

重点学术期刊专项基金管理中的期刊评价^①

——基于简化的区间数据主成分分析方法

张寅¹, 王岩², 王惠文¹

(1 北京航空航天大学经济管理学院, 北京 100191;

2 国家自然科学基金委员会计划局, 北京 100085)

摘要: 国家自然科学基金委设立“重点学术期刊专项基金”资助国内优秀学术期刊, 在优秀期刊的遴选和资助效果评价过程中, 存在期刊数量大, 评价指标多等问题。基于此, 提出一种针对大规模高维数据的简化的区间数据主成分分析方法 (simplified principal component analysis, SPCA)。该方法将区间主成分分析分解成 2 个基本阶段: 第一是如何更加简单和高精度地计算数据集合的主轴, 第二是如何绘制可视性与可解释性都更强的主平面图, 以增强研究人员对大规模数据主要特征的洞察能力。采用 SPCA 方法, 一方面能够从整体上研究各学科期刊的差异, 说明学科之间的不可比较性; 另一方面能够筛选出衡量期刊水平的关键指标, 遴选优秀期刊, 同时分析连续受资助期刊的动态资助效果。

关键词: 区间数据; SPCA; 主平面图; 期刊评价; 资助效果分析

中图分类号: G202 G312 **文献标识码:** A **文章编号:** 1007-9807(2010)07-0088-07

0 引言

国家自然科学基金委自 1999 年度起, 设立了“重点学术期刊专项基金”。该基金逢偶数年受理申请, 以两年为周期, 每次资助国内约 30 种优秀期刊。在优秀期刊的遴选过程中, 除了考虑量化指标外, 还要结合期刊的管理运营状况、国际化潜力及业内评审专家的意见, 本文仅从文献计量指标的角度, 从指标遴选和学科差异分析的角度来探讨期刊评价问题。

考虑到各学科期刊的差异, 目前对于期刊评价的研究, 多采用单一指标 (如影响因子) 或采用主成分分析、模糊评判等方法综合各种指标在学科内部进行评价。而由于期刊数量大, 评价指标复杂, 很难从整体上对学科差异给出综合判断, 并且在评价不同学科期刊时, 也存在较大困难。基于

此, 本文将提出一种针对大规模数据的简化的区间数据主成分分析方法, 更加简单而高精度地计算数据集合的主轴, 筛选评价期刊水平的关键指标, 并且通过绘制可视性与可解释性都更强的主平面图, 更加直观地从整体上反映学科差异。另外, 通过连续受资助期刊在关键指标上的动态变化, 可以很好地反映基金的资助效果。

1 区间数据的 SPCA 方法

1.1 区间数据主成分分析方法简介

区间数据是符号数据中应用最广泛的类型之一, 它通过将原始数据处理为若干个高维超矩形, 不仅极大地降低了样本容量, 同时又能够表现数据集合的集中趋势和离散程度。近年来, 人们提出了一系列的区间数据主成分分析方法, 其中最常

① 收稿日期: 2010-04-19; 修订日期: 2010-05-05.

基金项目: 国家自然科学基金创新研究群体科学基金资助项目 (70821061); 国家自然科学基金重点资助项目 (70531010); 国家自然科学基金资助项目 (70771004, J0824301).

作者简介: 张寅 (1986-), 女, 四川成都人, 硕士生, Email: zhangyin407@yahoo.com.cn.

用的是顶点法和中心点法主成分分析, 它们在实际应用中都存在一些不尽如人意的缺憾。

顶点法主成分分析 (vertices PCA, VPCA) 把区间数据的每一个样本点看作一个 p 维超立方体。然后, 以 p 维超立方体的各个顶点作为代表样本点, 对区间数据表做展平处理。这样, 就可以把区间数据表转换成普通数据表, 然后采用经典的主成分分析方法来计算主轴, 并用顶点投影的方法计算主成分。VPCA 方法最大的局限性是在处理高维数据时会出现维数灾难。这是由于, 对于一个 $n \times p$ 维的区间数据表, 如果用顶点法展成普通数据表, 其样本点的数量将是 $n \times 2^p$, 它将会随着变量个数 p 的增加呈指数膨胀的趋势。而另一方面, 由于区间数据采用“装箱”的方式对原始样本集合进行概括, 因此往往会不合理地扩大数据的取值范围, 并且可能引起主轴移动, 造成不必要的计算误差。

