主成分分析综合评价应该注意的问题*

林海明　杜子芳

内容提要: 将主成分分析用于多指标的综合评价较普遍,但因缺乏应用条件的考虑而导致评价结果不具合理性甚至错误,故应深入研究其应用条件。本文应用因子分析法因子载荷阵的简单结构、加权算术平均数的合理性,得出主成分分析综合评价的应用条件是: 指标是正向、标准化的,主成分载荷阵达到更好的简单结构,主成分正向,主成分与变量显著相关; 并结合 2010 年广东省各市对外贸易国际竞争力的评价实例提出了一些建议。

关键词: 主成分分析; 综合评价; 条件; 简单结构
中图分类号: C812　　**文献标识码:** A　　**文章编号:** 1002 - 4565(2013) 08 - 0025 - 07

Some Problems in Comprehensive Evaluation in the Principal Component Analysis

Lin Haiming & Du Zifang

Abstract: The principal component analysis is widely used in comprehensive evaluation, but sometimes, the results of the principal component analysis evaluation index is unreasonable and even wrong. Therefore, the construction conditions of the principal component analysis evaluation index needs to be further studied. Applying the simple structure of factor loading matrix in factor analysis method and the rationality of weighted arithmetic mean, we get the construction conditions for the principal component analysis evaluation index: the indicators are positive and standardized; the principle component loading matrix gets a better simple structure; the principle component is positive; the principle components and variables are significantly related. Also, we propose some suggestions through empirical study.

Key words: Principal Component Analysis; Comprehensive Evaluation; Conditions; Simple Structure

一、引言

将主成分分析用于多指标(变量) 的综合评价较为普遍。苏为华(2012) 认为,从评价方法看,应用与关注最多的方法是多元统计方法与运筹优化方法。文献数量最多的前十位分别是: 聚类分析、因子分析、主成分分析、AHP 法、模糊评价、判别分析、……。邱东(1990) 认为,主成分分析用于多指标综合评价的优点在于: 消除评价指标间的相关影响,有助于更客观地描述样品的相对地位,也消去了选择合成方法的工作。主成分分析法用于多指标综合评价时,伴随数学变换过程生成了信息量权数和系统效应权数,比人为确定权数的工作量少些,有助于保证客观性。信息量权数还有助于提高综合评价的区分效度。评价指标间相关程度较高时能得到较为理想的主成分结果。主成分分析综合评价的步骤可归纳为: ①原始指标数据的标准化; ②确定主成分个数; ③解释主成分含义; ④用主成分及其方差贡献率构造主成分综合评价函数 $Y_{综}$; ⑤计算 $Y_{综}$的样品值并给出样品的排序。

众所周知,合理的结果取决于方法应用条件的满足。如果不能满足方法的某一个应用条件,那么结果或结论可能都不具有合理性,相应的建议也就没有意义,故方法应用条件的明确及其满足是重要的。主成分分析综合评价因缺乏应用条件,导致了较多的质疑或争议,如: ①何平(2005) 认为,主成分的具体涵义是什么,许多评价没有给予较为清楚的解释,从而影响到评价结果的可信度。②王惠文

* 本文获教育部人文社会科学研究规划基金项目,项目号: 2009YJA910002; 教育部人文社会科学重点研究基地重大项目,项目号: 2009JJD910001; 教育部人文社会科学研究规划基金项目,项目号: 11YJA630026; 广东省普通高校人文社科研究项目,项目号: 10WYXM020; 广东商学院科学研究重点项目,项目号: 08ZD11001 的资助。

(1996) 认为,主成分存在无序变量时,主成分综合评价函数 $Y_{综}$ 会导致错误的结论,因此,在使用中必须格外谨慎。③Saporta(1989) 认为,如果想以一个综合变量来取代 p 维原始变量 X,则最好的选择便是第一主成分。但李靖华等(2002) 认为,$Y_{综}$ 的构造存在采用一个还是多个主成分的问题。只取第一主成分是一种极端的降维方法,其前提是第一主成分的方差贡献率足够大,但一般都较难满足这一条件。④王学民(2007) 认为,主成分综合评价函数 $Y_{综}$ 有什么样的实际含义,应用者都没有解释或做不出解释,用这种不知其实际含义的指标对样品进行排序说明不了什么问题。这些质疑或断言,不论指标是否标准化、是否非线性化等,都是存在的。苏为华(2000) 认为,正确认识多元统计方法的数学特性、经济意义及应用条件,是有效应用这些方法的前提。因此,根据上述质疑和争议,我们提出如下主成分分析综合评价的应用条件问题:

问题 1 主成分的具体涵义何时能给予较为清楚的解释?

问题 2 主成分存在无序变量时,主成分综合评价函数 $Y_{综}$ 的结果何时不出错?

问题 3 主成分个数如何确定是更好的?

问题 4 主成分综合评价函数 $Y_{综}$ 何时具有合理性? 如何解释其实际含义?

问题 5 如何给出含有方法应用条件及其措施的主成分分析综合评价步骤?

