

合理进行多元分析——主成分分析

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】 本文目的是介绍与主成分分析有关的基本概念、计算方法、两个实例以及 SAS 实现。基本概念包括相关矩阵、特征值与特征向量、主成分变量、主成分表达式和主成分的性质; 计算方法涉及特征值与特征向量的求法、主成分分析的计算原理以及系数估计和个数的确定; 两个实例中的资料分别为“20 例肝病患者的 4 项肝功能指标的测定结果”和“23 种肿瘤类期刊的文献计量学指标的调查结果”; 借助 SAS 对两个实例中的定量资料进行了主成分分析, 并基于主成分的计算结果分别实现了样品聚类和样品排序, 并对输出结果作出了解释。

【关键词】 特征值; 特征向量; 主成分分析; 样品聚类; 样品排序

中图分类号: R195.1

文献标识码: A

doi: 10.11886/scjsws20230605001

Reasonably carry out multivariate analysis: principal component analysis

Hu Chunyan¹, 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: lphu927@163.com)

【Abstract】 The purpose of this article was to introduce the basic concepts, calculation methods, two examples and SAS implementation related to the principal component analysis. Basic concepts included correlation matrix, eigenvalues and eigenvectors, principal component variables, principal component expressions and principal component properties. The calculation method involved the calculation of eigenvalues and eigenvectors, the calculation principle and the coefficient estimation and the number determination of the principal component. The data in the two examples were measurement results of 4 liver function indicators in 20 patients with liver disease and survey results of literature metrology indicators in 23 tumor journals. With the help of SAS software, the principal component analysis was carried out on the quantitative data in the two cases, and based on the calculation results of the principal components, the sample clustering and sample sorting were respectively realized, and a reasonable explanation was given for the output results.

【Keywords】 Eigenvalue; Eigenvector; Principal component analysis; Sample clustering; Sample sorting

在生物医学和临床研究中, 研究者经常收集到单组设计多元定量资料。如何选择合适的统计方法处理这种定量资料, 是研究者经常面临的一个棘手的统计问题, 因为可用于处理这种多元定量资料的多元统计分析方法约有十几种。本文将介绍一种最简单、最基础的多元统计分析方法, 即主成分分析。

1 基本概念

1.1 相关矩阵

设具有同质性的 n 个个体, 测量其 m 个定量指标 (记为 x_1, x_2, \dots, x_m) 的取值, 采用 Pearson 相关分析公式 [见式(1)] 计算出任意两个定量变量之间的相关系数 (记为 $r_{ij}, i, j=1, 2, \dots, m$)^[1], 将它们按一定顺序排

列成一个 $m \times m$ 方阵, 见式(2), 此方阵被称为相关矩阵。

$$r_{ij} = \frac{\sum_{k=1}^n (x_{ik} - \bar{x}_{ik})(x_{jk} - \bar{x}_{jk})}{\sum_{k=1}^n (x_{ik} - \bar{x}_{ik})^2 \sum_{k=1}^n (x_{jk} - \bar{x}_{jk})^2} \quad (1)$$

$$R = \begin{bmatrix} r_{11} & r_{12} & \cdots & r_{1m} \\ r_{21} & r_{22} & \cdots & r_{2m} \\ \vdots & \vdots & \vdots & \vdots \\ r_{m1} & r_{m2} & \cdots & r_{mm} \end{bmatrix} \quad (2)$$

式(1)中, k 代表观测或个体的编号。式(2)中, $r_{ij} = r_{ji}, i, j = 1, 2, \dots, m$, 即 R 是一个对称的矩阵。

