因子分析中公因子提取方法的比较与选择

王春枝

（内蒙古财经大学　统计与数学学院，内蒙古　呼和浩特　010070）

[摘　要]近代和现代统计分析方法中，因子分析是最重要的方法之一。因子分析中有 7 种提取公因子的方法，其中主成分法、极大似然法和主轴因子法是常用的方法。在着重解析这三种方法基本数学过程的基础上，对其适用条件和应用注意事项进行了比较，最后结合实例比较了不同公因子提取方法的结果，并对提取公因子过程中出现的问题给出了可能的解决办法。

[关键词]公因子；提取；因子分析

[中图分类号]O212　　[文献标识码]A　　[文章编号]2095－5871(2014)01－0090－05

一、因子分析的基本思想

因子分析，又称为探索性因素分析，1904 年首次由查尔斯·斯皮尔曼(Charles Spearman)提出，发展至今，该方法已经成为现代统计学的重要分支。因子分析是利用化简和降维的思想，对具有错综复杂关系的变量，根据其相关性对原始变量进行分组并根据分组的结果将多个变量综合成少数几个因子，以再现原始变量与因子之间的相互关系。其实质是探讨多个能够直接测量，并且具有一定相关性的实测指标如何受少数几个内在的独立因子所支配，同时以这些独立因子为框架分解原变量，并在一定条件下借以尝试对原变量进行分类。这些独立的因子又称为潜在因子，是不能观测的随机变量。

因子分析模型假定原始变量可以根据其相关性进行分组，即假设对于一个特定组内的所有变量彼此之间是高度相关的，而与不同组中的变量却有相对较小的相关性，这就意味着各组变量有一个潜在的结构或因子对该组变量观察到的相关性负责。例如，斯皮尔曼最初使用因子分析方法对学生的考试成绩进行研究时，发现学生的古典文学、法语、英语、数学、判别以及音乐测验成绩相关，这些成绩变量的相关性表明存在一个潜在的“智力”因子。因子分析方法就是要确认原始变量与潜在因子之间的这样一种结构是否存在。

二、因子分析模型

设 $x=(x_1,x_2,\cdots,x_p)'$ 是可观测的 P 个随机变量，当这 P 个随机变量之间存在较强的相关性时，因子分析模型会取得良好的结果。为了消除变量量纲的影响，需要对样本观测数据进行标准化处理。处理方法为：

$X_{ij}=\frac{x_{ij}-\bar{x}_j}{s_j}$，式中 $\bar{x}_j$、s_j 分别为指标 x_j 的样本平均值和样本标准差。将标准化后的原始向量记为 $X=(X_1,X_2,\cdots,X_p)'$，满足均值向量 $E(X)=\mu$，协方差矩阵 $Cov(X)=\Sigma$，并且协方差矩阵 Σ 与相关阵 R 相等。若

[收稿日期]2013－09－21

[作者简介]王春枝(1976－)，女，内蒙古巴彦淖尔人，内蒙古财经大学统计与数学学院副教授，硕士，从事应用统计研究.

$$\begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_p \end{bmatrix} = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_p \end{bmatrix} + \begin{bmatrix} \alpha_{11} & \alpha_{12} & \cdots & \alpha_{1m} \\ \alpha_{21} & \alpha_{22} & \cdots & \alpha_{2m} \\ \vdots & \vdots & & \vdots \\ \alpha_{p1} & \alpha_{p2} & \cdots & \alpha_{pm} \end{bmatrix} \begin{bmatrix} F_1 \\ F_2 \\ \vdots \\ F_p \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_p \end{bmatrix}$$

称 $F_1, F_2, \cdots, F_m (m < p)$ 为公共因子，是不可观测的变量，他们的系数称为因子载荷。ε 是特殊因子，是不能被前 m 个公共因子包含的部分，并且满足：

$Cov(F, \varepsilon) = 0$，即 F, ε 互不相关；

$D(F) = I$（单位矩阵），即 $F_1, F_2, \cdots, F_m$ 互不相关，方差为 1。

$$D(\varepsilon) = \begin{bmatrix} \sigma_1^2 & & & \\ & \sigma_2^2 & & \\ & & \ddots & \\ & & & \sigma_p^2 \end{bmatrix}$$，即 ε 的各分量之间互相独立。

