第21卷第5期 统计与信息论坛 Vol.21 No.5
2006年9月 Statistics & Information Forum Sept.,2006

【统计理论与方法】

多指标综合评价中主成分分析和因子分析方法的比较

王文博,陈秀芝
(西安交通大学 经济与金融学院,陕西 西安 710061)

摘要:文章通过对主成分分析和因子分析在研究目的、分析原理、SPSS软件实现过程方面的比较,指出在多指标综合评价时应用两种方法应该注意四个问题,以正确地进行实证研究。

关键词:主成分;因子;区别
中图分类号:O212.4 **文献标识码**:A **文章编号**:1007-3116(2006)05-0019-04

一、问题的提出

在多元统计分析中,主成分和因子分析是两种很重要的数据降维方法。随着两种方法不断被研究推广,它们越来越多地被应用于实证分析尤其是多指标综合评价中。但是由于对两种方法的原理理解不透彻,对统计软件SPSS中两种方法的实现过程不清楚,导致在实际应用时错误使用输出结果,使定量分析出现偏差。

二、主成分分析和因子分析的区别、联系

首先假定 $\boldsymbol{X}$ 是已经标准化的 p 维随机向量,因此 $\boldsymbol{X}$ 的协差阵和相关系数阵是同一的,用 $\boldsymbol{R}$ 表示;λ_i 是 $\boldsymbol{R}$ 的第 i 大特征根,λ_i 对应的单位特征向量 $\boldsymbol{e}_i$ 组成的矩阵为 $\boldsymbol{E}_{p\times p}=(\boldsymbol{e}_1,\boldsymbol{e}_2,\cdots,\boldsymbol{e}_p)$,$\boldsymbol{F}$ 是 m 维主成分向量,$\boldsymbol{Z}$ 是 m 维公共因子向量,$\boldsymbol{\xi}$ 是特殊因子向量。

主成分分析和 R 型因子分析都是对变量降维,分析的对象相同,所以本文主要讨论主成分分析和 R 型因子分析的区别和联系。

(一)主成分分析和因子分析简介

1.主成分分析

(1)发展历史和研究目的。主成分概念最早在1901年由Karl Pearson引进,对非随机变量讨论,1933年数学家Hotelling把它推广到随机向量[1-2]。这种方法的思想就是把相关联的原始变量用不相关的新变量表示,表示方式为线性组合;出于简化数据的目的,选取新变量个数少于原始变量个数,被选取的新变量就是"主成分"。选取的原则是不丢失主要信息,衡量信息量的指标是主成分的方差,方差越大,说明主成分包含的信息量越大。

(2)主成分分析的原理。理解主成分分析的原理需要特别注意以下两个方面:

其一,数学模型及其统计特征:

$\boldsymbol{F}_{m\times 1}=\boldsymbol{A}_{p\times m}'\boldsymbol{X}_{p\times 1}$,为保证线性组合的系数确定、唯一,模型需要满足:

(Ⅰ)$\mathrm{Cov}(F_i,F_j)=0$,即各个主成分之间互不相关;

(Ⅱ)$\mathrm{Var}(F_i)=\lambda_i$,$\lambda_i$ 是 $\boldsymbol{R}$ 的第 i 大特征根;主成分的形成顺序按照方差大小排列,即第一主成分方差最大,第二主成分方差第二大,依次类推;

(Ⅲ)$\boldsymbol{A}_{p\times m}'\boldsymbol{A}_{p\times m}=\boldsymbol{I}_{m\times m}$,即系数阵是单位阵。

其二,系数阵的特征和估计。应用主成分模型的关键是估计系数阵 $\boldsymbol{A}_{p\times m}'$。多元统计学已经证明:

$\boldsymbol{A}_{p\times m}'=(\boldsymbol{e}_1,\boldsymbol{e}_2,\cdots,\boldsymbol{e}_m)$,$\boldsymbol{e}_i$ 是 $\boldsymbol{R}$ 的特征根 λ_i 对应的单位特征向量。

收稿日期:2006-03-16

作者简介:王文博(1953-),男,陕西省西安市人,教授,硕士生导师,研究方向:计量经济模型与应用、经济预测与决策;
陈秀芝(1981-),女,天津市人,硕士生,研究方向:计量经济模型与应用。

(3) 主成分分析在 SPSS 中的实现。目前,SPSS 软件中并没有专门的主成分分析实现过程,需要借助因子分析间接得到系数阵 $\mathbf{A}_{p\times m}'$ 从而建立主成分模型。具体操作见下面因子分析的 SPSS 实现步骤。

2. 因子分析

(1) 发展历史和研究目的。因子分析方法最早是在 1904 年由 Charles Spearman 和 Karl Pearson 在一篇著名论文《对智力测验得分进行统计分析》中提出,之后被用于解决心理学和教育学方面的问题[3]。由于这种方法计算量大,到了 20 世纪 60 年代得益于计算机的应用才有新的发展。