1.2 特征值与特征向量

对于一个 m 阶矩阵 [见式(3)], 如果存在一个数 λ_0 和非零向量 X_0 使式(4)成立, 则称 λ_0 为矩阵 A 的

特征值,称 X_0 为矩阵 A 对应于特征值 λ_0 的特征向量^[2]。

$$A = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1m} \\ a_{21} & a_{22} & \cdots & a_{2m} \\ \vdots & \vdots & \vdots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mm} \end{bmatrix} \quad (3)$$

$$AX_0 = \lambda_0 X_0 \quad (4)$$

1.3 主成分变量

主成分变量(简称主成分)是一个不能被直接观测的隐变量。在一个单组设计 m 元定量资料中,有 m 个主成分,其中,每一个主成分都是由 m 个原变量线性组合而成的,但它们彼此互相独立;通常,它们所包含的信息量是不相等的,且满足如下关系,见式(5)。

$$P_m \leq P_{m-1} \leq \cdots \leq P_2 \leq P_1 \quad (5)$$

式(5)中, $P_i (i = 1, 2, \dots, m)$ 代表第 i 个主成分。主成分由原始变量的线性组合而成,故主成分也被称为线性主成分。当对数据进行非线性变换后再进行主成分分析时,就称为非线性主成分分析^[3]。

1.4 主成分表达式

主成分的常见表达式有以下 4 种,见式(6)、式(7)、式(8)、式(9)。

$$\begin{cases} Z_1 = a_{11}(X_1 - \bar{X}_1) + a_{12}(X_2 - \bar{X}_2) + \cdots + a_{1m}(X_m - \bar{X}_m) \\ Z_2 = a_{21}(X_1 - \bar{X}_1) + a_{22}(X_2 - \bar{X}_2) + \cdots + a_{2m}(X_m - \bar{X}_m) \\ \dots\dots \\ Z_m = a_{m1}(X_1 - \bar{X}_1) + a_{m2}(X_2 - \bar{X}_2) + \cdots + a_{mm}(X_m - \bar{X}_m) \end{cases} \quad (6)$$

$$\begin{cases} Z_1 = b_{11}x_1 + b_{12}x_2 + \cdots + b_{1m}x_m \\ Z_2 = b_{21}x_1 + b_{22}x_2 + \cdots + b_{2m}x_m \\ \dots\dots \\ Z_m = b_{m1}x_1 + b_{m2}x_2 + \cdots + b_{mm}x_m \end{cases} \quad (7)$$

$$\begin{cases} Z_1 = c_{11}x_1 + c_{12}x_2 + \cdots + c_{1m}x_m \\ Z_2 = c_{21}x_1 + c_{22}x_2 + \cdots + c_{2m}x_m \\ \dots\dots \\ Z_m = c_{m1}x_1 + c_{m2}x_2 + \cdots + c_{mm}x_m \end{cases} \quad (8)$$

$$\begin{cases} x_1 = c_{11}Z_1 + c_{21}Z_2 + \cdots + c_{m1}Z_m \\ x_2 = c_{12}Z_1 + c_{22}Z_2 + \cdots + c_{m2}Z_m \\ \dots\dots \\ x_m = c_{1m}Z_1 + c_{2m}Z_2 + \cdots + c_{mm}Z_m \end{cases} \quad (9)$$

式(6)、式(7)、式(8)、式(9)中, X_i 为第 i 个原始变量, x_i 为 X_i 的标准化变量,见式(10)。

$$x_i = \frac{X_i - \bar{X}_i}{S_i} \quad (10)$$

式(10)中, S_i 为第 i 个变量的样本标准差, $i = 1, 2, \dots, m$ 。

式(6)、式(7)、式(8)、式(9)中,各 a_{ij} 、 b_{ij} 、 c_{ij} 都是线性组合的系数,称为因子负荷量。其中, a_{ij} 、 b_{ij} 分别是基于原始数据与标准化数据计算所得到的第 i 个特征向量中的第 j 个元素。 c_{ij} 的计算见式(11)。

$$c_{ij} = \sqrt{\lambda_i} b_{ij} \quad (11)$$

式(11)中, λ_i 为第 i 个特征值; b_{ij} 为式(7)中的系数; $i, j = 1, 2, \dots, m$ 。

1.5 主成分的性质

各主成分之间互不相关,即任何两个主成分之间的相关系数为 0,若原变量服从正态分布,则各主成分之间互相独立;全部主成分所反映的信息等于全部原变量的总信息;各主成分的作用大小不等:第一主成分的作用大于等于第二主成分,第二主成分的作用大于等于第三主成分,以此类推,最后一个主成分的作用最小。因篇幅所限,主成分的其他性质从略。

