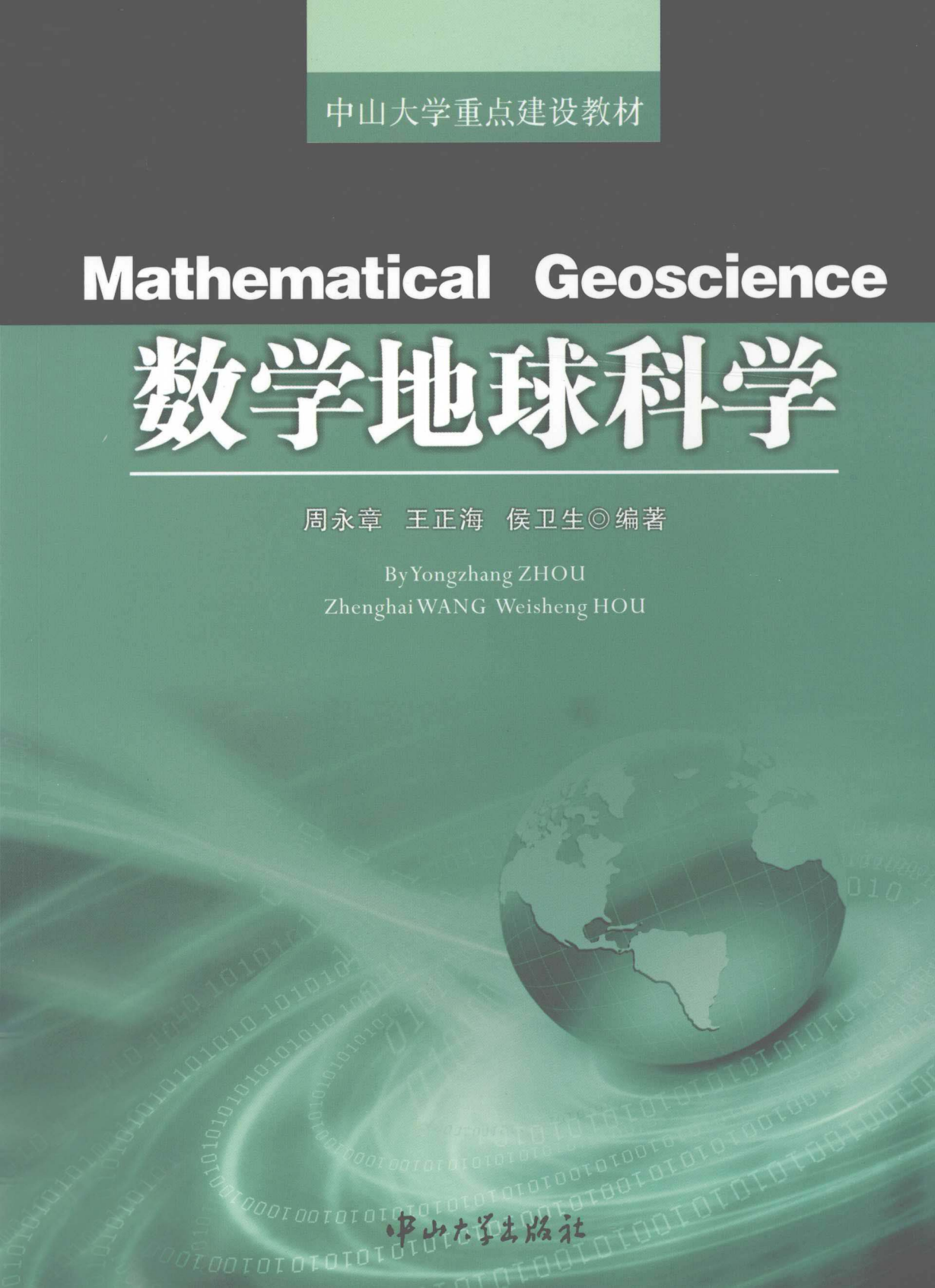
中山大学重点建设教材

Mathematical Geoscience

数学地球科学

周永章　王正海　侯卫生◎编著

By Yongzhang ZHOU
Zhenghai WANG Weisheng HOU



中山大学出版社

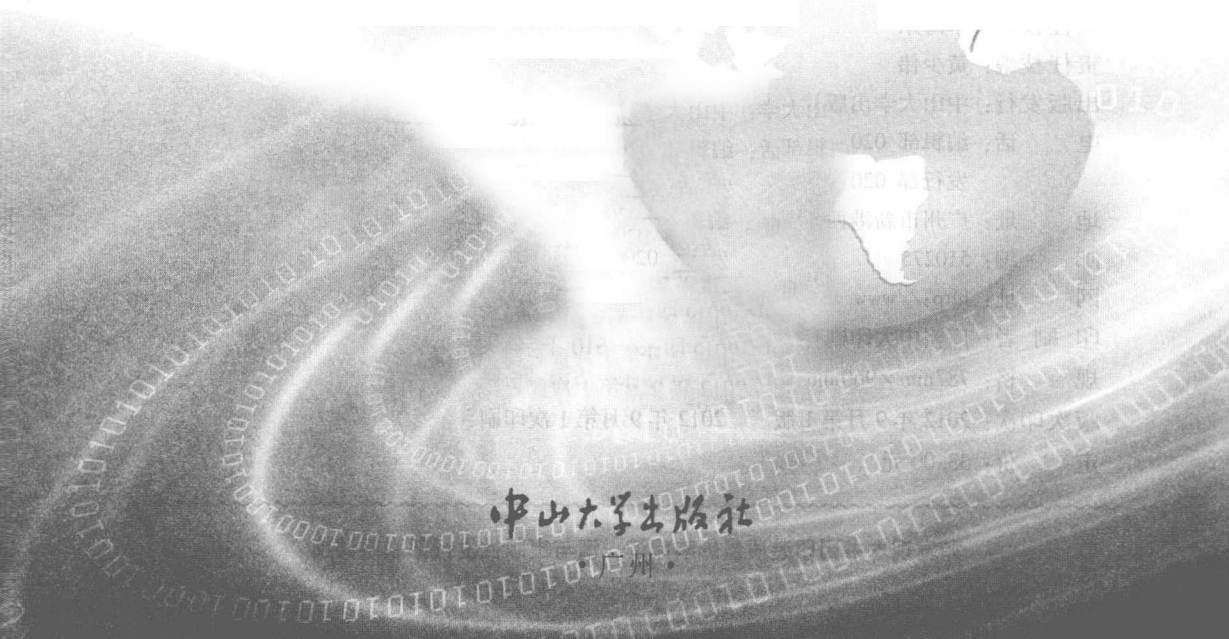
教材

Mathematical Geoscience

数学地球科学

周永章　王正海　侯卫生◎编著

By Yongzhang ZHOU
Zhenghai WANG Weisheng HOU



中山大学出版社
广州

图书在版编目（CIP）数据

数学地球科学/周永章，王正海，侯卫生编著．—广州：中山大学出版社，2012.9

ISBN 978-7-306-04320-7

Ⅰ．①数…　Ⅱ．①周…②王…③侯…　Ⅲ．①数字技术—应用—地球科学—高等学校—教材　Ⅳ．①P-39

中国版本图书馆 CIP 数据核字（2012）第 221304 号

出 版 人：祁　军
策划编辑：李海东
责任编辑：李海东
封面设计：曾　斌
责任校对：李海东
责任技编：黄少伟
出版发行：中山大学出版社
电　　话：编辑部 020-84114366，84111996，84113349，84111997，84110779
　　发行部 020-84111998，84111981，84111160
地　　址：广州市新港西路 135 号
邮　　编：510275　　传　真：020-84036565
网　　址：http://www.zsup.com.cn　　E-mail:zdcbs@mail.sysu.edu.cn
印 刷 者：广州中大印刷有限公司
规　　格：787mm×960mm　1/16　16 印张　310 千字
版次印次：2012 年 9 月第 1 版　　2012 年 9 月第 1 次印刷
定　　价：33.00 元

序　言

数学地球科学（Mathematical Geosciences），早期又称数学地质学（Mathematical Geology）。进入21世纪后，国际数学地质学会（International Association for Mathematical Geology）更名为国际数学地球科学协会（International Association for Mathematical Geosciences），故当下学界把数学地质学称为数学地球科学。