可见，因子分析中首要的步骤就是确定公共因子的系数，即确定因子载荷矩阵

$$A = \begin{bmatrix} \alpha_{11} & \alpha_{12} & \cdots & \alpha_{1m} \\ \alpha_{21} & \alpha_{22} & \cdots & \alpha_{2m} \\ \vdots & \vdots & & \vdots \\ \alpha_{p1} & \alpha_{p2} & \cdots & \alpha_{pm} \end{bmatrix}。$$

这一工作在实践中可以通过很多方法来完成，常用的主要有：主成分法、不加权的最小平方法、广义最小二乘法、最大似然法、主轴因子法、α 因子法、映像因子法。这些方法求解因子载荷的方法不同，提取的公共因子的结果也不完全相同，甚至在数据结构上产生较大的差异。

三、公因子提取方法的比较与选择

（一）主成分法

该方法假设变量是各因子的线性组合，从原始变量的总体方差变异出发，尽量使原始变量的方差能够被主成分（公因子）所解释，并且使得各公因子对原始变量方差变异的解释比例依次减少。这种方法是实践中最常用的方法。

设样本协方差矩阵的特征根 $\lambda_1 \geqslant \lambda_2 \geqslant \cdots \geqslant \lambda_p \geqslant 0$，对应的单位正交特征向量为 $U_1, U_2, \cdots, U_p$。

设矩阵 $B = (\sqrt{\lambda_1}U_1 \quad \sqrt{\lambda_2}U_2 \quad \cdots \quad \sqrt{\lambda_p}U_p)$，则 R 可以分解为：$R = BB' = BB' + 0$ 为一个精确的因子分解式，使用了主成分方法。事实上，由 $U'RU = \Lambda \Rightarrow R = U\Lambda U' = U\Lambda^{1/2}\Lambda^{1/2}U'$，令 $B = U\Lambda^{1/2}$，即可以实现。

同时，因子分析还要求只选择 $q (q < p)$ 个因子，因此当后 $p - q$ 个特征根较小时，去掉矩阵 B 的最后几列。

将 B 分块：$B = (A \quad C)$，则

$$R = BB' = (A \quad C)\begin{pmatrix} A \\ C \end{pmatrix}' = AA' + CC' \approx AA' + \varphi^{①},$$

其中，

$$A = (\sqrt{\lambda_1}U_1 \quad \sqrt{\lambda_2}U_2 \quad \cdots \quad \sqrt{\lambda_p}U_p)$$

$$C = (\sqrt{\lambda_{q+1}}U_{q+1} \quad \sqrt{\lambda_{q+2}}U_{q+2} \quad \cdots \quad \sqrt{\lambda_p}U_p)$$

$$\varphi = \begin{pmatrix} \varphi_1 & & & \\ & \varphi_2 & & \\ & & \ddots & \\ & & & \varphi_p \end{pmatrix} = diag(\varphi_1, \varphi_2, \cdots, \varphi_p),$$

$\varphi_i = 1 - \sum_{j=1}^{q} a_{ij}^{2②}$。

相比其他方法，主成分法比较简单，但是这种方法得到的特殊因子 ε 的各分量之间不独立，并不满足因子分析模型的前提条件，所得到的因子载荷可能会产生较大的偏差，只有当各公因子的共同度较大时，特殊因子的影响作用才可以忽略不计。因子，在实际使用中，往往将主成分提取方法与其他方法结合使用。

该方法在实际使用中会遇到 q 的选择问题，一般可以采取 3 种原则：如果取 q 个因子后，使残差矩阵 $R - (AA' + \varphi)$ 的元素绝对值都很小，则认为该 q 值合适；借鉴主成分分析方法确定主成分个数的准则，根据公共因子的累积方差贡献率需要达到一定

的比例来选择，实践中一般要求累积方差贡献率达到85%以上；选择 $\lambda>1$ 的个数为公共因子数 q。但对这些原则不应生搬硬套，而是具体问题具体对待，最终的目的还是保证公因子能够有效地反映原始变量的数据结构和关系，并且能够合理地解释因子模型。

