主成分分析中的统计检验问题

文/傅德印

摘 要：主成分分析已经越来越成为人们广泛应用的多元统计分析方法。但应用中盲目套用主成分分析方法的情况很多，而对主成分分析的适用性，主成分个数的合理性等问题重视不够，更谈不上对主成分分析进行统计检验。为此，为了更好应用主成分分析，就应对主成分分析结果进行统计检验并建立统计检验体系。主成分分析统计检验体系主要包括：主成分分析的适用性检验；等相关性检验；主成分方差的假设检验和选取主成分数目检验。

关键词：主成分分析；假设检验；巴特莱特检验

主成分分析已经越来越成为人们广泛应用的多元统计分析方法。特别是在系统综合评价，变量子集合的选择以及主成分回归中都取得了大量的应用成果。但是，应用中盲目套用主成分分析方法的情况很多，而对主成分分析的适用性，主成分个数的合理性等问题重视不够，更谈不上对主成分分析进行统计检验问题。为此，本文拟对主成分分析的有关统计检验问题，如主成分分析是否需要统计检验？若需要则进行哪些检验？检验怎样的假设？如何进行这些统计检验等问题进行探讨，以便抛砖引玉，供同行们做进一步研究。

一、主成分分析是否需要统计检验

主成分分析采取一种数学降维的方法，找出几个综合变量来代替原来众多的变量，使这些综合变量能尽可能地代表原来变量的信息量，而且彼此之间互不相关。这种将多个变量化为少数几个互不相关的综合变量的统计分析方法就叫做主成分分析或主分量分析。

从主成分的导出和计算上看，主成分是从原始数据的协方差矩阵或者相关系数矩阵出发，根据主成分应该满足的条件导出的。即主成分的协方差矩阵应该是一个对角矩阵，主成分表达式系数矩阵 A 应该是一个正交矩阵为条件，导出主成分的协方差矩阵的对角线元素是协方差矩阵或相关矩阵的特征值，主成分的方差就是原始数据协方差矩阵或相关矩阵的特征值，主成分表达式系数就是协方差矩阵或相关矩阵特征值对应的特征向量。主成分是原变量的线性组合，是对原变量信息的一种提取，主成分不增加总信息量，也不减少总信息量，只是对原信息进行了重新分配。从主成分的应用上看，求解主成分可以解决原始数据的相关性问题，可以实现数据的降维。正因为上述情况，所以常导致人们认为：主成分分析的主成分数和原始变量数相等，它是将一组具有相关性的变量变换为一组独立的变量，严格上不能作为一个模型来描述，它只能作为通常的变量变换，是一种变量变换行为，不涉及原来假设问题，所以不需要进行假设检验。

事实上，从理论上，主成分分析包括总体主成分分析和样本主成分分析，在实际问题中，总体协方差矩阵或相关矩阵都是未知的，都需要样本来估计，就必然涉及统计检验问题。而且在主成分分析的具体应用中，变量变换是一种手段，变量变换的最终目的是为了根据实际情况，最终要选择重要的信息量（即前几个主成分），以便在此基础上，进行进一步的分析。要进行这样的分析，实际上隐含了原始变量中存在着并且能够综合出重要信息的假设，为此就需要对相应的假设进行统计检验。

二、主成分分析的统计检验体系

（一）主成分分析适用性检验

并非所有的截面数据都适用于主成分分析的。主成分分析本身并不是目的，实际应用中主成分分析往往是一种手段。目的是通过主成分分析简化数据结构，在此基础上进行进一步的分析。因此，使用主成分分析的前提条件是原始数据各个变量之间应有较强的线性相关关系。如果原始变量之间的线性相关程度很小，它们之间不存在简化的数据结构，这时进行主成分分析实际是没有意义的。所以，应用主成分分析时，首先要对其适用性进行统计检验。主成分分析适用性检验的假设就是原始变量之间存在着较强的线性相关。具体检验方法有：

