

# 合理进行多元分析——广义主成分分析

胡纯严<sup>1</sup>, 胡良平<sup>1,2\*</sup>

(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】** 本文目的是介绍与广义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。基本概念包括合成资料、拟合成资料、部分合成资料和广义主成分分析; 计算方法涉及对数中心化, 构造协方差矩阵  $S$  以及求矩阵  $S$  的特征值和特征向量; 两个实例分别是“某矿石中 5 种成分的含量”和“1993 年我国 30 个地区农民家庭消费资料”。借助 SAS 对两个实例中的定量资料进行广义主成分分析, 只需要 1 个主成分就可包含多个原变量所包含的 85% 以上的信息, 取得了很好的降维效果; 在例 2 中, 基于广义主成分的计算结果, 还实现了对地区的排序和初步分档。

**【关键词】** 合成资料; 对数中心化; 协方差矩阵; 样品排序; 广义主成分

中图分类号: R195.1

文献标识码: A

doi: 10.11886/scjsws20230605002

## Reasonably carry out multivariate analysis: generalized principal component analysis

Hu Chunyan<sup>1</sup>, Hu Liangping<sup>1,2\*</sup>

(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 paper was to introduce the basic concepts, calculation methods, two examples and SAS implementation related to generalized principal component analysis. Basic concepts included synthetic data, quasi-synthetic data, partial synthetic data and generalized principal component analysis. The calculation method involved logarithmic centralization, constructing the covariance matrix  $S$  and finding the eigenvalues and eigenvectors of the matrix  $S$ . The data involved in the two examples were percentage content of 5 components in a certain ore and household consumption data of farmers in 30 regions of China in 1993. With the help of SAS software, generalized principal component analysis was carried out on the quantitative data in the two examples. Only one principal component was needed to contain more than 85% of the information contained in multiple original variables, and a good dimensionality reduction effect had been achieved. In example 2, based on the calculation results of generalized principal components, the sorting and preliminary classification of regions were also realized.

**【Keywords】** Synthetic data; Logarithmic centralization; Covariance matrix; Sample sorting; Generalized principal component

真正的多元定量资料是指需要同时分析的定量结果变量的个数 $\geq 2$ , 除此之外, 还需要看是否有影响因素或自变量。若没有, 就称为单组设计多元定量资料; 若有, 就属于非单组设计多元定量资料, 例如, 两因素析因设计多元定量资料。针对单组设计多元定量资料, 如何选择合适的多元统计分析方法, 需要结合实际问题和研究者的分析目的, 从十几种多元统计分析方法中选择一种或几种。本文将介绍类似于线性主成分分析的一种多元统计分析方法, 即非线性主成分分析, 也称为广义主成分分析<sup>[1]</sup>。

## 1 基本概念

### 1.1 合成资料

设具有同质性的  $n$  个样品, 当每个样品各指标

测定值之和为 100%、且每个测定值皆大于 0 时, 称这种资料为合成资料<sup>[1-2]</sup>。

### 1.2 拟合成资料

在一些实际问题中, 单组设计多元定量资料可能不满足前述关于合成资料的条件。此时, 可以通过下面的方法将其转变为合成资料, 它被称为拟合成资料。具体做法: 先求出每个样品中  $m$  个变量的和, 设第  $i$  行的和为  $R_i (i=1, 2, \dots, n)$ 。其次, 求出全部  $n$  个和的最大值, 记为  $\text{Max}R$ ; 再令  $L = \text{Max}R + a$  ( $a$  为任一正数), 并用此  $L$  值作为样本中第  $i$  行上的每个值的分母, 将求得的商记为  $y_{ij} (i=1, 2, \dots, n; j=1, 2, \dots, m)$ 。最后, 增加一个新变量  $y_{i(m+1)}$ , 它在第  $i$  行上的取值为  $y_{i(m+1)} = 1 - \sum y_{ij}$  (其中  $i=1, 2, \dots, n$ )。于是, 由  $\{y_{ij}\}$  组成的  $n \times (m+1)$  数据阵就是一个拟合成资料<sup>[2]</sup>。

