



地信网

<http://www.3s001.com>

点群分析

—地质学中多变量统计方法之六—

中国地质科学院地矿所方法组

一九七七年三月

新华书店
PDG



§1. 引言

点群分析 (cluster Analysis) 是一种分类方法。

自从地质学产生以来，分类就成为地质学研究的最基本方法而被长期沿用。例如，大到地质图上的各种构造带、地层、岩相、侵入岩体界线的划分，小至岩石、矿物类型和古生物种属的确定，都是地质分类的例子。有了分类这个工具，我们才能对地球进行大大小小的划分、对比和研究，建立起目前的传统地质学体系，在过去的一百多年中，分类对地质学作出的贡献是巨大的。

传统的分类方法主要有两种方式：第一是图解法，我们在下面将对它进行回顾。第二种方式是纯粹用人的头脑来综合大量资料，以划分出各种类型。例如在地质填图时，就是由地质人员在现场运用他头脑中贮存的各种地质知识对所观测到的地质现象进行综合分析，来决定应该如何分类和勾绘地质界线；我们知道，地质界线就是各种类型地质体之间的边界。由于分类需要考虑较多的地质特征，而人们在分类时往往只能顾及和强调某些使他印象深刻的特征，而实质上摒弃了其他的许多特征，因而这种分类总难免带有主观性和片面性。

长期以来，地质学中大量沿用一种三角图解分类法。例如，对碎屑岩，根据碎屑成分利用 Δ 的三角图解可以划分出石英砂岩、长石砂岩、硬砂岩，以及其间的一切过渡类型，如石英长石砂岩，长石石英砂岩，硬砂质石英长石砂岩-----等。这种三角图解分类，在地质学各领域中是屡见不鲜的。但是，当端元组分的数目超过三个时，图解分类就无能为力了，因为从几何作图来说，只能绘出一个三度空间的图形。

于是，人们开始寻求新的途径。我们在分类时能不能吸取几



何图解分类的基本思想，但却能避免使用图解的弊端呢？我们知道，所谓三角图解分类法，就是当一个样品的三个端元组分的数值确定时，这个样品在三角形中的位置就是完全确定的。因此，在三角形中那些位置相近的点，也就具有相近的端元组分特征，可以分归一类；另一些相近的点则构成另一类，等等。

点群分析，就是一种既体现了上述分类思想，又不受端元组分数目限制的数字分类技术。

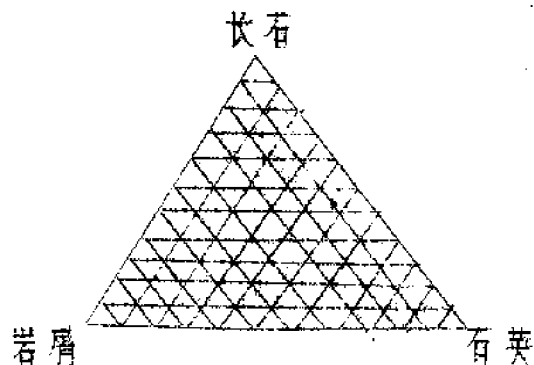


图 1.1 根据碎屑成分的
碎屑岩三角图解
分类示意图

它从属于多变量统计方法，可以用于研究对象的分类（组）。但是，它却不同于判别分析。不同之处在于：对判别分析来说，必须事先知道组数和各组的内容，并且必须事先具备来自各组的子样，才能建立判别分析的数学模型，并且将此模型应用于未知组别的研究对象的分组，判别其属于哪一组的的可能性最大。而对点群分析来说，在给定的一批需要分组的对象中存在的组数和各组的内容都是未知的，而且这正好是需要通过点群分析反其地质解释解决和确定的问题。因此，两者虽然都用于研究对象的分类，但是，却解决着不同范畴的分类问题。

点群分析的出发点是研究对象之间所存在的相似性或者‘亲疏’关系。我们假定给定的一批研究对象可以分成几组，那末显然，同一组研究对象之间的相似性要大于不同组研究对象之间的



相似性。这就构成了点群分析的依裾。我们可以把一些彼此之间相似性较大的研究对象分为一组，把另外一些彼此之间相似性较大的研究对象分为另外一组，-----。这样一来，就可以把不同的组一一地划分出来。以下我们将会看到，研究对象与欧氏空间的点可以一一对应起来。这些点在欧氏空间中形成一个一个的‘集团’，称之为点群。一个组实际上就是一个点群。所以，分组与这些代表研究对象的点在欧氏空间的自然聚集情况相一致。