1、巴特莱特球性检验

巴特莱特球性检验（Bartlett test of sphercity）是从整个相关矩阵出发进行的检验，检验的原假设是相关矩阵为单位矩阵，如果不能拒绝原假设，说明原始变量之间相互独立，不适合进行主成分分析。事实上，如果原始数据的相关

矩阵是一个单位矩阵，各个原始变量之间互不相关，这时进行主成分分析，则得到的主成分就是各个原始变量自身，显然是不适合进行主成分分析的。

巴特莱特球性检验的统计量。巴特莱特球性检验的理论依据源于多元正态总体协方差矩阵的检验理论。协方差矩阵的检验主要内容包括：对总体协方差矩阵 Σ 与已知矩阵 Σ_0 相等的检验，对总体协方差矩阵 Σ 中的元素是否均为已知协方差矩阵 Σ_0 中元素的 σ^2 倍的检验，以及检验多个总体的协方差矩阵都相等的检验等。其中，对总体协方差矩阵 Σ 中的元素是否均为已知协方差矩阵 Σ_0 中元素的 σ^2 倍的检验，其原假设为，$H_0:\Sigma=\sigma^2\Sigma_0$，$H_1:\Sigma\neq\sigma^2\Sigma_0$，其中，$\Sigma$ 为总体协方差矩阵，$\Sigma_0>0$ 为已知，σ^2 未知。这时检验的似然比统计量为：

$$\lambda=\frac{\left|\Sigma_0^{-1}A\right|^{n/2}}{\left[\mathrm{tr}(\Sigma_0^{-1}A)/p\right]^{pn/2}},\quad A=\sum_{i=1}^{n}(x_i-\bar{x})(x_i-\bar{x})'$$

或等价于 $W=(\lambda)^{2/n}=\dfrac{p^p\left|\Sigma_0^{-1}A\right|}{\left[\mathrm{tr}(\Sigma_0^{-1}A)\right]^p}$

应用中，当 n 很大时，则利用如下的近似公式进行的是 χ^2 检验：

$$[(n-1)-(2p^2+p+2)/6p]\ln W=\chi^2_{(p+2)(p-1)/2}$$

在上述检验的基础上，如果令 $\Sigma_0=I_p$ 时，则等价于检验相关矩阵为单位矩阵，即 $\rho=I_0$ 于是原假设为 $H_0:\Sigma=\sigma^2I$，$H_1:\Sigma\neq\sigma^2I$，这时，检验的似然比统计量为：

$$\Lambda=\frac{|S|^{n/2}}{\left[\mathrm{tr}(S)/p\right]^{n/2}}=\left[\frac{\prod_{i=1}^{p}\hat{\lambda}_i}{\left(\frac{1}{p}\sum_{i=1}^{p}\hat{\lambda}_i\right)^p}\right]^{n/2}=\left[\frac{\text{几何均值}\hat{\lambda}_i}{\text{算数均值}\hat{\lambda}_i}\right]^{np/2}$$

其中，S 是样本协方差矩阵，$S=A/(n-1)$，p 是变量个数，n 表示样本容量，λ_i 表示样本协方差矩阵 S 的特征值。

在假设成立下，对大容量样本，巴特莱特认为 $-2[1-(p^2+p+2)/6pn]\ln\Lambda$ 近似服从 $\chi^2_{(p+2)(p-1)/2}$，因此，该大样本的 α 临界值为 $\chi^2_{(p+2)(p-1)/2}(\alpha)$。

进行检验时，也可以根据检验统计量公式计算得概率 P 值，概率 P 值小于 0.05 时则拒绝原假设，认为原始数据适合进行主成分分析，相反，概率 P 值大于 0.05 时，将不适合进行主成分分析。由于当 $\Sigma=\sigma^2I$ 时，常数密度的轮廓线是球面，因此这个检验就称为巴特莱特球性检验。

