合理进行多元分析——定量资料对应分析

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】 本文目的是介绍与定量资料对应分析有关的基本概念、计算方法、两个实例以及SAS实现。基本概念包括变量与样品、显变量与隐变量、因子分析、R型分析与Q型分析、对应分析；计算方法涉及基本原理、变量变换、构建R型和Q型协方差矩阵、因子分析；两个实例分别为“某年中国10个省份农村居民家庭人均消费支出数据”和“不同民族的各种基因出现的频率”；借助SAS软件，对两个实例中的定量资料进行了定量资料对应分析，并对SAS输出结果做出了解释。

【关键词】 对应分析；因子分析；变量变换；协方差矩阵；特征值
中图分类号：R195.1 文献标识码：A doi：10.11886/scjsws20230726001

Reasonably carry out multivariate analysis: quantitative data correspondence 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 and SAS implementation related to the quantitative data correspondence analysis. The basic concepts included variable and sample, explicit variable and latent variable, factor analysis, R-type analysis and Q-type analysis, and correspondence analysis. The calculation methods involved the basic principles, variable transformation, construction of R-type and Q-type covariance matrices, and factor analysis. The data in the two examples were "per capita consumption expenditure data of rural households in 10 provinces in China in a certain year" and "frequency of occurrence of various genes of different ethnic groups". With the help of SAS software, the quantitative data in the two examples were analyzed by quantitative data correspondence analysis, and the SAS output results were given an explanation.

【Keywords】 Correspondence analysis; Factor analysis; Variable transformation; Covariance matrix; Eigenvalue

对于一个单组设计多元定量资料，通常各行代表样品或观测，各列代表定量变量或观测指标。对于以这种形式呈现的多元定量资料，大多数多元统计分析方法可以实现的统计分析目的包括如下两个：其一，研究全部定量变量之间的相互或依赖关系；其二，研究全部样品之间的相似或接近程度。本文将介绍一种非常特殊的多元统计分析方法，即定量资料对应分析。该法可以把变量与样品这两种截然不同的“属性”改造成同一种“属性”，从而将它们所代表的“数据点”呈现在同一个直角坐标系中，进而实现研究变量与样品之间关联关系的目的。

1 基本概念

1.1 变量与样品

在统计学上，最常见的一种数据形式就是对n个个体观测m个指标的取值，将所获得的结果写成n行m列的一个数据矩阵。每一列代表一个指标在n个个体上的全部信息，统计学上习惯将指标称为“变量”；每一行代表一个个体在m个指标上的全部信息，统计学上习惯将个体称为“样品”。需要说明的是，严格意义上的“指标”通常指“观测结果”，而“变量”可以包括“原因变量”与“结果变量”，故采用“变量”取代“指标”，更贴近统计资料的真实情况。

1.2 显变量与隐变量

通常的回归分析方法是以模型的形式呈现因变量与自变量之间的依赖关系，回归模型中可以直接观测其取值的变量被称为“显变量”，例如身高、体重、血压等；有些具有实际意义但无法直接测量其取值的变量，被称为“隐变量或潜变量”[1-2]，例如内向型性格、交感神经的状态、个人能力等。

1.3 因子分析

通过显变量的取值大小，找出在背后控制或影响它们的隐变量，并用模型将这两类变量互相表达出来（即以显变量线性表达隐变量，或以隐变量线性表达显变量）的一种多元统计分析方法，称为因子分析[3-4]。

1.4 R型分析和Q型分析

在统计学上，通常把以变量为研究对象的因子分析称为R型因子分析，简称为R型分析；而把以样品为研究对象的因子分析称为Q型因子分析，简称为Q型分析。

1.5 对应分析

对应分析是将变量与样品所反映的信息融合在一起进行联合分析，并以二维平面图形呈现它们之间关联关系的一种多元统计分析方法[5-6]。它主要解决以下两方面的问题：其一，根据R型分析和Q型分析的内在联系，可将变量与样品同时反映到同一个直角坐标系中，便于对实际问题进行分析，从而揭示变量与样品之间的联系；其二，从R型因子分析出发，能直接获得Q型因子分析的结果，从而克服了样品容量大、作Q型分析存在的计算上的困难。