因此，点群分析的基本思想与地质学中定性的或者几何图解的一些传统方法是一致的，只不过把它们定量化和解析化（公式化）而已。因此所考虑的地质特征（变量）或样品的数目将不受人的思维和作图的限制。这样一来就能更加充分的利用各种地质特征（变量）或各个样品所提供的相似信息，使得分类更客观和更细腻。

与因子分析相类似，点群分析可以分为Q型和R型两种。前者用于样品的分类，后者用于地质特征（变量）的分类。以后，我们把地质特征（变量）统统简称为变量。

本方法在编写过程中得到许多兄弟单位的大力支持。青海地质二队给我们的工作以很大的帮助；文中所引的地质实例均是我所二室超基性岩组提供的；国家地质总局一五〇工程指挥部、地质科学院水文所和北京工业大学计标站为我们提供了大量计标时间，使我们能顺利地完成DTS-6机点群分析程序，并试标了一定数量的问题，在此一并表示感谢。



§2. 相似性度量

假定有 N 个样品 $\mathbf{x}'_k = (x_{1k}, x_{2k}, \dots, x_{pk}) (k=1, 2, \dots, N)$ 其中 $x_{1k}, x_{2k}, \dots, x_{pk}$ 是第 k 个样品的变量 x_1, x_2, \dots, x_p 的取值, ' ' 表示转置。 N 个样品 $\mathbf{x}'_k = (x_{1k}, x_{2k}, \dots, x_{pk})$ 也可以看作是 p 维变量 $\mathbf{x}' = (x_1, x_2, \dots, x_p)$ 的 N 次观测。我们可以把这 N 个多变量观测值排成一个表或矩阵如下:

$$\mathbf{X} = (x_{ik}) = \begin{pmatrix} x_{11} & x_{12} & \dots & x_{1N} \\ x_{21} & x_{22} & \dots & x_{2N} \\ \dots & \dots & \dots & \dots \\ x_{p1} & x_{p2} & \dots & x_{pN} \end{pmatrix}, \quad (2.1)$$

其中元素的第 1 个下标 i ($i=1, 2, \dots, p$) 表示变量的序号, 第 2 个下标 k ($k=1, 2, \dots, N$) 表示样品的序号。由于第 k ($k=1, 2, \dots, N$) 个样品被一组数 $(x_{1k}, x_{2k}, \dots, x_{pk})'$, 即矩阵 \mathbf{X} 的第 k 列所描述; 第 i ($i=1, 2, \dots, p$) 个变量被一组数 $(x_{i1}, x_{i2}, \dots, x_{iN})$, 即矩阵 \mathbf{X} 的第 i 行所描述, 所以, N 个样品之间的相似性就寓于矩阵 \mathbf{X} 的 N 列之间; p 个变量之间的相似性就寓于矩阵 \mathbf{X} 的 p 行之间。

那末到底用什么来衡量样品或者变量之间的相似性大小呢? 两个样品或者变量之间的相似性, 如上所述, 无非表现在以下三个方面:

(1) 在矩阵 \mathbf{X} 中, 描述两个样品或者变量的两列或者两行对应的元素接近的程度;

(2) 在矩阵 \mathbf{X} 中, 描述两个样品或者变量的两列或者两行对应的元素成比例的程度;

(3) 在矩阵 \mathbf{X} 中, 描述两个样品或者变量的两列或者两行对应



的元素互长互消的关系密切的程度。

以上三个方面反映了三种不同意义下的相似。因此，用以下的三种统计量作为两个样品之间相似性大小的数量指标是合适的。（对于变量的情形，后面再讲）。

1) 距离函数

这里指的距离实际上是 *Mahalanobis* 距离（见地质学中多变量统计之五：判别分析与逐步判别）。但是，如果变量 x_1, x_2, \dots, x_p （在统计上）互不相关，那末，*Mahalanobis* 距离就成为我们所熟知的欧氏距离，当然，这种说法只有在 p 个变量都是标准化的时候是精确的。标准化见 § 3）。这时，第 k 个样品和第 l 个样品之间的距离：

$$\begin{aligned} d_{kl} &= \sqrt{(x_{1k} - x_{1l})^2 + (x_{2k} - x_{2l})^2 + \dots + (x_{pk} - x_{pl})^2} \\ &= \sqrt{\sum_{i=1}^p (x_{ik} - x_{il})^2} \end{aligned} \quad (2.2)$$