### 1.3 部分合成资料

在一个合成资料中,研究者并不关心所有  $m$  个变量,而只想考查其中  $k(k < m)$  个变量时,每个样品的  $k$  个变量的取值之和(设为  $W$ )总会小于 1。于是,可以产生一个“补足变量  $x_{k+1}$ ”,并令  $x_{k+1} = 1 - W$ ,这样形成的合成资料被称为“部分合成资料”<sup>[2]</sup>。

### 1.4 广义主成分分析

对于单组设计非合成多元定量资料所做的主成分分析属于线性主成分分析,而对于合成资料采用非线性降维法,且希望用尽可能少的主成分反映原来  $m$  个指标较多信息的方法,称为合成资料的主成分分析,也称为非线性主成分分析或广义主成分分析<sup>[1]</sup>。

## 2 计算方法

### 2.1 对数中心化

首先,取原变量的对数,见式(1)。再对行向量进行中心化变换,见式(2)。

$$y_{ij} = \ln(X_{ij}), \quad i = 1, 2, \dots, n; j = 1, 2, \dots, m \quad (1)$$

$$z_{ij} = y_{ij} - \sum_{i=1}^m y_{ij} / m, \quad i = 1, 2, \dots, n; j = 1, 2, \dots, m \quad (2)$$

### 2.2 构造协方差矩阵 $S$

求出  $Z = (z_{ij})_{n \times m}$  矩阵中  $m$  个变量的协方差矩阵  $S$ , 见式(3)。

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

式(3)中,第  $i$  行第  $j$  列上的元素  $s_{ij}$  为对数中心化后的数据的第  $i$  个变量与第  $j$  个变量之间的协方差。

### 2.3 求矩阵 $S$ 的特征值和特征向量

求矩阵  $S$  的特征值和特征向量涉及矩阵运算,通常可借助统计软件来实现<sup>[3]</sup>。基于求出的特征向量,主成分的表达式见式(4)。

$$\begin{cases} P_1 = a_{11}(Z_1 - \bar{Z}_1) + a_{12}(Z_2 - \bar{Z}_2) + \dots + a_{1m}(Z_m - \bar{Z}_m) \\ P_2 = a_{21}(Z_1 - \bar{Z}_1) + a_{22}(Z_2 - \bar{Z}_2) + \dots + a_{2m}(Z_m - \bar{Z}_m) \\ \dots \\ P_m = a_{m1}(Z_1 - \bar{Z}_1) + a_{m2}(Z_2 - \bar{Z}_2) + \dots + a_{mm}(Z_m - \bar{Z}_m) \end{cases} \quad (4)$$

式(4)中,  $P_j$  代表第  $j$  个主成分;  $a_{ij}$  代表基于式(3)求出的与第  $i$  个特征值对应的特征向量的第  $j$  个分量。

## 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】某矿石中 5 种成分的含量百分比见表 1<sup>[2]</sup>。共收集 25 个矿石样品,每个样品的 5 个成分的含量百分比之和为 100%。试寻找少数几个综合指标反映原来 5 个指标 85% 以上的信息,并基于分析结果,明确在综合指标中起主导作用的变量。

【例2】1993 年我国 30 个省、自治区和直辖市的农民家庭消费资料见表 2<sup>[1]</sup>。试基于食品、衣着、居住、家庭设备及有关服务、医疗保健、交通通讯、文教和娱乐消费、其他非商品及服务消费这 8 项消费指标的取值,对 30 个地区排序。

表 1 25 个矿石样品中 5 种成分的测定结果 (%)

Table 1 Determination results of 5 components in 25 ore samples

编号	X <sub>1</sub>	X <sub>2</sub>	X <sub>3</sub>	X <sub>4</sub>	X <sub>5</sub>
1	48.8	31.7	3.8	6.4	9.3
2	48.2	23.8	9.0	9.2	9.8
...	...	...	...	...	...
25	45.9	24.9	9.7	9.8	9.7

注: X<sub>1</sub>~X<sub>5</sub> 分别代表矿石中 5 种成分的含量

#### 3.1.2 对数据结构的分析

例 1 中, 25 个矿石样品中的每个样品都含有 5 种成分, 每个样品 5 种成分含量之和为 100%, 但各样品中 5 种成分的含量不完全相同, 故这是一个单组设计 5 元定量合成资料。

