对主成分分析法运用中十个问题的解析

林海明

(广东商学院 经济贸易与统计学院,广州 510320)

摘 要:主成分分析的应用十分广泛,但由于有关文献没有完整、系统地阐述主成分分析的综合评价步骤,以至应用主成分分析法进行综合评价时出现一些问题和困难。据归纳,有 10 个问题经常出现。本文对这些进行了逐一解析,提出了主成分分析法使用中的建议与综合评价步骤,并以实例说明它的有效性。

关键词:主成分分析;综合评价;步骤;问题;解析

中图分类号:O212 文献标识码:A 文章编号:1002-6487(2007)08-0016-03

1 主成分分析法在综合评价中出现的一些问题

在社会经济、管理、自然科学等众多领域的多指标体系中,如节约型社会指标体系、生态环境可持续型指标体系、和谐社会指标体系、投资环境指标体系等,主成分分析法常被应用于综合评价与监控。

主成分分析法的理论与计算是较成熟的,但在解决实际问题中,主成分分析法的应用并没有达到较成熟状态。据归纳,一些使用者在应用主成分分析法进行综合评价时,出现以下 10 个问题不明确:

①原始数据没有正向化,有何影响? 如何正向化?

②原始变量表示主成分的系数平方和不是 1 对吗?

③主成分分析法的主成分正交旋转后会怎样?

④主成分分析法的主成分有必要回归计算吗?

⑤主成分分析法与正交因子分析法能混合使用吗?

⑥何时使用主成分分析法?

⑦主成分分析法有时会丢失一些原始变量的原因是什么?

⑧主成分如何命名,并能保持原始变量与多个主成分的内在关系?

⑨前 m个主成分仍然是多因素,仅用综合主成分进行综合分析客观吗?

⑩综合评价结果,如何能深入到决策相关性程度?

有关文献并没有清楚地阐述上述问题,以至应用主成分分析法进行综合评价时,不易把握。本文除了逐一解析上述问题外,还给出了主成分分析法使用中的建议与综合评价步骤,并以实例说明它的有效性。

2 主成分分析法综合评价中 10 个问题的解析

问题①解析:主成分分析法是一种综合评价方法,是通过样品的相对位置,比较找出样品的优势、不足、差距状况及其原因,如果指标体系方向不是正向化的,便得不出有效结论。因此,分析中必须对指标体系中的强度逆向指标、适度指标进行正向化。

强度逆向指标 x_j 正向化公式[3]:

$$\begin{cases} 1/x_j & x_{ij}>0 \\ 1/(\max\limits_i |x_{ij}|+x_j+1) & x_{ij} \text{ 中有 } 0 \text{ 或有负数} \end{cases}$$

适度指标 x_j 正向化公式[3]:$1/(|x_j-E|+1)$,E 为理想值。这里 x_{ij} 为第 i 个样品第 j 个指标的观测值。

设 $X=(x_1,\cdots,x_p)^T$(T 为转置符号)为正向化、标准化随机变量向量($p\geq 2$),R 为相关系数矩阵,秩$(R)=r$(R 的非零特征根个数),R 的特征值为 λ_1、λ_2、…、λ_m、0,$\lambda_1\geq\lambda_2\geq\cdots\geq\lambda_r>0$,前 m 个单位正交特征向量矩阵 $A_m=(\alpha_1,\ldots,\alpha_m)=(\alpha_{ij})_{p\times m}$,主成分向量 $F_m=(f_1,\cdots,f_m)^T$。

性质[1] 变量 X 与主成分 f_i 的相关系数 $b_i^0=\sqrt{\lambda_i}\,\alpha_i$,即变量 X 与主成分 F_m 的相关系数阵(初始因子载荷阵):$B_0=(\sqrt{\lambda_1}\,\alpha_1,\ldots,\sqrt{\lambda_m}\,\alpha_m)=(b_1^0,\ldots,b_m^0)$。