当 $p=2$ 时，距离公式 (2.2) 就是著名的勾股定理， d_{kl} 就是直角三角形的斜边之长（见图 2.1）。一般地， d_{kl} 表示 p 维长方体的对角线之长。

为了使得 d_{kl} 在一定的范围内变化，通常我们是以下面的公式代替公式 (2.2)：

$$d_{kl} = \sqrt{\frac{1}{N} \sum_{i=1}^p (x_{ik} - x_{il})^2}, \quad (2.3)$$

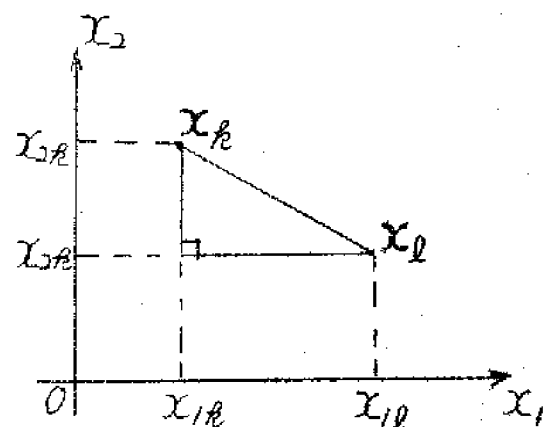


图 2.1



如果矩阵 X 的第 k 列和第 l 列对应的元素越接近，那末距离 d_{kl} 就越小。换言之，在第一种意义之下，如果距离 d_{kl} 越小，则第 k 个样品和第 l 个样品就越相似。

如果变量 x_1, x_2, \dots, x_p 之间彼此相关，那末，为了能够沿用公式 (2.3)，我们可以先对变量 x_1, x_2, \dots, x_p 进行因子分析（见地质学中多变量统计方法之一：因子分析与分量分析），求出主因子解或者其它正交因子解。然后只要以互不相关的因子去代替原始变量，以各个样品的因子得分去代替原始变量的取值即可。值得指出的是，这时因子的个数可以少于原始变量的个数，这是因为在那些方差贡献小的因子那里，各个样品之间显示不出多大的差异，所以这些因子对分析并不提供什么有用的信息，即使不考虑它们也无妨。

(ii) 夹角的余弦

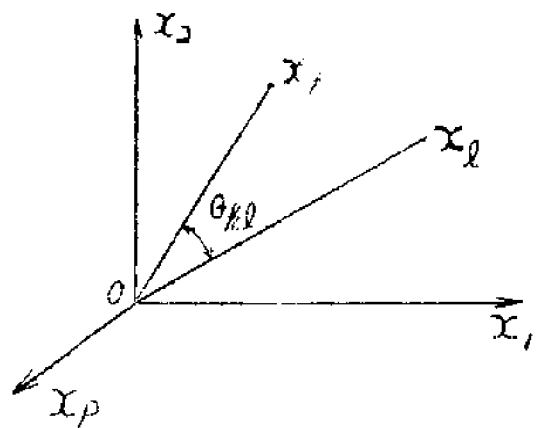
我们可以把第 k 个样品 $x'_k = (x_{1k}, x_{2k}, \dots, x_{pk})$ 和第 l 个样品 $x'_l = (x_{1l}, x_{2l}, \dots, x_{pl})$ 看成是 p 维欧氏空间的两个向量。于是在这两个向量之间就存在着一个夹角 θ_{kl} （见图 2.2）。这个夹角的余弦

$$\cos \theta_{kl} = \frac{\sum_{i=1}^p x_{ik} x_{il}}{\sqrt{\sum_{i=1}^p x_{ik}^2} \cdot \sqrt{\sum_{i=1}^p x_{il}^2}} \quad (2.4)$$

