合理进行多元分析——探索性因子分析

胡纯严[1]，胡良平[1,2*]

（1. 军事科学院研究生院，北京 100850；
2. 世界中医药学会联合会临床科研统计学专业委员会，北京 100029
*通信作者：胡良平，E-mail：lphu927@163.com）

【摘要】 本文目的是介绍与探索性因子分析有关的基本概念、计算方法、两个实例以及使用SAS实现计算的方法。基本概念包括公因子与特殊因子、因子载荷、公因子方差和因子碎石图；计算方法涉及因子分析的数学模型、主成分法、主因子法、质心法和最大似然法；两个实例的资料分别是“31个省级政府门户网站的评估数据”和“全国各地区基本建设投资来源”；借助SAS软件，对两个实例中的数据进行了探索性因子分析，并对SAS输出结果做出了解释。

【关键词】 因子模型；因子载荷；因子轴旋转；主成分法；最大似然法

中图分类号：R195.1 文献标识码：A doi：10.11886/scjsws20230925001

Reasonably carry out multivariate analysis：exploratory factor analysis

Hu Chunyan[1], *Hu Liangping*[1,2*]

（1. *Graduate School, Academy of Military Sciences PLA China, Beijing* 100850, *China*;
2. *Specialty Committee of Clinical Scientific Research Statistics of World Federation of Chinese Medicine Societies, Beijing* 100029, *China*
*Corresponding author: *Hu Liangping*, *E-mail*: *lphu*927@163.*com*）

【Abstract】 The purpose of this article was to introduce the basic concepts, calculation methods, two examples related to the exploratory factor analysis, and how to implement calculations using SAS. Basic concepts included common factors and special factors, factor loadings, common factor variances and factor scree plots. The calculation methods involved the mathematical model of factor analysis, principal component method, principal factor method, centroid method and maximum likelihood method. The data in the two examples were "evaluation data of 31 provincial government portal websites" and "sources of basic construction investment in various regions across the country". With the help of SAS software, the exploratory factor analysis was conducted on the data in the two examples, and an explanation was given to the SAS output results.

【Keywords】 Factor model; Factor loading; Factor axis rotation; Principal component method; Maximum likelihood method

因子分析就是找出某个问题中可直接测量的、具有一定相关性的诸指标（即变量）如何受少数几个在专业上有意义、又不可直接测量到、且相对独立的因子支配的规律，从而可用诸指标的测定值间接确定诸因子的状态。本文将介绍与因子分析有关的基本概念、计算方法、两个实例、使用SAS分析实例的方法以及需要注意的问题。

1 基本概念

1.1 公因子与特殊因子

公因子是隐藏在所有显变量的背后且支配着所有显变量的隐变量。实际上，公因子就是主成分分析中的“主成分变量（简称为主成分）”。特殊因子是每个显变量都受到一个特殊因子的控制，该因子与公因子和其他显变量之间互相独立。

1.2 因子载荷

在用若干个公因子线性表达每个显变量的表达式中，公因子之前的系数被称为因子载荷或权重。因子载荷的绝对值越大，表明所对应的公因子对显变量的控制力越强；因子载荷的正负号反映所对应的公因子对特定显变量的作用方向相同（正号）还是相反（负号）。

1.3 公因子方差

设有m个显变量受p个公因子控制（$p<m$），在初始因子模型中，用p个公因子线性表达每个显变量，p个公因子前面的系数（即因子载荷）的平方和被称为公因子方差。

1.4 因子碎石图

为了达到降维的目的，研究者希望尽可能减少公因子的个数。如何估计出比较合适的公因子数目，可在直角坐标系内，将横坐标轴上的变量设置为特征值的个数p；将纵坐标轴上的变量设置为特征值E。将数据点(p,E)描绘在直角坐标系内，用折线将所有散点依次相连，构成的折线图称为因子碎石图。此图为一条近似“L型”的曲线图，图中转折

点对应的横坐标轴上的特征值个数即为应当保留的公因子个数。

2 计算方法

2.1 因子分析的数学模型

设有 n 个样品，每个样品观测了 p 个变量（X_1，…，X_p），为了对变量进行比较，并消除数据量纲的差异以及数量级的影响，将观测变量进行标准化处理，$x_i = \frac{X_i - \bar{X}_i}{S_i}$，标准化后的变量均值为0，方差为1。

