DOI:10.13203/j.whugis1987.03.007

第12卷 第3期 武汉测绘科技大学学报 Vol.12 No.3
1987年 9月 Journal of Wuhan Technical University of Surveying and Mapping Sept. 1987

Q型聚类分析中变量相关性的处理方法分析

郭仁忠 张克权

摘 要

本文分析了斜交距离法、主成分分析法和马氏距离法等处理原始变量相关性的方法的原理，论述了Q型聚类分析相似性统计量的几种数据处理方法之间的特点及其等价关系，并且用实际算例验证了理论推导的正确性。

【关键词】 马氏距离；欧氏距离；斜交距离；主成分分析；Q型聚类分析

一、引 言

在进行Q型聚类分析的过程中，通常采用欧氏距离作为聚类统计量，这就要求变量是相互独立的。但是由于变量之间经常存在的不同程度的相关性，所以需要进行这种相关性的处理，以使分类结果更合理。目前各种文献介绍的变量间相关性的处理方法主要有三种：

1．斜交距离法；

2．主成分分析法；

3．马氏距离法。

以上三种方法各自的特点是什么，其效果如何，相互之间有何联系和差异？笔者迄今尚未见文献讨论过。事实上以上三种方法并不都是等价的，处理的结果有时具有截然相反的效果。本文的目的是对以上三种方法的原理进行分析，并以数据实例进行验证，指出各自的特点和功能，以便于在实际工作中针对需要有目的地选用。

二、斜交距离法

考虑到变量之间的相关性，坐标系不应为正交的，这就引入了斜交距离的概念。样品P_i、P_k的斜交距离的计算公式如下：

本文1986年11月收到。

$$D_{ik}=\left[\sum_{j=1}^{m}\sum_{l=1}^{m}(x_{ij}-x_{kj})(x_{il}-x_{kl})r_{jl}\right]^{\frac{1}{2}},\quad (i,\ k=1,\ 2,\ 3,\ \cdots,\ n) \tag{1}$$

上式中 n 为样品数，m 为变量数，r_{jl} 为变量间的相关（相似）系数。将 D_{ik} 平方后展开并以矩阵表示如下：

$$\begin{aligned}
D_{ik}^2 &= (x_{i1}-x_{k1})(x_{i1}-x_{k1})r_{11}+(x_{i1}-x_{k1})(x_{i2}-x_{k2})r_{12}+\cdots+\\
&\quad (x_{i1}-x_{k1})(x_{im}-x_{km})r_{1m}+(x_{i2}-x_{k2})(x_{i1}-x_{k1})r_{21}+\\
&\quad (x_{i2}-x_{k2})(x_{i2}-x_{k2})r_{22}+\cdots+(x_{i2}-x_{k2})(x_{im}-x_{km})r_{2m}+\\
&\quad +\cdots\cdots\cdots\cdots\\
&\quad +(x_{im}-x_{km})(x_{i1}-x_{k1})r_{m1}+(x_{im}-x_{km})(x_{i2}-x_{k2})r_{m2}+\cdots+\\
&\quad +(x_{im}-x_{km})(x_{im}-x_{km})r_{mm}\\
&=\begin{pmatrix}(x_{i1}-x_{k1})\\(x_{i2}-x_{k2})\\\cdots\cdots\\(x_{im}-x_{km})\end{pmatrix}'\begin{pmatrix}(x_{i1}-x_{k1})r_{11}+(x_{i2}-x_{k2})r_{12}+\cdots\cdots+(x_{im}-x_{km})r_{1m}\\(x_{i1}-x_{k1})r_{21}+(x_{i2}-x_{k2})r_{22}+\cdots\cdots+(x_{im}-x_{km})r_{2m}\\\cdots\cdots\cdots\cdots\\(x_{i1}-x_{k1})r_{m1}+(x_{i2}-x_{k2})r_{m2}+\cdots\cdots+(x_{im}-x_{km})r_{mm}\end{pmatrix}\\
&=\begin{pmatrix}(x_{i1}-x_{k1})\\(x_{i2}-x_{k2})\\\cdots\cdots\\(x_{im}-x_{km})\end{pmatrix}'\begin{pmatrix}r_{11}&r_{12}&\cdots\cdots&r_{1m}\\r_{21}&r_{22}&\cdots\cdots&r_{2m}\\\cdots\cdots&\cdots\cdots&\cdots\cdots&\cdots\cdots\\r_{m1}&r_{m2}&\cdots\cdots&r_{mm}\end{pmatrix}\begin{pmatrix}(x_{i1}-x_{k1})\\(x_{i2}-x_{k2})\\\cdots\cdots\\(x_{im}-x_{km})\end{pmatrix}\\
&=(P_i-P_k)'R(P_i-P_k).
\end{aligned} \tag{2}$$