R 型因子分析认为变量中存在一些不可观测的共同因素同时对原始变量产生影响,需要通过一定的方法提取"重要"的公共因子;"重要性"取决于因子对变量的影响程度,用二者之间的相关系数(因子载荷)表示。根据变量与各因子的"紧密"程度,把原始变量归结到各因子中,通过这些"精炼"的因子认识复杂现象。

(2) 因子分析的原理。理解因子分析方法需要特别注意三个方面:

其一,数学模型和特征。

$$\boldsymbol{X}_{p\times 1} = \boldsymbol{B}_{p\times m}\boldsymbol{Z}_{m\times 1} + \boldsymbol{\xi}_{p\times 1}$$

$\boldsymbol{Z}_{m\times 1}$ 用来解释影响原始变量的共同因素,$\boldsymbol{\xi}_{p\times 1}$ 表示影响原始变量但又不能由 $\boldsymbol{Z}_{m\times 1}$ 解释的特殊因素。这个数学模型需要满足以下条件[4]:

(Ⅰ)$m < p$ 即提取的公共因子个数少于原始变量个数;

(Ⅱ)$\mathrm{Cov}(Z,\xi) = 0$ 即公共因子和特殊因子不相关;

(Ⅲ)$\mathrm{Var}(Z) = I_{m\times m}$ 即各公共因子不相关且方差为 1;

(Ⅳ)$\mathrm{Cov}(\xi_i,\xi_j) = 0, \mathrm{Var}(\xi_i) = \delta_j$ 即各特殊因子不相关且方差不同。

其二,系数矩阵的估计方法、统计特征及其旋转。

系数阵 $\boldsymbol{B}_{p\times m}$ 是初始因子载荷阵,也是变量与公共因子之间的相关系数阵。估计 $\boldsymbol{B}_{p\times m}$ 有多种方法,其中主成分法应用最为广泛,它与主成分分析并没有原理上的实质联系,主要是外观的联系。因为用主成分法得到初始载荷阵 $\boldsymbol{B}_{p\times m} = (\sqrt{\lambda_1}\boldsymbol{e}_1, \sqrt{\lambda_2}\boldsymbol{e}_2, \cdots, \sqrt{\lambda_m}\boldsymbol{e}_m)$,$\boldsymbol{e}_i$ 是 R 的特征根 λ_i 对应的单位特征向量,它也是主成分分析系数阵 $A'_{p\times m}$ 第 i 个系数向量,所以 $\boldsymbol{B}_{p\times m}$ 第 i 列系数向量与 $\boldsymbol{A}'_{p\times m}$ 第 i 行系数向量仅相差倍数 $\sqrt{\lambda_i}$。

$\boldsymbol{B}_{p\times m}$ 有两个重要的统计意义:

(Ⅰ)$\sum_{i=1}^{p} b_{ij}^2 = g_j$,$\boldsymbol{B}_{p\times m}$ 各列系数相加,表示因子 Z_j 对所有变量的解释能力,即 Z_j 的"贡献"。如果 $\boldsymbol{B}_{p\times m}$ 用主成分法得到,那么,$\sum_{i=1}^{p} b_{ij}^2 = \lambda_j$,$\lambda_i$ 是 $\boldsymbol{R}$ 的特征根;

(Ⅱ)$\sum_{j=1}^{m} b_{ij}^2 = h_i$,$\boldsymbol{B}_{p\times m}$ 中各行系数相加,表示所有因子对变量 X_i 的解释程度,称为各变量在公共因素中的"共同度"。

多元统计中已经证明满足上述模型的系数阵 $\boldsymbol{B}_{p\times m}$ 不唯一,这成为因子载荷阵旋转的理论依据。一般情况下,初始因子载荷阵中各变量对因子的系数没有靠近两极数值"0"和"1",说明各变量在每个因子上"分量"差不多,各因子并不"偏向"某些变量,这样很难提炼公共因子的意义,因此要旋转 $\boldsymbol{B}_{p\times m}$,改变它的坐标系,使变量"偏向"不同的因子,并根据系数绝对值对变量归类命名。最常用的旋转方法是最大方差正交旋转。

(3) 因子得分。从模型分析看,因子分析到此就结束了,可是更多的时候需要直接看到因子的"具体表现",方法通常是以原始变量为自变量,以因子为因变量拟合回归模型。由于因子个数少于原始变量个数,即方程个数少于未知参数个数,所以回归方程的参数不能确定,只能依据一定方法估计,常用的是 Thompson 回归法和 Bartlett 加权最小二乘法。