因子分析是将实测变量表示成公因子的线性函数与特殊因子之和。公因子之间是独立的，特殊因子之间是独立的，公因子与特殊因子之间也是独立的，即各因子间正交，故称该因子模型为正交初始因子模型，建立 R 型正交因子分析模型（$m \leqslant p$）[1-2]。用矩阵表示，见式（1）。

$$\begin{pmatrix} x_1 \\ x_2 \\ \cdots \\ x_p \end{pmatrix} = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1m} \\ a_{21} & a_{22} & \cdots & a_{2m} \\ \cdots & \cdots & \cdots & \cdots \\ a_{p1} & a_{p2} & \cdots & a_{pm} \end{pmatrix} \begin{pmatrix} f_1 \\ f_2 \\ \cdots \\ f_m \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \cdots \\ \varepsilon_p \end{pmatrix} \quad (1)$$

简记为式（2）的形式。

$$\underset{(p\times 1)}{X} = \underset{(p\times m)}{A}\ \underset{(m\times 1)}{F} + \underset{(p\times 1)}{\varepsilon} \quad (2)$$

式（1）中，f_1，…，f_m 为公因子，是相互独立且不可观测的隐变量，它们之前的系数 a_{ij} 为因子载荷，是第 i 个指标所包含的信息分摊在第 j 个公因子上的数量；ε_i 为特殊因子，是不能被前 m 个公因子包含的部分信息，其并非普通意义上的残差，它只影响当前变量，与其他变量无关，表示该变量中特有的、不能被公因子解释的那部分信息[3-4]。

式（1）被称为初始因子模型，即采用公因子和特殊因子来线性表达每个显变量。但在实际使用过程中，研究者更希望利用原变量的取值预测公因子的取值，以便依据主要少数几个公因子的取值，对原始资料中的“样品”进行分类。为实现此目标，需构建因子得分模型。可直接对式（1）中的载荷矩阵进行变换，得到与因子得分模型相对应的系数矩阵。但这样做的结果并非十分理想。基于式（1）的公因子代表 m 维空间中的 m 个坐标轴，将它们按一定的规则旋转，可以改变全部数据所代表的观测点在 m 维空间中的相对位置。若能找到合适的旋转方向和角度，可使每个原变量只在少数公因子轴上有较大的载荷，换言之，式（1）中载荷矩阵中每列上的元素向0与1两极分化，但保持同一行中各元素平方和（称为各原始指标的公因子方差）不变，实现这一目的的变换方法称为因子轴的旋转，从而得到旋转后的因子模型。

从统计计算的角度看，探索性因子分析主要包括三个步骤：①估计初始因子模型的载荷矩阵中的元素；②估计旋转后因子模型的载荷矩阵中的元素；③估计因子得分模型中的系数，有回归分析法和来自最大似然估计的加权最小平方法[5-6]。

建立某实际问题的初始因子模型，关键是根据样本数据矩阵估计因子载荷矩阵 A。对 A 的估计方法包括主成分法、主因子法、重心法和最大似然法，其中主成分法和最大似然法应用较多[7-8]。若需初步估计因子模型中保留的公因子个数，主成分法较好，此法会产生公因子个数的最大值；最大似然法是一种比较客观的方法，可计算似然比检验统计量的值，以检验所选的公因子个数是否合适，但此方法倾向于产生有统计学意义但实际价值不大的公因子。

2.2 主成分法

主成分法就是在进行因子分析前，首先进行主成分分析，以 x_i（i=1，2，…，p）表示标准化的原变量，C_i（i=1，2，…，p）表示主成分，计算原变量的相关系数矩阵 R 的特征根和特征向量，得到以主成分变量线性表达原变量的多个表达式，称为主成分的联立方程组，见式（3）。

$$\begin{cases} C_1 = \gamma_{11}x_1 + \gamma_{12}x_2 + \cdots + \gamma_{1p}x_P \\ C_2 = \gamma_{21}x_1 + \gamma_{22}x_2 + \cdots + \gamma_{2p}x_P \\ \cdots \\ C_p = \gamma_{p1}x_1 + \gamma_{p2}x_2 + \cdots + \gamma_{pp}x_p \end{cases} \quad (3)$$

