1}{n}\sum_{k \in E} \frac{I_{Z(k)}(x)}{l(Z(k))} \tag{6}$$

其中,$I_A(\cdot)$ 是集合 A 的指示函数,$l(B)$ 为区间数 B 的宽度.

于是,可如下求得区间变量 Z 的均值

$$\bar{Z} = \int_{-\infty}^{\infty} x f_Z(x)\mathrm{d}x = \frac{1}{n}\sum_{k \in E}\int_{-\infty}^{\infty} x \frac{I_{Z(k)}(x)}{l(Z(k))}\mathrm{d}x = \frac{1}{n}\sum_{k \in E}\frac{1}{\overline{z_k} - \underline{z_k}}\int_{\underline{z_k}}^{\overline{z_k}} x\mathrm{d}x = \frac{1}{2n}\sum_{k \in E}(\overline{z_k} + \underline{z_k}) \tag{7}$$

同理,可求得 Z 的样本方差

$$S^2 = \frac{1}{4n}\sum_{k \in E}(\overline{z_k} + \underline{z_k})^2 - \frac{1}{4n^2}\left[\sum_{k \in E}(\overline{z_k} + \underline{z_k})\right]^2 \tag{8}$$

2.2 协方差和相关函数

采取与求解区间变量的经验均值和方差类似的方法,可进一步得到二维变量 (X_1, X_2) 的描述统计量.

设区间样本集合为 E,$k \in E$ 的样本观测值为 $X(k) = \{X_1(k), X_2(k)\}, k = 1,\cdots,n$. $X(k)$ 在矩形 $Z(k) = X_1(k) \times X_2(k) = ([\underline{a_{1k}}, \overline{a_{1k}}], [\underline{a_{2k}}, \overline{a_{2k}}])$ 上取值. 类似于单变量情形,仍然假设二维变量 (X_1, X_2) 在矩形 $Z(k)$ 上分别服从均匀分布,则可得到 (X_1, X_2) 的经验联合密度函数

$$f(\xi_1, \xi_2) = \frac{1}{n}\sum_{k \in E}\frac{I_{Z(k)}(\xi_1, \xi_2)}{a(Z(k))} \tag{9}$$

其中,$I_{Z(k)}(\xi_1, \xi_2)$ 表示 (ξ_1, ξ_2) 是否落在矩形 $Z(k)$ 的指示函数;$a(Z(k))$ 是矩形的面积.

于是,可如下推导得到 X_1 和 X_2 间的经验协方差

$$Cov(X_1, X_2) \equiv S_{12} = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}(\xi_1 - \bar{X}_1)(\xi_2 - \bar{X}_2) f(\xi_1, \xi_2)\mathrm{d}\xi_1, \mathrm{d}\xi_2 \tag{10}$$

将式(9)代入,可得

$$Cov(X_1, X_2) = \frac{1}{n}\sum_{k \in E}\frac{1}{(\overline{x_{k1}} - \underline{x_{k1}})(\overline{x_{k2}} - \underline{x_{k2}})} \times \iint_{(\xi_1, \xi_2) \in Z(k)} \xi_1 \xi_2 \mathrm{d}\xi_1 \mathrm{d}\xi_2 - \bar{X}_1 \bar{X}_2 \tag{11}$$

通过运算,最终求得协方差函数为

$$Cov(X_1,X_2) = \frac{1}{4n}\sum_{k\in E}(\bar{x}_{k1} + \underline{x}_{k1})(\bar{x}_{k2} + \underline{x}_{k2}) - \frac{1}{4n^2}[\sum_{k\in E}(\bar{x}_{k1} + \underline{x}_{k1})][\sum_{k\in E}(\bar{x}_{k2} + \underline{x}_{k2})] \quad (12)$$

在此基础上，可进一步求得 X_1 和 X_2 间的相关函数

$$r(X_1,X_2) = \frac{S_{12}}{\sqrt{S_1^2 S_2^2}} \quad (13)$$

其中，S_{12} 和 S_j^2 分别由式(12) 和式(8) 计算得到.

3　基于经验相关矩阵的主成分分析

基于上一节给出的区间数据经验相关函数的计算方法，可按照下述算法实施区间主成分分析.

算法 1　基于经验相关矩阵的主成分分析.

步骤 1　由式(13) 计算得到经验相关函数(矩阵)，按1.2的方法实施PCA，得到列正交的特征向量矩阵，记为 $\boldsymbol{B}$.

步骤 2　将 $\boldsymbol{B} = (b_{ij})_{p\times p}$ 按元素的正负符号分解为两个矩阵 $\boldsymbol{B}^+$ 和 $\boldsymbol{B}^-$

$$\boldsymbol{B}^+ = (b_{ij}^+)_{p\times p},\ b_{ij}^+ = \begin{cases} b_{ij}, \text{当 } b_{ij} \geqslant 0 \\ 0, \text{当 } b_{ij} < 0 \end{cases} \quad (14)$$

$$\boldsymbol{B}^- = (b_{ij}^-)_{p\times p},\ b_{ij}^- = \begin{cases} b_{ij}, \text{当 } b_{ij} \leqslant 0 \\ 0, \text{当 } b_{ij} > 0 \end{cases} \quad (15)$$

显然，$\boldsymbol{B}^+$ ($\boldsymbol{B}^-$) 包含了 $\boldsymbol{B}$ 的非负(非正) 元素，而其负(正) 元素被替换为0.