我们有

$$-1 \leq \cos \theta_{kl} \leq +1.$$

如果矩阵 X 的第 k 列和第 l 列对应的元素越接近于成比例，那末 $\cos \theta_{kl}$ 就越大。换言之，在第二种相似性意义之





下, 如果 $\cos \theta_{kl}$ 越大, 那末第 k 个样品和第 l 个样品就越相似。

(iii) 样品相关系数

第 k 个样品和第 l 个样品之间的样品相关系数规定为:

$$P_{kl} = \frac{\sum_{i=1}^p (x_{ik} - \bar{x}_k)(x_{il} - \bar{x}_l)}{\sqrt{\sum_{i=1}^p (x_{ik} - \bar{x}_k)^2 \cdot \sum_{i=1}^p (x_{il} - \bar{x}_l)^2}}, \quad (2.5)$$

其中

$$\bar{x}_k = \frac{1}{p} \sum_{i=1}^p x_{ik} \quad \text{和} \quad \bar{x}_l = \frac{1}{p} \sum_{i=1}^p x_{il}.$$

特别, 当 $\bar{x}_k = \bar{x}_l = 0$ 时, 我们有

$$P_{kl} = \cos \theta_{kl}.$$

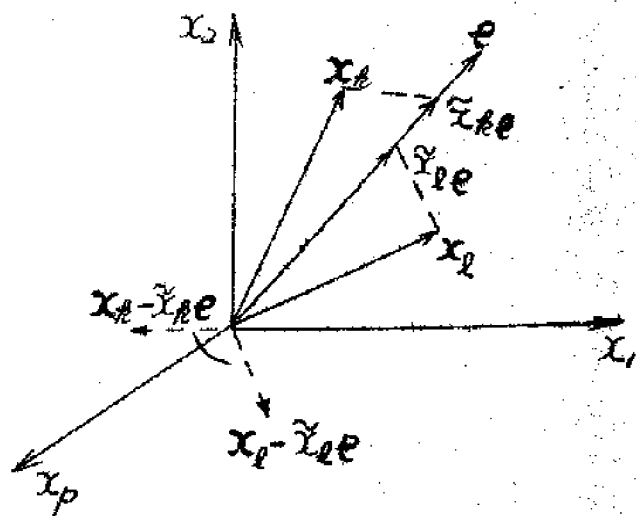
P_{kl} 满足不等式:

$$-1 \leq P_{kl} \leq +1.$$

如图 2.3 所示, 假定向量 $e' = (1, 1, \dots, 1)$, 则它位于等角线上。所谓等角线就是通过原点且与各个坐标轴具有相等的夹角的直线。这时 x_k 在 e 上的投影向量

$$\begin{aligned} x'_k(e) &= (\bar{x}_k, \bar{x}_k, \dots, \bar{x}_k) \\ &= \bar{x}_k e'. \end{aligned}$$

我们把 x_k 分解为两个向量 $\bar{x}_k e$ 和 $x_k - \bar{x}_k e$,





则前者就是上面已说过的 x_{el} 在 e 上的投影向量，后者就是 x_{el} 在过 e 且垂直于 e 的超平面 π 上的投影向量。

对 x_l 就不至复了。

由此可见，由 (2.5) 所确定的样品相关系数 p_{ell} 就是 x_{el} 和 x_l 在超平面上的投影向量 $x_{el} - \tilde{x}_{el} e$ 和 $x_l - \tilde{x}_l e$ 之间夹角 α 的余弦。

如果矩阵 X 的第 i 列和第 j 列对应的元素互长互消的关系越密切，那末 p_{ij} 的值就越大。换言之，在第三种相似性意义之下，如果 p_{ij} 的值越大，说明第 i 个样品和第 j 个样品就越相似。

在给出了样品之间相似性大小的数量指标之后，现在我们来定义变量之间相似性大小的数量指标。根据本节开头所述，变量之间的相似性和样品之间的相似性具有相同的含义，所不同的只是样品之间的相似性属于矩阵 X 的列之间，而变量之间的相似性却属于矩阵 X 的行之间。因此，第 i 个变量与第 j 个变量之间相似性大小的数量指标可以按照 d_{ij} ， $\cos \theta_{ij}$ 和 p_{ij} 依法炮制如下：

(i) 距离函数：

$$d_{ij}^* = \sqrt{\frac{1}{C} \sum_{k=1}^N (x_{ik} - x_{jk})^2},$$

其中 C 为一常数。

(ii) 夹角的余弦：

$$\cos \theta_{ij}^* = \frac{\sum_{k=1}^N x_{ik} x_{jk}}{\sqrt{\sum_{k=1}^N x_{ik}^2} \cdot \sqrt{\sum_{k=1}^N x_{jk}^2}},$$