根据主成分分析的特征根所对应的特征向量彼此正交的性质，通过转换，将原始变量表达为主成分，见以下方程组，即式（4）。

$$\begin{cases} x_1 = \gamma_{11}C_1 + \gamma_{21}C_2 + \cdots + \gamma_{p1}C_p \\ x_2 = \gamma_{12}C_1 + \gamma_{22}C_2 + \cdots + \gamma_{p2}C_p \\ \cdots \\ x_p = \gamma_{1p}C_1 + \gamma_{2p}C_2 + \cdots + \gamma_{pp}C_p \end{cases} \quad (4)$$

式（4）系数矩阵的行和列恰好是式（3）中系数矩阵的列和行。如果对上面每一等式保留 m 个主成分变量，而把后面的部分用 ε_i 代替，通过变换，可与因子分析的联立方程组[式（1）]形式一致，为了把联立方程组中的主成分变换为公因子，还需把主成分进行标准化变换。由主成分分析过程可知，各个主成分的均数为0，标准差为特征根的平方根 $\sqrt{\lambda_i}$，如果令 $f_i = C_i/\sqrt{\lambda_i}$，$a_{ij} = \sqrt{\lambda_i}\,\gamma_{ji}$，上述联立方程组就变为前面提及的因子分析的联立方程组，即式（1）。

因篇幅所限，主因子法、重心法和最大似然法的内容从略，详见文献[9]。

3 实例与SAS实现

3.1 问题与数据结构

3.1.1 2个实际问题及数据

【例1】对31个省级政府门户网站的绩效进行评估，收集相关资料：X_1为信息公开指数，X_2为在线办事指数，X_3为公众参与指数，X_4为每百万用户网站访问量，X_5为每个用户的页面浏览量，X_6为网站下载速度，X_7为被其他网站链接的数量，X_8为用户在网站停留时间。数据见表1[10]。

【例2】选取全国31个省市自治区为样本，5个基本建设资金来源为指标，以2000年全国各地区基本建设投资数据为基础，对全国各地区基本建设投资来源进行因子分析。数据见表2[10]。

表1 31个省级政府门户网站的评估数据

Table 1 Evaluation data of 31 provincial government portal websites

省 份	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8
北京	0.82	0.95	0.90	37.10	2.46	1.3	2 511	2.6
上海	0.84	0.83	0.76	89.00	3.37	2.8	1 819	4.0
…	…	…	…	…	…	…	…	…
海南	0.79	0.57	0.85	14.70	3.30	1.9	791	3.6
新疆	0.45	0.25	0.46	3.20	3.90	1.6	545	3.4

注：详细数据见后文例1的SAS过程步程序

表2 全国各地区基本建设投资来源

Table 2 Sources of basic construction investment in various regions of the country

地 区	国家预算内资金	国内贷款	利用外资	自筹资金	其他资金
北京	119.81	74.58	23.95	214.84	27.67
天津	9.28	76.41	48.68	79.52	20.31
…	…	…	…	…	…
宁夏	13.25	19.48	2.04	23.02	7.50
新疆	43.10	85.72	6.65	133.52	49.46

注：详细数据见后文例2的SAS过程步程序

3.1.2 对数据结构的分析

例1和例2的数据结构相同，它们都是单组设计多元定量资料。资料中的“省份”或“地区”相当于以“人”为受试对象的通常资料中的“个体”，在单组设计中，要求所有个体应具有同质性。在以“省份”为个体的资料中，“同质性”主要指政治和社会地位，而非其他方面（如经济实力和地理位置等）。

