总26期 第4期　　武汉测绘科技大学学报　　Sum26 №4
1986年12月　Journal of Wuhan Technical University of Surveying and Mapping　Dec. 1986

对应分析的原理和数据预处理

张克权　　郭仁忠

摘　要

本文对多元统计分析中目前应用的对应分析方法的原理公式作了重新推导，提出 $z_{ij}=(p_{ij}/p_{i.}p_{.j})-1$ 的数据变换公式；并对原始数据作了严格的预处理。最后应用一个实例计算和作图，与简单的因子分析及原用的对应分析成果作了比较。理论和算例都说明本文介绍的对应分析的数据预处理和分析方法较原用方法原理上更正确，应用上更合理，并克服了变量中对各种数据原有的局限性。

【关键词】 新对应分析方法；数据预处理；Q型因子分析；R型因子分析；极差正规化

一、对应分析简介

在很多学科中，所研究的变量往往复杂多样，当我们进行综合分析时，需要把它归结为影响这些变量的共同因子和特殊因子及其组合，然后通过数学方法把原来数目较多的、关系复杂的变量，采用转换运算找到数目较少的、相互独立的新的因子（原始变量的组合），使研究的任务大大简化，这种多元统计分析方法称为因子分析。根据所研究的目的不同，因子分析可分为研究各变量之间关系的R型因子分析和研究样品之间关系的Q型因子分析。

一九七〇年法国统计学者 Bezecri 教授首先提出了对应分析的原理和方法，随后该方法在地质、地理等不少学科得到应用并受到欢迎。事实上对应分析是在R型因子分析和Q型因子分析的基础上发展起来的。众所周知，一般地说进行R型因子分析比较容易，因为R型因子分析只要分析计算一个〔m×m〕的方阵（m为变量数目），m的数值通常是不大的，几个到几十个。但是Q型分析要分析计算一个〔n×n〕的方阵（n为样品数目），一般n较大，几十个到几百个甚至更大，随之数据量及计算量就相当大，这在进行Q型因子分析时很困难，有时甚至不可能。另外R型因子分析和Q型因子分析互相独立，无法综合分析样品和变量间的关系，从而损失不少信息。对应分析就是为了解决以上问题而提出的，其基本原理

本文1986年3月收到.

如下。

设原始数据矩阵为：$X=[x_{ij}]$ $(i=1,2,\cdots,m;\quad j=1,2,\cdots,n)$

进行如下数据变换：

$$\sum_{i=1}^{m} x_{ij}=x_{.j},\quad \sum_{j=1}^{n} x_{ij}=x_{i.},\quad \sum_{j=1}^{n} x_{.j}=\sum_{i=1}^{m} x_{i.}=T,$$

$$p_{.j}=x_{.j}/T,\quad p_{i.}=x_{i.}/T,\quad p_{ij}=x_{ij}/T。$$

得变换后的数据矩阵为：

$$P=[p_{ij}]=[x_{ij}/T]。$$

P是一个概率矩阵。在R型分析的情况下，n个样品可以表示为m维欧氏空间的n个向量，考虑到变量尺度和样品数量级（变幅）的差异，用以下坐标表示第j个样品在m维空间的向量：

$$\left(p_{1j}/\sqrt{p_{1.}\cdot p_{.j}},\ p_{2j}/\sqrt{p_{2.}\cdot p_{.j}},\ \cdots,\ p_{mj}/\sqrt{p_{m.}\cdot p_{.j}}\right)=\vec{p}(j)\quad (j=1,2,\cdots,n)$$

m个变量可以表示为n维欧氏空间的m个向量，类似地第i个变量在n维空间的坐标为：

$$\left(p_{i1}/p_{i.}\cdot\sqrt{p_{.1}},\ p_{i2}/p_{i.}\cdot\sqrt{p_{.2}},\cdots,p_{in}/p_{i.}\cdot\sqrt{p_{.n}}\right)=\vec{q}(i)\quad (i=1,2,\cdots,m)$$

利用概率方法根据样品坐标求出变量的均值为：

$$\overline{q}(i)=\sum_{j=1}^{n}\left(p_{ij}/\sqrt{p_{i.}\cdot p_{.j}}\right)\cdot p_{.j}=\sqrt{p_{i.}}$$

类似地，样品的均值为：

$$\overline{p}(j)=\sum_{i=1}^{m}\left(p_{ij}/p_{i.}\sqrt{p_{.j}}\right)\cdot p_{i.}=\sqrt{p_{.j}}$$