（二）主轴因子法

不同于主成分法从原始变量的变异出发，尽量使变量的方差能够被主成分解释，主轴因子法从变量的相关系数矩阵出发，使原始变量的相关程度尽可能地被公因子解释，该方法重在解释变量的相关性，确定内在结构，而对于变量方差的解释相对则不太重视，所以当研究的目的重在确定结构，而对变量方差的情况不太关心时可以使用此方法。该方法的基本步骤是：

1. 给出共同度 h_i^2 的初始估计值 $h_i^{*2}(i=1,2,\cdots,p)$；

2. 由 h_i^{*2} 求出 $\varphi_i^*=1-h_i^{*2}(i=1,2,\cdots,p)$，并求出约化相关系数矩阵 $R^*=R-\varphi^*$，其中

$$R^*=R-\varphi^*=\begin{pmatrix} h_1^{*2} & & & \\ r_{21} & h_2^{*2} & & \\ \vdots & \vdots & & \vdots \\ r_{p1} & r_{p2} & \cdots & h_p^{*2} \end{pmatrix};$$

3. 由方程 $|R^*-\lambda^* I|=0$ 求约化相关系数矩阵 R^* 的前 q 个特征根 $\lambda_1^*,\lambda_2^*,\cdots,\lambda_p^*>0$ 及对应的前特征向量 $U_1^*,U_2^*,\cdots,U_q^*$，于是有

$$R^*=R-\varphi^*$$

$$\approx(U_1^*,U_2^*,\cdots,U_q^*)\begin{pmatrix} \lambda_1^* & & & \\ & \lambda_2^* & & \\ & & \ddots & \\ & & & \lambda_p^* \end{pmatrix}\cdot$$

$$(U_1^*,U_2^*,\cdots,U_q^*)'$$

且令 $A_1=\Gamma_1\Lambda_1^{1/2}=(\sqrt{\lambda_1^*}U_1^* \quad \sqrt{\lambda_2^*}U_2^* \quad \cdots \quad \sqrt{\lambda_q^*}U_q^*)$；

4. 求出 φ 的估计 $\varphi^*(1)=R-A_1A_1'$；

5. 返回第二步用 $\varphi^*(1)$ 代 φ^*，直到 A 的值和 φ 的值达到稳定为止。

该方法在实际使用中也会遇到 q 的选择问题。尽管 R 是非负定阵，但由其得到的约化相关矩阵 R^* 却不一定是非负阵，可能会有负特征根。这时，正的特征根之和将超过总共同度（即 R^* 的迹），因为全部特征根之和就是总共同度。常用的确定 q 的办法是按特征根由大至小的次序抽取，直到 $\sum_{i=1}^{q}\lambda_i^*$ 与 $\sum_{i=1}^{p}\lambda_i^*=tr(R-\varphi^*)=tr(R^*)=\sum_{i=1}^{p}h_i^{*2}$ 接近为止。这样确定的 q 不超过正特征根的个数；

该方法另外一个关键是给出初始的共同度的估计值，然后才能据此计算出因子载荷矩阵。常用的方法有：

方法一：取 $\varphi_i^*=\frac{1}{r^{ii}}(i=1,2,\cdots,p)$，其中 r^{ii} 是相关系数矩阵 R 的逆矩阵 R^{-1} 中的主对角线元素。因此有 $h_i^{*2}=1-\varphi_i^*=1-\frac{1}{r^{ii}}$；

方法二：取 h_i^{*2} 为第 i 个变量 X_i 与其它所有变量 X_j 的复相关系数的平方，即 X_i 对其余的 $p-1$ 个 X_j 的回归方程的判定系数，这是因为 X_i 与公共因子的关系是通过其余的 $p-1$ 个的 X_j 线性组合联系起来的；