2 计算方法

2.1 基本原理

对应分析的关键是利用数据变换方法，使含有 n 个样品 m 个变量的原始数据矩阵 $X=(x)_{nm}$ 变成另一个矩阵 $Z=(z)_{nm}$，并使 $R=Z'Z$（分析变量之间关系的协方差矩阵）与 $Q=ZZ'$（分析样品之间关系的协方差矩阵）具有相同的非零特征根，它们相应的特征向量之间也有密切关系[7-8]。

对协方差矩阵 R 和 Q 进行加权主成分分析或因子分析，分别能提取两个最重要的公因子 R_1、R_2 与 Q_1、Q_2。由于采取的是一种特殊变换方法，公因子 R_1 与 Q_1 在本质上是相同的，R_2 与 Q_2 在本质上也是相同的，故可用 $\dim_1$ 作为 R_1 和 Q_1 的统一标志，用 $\dim_2$ 作为 R_2 和 Q_2 的统一标志。可将（R_1，Q_1）和（R_2，Q_2）两组数据点呈现在由（$\dim_1$，$\dim_2$）组成的同一个直角坐标系中，以便考查变量与样品之间的关联关系。

2.2 变量变换

设原始数据矩阵 $X=(X_{ij})_{nm}$，$i=1,2,\cdots,n$，其中 n 为样品数；$j=1,2,\cdots,m$，其中 m 为变量数。又设 $X_{i.}$ 为第 i 行的合计、$X_{.j}$ 为第 j 列的合计、$X_{..}$ 为全部数据的合计，则变量变换公式见式（1）。

$$Z_{ij}=\frac{X_{ij}-X_{i.}X_{.j}/X_{..}}{\sqrt{X_{i.}X_{.j}}} \quad (1)$$

由此变换产生出矩阵 Z，$Z=(Z_{ij})_{nm}$。

2.3 构建R型和Q型协方差矩阵

基于前述产生的数据矩阵 Z，将其转置得到矩阵 Z'，将这两个矩阵相乘，得到Q型协方差矩阵，见式（2）。

$$Q=ZZ' \quad (2)$$

再改变这两个矩阵的先后顺序，可得到R型协方差矩阵，见式（3）。

$$R=Z'Z \quad (3)$$

2.4 因子分析

分别对 $R=Z'Z$、$Q=ZZ'$ 实施因子分析，从每次因子分析的结果中，提取前两个公因子，将它们组建成一个二维直角坐标系，可以证明，这两个直角坐标系是完全重合的。因为R型协方差矩阵与Q型协方差矩阵具有相同的非零特征根。

由此可知，对于经式（1）变量变换后构建的 Z 矩阵，可构建出R型协方差矩阵和Q型协方差矩阵，再对它们分别实施因子分析，最终就实现了对应分析。

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】某研究者收集了我国部分省份的农村居民家庭人均消费支出的数据。选取7个变量：A为食品支出比重，B为衣着支出比重，C为居住支出比重，D为家庭设备及服务支出比重，E为医疗保健支出比重，F为交通和通讯支出比重，G为文教娱乐、日用品及服务支出比重。考查的地区（即样品）有10个，资料见表1[9]。

【例2】疾病与人的基因型密切相关，而不同民族的人群中各种基因出现的频率不同。表2是某研究者收集的不同民族各种基因出现频率的相关资料[9]。试分析各种基因出现的频率与民族的关系。

表 1 某年中国 10 个省份农村居民家庭人均消费支出数据

Table 1 Per capita consumption expenditure data of rural households in 10 provinces of China in a certain year