对于R型因子分析，同样利用样品点的坐标计算变量间的方差与协方差矩阵，得：

$$a_{ik}=\sum_{\alpha=1}^{n}\left(\frac{p_{i\alpha}}{\sqrt{p_{i.}\cdot p_{.\alpha}}}-\sqrt{p_{i.}}\right)\left(\frac{p_{k\alpha}}{\sqrt{p_{k.}\cdot p_{.\alpha}}}-\sqrt{p_{k.}}\right)\cdot p_{.\alpha}$$

$$=\sum_{\alpha=1}^{n}\left(\frac{p_{i\alpha}-p_{i.}\cdot p_{.\alpha}}{\sqrt{p_{i.}\cdot p_{.\alpha}}}\right)\left(\frac{p_{k\alpha}-p_{k.}\cdot p_{.\alpha}}{\sqrt{p_{k.}\cdot p_{.\alpha}}}\right)=\sum_{\alpha=1}^{n} z_{i\alpha}z_{k\alpha}$$

$$A=[a_{ik}]=\left[\sum_{\alpha=1}^{n} z_{i\alpha}z_{k\alpha}\right]=ZZ'\quad (i,k=1,2,\cdots,m)$$

对于Q型因子分析，利用变量点的坐标计算样品间的协方差矩阵，得：

$$b_{jl}=\sum_{\beta=1}^{m}\left(\frac{p_{\beta j}}{p_{\beta.}\cdot\sqrt{p_{.j}}}-\sqrt{p_{.j}}\right)\left(\frac{p_{\beta l}}{p_{\beta.}\cdot\sqrt{p_{.l}}}-\sqrt{p_{.l}}\right)\cdot p_{\beta.}$$

$$=\sum_{\beta=1}^{m}\left(\frac{p_{\beta j}-p_{\beta.}\cdot p_{.j}}{\sqrt{p_{\beta.}\cdot p_{.j}}}\right)\left(\frac{p_{\beta l}-p_{\beta.}\cdot p_{.l}}{\sqrt{p_{\beta.}\cdot p_{.l}}}\right)=\sum_{\beta=1}^{m} z_{\beta j}z_{\beta l}$$

$$B=[b_{jl}]=\left[\sum_{\beta=1}^{m} z_{\beta j}z_{\beta l}\right]=Z'Z\qquad (j,l=1,2,\cdots,n)$$

根据线性代数理论，A·B 的非零特征值对应相等，特征向量具有对应性。设 A 的特征值为 $\lambda_i > 0$ ($i=1, 2, \cdots, r$; $r \leqslant m$)，相应的特征向量为 u_i，则 λ_i 也是B的非零特征值，$v_i = Z'u_i$ 是 B 的相应于 λ_i 的特征向量，所以从 A 进行 R 型因子分析可以推得 Q 型因子分析的结果，同时由于特征值对应相等，可以将 R 型因子分析和 Q 型因子分析的结果置于同一因子空间进行联合分析，这就是对应分析的优越性。

从以上介绍，我们可以发现现有的对应分析尚有以下几个方面值得进一步探讨：

1．现有方法是根据 m 维空间的样品点坐标来计算变量的均值和协方差，这在逻辑上有一定的矛盾性，因为把样品表示为 m 维空间的点就是基于变量正交这一假设之上的，没有这一假设，样品就不能表示为 m 维欧氏空间的向量。而计算变量的协方差就意味着变量并非正交而是相关的。在进行样品协方差的计算中也存在类似的逻辑上的矛盾性，虽然这样计算总是能得到结果的，但概念却是前后不一致的。

2．从 R 型分析推求 Q 型分析结果时，求得 $v_i = Z'u_i$ 后即对 v_i 进行单位化。然而 v_i 为 n维向量，u_i 为 m 维向量，一般地 $n \gg m$。因为 v_i、u_i 都要单位化，就必然有 v_i 的元素的绝对值远小于 u_i 的元素的绝对值，当进行Q、R 型联合分析时，样品点总是趋于坐标原点，变量点则正常分布，样品点与变量趋于分离，联合分析就比较困难，也难以揭示样品与变量的关系。

3．原始变量往往复杂多样，不仅量纲不一，数据变化的幅度也很悬殊，且经常既有正值又有负值，这样就使上述对应分析中的 $p_{i\cdot}$、$p_{\cdot j}$ 和 T 参与计算很不合理，甚至是不可能的。

下面提出对应分析的一个新公式推导并进行必要的数据预处理，该方法对解决以上问题具有较好作用。

二、对应分析方法的重新推导

类似于现有方法，设原始数据为：

$$X=[x_{ij}] \quad (i=1, 2, \cdots, m;\ j=1, 2, \cdots, n)$$

按原有方法的数据变换公式计算得 $x_{i\cdot}$、$x_{\cdot j}$、T 以及 p_{ij}、$p_{i\cdot}$ 和 $p_{\cdot j}$。考虑到我们主要关心的是同一样品内部的各变量的相对比例，以 $p_{\cdot j}$ 除 p_{ij} 得 $p_{ij}/p_{\cdot j}$，又考虑到变量尺度的差异，需进行指标处理，以 $p_{i\cdot}$ 除 $p_{ij}/p_{\cdot j}$ 得 $p_{ij}/p_{i\cdot}\cdot p_{\cdot j}$。令 $w_{ij}=p_{ij}/p_{i\cdot}\cdot p_{\cdot j}$，得：

$$W=[w_{ij}]=\left[\frac{p_{ij}}{p_{i\cdot}\cdot p_{\cdot j}}\right]=\left[\frac{x_{ij}/T}{(x_{i\cdot}/T)(x_{\cdot j}/T)}\right] \quad (i=1,2,\cdots,m;\ j=1,2,\cdots,n) \tag{1}$$