它满足

$$-1 \leq \cos \theta_{ij}^* \leq +1.$$



$$p_{ij}^* = \frac{\sum_{k=1}^N (x_{ik} - \bar{x}_i)(x_{jk} - \bar{x}_j)}{\sqrt{\sum_{k=1}^N (x_{ik} - \bar{x}_i)^2 \cdot \sum_{k=1}^N (x_{jk} - \bar{x}_j)^2}},$$

它满足

$$-1 \leq p_{ij}^* \leq +1,$$

其中

$$\bar{x}_i = \frac{1}{N} \sum_{k=1}^N x_{ik} \quad \text{和} \quad \bar{x}_j = \frac{1}{N} \sum_{k=1}^N x_{jk}.$$

以上三个用来度量变量之间相似性大小的统计量, 其几何解释可以雷同 α_{kl} , $\cos \theta_{kl}$ 和 ρ_{kl} , 所不同的只是现在是在一个 N 维的欧氏空间中, N 个坐标轴分别表示 N 次观测; 所考虑的点或者向量 $(x_{i1}, x_{i2}, \dots, x_{iN})$ 和 $(x_{j1}, x_{j2}, \dots, x_{jN})$ 是矩阵 X 的第 i 行和第 j 行。

§3. 数据的标准化和正规化

由于下述两方面的原因:

(i) 各个变量的测量单位不一样;

(ii) 各个变量的测量单位虽然一样, 但是测量的结果数量级不一样。如果我们直接利用原始数据, 根据在第二节中例举的相似性度量进行点群分析, 那末各个变量就以不等的权参加了进来, 或者说对各个变量就不能一视同仁。为了使得各个变量有相等的权, 或者说能一视同仁各个变量, 就要对原始数据进行下述的标



(1) 标准化

对第 i 个变量进行标准化, 那就是令

$$x_{ih}^* = \frac{x_{ih} - \bar{x}_i}{\sigma_i}, \quad (i=1, 2, \dots, p; h=1, 2, \dots, N) \quad (3.1)$$

其中

$$\bar{x}_i = \frac{1}{N} \sum_{h=1}^N x_{ih} \quad \text{和} \quad \sigma_i = \sqrt{\frac{1}{N-1} \sum_{h=1}^N (x_{ih} - \bar{x}_i)^2}$$

通过变换 (3.1) 之后, 我们有

$$\bar{x}_i^* = \frac{1}{N} \sum_{h=1}^N x_{ih}^* = 0$$

和

$$\sigma_i^* = \sqrt{\frac{1}{N-1} \sum_{h=1}^N (x_{ih}^* - \bar{x}_i^*)^2} = 1,$$

也就是说, 新的变量 x_i^* 的数学期望和标准差分别等于 0 和 1。

(2) 正规化

对第 i 个变量进行正规化, 那就是令

$$x_{ih}^* = \frac{x_{ih} - \min_{1 \leq h \leq N} x_{ih}}{\max_{1 \leq h \leq N} x_{ih} - \min_{1 \leq h \leq N} x_{ih}}, \quad (i=1, 2, \dots, p; h=1, 2, \dots, N) \quad (3.2)$$

其中 $\min_{1 \leq h \leq N} x_{ih}$ 表示 $x_{i1}, x_{i2}, \dots, x_{iN}$ 中最小的一个, $\max_{1 \leq h \leq N} x_{ih}$

表示 $x_{i1}, x_{i2}, \dots, x_{iN}$ 中最大的一个。

通过变换 (3.2) 之后, 我们有

$$0 \leq x_i^* \leq 1.$$



§4. 谱系图的形成

互群分析的最终结果以谱系图（树状图）的形式直观地给出，使我们能一目了然。那末什么是谱系图，它又是怎样形成的呢？下面我们就来回答这个问题，为此，我们以仅型互群分析为例，并假定 $N=5$ ，相似性度量选用样品相关系数。