而中心法主成分分析 (centers PCA, CPCA) 采用每一类样本点的重心作为该类的代表, 构成普通数据表, 然后再结合经典的主成分分析方法计算主轴和主成分。CPCA 确实解决了 VPCA 中的维数灾难问题, 但是由于仅使用一个重心点来代替整个类中的所有样本点, 必然会丢失很多数据信息, 在应用中甚至会歪曲主轴的本来方向。

李汶华、郭均鹏比较了 VPCA 和 CPCA, 发现其协方差矩阵具有极大的相似性, 随着变量数和样本数的增加, 两种方法的效度均有下降^[1]。而王惠文、关蓉指出 VPCA 与 CPCA 中计算主成分的方法是一致的。

针对区间数据主成分分析现有方法的种种局限性, 本文将提出一种新的区间数据主成分分析方法——simplified principal component analysis (SPCA)。与现有的几种区间数据主成分分析方法相比, 该方法计算过程简单, 并且由于在计算主轴时完全采用原始数据信息, 因此计算精度更高, 所得到的主轴也更加符合原始数据的规律特征。

1.2 区间数据 SPCA 方法的计算步骤

设有一个由 N 个样本点和 p 个变量组成的数据表 $X_{N \times p} = (x_{ij})_{N \times p}$ 。根据数据系统中的某些重要属性, 将这 N 个样本点划分成 n 类子集合: S_1, S_2, \dots, S_n 。如果用区间数据 e_k 来概括 S_k , 常用的方

法有

$$e_k = \{ [\underline{x}_{kj}, \bar{x}_{kj}] = [\min_{i \in S_k} (x_{ij}), \max_{i \in S_k} (x_{ij})] \}$$

$$j = 1, 2, \dots, p \quad (1)$$

在实际应用中, 还可以采用适当的分位点作为每个区间的上下限。

一般地, n 个数据集合 $S_k (k = 1, \dots, n)$ 可以被概括成一个 $n \times p$ 维的“样本点 \times 定量区间变量”类型的数据表, 其形式如下

$$X_{n \times p} = \begin{bmatrix} [\underline{x}_{11}, \bar{x}_{11}] & [\underline{x}_{12}, \bar{x}_{12}] & \dots & [\underline{x}_{1p}, \bar{x}_{1p}] \\ [\underline{x}_{21}, \bar{x}_{21}] & [\underline{x}_{22}, \bar{x}_{22}] & \dots & [\underline{x}_{2p}, \bar{x}_{2p}] \\ \vdots & \ddots & \ddots & \vdots \\ [\underline{x}_{n1}, \bar{x}_{n1}] & [\underline{x}_{n2}, \bar{x}_{n2}] & \dots & [\underline{x}_{np}, \bar{x}_{np}] \end{bmatrix} = \begin{bmatrix} e'_1 \\ e'_2 \\ \vdots \\ e'_n \end{bmatrix} \quad (2)$$

式中, $e'_i = ([\underline{x}_{i1}, \bar{x}_{i1}] [\underline{x}_{i2}, \bar{x}_{i2}] \dots [\underline{x}_{ip}, \bar{x}_{ip}])$, 被称为区间数据表的样本点。

则区间数据的 SPCA 方法的计算过程如下:

步骤 1 采用普通的主成分分析方法, 对原始数据表 $X_{N \times p} = (x_{ij})_{N \times p}$ 的相关系数矩阵 R 求特征值和特征向量。得到特征值为 $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_p \geq 0$; 对应的特征向量为 u_1, u_2, \dots, u_p , 其中, $u_h = (u_{h1}, u_{h2}, \dots, u_{hp})'$; 向量 u_1, u_2, \dots, u_p 之间是标准正交的;