据查阅,一些经济管理类重要期刊的论文,如"中国社会科学"陈秀山和徐瑛(2004) 的论文"中国区域差距影响因素的实证研究',《经济研究》鲁明泓(1997) 的论文"外国直接投资区域分布与中国投资环境评估"、钞小静和任保平(2011) 的论文"中国经济增长质量的时序变化与地区差异分析',《管理世界》颜莉(2012) 的论文"我国区域创新效率评价指标体系实证研究',《统计研究》陈述云和张崇甫(1995) 的论文"对多指标综合评价的主成分分析方法的改进"等,都存在上述 5 个问题(甚至一些论文混淆了主成分与因子),即这些问题具有一定范围的共性,且近期看到其中的一些论文已被其他论文引用 473 篇,影响力较大。故重视主成分分析综合评价的应用条件,能对该方法的应用者带来一些有益的帮助。

关于这些问题的研究进展如下: Johnson 和 Wichern(2007) 认为,既考察主成分的变量系数又考察变量与主成分的相关系数,有助于解释主成分,但没有给出指标标准化下的结果。阎慈琳(1998) 认为,若主成分 y_i 的系数绝对值较大者都为负值,此时应把主成分系数变为其相反数,使主成分成为 $-y_i$。胡永宏(2012) 认为,第一主成分后面的主成分出现较多系数为负时,是否反向没有公认的准则。邱东(1990) 归纳了一些主成分个数的确定方法: 依据前几个主成分累计方差贡献率 $\geqslant 85\%$ 确定; 依据 $\lambda_i \geqslant 1$ 的个数确定; 或依据斯格里准则和巴特莱特检验准则确定等,但理论上无法判定哪个更好(张尧庭和方开泰,1982)。郭亚军和于兆吉(2002) 基于原始指标,将综合评价的合理性归结为综合评价中 15 个环节的合理性,但没有给出主成分分析综合评价的合理性。阎慈琳(1998) 将主成分分析法进行改进后,用因子分析法进行综合评价,但没有解决上述问题。据查阅,其他有关文献也没有解决上述问题。

本文参照因子分析法的因子解释更精细,根据主成分载荷阵与因子载荷阵的关系,应用因子分析中因子载荷阵的简单结构、变量与主成分的相关性、加权算术平均数的合理性,解决主成分分析综合评价中存在的上述 5 个问题,并结合 2010 年广东省各市对外贸易国际竞争力评价的实例提出一些建议。

二、问题解析

设 p 维变量 $X=(x_1,\cdots,x_p)'$,X 是正向的(X 的每个变量越大越好)、标准化的①,记前 k 个主成分为 y_1、$\cdots$、y_k,$\lambda_j=Var(y_j)$,变量 X 与主成分 y_1、$\cdots$、y_k 的相关阵为 $B_k^0=(b_1^0,\cdots,b_k^0)=(b_{ij}^0)_{p\times k}$,称 B_k^0 为主成分载荷阵,则有: b_j^0 是变量 X 与主成分 y_j 的相关系数列,载荷 b_{ij}^0 是变量 x_i 与主成分 y_j 的相关系数,主成分 $y_j=(\lambda_j^{-1/2}b_j^0)'X$(任雪松和于秀林,2011)。以下给出上述 5 个问题的解答。

(一) 问题 1 解答

主成分 y_j 是综合指标(任雪松和于秀林,

① 选择 X 是正向的原因是: 便于综合中确认变量的方向(正向化措施见: 郭亚军和于兆吉,2002); 选择 X 是标准化的原因是: 标准化是等价的线性变换,能消除量纲的影响,能保持变量间的相关性和方向,使样品间有相对可比性(曾五一和肖红叶,2007),又因为主成分 $y_j=(\lambda_j^{-1/2}b_j^0)'X$,所以主成分有时能直接反映变量、主成分之间的相关性,主成分对变量的解释较直接,且能解释方差小的原始变量。其他变换不一定有这些优点。后面结论 3 将给出该条件的合理性。

2011),故其是有涵义和方向的。主成分 y_1、…、y_k要有较为清楚的解释,取决于主成分载荷阵 B_k^0的结构①。主成分载荷阵 B_k^0是因子分析主成分法的前 k 列初始因子载荷阵(任雪松和于秀林,2011),因子分析中,因子的解释较为精细,故我们结合因子分析解决此问题。设因子载荷阵 $B=(b_{ij})_{p\times k}$,载荷 b_{ij}是变量 x_i与因子 f_j的相关系数,故 $|b_{ij}|\leqslant 1$;设因子分析主成分法的前 l 列初始因子载荷阵为 B_l^0。因子要有较为清楚的解释,取决于因子载荷阵的简单结构:因子载荷阵使各变量在某单个因子上有高额载荷(Johnson 和 Wi-chern,2007)。考虑到降维,笔者给出定义:如果因子载荷阵 B 每行载荷有最大绝对值较靠近 1,且列数较小,则称因子载荷阵达到简单结构(见第 3 部分表 2 B_3^0 的第 1 列 B_1^0)。Johnson 和 Wichern(2007)认为,要得到简单结构,并不总是可能的。为此,我们用穷举法来实现这个结果。记达到简单结构的初始因子载荷阵为 B_s^0(如表 2 B_3^0 的第 1 列 B_1^0),将 B_s^0、B_{s+1}^0、…、B_p^0进行方差最大化正交旋转,记旋转后因子载荷阵为 $B_s^0\Gamma_s$、$B_{s+1}^0\Gamma_{s+1}$、…、$B_p^0\Gamma_p$,这里 Γ_l是使 $B_l^0\Gamma_l$达到方差最大化的正交旋转矩阵(任雪松和于秀林,2011),用"因子载荷阵每行元素最大绝对值靠近 1 频数表"(如表 3)进行比较,从中选出达到简单结构的旋转后因子载荷阵,记为 $B_m^0\Gamma_m$(m 列),$s\leqslant m$,$m=1$ 时,定义 $\Gamma_1=1$。据此,有如下结论和推论:

结论 1:B_s^0、$B_m^0\Gamma_m$进行比较②,如果前 s 列主成分载荷阵 B_s^0达到更好的简单结构或 B_s^0、$B_m^0\Gamma_m$都是差异不大的简单结构,则主成分有较为清楚的解释(证明可向作者索取)。

需要指出的是,主成分值与因子值不可混淆(林海明,2009)。结论 1 用到了主成分载荷阵 B_s^0、旋转后因子载荷阵 $B_m^0\Gamma_m$,是为了比较说明主成分是否有较为清楚的解释,没有用到因子值,故没有混淆主成分值和因子值。以下主成分的命名、正向化、个数确定同样没有混淆。

由主成分的表达式 $y_j=(\lambda_j^{-1/2}b_j^0)'X$ 可知,主成分 y_j的涵义和方向由 b_j^0和 X 决定。由此在结论 1 条件下有:

主成分 y_j的命名(解释)步骤:①设显著相关的临界值为 $r(n-2)$(大样本时或取 0.5),如果 b_j^0中的载荷有 $|b_{i_1j}^0|$、…、$|b_{i_cj}^0|\geqslant r(n-2)$,则变量 x_{i_1}、…、x_{i_c}与 y_j显著相关;②对载荷 $b_{i_1j}^0$、…、$b_{i_cj}^0$进行降序排列,因为 X 正向,这些载荷同号的对应变量是相互协作关系、不同号的对应变量是相互制约关系;③如果载荷 $b_{i_1j}^0$、…、$b_{i_cj}^0$差异不大,符号相同,则将这些变量综合起来命名,如"y_j是反映 x_{i_1}、…、x_{i_c}综合影响的成分";④如果这些变量排序后,前面的部分变量附带了后面部分变量的影响,这些载荷符号相同,则用前面的部分变量对 y_j命名;⑤如果载荷 $b_{i_1j}^0$、…、$b_{i_cj}^0$同时有正号、负号,用这些变量内在的协作性、制约性对 y_j命名,如"y_j是反映某些变量与余下变量对比的成分"。

主成分 y_j的方向:由 b_j^0和显著相关的临界值 $r(n-2)$,明确与 y_j显著相关的变量 x_{i_1}、…、x_{i_c},如果用载荷 $b_{i_1j}^0$、…、$b_{i_cj}^0$,指标 x_{i_1}、…、x_{i_c}(含指标的专业知识)确定主成分 y_j的影响是越大越好,则称主成分 y_j正向;否则,称主成分 y_j非正向。y_j非正向时,正向的措施是:

推论 1:如果主成分 y_j非正向,则主成分 $-y_j$正向(证明可向作者索取)。

(二)问题 2 解答

众所周知,指标体系是由一系列相互联系的指标所组成的有机整体。用以反映所研究现象各方面相互促进、相互制约的关系(曾五一和肖红叶,2007)。但这些关系并不是已知的,主成分分析在结论 1 条件下能反映这些关系:主成分之间不相关。因为 X 正向,主成分中载荷较大的对应指标同符号是相互促进关系、不同符号是相互制约关系。即主成分变量系数中部分有正号、部分有负号(无序变量)是相应指标存在有机联系的正常表现,是一种必须面对的常见现象。我们认为,王惠文(1996)的断言应该是:主成分 y_1、…、y_k存在不同方向时,主成分综合评价函数 $Y_{综}$会

① 设显著相关的临界值为 $r(n-2)$,则 b_j^0中 $|b_{ij}^0|\geqslant r(n-2)$ 的对应变量与主成分 y_j显著相关,由相关的传递性,这些变量有相关关系,又因为 $y_j=(\lambda_j^{-1/2}b_j^0)'X$,不同 y_j不相关,所以 y_j有时能较直接地解释这些变量的相关关系,且 b_1^0,…,b_k^0将这些变量进行了相关结构的分组,故 B_k^0的结构对主成分 y_1,…,y_k解释 X 有时能起到决定性的作用。

② $B_m^0\Gamma_m$是逐次对 B_m^0每两列元素进行方差最大化正交旋转的结果,B_m^0是列元素平方和(因子方差贡献)降序排列达到最大化的结果(张尧庭和方开泰,1982)。即 B_m^0、$B_m^0\Gamma_m$的最大化方向不同,故一般情况下,B_m^0、$B_m^0\Gamma_m$达到简单结构的结果是不同的。