当变量互不相关时，R = E 为单位矩阵，（2）式变为：

$$D_{ik}^2=(P_i-P_k)'E(P_i-P_k)=(P_i-P_k)'(P_i-P_k). \tag{3}$$

（3）式为欧氏距离的计算公式，所以欧氏距离即为假定变量不相关时的空间距离。

斜交距离 D_{ik} 可以用几何图形直观地显示出其与欧氏距离之间的差异。设 m = 2，即变量为 x_1、x_2 时，任意两样品 $P_i=\{x_{i1},\ x_{i2}\}$，$P_k=\{x_{k1},\ x_{k2}\}$，当 x_1，x_2 之间的相关系数 $r_{12}>0$ * 时，斜交距离 D_{ik} 如图 1 (a) 所示，$r_{12}=0$ 时如图 1 (b) 所示，$r_{12}<0$ 时则如

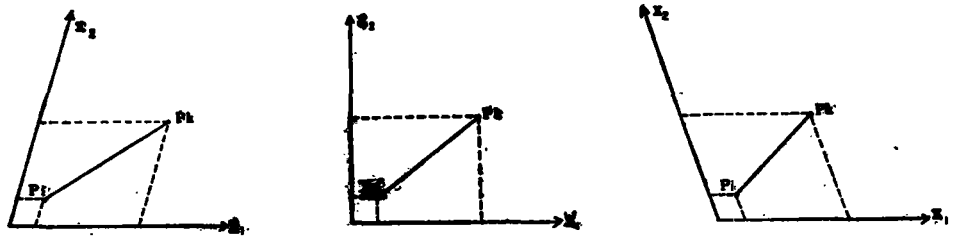


图 1

* 本文论证均是对几种处理相关性变量方法的原理上的推导，实际工作中能否采用欧氏距离，需要事先对变量间相关系数进行显著性检验后确定。

图 1 (c) 所示。P_i、P_k 之间的欧氏距离则无论 r_{12} 为何值，都如图 1（b）所示。这就是说欧氏距离没有考虑变量间的相关性，或者说欧氏距离不受变量间相关性的影响。

关于斜交距离与欧氏距离的差异从图 1 尚不能充分表达出来，现进一步分析如下。

给定一组变量 x_1，x_2，x_3，……，x_m，不失一般性，

假设 $r_{12}=r_{21}\neq 0$；

$r_{jl}=0$。($j\neq l$，$jl\neq 21$或12，j，$l=1$，2，3，……，m)

就是说除了 x_1 和 x_2 相关外，其余变量之间都互不相关，我们来探讨变量相关的影响如何。由斜交距离的计算公式知，变量 x_1，x_2 对 P_i，P_k 之间的距离的贡献为：

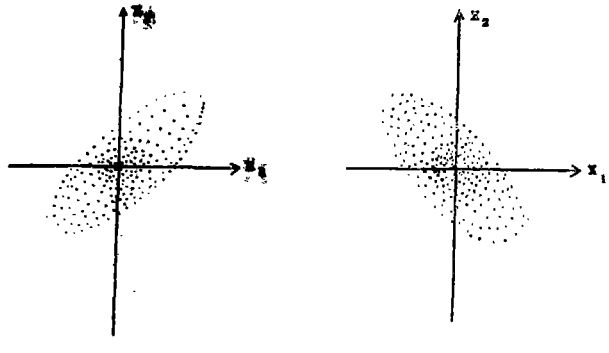
$$\Delta D^2_{ik}(1,2)=(x_{i1}-x_{k1})^2+(x_{i2}-x_{k2})^2+2r_{12}(x_{i1}-x_{k1})(x_{i2}-x_{k2}), \tag{4}$$

而 x_1、x_2 对欧氏距离平方的贡献为：

$$\Delta d^2_{ik}(1,2)=(x_{i1}-x_{k1})^2+(x_{i2}-x_{k2})^2, \tag{5}$$

两者之差为 $2r_{12}(x_{i1}-x_{k1})(x_{i2}-x_{k2})$。显然当 $|r_{12}|$ 较小时，两者差异是较小的，随着 x_1，x_2 之间的相关系数 r_{12} 的绝对值的增大，两种距离的差异显然亦将增大。当 $r_{12}>0$ 时，样品在 x_1，x_2 的坐标空间的分布如图 2 (a) 所示，由于 x_1，x_2 的正相关性，由图 2 (a) 知通常有 $(x_{i1}-x_{k1})(x_{i2}-x_{k2})>0$，$r_{12}$ 越大，$(x_{i1}-x_{k1})(x_{i2}-x_{k2})$ 的值大于零的可能性亦越大，当 $r_{12}=1$，这就是说所有样品位于 $x_2=ax_1+b$ ($a>0$) 上，必然有 $(x_{i1}-x_{k1})(x_{i2}-x_{k2})>0$。因此如果 x_1，x_2 为正相关，尤其当相关相当紧密时 ($r_{12}\approx 1$)，x_1，x_2 对斜交距离的贡献要比互不相关时为大。类似地，当 $r_{12}<0$ 时，由图 2 (b) 知，由于 x_1，x_2 之间的负相关性，常有 $(x_{i1}-x_{k1})(x_{i2}-x_{k2})<0$，当 $|r_{12}|$ 越大时，这种小于零的可能性越大，当 $r_{12}=-1$ 时，$(x_{i1}-x_{k1})(x_{i2}-x_{k2})<0$ 必成立。所以无论 $r_{12}>0$ 还是 $r_{12}<0$，通常都有：

$$r_{12}(x_{i1}-x_{k1})(x_{i2}-x_{k2})>0,$$



(a) (b)