步骤 2 根据累计贡献率确定保留主成分的个数 m 。累计贡献率的计算方法见式 (3)

$$Q_m = \frac{\sum_{h=1}^m \lambda_h}{\sum_{j=1}^p \lambda_j} \quad (3)$$

步骤 3 求主成分 F_1, F_2, \dots, F_m 。若记 $F_h(i) = [f_{ih}, \bar{f}_{ih}]$ 是一个区间数据, 它是第 i 个区间数据的样本点在第 h 主成分上的取值。按照 VPCA 和 CPCA 中的主成分计算方法, 可以得到 $F_h(i) = [f_{ih}, \bar{f}_{ih}]$ 的计算公式如下

$$f_{ih} = \sum_{j=1}^p u_{hj} [\tau \underline{x}_{ij} + (1 - \tau) \bar{x}_{ij}] \quad (4)$$

$$\bar{f}_{ih} = \sum_{j=1}^p u_{hj} [(1 - \tau) \underline{x}_{ij} + \tau \bar{x}_{ij}] \quad (5)$$

$$\text{其中, } \tau = \begin{cases} 0 & u_j \leq 0 \\ 1 & u_j > 0 \end{cases}$$

由此得到主成分 $F_h = (F_h(1), F_h(2), \dots, F_h(n))'$, $h = 1, 2, \dots, m$.

2 SCI期刊学科特点分析

在学术期刊评价问题中, 不同学科期刊由于社会因素(如学科自身的特色、学科发展阶段和速度等)和统计因素(如统计源的学科结构等), 在文献计量指标上具有显著的差异. 而由于学术期刊数量庞大, 评价指标也很多, 很难从宏观上对某学科的整体水平给出科学判断. 本节中将以SCI期刊数据为例, 对比普通PCA、VPCA、CPCA在可视化、计算复杂度和精确度等方面存在的局限性, 采用本文所提出的SPCA方法, 首先寻找评价期刊发展水平的关键因素, 然后利用区间数据主平面图, 直观地表现各学科在关键因素上的差异, 说明不同学科期刊之间的不可比性.

2.1 数据来源

本文以JCR网络数据库(2007)提供的6337种SCI收录期刊为例, 选用总被引频次(Total Cites)、影响因子(Impact Factor)、五年影响因子(5-Year Impact Factor)、即年指标(Immediacy Index)、载文量(Articles)、特征因子分值(EigenfactorTM Score)、文章影响分值(Article InfluenceTM Score)等7个指标进行分析.

2.2 普通因子分析

剔除数据缺失的样本点后进行因子分析, 前两个主因子的累积贡献率达到89.1%, 因此选用前两个主因子就基本能够代表原始变量的信息. 图1为因子载荷图, 可以看到, 第一主因子与影响因子、五年影响因子、文章影响分值、即年指标等指标呈强正相关, 均为反映期刊影响力的相对引用指标, 而第二主因子与总被引频次、特征因子分值、载文量等指标呈强, 反映期刊的信息含量. 图2为期刊样本主平面图, 可以看到, 大部分期刊的第一和第二主因子得分处于-5到5之间, 也有一定数量的期刊表现突出, 甚至有期刊远远超过这个水平. 由于样本点数量太多, 可视性很差, 并且由于期刊之间的表现差异巨大, 很难通过该图来

反映各学科期刊的整体水平.

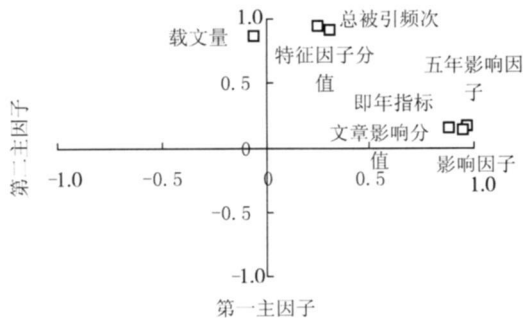


图1 普通因子分析因子载荷图

Fig. 1 Loading plot for factor analysis

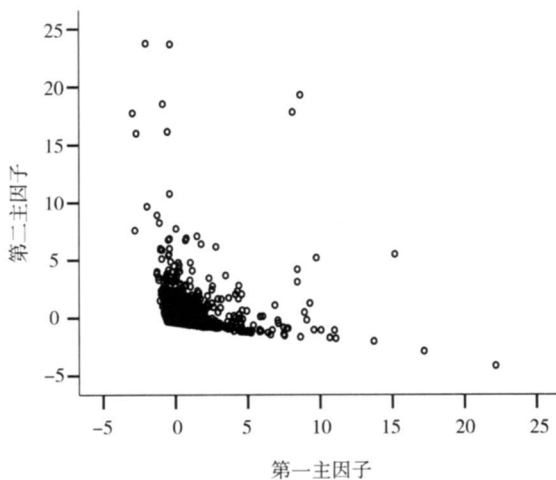


图2 普通因子分析样本主平面图

Fig. 2 The plane of principal axes for factor analysis

2.3 顶点法主成分分析(VPCA)