主成分分解(Hotelling,1933):$f_i=\alpha_i'X$,$Varf_i=\lambda_i$,$i=1,\cdots,m$。

问题②解析:主成分 f_i 中变量 x 的系数向量 α_i' 是(R 的特征值 λ_i 的相应)单位正交特征向量,即主成分中变量 X 的系数平方和全部是 1,$\alpha_i'\alpha_j=0$,$i\neq j$,如果不符合这个条件就是错的,同时有:

结论 1 $f_i=(b_i^0/\sqrt{\lambda_i})'X$,$i=1,\cdots,m$。

问题③解析:主成分分解的公式对主成分是无旋转的,即主成分分析法中对主成分没有旋转。如果对主成分进行正交旋转,原始变量的线性组合会发生改变,该线性组合不能达到方差的最大化,这已不是主成分分析的结果了。

问题④解析:主成分分解的公式是直接的表达式,主成分

基金项目:广州市哲学社会科学规划资助项目(06YZ140);广东商学院经济贸易与统计学院 2006 年资助课题

分析法中的主成分分解是完全没有必要进行回归计算的。

问题⑤解析：主成分 f_i 与正交因子 z_i 有，$Varf_i=\lambda_i$，$Varz_i=1$，主成分分析法中没有旋转，即主成分 f_i 与正交因子 z_i 的取值范围、旋转方向不同，故样品计量值不相等、两种方法应用条件不相同，混淆在一起是样品计量值、旋转方向交替错误(具体异同见文献[4])，故不论何条件，有：

结论 2　主成分分析法与正交因子分析法不能混淆使用。

问题⑥解析：主成分分析法的优点是，对原始变量具有综合性的降维能力；如果 B_0 中每行的系数绝对值往 0 或 1 靠近得较多(与旋转后因子载荷阵 B_0C[1]比较)，即主成分命名、解释原始变量清晰，同时主成分 F_m 解释原始变量 X 的信息误差($\sum_{i=m+1}^{r}\lambda_i$)达到最小，使用主成分分析法最好。

结论 3　当原始变量之间有相关性，B_0 中每行的系数绝对值往 0 或 1 靠近得较多 (与旋转后因子载荷阵 B_0C 比较)，则使用主成分分析法。

问题⑦解析：m按某个累积贡献率确定，当 $\lambda_m>1$、第 m+1 个单位特征向量 α_{m+1} 的第 t 个元素 $a_{tm+1}\geq0.9$ 时，则 F_m 中不能解释原始变量 x_t，这是主成分分析法有时会丢失一些原始变量解释的主要原因。因为初始因子载荷阵 B_0 是变量 X 与主成分 F_m 的相关系数阵，如果 B_0 每行中至少有一个系数绝对值足够大(≥0.5)，则主成分 F_m 不会丢失原始变量的解释，故有：

结论 4　如果 B_0 每行中至少有一个系数绝对值足够大(≥0.5)，m便是主成分的确定个数，此时主成分 F_m 不会丢失原始变量，能达到最大限度降维的目的。

问题⑧解析：B_0 的第 i 列 b_i^0 是原始变量 X 与主成分 f_i 的相关系数，绝对值大(≥0.5)的对应变量与 f_i 相关性高，而以 f_i 中 X 的系数向量 α_i 对主成分 f_i 进行命名不能判断出原始变量 X 与主成分 f_i 的相关性，这样主成分分析法有时会失去一些原始变量与多个主成分的内在关系，因此有：

结论 5　B_0 的第 i 列绝对值大(≥0.5)的对应原始变量归为主成分 f_i 一类，并由这些变量对 f_i 命名，这样主成分分析法不会失去一些原始变量与多个主成分的内在关系。

问题⑨解析：前 m个主成分的样品值反映的是 n个样品在 m个主成分中的相对位置，表现出样品的优势、劣势、差距状况等，且没有相关性，分析问题可靠性高，仅用综合主成分进行综合分析失去的就是这些内在因素，以致不客观，因此，应将前 m个主成分、综合主成分的样品值结合起来分析才是客观的、可靠的。但样品数量较多，逐个分析看不出共性规律，为此，对无相关性的前 m个主成分样品值进行聚类分析，并按综合主成分值相应顺序给出分类，便找出了样品之间具有可靠性的共性规律，故有：