方法三：取 h_i^{*2} 为第 i 个变量 X_i 与其它所有变量 X_j 的相关系数的最大值（绝对值），即

$h_i^{*2}=\max_{j\neq i}|r_{ij}|$，其中 r_{ij} 为变量 X_i 与 X_j 的相关系数；

方法四：取 $h_i^{*2}=1$，在这种情况下主轴因子解与主成分解等价；

（三）极大似然法

极大似然估计法要求公共因子和特殊因子服从正态分布。设 $X=(X_1,X_2,\cdots,X_p)'$ 为来自正态总体 $N_p=(\mu,\Sigma)$ 的随机样本，$\Sigma=AA'+\Sigma_\varepsilon$。根据似然函数的理论有：

$$L(\hat{m},\hat{A},\hat{D})$$

$$=\prod_{i=1}^{n}(2\pi)^{-p/2}|\Sigma|^{1/2}\exp[-\frac{1}{2}(x_i-\mu)'\Sigma^{-1}(x_i-\mu)]$$

$$=[(2\pi)^p|\Sigma|]^{-n/2}\exp[-\frac{1}{2}\sum_{i=1}^{n}(X_i-\mu)'\Sigma^{-1}(X_i-\mu)]$$

它通过依赖于 A 和 Σ_ε，但上式并不能唯一确定 A，因此添加条件 $A'\Sigma_\varepsilon^{-1}A=\Lambda$，其中 Λ 是一个对角阵，数值极大化的方法可以得到极大似然估计 $\hat{A}$ 和 $\hat{\Sigma}_\varepsilon$。$\hat{A}$、$\hat{\Sigma}_\varepsilon$ 和 $\hat{\mu}=\bar{x}$ 将是矩阵 $\hat{A}'\hat{\Sigma}_\varepsilon^{-1}\hat{A}=\hat{\Lambda}$ 为对角矩阵，且似然函数达到最大。相应的共同度的似然估计为：$\hat{h}_i^2=\hat{a}_{i1}^2+\hat{a}_{i2}^2+\cdots+\hat{a}_{im}^2$，第 j 个因子对总方差的贡献为 $S_j^2=\hat{a}_{1j}^2+\hat{a}_{2j}^2+\cdots+\hat{a}_{pj}^2$。

该方法不要求数据服从正态分布，在样本量较大时使用效果较好，特别是当样本量极大（1500 以上），结果会更为精确。

（四）其他方法

因子分析方法中，除了上述 3 种常用的公因子提取方法之外，还有广义最小二乘法、未加权最小平方法、α 因子法、映像因子法。其中广义最小二乘法法根据变量值进行加权，使实际的相关阵和再生的相关阵之差的平方和最小；未加权最小平方法不对变量值进行加权，使实际的相关阵和再生的相关阵之差的平方和最小；α 因子法将变量看成是从潜在变量空间中抽取出来的样本，在计算中尽量使变量的 α 信度达到最大；映像因子法把一个变量表示成是其他变量的多元回归方程，据此提取公因子。

实际上，如果进行因子分析所采用的变量数和样本量都比较大，而且各原始变量之间的相关性较强，各种因子法提取的结果基本相同，区别仅仅在于各种方法分析的思想存在差异。主成分法在大多数情况下分析的效果是最佳的；如果样本少、变量少，α 因子法或映像因子法是可取的选择。

四、不同公因子提取方法的实例比较

本文以参考文献[2]的数据资料进行因子分析，该资料选取了我国 30 个省市自治区 2008 年反映经济发展基本情况的八项指标，分别是国内生产总值 X_1、居民消费水平 X_2、固定资产投资 X_3、职工平均工资 X_4、货物周转量 X_5、居民消费价格指数 X_6、商品零售价格指数 X_7、工业总产值 X_8，利用 SPSS 软件中 Factor 过程对各地区经济发展水平进行因子分析，结果如表 1 所示。