2 计算方法

2.1 特征值与特征向量的求法

第一步,构造一个特征矩阵,见式(12):

$$\lambda E - A = \begin{bmatrix} \lambda - a_{11} & -a_{12} & \cdots & -a_{1m} \\ -a_{21} & \lambda - a_{22} & \cdots & -a_{2m} \\ \vdots & \vdots & \vdots & \vdots \\ -a_{m1} & -a_{m2} & \cdots & \lambda - a_{mm} \end{bmatrix} \quad (12)$$

式(12)中, λ 是一个未知量; E 是一个对角线上元素全为 1 的 $m \times m$ 对角矩阵;矩阵 A 的定义见前文式(3)。

第二步,构造一个 λ 的 n 齐次多项式,见式(13)。

$$f(\lambda) = |\lambda E - A| = \begin{vmatrix} \lambda - a_{11} & -a_{12} & \cdots & -a_{1m} \\ -a_{21} & \lambda - a_{22} & \cdots & -a_{2m} \\ \vdots & \vdots & \vdots & \vdots \\ -a_{m1} & -a_{m2} & \cdots & \lambda - a_{mm} \end{vmatrix} \quad (13)$$

式(13)中的“ $|*|$ ”代表由“ $*$ ”确定的行列式,它是一个具体的数值。

第三步,构造一个矩阵 A 的特征方程组,见式(14)。

$$f(\lambda) = |\lambda E - A| = 0 \quad (14)$$

第四步,求出特征方程组的全部根,即特征值 $\lambda_k (k = 1, 2, \dots, m)$ 。

第五步,把特征值逐个代入齐次线性特征方程组,见式(15)。

$$(\lambda_k E - A)X = 0 \quad (15)$$

式(15)中, X 是由全部定量变量构成的向量, 见式(16)。

$$X = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_m \end{bmatrix} \quad (16)$$

求出方程组的基础解系, 则这个基础解系的非零线性组合就是矩阵 A 对应于特征值 $\lambda_k (k=1, 2, \dots, m)$ 的全部特征向量。因篇幅所限, 具体求解过程见文献[2,4]。

2.2 主成分分析的计算原理

设有 n 例儿童的身高(X_1)和体重(X_2)两个观测指标, 显然它们之间有较强的相关性。以 X_1 为横轴, 以 X_2 为纵轴, 绘制散点图如图 1 所示。可见这 n 个点的分布在一条直线的近旁, 呈现出直线化趋势。它们沿 X_1 轴和 X_2 轴方向都具有较大的变异度, 个体在某个方向上的变异度可用该方向上相应观测变量的方差来表示。以此直线作为新的横轴 Z_1 , 再作一条垂直于 Z_1 的直线作为纵轴 Z_2 , 在平面上, 这 n 点的变异主要集中在 Z_1 方向上, 在 Z_2 方向上变异很小。所以, 研究这 n 个对象的变异, 可以只考虑 Z_1 值的大小, 忽略 Z_2 值的大小。也就是说, 若取 Z_1 作为第一主成分, 则 Z_1 就反映了原始指标 X_1 和 X_2 所包含的大部分信息。若将 X_1 和 X_2 标准化后的指标记为 Y_1, Y_2 , 则 Z_1, Z_2 与 Y_1, Y_2 有以下关系, 见式(17)和式(18)。

$$Z_1 = l_{11}Y_1 + l_{12}Y_2 \quad (17)$$

$$Z_2 = l_{21}Y_1 + l_{22}Y_2 \quad (18)$$

Z_1, Z_2 是 Y_1, Y_2 的线性函数, 显然也是 X_1, X_2 的线性函数, 且 Z_1, Z_2 不相关。称 Z_1 为第一主成分, Z_2 为第二主成分, 并称这种分析方法为主成分分析法。