结论 6　对无相关性的前 m个主成分样品值进行聚类分析，按综合主成分值相应顺序给出分类，能可靠地反映样品之间的共性规律，便于客观、可靠地进行样品共性的分析。

问题⑩解析：主成分分析、聚类分析给出了样品客观、可靠的个性与共性特征，但主成分有综合性，决策相关性有待与原始指标结合起来。注意到主成分是按相关性高的原始变量进行归类命名的，故将相应原始变量对应替换为相应主成分进行分析，便得出了可靠的决策相关性分析，达到了数据分析的目的。

结论 7　将主成分对应替换为相应原始变量进行数据分析，得出的就是客观、可靠的决策相关性分析。

3　主成分分析法综合评价步骤

①指标的正向化(单独计算)；

②指标数据标准化(SAS 软件自动执行)；

③指标之间的相关性判定：用 SAS 软件的 Correlations (相关系数矩阵 R) 判定，若变量间有相关性，主成分分析继续；否则，直接进行逐个指标分析，用 $\sum_{i=1}^{p}$ 进行综合分析(x_i 是正向化、标准化的)；

④求相关系数矩阵 R 的特征值 λ_i、单位正交特征向量矩阵 α_i，变量 X 与主成分 F_m 的相关系数阵 (初始因子载荷阵，SAS 软件因子分析过程命令中的 Factor Pattern)；

⑤与旋转后因子载荷阵 B_0C (SAS 软件因子分析过程命令中的 Rotated Factor Pattern)比较，若 B_0 中每行的系数绝对值往 0、1 靠近较多，则用主成分分析法(结论 3)；

⑥确定主成分个数 m：以 B_0 每行中至少有一个系数绝对值足够大(≥0.5)确定(结论 4)；

⑦主成分 f_i 的命名：将 B_0 的第 i 列 b_i^0 绝对值大(≥0.5)的对应变量归为 f_i 一类，由这些变量对主成分 f_i 进行命名(结论 5)；

⑧前 m个主成分函数：$f_i=\alpha_i'X=(b_i^0/\sqrt{\lambda_i})'X$，i=1,…,m(主成分 X 的系数平方和是 1、无旋转、无回归，z_i^0 为未旋转因子得分，结论 1)；

⑨综合主成分函数 $F_{综}=\sum_{i=1}^{m}(\lambda_i/p)f_i$；

⑩对前 m个主成分的样品值进行排序，用 SAS 软件 iml 模块计算综合主成分 $F_{综}$ 的样品值并排序；

⑪用前 m个主成分的样品值做聚类分析，按综合主成分值相应顺序给出分类结果(结论 5)；

⑫结合前 m个主成分样品值的聚类分析结果，主成分、综合主成分样品值和排序，主成分、综合主成分与原始变量的对应关系，进行优势、劣势、潜力、差距状况和原因等的综合评价，给出决策相关性建议(结论 5、结论 6)。

4　实证应用：安徽省各地市经济发展综合评价与建议

现以文献[2]数据为例，指标选取为：X_1- 城镇单位在岗职工平均工资(元)，X_2- 固定资产投资(万元)，X_3- 进口总额(万美元)，X_4- 社会消费品零售总额(万元)，X_5- 工业增加值(亿元)，X_6- 财政收入(亿元)；城市为 17 个：合肥市、淮北市、亳州市、宿州市、蚌埠市、阜阳市、淮南市、滁州市、六安市、马鞍山市、巢湖市、芜湖市、宣城市、铜陵市、池州市、安庆市和黄山市。原始数据见文献[2]。