3.1.3 创建SAS数据集

3.1.3.1 分析例1的资料

设所需SAS数据步程序如下：

```
data a1;
length regine $ 6;
input regine $ X1-X8;
cards;
北京 0.82 0.95 0.90 37.10 2.46 1.3 2511 2.6
上海 0.84 0.83 0.76 89.00 3.37 2.8 1819 4.0
天津 0.64 0.57 0.44 5.10 1.80 2.0 890 1.9
重庆 0.47 0.32 0.54 25.80 3.90 1.2 1186 4.1
河北 0.55 0.47 0.50 5.60 1.80 2.4 882 1.8
河南 0.54 0.36 0.56 10.70 3.30 3.3 1000 3.4
山东 0.41 0.23 0.40 1.10 4.20 2.3 560 2.6
山西 0.57 0.36 0.47 0.45 1.20 3.6 481 1.2
内蒙古 0.52 0.32 0.48 3.30 1.90 1.1 559 2.3
辽宁 0.66 0.45 0.40 7.00 2.40 2.5 819 2.3
吉林 0.58 0.36 0.51 9.00 3.10 1.7 708 2.7
黑龙江 0.63 0.28 0.65 4.80 2.60 0.5 824 1.9
江苏 0.63 0.39 0.60 5.50 2.30 1.6 778 2.4
浙江 0.78 0.54 0.85 10.40 3.50 2.4 187 3.6
安徽 0.65 0.49 0.65 8.90 2.00 1.7 1060 2.0
福建 0.66 0.52 0.57 7.10 3.00 0.6 860 2.4
江西 0.61 0.30 0.46 6.40 3.20 2.5 824 3.3
湖北 0.59 0.38 0.49 6.70 2.00 1.4 782 2.2
湖南 0.65 0.53 0.57 10.40 3.10 1.0 1050 3.4
广东 0.69 0.60 0.60 16.00 2.30 1.4  201 2.8
广西 0.46 0.25 0.33 3.50 2.50 2.1 472 2.2
四川 0.58 0.56 0.66 0.14 1.40 1.4 268 1.5
贵州 0.44 0.13 0.35 2.60 2.90 1.1 803 3.0
云南 0.63 0.34 0.54 23.60 2.37 0.6 1002 2.4
陕西 0.72 0.61 0.80 12.40 3.60 1.2 832 4.6
西藏 0.42 0.22 0.21 0.80 4.90 1.7 97 1.8
甘肃 0.26 0.14 0.41 9.30 3.60 1.1 140 4.0
青海 0.49 0.41 0.36 1.90 1.70 2.3 298 2.4
宁夏 0.37 0.19 0.42 0.80 2.20 1.7 195 2.0
海南 0.79 0.57 0.85 14.70 3.30 1.9 791 3.6
新疆 0.45 0.25 0.46 3.20 3.90 1.6 545 3.4
;
run;
```

3.1.3.2 分析例2的资料

设所需SAS数据步程序如下：

```
data a2;
length regine $ 6;
input regine $ X1-X5;
cards;
北京 119.81 74.58 23.95 214.84 27.67
```

```
天津 9.28 76.41 48.68 79.52 20.31
河北 36.49 128.18 32.05 254.43 101.26
山西 15.83 79.79 54.28 69.59 44.92
内蒙古 29.39 37.40 12.44 52.86 35.61
辽宁 63.55 107.09 32.36 165.51 39.01
吉林 32.12 42.24 5.65 92.82 37.31
黑龙江 38.53 66.30 13.99 129.75 77.86
上海 33.44 131.11 121.34 352.64 67.42
江苏 52.28 142.17 49.04 376.01 51.26
浙江 59.91 174.87 60.61 313.11 53.85
安徽 34.74 50.44 4.81 107.67 37.41
福建 33.34 66.49 42.29 126.12 45.78
江西 33.59 42.56 5.82 59.48 25.63
山东 48.74 193.71 46.69 299.76 64.58
河南 77.34 120.98 31.58 190.61 90.27
湖北 90.09 119.13 16.91 194.80 97.76
湖南 55.73 61.45 10.81 133.15 55.55
广东 39.63 260.53 167.29 570.33 78.54
广西 30.50 46.29 25.91 104.32 49.44
海南 12.54 33.60 20.31 58.59 16.98
重庆 38.08 52.60 4.64 58.78 28.96
四川 34.11 152.75 34.51 218.96 81.20
贵州 21.92 38.89 8.66 46.13 15.01
云南 28.32 89.04 5.57 131.33 39.79
西藏 24.28 1.84 0.02 16.45 5.32
陕西 47.99 70.94 6.93 86.81 53.57
甘肃 25.07 53.85 5.13 56.26 35.60
青海 10.99 27.25 0.68 19.96 12.14
宁夏 13.25 19.48 2.04 23.02 7.50
新疆 43.10 85.72 6.65 133.52 49.46
;
run;
```

3.2 用SAS实现统计分析

3.2.1 分析例1的资料

设所需SAS过程步程序如下[11]：

```
proc factor method=principal
priors=one c scree n=5 rotate=varimax score;
var X1-X8;
run;
proc factor method=ml priors=s heywood n=2 reorder
score out=scoredata rotate=varimax;
var X1-X8;
run;
proc print data=scoredata;
var factor1 factor2;
run;
```

