基于主成分分析法的科技期刊评价方法

张 弘 赵惠祥 刘燕萍 陶文文
(同济大学学报(自然科学版)编辑部,200092,上海)

摘 要 针对目前科技期刊评价方法中存在的指标之间相关性问题和指标权重选取问题,提出了基于主成分分析的科技期刊评价方法。该评价方法通过相关系数矩阵的特征向量将评价指标线性变化成彼此独立的主成分,根据主成分累计贡献值确定主成分的取用维数,由主成分方差确定权重。其优点是:可以消除由于指标间的相关性带来的偏差,降低计算维数,从而降低指标选择的难度,提高评价结果的可信度;此外,可以消除人为确定指标权重引起的弊病,使评价结果更具客观性和准确性。

关键词 科技期刊;期刊评价;主成分分析;相关系数矩阵;特征向量

Evaluation method on sci-tech journals based principle component analysis // ZHANG Hong, ZHAO Huixiang, LIU Yanping, TAO Wenwen

Abstract An evaluation method on sci-tech journals based on principal component analysis is proposed to deal with the relevance between various indices and the selection of index weight. According to this evaluation method, the evaluation index is turned into mutually independent principal components on the basis of linear transformation through the eigenvector of correlation coefficient matrix. The dimension selection of the principal components is determined according to the accumulated contributing value of the principal components and the weight is determined by the variance of the principal components. When this method is employed to evaluate the sci-tech journals, the deviation brought about by the relevance between various indices can be eliminated and the calculating dimension can be lowered, hence making the selection of index easier and improving the credibility of the evaluation results. In addition, the problems arising from subjective determination of index weight can be solved, and the evaluation results tend to be more objective and accurate.

Key words sci-tech journal; journal evaluation; principle component analysis; correlation coefficient matrix; eigenvector

Author's address Editorial Department of Journal of Tongji University, 200092, Shanghai, China

采用计量统计指标对科技期刊进行评价是当前科技期刊研究的一个热点,我国近年来不断有关于科技期刊评价方法和评价指标体系研究的报道,有些已在一些科技期刊评价中得到应用[1-5]。但是,我国目前采用的一些评价方法并不完善,其中最明显的有 2 个缺陷:一是需要人为确定指标权重,由此会产生主观偏差;二是对于指标之间的相关性未给予充分的考虑,造成所确定的权重并不是实际计算过程中所体现的真实权重。

科技期刊评价属于多因素综合评价,目前各种评价体系一般采用线性函数模型,其通用表达式如下:

$$F = a_0 + \sum a_j x_j + \varepsilon。$$

其中:a_0 为常数项(如政治标准合格分);a_j 为指标权重;x_j 为评价指标;ε 为误差修正系数。2 个缺陷即 a_j 的准确选取问题和 x_j 的线性相关问题。

本文采用主成分分析法,构建一种新的科技期刊评价方法,可以克服以上缺点,并且可以用于科技期刊的动态评价。

1 主成分分析法

主成分分析法是数理统计学中一种多元分析方法,是因子分析法的一种特殊形态。其基本原理是在一群具有相关性的统计数据中找出彼此间趋向独立的并且足以反映原始数据的共同因素,用少于原来变量维数且互不相关的主成分替代原来的变量,其权重由方差贡献率计算得出。主成分分析不仅可以反映原有指标的信息量,而且可以解决指标之间信息重叠问题和权重选取问题,并且可以进行降维计算以减少计算量。

1.1 原始数据标准化 设有 n 个样本,每个样本有 P 项指标,则原始样本矩阵为

$$X = (X_{ij})_{n\times p} \quad i = 1,2,\cdots,n; j = 1,1,\cdots,p。$$

式中 X_{ij} 表示第 i 个样本的第 j 项指标。考虑指标的趋势和量纲问题,可采用倒数的方法对指标进行同趋势化处理,然后用 Z-score 法对样本进行标准化变换[6],即

$$Z_{ij} = (x_{ij} - \bar{x}_j)/S_j。 \tag{1}$$

其中:$\bar{x}_j = \sum_{i=1}^{n} x_{ij}/n$, $S_j^2 = [\sum_{i=1}^{n}(x_{ij} - \bar{x}_j^2]/(n=1)$,则可得标准化样本矩阵

$$\boldsymbol{Z} = (Z_{ij})_{n\times p} \quad i = 1,2,\cdots,n; j = 1,1,\cdots,p。$$