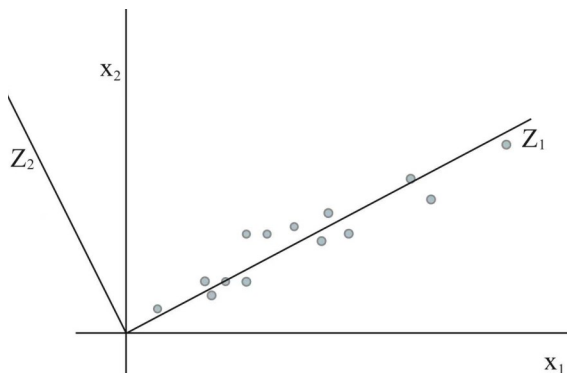


图 1 主成分分析示意图

Figure 1 Schematic diagram of principal component analysis

2.3 主成分表达式中系数的估计

为了计算简便, 主成分的计算一般从求相关矩阵出发^[2,5]。以表 1 为例, 由 m 个变量 X_1, X_2, \dots, X_m 的 n 个样本观测值求出主成分。计算步骤如下。

第一步, 对各原始指标数据进行标准化, 见式(19)。然后用标准化的数据来计算主成分。记 X 为标准化后的数据矩阵, 见式(20)。

$$x_{ij} = \frac{X_{ij} - \bar{X}}{S_j}, \quad j = 1, 2, \dots, m \quad (19)$$

$$X = \begin{bmatrix} x_{11} & x_{12} & \dots & x_{1m} \\ x_{21} & x_{22} & \dots & x_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \dots & x_{nm} \end{bmatrix} \quad (20)$$

第二步, 求出 X 的相关矩阵 R [标准化后, X 的相关矩阵即为协方差矩阵 $\text{Cov}(X)$], 见式(21)。

$$R = \text{Cov}(X) = \begin{bmatrix} r_{11} & r_{12} & \dots & r_{1m} \\ r_{21} & r_{22} & \dots & r_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ r_{m1} & r_{m2} & \dots & r_{mm} \end{bmatrix} = \begin{bmatrix} 1 & r_{12} & \dots & r_{1m} \\ r_{21} & 1 & \dots & r_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ r_{m1} & r_{m2} & \dots & 1 \end{bmatrix} \quad (21)$$

第三步, 求出相关矩阵的特征值和特征值所对应的特征向量。求主成分问题, 实际上就是要求出 X 的协方差矩阵 $\text{Cov}(X)$ (这里即为 X 的相关矩阵 R) 的特征值和特征向量。由于 R 为半正定矩阵, 故可由 R 的特征方程 $|R - \lambda I| = 0$ 求得 m 个非负特征值, 这些特征值按从大到小排序为: $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_m$ 。特征方程见式(22)。

$$\begin{cases} (R - \lambda_i I)a_i = 0 \\ a_i' a_i = 1 \end{cases} \quad i = 1, 2, \dots, m \quad (22)$$

求出式(22)的解, 得每一个特征值对应的单位特征向量 $a_i = (a_{i1}, a_{i2}, \dots, a_{im})'$, 可写出主成分的表达式, 见式(23)。

$$Z_i = a_i' X = a_{i1}x_1 + a_{i2}x_2 + \dots + a_{im}x_m, \quad i = 1, 2, \dots, m \quad (23)$$

2.4 主成分个数的确定

主成分分析的目的是采用降维的思想, 降低数据的维数^[2,5]。故一般只取前 $k (k < m)$ 个主成分。通常采用累计贡献率(c)的方法确定 k 值, 即根据实际问题给出一个 c 值, 一般 c 在 70%~85% 之间。若前 $k-1$ 个主成分的累计贡献率小于 c , 而前 k 个主成分的累计贡献率 $\geq c$, 则取 k 个主成分。

3 实例与 SAS 实现

3.1 问题与数据结构

3.1.1 两个实际问题及数据

【例1】某医学院的研究者测得 20 例肝病患者的 4 项肝功能指标的具体值,各指标分别为:转氨酶(X_1)、肝大指数(X_2)、硫酸锌浊度(X_3)和胎甲球(X_4)。试采用某种多元统计分析方法处理资料,并依据一定的理由对 20 例肝病患者进行分类。数据见表 1^[6]。