导致错误的结论。故有,对主成分 y_1、…、y_k 中非正向的主成分取负号(正向化)后,主成分正向,此时,主成分综合评价函数 $Y_{综}$ 方向综合上的结果是合理的,不能认为其方向综合上存在错误。

(三)问题 3 解答

我们给出定义:通过主成分载荷阵 $B_p^0=(b_{ij}^0)_{p\times p}$ 中列元素最大值和显著相关的临界值 $r(n-2)$(大样本时或取 0.5)判断,如果 B_p^0 列元素的 $\max\limits_{1\leqslant i\leqslant p}\{|b_{ij}^0|\}\geqslant r(n-2)$,$j=1,\cdots,k$;$\max\limits_{1\leqslant i\leqslant p}\{|b_{ij}^0|\}<r(n-2)$,$j>k$,$k\geqslant s$,称主成分与变量显著相关(或主成分 y_1、y_2、…、y_k 与 X 显著相关)。据此,有:

推论 2:结论 1 条件下,如果主成分与变量显著相关,则主成分 y_1、…、y_k 解释变量 X 时,不会遗失 X 中每个变量的主要信息,故主成分个数确定为 k 是更好的(证明可向作者索取)。

(四)问题 4 解答

这里先给出多变量加权算术平均数的合理性条件。众所周知,学生语文、数学、外语课命题和考试都规范的卷面考试成绩,正向和百分制下记为 x_1、x_2、x_3,假定 x_1、x_2、x_3 同等重要、不相关时,卷面总成绩为 $Y_{总}=x_1+x_2+x_3$,人们通常用 $Y_{总}$ 来反映学生这三门课卷面考试成绩的综合评价结果及其水平程度。$Y_{总}$ 归一化后有加权算术平均数:$Y_{综}=(x_1+x_2+x_3)/3$,$Y_{总}$ 与 $Y_{综}$ 的评价结果是一样的,此时,人们将加权算术平均数 $Y_{综}$ 当成了综合评价函数。

参照此做法设 p 维变量 $X=(x_1,\cdots,x_p)'$(这里不要求 X 标准化),对 X 的观测数据,有:

结论 2:综合评价函数 $Y_{综}=\alpha_1x_1+\cdots+\alpha_px_p$ 合理的构造条件是:①X 正向;②X 量纲制相同;③X 不相关;④合理权重 $\alpha_i\geqslant 0$,$\alpha_1+\cdots+\alpha_p=1$($x_i$ 同等重要时,$\alpha_i=1/p$)。

满足结论 2 的 4 个条件时,人们认为,$Y_{综}=\alpha_1x_1+\cdots+\alpha_px_p$ 具有加权算术平均数的合理性。而实际上,多数情况下结论 2 的条件②、条件③不成立,这使得变量不能综合,且有信息的重叠,从而需要用少数几个量纲制相同、不相关的综合变量取代 X。主成分分析有时便是取代方法之一。

结论 3:主成分综合评价函数 $Y_{综}=\alpha_1y_1+\cdots+\alpha_ky_k$($\alpha_i=\lambda_i/p$)合理的构造条件是:①$X$ 是正向、标准化的;②主成分载荷阵 B_s^0 达到更好的简单结构或 B_s^0、$B_m^0\Gamma_m$(旋转后因子载荷阵)都是差异不大的简单结构;③主成分 y_1、…、y_k 正向;④主成分与变量显著相关(证明可向作者索取)。

结论 3 的条件满足时,将主成分 $y_j=u_j'X$(u_j 为主成分 y_j 中变量 X 的系数向量)代入主成分综合评价函数 $Y_{综}=\alpha_1y_1+\cdots+\alpha_ky_k$ 后,记 $Y_{综}=u'_{综}X$($u_{综}$ 为 $Y_{综}$ 中变量 X 的系数向量),故 $Y_{综}$ 的实际含义可根据 $u'_{综}$ 中元素的大小和 X 作出解释。结合结论 3 便解答了问题 4。

结论 3 的条件是主成分分析综合评价较好的数学特性和应用条件,此时,主成分的解释(或经济意义)是较为清楚的。用不适合结论 3 条件的数据作主成分分析综合评价,结果不合理的例子是屡见不鲜的,比如,人们是千篇一律地采用传统的变量型主成分评价技术(苏为华,2012),或即便是适合这些条件,也有待检验和分析的深入。因此,重视结论 3 及其条件的应用,使主成分分析综合评价的结果不出错或更合理,是必要的。据我们对现有文献的抽样,符合结论 3 条件的多元数据占 33% 左右。

(五)问题 5 解答

根据结论 3 等,我们给出一个含有方法应用条件及其措施(深化)的主成分分析综合评价步骤:

(1)数据的预处理:对原始指标进行正向化、标准化(结论 3),记为 $X=(x_1,\cdots,x_p)'$。

(2)指标 X 可降维的判定:如果变量间有相关系数的绝对值 $\geqslant 0.8$,则指标 X 可降维。

(3)选出简单结构的初始、旋转后因子载荷阵:因子分析主成分法下,设达到简单结构的初始因子载荷阵(主成分载荷阵)为 B_s^0(s 列),将 B_s^0、B_{s+1}^0、…、B_p^0 进行方差最大化正交旋转,记得出的旋转后因子载荷阵为 $B_s^0\Gamma_s$、$B_{s+1}^0\Gamma_{s+1}$、…、$B_p^0\Gamma_p$,用“因子载荷阵每行元素最大绝对值靠近 1 频数表”(如表 3)进行比较,从中选出达到简单结构的旋转后因子载荷阵(穷举法),记为 $B_m^0\Gamma_m$(m 列)。

(4)主成分有较为清楚解释的判定:用“因子载荷阵每行元素最大绝对值靠近 1 频数表”进行比较,若主成分载荷阵 B_s^0 达到更好的简单结构或 B_s^0、$B_m^0\Gamma_m$ 都是差异不大的简单结构,则主成分有较为清楚的解释(结论 3)。

(5)确定主成分个数:主成分 y_1、y_2、…、y_k 与 X 显著相关时,则主成分个数确定为 k(推论 2),相应的主成分载荷阵记为 $B_k^0=(b_1^0,\cdots,b_k^0)=(b_{ij}^0)_{p\times k}$。

(6)主成分的正向化、命名:在 B_k^0 的第 j 列 $b_j{}^0$ 的

载荷中,选出绝对值大于显著相关临界值的载荷 $b_{i_1j}^0$、…、$b_{i_cj}^0$,变量 x_{i_1}、…、x_{i_c} 归为主成分 y_j 一组,如果用载荷 $b_{i_1j}^0$、…、$b_{i_cj}^0$、变量 x_{i_1}、…、x_{i_c}(含指标的专业知识)确定 x_{i_1}、…、x_{i_c} 的综合影响是越大越好,则主成分 y_j 取正号,否则,主成分取负号成为 $-y_j$(推论 1);之后对归为正向化主成分 y_j 一组的变量,按载荷 $b_{i_1j}^0$、…、$b_{i_cj}^0$ 降序排列后,结合载荷 $b_{i_1j}^0$、…、$b_{i_cj}^0$,变量 x_{i_1}、…、x_{i_c}(含指标的专业知识)对主成分 y_j 进行命名。正向化后的主成分仍记为 y_1、y_2、…、y_k,方差为 λ_1、λ_2、…、λ_k。

(7) 构造主成分综合评价函数: $Y_{综}=\alpha_1 y_1+\alpha_2 y_2+\cdots+\alpha_k y_k=u'_{综}X$,$\alpha_i=\lambda_i/p$(结论 3)。

(8) 样品值及其排序: 计算 y_1、…、y_k,$Y_{综}$ 的样品值,并给出其排序。

(9) 样品分类: 对主成分 y_1、…、y_k 样品值进行标准化后(任雪松和于秀林,2011),做系统聚类分析,按主成分综合评价函数 $Y_{综}$ 中样品值的排序给出 n 个样品的分类结果。

(10) 评价与建议: 结合样品的分类结果,y_1、…、y_k、$Y_{综}$ 的样品值及其排序,y_1、…、y_k、$Y_{综}$ 中变量 X 及其系数确定指标体系的内在促进或制约关系,进行综合评价,给出较客观、可靠的决策建议。

三、应用实例

为验证上述主成分分析综合评价步骤的有效性,以 2010 年广东省各市对外贸易国际竞争力的数据进行评价。指标体系为: X_1 为地区生产总值(亿元),X_2 为年均从业人员数(万人),X_3 为从业人员的平均劳动报酬(元),X_4 为国际市场占有率(‰),X_5 为城镇居民人均可支配收入(元),X_6 为工业企业新产品出口(万美元),X_7 为工业企业 R&D 活动人员数(人),X_8 为净出口量(万美元),X_9 为对外贸易依存度(%),X_{10} 为实际利用外资额(万美元),X_{11} 为合同外资额(万美元),X_{12} 为金融机构储蓄存款(百亿元),X_{13} 为社会固定资产投资额(百亿元),X_{14} 为第三产业增加值(百亿元),指标个数 $p=14$,样本容量 $n=21$。

(1) 数据的预处理: 指标都是正向的,只对变量 X_1-X_{14} 进行标准化,记为 x_1-x_{14};

(2) 指标可降维的判定: 用 SPSS 软件计算,由变量相关阵 R 得,x_1 与 x_{11}、x_{12}、x_{13}、x_{14} 的相关系数分别为 0.951、0.993、0.946、0.986 等,即变量之间有高度相关性,故变量可降维;

(3) 选出简单结构的初始、旋转后因子载荷阵: 因子分析主成分法下,列数 $s=1$ 时,初始因子载荷阵 B_1^0(表 1 B_3^0 的第 1 列)达到简单结构,从多个不同列的旋转后因子载荷阵中挑选(表 2 频数的第 2 – 4 列),$m=1$ 时,旋转后因子载荷阵 $B_1^0\Gamma_1$(表 1)达到简单结构(此时,$\Gamma_1=1$);