图 2

亦即斜交距离大于欧氏距离，$|r_{12}|$ 越大，这种可能性亦越大。如果我们将（4）、（5）作一比较，令 $K=(x_{i1}-x_{k1})/(x_{i2}-x_{k2})$，得：

$$\frac{\Delta D^2_{ik}(1,2)}{\Delta d^2_{ik}(1,2)} = \frac{(x_{i1}-x_{k1})^2+(x_{i2}-x_{k2})^2+2r_{12}(x_{i1}-x_{k1})(x_{i2}-x_{k2})}{(x_{i1}-x_{k1})^2+(x_{i2}-x_{k2})^2}$$

$$= 1 + \frac{2r_{12}(r_{i1}-x_{k1})(x_{i2}-x_{k2})}{(x_{i1}-x_{k1})^2+(x_{i2}-x_{k2})^2}$$

$$= 1 + \frac{2r_{12}}{K+\frac{1}{K}}.$$

因为 $\left|K+\frac{1}{K}\right| \geqslant 2$，所以当 $r_{12}>0$，$K+\frac{1}{K}=2$ 时，亦即 P_i，P_k 处于 $x_2=x_1+c$ 的位置上时，斜交距离达最大。同样地，当 $r_{12}<0$，$K+\frac{1}{K}=-2$ 时，亦即当 P_i，P_k 处于 $x_1=-x_2+c$ 上时，斜交距离达最大。对照图 2，可得出如下结论：当样品沿分布椭圆的长轴方向时，则斜交距离大于欧氏距离，当沿分布椭圆的短轴方向时，则小于欧氏距离，这是一个连续变化过程。总之斜交距离给予相关变量以大于其它互不相关变量的权重，而欧氏距离相当于给所有变量以相等的权重。欧氏距离与斜交距离的差异将在本文第六部分算例中得到进一步验证。

三、 主成分分析法

主成分分析法的基本思想是先将原始变量转换为互不相关的假设“正交”变量，然后采用欧氏距离作为聚类统计量。这互不相关的正交变量就是主成分。主成分分析的数学模型如下：

$$X = AF. \qquad (6)$$

（6）中 X 为经规格化了的原始变量，A 为因子载荷矩阵，F 为主成分。A 的解如下：

$$XX' = R = U\Lambda U' = U\Lambda^{\frac{1}{2}}\Lambda^{\frac{1}{2}}U' = AA',$$

故

$$A = U\Lambda^{\frac{1}{2}},$$

这里 R 为相关矩阵，U 为 R 的特征向量矩阵，Λ 为以 R 的特征值 λ_1，λ_2，……，λ_m 为元素的对角矩阵，$\Lambda^{\frac{1}{2}}$ 则是以 $\lambda_1^{\frac{1}{2}}$，$\lambda_2^{\frac{1}{2}}$，……，$\lambda_m^{\frac{1}{2}}$ 为元素的对角矩阵。则 m 个主成分的解是以主成分为坐标轴的空间中 n 个样品的坐标。因为 n 个样品在以 m 个原始变量 x_1，x_2，……，x_m 为坐标轴的空间中的坐标为

$$X' = \begin{pmatrix} x_{11} & x_{12} & \cdots\cdots & x_{1m} \\ x_{21} & x_{22} & \cdots\cdots & x_{2m} \\ & \cdots\cdots\cdots & & \\ x_{n1} & x_{n2} & \cdots\cdots & x_{nm} \end{pmatrix},$$

根据（6），X = AF，所以主成分的解为 $X'A = X'U\Lambda^{\frac{1}{2}}$，展开即为：

$$\begin{pmatrix} x_{11} & x_{12} & \cdots\cdots & x_{2m} \\ x_{21} & x_{22} & \cdots\cdots & x_{2m} \\ \cdots\cdots & & & \\ x_{n1} & x_{n2} & \cdots\cdots & x_{nm} \end{pmatrix} \begin{pmatrix} u_{11} & u_{12} & \cdots\cdots & u_{1m} \\ u_{21} & u_{22} & \cdots\cdots & u_{2m} \\ \cdots\cdots & & & \\ u_{m1} & u_{m2} & \cdots\cdots & u_{mm} \end{pmatrix} \begin{pmatrix} \lambda_1^{\frac{1}{2}} & & & \\ & \lambda_2^{\frac{1}{2}} & & \\ & & \ddots & \\ & & & \lambda_m^{\frac{1}{2}} \end{pmatrix},$$