1.2 相关系数矩阵 计算标准化样本每 2 指标间的相关系数,得出相关系数矩阵 R[7]

$$\boldsymbol{R} = \frac{1}{n-1}\boldsymbol{Z}\boldsymbol{Z} = (r_{uv})_{p\times p}$$

$$u = 1,2,\cdots,p; v = 1,1,\cdots,p。 \tag{2}$$

其中

$$r_{uv}=\frac{1}{n-1}=\sum_{i=1}^{n}Z_{iu}Z_{iv}=\frac{1}{n-1}=$$

$$\sum_{i=1}^{n}[(x_{iu}-\bar{x}_u)^2/S_u][(x_{iv}-\bar{x}_v)^2/S_v]。\qquad(3)$$

1.3 计算主成分 由特征式$|\lambda I-R|=0$求得p个特征根,将其按大小排列为$\lambda_1\geqslant\lambda_2\geqslant\cdots\geqslant\lambda_p\geqslant0$。它们是主成分的方差,其大小描述了各对应的主成分对原始样本的权重。由特征方程式($|\lambda I-R|=0$)求得每个特征根对应的特征向量$l_{g1},l_{g2},\cdots,l_{gp}{}^{\mathrm{T}}$。

通过特征向量将标准化的指标转化为主成分[7]:

$$F_g=Z\times L_g\qquad g=1,2,\cdots,p。\qquad(4)$$

F_1为第1主成分,$\cdots$,F_p为第p主成分。

1.4 确定主成分个数 主成分个数等于原始指标个数。为减少计算量并降低维数,一般根据主成分方差累计贡献率大于80% ~90%确定取用的主成分个数k,即

$$\sum_{g=1}^{k}\lambda_g/\sum_{g=1}^{p}\lambda_g\geqslant80\%\sim90\%。\qquad(5)$$

1.5 对k个主成分综合评价 对于每一个样本,先求前k个主成分的值,即

$$F_{ig}=Z_{i1}l_{g1}+Z_{i2}l_{g2}+\cdots+Z_{ip}l_{gp}$$

$$i=1,2,\cdots,n;g=1,2,\cdots,k。\qquad(6)$$

然后对前k个主成分进行加权求和,即得每个样本的综合主成分值,即综合评价值。每个主成分的权重为该主成分的方差贡献率,即$\lambda_g/\sum_{g=1}^{p}\lambda_g$,最终对样本$i$的综合评价值为

$$F_i=\sum_{g=1}^{k}(\lambda_g/\sum_{g=1}^{p}\lambda_g)\ F_{ig}。\qquad(7)$$

2 评估实例

采用主成分分析法既可以对不同的期刊进行横向比较评价,也可以对某一期刊进行动态评价。

本文以2006年版《中国科技期刊引证报告(核心版)》[8]中"理工大学,工业综合类期刊"的总被引频次前10名及影响因子前10名的期刊为样本,对共计16种学术类科技期刊进行评价,具体数据见表1。采用SPSS 13 for Windows分析软件做分析计算[9]。

2.1 计算相关系数矩阵与特征根 由于给出的12项

表1 2005年16种科技期刊主要被引用指标和来源指标

期刊编号 i	总被引频次 x_{i1}	影响因子 x_{i2}	即年指标 x_{i3}	他引率 x_{i4}	引用刊数 x_{i5}	扩散因子 x_{i6}
01(Z17/Y5)	491	0.454	0.055	0.85	194	39.51
02(Z13/Y1)	638	0.801	0.130	0.72	159	24.92
03(Z8/Y9)	740	0.386	0.038	0.94	347	46.89
04(Z5/Y4)	882	0.492	0.078	0.77	301	34.13
05(Z10/Y18)	706	0.317	0.028	0.88	319	45.18
06(Z30/Y2)	339	0.588	0.087	0.40	95	28.02
07(Z3/Y37)	997	0.246	0.046	0.91	390	39.12
08(Z1/Y16)	1 589	0.328	0.011	0.96	528	33.23
09(Z2/Y19)	1 194	0.316	0.052	0.88	474	39.70
10(Z6/Y28)	789	0.268	0.040	0.93	330	41.83
11(Z11/Y3)	690	0.535	0.034	0.79	173	25.07
12(Z7/Y7)	753	0.413	0.048	0.75	322	42.76
13(Z4/Y7)	996	0.413	0.048	0.93	383	38.45
14(Z15/Y6)	514	0.436	0.058	0.67	217	42.22
15(Z9/Y28)	734	0.268	0.027	0.85	333	45.37
16(Z36/Y10)	274	0.385	0.053	0.42	79	28.83