以上的数据变换对变量和样品是完全对等的。对于任一变量 i 我们将其表示为 n 维空间的一个向量：

$$[p_{i1}/p_{i\cdot}\cdot p_{\cdot 1},\ p_{i2}/p_{i\cdot}\cdot p_{\cdot 2},\ \cdots,\ p_{in}/p_{i\cdot}\cdot p_{\cdot n}]$$

任一样品 j 可以表示为 m 维空间的一个向量：

$$[p_{1j}/p_{1\cdot}\cdot p_{\cdot j},\ p_{2j}/p_{2\cdot}\cdot p_{\cdot j},\ \cdots,\ p_{mj}/p_{m\cdot}\cdot p_{\cdot j}]$$

任一变量（变量空间点坐标）的均值的计算仿照原有方法按概率 $p_{\cdot j}$ 计算得：

$$\sum_{i=1}^{n}\left(p_{ij}/p_{i\cdot}\cdot p_{\cdot j}\right)\cdot p_{\cdot j}=\sum_{j=1}^{n}p_{ij}/p_{i\cdot}=1 \tag{2}$$

任一样品（样品空间点坐标）的均值类似地有：

$$\sum_{i=1}^{m}\left(p_{ij}/p_{i\cdot}\cdot p_{\cdot j}\right)\cdot p_{i\cdot}=\sum_{i=1}^{m}p_{ij}/p_{\cdot j}=1 \tag{3}$$

以上我们从样品点的坐标计算样品的均值，从变量点的坐标计算变量的均值，这在逻辑上是合理的。由于数据变换的结果，样品点和变量点的空间坐标的对应值完全相等，这就使我们能较容易地解决前述第 1 个问题。对于 R 型因子分析，变量协方差矩阵元素计算如下：

$$\begin{aligned}a_{ik}&=\frac{1}{n}\sum_{\alpha=1}^{n}\left[(p_{i\alpha}/p_{i\cdot}\cdot p_{\cdot\alpha})-1\right]\left[(p_{k\alpha}/p_{k\cdot}\cdot p_{\cdot\alpha})-1\right]\\&=\frac{1}{n}\sum_{\alpha=1}^{n}(w_{i\alpha}-1)(w_{k\alpha}-1)=\frac{1}{n}\sum_{\alpha=1}^{n}z_{i\alpha}z_{k\alpha}\end{aligned} \tag{4}$$

类似地对于 Q 型因子分析，样品协方差矩阵的元素计算如下：

$$\begin{aligned}b_{jl}&=\frac{1}{m}\sum_{\beta=1}^{m}\left[(p_{\beta j}/p_{\beta\cdot}\cdot p_{\cdot j})-1\right]\left[(p_{\beta l}/p_{\beta\cdot}\cdot p_{\cdot l})-1\right]\\&=\frac{1}{m}\sum_{\beta=1}^{m}(w_{\beta j}-1)(w_{\beta l}-1)=\frac{1}{m}\sum_{\beta=1}^{m}z_{\beta j}z_{\beta l}\end{aligned} \tag{5}$$

从而得到：

$$Z=[z_{ij}]=[w_{ij}-1]=[(p_{ij}/p_{i\cdot}\cdot p_{\cdot j})-1]\qquad(i=1,2,\cdots,m;\ j=1,2,\cdots,n) \tag{6}$$

$$\begin{cases}A=[a_{ik}]=\dfrac{1}{n}\left[\sum_{\alpha=1}^{n}z_{i\alpha}z_{k\alpha}\right]=\dfrac{1}{n}ZZ' & (i,k=1,2,\cdots,m)\\[2ex] B=[b_{jl}]=\dfrac{1}{m}\left[\sum_{\beta=1}^{m}z_{\beta j}z_{\beta l}\right]=\dfrac{1}{m}Z'Z & (j,l=1,2,\cdots,n)\end{cases} \tag{7}$$