任意两样品 P_i、P_k 的坐标差为：

$$\begin{pmatrix} (x_{i1}-x_{k1}) \\ (x_{i2}-x_{k2}) \\ \cdots\cdots \\ (x_{im}-x_{km}) \end{pmatrix} \begin{pmatrix} u_{11} & u_{12} & \cdots\cdots & u_{1m} \\ u_{21} & u_{22} & \cdots\cdots & u_{2m} \\ \cdots\cdots & & & \\ u_{m1} & u_{m2} & \cdots\cdots & u_{mm} \end{pmatrix} \begin{pmatrix} \lambda_1^{\frac{1}{2}} & & & \\ & \lambda_2^{\frac{1}{2}} & & \\ & & \ddots & \\ & & & \lambda_m^{\frac{1}{2}} \end{pmatrix}. \tag{7}$$

记 $(x_{ij}-x_{kj})=d_j$，$d'=(d_1, d_2, \cdots\cdots, d_m)'$，则（7）为

$$d'U\Lambda^{-\frac{1}{2}},$$

P_i，P_k 的欧氏距离的平方为：

$$d_{ik}^2 = d'U\Lambda^{-\frac{1}{2}}\Lambda^{-\frac{1}{2}}U'd = d'AA'd = d'Rd. \tag{8}$$

将（8）与（2）对比，我们发现两式完全一致。这就是说主成分分析法与斜交距离法是完全等价的。当然这里假定了变量是经过标准化的，然而标准化与否只是数据预处理方法的差异，本文第六部分将以算例说明主成分分析法和斜交距离的等价性。

主成分分析法与斜交距离的等价性从几何上可以得到直观的说明。因为主成分解就是样品以相互正交的主成分为坐标轴空间的坐标，这只是一个坐标的线性变换，当然是不改变空间相互关系的。

四、 馬氏距离法

多元正态随机变量 $X=(x_1, x_2, \cdots\cdots, x_m)$ 的分布密度函数为：

$$f(x_1,x_2,\cdots\cdots,x_m)=(2\pi)^{-\frac{1}{2}m}|\Sigma|^{-\frac{1}{2}}e^{-\frac{1}{2}(x-\mu)'\Sigma^{-1}(x-\mu)}$$

式中 μ 为 X 的均值向量，Σ 为 X 的协方差矩阵，由正态分布的性质知，样品与均值的差异越大，其发生的概率越小，这样就认为两者的距离越大。马氏距离就是采用上述密度函数式中的指数部分的 $(X-\mu)\Sigma^{-1}(X-\mu)$ 作为样品间距离的度量基础的，若 Σ 以样本协方差矩阵 S 代替，则任意两样品 P_i，P_k 之间的马氏距离为

$$D_{ik}=[(P_i-P_k)'S^{-1}(P_i-P_k)]^{\frac{1}{2}}. \tag{9}$$

不失一般性，仍设

$$r_{12}=r_{21}\neq 0,$$

$$r_{jl}=0, \quad (j\neq l,\ jl\neq 12或21,\ j,l=1,2,\cdots\cdots,m),$$

则有：

$$S=\begin{pmatrix} s_1^2 & s_{12} & & & 0 \\ s_{21} & s_2^2 & & & \\ & & s_3^2 & & \\ & & & \ddots & \\ 0 & & & & s_m^2 \end{pmatrix},$$

$$S^{-1}=\begin{pmatrix} \dfrac{-s_2^2}{s_{12}^2-s_1^2s_2^2} & \dfrac{s_{12}}{s_{12}^2-s_1^2s_2^2} & & & 0 \\ \dfrac{s_{21}}{s_{12}^2-s_1^2s_2^2} & \dfrac{-s_1^2}{s_{12}^2-s_1^2s_2^2} & & & \\ & & s_3^{-2} & & \\ & & & \ddots & \\ 0 & & & & s_m^{-2} \end{pmatrix}.$$

令 $$[S^{-1}]=\begin{pmatrix} \dfrac{-s_2^2}{s_{12}^2-s_1^2s_2^2} & \dfrac{s_{12}}{s_{12}^2-s_1^2s_2^2} \\ \dfrac{s_{21}}{s_{12}^2-s_1^2s_2^2} & \dfrac{-s_1^2}{s_{12}^2-s_1^2s_2^2} \end{pmatrix},$$

则变量 x_1, x_2 对马氏距离的贡献为：

$$\Delta D_{ik}^2(1,2)=[(x_{i1}-x_{k1})(x_{i2}-x_{k2})]\begin{Bmatrix} s^{-1} \end{Bmatrix}\begin{Bmatrix} (x_{i1}-x_{k1}) \\ (x_{i2}-x_{k2}) \end{Bmatrix}$$

$$=\frac{1}{s_{12}^2-s_1^2s_2^2}[-s_2^2(x_{i1}-x_{k1})^2+2s_{12}(x_{i1}-x_{k1})(x_{i2}-x_{k2})$$

$$-s_1^2(x_{i2}-x_{k2})^2]. \qquad (10)$$

(10) 中 $S_{12}=r_{12}S_1S_2$，又令 $x_{i1}-x_{k1}=d_1$，$x_{i2}-x_{k2}=d_2$，代入得：

$$\Delta D_{ik}^2(1,2)=\frac{1}{s_1^2s_2^2(r_{12}^2-1)}[-s_2^2d_1^2+2r_{12}s_1s_2d_1d_2-s_1^2d_2^2]$$

$$=\frac{1}{r_{12}^2-1}\left[-\left(\frac{d_1}{s_1}\right)^2+2r_{12}\left(\frac{d_1d_2}{s_1s_2}\right)-\left(\frac{d_2}{s_2}\right)^2\right]. \qquad (11)$$