期刊编号 i	来源文献量 x_{i7}	参考文献量 x_{i8}	平均引文数 x_{i9}	地区分布数 x_{i10}	机构分布数 x_{i11}	基金论文比 x_{i12}
01(Z17/Y5)	183	1 566	8.56	7	17	0.81
02(Z13/Y1)	161	1 170	7.27	11	23	0.47
03(Z8/Y9)	183	1 528	8.35	3	7	0.68
04(Z5/Y4)	319	3 267	10.24	2	2	1.00
05(Z10/Y18)	503	3 638	7.23	20	95	0.64
06(Z30/Y2)	173	2 044	11.82	21	60	0.94
07(Z3/Y37)	461	2 649	5.75	16	50	0.74
08(Z1/Y16)	438	3 438	7.85	17	36	0.78
09(Z2/Y19)	481	3 855	8.01	15	45	0.66
10(Z6/Y28)	354	2 935	8.29	6	10	0.49
11(Z11/Y3)	264	2 203	8.34	15	53	0.78
12(Z7/Y7)	400	2 637	6.59	22	81	0.69
13(Z4/Y7)	333	2 640	7.93	13	19	0.80
14(Z15/Y6)	417	3 752	9.00	3	7	0.66
15(Z9/Y28)	482	4 443	9.22	13	50	0.75
16(Z36/Y10)	150	1 066	7.11	12	20	0.29

注:Z8表示在该类期刊中总被引频次排名第8位,Y3表示在该类期刊中影响因子排名第3位。

指标都是越大越好,所以不需要同趋势化处理。SPSS 在调用 Factor Analyze 过程进行分析时,SPSS 会自动对原始数据进行标准化处理,所以,在得到计算结果后的变量都是经过标准化处理后的变量,但 SPSS 不会直接给出标准化后的数据,如需要得到标准化数据,则需调用 Descriptives 过程进行计算[10]。

通过 SPSS 的 Factor Analyze 计算,可得到相关系数矩阵和特征根。相关系数矩阵见表 2,由大到小排列的主成分特征根值、方差(贡献率)、累计方差(累计贡献率)见表 3。

表 2　相关系数矩阵

指标	Z_1	Z_2	Z_3	Z_4	Z_5	Z_6	Z_7	Z_8	Z_9	Z_{10}	Z_{11}	Z_{12}
Z_1	1.000	-0.392	-0.422	0.722	0.923	0.165	0.592	0.491	-0.275	0.144	0.040	0.254
Z_2	-0.392	1.000	0.827	-0.472	-0.621	-0.694	-0.672	-0.558	0.294	-0.061	-0.150	0.057
Z_3	-0.422	0.827	1.000	-0.505	-0.551	-0.517	-0.547	-0.467	0.234	-0.162	-0.247	-0.060
Z_4	0.722	-0.472	-0.505	1.000	0.811	0.545	0.518	0.392	-0.378	-0.152	-0.043	0.165
Z_5	0.923	-0.621	-0.551	0.811	1.000	0.521	0.722	0.607	-0.309	0.091	0.085	0.213
Z_6	0.165	-0.694	-0.517	0.545	0.521	1.000	0.556	0.505	-0.168	-0.175	0.081	0.037
Z_7	0.592	-0.672	-0.547	0.518	0.722	0.556	1.000	0.903	-0.274	0.282	0.431	0.145
Z_8	0.491	-0.558	-0.467	0.392	0.607	0.505	0.903	1.000	0.129	0.095	0.258	0.313
Z_9	-0.275	0.294	0.234	-0.378	-0.309	-0.168	-0.274	0.129	1.000	-0.234	-0.225	0.562
Z_{10}	0.144	-0.061	-0.162	-0.152	0.091	-0.175	0.282	0.095	-0.234	1.000	0.893	0.053
Z_{11}	0.040	-0.150	-0.247	-0.043	0.085	0.081	0.431	0.258	-0.225	0.893	1.000	0.094
Z_{12}	0.254	0.057	-0.060	0.165	0.213	0.037	0.145	0.313	0.562	0.053	0.094	1.000

表 3　特征根、方差和累计方差

主成分序号	特征根值	方差	累计方差
		%	
1	5.292	44.101	44.101
2	2.046	17.050	61.151
3	1.721	14.340	75.491
4	1.242	10.351	85.842
5	0.650	5.415	91.257
6	0.540	4.498	95.754
7	0.225	1.874	97.629
8	0.180	1.501	99.130
9	0.071	0.592	99.722
10	0.028	0.236	99.958
11	0.004	0.033	99.990
12	0.001	0.010	100.000