设 ZZ' 的特征值为 $\lambda_1,\ \lambda_2,\ \cdots,\ \lambda_r>0\ (r\leqslant m)$，相应的特征向量为 $u_1,\ u_2,\ \cdots,\ u_r$，则我们有

$$ZZ'\lambda_i=u_i\lambda_i\qquad(i=1,2,\cdots,r)$$

设：

$$U=[u_1,\ u_2,\ \cdots,\ u_r],\ \lambda=\begin{bmatrix}\lambda_1 & & & 0\\ & \lambda_2 & & \\ & & \ddots & \\ 0 & & & \end{bmatrix}$$

则：

$$ZZ'U=U\lambda$$

两边同乘 $\frac{1}{n}$ 有：

$$\frac{1}{n}ZZ'U=AU=\frac{1}{n}U\lambda=U\left(\frac{1}{n}\lambda\right)$$

可见 ZZ' 的特征向量与 A 的完全相同，特征值差一个常数$\frac{1}{n}$。类似的推导可知 $Z'Z$和 B 的特征向量相同而特征值相差一个常数$\frac{1}{m}$。所以 A 和 B 的特征向量的对应性不变，特征值相差一个常数$\frac{n}{m}$，亦即 B 的特征值是 A 的相应特征值的 n/m 倍。由此可见从 R 型因子分析不难推导 Q 型因子分析的结果。另外，由于 A 与 B 的非零特征值不再完全对应相应，但统一相差一个常数 n/m，特征值之间的相对关系没有变，这就保证了进行 R、Q型因子分析结果联合解释的可能性，所以对应分析的两大优越性在这里都得以保持。由于 B 的特征值是 A 的相应特征值的 n/m 倍，这就保证了无论样品数 n 和变量数 m 如何变化，由于特征向量单位化造成的 Q 型因子载荷和 R 型因子载荷的差异都在此得到补偿，保证了因子载荷的稳定性，这就克服了前述的第 2 个问题。

三、数据预处理的合理方法

原有的对应分析中，对所有原始数据除以总和值 T 予以变换处理；并在设定和推导变量或样品的点位坐标、均值和协方差时，都以 $p_{i.}$ 或 $p_{.j}$ 值予以相对比例的处理。这样用原始数据进行对应分析有两点是不合理的，甚至不可能的：

①由于各变量原始数据的量纲不一，变幅差异很大，各样品的 $p_{.j}$ 值和数据总和值 T，从原理上讲并无实际意义，不能起到正确分析对应的作用；

②原始数据中往往出现有负值，则 T 和 $p_{.j}$ 值都是不正确的，无法进行对应分析计算。

因此，在进行对应分析计算前，必须对原始数据进行规格化预处理，同时还可以达到适当简化计算的目的。根据对应分析原理和计算的特点，提出下列三点规格化的要求：

①规格化后数据 x'_{ij} 的 $x'_{.j}$ 和 $x'_{i.}$ 均不能为零，否则计算公式中某些值为无穷大，无法进行对应分析计算；

②应使$x'_{ij} \geqslant 0$，否则x'_{ij}值无意义；

③为使 $p_{i.}$、$p_{.j}$ 和 T 值的量纲和变幅更为合理，要求规格化后的 $x'_{i.}$为一常数，为计算方便可令 $x'_{i.} = 1$。

另外，根据原有对应分析方法算得的协方差矩阵元素绝对值都很小，计算过程中误差较大，解的精度低，通过数据预处理，可以使协方差矩阵元素绝对值增大，有利于增加其后较复杂数值运算的稳定性，提高解的精度。

根据这些要求，规格化分两步进行：

1．变量量纲和负值的规格化处理。为了满足上述第①、②条要求，不能采用通常的规格化方法，即$x''_{ij} = \frac{x_{ij} - \bar{x}_i}{S_i}$ $(i = 1, 2, \cdots, m;\ j = 1, 2, \cdots, n;\ \bar{x}_i$为第 i 个变量各样品的均值；$S_i$为变量的标准差)，也不能采用极差规格化，即$x''_{ij} = \frac{x_{ij} - \bar{x}_i}{\max\limits_{1 \leqslant j \leqslant n} x_{ij} - \min\limits_{1 \leqslant j \leqslant n} x_{ij}}$。

因为，原始数据经过这样规格化后，前者的 $x''_{i.}$ 为零而后者的 x''_{ij} 会出现负值，均不能进行对应分析计算。为此，应该用极差正规化处理，即：

$$x''_{ij}=\frac{x_{ij}-\min\limits_{1\leqslant j\leqslant n}x_{ij}}{\max\limits_{1\leqslant j\leqslant n}x_{ij}-\min\limits_{1\leqslant j\leqslant n}x_{ij}} \tag{8}$$