【例2】某研究者提供了某年我国 23 种肿瘤类期刊文献计量学方面 8 个定量指标的调查数据,这些指标分别为:载文量(X_1)、基金论文比(X_2)、总被引

频次(X_3)、影响因子(X_4)、5 年影响因子(X_5)、即年指标(X_6)、被引半衰期(X_7)和 Web 即年下载率(X_8)。试采用某种多元统计分析方法处理资料,并依据一定的理由对 23 种肿瘤类期刊进行排序。各指标的具体数据见表 2^[7]。

表 1 20 例肝病患者的 4 项肝功能指标的测定结果

Table 1 Measurement results of four liver function indicators in 20 patients with liver disease

患者编号	转氨酶 (X_1)	肝大指数 (X_2)	硫酸锌 浊度(X_3)	胎甲球 (X_4)
1	40	2.0	5	20
2	10	1.5	5	30
...
19	20	1.0	12	60
20	120	2.0	20	0

表 2 23 种肿瘤类期刊的文献计量学指标的调查结果

Table 2 Survey results of literature metrology indicators of 23 tumor journals

刊名	载文量 (X_1)	基金论文比 (X_2)	总被引频次 (X_3)	影响因子 (X_4)	5 年影响因子 (X_5)	即年指标 (X_6)	被引半衰期 (X_7)	Web 即年 下载率(X_8)
中华肿瘤杂志	234	0.35	2 705	1.415	1.394	0.120	6.0	38.900
癌症	316	0.49	1 935	0.742	0.879	0.104	4.3	35.200
中国肿瘤临床	507	0.33	1 710	0.420	0.673	0.026	5.3	22.300
中华放射肿瘤学杂志	102	0.17	942	1.011	1.290	0.029	5.3	7.500
肿瘤	191	0.40	702	0.470	0.525	0.021	4.7	25.000
中国肿瘤	243	0.13	660	0.358	0.367	0.058	3.7	15.500
中国肿瘤临床与康复	255	0.05	595	0.206	0.228	0.020	4.1	7.000
肿瘤防治研究	302	0.25	585	0.280	0.332	0.023	4.7	24.700
实用肿瘤杂志	198	0.17	566	0.326	0.332	0.035	5.3	24.100
实用癌症杂志	251	0.15	546	0.296	0.294	0.012	3.9	19.300
中国癌症杂志	188	0.14	526	0.355	0.419	0.032	3.8	25.400
肿瘤防治杂志	509	0.18	476	0.230	0.244	0.024	3.0	15.800
中国肺癌杂志	172	0.24	412	0.603	0.643	0.058	2.9	21.700
癌变·畸变·突变	120	0.45	341	0.406	0.452	0.167	5.3	31.200
实用肿瘤学杂志	233	0.09	302	0.137	0.220	0.009	5.5	8.600
临床肿瘤学杂志	325	0.06	298	0.318	0.262	0.037	2.8	15.100
中国肿瘤生物治疗杂志	82	0.70	296	0.387	0.459	0.037	4.4	11.200
肿瘤研究与临床	256	0.07	246	0.163	0.163	0.008	3.9	18.000
现代肿瘤医学	336	0.09	243	0.259	0.197	0.063	2.5	12.400
白血病·淋巴瘤	200	0.11	231	0.159	0.202	0.025	4.4	15.800
河南肿瘤学杂志	274	0.04	230	0.100	0.097	0.000	4.6	10.200
肿瘤学杂志	188	0.13	207	0.233	0.186	0.005	3.5	15.400
四川肿瘤防治	143	0.04	110	0.110	0.132	0.000	4.4	12.300

3.1.2 对数据结构的分析

在例 1 中,20 例患者均为肝病患者,研究者获取了这些患者肝功能状态的 4 项定量指标的值,故这是一个单组设计 4 元定量资料。

在例 2 中,23 种期刊均为肿瘤类期刊,研究者收集了关于这些期刊文献计量学方面的 8 项定量指标的数值,故这是一个单组设计 8 元定量资料。

3.1.3 创建 SAS 数据集

分析例 1 资料,设所需要的 SAS 数据步程序如下:

```
data a1;
input X1-X4;
obs=_n_;
cards;
```

```

40 2.0
5 20
10 1.5
5 30
120 3.0 13 50
250 4.5 18 0
120 3.5 9 50
10 1.5 12 50
40 1.0 19 40
270 4.0 13 60
280 3.5 11 60
170 3.0 9 60
180 3.5 14 40
130 2.0 30 50
220 1.5 17 20
160 1.5 35 60
220 2.5 14 30
140 2.0 20 20
220 2.0 14 10
40 1.0 10 0
20 1.0 12 60
120 2.0 20 0
;
run;