另外，巴特莱特球性检验统计量的计算公式有时还采用如下近似公式：

$$\chi^2=\frac{(11+2p-6n)}{6}\ln|R|$$

这里，自由度为 $p(p-1)/2$，p 是变量个数，n 表示样本容量，$|R|$ 为相关系数矩阵。应用时，为配合巴特莱特球性检验，有时还需要直接计算 $\ln|R|$ 值，根据 $\ln|R|$ 值的大小进行检验判断主成分分析的适用性。若进行主成分分析使用的是协方差矩阵，则 $\ln|R|$ 就取协方差矩阵行列式的自然对数值。

2、相关系数矩阵的直观检验

相关系数矩阵的直观检验是直接根据相关系数矩阵中所反映出的原始变量之间的线性相关大小来检验主成分分析的适用性。具体使用有两种矩阵：

(1) 根据简单相关矩阵进行直观检验。计算出简单相关矩阵后，对各个变量之间的简单相关系数进行一般的分析观察，如果相关矩阵的大部分相关系数都小于 0.3，原始数据之间的相关关系不大，则不适合进行主成分分析。相反，则适合进行主成分分析。

(2) 根据反映像相关矩阵进行直观检验。反映像相关矩阵(Anti-image correlation matrix) 检验是通过偏相关系数矩阵进行的检验。反映像相关矩阵是由元素等于负的偏相关系数形成的矩阵。偏相关系数是控制其它变量不变，来测量一个自变量对因变量的独特解释作用的相关系数指标。如果反映像相关矩阵中的很多元素值比较大时，应该考虑该原始变量不适合进行主成分分析。

3、KMO 检验和 MSA

KMO (Kaiser-Meyer-Olkin-Measure of Sampling Adequacy) 检验是从比较原始变量之间的简单相关系数和偏相关系数的相对大小出发来进行的检验。当所有变量之间的偏相关系数的平方和，远远小于所有变量之间的简单相关系数的平方和时，变量之间的偏相关系数很小，KMO 值接近 1，变量适合进行主成分分析。KMO 值的计算公式为：

$$KMO=\frac{\sum\sum_{i\neq j}r_{ij}^2}{\sum\sum_{i\neq j}r_{ij}^2+\sum\sum_{i\neq j}\alpha_{ij}^2}$$

这里 r_{ij} 表示简单相关系数，$\alpha^2_{ij.1,2,3,\cdots k}$ 表示偏相关系数。显然，当 $\alpha^2_{ij.1,2,3,\cdots k}\approx 0$ 时，$KMO\approx 1$；当 $\alpha^2_{ij.1,2,3,\cdots k}\approx 1$ 时，$KMO\approx 0$，KMO 的取值介于 0 和 1 之间。Kaiser 给出了一个 KMO 的度量标准。

表 1　　KMO 度量标准表

KMO 值	分析的适用性
0.90～1.00	非常好
0.80～0.89	好
0.70～0.79	一般
0.60～0.69	差
0.50～0.59	很差
0.00～0.49	不能进行分析的

实际应用中，还可以计算与 KMO 值类似的指标 MSA (Measures of Sampling Adequacy)，如在 SPSS 软件中，MSA 值计算结果显示在反映象相关矩阵的主对角线上。MSA 计算公式为：$MSA_i=\dfrac{\sum_{i\neq j}r_{ij}^2}{\sum_{i\neq j}r_{ij}^2+\sum_{i\neq j}\alpha_{ij}^2}$　i=1,2,…p

其中 p 表示自变量个数。

与 KMO 值比较，MSA 值的求和项中并没有包含所有变量的相关系数和偏相关系数，而只包含了某个变量所涉及到的另外的(p-1)个变量。即根据每个原始变量与其它变量