按照国家自然科学基金委的学部分类方法, 将SCI期刊分为数理科学、化学科学、工程材料科学、地球科学、生命科学、信息科学、管理科学、综合等8类, 采用上下四分位点作为区间端点, 利用顶点法展平区间数据表, 并进行主成分分析. 前两个主成分的累积贡献率仅43.4%, 因子载荷图如图3所示, 计算结果与对所有原始数据进行主成分分析的结果相差较大. 分析结果很好地说明了, 由于VPCA采用“装箱”的方式, 用顶点代替整个区间, 会不合理地扩大数据的取值范围, 造成主轴移动, 引起计算误差.

2.4 中心点法主成分分析(CPCA)

采用中心点法主成分分析得到前两个主因子的累积贡献率为88.1%, 因子载荷图如图4所示, 与原始数据的分析结果有一定的差距, 说明由于CPCA采用中心点代替整个超立方体, 丢失了大

量原始信息, 会造成主轴移动.

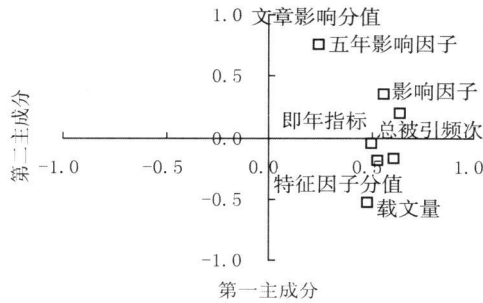


图 3 VPCA 因子载荷图

Fig. 3 Loading plot for VPCA

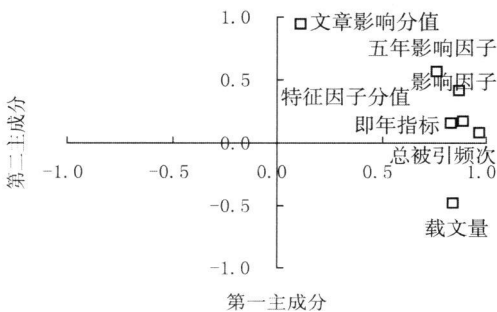


图 4 CPCA 因子载荷图

Fig. 4 Loading plot for CPCA

2.5 简化区间主成分分析 (SPCA)

根据原始数据的因子分析结果 (如图 1 所示), 全部 7 个描述期刊情况的文献计量指标可以分为两个维度: 第一主因子以影响因子、五年影响因子、文章影响分值、即年指标等 4 个指标为代表, 第二主因子以总被引频次、特征因子分值、载文量等 3 个指标为代表. 采用简化区间主成分分析方法 SPCA, 得到的区间数据主平面图如图 5 所示. 以“十”字表示 8 类期刊的区间数据, 其中, 每一个“十”字的交叉点表示相应学科期刊的中心位置, 而十字的长短则分别表示该类期刊在横纵轴上的离散范围. 由于采用四分位点为端点绘制十字图, 能够减小极端值的影响, 更客观地反映各学科的总体水平. 在这两个维度上, 各学科的差异明显, 如化学、生命类期刊在这两个主因子上得分均较高, 而工材、管理类期刊在这两个因子上得分均较低.

造成这种现象的主要原因有两个. 一是学科自身的发展特点, 如科学家研究行为的社会性、学科发展所处的阶段、学科间交叉渗透的程度等. 二是统计源的学科结构, 如在来源期刊中, 某些学科在整体学科结构中占有较高比例, 其影响因子等

指标必然偏高^[2].

可见, 由于学科自身的特点不同, 不同学科期刊在文献指标上具有显著差异. 在期刊评价的研究中, 如果把不同学科的期刊放在一起比较, 对于某些学科的期刊, 是非常不公平的.

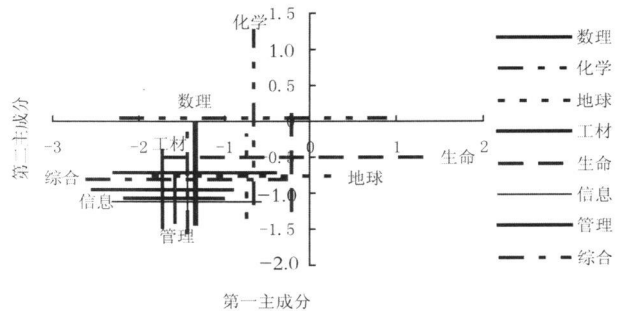


图 5 SPCA 样本主平面图

Fig 5 The plane of principal axes for SPCA

2.6 各种方法的比较

采用 SPCA 的计算过程非常简单, 对原始数据的代表性和结果的可视性均很强. 表 1 比较了本文提出的 SPCA 方法与各种常用区间主成分分析方法的优劣.