表 1　　因子载荷阵

变量	B_3^0(初始因子载荷阵、主成分载荷阵)			$B_1^0\Gamma_1$(旋转后因子载荷阵)
	1	2	3	1
x_1	0.959*	−0.238	−0.124	0.959
x_2	0.910*	−0.227	−0.228	0.910
x_3	0.908*	−0.208	0.273	0.908
x_4	0.914*	0.397	0.018	0.914
x_5	0.842*	−0.126	0.506*	0.842
x_6	0.809*	0.550*	−0.168	0.809
x_7	0.931*	0.348	−0.073	0.931
x_8	0.732*	0.653*	0.061	0.732
x_9	0.931*	0.345	0.010	0.931
x_{10}	0.969*	−0.133	0.120	0.969
x_{11}	0.976*	−0.102	−0.003	0.976
x_{12}	0.932*	−0.325	−0.108	0.932
x_{13}	0.861*	−0.458*	−0.082	0.861
x_{14}	0.934*	−0.295	−0.167	0.934

(4) 主成分有较为清楚解释的判定: B_1^0 同 $B_1^0\Gamma_1$ 比较: 由表 1 的 B_1^0 得表 2 频数的第 1 列,表 2 频数的第 1 – 2 列表明,B_1^0、$B_1^0\Gamma_1$ 是一致的简单结构,故主成分有较为清楚的解释;

表 2　因子载荷阵每行元素最大绝对值靠近 1 频数表

每行因子载荷最大绝对值区间	频数			
	B_1^0	$B_1^0\Gamma_1$	$B_2^0\Gamma_2$	$B_t^0\Gamma_t$,$t=3\sim14$
0.9 以上	10	10	6	5
0.8 ~ 0.9	3	3	7	5
0.7 ~ 0.8	1	1	1	3
0.6 ~ 0.7	0	0	0	1
合计	14	14	14	14

(5) 确定主成分个数 k: 变量正态分布下,取显著水平为 5%,显著相关的临界值是 $r(19)=0.443$①,$p=14$,由 B_{14}^0 有: $\max\limits_{1\leqslant i\leqslant 14}\{|b_{i1}^0|\}=b_{11\,1}^0=0.976$、$\max\limits_{1\leqslant i\leqslant 14}\{|b_{i2}^0|\}=b_{82}^0=0.653$、$\max\limits_{1\leqslant i\leqslant 14}\{|b_{i3}^0|\}=b_{53}^0=0.506\geqslant r(19)$;$\max\limits_{1\leqslant i\leqslant 14}\{|b_{ij}^0|\}<r(19)$,$j>3$,即主成分 y_1、y_2、y_3 与 X 显著相关,故 $k=3$,累计方差率为 97.4%;

(6) 主成分正向化与命名: 由 B_3^0 得表 3,主成分

① 茆诗松等编著. 概率论与数理统计 [M]. 北京: 中国统计出版社,2000: 106,420.

y_1 与全部指标 $x_1 - x_{14}$ 显著正相关，y_1 中 $x_1 - x_{14}$ 的综合影响越大越好，故 y_1 是正向的，y_1 称为外贸国际竞争力水平成分；主成分 y_2 与 x_6（工业企业新产品出口），与 x_8（净出口量）显著正相关，与 x_{13}（社会固定资产投资额）显著负相关，这反映了出口与社会投资的内在关系：社会固定资产投资对出口有一些负影响，原因是国内的社会固定资产投资主要是内需拉动的，不是出口拉动的，但从总体上讲，参与出口及其竞争与坚持对外开放是一个方向，故 y_2 是正向的，y_2 称为出口与社会投资对比成分；y_3 与 x_5（城镇居民人均可支配收入）显著正相关，x_5 的影响越大越好，故 y_3 是正向的，由于 x_5 同时与 y_1、y_3 显著正相关，相关系数分别为 0.842、0.506，故 y_3 称为城镇居民人均可支配收入补充成分。$\lambda_1 = 11.415$、$\lambda_2 = 1.725$，$\lambda_3 = 0.496$，主成分 y_1、y_2、$y_3$① 如下：

表 3 主成分命名

主成分	与主成分显著相关的指标及其载荷	命名
y_1	$X_1 - X_{14}$；载荷范围 0.732 ~ 0.976，全部正号。	外贸国际竞争力水平成分
y_2	X_8（净出口量），X_6（工业企业新产品出口），X_{13}（社会固定资产投资额），载荷分别为 0.653、0.550、-0.458。	出口与社会投资对比成分
y_3	X_5（城镇居民人均可支配收入），载荷为：0.506。	城镇居民人均可支配收入补充成分

$$y_1 = 0.284x_1 + 0.269x_2 + 0.269x_3 + 0.271x_4 + 0.249x_5 + 0.239x_6 + 0.276x_7 + 0.217x_8 + 0.276x_9 + 0.287x_{10} + 0.289x_{11} + 0.276x_{12} + 0.255x_{13} + 0.276x_{14}$$