地 区	A	B	C	D	E	F	G
1 山西	0.583 910	0.111 480	0.092 473	0.050 073	0.038 193	0.018 803	0.079 946
2 内蒙古	0.581 218	0.081 315	0.112 380	0.042 396	0.043 280	0.040 004	0.083 339
3 辽宁	0.565 036	0.100 121	0.123 970	0.041 121	0.043 429	0.031 328	0.078 919
4 吉林	0.530 918	0.105 360	0.116 952	0.045 064	0.043 735	0.038 508	0.095 256
5 黑龙江	0.555 201	0.096 500	0.143 498	0.037 566	0.052 111	0.026 267	0.072 829
6 海南	0.654 952	0.047 852	0.095 238	0.047 945	0.022 134	0.018 519	0.096 844
7 四川	0.640 012	0.061 680	0.116 677	0.048 471	0.033 529	0.017 439	0.072 043
8 贵州	0.725 239	0.056 362	0.073 262	0.044 388	0.016 366	0.015 720	0.057 261
9 甘肃	0.678 630	0.058 043	0.088 316	0.038 100	0.039 794	0.015 167	0.067 999
0 青海	0.665 913	0.088 508	0.096v899	0.038 191	0.039 275	0.019 243	0.033 801

表 2 不同民族的人群各基因出现的频率

Table 2 Frequency of various genes appearing in different ethnic groups

基因型	频 率				基因型	频 率			
	藏族	尼泊尔	印度	汉族		藏族	尼泊尔	印度	汉族
A1	0.030 8	0.018 0	0.119 0	0.014 9	B38	0.046 5	0.047 0	0.003 0	0.001 5
A2	0.333 3	0.107 0	0.148 0	0.349 2	B39	0.010 2	0	0.009 0	0.017 6
A3	0.020 4	0.019 0	0.101 0	0.017 6	B46	0.010 2	0.009 0	0	0.181 3
A9	0.303 7	0.279 0	0.156 0	0.141 4	B48	0.057 2	0.150 0	0.003 0	0.010 8
A10	0.040 9	0.018 0	0.039 0	0.031 3	B50	0.010 2	0.018 0	0.037 0	0
A11	0.135 4	0.422 0	0.126 0	0.297 7	B53	0.005 0	0	0.006 0	0
A28	0	0.018 0	0.083 0	0.009 4	B54	0.015 3	0	0	0.017 6
A30	0.041 3	0	0	0.021 7	B55	0.057 2	0.028 0	0.026 0	0.0217
A31	0.051 8	0.037 0	0.022 0	0.012 1	B56	0.010 2	0.009 0	0.006 0	0.004 0
A32	0	0.019 0	0.039 0	0.001 3	B57	0.005 0	0.018 0	0.039 0	0.034 1
A33	0	0.067 0	0.083 0	0.060 8	B58	0	0.067 0	0.033 0	0.013 9
B5	0.282 8	0.118 0	0.134 0	0.082 5	B60	0.062 6	0.028 0	0.0220	0.072 3
B7	0	0.019 0	0.080 0	0.024 4	B61	0.089 9	0	0.083 0	0.108 0
B8	0.0102	0.0118 0	0.045 0	0.009 4	B70	0.005 0	0	0.008 0	0
B12	0.0102	0.037 0	0.066 0	0.012 1	C1	0.089 9	0.037 0	0.023 0	0.171 6
B13	0.0102	0.077 0	0.006 0	0.065 0	C2	0.020 4	0	0.073 0	0.039 7
B14	0	0	0.006 0	0.001 3	C3	0.179 8	0.107 0	0.083 0	0.326 9
B15	0.192 3	0.254 0	0.096 0	0.109 2	C4	0.165 1	0.077 0	0.134 0	0.049 5
B18	0.005 0	0.028 0	0.022 0	0	C5	0	0.009 0	0.016 0	0.005 4
B27	0.1067	0	0.026 0	0.020 4	C6	0.025 6	0.245 0	0.045 0	0.008 1
B35	0.062 6	0.057 0	0.148 0	0.034 2	C7	0.171 2	0.218 0	0.119 0	0.1152
B37	0.010 2	0.018 0	0.009 0	0.006 7	C8	0.005 0	0	0.004 0	0.002 7