```

分析例 2 资料, 设所需要的 SAS 数据步程序如下:

```

data a1;
input X1-X8;
Jour=_n_;
cards;
234 0.35 2705 1.415 1.394 0.120 6.0 38.900
316 0.49 1935 0.742 0.879 0.104 4.3 35.200
... ..
188 0.13 207 0.233 0.186 0.005 3.5 15.400
143 0.04 110 0.110 0.132 0.000 4.4 12.300
;
run;

```

3.2 用 SAS 实现统计分析

3.2.1 分析例 1 中的资料

设所需要的 SAS 程序如下^[8]:

```

proc princomp out=aaa prefix=z;
var X1-X4;
run;

```

```

data a2; set aaa;
maxz=max(of z1-z4);
if maxz=z1 then do; zz='z1'; c1=obs; end;
if maxz=z2 then do; zz='z2'; c2=obs; end;
if maxz=z3 then do; zz='z3'; c3=obs; end;
if maxz=z4 then do; zz='z4'; c4=obs; end;
maxz=round(maxz, 0.001);
proc print data=a2;
var maxz c1-c4;
run;

```

【SAS 输出结果及解释】由 4 个变量两两之间的相关系数按一定规律排列出来, 所形成的方阵称为相关矩阵。见表 3。X₁ 与 X₂ 之间的相关性最高, r 值为 0.695; X₁ 与 X₃ 之间的相关性次之, r 值为 0.220。

表 3 相关矩阵

Table 3 Correlation matrix

矩 阵	X ₁	X ₂	X ₃	X ₄
X ₁	1.000			
X ₂	0.695	1.000		
X ₃	0.220	-0.148	1.000	
X ₄	0.025	0.135	0.071	1.000

注: X₁, 转氨酶; X₂, 肝大指数; X₃, 硫酸锌浊度; X₄, 胎甲球

相关矩阵的特征值见表 4。若希望累计贡献率在 85% 以上, 至少需要前 3 个主成分。

表 4 相关矩阵的特征值

Table 4 Eigenvalues of the correlation matrix

特征值编号	特征值	相邻特征值之差	比 例	累计贡献率
1	1.718	0.625	0.430	43.0%
2	1.094	0.112	0.273	70.3%
3	0.981	0.774	0.245	94.8%
4	0.207	-	0.052	100.0%

与 4 个特征值对应的特征向量的计算结果见表 5。

表 5 与特征值对应的特征向量

Table 5 Eigenvectors corresponding to the eigenvalues

变量名称	特征变量 z ₁	特征变量 z ₂	特征变量 z ₃	特征变量 z ₄
X ₁	0.700	0.095	-0.240	-0.666
X ₂	0.699	-0.284	0.058	0.664
X ₃	0.083	0.904	-0.270	0.319
X ₄	0.163	0.305	0.931	-0.121

由表 5 可知, 4 列上的数值代表 4 个特征向量的元素, 按前文的式 (7) 可写出 4 个主成分表达式, 下面仅呈现第 1 个主成分的表达式, 见式 (24)。

$$z_1 = 0.700x_1 + 0.699x_2 + 0.083x_3 + 0.163x_4 \quad (24)$$

值得一提的是: SAS 输出结果中的英文字母都是大写的, 而且是正体; 但表 5 中的系数是基于标准

化变量计算的结果,故呈现在表达式中,应将英文字母写成小写形式。

基于临床专业知识,结合各主成分表达式中系数的绝对值和正负号,解读各主成分所代表的含义:第一个主成分表达式中,变量 x_1 和 x_2 的系数最大,说明第一主成分受控于转氨酶和肝大指数,这两项指标的数值大,意味着肝病患者处于急性炎症状态。第二个主成分表达式中,变量 x_3 的系数最大,说明第二主成分受控于硫酸锌浊度,这项指标的数值大,意味着肝病患者处于慢性炎症状态。第三个主成分表达式中,变量 x_4 的系数最大,说明第三主成分受控于胎甲球,这项指标的数值大,意味着肝病