表 1 SPCA 与其它区间主成分分析方法的优劣

Table 1 Comparison of SPCA and other PCA methods based on interval data

	因子分析	VPCA	CPCA	SPCA
计算主轴的数据表维度	n	$n \times 2^p$	n	n
对原始数据的概括精度	利用原始数据计算主轴	比较差	非常差	利用原始数据计算主轴
大规模数据主平面图的可视性	非常差	好	好	好
计算过程的复杂程度	最简单	复杂	较复杂	简单

3 基金资助效果评价

3.1 优秀期刊遴选

上节的分析已经指出了不同学科期刊之间的不可比性, 根据本文提出的 SPCA 方法, 还可以判断我国 SCI 期刊在国际上的位置, 在学科内部遴选优秀期刊 (本文的阐述仅限 SCI 期刊).

采用因子分析计算出的两个主轴, 仍然用上下四分位点作为端点, 确定该学科大部分期刊的

整体水平,然后将中国该学科 SCI期刊数据投影到主轴上,可以得出我国 SCI期刊的国际地位,如图 6所示为我国数学学科 SCI期刊的国际水平的散点图.

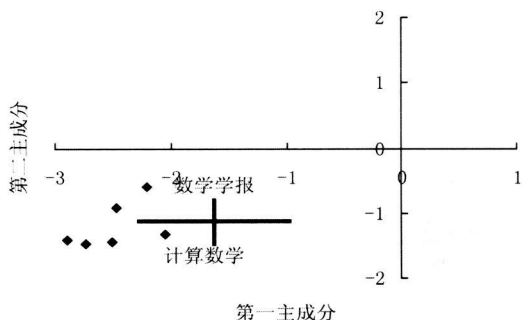


图 6 我国数学类 SCI期刊国际地位散点图

Fig. 6 Scatter plot for international position of Chinese mathematical SCI journals

由图 6 在我国数学类 SCI期刊中,《数学学报》《计算数学》在两方面指标上均已处于国际中游水平以上,属于我国 SCI期刊中的佼佼者.而 2008 年专项基金资助的数学类期刊有两种,正好为《数学学报》和《计算数学》也反映了基金资助的科学性和合理性.

同样的,图 7 显示我国医药类 SCI期刊的国际地位,可以看到,《中国药理学报》《亚洲男性学》在两个主成分得分上均已处于国际中游水平,而《中华医学杂志》的第二主成分得分达到了国际领先水平,这三种期刊均属于我国医药类 SCI期刊中较优秀的.而 2008 年专项基金资助的医药类期刊有 3 种,正好为《中国药理学报》《亚洲男性学》和《中华医学杂志》这一结果也说明本文的分析和基金的资助相吻合.

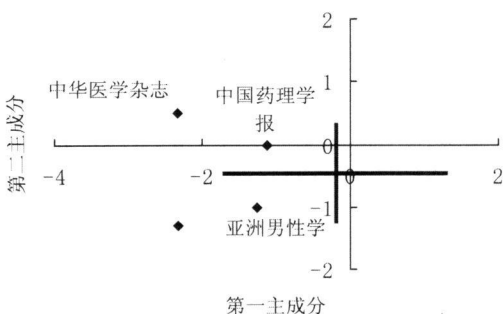


图 7 我国医药类 SCI期刊国际地位散点图

Fig. 7 Scatter plot for international position of Chinese medical SCI journals

图 8 显示我国综合类 SCI期刊的国际地位,可以看到,《科学通报》的第一主成分得分达到国

际中游水平,而第二主成分得分则遥遥领先于该类大部分国际期刊,而《科学通报》也正好是 2008 年专项基金资助的唯一综合类期刊.