(11) 中 d_1/s_1，d_2/s_2 相当于消除变量尺度之影响，等价于原始变量的规格化，故 (11) 可写为：

$$\Delta D_{ik}^2(1,2)=\frac{1}{1-r_{12}^2}[(x_{i1}-x_{k1})^2-2r_{12}(x_{i1}-x_{k1})(x_{i2}-x_{k2})$$

$$+(x_{i2}-x_{k2})^2]. \qquad (12)$$

将 (12) 与 (15) 作一比较得：

$$\frac{\Delta D_{ik}^2(1,2)}{\Delta d_{ik}^2(1,2)}=\frac{1}{1-r_{12}^2}\left\{1-\frac{2r_{12}(x_{i1}-x_{k1})(x_{i2}-x_{k2})}{(x_{i1}-x_{k1})^2+(x_{i2}-x_{k2})^2}\right\}$$

$$= \frac{1}{1-r_{12}^2}\left(1-\frac{2r_{12}}{K+1/K}\right). \tag{13}$$

式中 $K=(x_{i1}-x_{k1})/(x_{i2}-x_{k2})$，由前述知当 $r_{12}>0$，常有 $K>0$；$r_{12}<0$，常有 $K<0$。所以 $2r_{12}/(K+\frac{1}{K})$ 通常大于零。因为 $|K+\frac{1}{K}|\geqslant 2$，当 $r_{12}\geqslant 0$ 时，若 $K+\frac{1}{K}=2$，说明样品 P_i，P_k 处于 $x_1=x_2+c$ 的直线位置上，此时，马氏距离达极小值，原式等于 $(1-r)/(1-r^2)=1/(1+r)<1$，说明此时马氏距离小于欧氏距离。当 P_i，P_k 偏离 $x_1=x_2+c$ 时，马氏距离逐渐增大，当样品 P_i，P_k 处于 $x_1=-x_2+c$ 的位置上时，马氏距离达到极大，且大于欧氏距离。当 $r_{12}<0$ 时，则有完全对应的类似结论。对照图 2 可以得出如下结论：当样品沿分布椭圆的长轴方向时，则马氏距离小于欧氏距离，沿短轴方向则马氏距离大于欧氏距离。这一结论说明了在处理变量相关性影响方面，斜交距离和马氏距离的功能正好相反，对于相关变量，前者通常给予较大的权值而后者给予较小的权值，欧氏距离则介于两者之间。

五、馬氏距离的进一步分析

由因子分析知：

$$X=AF,\quad A=U\Lambda^{\frac{1}{2}},\quad U\Lambda^{\frac{1}{2}}\Lambda^{\frac{1}{2}}U'=R. \tag{14}$$

(14) 中各符号的意义均如前所述。由 (14) 得：

$$F=A^{-1}X=(U\Lambda^{\frac{1}{2}})^{-1}X=\Lambda^{-\frac{1}{2}}U'X, \tag{15}$$

转置 (15) 得：

$$F'=X'U\Lambda^{-\frac{1}{2}}, \tag{16}$$

展开 (16) 得：

$$\begin{pmatrix} f_{11} & f_{12} & \cdots & f_{1m} \\ f_{21} & f_{22} & \cdots & f_{2m} \\ \cdots & \cdots & \cdots & \cdots \\ f_{n1} & f_{n2} & \cdots & f_{nm} \end{pmatrix} = \begin{pmatrix} x_{11} & x_{22} & \cdots & x_{1m} \\ x_{21} & x_{22} & \cdots & x_{2m} \\ \cdots & \cdots & \cdots & \cdots \\ x_{n1} & x_{n2} & \cdots & x_{nm} \end{pmatrix} \begin{pmatrix} u_{11} & u_{12} & \cdots & u_{1m} \\ u_{21} & u_{22} & \cdots & u_{2m} \\ \cdots & \cdots & \cdots & \cdots \\ u_{m1} & u_{m2} & \cdots & u_{mm} \end{pmatrix} \cdot$$

$$\begin{pmatrix} \lambda_1^{-\frac{1}{2}} & & & 0 \\ & \lambda_2^{-\frac{1}{2}} & & \\ & & \ddots & \\ 0 & & & \lambda_m^{-\frac{1}{2}} \end{pmatrix}. \tag{17}$$