3.1.2 对数据结构的分析

例 1 中，地区可以被视为“样品”，A~G 代表 7 个变量，故这是一个单组设计 7 元定量资料。

例 2 中，4 个民族可以被视为“样品”，基因型可以被视为“变量”，故这是一个单组设计 44 元定量资料。

3.1.3 创建 SAS 数据集

分析例 1 资料，设所需 SAS 数据步程序如下：

```
data a1;
input regine $ 7. A B C D E F G;
cards;
```

（此处输入表 1 中的 10 行数据，包括第 1 列“地区”）

```
;
run;
```

分析例 2 资料，设所需 SAS 数据步程序如下：

```
data a2;
input gen $ Z N Y H @@;
cards;
```

（此处输入表2中的全部数据，包括“基因型”）
;
run;

3.2 用SAS实现统计分析

3.2.1 分析例1的资料

设所需要的SAS过程步程序如下[10]：

```
proc corresp data= a1 OUTC=aaa;
var A B C D E F G;
id regine;
run;
%plotit(data=aaa, datatype=corresp, tsize=0.5,
color='black', href=0, vref=0)
```

【SAS输出结果及解释】各地区在两个公因子上的负荷见表3。在以dim1为横轴、以dim2为纵轴的直角坐标系内，每个地区就是一个点，如“山西”点的坐标为(0.058，-0.028)，显然，该点在第四象限。这10个点中任何两点之间的欧氏距离都可以计算出来，并被标记在直角坐标系相应的位置上。

表3 各地区在两个公因子上的负荷

Table 3 Load of each region on two common factors

地 区	dim1	dim2	地 区	dim1	dim2
1山西	0.058	-0.028	6海南	-0.117	0.134
2内蒙古	0.091	0.042	7四川	-0.060	0.024
3辽宁	0.126	-0.010	8贵州	-0.228	-0.007
4吉林	0.184	0.047	9甘肃	-0.125	-0.008
5黑龙江	0.148	-0.041	0青海	-0.073	-0.153

与各地区对应的3个统计量的计算结果见表4。表4中，“质量”为每个地区上两个公因子贡献率之和，此值接近1，表明对应的地区信息由两个公因子就可很好地反映出来；“密度”为原始数据中各行数据之和占总合计的百分比；“惯量”为每个地区对总特征值0.026贡献的比例。

表4 与各地区对应的3个统计量的计算结果

Table 4 Calculation results of three statistics corresponding to each region

地 区	质 量	密 度	惯 量
1山西	0.224	0.099	0.070
2内蒙古	0.646	0.100	0.059
3辽宁	0.981	0.100	0.061
4吉林	0.944	0.099	0.143
5黑龙江	0.778	0.100	0.115
6海南	0.985	0.100	0.122
7四川	0.456	0.101	0.035
8贵州	0.949	0.101	0.208
9甘肃	0.780	0.100	0.076
0青海	0.980	0.100	0.110

每个公因子在每个地区上的贡献率见表5。各列数值之和为1。显然，“贵州”“吉林”和“黑龙江”对第一公因子贡献最大；“青海”和“海南”对第二公因子贡献最大。

表5 各公因子在各地区上的贡献率

Table 5 Contribution rates of common factors in different regions

地 区	dim1	dim2	地 区	dim1	dim2
1山西	0.019	0.016	6海南	0.079	0.372
2内蒙古	0.048	0.037	7四川	0.021	0.011
3辽宁	0.090	0.002	8贵州	0.301	0.001
4吉林	0.194	0.045	9甘肃	0.090	0.001
5黑龙江	0.126	0.035	0青海	0.031	0.479

各地区的坐标对特征值贡献最多的标志见表6。贡献少、中、多分别用0、1、2表示。

表6 各地区的坐标对特征值贡献最多的标志

Table 6 Indicators with the most contribution of coordinates to eigenvalues by region

地 区	dim1	dim2	最 佳	地 区	dim1	dim2	最 佳
1山西	0	0	1	6海南	0	2	2
2内蒙古	0	0	1	7四川	0	0	1
3辽宁	1	0	1	8贵州	1	0	1
4吉林	1	0	1	9甘肃	1	0	1
5黑龙江	1	0	1	0青海	0	2	2