2.2　确定主成分个数及表达式　由表 2 可看出,前 4 个主成分方差累计贡献率达 85.842,故本文选取 4 个主成分,即 $k=4$。由式(2)可计算得 4 个主成分的特征向量,再由式(6)可得如下 4 个主成分的表达式:

$$F_1 = 0.326Z_1 - 0.353Z_2 - 0.33Z_3 + 0.333Z_4 + 0.393Z_5 + 0.294Z_6 + 0.386Z_7 + 0.331Z_8 - 0.148Z_9 + 0.082Z_{10} + 0.13Z_{11} + 0.076Z_{12}; \quad (8)$$

$$F_2 = -0.082Z_1 + 0.034Z_2 - 0.031Z_3 - 0.243Z_4 - 0.117Z_5 - 0.172Z_6 + 0.129Z_7 - 0.004Z_8 - 0.185Z_9 + 0.657Z_{10} + 0.633Z_{11} - 0.078Z_{12}; \quad (9)$$

$$F_3 = 0.064Z_1 + 0.155Z_2 + 0.087Z_3 - 0.108Z_4 + 0.011Z_5 - 0.091Z_6 + 0.095Z_7 + 0.319Z_8 + 0.617Z_9 + 0.08Z_{10} + 0.102Z_{11} + 0.66Z_{12}; \quad (10)$$

$$F_4 = 0.546Z_1 + 0.358Z_2 + 0.244Z_3 + 0.289Z_4 + 0.298Z_5 - 0.444Z_6 - 0.088Z_7 - 0.201Z_8 - 0.234Z_9 + 0.137Z_{10} - 0.07Z_{11} + 0.135Z_{12}。 \quad (11)$$

2.3　计算主成分与综合评价值　将标准化后的样本数据代入式(8)~(11),可得各样本的 4 个主成分值 F_{i1},F_{i2},F_{i3},F_{i4},由式(7)可计算出各样本的综合主成分值 F_i,即评价排名。计算结果见表 4。

表 4　主成分计算结果

样本编号$_i$	F_{i1}	F_{i2}	F_{i3}	F_{i4}	F_i
01(Z17/Y5)	-1.444	-1.178	-0.105	-0.325	13
02(Z13/Y1)	-4.127	0.187	-1.106	2.154	15
03(Z8/Y9)	-0.025	-2.23	-0.998	-0.285	12
04(Z5/Y4)	-0.729	-2.174	2.166	0.451	10
05(Z10/Y18)	2.326	2.14	-0.257	-1.011	2
06(Z30/Y2)	-3.728	1.775	2.896	-0.585	14
07(Z3/Y37)	2.102	0.735	-0.972	0.614	4
08(Z1/Y16)	3.127	0.052	0.423	1.884	1
09(Z2/Y19)	2.425	0.191	0.211	0.596	3
10(Z6/Y28)	0.961	-1.493	-1.172	-0.705	8
11(Z11/Y3)	-1.232	1.017	0.457	0.671	9
12(Z7/Y7)	0.971	2.211	-0.469	-0.092	6
13(Z4/Y7)	0.885	-0.683	0.126	0.905	7
14(Z15/Y6)	-0.383	-1.421	0.363	-1.557	11
15(Z9/Y28)	2.221	0.076	0.963	-1.544	5
16(Z36/Y10)	-3.348	0.795	-2.527	-1.172	16

3　评价结果分析

3.1　评价结果　从表 2 的数据 F_i 可以看出,16 种期刊的学术综合评价排名为:期刊 8,5,9,7,15,12,13,10,11,4,14,3,1,6,2,16。由此可以看出,总被引频次的排名比影响因子排名更接近于主成分分析法排名,即绝对量引证指标比相对量引证指标的评价作用大。

3.2 因子负荷分析 利用主成分因子负荷矩阵可以分析主成分与各指标的相关性。主成分因子负荷矩阵由式(12)确定[10-11]:

$$P(F_g,x_j)=(\lambda_g)^{1/2}l_{gj}\quad g=1,2,\cdots,k;j=1,2,\cdots,p。\ (12)$$

计算实例得出的主成分因子负荷矩阵如表 5 所示。

表 5 主成分因子负荷矩阵

主成分	总被引频次 x_{i1}	影响因子 x_{i2}	即年指标 x_{i3}	他引率 x_{i4}	引用刊数 x_{i5}	扩散因子 x_{i6}
1	0.749	-0.811	-0.759	0.765	0.903	0.676
2	-0.118	0.048	-0.044	-0.347	-0.167	-0.246
3	0.084	0.203	0.114	-0.142	0.015	-0.120
4	0.608	0.399	0.272	0.322	0.332	-0.495
主成分	来源文献量 x_{i7}	参考文献量 x_{i8}	平均引文数 x_{i9}	地区分布数 x_{i10}	机构分布数 x_{i11}	基金论文比 x_{i12}
1	0.889	0.762	-0.341	0.188	0.298	0.174
2	0.184	-0.006	-0.265	0.940	0.905	-0.111
3	0.125	0.418	0.810	0.105	0.134	0.886
4	-0.098	-0.224	-0.261	0.153	-0.078	0.151

由表 5 可以看出:与总被引频次、他引率、引用刊数、来源文献量及参考文献量在第 1 主成分有较高载荷,说明第 1 主成分反映了引用与被引用总量方面的信息;地区分布数和机构分布数在第 2 主成分上有较高载荷,说明第 2 主成分基本反映了作者扩散方面的信息;平均引文数和基金论文比在第 3 主成分上有较高载荷,说明第 3 主成分反映了文献质量方面的信息;总被引频次在第 4 主成分上有较高载荷,说明第 4 主成分基本反映了被引总量方面的信息。这 4 个主成分基本包含了原指标的信息。

影响因子在第 1 主成分上有较高的负载荷,说明与总被引频次、他引率、引用刊数及扩散因子等是负相关的,这似乎与常规解释有矛盾;但分析期刊样本的指标数据可以看出,影响因子最高的几个刊物的总被引频次和他引率反而较低,自引较高导致了影响因子的升高。从 4 个主成分综合来看,影响因子所占的权重很小,评价结果更接近于总被引频次的排名,说明主成分分析法可以有效的消除自引较高带来的评价失真问题,更适合我国科技期刊的现状。

4 结论

主成分评价方法通过相关系数矩阵的特征向量将评价指标线性变化成彼此独立的主成分,通过特征值确定主成分的取用维数和权重,对主成分加权求和得到评价值。使用该评价方法对科技期刊进行评价的优点是:可以消除由于指标间的相关性带来的评价偏差,降低计算维数,从而降低指标选择难度;此外,可以消除人为确定指标权重引起的弊病,使评价结果更具客观性和准确性。

采用主成分分析法对科技期刊进行评估,其优点是可以考虑尽可能多的计量指标,不用关心指标之间的相关重叠性,也不用人为去确定指标的权重,通过主成分分析可以降维计算,去除指标之间的相关重叠因素,根据指标的变化幅度确定权重。通过主成分因子负荷矩阵分析,可以看出各指标对评价值的影响程度,这对今后评价指标的选取可以提供定量的参考。另外,采用主成分分析可以有效消除自引过高导致影响因子失真对评价带来的负面影响。

5 参考文献

[1] 赵惠祥,曲俊延,张全福. 论我国科技期刊评估的现状与发展[J]. 编辑学报,2000,12(2):90-93

[2] 张玉华,潘云涛,马峥. 科技论文评估方法研究[J]. 编辑学报,2004,16(4):243-244

[3] 刘明寿,马峥,潘云涛. 学术类科技期刊影响力归一化评判定量模型的构建[J]. 编辑学报,2004,16(6):405-406

[4] 何学锋,彭超群. 科技期刊学术影响力的动态评估模型[J]. 编辑学报,2002,14(4):238-240

[5] 金碧辉,汪寿阳,任胜利,等. 论期刊影响因子与论文学术质量的关系[J]. 中国科技期刊研究,2000,11(4):202-205

[6] 孙文爽,陈兰祥. 多元统计分析[M]. 北京:高等教育出版社,1994

[7] 韩伟,李钢. 主成分分析在地区科技竞争力评测中的应用[J]. 数理统计与管理,2006,25(5):512-517

[8] 中国科学技术信息研究所. 2006 年版中国科技期刊研究引证报告:核心版[R]. 北京:科学技术文献出版社,2007

[9] 张力. SPSS 13.0 在生物统计中的应用[M]. 厦门:厦门大学出版社,2006

[10] 刘顺忠. 数理统计理论、方法、应用和软件计算[M]. 武汉:华中科技大学出版社,2005

[11] 杨宇音,赵雅明,曲立敏. 因子分析法在大学生综合排名中的应用[J]. 贵州工业大学学报:自然科学版,2005,34(1):9-13

(2007-09-02 收稿;2007-10-10 修回)