的相关系数和偏相关系数计算一个 MSA 值，共有 p 个 MSA 值，而 KMO 值是根据所有变量的相关系数和偏相关系数计算的，只有一个总比较值。因此 MSA 值反映的内容是按照各个变量进行的，更为详细。MSA 值的使用标准同 KMO 值，越接近1，说明原始变量越适合做因子分析，一般应在 0.50 以上。

4、φ 检验

φ 值是关于多个变量之间线性相关性大小的 Gleason-Staelin 的测度。具体计算公式为，

$$\varphi=\sqrt{\frac{\sum_{i=1}^{p}\sum_{j=1}^{p}r_{ij}^{2}-p}{p(p-1)}}$$

当 φ=0 时，表明多个变量之间没有相关性，当 φ=1 时，表明多个变量之间存在完全线性相关。因此，φ 值越接近 1，越适合做主成分分析。但应用 φ 值时需要注意，一是有时各个变量之间存在明显的相关性时，φ 值也可能会小于 0.5；二是，一般是在两个以上的变量时计算并使用 φ 值时，使用效果会更好。

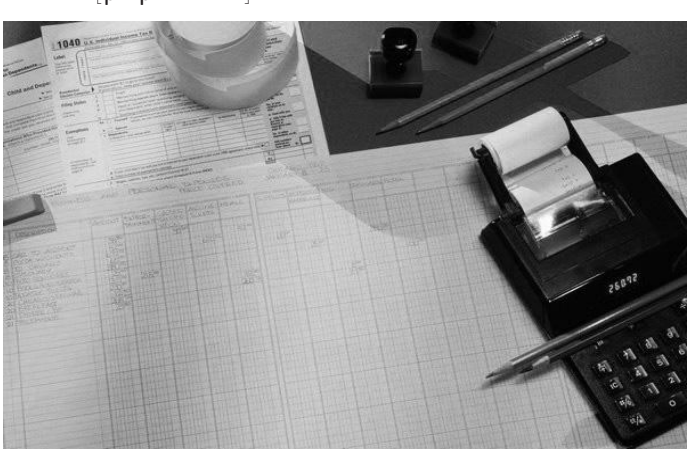
5、碎石图的直观检验

碎石图是根据原始数据相关矩阵特征值的大小即主成分方差大小的顺序，画出的主成分方差随主成分个数变化的散点图。根据碎石图的形状也可以对主成分分析的适用性进行判断。碎石图的形状理论上应该像个山崖，从第一个主成分开始，曲线迅速下降，然后下降变得平缓，最后变成为近似一条直线，近似直线上的散点就像山脚下的碎石，该图因此得名。显然，碎石图的弯曲的程度越明显，越像个山崖，越适合进行主成分分析，相反，碎石图从开始就近似为一条直线，则说明不适合进行主成分分析。

（二）等相关性检验

等相关性检验是检验各变量间的相关系数不全相等或不相等的假设。如果 p 个变量间的相关系数都相等，这时的相关系数矩阵的特征值不是不同的，则无法求出 p 个主成分，也无法进行主成分分析。即检验：

$$H_0:R=R_0=\begin{bmatrix}1 & \rho & \cdots & \rho\\ \rho & 1 & \cdots & \rho\\ & \cdots & \cdots & \\ \rho & \rho & \cdots & 1\end{bmatrix},\ H_0:R\neq R_0$$



该检验的理论依据仍然是源于多元正态总体协方差矩阵的检验理论，用似然比统计量导出一个检验。Lawley (1963) 年提出一个等价的检验方法，它是从相关矩阵的非对角线元素导出的。记：

$$\bar{r}=\frac{2}{p(p-1)}\sum_{i<k}\sum r_{ik},$$

$$\gamma=\frac{(p-1)^2[1-(1-\bar{r})^2]}{p-(p-2)(1-\bar{r})^2}$$

这里 i 表示相关矩阵的行，k 表示相关矩阵的列，$\bar{r}_k$ 是相关矩阵第 k 列对角线元素的平均值，而 $\bar{r}$ 是相关矩阵中所用非对角线元素的平均值。检验的统计量为：

$$T=\frac{n-1}{(1-\bar{r})^2}[\sum_{i<k}\sum(r_{ik}-\bar{r})^2-\gamma\sum_{k=1}^{p}(\bar{r}_k-\bar{r})^2]$$