例 2 中, 30 个地区中的每个地区 8 项指标取值之和近似 100%, 故这是一个单组设计 8 元定量合成资料。

表 2 1993 年我国 30 个地区农民家庭消费资料 (%)

Table 2 Household consumption data of farmers in 30 regions of China in 1993

地区	食品(X <sub>1</sub> )	衣着(X <sub>2</sub> )	居住(X <sub>3</sub> )	家庭设备及有关服务(X <sub>4</sub> )	医疗保健(X <sub>5</sub> )	交通通讯(X <sub>6</sub> )	文教和娱乐消费(X <sub>7</sub> )	其他非商品及服务消费(X <sub>8</sub> )
北京	48.16	10.58	11.29	9.26	4.62	3.04	11.61	1.40
天津	53.31	9.54	15.98	6.10	3.39	2.54	7.59	1.63
...	...	...	...	...	...	...	...	...
新疆	52.15	12.94	14.70	5.37	4.22	2.77	6.24	1.61

注: 因仅保留两位小数和四舍五入等原因, 各行数据之和近似 100%

### 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(编号为: PRG1)如下。

```
data a1;
input X1-X5;
ore=_n_;
cards;
48.8 31.7 3.8 6.4 9.3
48.2 23.8 9.0 9.2 9.8
37.0 9.1 34.2 9.5 10.2
50.9 23.8 7.2 10.1 8.0
44.2 38.3 2.9 7.7 6.9
52.3 26.2 4.2 12.5 4.8
44.6 33.0 4.6 12.2 5.6
34.6 5.2 42.9 9.6 7.7
41.2 11.7 26.7 9.6 10.8
42.6 46.6 0.7 5.6 4.5
49.9 19.5 11.4 9.5 9.7
45.2 37.3 2.7 5.5 9.3
32.7 8.5 38.9 8.0 11.9
41.4 12.9 23.4 15.8 6.5
26.2 17.5 15.8 28.3 12.2
32.3 7.3 40.9 12.9 6.6
43.2 44.3 1.0 7.8 3.7
49.5 32.3 3.2 8.7 6.3
42.3 15.8 20.4 8.3 13.2
44.6 11.5 23.8 11.6 8.5
45.8 16.6 16.8 12.0 8.8
49.9 25.0 6.8 10.9 7.4
48.6 34.0 2.5 9.4 5.5
45.5 16.6 17.6 9.6 10.7
45.9 24.9 9.7 9.8 9.7
;
```

分析例 2 资料, 设所需要的 SAS 数据步程序(编号为: PRG2)如下。

```
data b1;
input X1-X8;
rig=_n_;
cards;
48.16 10.58 11.29 9.26 4.62 3.04 11.61 1.40
53.31 9.54 15.98 6.10 3.39 2.54 7.59 1.63
58.39 7.50 13.86 5.16 6.32 1.63 5.90 1.23
57.36 11.20 11.89 5.80 3.57 2.00 6.74 1.44
```

```
58.40 8.14 14.04 4.56 3.76 1.83 8.15 1.12
55.14 10.47 12.34 4.79 4.30 1.98 8.91 2.07
60.63 9.70 9.99 4.03 4.44 1.98 8.91 2.07
61.40 8.88 15.25 3.88 3.38 1.56 5.04 0.97
46.43 7.14 16.24 11.75 2.28 2.98 9.99 3.19
50.21 7.40 17.89 8.58 3.32 2.81 7.76 2.03
50.16 7.46 17.35 7.63 3.97 3.05 6.83 3.54
63.90 6.95 10.09 5.41 3.31 1.87 6.84 1.62
60.60 4.99 13.44 5.04 2.43 2.90 7.46 3.13
61.35 5.94 13.36 5.00 3.45 2.41 7.27 1.22
57.36 8.31 16.64 5.60 3.41 2.04 8.32 1.31
59.21 8.15 13.27 5.32 4.01 1.55 6.94 1.54
61.85 6.46 10.75 5.43 3.07 1.51 9.88 1.05
61.10 5.56 13.23 4.91 2.98 1.98 9.08 1.14
52.77 3.73 16.98 6.69 3.17 4.28 9.41 2.97
64.33 4.31 10.70 6.24 2.77 2.01 9.89 0.75
67.17 4.67 7.06 4.87 2.55 2.66 9.08 1.94
63.29 6.33 12.20 5.23 3.13 1.95 6.99 0.88
70.98 6.46 8.37 4.32 1.87 1.49 4.92 1.60
61.20 6.92 13.20 6.50 3.22 1.65 5.96 1.35
66.53 8.78 18.21 3.40 0.49 0.98 0.62 1.00
57.13 7.75 15.39 5.50 4.58 1.33 7.08 1.24
55.36 7.35 19.24 5.13 4.01 1.62 6.02 1.28
55.50 9.54 20.86 4.43 3.84 1.75 3.01 1.08
58.75 8.87 12.00 5.57 4.67 2.04 7.11 0.95
52.15 12.94 14.70 5.37 4.22 2.77 6.24 1.61
;
```