(二) 主成分分析和因子分析的区别

两种方法的具体区别,详见表 1。

(三) 两种方法的主要联系以及 SPSS 输出结果的应用

1. 可以从两个方面认识两种方法的联系

(1) 系数阵。前面已经提到,主成分法估计得到初始载荷阵 $\boldsymbol{B}_{p\times m}$ 的第 i 列向量 $\sqrt{\lambda_i}\boldsymbol{e}_i$ 与第 i 个主成分的系数向量 $\boldsymbol{e}_i$ 仅相差倍数 $\sqrt{\lambda_i}$。

(2) 贡献率。各主成分的贡献率是各主成分的方差占总方差的比重,用 w_i 表示,因为 $\mathrm{Var}(F_i) = \lambda_i$,所以 $w_i = \lambda_i / \sum_{i=1}^{m}\lambda_i$。初始因子载荷阵 $\boldsymbol{B}_{p\times m} = (b_{ij})_{p\times m}$,各因子贡献用 $\sum_{i=1}^{p} b_{ij}^2 = \lambda_j$ 表示,相应贡献率 $w_i = \lambda_i / \sum_{i=1}^{m}\lambda_i$。比较主成分的贡献率看到,由初

始因子载荷阵得到的第 i 个因子贡献率与主成分分析第 i 个主成分贡献率相同,这说明可以借用初始因子载荷阵选取因子的方法选取主成分。

如果因子载荷阵经过旋转,那么 $\boldsymbol{B}_{p\times m}$ = $(b_{ij})_{p\times m}$ 就会改变,这时相应因子的贡献率与相应主成分贡献率就不同了。SPSS 软件能显示旋转后因子载荷阵的结果,可以专门用于计算旋转后因子的贡献率。

表 1　主成分分析和因子分析的区别表

比较项目	主成分分析	因子分析
数学模型	$\boldsymbol{F}_{m\times 1} = \boldsymbol{A}_{p\times m}{}' \boldsymbol{X}_{p\times 1}$	$\boldsymbol{X}_{p\times 1} = \boldsymbol{B}_{p\times m}\boldsymbol{Z}_{m\times 1} + \boldsymbol{\xi}_{p\times 1}$
系数矩阵	$\boldsymbol{A}_{m\times p} = (\boldsymbol{e}_1, \boldsymbol{e}_2, \cdots, \boldsymbol{e}_p)$ $\boldsymbol{A}_{m\times p}\boldsymbol{A}_{m\times p}{}' = \boldsymbol{I}_m$	初始因子载荷阵 $\boldsymbol{B}_{p\times m} = (\sqrt{\lambda_1}\boldsymbol{e}_1, \sqrt{\lambda_2}\boldsymbol{e}_2, \cdots, \sqrt{\lambda_m}\boldsymbol{e}_m)$
方差	$\mathrm{Var}(F_i) = \lambda_i$,$\lambda_i$ 是 $\boldsymbol{R}$ 的第 i 大特征根	$\mathrm{Var}(Z_j) = 1$
主成分或因子对 X 的贡献	特征根 λ_i	旋转后因子 Z_j 的贡献率 $v_j = \sum_{i=1}^{p} c_{ij}^2$,$c_{ij}$ 是旋转因子载荷阵系数,通常 $v_j < \lambda_i$
综合评价函数	$F_{综} = \sum_{i=1}^{m}(\lambda_i/k)F_i, k = \lambda_1 + \lambda_2 + \cdots + \lambda_m$	$F_{综} = \sum_{j=1}^{m}(v_j/k)Z_j, k = v_1 + v_2 + \cdots + v_m$
SPSS 操作	没有直接操作结果,需要根据因子分析结果推算系数 a_{ij}	可以直接得到结果,包括因子得分

2. SPSS 实现因子分析主要步骤及结果说明[5]

(1) 数据正向化、标准化。

(2) 在"Analyze"菜单中选择"Data Reduction… factor",把变量选入"variables"栏。

(3)"Extraction"按钮:选择主成分法为系数矩阵计算方法;确定以相关系数阵(Correlation Matrix)为分析对象;根据特征值超过临界值(通常为"1")确定提取的因子数目或者自行决定因子个数。

(4)"Rotation"按钮:旋转方法一般选择最大方差旋转(Varimax)。

(5)"Score"按钮提示得分模型系数估计方法:回归法、Bartllet 法还是 Anderson - Rubin 法以及是否显示得分系数阵("Factor Score Coefficient Matrix")。

借助 SPSS 软件用主成分法进行因子分析时可以实现主成分分析过程,主要应用两部分结果:

(1) 应用"累计贡献率"(Total Variance Explained)表中"提取后的因子载荷平方和"(Extraction Sums of Squared Loadings)部分,包括三栏:"Total"栏表示 $\boldsymbol{B}_{p\times m}$ 各列系数平方和 $\sum_{i=1}^{p} b_{ij}^2 = \lambda_j (j = 1,2,\cdots,m)$,也是各主成分的方差,"% of variance"和"Cumulative variance"分别表示各主成分的贡献率和累计贡献率。另外表中"旋转后的因子载荷平方和"(Rotation Sums of Squared Loadings)用于因子分析。