(17) 为因子得分的计算式，根据因子得分计算 P_i，P_k 之间的欧氏距离则有：

$$D_{ik}=\left[\sum_{j=1}^{m}(f_{ij}-f_{kj})^2\right]^{\frac{1}{2}}$$

$$=\{[(x_{i1}-x_{k1})(x_{i2}-x_{k2})\cdots\cdots(x_{im}-x_{km})]U\Lambda^{-\frac{1}{2}}\Lambda^{-\frac{1}{2}}U'\cdot$$
$$[(x_{i1}-x_{k1})(x_{i2}-x_{k2})\cdots\cdots(x_{im}-x_{km})]'\}^{\frac{1}{2}}$$

$$=[(P_i-P_k)'(A^{-1})'A^{-1}(P_i-P_k)]^{\frac{1}{2}}$$

$$=[(P_i-P_k)'R^{-1}(P_i-P_k)]^{\frac{1}{2}} \quad (18)$$

(18) 与 (9) 基本相同，只是中间的逆矩阵不同，前者为 R^{-1}，后者为 S^{-1}。由 (12) 式我们知道变量规格化与否，(9) 的计算结果不变；而采用规格化数据时 $S^{-1}=R^{-1}$，则 (9) 与 (18) 完全相同。这就是说采用因子得分计算的样品间距离与马氏距离完全等价，这一等价关系与斜交距离和主成分分析法的等价关系是对应的，而对于实际工作中灵活选取聚类统计量是很有好处的，这将在第七部分进一步说明。

六、算　例

下面选用一个数据计算实例来分别说明前述原理推导的论证结果。

假设有样品单元 n = 10，指标变量 m = 5 的一组数据，其几种距离的计算过程和结果如下：

1. 变量数据矩阵和相关矩阵

1)　设给定的原始数据矩阵为 $X=[x_{ij}]$ (见表 1)。

表1

i	j				
	1	2	3	4	5
1	3.76	3.66	0.54	5.27	9.76
2	8.59	4.99	1.34	10.22	7.50
3	6.52	6.14	4.52	9.84	2.17
4	7.57	7.28	7.07	12.66	1.79
5	9.03	7.08	2.59	11.80	4.54
6	5.51	3.98	1.30	6.90	5.32
7	3.27	0.62	0.44	3.35	7.63
8	8.74	7.00	3.31	11.60	3.53
9	9.64	9.49	1.03	13.00	13.13
10	9.73	1.38	1.00	9.80	9.87

2)　根据 X 矩阵按常用的数据规格化方法，即按式 $x'_{ij}=\frac{x_{ij}-\overline{x_j}}{\sigma_j}$ 计算得规格化后的数据矩阵 X′ (见表 2)。

3)　根据原始变量数据矩阵 X 计算各变量间的相关矩阵 R (见表 3)；并计算其逆矩阵 R^{-1} (见表 4)。

表2

i	j				
	1	2	3	4	5
1	−1.54	−0.56	−0.88	−1.34	0.92
2	0.60	−0.06	−0.48	0.23	0.28
3	−0.32	0.37	1.10	0.11	−1.24
4	0.15	0.80	2.36	1.00	−1.35
5	0.80	0.73	0.14	0.73	−0.57
6	−0.77	−0.44	−0.50	−0.82	−0.34
7	−1.76	−1.71	−0.93	−1.95	0.32
8	0.67	0.70	0.49	0.66	−0.85
9	1.07	1.64	−0.64	1.30	1.88
10	1.11	−1.44	−0.65	0.09	0.95

表3

i	j				
	1	2	3	4	5
1	1.00	0.54	0.21	0.89	0.07
2	0.54	1.00	0.50	0.83	−0.18
3	0.21	0.50	1.00	0.53	−0.79
4	0.89	0.83	0.53	1.00	−0.14
5	0.07	−0.18	−0.79	−0.14	1.00

表4

i	j				
	1	2	3	4	5
1	1031.56	578.92	572.76	−1665.03	247.12
2	578.92	328.10	321.17	−936.94	138.68
3	572.76	321.17	323.24	−926.49	141.01
4	−1665.03	−936.94	−926.49	2691.45	−400.34
5	247.12	138.68	141.01	−400.34	63.00

2. 求因子载荷矩阵、主成分值和因子得分

1） 解算相关矩阵 R 后得特征值矩阵 Λ 和特征向量矩阵 U（见表 5）。

2） 求得因子载荷矩阵 A（见表 6）及其逆矩阵 A^{-1}（见表 7）。

表5

u_{ji}	λ_i				
	1	2	3	4	5
	2.920	1.528	0.416	0.136	0.000
u_{1i}	0.436	0.437	−0.598	−0.170	−0.482
u_{2i}	0.498	0.128	0.759	−0.292	−0.271
u_{3i}	0.436	−0.498	−0.028	0.699	−0.269
u_{4i}	0.556	0.247	−0.079	0.126	0.780
u_{5i}	−0.252	0.695	0.243	0.617	−0.116

表6

j	i				
	1	2	3	4	5
1	0.745	0.540	−0.385	−0.063	−0.007
2	0.850	0.159	0.490	−0.108	−0.004
3	0.744	−0.616	−0.018	0.258	−0.004
4	0.950	0.305	−0.051	0.047	0.012
5	−0.430	0.860	0.157	0.228	−0.002

表7

j	i				
	1	2	3	4	5
1	0.26	0.29	0.25	0.33	−0.15
2	0.35	0.10	−0.40	0.20	0.56
3	−0.93	1.18	−0.04	−0.12	0.38
4	−0.46	−0.79	1.90	0.34	1.67
5	−32.11	−18.06	−17.88	51.90	−7.73