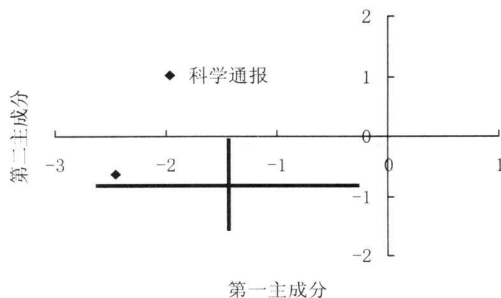


图 8 我国综合类 SCI期刊国际地位散点图

Fig. 8 Scatter plot for international position of Chinese multidisciplinary SCI journals

3.2 基金资助效果动态分析

在本节中,将以某连续受到专项基金资助的生物科学类期刊 A 为例,研究其在国际上地位的动态变化情况(选取的时间段为 2004—2007 年),说明 SPCA 方法在期刊资助后评估中的应用.

由于论文所用的 7 个文献指标有些是 2007 年新增加的,无法获得其动态数据.而根据上节的分析结果,对于 SCI期刊来说,反映其水平的指标分为两个维度:一是以影响因子为代表的反映期刊影响力的指标,二是以总被引频次为代表的反映期刊信息含量的指标,用“影响因子”和“总被引频次”这两个指标就可以涵盖原始变量的绝大部分信息,大大简化了期刊评价的指标维度.采用这两个指标分别作为横纵轴,绘制各学科国际整体水平的“十字图”.为了方便比较,认为这些指标的国际整体水平不会有太大变化,采用 2007 年的数据作为衡量国际水平的基础,如图 9 所示.

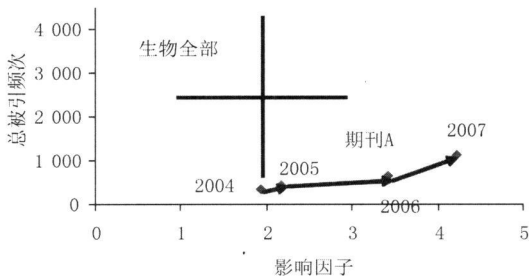


图 9 期刊 A 的动态变化情况

Fig. 9 Dynamic funding effect evaluation of journal A

由上图 9 看到,该期刊近年来进步非常明显,在影响因子和总被引频次上均有较大幅度的提升,2004 年时还处于国际中游水平,到 2007 年已

经达到国际领先, 提高非常迅速, 表现很突出, 说明专项基金的资助取得了很好的效果。

除了文献指标, 基金的资助在期刊的管理运营水平、办刊条件的提高等方面也起到了重要的作用, 由于非本文重点, 在此不一一叙述。

需要强调的是, 由于期刊的发展思路和政策有所不同, 如果仅通过短期内的文献指标来对其水平给出判断, 难免有失偏颇。文献指标由于其计算方式各有特点, 如影响因子存在时滞, 它可能在期刊因为要扩大影响而提高载文量时, 在短期内出现倒退。这告诉我们不能仅仅从数据看表面现象, 而应该通过更加深入的研究, 观察期刊的动态发展变化情况, 更加系统、科学地对其水平给出判断, 并从他们的成功和失误中吸取经验和教训。

4 结 论

针对优秀期刊遴选和资助效果评价过程中存在的期刊数量大、计量指标复杂等问题, 本文提出一种简化的区间主成分分析方法 (SPCA)。该方法分为 2 个基本阶段: 第一阶段利用原始数据简单而高精度地获取数据集的主轴; 第二阶段利

用区间投影绘制可视性与解释性都更高的主平面图。

为了进一步说明该方法的工作过程, 本文以 JCR 网络数据库 (2007) 提供的 6 337 种 SCI 期刊为案例, 首先分别应用普通因子分析、VPCA、CPCA 等方法分析, 所得到的结果存在可视性不强、对原始数据代表性不强、计算复杂等各种问题, 而采用本文提出的 SPCA 方法, 分析结果能够很好地解释大规模期刊数据蕴含的关键信息, 反映不同学科期刊的整体特点, 说明学科间的不可比性。该方法计算过程简单, 精度高, 结果的可解读性强, 能够有效辅助研究人员认识系统的主要特征, 在实际中有非常好的应用前景。

通过 SPCA 方法, 还能够直观地反映某一期刊在国际上所处的位置, 从而遴选出具有国际化潜力的优秀期刊。在实际中, 还可以利用对主轴代表性最强的指标来概括原始变量的信息, 作为期刊遴选的关键指标, 这样的处理大大降低了变量的维度, 且可操作性和解释性都更强。借助关键指标和区间数据的思想, 还可以尝试分析连续受到基金资助期刊的动态变化, 评价基金资助的效果, 做好基金资助的后评估工作。