这样规格化后的数据为 $0\leqslant x''_{ij}\leqslant 1$。

2．数据变幅的一致性处理。为了满足第三点要求，将上面经极差正规化后的数据均除以 $x''_{i.}$，即：

$$x'_{ij}=\frac{x''_{ij}}{x''_{i.}} \tag{9}$$

这样就使各变量的 $x'_{i.}$ 均为 1（也可化为其它同一常数）。

经过上述两步数据预处理之后，$x'_{i.}=1$ 和 $T=m$，则公式（1）、（4）、（5）、（6）可简化为：

$$W=[w_{ij}]=\left[\frac{x'_{ij}/m}{\left(\frac{1}{m}\right)(x'_{.j}/m)}\right]=\left[\frac{x'_{ij}}{x'_{.j}}\right]\cdot m \tag{10}$$

$$Z=[z_{ij}]=\left[\frac{x'_{ij}}{x'_{.j}}\cdot m-1\right] \tag{11}$$

$$a_{ik}=\frac{1}{n}\sum_{\alpha=1}^{n}\left[\frac{x'_{i\alpha}}{x'_{.\alpha}}\cdot m-1\right]\left[\frac{x'_{k\alpha}}{x'_{.\alpha}}\cdot m-1\right]=\frac{1}{n}\sum_{\alpha=1}^{n}z_{i\alpha}z_{k\alpha} \tag{12}$$

$$b_{jl}=\frac{1}{m}\sum_{\beta=1}^{m}\left[\frac{x'_{\beta j}}{x'_{.j}}\cdot m-1\right]\left[\frac{x'_{\beta l}}{x'_{.l}}\cdot m-1\right]=\frac{1}{m}\sum_{\beta=1}^{m}z_{\beta j}z_{\beta l} \tag{13}$$

四、计算步骤和算例解释

将上述推导的公式和数据预处理的方法，结合一个实例来介绍计算步骤；并根据计算成果画出的二维平面投影图与简单的 R 型因子分析、Q 型因子分析及原有对应分析的投影图比较，并作简要的解释。

1、计 算 步 骤

设 $n=17$ 个样品（区域），每个样品有 $m=7$ 个指标（为农作物生长期日数、三个日照指标和三个积温指标），来分析某些气候指标对各区域农业影响的情况。其原始数据列于表 1。

（1）将原始数据规格化（极差正规化、每一变量各样品之和处理为 $x'_{i.}=1$），并按行、列求和及总和值，见表 2。

表 1　原 始 数 据 表

j \ i	1	2	3	4	5	6	7
1	230.90	323.50	500.00	479.50	997.00	1629.00	1563.00
2	232.00	300.70	445.80	423.10	1006.00	1650.00	1587.00
3	237.00	309.00	464.50	471.00	1015.00	1650.00	1608.00
4	235.00	324.00	476.00	466.00	990.00	1650.00	1590.00
5	230.00	328.00	456.00	456.00	1001.00	1638.00	1566.00
6	236.30	295.60	429.80	452.60	1026.00	1647.00	1614.00
7	234.30	302.90	457.10	471.60	1013.00	1641.00	1596.00
8	239.00	300.30	461.60	436.20	1007.00	1638.00	1599.00
9	236.20	300.40	433.70	452.70	1007.00	1623.00	1608.00
10	231.00	343.50	494.40	473.10	990.00	1623.00	1545.00
11	231.80	318.80	462.00	453.40	995.00	1638.00	1569.00
12	230.30	302.20	465.70	448.90	974.00	1626.00	1527.00
13	235.20	279.00	442.20	418.70	1017.00	1644.00	1581.00
14	235.00	326.50	473.10	470.10	1015.00	1650.00	1599.00
15	234.00	293.10	460.30	462.10	1009.00	1650.00	1587.00
16	234.00	277.20	425.00	398.90	1021.00	1653.00	1578.00
17	237.00	286.80	437.20	430.20	1021.00	1623.00	1572.00

（2）根据表 2 按公式（11）计算矩阵 Z 后得协方差矩阵 ZZ′。

$$ZZ' = \begin{pmatrix} 6.97978 & -5.34131 & -5.28607 & -2.85390 & 3.61594 & -0.01125 & 2.89681 \\ -5.34131 & 6.74216 & 5.82537 & 3.66321 & -4.22370 & -3.67057 & -2.99516 \\ -5.28607 & 5.82537 & 8.17098 & 4.18449 & -4.95029 & -3.76180 & -4.18268 \\ -2.85390 & 3.66321 & 4.18449 & 3.40738 & -2.97997 & -3.40524 & -2.01598 \\ 3.61594 & -4.22370 & -4.95029 & -2.97997 & 4.23392 & 1.98789 & 2.31621 \\ -0.01125 & -3.67057 & -3.76180 & -3.40524 & 1.98789 & 7.38193 & 1.47904 \\ 2.89681 & -2.99516 & -4.18268 & -2.01598 & 2.31621 & 1.47904 & 2.50174 \end{pmatrix}$$