每个地区对两个公因子各自的贡献率见表7。各行数值之和近似为1，因为只用了两个主要的公因子。由各行数值可看出，只有“辽宁”“吉林”“海南”“贵州”和“青海”这5个地区可以由这两个公因子较好地反映出来。

表7 每个地区对两个公因子各自的贡献率

Table 7 Contribution rates of each region to two common factors

地 区	dim1	dim2	地 区	dim1	dim2
1山西	0.182	0.042	6海南	0.425	0.560
2内蒙古	0.532	0.114	7四川	0.396	0.060
3辽宁	0.975	0.006	8贵州	0.948	0.001
4吉林	0.886	0.058	9甘肃	0.776	0.003
5黑龙江	0.721	0.057	0青海	0.183	0.798

同理，可以解释关于列变量（本例为“农村居民家庭人均消费支出项目”）的类似输出结果。因篇幅所限，此处从略，仅扼要说明如下：项目F、E、B对第一公因子贡献较大；项目G对第2公因子贡献最大。绘制出反映本资料中各“地区”与各“农村居民家庭人均消费支出”之间关联性的二维图形，见图1。由图1可看出，“辽宁”“山西”和“黑龙江”在项目“C（居住支出比重）”上比较接近，“甘肃”和“四川”在项目A（食品支出比重）和D（家庭设备及服务支出比重）上比较接近。

3. 2. 2　分析例 2 的资料

设所需要的SAS过程步程序如下[10]：

```
proc corresp data= a2 OUTC=aaa;
var Z N Y H;
id gen;
run;
%plotit(data=aaa, datatype=corresp, tsize=0. 5,
color='black', href=0, vref=0)
```

【SAS输出结果及解释】因SAS输出结果很多，为节省篇幅，下面仅给出反映“民族”与“基因型”之间关联性的图形，见图2。图2中，H(汉族)与Z(藏族)距离非常近，而它们与Y(印度)和N(尼泊尔)相距都很远。H(汉族)与Z(藏族)周围有如下基因型：A30、C1、C3、A2、B60、B55、B5、B27。也就是说，H(汉族)与Z(藏族)前述8种基因型出现的频率比较接近。从图2显示的坐标点的位置直观给出结论，其精确度较低。若基于输出坐标点的坐标，按欧式距离公式计算出来，就更精确了。

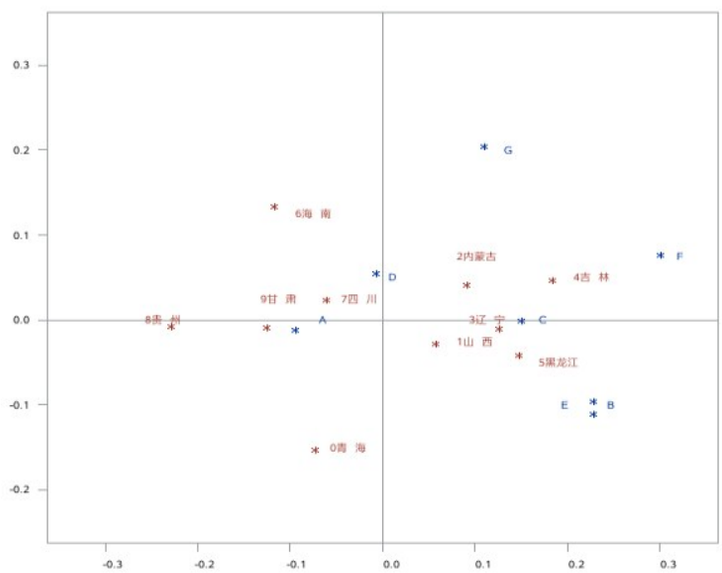

图1　“地区”与“支出”之间的关联性

Figure 1　Association between "region" and "expenditure"