参 考 文 献:

- [1] 李汶华, 郭均鹏. 区间主成分分析方法的比较 [J]. 系统管理学报, 2008, 17(1): 94-98
Li W enhua Guo Junpeng Comparative study on the PCA for interval data [J]. Journal of Systems Engineering 2008 17 (1): 94-98 (in Chinese)
- [2] 金碧辉, 汪寿阳, 等. 国际论文与国内论文合一统计方法研究 [J]. 管理科学学报, 1999, 2(3): 59-65
Jin B ihui Wang Shouyang et al A united method of counting international and domestic articles [J]. Journal of Management Sciences in China 1999, 2(3): 59-65 (in Chinese)
- [3] Bock H, Diday E. Analysis of Symbolic Data [M]. Berlin Springer 2000
- [4] 汪寿阳, 金碧辉. SCI、SSCI 与管理科学期刊 [J]. 管理科学学报, 2000, 3(4): 1-6
Wang Shouyang Jin B ihui SCI, SSCI and journals of management science [J]. Journal of Management Sciences in China 2000 3(4): 1-6 (in Chinese)
- [5] 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, 117(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, 117(2): 40-44 (in Chinese)
- [6] Moore R E, Kearott R B, Cloud M J. Introduction to Interval Analysis [M]. Philadelphia Siam, 2008
- [7] Moore R E. Interval Analysis [M]. N.J. Prentice Hall 1966
- [8] Diday E. The symbolic approach in clustering and related method of data analysis: The basis choices [C] // In Bock H H.

- [9] Lauro N C, Palumbo F. Principal components analysis of interval data: A symbolic data analysis approach[J]. *Computational Statistics*, 2000, 15(1): 73–87.
- [10] Góia F, Lauro N C. Principal component analysis on interval data[J]. *Computational Statistics*, 2006, 21(2): 343–363.
- [11] Bock H, Dilla E. Analysis of symbolic data[C] // *Experimental Methods For Extracting Statistical Information From Complex Data*. Heidelberg: Springer Verlag, 2000.
- [12] 李大鹏. 因素区间数据的分析方法及其在股票市场研究中的应用[D]. 北京: 北京航空航天大学, 2005.
Li Dapeng. Factor Interval Analysis and Its Application to the Chinese Stock Markets[D]. Beijing: Beijing University of Aeronautics and Astronautics, 2005. (in Chinese)
- [13] 王亚楠. 基于区间数据分析的中国股票市场运行特征研究[D]. 北京: 北京航空航天大学, 2007.
Wang Yanan. Research on the Operation of China Stock Market Based on Interval Data Analysis[D]. Beijing: Beijing University of Aeronautics and Astronautics, 2007. (in Chinese)
- [14] 何颖, 官建成. 国际国内期刊论文的综合评价方法及应用[J]. *科学学研究*, 2007, 25(2): 215–221.
He Ying, Guan Jiancheng. A combinative method for evaluating international and domestic scientific papers and its application[J]. *Studies in Science of Science*, 2007, 25(2): 215–221. (in Chinese)

Evaluating of academic journals in management of key academic journal fund: An application of simplified principal component analysis based on interval data

ZHANG Yin¹, WANG Yan², WANG Huiwen¹

1. School of Economics and Management, Beihang University, Beijing 100191, China

2. National Natural Science Foundation of China, Beijing 100085, China

Abstract National Natural Science Foundation of China (NSFC) encountered many problems such as an increasing number of applicants and a great many indices when selecting excellent academic journals as grantees of “Key Academic Journal Fund” and evaluating funding effects. This paper puts forward a method for Principal Component Analysis on interval data referred to as Simplified Principal Component Analysis (SPCA), specific for large-scaled and high-dimensional data. The proposed method involves two steps: first look for factor axes of the original large-scaled data and then illustrate the analytical results on a low-dimensional space with strengthened visibility and high interpretation. Performing SPCA on academic journal dataset helps to understand the differences in all disciplines and to choose key criteria for selecting excellent journals and measuring funding effect dynamically.

Key words interval data; SPCA; the plane of principal axes; journal evaluation; funding effect analysis