第一步：根据标准化或者正规化数据矩阵

$$X_{(i)}^* = \begin{pmatrix} x_{11}^* & x_{12}^* & \cdots & x_{15}^* \\ x_{21}^* & x_{22}^* & \cdots & x_{25}^* \\ \vdots & \vdots & \ddots & \vdots \\ x_{p1}^* & x_{p2}^* & \cdots & x_{p5}^* \end{pmatrix} \quad (4.1)$$

可以计算得初始样品相关矩阵

$$\rho^{(1)} = \begin{pmatrix} \rho_{11}^{(1)} & \rho_{12}^{(1)} & \cdots & \rho_{15}^{(1)} \\ \rho_{21}^{(1)} & \rho_{22}^{(1)} & \cdots & \rho_{25}^{(1)} \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{51}^{(1)} & \rho_{52}^{(1)} & \cdots & \rho_{55}^{(1)} \end{pmatrix}$$

其中 $\rho_{kl}^{(1)}$ ($k, l=1, 2, \dots, 5$) 表示第 k 个样品和第 l 个样品的样品相关系数。因为 $\rho^{(1)}$ 是一个对称矩阵，并且诸对角元素所表示的是相关（它们恒等于 1），所以，实际上只要计算 $\rho^{(1)}$ 的不包括对角线在内的上三角即可，也即

$$\rho^{(1)} = \begin{pmatrix} \rho_{12}^{(1)} & \rho_{13}^{(1)} & \rho_{14}^{(1)} & \rho_{15}^{(1)} \\ & \rho_{23}^{(1)} & \rho_{24}^{(1)} & \rho_{25}^{(1)} \\ & & \rho_{34}^{(1)} & \rho_{35}^{(1)} \\ & & & \rho_{45}^{(1)} \end{pmatrix}.$$



在以后各步中出现的样品相关矩阵也将与此相类似，都是些不包括对角线在内的上三角。

在 $\rho^{(1)}$ 的诸元素中选取最大者，假定为 $\rho_{24}^{(1)}$ 。那末说明在 5 个样品 x_1, x_2, \dots, x_5 中，样品 x_2 与 x_4 的相似性为最大。我们把样品 x_2 与 x_4 如图 4.1 那样用‘ \sqcap ’型的线联结起来，并且把它们组合成为一个新的样品，叫做组合样品，用 $x_2^{(2)} = (x_{12}^{(2)}, x_{22}^{(2)}, \dots, x_{p2}^{(2)})'$ 表示。我们有 $x_2^{(2)}$ 等于 x_2 与 x_4 的算术平均：

$$x_2^{(2)} = \frac{x_2 + x_4}{2}, \quad (4.2)$$

或者

$$x_{\lambda 2}^{(2)} = \frac{x_{\lambda 2}^* + x_{\lambda 4}^*}{2}, \quad (\lambda = 1, 2, \dots, p) \quad (4.3)$$

在记号 $x_2^{(2)}$ 中，下标表示组合样品的序号，这序号与组成它的两个样品中的较小的一个序号相同，上标表示组成组合样品的原样品的数目。

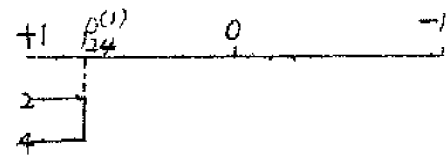


图 4.1

第二步：根据数据矩阵

$$X_{(2)}^* = \begin{pmatrix} x_{11}^* & x_{12}^{(2)*} & x_{13}^* & x_{15}^* \\ x_{21}^* & x_{22}^{(2)*} & x_{23}^* & x_{25}^* \\ \vdots & \vdots & \vdots & \vdots \\ x_{p1}^* & x_{p2}^{(2)*} & x_{p3}^* & x_{p5}^* \end{pmatrix}$$

可以计算得第二步样品相关矩阵

$$\rho = \begin{pmatrix} \rho_{12}^{(2)} & \rho_{13}^{(2)} & \rho_{15}^{(2)} \\ & \rho_{23}^{(2)} & \rho_{25}^{(2)} \\ & & \rho_{35}^{(2)} \end{pmatrix}$$



在 $\rho^{(2)}$ 的诸元素中选取最大者，假定的 $\rho_{12}^{(2)}$ 。那末说明在 4 个样品 $x_1, x_2^{(2)}, x_3, x_4$ 中，样品 x_1 与 $x_2^{(2)}$ 的相似性为最大。我们把 x_1 与 $x_2^{(2)}$ 如图 4.2