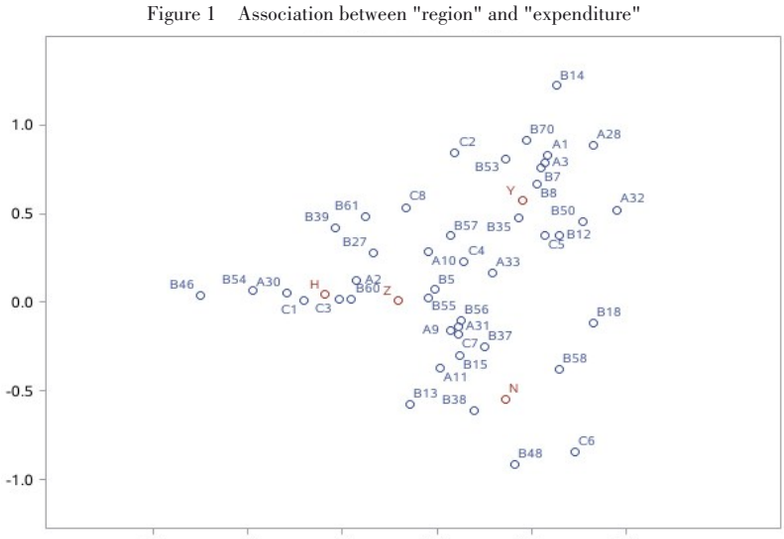

图2　反映“民族”与“基因型”之间关联性的图形

Figure 2　Showing the association between "ethnicity" and "genotype"

4　讨论与小结

4. 1　讨论

对应分析生成的二维图上的各状态点，实际上是两个多维空间上的点的投影，在某些特殊情况下，在多维空间中相距较远的点，在二维平面上的投影却很接近。此时，需借助公因子的贡献大小等信息对二维图上的各点进行深入的了解。另外，对应分析只能用图形的方式呈现变量与样品之间的关联关系，不能给出具体的统计量来度量这种关联程度的高低[8]。

4.2 小结

本文介绍了与对应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。基本概念包括变量与样品、显变量与隐变量、因子分析、R型分析与Q型分析、对应分析；计算方法涉及基本原理、变量变换、构建R型和Q型协方差矩阵、因子分析；两个实例分别为“某年中国10个省份农村居民家庭人均消费支出数据”和“不同民族的各种基因出现的频率”；借助SAS软件，对两个实例中的数据进行了对应分析，并对SAS输出结果做出了解释。

参考文献

[1] 张岩波. 潜变量分析[M]. 北京：高等教育出版社，2009：1-13.
Zhang YB. Latent variables analysis [M]. Beijing: Higher Education Press, 2009: 1-13.

[2] Armitage P, Colton T. Encyclopedia of biostatistics [M]. 2nd edition. New York: John Wiley & Sons, Inc, 2005: 4300-4304.

[3] 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.

[4] 王静龙. 多元统计分析[M]. 北京：科学出版社，2008：360-375.
Wang JL. Multivariate statistical analysis[M]. Beijing: Science Press, 2008: 360-375.

[5] 李卫东. 应用多元统计分析[M]. 北京：北京大学出版社，2008：239-258.
Li WD. Applied multivariate statistical analysis [M]. Beijing: Peking University Press, 2008: 239-258.

[6] 高惠璇. 应用多元统计分析[M]. 北京：北京大学出版社，2005：324-342.
Gao HX. Applied multivariate statistical analysis [M]. Beijing: Peking University Press, 2005: 324-342.

[7] 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，2005：232-250.
Yu JH, Yang WQ. Multivariate statistical analysis and application [M]. Guangzhou: Sun Yat-sen University Press, 2005: 232-250.

[8] 何晓群. 多元统计分析[M]. 2版. 北京：中国人民大学出版社，2008：227-254.
He XQ. Multivariate statistical analysis[M]. 2nd edition. Beijing: China Renmin University Press, 2008: 227-254.

[9] 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，2012：275-285.
Hu LP. Problem-oriented statistics: (3) experimental design and multivariate statistical analysis [M]. Beijing: People's Mental Publishing House, 2012: 275-285.

[10] SAS Institute Inc. SAS/STAT®15.1 user's guide[M]. Cary, NC: SAS Institute Inc, 2018: 2514-2578.

（收稿日期：2023-07-26）
（本文编辑：陈 霞）