步骤 3　分别由下述式(16) 和式(17) 计算得到区间主成分

$$\underline{\boldsymbol{Y}} = \overline{\boldsymbol{X}}\boldsymbol{B}^- + \underline{\boldsymbol{X}}\boldsymbol{B}^+ \quad (16)$$

$$\overline{\boldsymbol{Y}} = \overline{\boldsymbol{X}}\boldsymbol{B}^+ + \underline{\boldsymbol{X}}\boldsymbol{B}^- \quad (17)$$

其中，$\underline{\boldsymbol{X}}$ 和 $\overline{\boldsymbol{X}}$ 分别为样本数据矩阵(1) 的下限值和上限值；$\underline{\boldsymbol{Y}}$ 和 $\overline{\boldsymbol{Y}}$ 分别为区间主成分的下限值和上限值.

4　算例分析

以文献[1] 中的数据表为例，进行分析. 表1为区间样本数据.

表 1　区间样本数据表

Table 1 Interval sample data

样本	变量			
	X_1	X_2	X_3	X_4
S_1	[0.93, 0.94]	[-27, -18]	[170, 204]	[118,196]
S_2	[0.93, 0.94]	[-5, -4]	[192, 208]	[188,197]
S_3	[0.92, 0.92]	[-6, -1]	[99, 113]	[189,198]
S_4	[0.92, 0.93]	[-6, -4]	[104, 116]	[187,193]
S_5	[0.92, 0.92]	[-21, -15]	[80, 82]	[189,193]
S_6	[0.91, 0.92]	[0, 6]	[79, 90]	[187,196]
S_7	[0.86, 0.87]	[30, 38]	[40, 48]	[190,199]
S_8	[0.86, 0.86]	[22, 32]	[53, 77]	[190,202]

表2为由算法1得到的特征值及其累积贡献率.

表 2　特征值及其累积贡献率

Table 2 Eigenvalues and their cumulative contribution proportions

特征值	2.871 3	0.708 0	0.363 5	0.057 2
累积贡献率	0.717 8	0.894 8	0.985 7	1.000 0

选取前两个主成分，分别记为 PC_1、PC_2，累积贡献率达到89.48%. 表3为各样本在 PC_1 和 PC_2 上的区间值. 将表3的数据绘制成图，如图1所示.

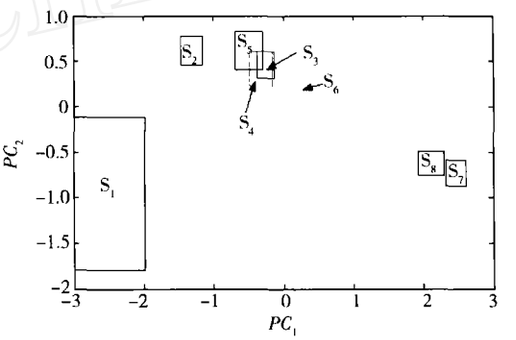

图 1　各样本的前两个主成分的区间值示意图

Fig. 1 Interval values of the first two principals of all the individuals

表 3　前两个主成分的区间值

Table 3 Interval values of the first two principals

样本	主成分	
	PC_1	PC_2
S_1	[-2.99, -1.98]	[-1.79, -0.12]
S_2	[-1.46, -1.16]	[0.46, 0.77]
S_3	[-0.37, -0.13]	[0.31, 0.61]
S_4	[-0.48, -0.15]	[0.22, 0.60]
S_5	[-0.69, -0.30]	[0.40, 0.83]
S_6	[0.18, 0.38]	[0.07, 0.33]
S_7	[2.32, 2.61]	[-0.87, -0.59]
S_8	[1.93, 2.30]	[-0.76, -0.50]

基于上述结果可对其作进一步的分析，但已不是本文重点，故不再赘述. 与文献[1] 中的顶点法和中点法相比较，计算结果相似，前两个主成分的累计贡献率均接近 90%. 但文中方法基于区间数据直接计算相关矩阵进行主成分分析，方法简单直观且不丢失信息，克服了顶点法计算量大与中点法损失数据信息的缺点.

5 结 论

通过计算区间数据的相关矩阵，提出了一种针对区间数据的 PCA 的新的方法，给出了具体的实施算法. 将来可进一步研究区间数据的其它统计方法的求解.

参 考 文 献:

[1] Bock H H, Diday E. Analysis of Symbolic Data[M]. Berlin, New York: Springer-Verlag, 2000.

[2] Cazes P, Chouakria A, Diday E. Symbolic principal components analysis[A]. In Bock H H, Diday E. (Eds.) Analysis of Symbolic Data[M]. Berlin, New York: Springer-Verlag, 2000.