3） 由表 2 和表 6 计算得主成分值矩阵 W（见表 8）。

4） 根据表 7 和表 2 求得因子得分值 F（见表 9）。

表 8

样 品	成 分				
	1	2	3	4	5
1	−3.9563	0.0019	0.5469	0.0778	−0.0003
2	0.1305	0.9207	−0.2217	−0.0819	0.0000
3	1.5279	−1.8209	0.0844	−0.0148	−0.0002
4	4.0784	−2.1006	0.0300	0.2533	0.0002
5	2.2465	0.1973	−0.0797	−0.1877	0.0002
6	−1.9595	−0.7212	0.0751	−0.1505	0.0003
7	−5.4477	−0.9744	0.0056	0.0359	0.0001
8	2.4544	−0.3638	−0.0553	−0.1525	−0.0004
9	2.1348	3.2424	0.6301	0.0810	0.0001
10	−1.2090	1.6185	−0.9774	0.1394	−0.0001

表 9

样 品	因 子				
	1	2	3	4	5
1	−1.36	0.00	1.32	0.57	−1.21
2	0.04	0.60	−0.53	−0.60	0.10
3	0.52	−1.19	0.20	−0.11	−0.96
4	1.40	−1.37	0.07	1.06	0.85
5	0.77	0.13	−0.19	−1.38	0.97
6	−0.67	−0.47	0.18	−1.10	1.48
7	−1.87	−0.64	0.01	0.26	0.52
8	0.84	−0.24	−0.22	−1.15	−1.80
9	0.73	2.12	1.52	0.59	0.36
10	−0.41	1.06	−2.35	1.02	−0.31

3．求各种数据处理方法所得的距离系数

1） 根据规格化数据矩阵（表2）计算一般的欧氏距离系数矩阵D（不考虑变量间相关的影响）（见表10）。

表10

j	i									
	1	2	3	4	5	6	7	8	9	10
1	0.00	7.88	13.05	25.87	14.67	2.63	2.10	15.54	19.60	9.91
2	7.88	9.00	5.85	12.29	2.01	3.51	13.23	3.01	6.84	2.67
3	13.05	5.85	0.00	2.81	3.13	5.09	17.17	1.90	17.69	13.19
4	25.87	12.29	2.81	0.00	6.00	14.93	32.20	4.13	21.00	21.16
5	14.67	2.01	3.13	6.00	0.00	5.68	21.50	0.23	7.81	8.13
6	2.63	3.51	5.09	14.93	6.68	0.00	4.48	6.83	17.16	7.06
7	2.10	13.23	17.17	32.20	21.50	4.48	0.00	21.92	32.28	12.96
8	15.54	2.01	1.90	4.13	0.23	6.83	21.92	0.00	10.21	9.67
9	19.60	6.84	17.69	21.00	7.81	17.16	32.28	10.21	0.00	11.79
10	9.91	2.67	13.19	21.16	8.13	7.06	12.96	9.67	11.79	0.00

表 11

j	i									
	1	2	3	4	5	6	7	8	9	10
1	0.00	18.16	33.62	69.27	38.98	4.78	3.47	41.69	47.61	12.49
2	18.16	0.00	9.57	24.89	5.03	7.16	34.77	7.07	10.16	2.90
3	33.62	9.57	0.00	6.66	4.65	13.39	49.38	3.03	26.31	20.47
4	69.27	24.89	6.66	0.00	8.84	38.52	92.06	5.83	32.71	42.82
5	38.98	5.03	4.65	8.84	0.00	18.56	60.63	0.36	9.86	14.87
6	4.78	7.16	13.39	38.52	18.56	0.00	12.27	19.64	32.88	7.23
7	3.47	34.77	49.38	92.06	60.63	12.27	0.00	62.86	75.67	25.67
8	41.69	7.07	3.03	5.83	0.36	19.64	62.86	0.00	13.69	18.22
9	47.61	10.16	26.31	32.71	9.86	32.83	75.67	13.69	0.00	16.41
10	12.49	2.90	20.47	42.82	14.87	7.23	25.67	18.22	16.41	0.00