（3）计算 ZZ′ 的特征值和特征向量，$\lambda_1 \geqslant \lambda_2 \geqslant \cdots \geqslant \lambda_m \geqslant 0$，得

表 2　规格化数据表

j \ i	1	2	3	4	5	6	7	$x'_{.j}$
1	0.013043	0.092767	0.134072	0.091300	0.042124	0.021277	0.038710	0.433293
2	0.028986	0.047085	0.037183	0.027413	0.058608	0.095745	0.064516	0.359536
3	0.101449	0.063715	0.070611	0.081672	0.075092	0.095745	0.087097	0.575381
4	0.072464	0.093769	0.091169	0.076008	0.029304	0.095745	0.067742	0.526201
5	0.000000	0.101783	0.055417	0.064681	0.049450	0.053191	0.041935	0.366457
6	0.091304	0.036866	0.008581	0.060829	0.095238	0.085106	0.093548	0.471472
7	0.062319	0.051493	0.057383	0.082352	0.071429	0.063830	0.074194	0.463000
8	0.130435	0.046283	0.065427	0.042252	0.060439	0.053191	0.077419	0.475446
9	0.089855	0.046484	0.015552	0.060942	0.060439	0	0.087097	0.360369
10	0.014493	0.132839	0.124062	0.084051	0.029304	0	0.019355	0.404104
11	0.026087	0.083350	0.066142	0.061735	0.038461	0.053191	0.045161	0.374127
12	0.004348	0.050090	0.072757	0.056638	0	0.010638	0	0.194471
13	0.075362	0.003607	0.030747	0.022429	0.078755	0.074468	0.058065	0.343433
14	0.072464	0.098778	0.085985	0.080652	0.075092	0.095745	0.077419	0.586135
15	0.057971	0.031857	0.063103	0.071590	0.064103	0.095745	0.064516	0.448885
16	0.057971	0	0	0	0.086081	0.106383	0.054839	0.305274
17	0.101449	0.019235	0.021809	0.035455	0.086081	0	0.048387	0.312416
$x'_{i.}$	1.000000	1.000001	1.000000	0.999999	1.000000	1.000000	1.000000	T = 7.000000

$$\lambda = \begin{pmatrix} 27.47326 & & & & & & \\ & 7.52360 & & & 0 & & \\ & & 2.12145 & & & & \\ & & & 1.21081 & & & \\ & 0 & & & 0.81056 & & \\ & & & & & 0.27835 & \\ & & & & & & 0 \end{pmatrix}$$

$$U=\begin{pmatrix}
-0.39411 & +0.54464 & -0.36574 & +0.23798 & -0.35929 & +0.29274 & -0.37797\\
+0.46038 & -0.02297 & 0.56532 & 0.01806 & -0.52487 & 0.22195 & -0.37797\\
+0.51554 & -0.02173 & -0.59127 & -0.23576 & -0.09717 & -0.41984 & -0.37796\\
+0.31168 & +0.16000 & 0.02702 & 0.24082 & 0.73314 & 0.37173 & -0.37797\\
-0.34536 & +0.07784 & 0.20689 & -0.81049 & 0.17571 & 0.03525 & -0.37796\\
-0.28666 & -0.81529 & -0.18669 & 0.17065 & -0.05031 & 0.20920 & -0.37797\\
-0.26146 & 0.07752 & 0.34447 & 0.37847 & 0.12281 & -0.71104 & -0.37796
\end{pmatrix}$$

（4）计算特征值累积百分比，按 $\sum_{a=1}^{k}\lambda_a \Big/ \sum_{a=1}^{m}\lambda_a \geqslant 85\%$ 取出前 k 个特征值 $\lambda_1,\lambda_2,\cdots,\lambda_k$。

表 3　特征值累积百分比表

序　号	特 征 值 λ	累　积　值	累 积 百 分 比
1	27.47326	27.47326	69.70
2	7.52360	34.99686	88.78
3	2.12145	37.11831	94.17
4	1.21081	38.32912	97.24
5	0.81056	39.13968	99.29
6	0.27835	39.41803	100.00
7	0	39.41803	100.00

从上表知，取前两个特征值 $\lambda_1\cdot\lambda_2$ 所代表的方差已占总方差的 88.78%。因此，选用前两个主因子已足够精确代表整个数据的变化。