[3] 胡 艳, 王惠文. 一种海量数据的分析技术—符号数据分析及应用[J]. 北京航空航天大学学报(社会科学版), 2004, 17(2): 40—44.
Hu Yan, Wang Huiwen. A new data mining method based on huge data and its application[J]. Journal of Beijing University of Aeronautics and Astronautics (Social Sciences Edition), 2004, 17(2): 40—44. (in Chinese)

[4] Pierpaolo D, Paolo G. A least squares approach to principal component analysis for interval valued data[J]. Chemometrics and Intelligent Laboratory Systems, 2004, 70(2): 179—192.

[5] Antonio I. Spaghetti PCA analysis—An extension of principal components analysis to time dependent interval data[J]. Pattern Recognition Letters, 2006, 27(5): 504—513.

[6] 郭均鹏, 李汶华. 区间线性规划的标准型及其最优值区间[J]. 管理科学学报, 2004, 7(3): 59—63.
Guo Junpeng, Li Wenhua. Standard form of interval linear programming and its optimal objective interval value[J]. Journal of Management Sciences in China, 2004, 7(3): 59—63. (in Chinese)

[7] 徐泽水, 孙在东. 一类不确定型多属性决策问题的排序方法[J]. 管理科学学报, 2002, 5(3): 35—39.
Xu Zeshui, Sun Zaidong. Priority method for a kind of multi-attribute decision-making problems[J]. Journal of Management Sciences in China, 2002, 5(3): 35—39. (in Chinese)

[8] 郭均鹏, 吴育华, 李汶华. 基于标准化区间权重向量的层次分析法研究[J], 系统工程与电子技术, 2004, 26(7): 900—902.
Guo Junpeng, Wu Yuhua, Li Wenhua. On the analytic hierarchy process with normalized interval weight vector[J]. Systems Engineering and Electronics, 2004, 26(7): 900—902. (in Chinese)

[9] Sugihara K, Tanaka H. Interval evaluations in the analytic hierarchy process by possibility analysis[J]. Computational Intelligence, 2001, 17(3): 567—579.

[10] 王慧英. 基于主成分分析的制造企业信息化评价方法研究[J]. 天津大学学报: 社会科学版, 2005, 7(2): 106—109.
Wang Huiying. On evaluation method of manufacturing enterprise informatization based on principal component analysis [J]. Journal of Tianjin University (Social Sciences Edition), 2005, 7(2): 106—109. (in Chinese)

[11] 刘金兰, 朱晓旸. 顾客满意度指标重要性测量的主成分分析与多元回归方法[J]. 天津大学学报: 社会科学版, 2004, 6(2): 159—163.
Liu Jinlan, Zhu Xiaoyang. Multivariate regression combined with principal components analysis used to measure the importance of customer satisfaction indicators[J]. Journal of Tianjin University (Social Sciences Edition), 2004, 6(2): 159—163. (in Chinese)

[12] 王明进, 陈奇志. 基于独立成分分解的多元波动率模型[J]. 管理科学学报, 2006, 9(5): 56—64.
Wang Mingjin, Chen Qizhi. Multivariate volatilities modeling based on independent components[J]. Journal of Management Sciences in China, 2006, 9(5): 56—64. (in Chinese)

(下转第 95 页)

Study on asymmetric information screening in supply chain with long-term cooperation prospect

HUO Jia-zhen, ZHANG Jian-jun, ZHAO Jin
School of Economics and Management, Tongji University, Shanghai 200092, China

Abstract: This thesis firstly studies one-off supply chain cooperation involving screening asymmetric information of common hypothesis, and constructs a Second-Best coordination model to prove that it is impossible to realize the absolute optimal coordination in such an occasion. Then, extending the one-off relationship to a long-term cooperation, the thesis reveals that repeated Stackelberg Games between the upper and the lower have an influence on supply chain coordination with asymmetric-information in a long-term background. If the agent's (retailer's) private information of one season is associated with that of other seasons, and cannot be testified in the next season, ratchet effect will be resulted in information screening. However, if the private information is likely to be identified, whether the supply chain, with an incentive of the principal's (manufacturer's) trigger strategy, can be optimally coordinated depends on the probability of the private information being identified, retailer's information rent, his income in symmetric information background and reservation utility when there is no cooperation. This thesis provides corresponding quantitative descriptions for all these conclusions respectively.

Key words: supply chain coordination; information asymmetry; information screening; long-term cooperation; repeated games; ratchet effects; trigger strategy

(上接第52页)

Interval PCA based on empirical correlation matrix

GUO Jun-peng, LI Wen-hua
School of Management, Tianjin University, Tianjin 300072, China

Abstracts: A methodology of principal component analysis (PCA) for interval data is proposed. Empirical descriptive statistics for interval data is first studied including mean and the variance of univarible and the covariance and correlation coefficient of bivarible. Based on the empirical correlation matrix, an arithmetic of PCA for interval data is put forward which gives interval principal values. An example is illustrated. It indicates that the given method is simple and can overcome the shortcoming of the existing methods of PCA for interval data.

Key words: principal component analysis; interval; correlation matrix