(2) 用初始因子载荷阵(Component Matrix)$\boldsymbol{B}_{p\times m} = (\sqrt{\lambda_1}\boldsymbol{e}_1, \sqrt{\lambda_2}\boldsymbol{e}_2, \cdots, \sqrt{\lambda_m}\boldsymbol{e}_m)$ 第 i 列向量除以 $\sqrt{\lambda_i}$ 就得到第 i 个主成分的系数向量。

另外根据旋转的因子载荷阵(Rotated Component Matrix)可以写出因子分析模型;根据因子得分系数矩阵(Factor Score Coefficient Matrix)可以写出因子得分模型,SPSS 软件直接给出因子得分并以变量形式保存。

三、两种方法在综合评价中的应用

(一) 两种方法都有在不丢失主要信息的情况下简化数据的特点,所以在多指标综合评价中都可以用来简化指标同时对筛选出来的指标赋权。主成分模型中的系数 a_{ij} 和因子得分模型中的系数 b_{ij} 可以看作指标 X_i 在第 j 个主成分或公共因子上的权重。只是应用的前提是原始指标的个数要足够多,并且有一定的相关性,这样才有必要、有可能进行主成分分析和因子分析。

(二) 在进行这两种分析之前都要进行前期处理,一般是数据正向化和标准化,处理之后可以选择从协差阵或相关系数阵开始分析;如果不处理原始数据就必须选择分析相关系数阵,否则协差阵会受指标数值或量纲不一致的影响[6]。

对于因子分析还要检验共同度(Communalities),只有公共因子对各变量的解释能力都很强的情况下,也就是"Communalities"表中"Extraction"栏中数值接近于"1",才能提取公共因子,否则就是变相丢失指标。

(三)命名问题。对主成分命名,应该根据$A_{p\times m}'$中系数绝对值大的变量,或直接依据初始因子载荷阵中系数绝对值大的变量。因子分析中应该把旋转因子载荷阵中第j列系数绝对值大的变量归为一组,并根据这组变量的含义对第j个因子命名。

(四)计算综合评价值。主成分分析$F_{综}=\sum_{i=1}^{m}(\lambda_i/k)F_i, k=\lambda_1+\lambda_2+\cdots+\lambda_m$;因子分析$F_{综}=\sum_{i=1}^{m}(v_i/k)Z_i, k=v_1+v_2+\cdots+v_m, v_j=\sum_{i=1}^{p}c_{ij}^2$,$c_{ij}$是旋转后的因子载荷[7]。综合评价值就是把主成分或公共因子的数值加权平均,权重是主成分或因子的贡献率。

四、结 论

通过比较主成分和因子分析在研究目的、数学模型、SPSS中的实现过程以及综合评价中的具体应用,看到两种方法的原理不同:主成分分析是把原来相关联的指标进行坐标轴旋转,使之不相关,且根据主成分方差衡量信息量并依此选取若干主成分;因子分析是把影响变量的公共因子按照重要程度提炼出来。二者的具体模型特征及许多细节也是不同的。

两种方法的联系在于:都通过对变量的提炼,以更简洁的方式认识复杂现象,在多指标综合评价中可以对变量赋权,计算综合评价值。在具体运算方面二者模型中的系数阵相互关联,并且主成分贡献率和初始因子贡献率相同。

参考文献:

[1] 于秀林,任雪松.多元统计分析[M].北京:中国统计出版社,1999:189-217.

[2] 王学民.应用统计分析[M].上海:上海财经大学出版社,2004:169-206.

[3] 王文博.计量经济学[M].西安:西安交通大学出版社,2004:202-205.

[4] 杨维权,刘兰亭,林鸿洲.多元统计分析[M].北京:高等教育出版社,1989.

[5] 卢纹岱.SPSS FOR WINDOWS 统计分析[M].北京:电子工业出版社,2002.

[6] 阮敏.主成分方法在经济管理综合评价应用中的误区[J].统计与决策,2005(4):23-24.

[7] 林海明,张文霖.主成分分析和因子分析的异同和SPSS软件[J].统计研究,2005(3):65-69.

(责任编辑:崔国平)

Comparison of Principal Component Analysis with Factor Analysis in Comprehensive Multi-indicators Scoring

WANG Wen-bo, CHEN Xiu-zhi

(School of Economics and Finance, Xi'an Jiaotong University, Xi'an 710061, China)

Abstract: The article compares the method of principal component analysis and factor analysis on with respect to aims, principle and realization in statistical software SPSS. Especially, it explains four points in comprehensive multi-indicators scoring in order to do practical analysis.

Key words: principal components; factor; difference