【SAS程序说明】第一个过程步中，“method=principal priors=one”表示选用主成分分析方法进行因子分析，“c”表示输出变量之间的相关系数矩阵，“scree”为输出因子碎石图，用于选择公因子个数，“n=5”指定公因子数为5，“rotate=varimax”执行方差最大正交旋转；第二个过程步中，“method=ml”表示选用最大似然法进行因子分析，选项“heywood”的作用是当公因子方差大于1时，令其为1，并允许迭代继续进行；第三个过程步是输出每个观测（或个体）在第一和第二公因子上的得分。

【SAS输出结果及解释】基于第一个过程步程序输出的第一部分结果见表3。由表3可知，欲使累积贡献率达到90%，需要前5个特征值。由因子碎石图（因所占篇幅过大，此处从略）也可知，在5个公因子之前，曲线以很陡峭的趋势下降，而之后呈现几乎平缓的趋势下降。因此，保留5个公因子是合适的。

基于第一个过程步程序输出的第二部分结果见表4。

表3 基于相关矩阵计算得到的特征值和累积贡献率

Table 3 Eigenvalues and cumulative contribution rates calculated based on correlation matrix

编 号	特征值	差 分	比 例	累积(%)
1	3.781	2.045	0.473	0.473
2	1.736	0.708	0.217	0.690
3	1.028	0.381	0.129	0.818
4	0.647	0.332	0.081	0.899
5	0.315	0.086	0.039	0.938
6	0.229	0.087	0.029	0.967
7	0.142	0.020	0.018	0.985
8	0.122	-	0.015	1.000

表4 初始因子模型中载荷的计算结果

Table 4 Calculation results of the load in the initial factor model

变 量	Factor 1	Factor 2	Factor 3	Factor 4	Factor 5
X_1	0.882	-0.246	-0.015	0.246	0.173
X_2	0.906	-0.268	0.021	0.069	0.095
X_3	0.854	-0.019	-0.168	0.407	-0.086
X_4	0.788	0.233	0.205	-0.405	-0.132
X_5	0.003	0.913	0.095	0.048	0.388
X_6	0.007	-0.202	0.969	0.109	0.003
X_7	0.819	-0.047	-0.071	-0.459	0.075
X_8	0.402	0.820	0.068	0.167	-0.309

基于表4各行的结果，可以写出初始因子模型，现以第一行为例，即有如下表达，见式(5)。

$$x_1 = 0.882f_1 - 0.246f_2 - 0.015f_3 + 0.246f_4 + 0.173f_5 \quad (5)$$

仿照上式，可以写出基于5个公因子线性表达X_2到X_8的表达式。因篇幅所限，此处从略。

基于第一个过程步程序输出的第三部分结果见表5。由表5可知，第一公因子主要控制着X_1、X_2和X_3(因为这3个变量前的系数接近于1)；而第二、三、四、五个公因子分别控制着(X_4和X_7)、X_5、X_6、X_8。基于表5的结果，可写出旋转后因子模型，现以第一行为例，即有如下表达，见式(6)。

$$x_1 = 0.911f_1 + 0.306f_2 - 0.034f_3 + 0.059f_4 - 0.028f_5 \quad (6)$$

仿照上式，可以写出基于5个公因子线性表达X_2到X_8的表达式。因篇幅所限，此处从略。

基于第一个过程步程序输出的第四部分结果见表6。

利用表6中各列数据可以写出5个公因子表达式，见式(7)。

结合式(7)和例1中各变量的专业含义，可以给5个公因子分别取一个合适的名称，例如，公因子f_3、f_4、f_5可分别被称为“每个用户的页面浏览量因子”“网站下载速度因子”和“用户在网站停留时间因子”。因篇幅所限，基于第二和第三个过程步程序输出的结果，此处从略。

表5 旋转后因子模型的载荷

Table 5 Load of factor model after rotation

变 量	Factor 1	Factor 2	Factor 3	Factor 4	Factor 5
X_1	0.911	0.306	−0.034	0.059	−0.028
X_2	0.820	0.465	−0.111	0.076	−0.019
X_3	0.887	0.167	−0.037	−0.109	0.319
X_4	0.289	0.831	0.099	0.119	0.317
X_5	−0.120	0.026	0.950	−0.041	0.277
X_6	0.006	0.007	−0.039	0.994	−0.027
X_7	0.422	0.838	−0.009	−0.113	−0.027
X_8	0.139	0.178	0.444	−0.041	0.844