①指标都是正向的，直接使用；

②调用 SAS 软件的主成分分析过程命令，输入原始数

据，数据标准化自动执行；

③变量有相关性(相关系数矩阵 R 略)，继续；

④相关系数矩阵 R 的特征值：λ_1=4.6412321、λ_2=1.1006631、…，相应单位正交特征向量矩阵(见第 8 步 f_1、f_2、…表达式中 X 的系数)，初始因子载荷矩阵 B_0(表 1)；

⑤与旋转后因子载荷阵 B_0C（SAS 软件因子分析过程命令中的 Rotated Factor Pattern)比较，B_0 中每行的系数绝对值往 0、1 靠近较多，故用主成分分析法；

⑥B_0 每行中至少有一个系数绝对值足够大（≥0.5)，所以 m=2，前两个主成分的累计方差贡献率已达到 95.7%；

表 1 初始因子载荷阵

	Factor Pattern	
	Factor1	Factor2
x_1	0.57994	0.79515
x_2	0.98026	- 0.04923
x_3	0.95613	- 0.16804
x_4	0.76168	- 0.60047
x_5	0.92908	0.27340
x_6	0.99320	- 0.04919

⑦第一个主成分 f_1 与 X_2- 固定资产投资，X_3- 进口总额，X_4- 社会消费品零售总额，X_5- 工业增加值，X_6- 财政收入十分显著地正相关，故称 f_1 为生产总量成分；第二个主成分 f_2 与 X_1- 城镇单位在岗职工平均工资十分显著地正相关，注意到它受 X_4- 社会消费品零售总额的负影响也很大，故称 f_2 为生活成分。内在关系：社会消费品零售总额 X_4 对生产总量成分 f_1 有较大的促进作用、对生活成分 f_2 有较大的负影响，城镇单位在岗职工平均工资 X_1 对经济总量成分 f_1 有正常的促进作用、对生活成分 f_2 是直接的正的影响；

⑧主成分函数(x_i 为 X_i 的标准化变量)：

f_1=0.269x_1+0.455x_2+0.444x_3+0.354x_4+0.431x_5+0.461x_6

f_2=0.758x_1- 0.047x_2- 0.16x_3- 0.572x_4+0.26x_5- 0.047x_6

⑨综合主成分函数：$F_{综}$=(4.6412321f_1+1.1006631f_2) /6；

⑩主成分、综合主成分样品值及排序见表 3(综合主成分值 SAS 软件 iml 模块计算)；

⑪调用 SAS 软件的聚类分析过程命令，选用欧氏距离和类平均法，通过表 2 两个主成分 f_1、f_2 的样品值对 17 个城市进行聚类。取分类阈值为 1.5 时，分成五类，聚类结果如下：

第一类：合肥市；第二类：马鞍山市；第三类：芜湖市；第四类：淮南市，淮北市，宣城市，铜陵市，黄山市和池州市；第五类：蚌埠市，安庆市，滁州市，巢湖市，六安市，阜阳市，宿州市和亳州市。

⑫结合前 2 个主成分样品值的聚类分析结果，主成分、综合主成分样品值和排序，主成分、综合主成分与原始变量的对应关系，进行优势、劣势、潜力、差距状况和原因等的综合评价，给出决策相关性建议。评价中注意：主成分函数 f_1、f_2 表明了 X_4- 社会消费品零售总额一方面对总量成分 f_1 有促进作用(影响系数为 0.354)，另一方面对工资与消费成分 f_2 有负影响作用(影响系数为- 0.572)。

第一类的合肥市综合主成分 $F_{综}$值排第 1(5.307)。其生产总量成分 f_1 得分值排第 1(7.113)，优势相当明显，可生活成分 f_2 排在倒数第 2(- 1.058)。原因及问题：生产总量成分 f_1 中 X_4- 社会消费品零售总额为 2397739 万元列第 1，生活成分 f_2 中 X_1- 城镇单位在岗职工平均工资为 162369 元列第 3，即合肥市是生产总量、消费高但平均工资不是太高的城市。综合函数值中，生产总量成分 f_1 综合值为 5.502，而生活成分 f_2 有综合抵减值 0.194（抵减率 3.526%），带来了不良影响。