### 3.2 用 SAS 实现统计分析

#### 3.2.1 分析例 1 中的资料

进行线性主成分分析, 设所需 SAS 程序如下<sup>[3]</sup>:

```
proc princomp data=a1 out=aaa prefix=z;
var X1-X5;
run;
```

【SAS 输出结果及解释】求出的特征值等见表 3。由表 3 可知, 欲使累计贡献率大于 85%, 至少需取 3 个主成分, 降维的效果不理想。因此, 应选用广义主成分分析。

进行广义主成分分析, 设所需要的 SAS 程序(编号为: PRG3)如下。

```
data a2;
%let m=5;
%let n=25;
array a(&n);
```

表 3 相关矩阵的特征值

Table 3 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	2.932	1.979	0.586	58.641%
2	0.953	0.276	0.191	77.698%
3	0.677	0.239	0.135	91.243%
4	0.438	0.438	0.088	100.000%
5	0.000	-	0.000	100.000%

```
array y(&m);
array x(&n,&m);
do i=1 to &n;
a(i)=0;
do j=1 to &m;
input x(i,j) @@;
x(i,j)=log(x(i,j));
end;
do j=1 to &m;
a(i)=a(i)+x(i,j);
end;
a(i)=a(i)/&m;
do j=1 to &m;
x(i,j)=x(i,j)-a(i);
end;
do j=1 to &m;
y(j)=x(i,j);
end;
output;
end;
cards;
(此处输入前文“3.1.3节”中的25行5列数据)
run;
data a3;
set a2;
proc princomp data=a3 covariance prefix=z;
var y1-y&m;
run;
【SAS输出结果及解释】求出的特征值等见表4。
```

由表4可知,欲使累计贡献率大于85%,只需取

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	1.596	1.498	0.319	91.969%
2	0.098	0.061	0.020	97.641%
3	0.037	0.033	0.007	99.760%
4	0.004	0.004	0.001	100.000%
5	0.000	-	0.000	100.000%

第一个主成分。

5个指标对数中心化变换后的均值向量  $z=(1.105, 0.326, -0.471, -0.372, -0.588)'$ 。与第一个特征值对应的特征向量  $a=(-0.205, -0.289, 0.781, -0.016, 0.030)'$ 。将以上求出的均值向量和特征向量代入式(4),可得第一主成分的表达式,见式(5)。

$$P_1 = -0.205(Z_1 - 1.105) - 0.289(Z_2 - 0.326) + 0.781(Z_3 + 0.471) - 0.016(Z_4 + 0.372) + 0.030(Z_5 + 0.588) \quad (5)$$

从主成分  $P_1$  表达式的系数可看出:  $P_1$  主要受  $Z_3$  与  $Z_2$  的影响,而且,它们的符号相反。即当这种矿石的样品中  $X_3$  的含量高时,  $X_2$  的含量就低,反之亦然。所以该主成分可以取名为该矿石中  $X_3$  与  $X_2$  含量的对比因子。

结论:对于合成资料,直接进行线性主成分分析<sup>[4-5]</sup>,效果不好;而进行广义主成分分析,效果比较理想。就本例而言,广义主成分分析的结果表明:只需要第一主成分,就可反映原先5个定量指标91.969%的信息。基于该主成分表达式中绝对值最大的两个系数,可将该主成分命名为“矿石中  $X_3$  与  $X_2$  含量的对比因子”。

### 3.2.2 分析例2中的资料

进行线性主成分分析,设所需SAS程序如下:

```
proc princomp data=a1 out=aaa prefix=z;
var X1-X8;
run;
```