表6 标准化评分系数的计算结果

Table 6 Calculation results of standardized scoring coefficients

变 量	Factor 1	Factor 2	Factor 3	Factor 4	Factor 5
X_1	0.530	−0.188	0.283	0.059	−0.347
X_2	0.338	0.041	0.095	0.060	−0.234
X_3	0.480	−0.376	−0.151	−0.083	0.384
X_4	−0.312	0.662	−0.192	0.097	0.292
X_5	0.152	−0.045	1.208	0.026	−0.567
X_6	0.018	−0.040	0.031	0.962	0.031
X_7	−0.152	0.676	0.079	−0.147	−0.323
X_8	−0.093	−0.076	−0.252	0.016	1.090

$$\begin{cases} f_1 = 0.503x_1 + 0.338x_2 + 0.480x_3 - 0.312x_4 + 0.152x_5 + 0.018x_6 - 0.152x_7 - 0.093x_8 \\ f_2 = -0.188x_1 + 0.041x_2 - 0.376x_3 + 0.662x_4 - 0.045x_5 - 0.040x_6 + 0.676x_7 - 0.076x_8 \\ f_3 = 0.283x_1 + 0.095x_2 - 0.151x_3 - 0.192x_4 + 1.208x_5 + 0.031x_6 + 0.079x_7 - 0.252x_8 \\ f_4 = 0.059x_1 + 0.060x_2 - 0.083x_3 + 0.097x_4 + 0.026x_5 + 0.962x_6 - 0.147x_7 + 0.016x_8 \\ f_5 = -0.347x_1 - 0.234x_2 + 0.384x_3 + 0.292x_4 - 0.567x_5 + 0.031x_6 - 0.323x_7 + 1.090x_8 \end{cases} \quad (7)$$

3.2.2 分析例2的资料

设所需要的SAS过程步程序如下[11]：

```
proc factor method=principal
priors=one c scree n=3 rotate=varimax score;
var X1-X5;
run;
proc factor method=ml priors=s heywood n=3 reorder
score out=scoredata rotate=varimax;
var X1-X5;
run;
proc print data=scoredata;
var factor1 factor2;
run;
```

【SAS输出结果及解释】为节省篇幅，下面仅输出基于第一个过程步程序输出的最后一部分结果。见表7。利用表7中各列数据，可以写出3个公因子表达式，见式(8)。

表7 标准化评分系数的计算结果

Table 7 Calculation results of standardized scoring coefficients

变 量	Factor 1	Factor 2	Factor 3
X_1	−0.050	−0.283	1.036
X_2	0.274	0.137	−0.008
X_3	0.565	−0.339	−0.143
X_4	0.388	−0.139	0.140
X_5	−0.337	1.173	−0.249

$$\begin{cases} f_1 = -0.050x_1 + 0.274x_2 + 0.565x_3 + 0.388x_4 - 0.337x_5 \\ f_2 = -0.283x_1 + 0.137x_2 - 0.339x_3 - 0.139x_4 + 1.1735x_5 \\ f_3 = 1.036x_1 - 0.008x_2 - 0.143x_3 + 0.140x_4 - 0.249x_5 \end{cases} \quad (8)$$

结合式(8)和例2中各变量的专业含义，可以给3个公因子分别取一个合适的名称，例如，公因子f_1、f_2、f_3可分别被称为“利用外资因子”“其他资金因子”

和“国家预算内资金因子”。因篇幅所限，基于第二和第三个过程步程序输出的结果，此处从略。

4 讨论与小结

4.1 讨论

当研究者采用探索性因子分析处理一个非单组设计多元混合型资料时，如果计算结果无法做出符合实际的解释，原因可能如下：其一，非单组设计，这就意味着全部样本可能受到一个$k(k\geq 2)$水平因素的影响或多因素的影响，原本是希望研究相同条件（单组设计）下定量变量之间的相互关系，而实际上处理的却是不同条件下定量变量之间的相互关系，但又未消除“不同条件”造成的干扰影响，这就不可避免地得出错误的结论；其二，多元混合型资料，这就意味着全部变量中有些是定量的、有些是定性的，特别是二值或多值名义变量，其所承载的信息具有“凝集性”，它们与定量变量在数量关系上是不对等的，而在进行探索性因子分析的过程中，并未对此采取相应的校正，故计算结果是不正确的。