那样用 '□' 型的线联结起来，在这里，因为组合样品 $x_2^{(2)}$ 代表着两个原样品 x_2 和 x_4 ，所以在图 4.2 中实际上是把 x_1 与 x_2 ， x_4 联结在一起。并且由 x_1 与 $x_2^{(2)}$ 可得组合样品 $x_{1'}^{(3)} = (x_{1'}^{(3)}, x_{2'}^{(3)}, \dots, x_{p'}^{(3)})'$ ，我们有 $x_{(1)}^{(3)}$ 等于 x_1, x_2 和 x_4 的算术平均：

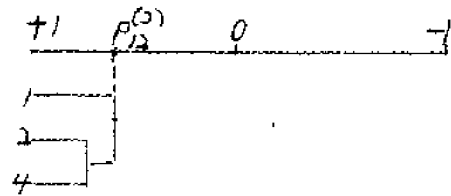


图 4.2

$$x_{1'}^{(3)} = \frac{x_1 + x_2 + x_4}{3} \quad (4.4)$$

或者

$$x_{\lambda'}^{(3)} = \frac{x_{\lambda 1}^* + x_{\lambda 2}^* + x_{\lambda 4}^*}{3}, \quad (\lambda = 1, 2, \dots, p) \quad (4.5)$$

公式 (4.4) 和 (4.5) 也可以分别写成以下的形式：

$$x_{1'}^{(3)} = \frac{x_1 + 2x_2^{(2)}}{3}, \quad (4.6)$$

$$x_{\lambda'}^{(3)} = \frac{x_{\lambda 1}^* + 2x_{\lambda 2}^{(2)}}{3}, \quad (\lambda = 1, 2, \dots, p) \quad (4.7)$$

由 (4.6) 和 (4.7) 两公式可知，如果原样品都赋权为 1，组合样品 $x_2^{(2)}$ 赋权为 2，这权数等于组成它的样品的权数之和或者组成它的原样品的数目。那末， $x_{1'}^{(3)}$ 就等于 x_1 与 $x_2^{(2)}$ 的加权平均，或者说 $x_{1'}^{(3)}$ 的第 λ 个分量 $x_{\lambda 1'}^{(3)}$ 就等于 x_1 的第 λ 个分量 $x_{\lambda 1}^{(2)}$ 与 $x_2^{(2)}$ 的第 λ 个分量 $x_{\lambda 2}^{(2)}$ 的加权平均。从这个意义上来说，公式 (4.2) 和 (4.3) 也可以看作是加权平均，只不过在那里， x_2 和 x_4 的权都是 1 罢了。显



然，在组合样品是由较多的原样品组成的情况下，进行再一次组合的时候，采用加权平均公式要比采用算术平均公式来的简单。

在 $x^{(3)}$ 中，下标和上标的意义同 $x^{(2)}$ 。

第三步：根据数据矩阵

$$x_{(3)}^* = \begin{pmatrix} x_{11}^{(3)} & x_{13}^* & x_{15}^* \\ x_{21}^{(3)} & x_{23}^* & x_{25}^* \\ \vdots & \vdots & \vdots \\ x_{p1}^{(3)} & x_{p3}^* & x_{p5}^* \end{pmatrix}$$

可以计算得第三步样品相关矩阵

$$\rho^{(3)} = \begin{pmatrix} \rho_{13}^{(3)} & \rho_{15}^{(3)} \\ & \rho_{35}^{(3)} \end{pmatrix}$$

在 $\rho^{(3)}$ 的诸元素中选取最大者，假定为 $\rho_{35}^{(3)}$ 。那么说明在 3 个样品 $x^{(3)}$ ， x_3 和 x_5 中，样品 x_3 与 x_5 的相似性为最大。我们把 x_3 与 x_5 如图 4.3 那样联结起来。并且由 x_3 与 x_5 可以得组合样品 $x_3^{(2)} = (x_{13}^{(2)}, x_{23}^{(2)}, \dots, x_{p3}^{(2)})'$ ，我们有 $x_3^{(2)}$ 等于 x_3 与 x_5 的加权平均(这时因为 x_3 与 x_5 的权都是 1，所以加权平均就是算术平均)：