患者可能处于肝癌可疑状态。第四个主成分表达式中,变量 x_1 和 x_2 的系数的绝对值最大,但符号相反,由于第四个主成分的贡献率很小(5.170%),仅供参考,临床上认为可能指向急性肝萎缩。

患者分类的主要输出结果见表 6。由 1 类~4 类列可知,第一类包含编号为 4、8、9、11、15、17 这 6 位肝病者,他们属于急性炎症肝病者;第二类包含编号为 7、12、13、14、16 这 5 位肝病者,他们属于慢性炎症肝病者;第三类包含编号为 2、3、5、6、10、19 这 6 位肝病者,他们属于疑似肝癌患者;第四类包含编号为 1、18、20 这 3 位肝病者,他们属于疑似肝萎缩肝病者。

表 6 20 例肝病者分成 4 类的结果

Table 6 Results of 20 liver disease patients divided into 4 categories

患者编号	最大值	1类	2类	3类	4类	患者编号	最大值	1类	2类	3类	4类
1	0.185	-	-	-	1	11	1.121	11	-	-	-
2	0.430	-	-	2	-	12	2.109	-	12	-	-
3	0.776	-	-	3	-	13	0.337	-	13	-	-
4	2.076	4	-	-	-	14	3.024	-	14	-	-
5	0.949	-	-	5	-	15	0.707	15	-	-	-
6	1.026	-	-	6	-	16	0.483	-	16	-	-
7	0.802	-	7	-	-	17	0.232	17	-	-	-
8	2.293	8	-	-	-	18	-0.118	-	-	-	18
9	2.022	9	-	-	-	19	1.397	-	-	19	-
10	1.212	-	-	10	-	20	0.341	-	-	-	20

注:每位患者可基于 4 个主成分计算出 4 个得分值,取其中最大值作为该患者的最终得分值

结论:在本例中,求出了 4 个主成分,从第一到第四个主成分的贡献率依次为 43.0%、27.3%、24.5% 和 5.2%。将每位肝病者在 4 项定量指标上的取值分别代入所求出的 4 个主成分表达式中进行计算(注意:对原变量需要做标准化变换),并按绝对值最大作为分类的“标准”,可得到 4 类分类结果。

3.2.2 分析例 2 中的资料

设所需要的 SAS 过程步程序如下:

```
proc princomp out=aaa prefix=z;
var X1-X8;run;
data a2;set aaa;
zt=0.531*z1+0.157*z2+0.121*z3+0.086*z4;
zt=round(zt,0.001);
proc rank descending data=a2 out=bbb;
var zt;ranks order;run;
proc print data=bbb noobs;
var jour zt order;run;
```

【SAS 程序说明】由 proc princomp 过程计算的结果可知,前 4 个主成分的累计贡献率为 89.45%,大

于 85.00%,故取前 4 个主成分,并以它们各自的贡献率为权重,进行加权平均,求出与每种期刊对应的一个综合评价指标 z_t 的值。实现的语句为:“ $z_t=0.531*z_1+0.157*z_2+0.121*z_3+0.086*z_4$ ”,该语句中的系数分别代表前 4 个主成分的贡献率。

【SAS 输出结果及解释】因篇幅所限,由 proc princomp 过程输出的结果从略。下面仅输出基于主成分计算结果进行综合评价的结果,见表 7。两列“总分”代表各期刊的综合评分值;两列“排序”代表期刊的排序位次,数值越小,表明综合评分值越大。

结论:在本例中,基于 8 项文献计量学指标的数值实施了主成分分析,获得了累计贡献率达到 89.45% 的前 4 个主成分;再基于这 4 个主成分以及它们各自的贡献率,采取加权算法获得综合评价指标 z_t ;进而,基于 z_t 的表达式,算出每种期刊的综合得分值,并对综合得分值排序,最终获得了全部 23 种肿瘤类期刊的排列顺序。这是将“无序样品(本文指‘期刊’)”转化成“有序样品”的过程,属于传统综合评价方法的一种扩展^[9]。