（5）R 型因子载荷矩阵的计算。取前两个特征值 $\lambda_1=27.47326$ 和 $\lambda_2=7.52360$ 的 $\frac{1}{n}$ 及其相应的单位特征向量 u_1 和 u_2，按下式计算因子载荷：

$$F=\begin{pmatrix} u_{11}\sqrt{\frac{\lambda_1}{n}} & u_{12}\sqrt{\frac{\lambda_2}{n}} \\ u_{21}\sqrt{\frac{\lambda_1}{n}} & u_{22}\sqrt{\frac{\lambda_2}{n}} \\ u_{31}\sqrt{\frac{\lambda_1}{n}} & u_{32}\sqrt{\frac{\lambda_2}{n}} \\ u_{41}\sqrt{\frac{\lambda_1}{n}} & u_{42}\sqrt{\frac{\lambda_2}{n}} \\ u_{61}\sqrt{\frac{\lambda_1}{n}} & u_{62}\sqrt{\frac{\lambda_2}{n}} \\ u_{61}\sqrt{\frac{\lambda_1}{n}} & u_{62}\sqrt{\frac{\lambda_2}{n}} \\ u_{71}\sqrt{\frac{\lambda_1}{n}} & u_{72}\sqrt{\frac{\lambda_2}{n}} \end{pmatrix}$$

得因子载荷矩阵如表 4：

表 4

序　号	变　量	F_1	F_2
1	Ⅰ	−0.50101	0.36232
2	Ⅱ	+0.58525	−0.01528
3	Ⅲ	0.65538	−0.01446
4	Ⅳ	0.39623	0.10644
5	Ⅴ	−0.43904	0.05178
6	Ⅵ	−0.36442	−0.54238
7	Ⅶ	−0.33239	0.05157
	方差贡献	1.61607	0.44256
	累积方差贡献%	69.70	88.78

（6）Q 型因子载荷矩阵的计算。取前两个特征值 $\lambda_1 = 27.47326$ 和 $\lambda_2 = 7.52360$的$\frac{1}{m}$，并计算对应于矩阵 $Z'Z$ 的单位特征向量 $Z'u_1 = v_1$、$Z'u_2 = v_2$，从而按 $G=(g_{jk})=\left(v_{jk}\sqrt{\frac{\lambda_k}{m}}\right)$ 得 Q 型前两个主因子的载荷表（表 5）。

（7）作因子平面投影图。根据表 4 和表 5，可作出各样品及各变量在因子平面 G_1—G_2

表 5

样品序号	G_1	G_2
1	0.63733	0.03419
2	−0.19567	−0.36951
3	−0.11470	−0.00065
4	0.15096	−0.11582
5	0.37757	−0.21270
6	−0.43421	0.02080
7	−0.04554	0.02262
8	−0.22016	0.23736
9	−0.22500	0.50467
10	0.85388	0.12685
11	0.29060	−0.11413
12	0.99940	0.00034
13	−0.53100	−0.04776
14	0.05760	−0.08116
15	−0.11663	−0.15999
16	−0.84423	−0.38315
17	−0.43371	0.59672
方差贡献	3.92475	1.07480
累积方差贡献%	69.70	88.78