【SAS输出结果及解释】求出的特征值等内容见表5。

表 5 相关矩阵的特征值

Table 5 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	3.170	1.315	0.396	39.630%
2	1.855	0.485	0.232	62.820%
3	1.370	0.789	0.171	79.940%
4	0.581	0.017	0.073	87.200%
5	0.564	0.281	0.071	94.250%
6	0.283	0.112	0.035	97.790%
7	0.171	0.165	0.021	99.930%
8	0.006	-	0.000	100.000%

由表5可知,欲使累计贡献率大于85%,至少需取4个主成分,降维的效果不明显。因此,应选用广义主成分分析。

进行广义主成分分析,设所需要的SAS程序(编号为:PRG4)如下。

与前面“3.1.3节”中的程序“PRG3”基本相同,其中,第二句和第三句需修改为“%let m=8; %let n=30;”。另外,“cards;”语句后面的数据需要用“3.1.3

节”中的程序“PRG2”中的数据替换;“proc princomp”语句需修改为“proc princomp data=a3 noint covariance prefix=p;”。

【SAS 输出结果及解释】求出的特征值等见表 6。

表 6 相关矩阵的特征值  
Table 6 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	9.464	1.315	0.396	94.360%
2	0.274	0.485	0.232	97.090%
3	0.157	0.789	0.171	98.650%
4	0.055	0.017	0.073	99.200%
5	0.041	0.281	0.071	99.610%
6	0.025	0.112	0.035	99.860%
7	0.014	0.165	0.021	100.000%
8	0.000	-	0.000	100.000%

由表 6 可知,欲使累计贡献率大于 85%,只需取第一个主成分。

5 个指标对数中心化变换后的均值向量  $z=(2.239, 0.185, -0.114, -0.639, -1.108, 0.085, -1.428)'$ 。与第一个特征值对应的特征向量  $a=(-0.731, -0.066, 0.257, -0.037, -0.208, -0.362, 0.021, -0.468)'$ 。将以上求出的均值向量和特征向量代入式(4),可得第一主成分的表达式,见式(6)。

$$P_1=0.731(Z_1-2.239)+0.066(Z_2-0.185)+0.257(Z_3-0.779)-0.037(Z_4+0.114)-0.208(Z_5+0.639)-0.362(Z_6+1.108)+0.021(Z_7-0.085)-0.468(Z_8+1.428) \quad (6)$$

基于第一主成分对地区排序,所需要的 SAS 程序如下:

```
data a4;
set aaa;
zt=0.731*(p1-2.239)+0.066*(p2-0.185)+0.257*(p3-0.779)-0.037*(p4+0.114)-0.208*(p5+0.639)-0.362*(p6+1.108)+0.021*(p7-0.085)-0.468*(p8+1.428);
zt=round(zt,0.001);
proc rank descending data=a4 out=bbb;
var zt;
ranks order;
run;
proc sort data=bbb out=ccc;
key zt/ascending;
run;
data ddd;
set ccc;
```

```
order=int(order);
length name $10;
if rig=1 then name='北京';
else if rig=2 then name='天津';
else if rig=3 then name='河北';
.....
else if rig=29 then name='宁夏';
else if rig=30 then name='新疆';
proc print data=ddd; var name order;run;
data e1 e2 e3;
set ddd;
if order>20 then output e1;
else if 10<order<21 then output e2;
else if 0<order<11 then output e3;
keep rig name order;
run;
data e1(keep=id rig1 name1 order1);
set e1;id=_n_;
length name1 $10;
rename rig=rig1 name=name1 order=order1;
run;
data e2(keep=id rig2 name2 order2);
set e2;id=_n_;
length name2 $10;
rename rig=rig2 name=name2 order=order2;
run;
data e3(keep=id rig3 name3 order3);
set e3;id=_n_;
length name3 $10;
rename rig=rig3 name=name3 order=order3;
run;
data eee(drop=id);
merge e1 e2 e3;
by id;
ods rtf;
proc print data=eee noobs;
run;
```