探索性因子分析应用是否正确，取决于“试验设计”是否科学完善，包括以下三个方面：其一，待处理的样本应取自与所研究问题相符合的、具有同质性的总体，这是“单组设计”的真正含义，例如，当研究者希望考查正常成年人多项血脂指标之间的关系时，若样本中既包含冠心病和糖尿病患者，还包含未成年人，显然，他们与“正常成年人”是不同质的。其二，所观测的变量应均为定量的、与研究目的完全吻合的，并涵盖了主要相关变量。仍以“血脂指标”为例，若全部血脂指标有50项，而研究者仅观测了其中10项，且这10项并不是全部50项中最主要的，故由此得到的结论可能是不正确的。其三，应具有足够大的样本含量。在生物医学研究领域里，即使样本的同质性和代表性很好，但由于个体差异往往很大，在样本含量并非足够大的条件下得出的结论，其可信度也不会很高。样本含量一般应为全部定量变量数量的20倍以上，最低也应在10倍以上。

4.2 小结

本文介绍了与探索性因子分析有关的基本概念、计算方法、两个实例以及使用SAS实现计算的方法。基本概念包括公因子与特殊因子、因子载荷、公因子方差和因子碎石图；计算方法涉及因子分析的数学模型、主成分法、主因子法、重心法和最大似然法；两个实例的资料分别是“31个省级政府门户网站的评估数据”和“全国各地区基本建设投资来源”；借助SAS软件，对两个实例中的数据进行了探索性因子分析，并对SAS输出结果做出了解释。

参考文献

[1] Johnson RA, Wichern DW. 实用多元统计分析[M]. 6版. 北京：清华大学出版社，2008：481-538.
Johnson RA, Wichern DW. Applied multivariate statistical analysis [M]. 6th edition. Beijing: Tsinghua University Press, 2008: 481-538.

[2] Armitage P, Colton T. Encyclopedia of biostatistics [M]. 2nd edition. New York: John Wiley & Sons, Inc, 2005: 2053-2072.

[3] 张岩波. 潜变量分析[M]. 北京：高等教育出版社，2009：35-59.
Zhang YB. Latent variables analysis [M]. Beijing: Higher Education Press, 2009: 35-59.

[4] 李卫东. 应用多元统计分析[M]. 北京：北京大学出版社，2008：207-238.
Li WD. Applied multivariate statistical analysis [M]. Beijing: Peking University Press, 2008: 207-238.

[5] 余锦华，杨维权. 多元统计分析与应用[M]. 广州：中山大学出版社，2005：210-231.
Yu JH, Yang WQ. Multivariate statistical analysis and application [M]. Guangzhou: Sun Yat-sen University Press, 2005: 210-231.

[6] 何晓群. 多元统计分析[M]. 2版. 北京：中国人民大学出版社，2008：192-226.
He XQ. Multivariate statistical analysis[M]. 2nd edition. Beijing: China Renmin University Press, 2008: 192-226.

[7] 王静龙. 多元统计分析[M]. 北京：科学出版社，2008：360-391.
Wang JL. Multivariate statistical analysis[M]. Beijing: Science Press, 2008: 360-391.

[8] 高惠璇. 应用多元统计分析[M]. 北京：北京大学出版社，2005：293-323.
Gao HX. Applied multivariate statistical analysis[M]. Beijing: Peking University Press, 2005: 293-323.

[9] 张润楚. 多元统计分析[M]. 北京：科学出版社，2006：190-217.
Zhang RC. Multivariate statistical analysis[M]. Beijing: Science Press, 2006: 190-217.

[10] 胡良平. 面向问题的统计学：(3)试验设计与多元统计分析[M]. 北京：人民卫生出版社，2012：57-83.
Hu LP. Problem-oriented statistics: (3) experimental design and multivariate statistical analysis [M]. Beijing: People's Medical Publishing House, 2012: 57-83.

[11] SAS Institute Inc. SAS/STAT®15.1 user's guide[M]. Cary, NC: SAS Institute Inc, 2018: 2715-2824.

（收稿日期：2023-09-25）
（本文编辑：陈 霞）