表 2 主成分值、综合主成分值及排序

城市	f1	序	f2	序	F 综	序
合肥市	7.11276	1	- 1.05753	16	5.3071	1
马鞍山	1.84384	3	2.78539	1	1.937244	2
芜湖市	2.30617	2	0.31395	7	1.841504	3
淮南市	- 0.12616	6	1.50569	2	0.17862	4
安庆市	0.42001	4	- 0.84506	13	0.16987	5
蚌埠市	0.22308	5	- 0.39238	10	0.100581	6
宣城市	- 0.52566	7	0.09539	8	- 0.38912	7
淮北市	- 0.77301	9	0.79739	3	- 0.45168	8
铜陵市	- 0.80935	11	0.62634	4	- 0.51117	9
滁州市	- 0.61361	8	- 0.71968	12	- 0.60667	10
巢湖市	- 0.86366	12	- 0.33622	9	- 0.72975	11
阜阳市	- 0.78570	10	- 1.35248	17	- 0.85587	12
六安市	- 0.91152	13	- 0.88723	15	- 0.86785	13
黄山市	- 1.48609	14	0.39334	5	- 1.07739	14
宿州市	- 1.54819	15	- 0.44436	11	- 1.2791	15
池州市	- 1.79223	17	0.36983	6	- 1.31852	16
亳州市	- 1.67067	16	- 0.85238	14	- 1.44869	17

合肥市在保持生产总量成分 f_1 中 X_2- 各市固定资产投资(第 1)，X_3- 各市进口总额（第 1)，X_4- 社会消费品零售总额(第 1)，X_5- 各市工业增加值(第 1)，X_6- 财政收入(第 1)优势的同时，如果能够进一步结合劳动生产率、成本费用利润率(此说明文献[2]漏选了指标：劳动生产率、成本费用利润率，致使工资与消费的协调性无法分析）协调生活成分 f_2 中 X_1- 城镇单位在岗职工平均工资与 X_4- 社会消费品零售总额的良性关系，将对经济有更大的促进作用。

第二类的马鞍山市、第三类的芜湖市综合评价、建议方法与第一类的合肥市类似，此略。

第四类城市淮南市、宣城市、淮北市、铜陵市、黄山市和池州市综合主成分 F 综值排名依次是 4、7、8、9、14 和 16。它们的生产总量成分 f_1 排名依次是 6、7、9、11、14、17，均低于平均水平，生活成分 f_2 排名依次是 2、8、3、4、5、6，均高于平均水平。共性原因为该类城市生活成分 f_2 中 X_1- 城镇单位在岗职工平均工资列前 10 名、生产总量成分 f_1 中 X_4- 社会消费品零售总额列第 10 之后，即该类城市是工资较高、生产总量水平低、消费不足的城市。个性原因及问题：如淮北市生活成分 f_2 中 X_1- 城镇单位在岗职工平均工资(第 4：13379 元)、X_4- 社会消费品零售总额(第 14：456100 万元)，生产总量成分 f_1 排名差异大（第 9），其中 X_2- 各市固定资产投资为（第 14：566257 万元)、X_3- 各市进口总额为（第 13：4744 万美元)、X_6- 财政收入为(第 9：202637 亿元)等。

以上分析及结论，找到了研究对象的优势、不足、差距状况和原因等，用具有可控性的原始指标给出了可靠的决策相关性建议，对指标体系选取的代表性具有可验证性，使主成分分析法的应用得到深入。

参考文献：

[1]于秀林，任雪松编著.多元统计分析[M].北京：中国统计出版社，1999.5.

[2]宋马林.安徽省各地市经济发展评价[J].统计教育，2006,(4).

[3]陈迪红,李华中,杨湘豫.行业景气指数建立的方法选择及实证分析[J].系统工程，2003,(4).

[4]林海明,张文霖.主成分分析与因子分析的异同和 SPSS 软件[J].统计研究，2005,(3).

（责任编辑/李友平）