在样本容量 n 较大时，统计量 T 近似服从自由度为 (p+1)(p-2)/2 的 χ^2 分布。T 值大于给定的临界值，则拒绝原假设，认为各变量间的相关系数不全相等或不相等的假设，适合进行主成分分析。

（三）主成分方差的假设检验和置信区间

主成分方差的假设检验实际是对相关矩阵特征值的假设检验。检验的假定是相关矩阵的特征值不同且皆为正值，即 $\lambda_1>\lambda_2>\cdots>\lambda_p>0$。Anderson(1963) 建立了特征值的大样本性质。样本含量较大时，各个特征值 λ_i 相互独立，且近似服从正态分布 $N(\lambda_i,2\lambda_i/n)$。于是利用正态分布的性质，可以对特征值进行相应检验，而且得到 $100(1-\alpha)\%$ 的置信区间为：

$$\left(\frac{\lambda_i}{1+u_{\alpha/2}\sqrt{2/n}},\frac{\lambda_i}{1-u_{\alpha/2}\sqrt{2/n}}\right)$$

其中，$u_{\alpha/2}$ 为标准正态分布的分位数。若用 $u_{\alpha/2p}$ 代替上式中的 $u_{\alpha/2}$，就得到 p 个 λ_i 的庞费罗尼 (Bonferroni) 型 $100(1-\alpha)\%$ 联合置信区间。

（四）主成分个数的选取和检验

1、巴特莱特检验（主成分相等的检验）

在判断取几个主成分时，经常要检验最后 (p-k) 个主成分是否为 0，如果后边的 (p-k) 个特征值均与 0 无统计显著差异则不考虑，而只取特征值不为 0 的前 k 个主成分。为此：

$H_0:\lambda_p=\lambda_{p-1}=\cdots=\lambda_{k+1}=0$

$H_1:\lambda_p,\lambda_{p-1},\cdots,\lambda_{k+1}$ 不为 0 或不全为 0。

或更广义的：

$H_0:\lambda_p=\lambda_{p-1}=\cdots=\lambda_{k+1}=\lambda_0$

$H_1:\lambda_p,\lambda_{p-1},\cdots,\lambda_{k+1}$ 不为 0 或不全为 0。

记：$Q=(\prod_{i=k+1}^{p}\lambda_i).[\prod_{i=k+1}^{p}\lambda_i/(p-k)^{-(p-k)}]$

当假设成立时，$-n\ln Q$ 近似服从自由度为 $(p-k-1)(p-k+2)/2$ 的 χ^2 分布，即

$$\chi^2=-n\ln Q\sim\chi^2_{(p-k-1)(p-k+2)/2}$$

或令：

$$\tilde{Q}=-\{n-k-\frac{1}{6}[2(p-k)+1\frac{2}{p-k}]+\bar{\lambda}^{-2}\sum_{i=1}^{k}(\lambda_i-\bar{\lambda})^2\}.\ln Q$$

式中，$\bar{\lambda}$ 是 p-k 后个特征值的均数。则 $\tilde{Q}$ 近似服从

自由度为(p- k)(p- k+1)/2- 1 的 χ^2 分布，即$\tilde{Q}\sim\chi^2_{(p-k)(p-k+1)/2-1}$。

2、累计贡献率法

根据经验，主成分的累计贡献率达到 80%或 85%以上时，主成分数目 k 就可以了。累计贡献率为：

累计贡献率$=\sum_{i=1}^{k}\lambda_i/\sum_{i=1}^{p}\lambda_i$，这里 k 为主成分数目。

3、特征值平均数法

计算特征值的均数$\bar{\lambda}$，选择大于$\bar{\lambda}$之的特征值对应的主成分。其中，当基于相关矩阵进行主成分分析时，由于得到的特征值其平均数等于 1，所以保留大于 1 的特征值对应的主成分。