（x_i 是 X_i 的标准化）

$$y_2 = -0.181x_1 - 0.173x_2 - 0.158x_3 + 0.302x_4 - 0.096x_5 + 0.419x_6 + 0.265x_7 + 0.497x_8 + 0.263x_9 - 0.101x_{10} - 0.078x_{11} - 0.247x_{12} - 0.349x_{13} - 0.225x_{14}$$

$$y_3 = -0.176x_1 - 0.323x_2 + 0.387x_3 + 0.026x_4 + 0.718x_5 - 0.239x_6 - 0.104x_7 + 0.086x_8 + 0.015x_9 + 0.171x_{10} - 0.004x_{11} - 0.154x_{12} - 0.117x_{13} - 0.237x_{14}$$

（7）构造主成分综合评价函数：

$$Y_{综} = (11.415\ y_1 + 1.725y_2 + 0.496y_3)/14 = 0.259x_4 + 0.259x_9 + 0.255x_7 + 0.241x_8 + 0.239x_6 + 0.228x_{10} + 0.227x_{11} + 0.217x_5 + 0.215x_3 + 0.203x_1 + 0.190x_{14} + 0.190x_{12} + 0.188x_2 + 0.159x_{13}$$

$Y_{综}$ 的含义：$Y_{综}$ 按系数大小对变量排序是 x_4、x_9、x_7、x_8、x_6、x_{10}、x_{11}、x_5、x_3、x_1、x_{14}、x_{12}、x_2、x_{13}，该评价指数前 8 个指标注重 x_4（国际市场占有率）、x_9（对外贸易依存度）、x_7（工业企业 R&D 活动人员）、x_8（净出口量）、x_6（工业企业新产品出口）、x_{10}（实际利用外资）、x_{11}（合同外资额）、x_5（城镇居民人均可支配收入），故 $Y_{综}$ 的评估与对外贸易国际竞争力的目标基本相符。

（8）样品值及排序：主成分 y_1、y_2、y_3、综合评价函数 $Y_{综}$ 样品值及其排序见表 4。

表 4 主成分、主成分综合评价函数样品值及其排序

城市	y_1	序	y_2	序	y_3	序	$Y_{综}$	序
深圳	10.789	1	3.577	1	-0.693	19	9.213	1
广州	7.068	2	-4.483	21	-0.773	21	5.182	2
东莞	3.473	3	-0.292	18	1.812	1	2.860	3
佛山	2.645	4	-0.742	20	0.143	7	2.070	4

（9）样品分类：用表 5 中 y_1、y_2、y_3 样品值的标准化值作系统聚类分析，采用欧氏距离，选取类平均法，分类阈值取为 1.68 时，分成 4 类，结合 $Y_{综}$ 样品值排名顺序给出样品分类结果：

第一类：深圳；第二类：广州；第三类：东莞、中山、珠海；第四类：佛山、惠州、江门等。

（10）评价：以第二类的广州为例，$Y_{综}$ 值（5.182）排第 2 位，远高于平均水平，优势明显。其中外贸国际竞争力水平成分 y_1 值（7.068）排第 2 位，远高于平均水平，优势明显；出口与社会投资对比成分 y_2 值（-4.483）倒数第 1 位；城镇居民人均可支配收入补充成分 y_3 值（-0.544）倒数第 1 位。即广州是国际竞争力水平优势明显，但出口与社会投资、城镇居民人均可支配收入补充方面有待协调。

建议：①因为 $x_1 - x_{14}$ 与 y_1 显著正相关，故 y_1 中 $x_1 - x_{14}$ 是相互促进的变量，因此，广州市在继续保持外贸国际竞争力水平成分 y_1 中 x_1（地区生产总值）、x_2（从业人员年末人数）、x_3（从业人员的平均劳动报酬），x_{12}（金融机构储蓄存款）、x_{13}（社会固定资产投资额）（如亚运会场馆、配套城市建设等）排序均为 1，x_7（工业企业 R&D 活动人员）、x_{10}（实际利用外资）、x_{11}（合同外资额）排序均为 2 的前提下，促进了 x_4（国际市场占有率）、x_5（城镇居民人均可支配收入）、x_9（对外贸易依存度）的发展和提高；②因为 x_6、x_8 与 y_2 显著正相关、x_{13} 与 y_2 显著负相关，故 y_2 中的 x_6、x_8 与 x_{13} 是相互制约的变量，因此，广州市在保持出口与社会投资对比成分 y_2 中 x_{13}（社会固定资产投资额）排序均为 1 的有利条件下，必须

① 林海明. 如何用 SPSS 快速计算主成分的结果［J］. 统计与决策，2011(6)：16 - 18.