【SAS 输出结果及解释】因篇幅所限,由 proc princomp 过程输出结果从略。下面仅输出基于第一主成分取值对 30 个地区进行排序的结果,见表 7。

在表 7 中,前 3 列的 10 个地区属于“高分地区”;中间 3 列的 9 个地区属于“中分地区”;最后 3 列的 11 个地区属于“低分地区”。“高分地区”的特点是:食品、居住、文教和娱乐消费占比较大;“中分地区”的特点

表7 基于第一主成分取值对30个地区排序的结果

Table 7 Sorting results of 30 regions based on the values of the first principal component

编号	分值	地区	编号	分值	地区	编号	分值	地区
19	30	广东	6	20	辽宁	22	10	四川
9	29	上海	14	19	江西	24	10	云南
11	28	浙江	12	18	安徽	5	9	内蒙古
13	27	福建	4	17	山西	16	8	河南
21	26	海南	23	16	贵州	17	7	湖北
1	25	北京	18	15	湖南	3	6	河北
10	24	江苏	20	14	广西	27	5	甘肃
30	23	新疆	15	13	山东	28	4	青海
2	22	天津	29	12	宁夏	8	3	黑龙江
7	21	吉林	-	-	-	26	2	陕西
-	-	-	-	-	-	25	1	西藏

注：“编号”代表地区的编号；“分值”代表基于第一主成分计算所得的分值

是：食品、衣着、居住占比较大；而“低分地区”的特点是：食品、衣着、居住、文教和娱乐消费占比较大。

结论：对于合成资料，直接进行线性主成分分析，效果不好；而进行广义主成分分析，效果很明显。就本例而言，广义主成分分析的结果表明：只需要第一主成分，就可反映原先8个定量指标94.360%的信息。基于该主成分表达式中绝对值最大的4个系数（对应的变量分别为 $X_1$ 、 $X_3$ 、 $X_6$ 、 $X_8$ ），可将该主成分命名为“农民家庭消费中食品（ $X_1$ ）和居住（ $X_3$ ）”与“交通通讯（ $X_6$ ）和其他非商品及服务消费（ $X_8$ ）的对比因子”。

## 4 讨论与小结

### 4.1 讨论

采用线性主成分分析处理合成资料，其信息较分散，表现为大多数特征值彼此相差不大，使主成分分析法降维的效果欠佳；采用广义主成分分析处理合成资料，若降维效果不够明显，改进的做法是基于未对均值校正的协方差矩阵求特征值和特征向量。例如，在本文例2中，若基于对均值校正的协方差矩阵进行计算，得到的结果是：所求得特征值均小于1，且特征值之间的差距非常小。为了解决这个问题，在调用princomp过程时，需要在“proc princomp”语句中增加一个选项“noint”。此时，可以获得降维效果很好的计算结果。

### 4.2 小结

本文介绍了与广义主成分分析有关的4个基本概念，计算方法和两个实例以及SAS实现。4个基本概念分别是合成资料、拟合成资料、部分合成资料和广义主成分分析；计算方法涉及对数中心化、构造协方差矩阵 $S$ 、求矩阵 $S$ 的特征值和特征向量；两个实例分别是“某矿石中5种成分的含量百分比”和“1993年我国30个地区农民家庭消费资料”。基于两个实例，实现了仅需要1个主成分就可反映多个原变量所包含的85%以上的信息；在例2中，还实现了对地区的排序和初步分档<sup>[6-7]</sup>。

### 参考文献

- [1] 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.
- [2] 孙尚拱. 实用多变量统计方法与计算程序[M]. 北京: 北京医科大学、中国协和医科大学联合出版社, 1990: 126-137.  
Sun SG. Practical multivariate statistical methods and calculation programs [M]. Beijing: Beijing Medical University and China Union Medical University Joint Publishing House, 1990: 126-137.
- [3] SAS Institute Inc. SAS/STAT®15.1 user's guide[M]. Cary, NC: SAS Institute Inc, 2018: 7891-7932.
- [4] 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.
- [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]. 6<sup>th</sup> edition. Beijing: Tsinghua University Press, 2008: 430-480.
- [6] 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.
- [7] 孙振球. 医学统计学[M]. 北京: 人民卫生出版社, 2002: 373-396.  
Sun ZQ. Medical statistics [M]. Beijing: People's Mental Publishing House, 2002: 373-396.

(收稿日期: 2023-06-05)

(本文编辑: 陈霞)