上述确定和检验主成分数目的方法中巴特莱特检验方法是最具有推理性的，但是，在实际应用中，还没有一个统一的标准或原则，为此，最好是把上述几种方法结合起来应用更为合适。

三、几点讨论

将上述主成分分析的适用性检验、等相关性检验、主成分方差的假设检验和选取主成分数目检验结合起来，则可以形成主成分分析统计检验体系的基本框架。为了恰当应用主成分分析，充分发挥主成分分析的作用，建议在应用时应按照上述框架进行相应的统计检验。当然，上述体系一方面还需要有待完善之处，另一方面，在实际应用中，即便是根据上述内容都进行了统计检验，但也还会存在许多考虑不周之处，还要注意以下问题：

1、关于原始数据变量的总体分布问题。主成分分析主要依赖于变量的协方差矩阵或者相关矩阵，一般与变量的分布无关，因此对总体的分布没有特殊的要求。

2、主成分的计算是根据协方差矩阵还是相关矩阵进行。主成分分析可以根据协方差矩阵进行，也可以根据相关矩阵进行，但二者的主成分计算结果往往是不同的，特别是当各个变量的变异相差较大时。但根据经验，实际应用中，一般都是根据相关矩阵来进行计算主成分，除非是专门针对协方差矩阵进行的分析。

3、原始变量之间的非线性关系问题。主成分分析基于相关矩阵来计算时，这里的相关矩阵实际上是 Pearson 的积差相关，反映的是变量之间的线性相关关系，但是，如果变量之间的关系不是线性的，而是非性相关关系，这时，由于 Pearson 的积差相关矩阵已经无法准确表达原始变量之间的关系，于是，所进行的分析以及结论也就失去应有的意义了。

4、样本容量 n 问题

进行主成分分析时，样本容量达到多少为宜，目前尚没有统一的结论。有的认为样本容量应是变量个数的 10~20 倍，有的认为样本容量要在 100 以上比较合适，有的认为进行巴特莱特检验时的样本容量应该大于 150 方可，也有的认为不必苛求太多的样本容量，当原始变量之间的相关性很小时，即使再扩大样本容量，也难以得到主成分分析的满意效果。但有一点这里必须要肯定的，即样本容量一定要大于原始变量个数。

总之，主成分分析本身不是目的，往往是为了达到分析目的的一种手段。主成分分析用于多元回归，可以解决变量的多重共线问题，用于因子分析、聚类分析、判别分析等中可以实现降维，减少变量个数，用于综合评价除了解决变量相关和降维问题外，还可以提供综合变量的权数，应用中要针对目的加以选择的同时要进行必要的统计检验。

参考文献：

[1]罗积玉，邢　瑛.经济统计分析及预测[M].北京：清华大学出版社，1987:P110- 115.

[2]张尧庭，方开泰.多元统计分析引论[M].北京：科学出版社，1982:P322- 327.

[3]方开泰编著.实用多元统计分析[M].上海：华东师范大学出版社，1989:P302- 312.

[4]于秀林.多元统计分析及程序[M].北京：中国统计出版社，1993:P158- 161.

[5]Richard A Johnson, Dean W Wichern. Applied Multivariate Statistical Analysis. 4th Edition. Englewood Cliffs,N J:Prentice-Hall,Inc,1998,p452- 455.

[6]陈峰编.医用多元统计分析方法(第 1 版)[M].北京：中国统计出版社，2000:P160- 191.

[7]高惠璇.应用多元统计分析[M].北京：北京大学出版社，2005:P216- 264.

作者单位：兰州商学院

（责任编辑：陈晓卫）