协调促进 x_6(工业企业新产品出口) 、x_8(净出口量) 的增加; ③因为 x_5(城镇居民人均可支配收入) 与城镇居民人均可支配收入补充成分 y_3 显著正相关,因此,广州市应发挥好城镇居民人均可支配收入补充的协调作用。这三方面工作的共同发展和提高,将会使广州市有更高水平的外贸国际竞争力。

其他类各样品的评价与建议略。

四、结论与建议

本文给出了主成分分析综合评价的应用条件: 变量 X 是正向、标准化的; 主成分载荷阵 B_s^0 达到更好的简单结构或 B_s^0、$B_m^0\Gamma_m$(达到简单结构的旋转后因子载荷阵) 是差异不大的简单结构; 主成分 y_1、…、y_k 正向; 主成分与变量显著相关。满足这些条件时,主成分综合评价函数在加权算术平均数意义下是合理的,并给出了一个含有方法应用条件及其措施(深化) 的主成分分析综合评价步骤及其应用实例。以下几个要点应特别注意:

(1) 主成分有较为清楚解释的判定。建议在因子分析主成分法下,计算并给出达到简单结构的初始因子载荷阵(主成分载荷阵) B_s^0,从多个不同列旋转后因子载荷阵中,选出达到简单结构的旋转后因子载荷阵,记为 $B_m^0\Gamma_m$,若主成分载荷阵 B_s^0 达到更好的简单结构或 B_s^0、$B_m^0\Gamma_m$ 是差异不大的简单结构,则主成分有较为清楚的解释。

(2) 确定主成分个数。若主成分 y_1、…、y_k 与 X 显著相关,建议主成分个数确定为 k。

(3) 主成分的正向化。建议与主成分 y_j 显著相关的变量归为 y_j 一组,如果这组变量及其载荷的综合影响是越大越好,则主成分 y_j 取正号,否则,该主成分正向化取负号成为 $-y_j$。

(4) 主成分的命名(解释) 。建议与正向主成分 y_j 显著相关的变量归为 y_j 一组,对这组变量的载荷降序排列后,用这组变量及其载荷的综合影响对主成分 y_j 进行命名。

⑸如果旋转后因子载荷阵 $B_m^0\Gamma_m$ 达到更好的简单结构,主成分综合评价函数可能得不到较满意的结果,建议用因子分析主成分法的因子分析综合评价解决问题(林海明,2009) 。

对适合结论 3 条件的数据,除了能使用主成分分析综合评价外,还可使用初始因子分析综合评价(林海明,2009) ,但哪种方法的结果更好有待讨论。

参考文献

[1] 苏为华. 我国多指标综合评价技术与应用研究的回顾与认识[J]. 统计研究,2012(8) : 98 - 107.

[2] 邱东. 多指标综合评价方法[J]. 统计研究,1990(6) : 43 - 51.

[3] 何平. 我国综合评价活动发展述评[EB/OL]. http: //www. sts. org. cn/fxyj/zbtx /documents/zhps. htm,2005.

[4] 王惠文. 用主成分分析法建立系统评估指数的限制条件浅析[J]. 系统工程理论与实践,1996(9) : 25 - 28.

[5] Saporta. M. , Analyse des Donnees, EN SAE,1989.

[6] 李靖华,郭耀煌. 主成分分析用于多指标评价的方法研究——主成分评价[J]. 管理工程学报,2002,16(1) : 39 - 43.

[7] 王学民. 对主成分分析中综合得分方法的质疑[J]. 统计与决策,2007(4) : 31 - 32.

[8] 苏为华. 多指标综合评价理论与方法问题研究[D]: [博士学位论文]. 厦门: 厦门大学,2000.

[9] Johnson, R. A. , Wichern, D. W. Applied Multivariate Statistical Analysis [M]. 6th ed. Published by Pearson Educatin, Inc, publishing as Prentice Hall, Copyritght, 2007: 430 - 538.

[10] 阎慈琳. 关于用主成分分析做综合评价的若干问题[J]. 数理统计与管理,1998(2) : 22 - 25.

[11] 胡永宏. 对统计综合评价中几个问题的认识与探讨[J]. 统计研究,2012(1) : 26 - 30.

[12] 张尧庭,方开泰. 多元统计分析引论[M]. 北京: 科学出版社,1982.

[13] 郭亚军,于兆吉. 综合评价的合理性问题[J]. 东北大学学报,2002(9) : 844 - 847.

[14] 曾五一,肖红叶主编. 统计学导论[M]. 北京: 科学出版社,2007.

[15] 任雪松,于秀林. 多元统计分析[M]. 北京: 中国统计出版社,2011: 184 - 231.

[16] 林海明. 因子分析模型的改进与应用[J]. 数理统计与管理,2009,28(6) : 998 - 1012.

作者简介

林海明,男,1959 年生,湖南省宁乡县人,1988 年获湖南大学应用数学专业理学硕士学位,现为广东财经大学经济贸易学院、华商学院统计学教授,中国人民大学应用统计科学研究中心兼职研究员,广东省现场统计学会常务理事。研究方向为多元统计模型和应用。

杜子芳,男,1958 年生,山东省文登市人,1988 年毕业于中国人民大学统计系,获经济学硕士学位,现为中国人民大学统计学院教授,博士生导师,中国现场统计学会抽样分会常务理事,国务院反垄断委员会专家组专家。研究方向为抽样调查、数据处理、宏观概算。

(责任编辑: 麦　芒)