2） 根据表2和表3计算斜交距离系数矩阵$D_{斜}$（见表11）。

3） 根据表8按计算欧氏距离公式得主成分值距离系数矩阵$D_{成}$，其结果与斜交距离（表11）完全一致。

4） 根据表4和表2计算得马氏距离系数矩阵$D_{马}$（见表12）。

5） 根据表9按计算欧氏距离公式得因子得分距离系数矩阵$D_{因}$，其结果与马氏距离（表12）完全一致。

表12

j	i									
	1	2	3	4	5	6	7	8	9	10
1	0.00	8.83	6.71	16.92	15.37	12.05	5.44	10.45	11.37	16.45
2	8.33	0.00	5.36	12.72	2.24	4.35	6.41	5.31	8.48	6.53
3	6.71	5.36	0.00	7.97	7.32	8.91	8.38	2.92	15.00	14.16
4	16.92	12.72	7.97	0.00	13.22	14.29	13.85	17.59	16.59	17.12
5	15.37	2.24	7.32	13.22	0.00	2.91	10.48	7.91	11.15	14.35
6	12.05	4.35	8.91	14.29	2.91	0.00	4.29	13.29	14.62	16.57
7	5.44	6.41	8.38	13.85	10.48	4.29	0.00	14.82	16.75	11.84
8	10.45	5.31	2.92	17.59	7.91	13.29	14.82	0.00	16.23	14.58
9	11.37	8.48	15.00	16.59	11.15	14.62	16.75	16.23	0.00	18.03
10	16.45	6.53	14.16	17.12	14.35	16.57	11.84	14.58	18.03	0.00

七、 几 点 結 論

根据前面的推导分析和计算验证，可以得出以下几点结论：

1．斜交距离和用主成分处理计算的距离完全等价，马氏距离和用因子得分处理计算的距离完全等价；主成分分析法和因子得分法在计算上都要比斜交距离和马氏距离复杂，但主成分分析和因子得分方法具有运用上的灵活性，在保证不丢失过多信息的前提下，可以选取少量主因子（主成分）参与聚类。

2．斜交距离和马氏距离对变量相关处理的效果相反，对于相关紧密的变量，斜交距离赋以较大的权重，欧氏距离赋以等权，而马氏距离赋以较小的权重。从统计学的观点看，两变量相关是说明彼此携带有相同信息，具有重复性，所以应当具有较小权重，故马氏距离是合适的；从应用的观点来看，斜交距离强调了样品间的差异关系，有利于反映样品间的集群性。

3．根据本文算例中欧氏距离、斜交距离和马氏距离的计算成果，用最短距离法进行聚类。按斜交距离聚类的结果与用欧氏距离的一致，并且类间距离是欧氏距离聚类的两倍；而按马氏距离聚类的结果有明显的差异。为此建议，当变量之间虽有相关，但其信息都需充分使用时，采用斜交距离法处理较好；如果信息之间可以互相代替时，采用马氏距离法处理为宜。

参 考 文 献

〔1〕 张尧庭、方开泰，多元统计分析引论，科学出版社，1982.

〔2〕 王学仁，地质数据的多变量统计分析，科学出版社，1982.

〔3〕 於崇文等，数学地质的方法与应用，冶金工业出版社，1980.

〔4〕 Жуков В.Т.、Сербенюк С.Н.、Тикунов В.С.，Математико—картографическое Моделирование в Географий, Москва, "Мысль", 1980.

Analysis and Comparison of the Techniques for Treating Correlativity Between Variables in Q-mode Cluster Analysis

Guo Renzhong　　*Zhang Kequan*

Abstract

The Euclidean distance, which is most often used in Q-mode cluster analysis, is unable to reflect the influence of correlativity between variables on the results of cluster analysis, so it cannot fully reveal the clustering situation of samples. For this reason three techniques are proposed in a lot of literature to solve this problem; they are:

1. the oblique distance method,

2. the principle component analysis method, and

3. the Mahalanobis distance method.

But there has been no paper researching on the characteristics of and the relationship between these methods. In practice, these methods are chosen quite at random and without rules. This paper, in both theory and practice, analyses and compares the three methods, and gives the following conclusions:

1. the oblique distance method functions the same as the principle component

analysis method, because

$$\sum_{j=1}^{m}\sum_{l=1}^{m}(x_{ij}-x_{kj})(x_{il}-x_{kl})r_{jl}=d'U\Lambda^{\frac{1}{2}}\Lambda^{\frac{1}{2}}U'd,$$

in the above equation, the left part is the expression for computing oblique distance, and the right part is the Euclidean distance from the principle component analysis method.

2. the Mahalanobis distance is equal to the Euclidean distance computed from factor scores; this is because, after data standardization,

$$(P_i-P_k)'S^{-1}(P_i-P_k)=(P_i-P_k)'R^{-1}(P_i-P_k)$$
$$=(P_i-P_k)'U\Lambda^{-\frac{1}{2}}\Lambda^{-\frac{1}{2}}U'(P_i-P_k),$$

in this equation, the first part is the expression for computing Mahalanobis distance, and the last part is the expression for computing Euclidean distance from factor scores.

3. the Mahalanobis distance and the oblique distance deal with correlativity in two opposite ways. Generally speaking, the Mahalanobis distance gives correlated variables smaller weight values, and the oblique distance gives them larger weight values, while the Euclidean distance gives equal weights to all variables. If any two samples are located along the direction of major axis of the distribution ellipse, the Mahalanobis distance between them is smaller than the Euclidean distance, while the oblique distance is larger than the Euclidean distance; if located along the direction of minor axis, it will be an opposite conclusion.

The three conclusions stated above are instructive in choosing statistics for clustering in practical work, and helpful in avoiding blindness, and studying the computational results of cluster analysis.

[Key words] Mahalanobis distance; Euclidean distance; oblique distance; principle component analysis; Q-mode cluster analysis