和 F_1—F_2 上的投影图（见图 1，图中△表示变量点，·表示样品点）。

2、算例简要解释

为了与简单的 R 型、Q 型因子分析及原有对应分析计算成果作比较，在此引入根据同样实例数据计算的这些成果的平面投影图（见图 2 、 3 、 4）。

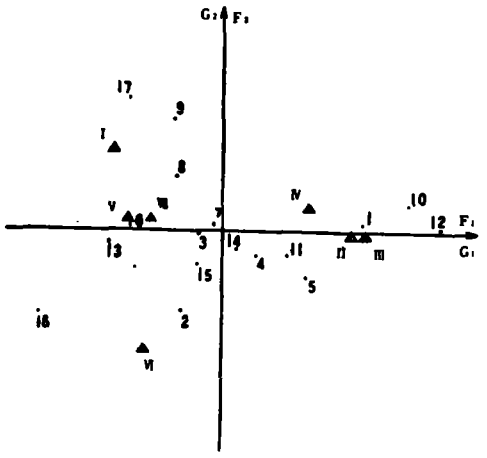

图 1　对应分析因子一和因子二载荷平面聚点图（根据本文所提方法）

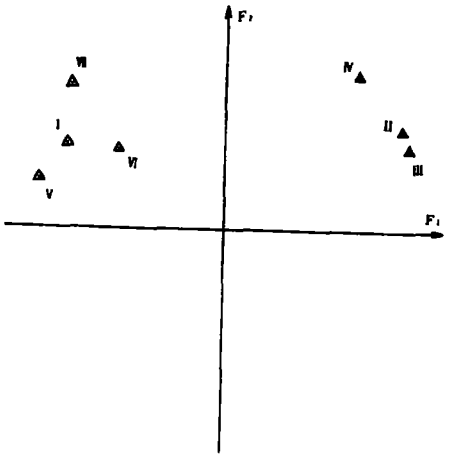

图 2　R 型分析因子一和因子二载荷平面聚点图

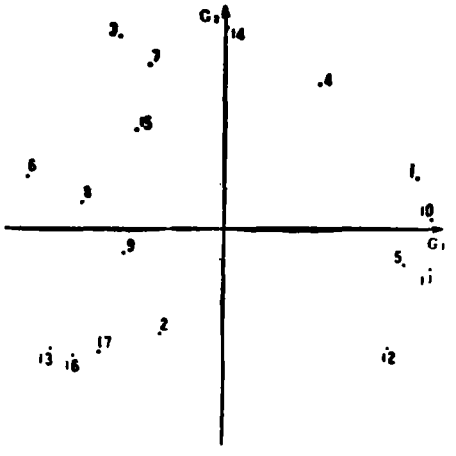

图 3　Q 型分析因子一和因子二载荷平面聚点图

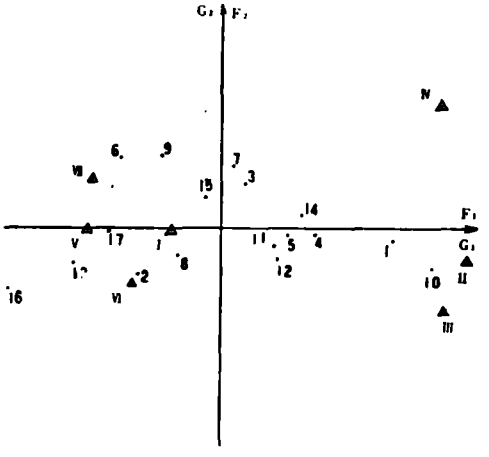

图 4　对应分析因子一和因子二载荷平面聚点图（根据原有对应分析方法）

从这些投影图及原始数据来分析，可得到如下简要的结论：

1．由简单的因子分析计算成果知，影响该地区的气候条件（部分指标），可把变量组合成 2～3 个主因子。第一主因子主要由变量指标Ⅱ、Ⅲ、Ⅳ（日照指标）组成，第二主因子主要由变量指标Ⅰ、Ⅴ、Ⅶ、Ⅵ（生长期日数和积温）组成；其中，变量指标Ⅵ（夏季积温）可以是第三个因子的主要组成部分。在这些平面投影图上都得到了反映，但在图 1 上反映最为明显。

2．样品分布的总规律是类似的，但在简单的因子分析图上，由于 R 型和 Q 型因子分析相互独立，单独作图，对变量和样品间的连系，不能综合地得到充分体现。在原有对应分析图上（图 4），反映了相互的对应关系，但样品点的位置受到向量单位化的影响；只有在图 1 上才能得到明显而正确的反映。

3。在原有的对应分析方法的平面投影图上（图4），还有不少样品分布的点位与变量分布的点位的相对关系并不正确，而在本文所提出的方法的分析结果图上（图1），得到了明显的改善，充分体现了各样品及变量组合（主因子）相互作用的密切关系；同时，图1上还正确反映出样品在各变量组合之间的差异程度。例如，样品12明显受到第一变量组合的影响，并且不同变量组合对它影响的差异程度最大，应处于点位的极端位置；样品5和11与样品4和14对于变量组合的相对位置得到了调整；样品17应与变量Ⅰ、Ⅴ、Ⅶ的组合因子接近，而与变量Ⅵ远离。这些特征，在图1上得到了满意的成果。

总之，本文所提出的对应分析原理公式的重新推导和应用中的必要的原始数据预处理方法，能比较理想的改善原有对应分析方法的合理性和应用的广泛性。

参 考 文 献

〔1〕 王学仁，地质数据的多变量统计分析，科学出版社，1982。

〔2〕 於崇文等，数学地质的方法与应用，冶金工业出版社，1980。

Improvements on the Existing Correspondence Analysis and Its Data Pre-processing

Zhang Kequan *Guo Renzhong*

Abstract

In this paper, a new algorithm is put forward as an alternative to the existing correspondence analysis algorithm in multivariate statistical inference, which differs from the existing algorithm mainly in the following two aspects: (1) $z_{ij} = \frac{p_{ij}}{p_{i\cdot} \cdot p_{\cdot j}} - 1$ instead of $\frac{p_{ij} - p_{i\cdot} \cdot p_{\cdot j}}{\sqrt{p_{i\cdot} \cdot p_{\cdot j}}}$, and (2) performing necessary data pre-processing through the transformation formula $x'_{ij} = \frac{x_{ij} - \min x_{ij}}{\max x_{ij} - \min x_{ij}}$ and the constraint $x'_{i\cdot} = 1$ (or other constant). Moreover, a practical experimental example is given to make a brief comparison between the here-stated algorithm and the existing one as well as the individual R-mode and Q-mode factor analyses. Both theory and example prove that the algorithm stated in this paper is more logical in theory and applicable in practice than the existing one and overcomes the limitation of various kinds of geo-data on the existing one.

[Key Words] adapted correspondence analysis; pre-processing of raw data; Q-mode factor analysis; R-mode factor analysis; standardization with extreme difference