$$x_3^{(2)} = \frac{x_3 + x_5}{2}$$

或者

$$x_{i3}^{(2)} = \frac{x_{i3}^* + x_{i5}^*}{2}$$

$$(i=1, 2, \dots, p)$$

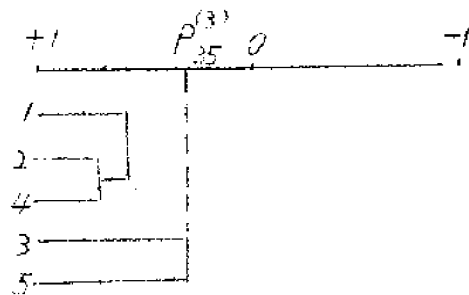


图 4.3



地信网

http://www.3s001.com

在 $X_3^{(2)}$ 中，下标和上标前意义同 $X_2^{(2)}$ 。

第四步：根据数据矩阵

$$X_{(4)}^* = \begin{bmatrix} X_{11}^{(3)} & X_{13}^{(2)} \\ X_{21}^{(3)} & X_{23}^{(2)} \\ \vdots & \vdots \\ X_{p1}^{(3)} & X_{p3}^{(2)} \end{bmatrix}$$

可以计算得第四步样品相关矩阵

$$\rho^{(4)} = \begin{bmatrix} & \rho_{13}^{(4)} \end{bmatrix}$$

在 $\rho^{(4)}$ 的不包括对角线在内的上三角中仅有唯一的元素 $\rho_{13}^{(4)}$ ，所以只要把 $X_1^{(3)}$ 与 $X_3^{(2)}$ 最后如叠 4.4 那样连结起来，就可以得到一张所谓的完整的谱系图。

如果我们选取的相似性水平（分类单位）为 ρ_0 ，它满足

$$\rho_{35}^{(2)} < \rho_0 < \rho_{13}^{(4)}$$

那末从叠 4.4 可以看出，这时联结样品 X_1, X_2, X_4 和 X_3, X_5 的线就断开，也就是说，点群分析的结果

，把样品 X_1, X_2, X_4 分为一组， X_3, X_5 分为另外一组。

现在我们把谱系图的形成过程其主要归纳如下：

(1) 原始数据前标准化或者正规化；

(2) 每一步都比前一步少了一个对象（样品或变量），每一步相似性矩阵的阶数都比前一步少 1；

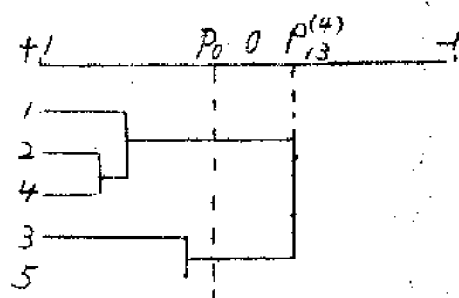


图 4.4



(3) 属对象的权数为 1，组合对象的权数等于组成它的两个对象的权数之和或者组成它的属对象的数目；

(4) 组合对象等于组成它的两个对象（它们本身可能是属对象，也可能是较早阶段形成的组合对象）的加权平均；

(5) 当把最后两个对象联结在一起，也即当所有的属对象都联结在一起时，谱系图便告完成。所以在一般情况下，要经过 $(N-1)$ 步。

应当指出，在互群分析中，由于相似性水平（分类单位）可大可小，所以分组可细可粗。这样一来，不仅可以得到组，还可以得到亚组，亚亚组（如果它们存在）。

另外还应当指出，相似性水平的使用不一定那末机械，主要还是要看谱系图的自然趋势。有的时候，从谱系图上看，存在若干组是很明显的，同时也能得到很好的地质解释。但是，如果硬用某一个相似性水平去机械地划分，却无论如何也不可能同时得到若干组。

5.5. 地质实例

（一）北祁连某超基性岩体铂族元素的 R 型互群分析。

为了探索北祁连某超基性岩体中铂族元素的赋存状态与聚集规律，我们共取用了 90 个样品，每个样品均作 Cr_2O_3 、 Ni 、 Au 、 Pd 、 Pt 、 Ir 、 Rh 、 Os 、 Ru 的试金分析，共 10 项。这样我们获得了一个 90×10 的原始数据矩阵，以它为基础进行 R 型互群分析。



地信网

<http://www.3s001.com>