表7 对23种肿瘤期刊进行综合评价的结果

Table 7 Comprehensive evaluation results of 23 cancer journals

期刊编号	总分	排序	期刊编号	总分	排序
1	3.546	1	13	-0.074	9
2	2.044	2	14	0.749	5
3	1.393	3	15	-0.655	15
4	0.879	4	16	-0.756	17
5	0.398	6	17	-0.182	10
6	-0.273	12	18	-0.813	19
7	-0.767	18	19	-0.892	20
8	0.110	7	20	-0.738	16
9	0.092	8	21	-0.964	21
10	-0.391	13	22	-0.969	22
11	-0.184	11	23	-1.154	23
12	-0.399	14	-	-	-

4 讨论与小结

4.1 讨论

主成分分析方法是众多多元统计分析方法的基础,虽然计算过程和方法比较简单,但从它的计算结果中可以引申出多种非常有价值的应用。由本文的例1可知,主成分分析可以被用来实现“样品聚类”;由本文的例2可知,主成分分析可以被用来实现“传统综合评价”,使无序样品变成有序样品,从而实现样品排序。事实上,主成分分析还有许多其他应用,因篇幅所限,不再赘述。

4.2 小结

本文介绍了与主成分分析有关的5个基本概念、计算方法和两个实例以及SAS实现。5个基本概念包括相关矩阵、特征值与特征向量、主成分变量、主成分表达式和主成分的性质;计算方法涉及特征值与特征向量的求法、主成分分析的计算原理、系数估计和个数确定;两个实例涉及的资料分别是“20例肝病患者的4项肝功能指标的测定结果”和“23种肿瘤类期刊的文献计量学指标的调查结果”。基于两个实例,分别实现了样品聚类和样品排序。

参考文献

- [1] Rice JA. Mathematical statistics and data analysis [M]. 2版. 北京: 机械工业出版社, 2003: 129-134.
Rice JA. Mathematical statistics and data analysis [M]. 2nd edition. Beijing: China Machine Press, 2003: 129-134.
- [2] Johnson DE. 应用多元统计分析方法[M]. 北京: 高等教育出版社, 2005: 77-92, 93-146.
Johnson DE. Applied multivariate methods for data analysis [M]. Beijing: Higher Education Press, 2005: 77-92, 93-146.
- [3] 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: 189-209.
Yu JH, Yang WQ. Multivariate statistical analysis and application [M]. Guangzhou: Sun Yat-sen University Press, 2005: 189-209.
- [4] 郭大钧. 大学数学手册[M]. 济南: 山东科学技术出版社, 1985: 300-302.
Guo DJ. Handbook of college mathematics [M]. Ji'nan: Shandong Science and Technology Press, 1985: 300-302.
- [5] 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: 430-480.
Johnson RA, Wichern DW. Applied multivariate statistical analysis [M]. 6th edition. Beijing: Tsinghua University Press, 2008: 430-480.
- [6] 胡良平. 现代统计学与SAS应用[M]. 北京: 军事医学科学出版社, 1996: 316-323.
Hu LP. Modern statistics and SAS applications [M]. Beijing: Military Medical Science Press, 1996: 316-323.
- [7] 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: 19-39.
Hu LP. Problem-oriented statistics: (3) experimental design and multivariate statistical analysis [M]. Beijing: People's Mental Publishing House, 2012: 19-39.
- [8] SAS Institute Inc. SAS/STAT®15.1 user's guide[M]. Cary, NC: SAS Institute Inc, 2018: 7891-7932.
- [9] 胡良平, 黄国平. 医学科研设计与关键技术[M]. 成都: 四川大学出版社, 2017: 349-360.
Hu LP, Huang GP. Medical research design methods and key technologies [M]. Chengdu: Sichuan University Press, 2017: 349-360.

(收稿日期:2023-06-05)

(本文编辑:吴俊林)