表 1　　不同公因子提取方法结果对比

	公因子共同度						
	主成分法	主轴因子法	极大似然法	广义最小二乘法	未加权最小平方法	α 因子法	映像因子法
国内生产总值 X_1	0.945	0.965	0.999	0.975	0.965	0.955	0.966
居民消费水平 X_2	0.799	0.646	0.618	0.757	0.651	0.648	0.603
固定资产投资 X_3	0.902	0.874	0.966	0.964	0.872	0.855	0.947
职工平均工资 X_4	0.873	0.838	0.831	0.737	0.837	0.838	0.702
货物周转量 X_5	0.857	0.730	0.707	0.737	0.724	0.730	0.681
居民消费价格指数 X_6	0.957	0.901	0.718	0.779	0.999	0.898	0.708
商品零售价格指数 X_7	0.928	0.886	0.999	0.816	0.835	0.894	0.770
工业总产值 X_8	0.904	0.851	0.767	0.898	0.853	0.874	0.783
大于 1 的特征根个数	3	3	3	3	3	3	3
累计方差贡献率	89.551	83.649	82.564	89.550	84.194	83.650	77.004

表 1 中，公因子共同度反映了公因子对原始指标信息的提取程度，按照特征根大于 1 原则提取公因子，可以看到各种方法大于 1 的特征根个数均为 3，所以均提取 3 个公因子。将原始指标的信息看作

是100%，对第一个原始指标国内生产总值 X_1 的信息提取度最高的是极大似然法，3 个公因子反映了原始指标国内生产总值 $X_1$99.9% 的信息。累计方差贡献率反映了公因子对原始指标总体信息的提取程度。以主成分法为例，3 个公因子共提取了原始 8 个指标 89.551% 的信息，其中提取了国内生产总值 $X_1$94.5% 的信息，提取了居民消费水平 $X_2$79.9% 的信息，提取了固定资产投资 $X_3$90.2% 的信息，其余指标以此类推。

从表 1 还可以看到各种公因子提取方法对第二个原始指标居民消费水平 X_2 的信息提取都是比较低的，最高主成分法仅有 79.9%，映像因子法仅提取了 60.3% 的信息，这可能是由于居民消费水平 X_2 与其他变量的相关性较低造成的，利用 SPSS 输出相关系数矩阵如表 2 所示。

从表 2 可以看到，居民消费水平与国内生产总值、货物周转量、居民消费价格指数在 95% 的置信度下不存在显著相关性，这时可以考虑去掉或者替换该指标。

公因子提取完成后，可以进行旋转对各变量之间的结构关系进行探索，也可以根据因子得分对各地区经济发展水平进行综合评价，本文不再赘述。

表 2　居民消费水平与其他变量的相关系数

	居民消费水平 X_2
国内生产总值 X_1	0.267
固定资产投资 X_3	0.426*
职工平均工资 X_4	0.716**
货物周转量 X_5	-0.151
居民消费价格指数 X_6	-0.235
商品价格指数 X_7	-0.593**
工业总产值 X_8	0.363*

注：用“**”标识的相关系数在 99% 的置信度下显著，用“*”标识的相关系数在 95% 的置信度下显著.

[注　　释]

① 该式即为因子分析模型的主成分解. 与 φ 的主对角线元素是相等的，而非主对角线上的元素却不相等.

② 与 φ 的最大差异就是忽略了其中的非主对角线元素.

[参考文献]

[1] 何晓群. 多元统计分析[M]. 北京：中国人民大学出版社，2012：165.

[2] 张文彤. SPSS 统计分析高级教程[M]. 北京：高等教育出版社，2004：226－227.

[3] 朱星宇，陈勇强. SPSS 多元统计分析方法及应用[M]. 北京：清华大学出版社，2011：251.

[责任编辑：姚志峰]

The Comparison and Selection of the Common Factors Extraction Method in Factor Analysis

WANG Chun－zhi

(Inner Mongolia University of Finance and Economics, Hohhot 010070, China)

Abstract: Factor Analysis is one of the most important methods in modern and contemporary statistical analysis methods. Generally, seven statistical methods are used for extracting the common factors in factor analysis, and the principal component analysis, the maximum likelihood method and the principal axis factoring method are the most commonly used. The paper mainly analyzes the basic Mathematical process of the three methods, and compares the conditions of application of these methods. Finally, the paper compares the results of the seven methods extracting the common factors, and gives the possible solutions to the problems In the application of factor analysis process.

Key words: the common factors; extraction; factor analysis