数学地球科学是20世纪50年代以来迅速形成的一门交叉学科。1968年，国际数学地质学会（IAMG）正式成立；1969年，学术期刊 *International Journal of Mathematical Geology* 正式出版。至1978年，每年的数学地球科学论文超过1000篇，这说明数学地球科学发展迅速，标志着数学地球科学作为一门新兴学科在国际上得到认可。

数学地球科学是数学及电子计算机与地球科学相结合的产物，目的是从量的方面研究和解决地球科学问题。它的出现适应了地球科学从定性的描述阶段向定量研究发展的需求，为地球科学开辟了新的发展途径。数学地球科学方法的应用范围极其广泛，几乎渗透到地球科学的各个领域。

目前，数学地球科学的基本内容或方法有：①地质数据的统计分析，其中常用的有趋势面分析、回归分析、因子分析、判别分析、聚类分析、典型相关分析、克里格法、时间序列分析、数字滤波等；②地质过程的数字模拟，如地下水运动过程模拟、构造断裂模拟、矿物地球化学模拟、概率性数学模拟（如地层剖面的马尔科夫过程模拟）等；③地质数据储存、索取、自动处理和显示，如野外地质数据处理系统，矿产资源、地下水资源数据处理系统，各种专用地质数据处理系统，自动绘图系统等。

早期优先发展的数学地球科学是狭义的，主要指建立、检验和解释地质过程概念的随机模型的总称。目前数学地球科学主要以地球科学为基础，以数学为工具，以电子计算机为技术手段，以解决地质问题为目的。而广义的数学地球科学指数学在地球科学中的应用，即用数学的理论、方法、思维方式研究和解决地质问题。

本书从地学信息的处理、应用和表达三个方面来阐述数学地球科学领域的相关知识。其中，第1章至第5章介绍多元统计分析，第6章介绍地质统计学，第7章和第8章分别介绍高光谱遥感岩矿信息定量提取和基于GIS的综合

信息定量预测，第9章至第13章介绍覆盖数据管理、数据处理分析、数据应用、空间建模及可视化的整个地质数据生命周期的相关理论和技术。

本书在中山大学地球科学系以及国际数学地球科学协会中山大学学生分会（IAMG Student Chapter at SYSU）多年内部试用教材的基础上编写而成。本书的主要分工如下：周永章负责总体框架设定、审定及第1章至第5章的撰写，王正海负责第6章至第8章的撰写，侯卫生负责第9章至第13章的撰写。在试教和写作过程中，参考了大量自20世纪70年代以来出版的经典数学地质教材，并曾与赵鹏大团队以及侯景儒、吴冲龙、鲍征宇等同行专家开展过有意义的交流。SPSS、Surfer、MATLAB 、Excel 等常用软件的普及大大提高了本书的实用价值和学生学习的热情。本书适用于地球科学类专业高年级本科生、各级研究生及科研人员学习和参考。

目　录

第1章　地球化学变量

地学变量是地学研究对象中能够观察到的某种可以度量的特征。地学数据是地学变量的观测值，具有来源丰富、数据量庞大、数据结构复杂、主题多样、格式多样的特点。为了便于学习，本章仅仅讨论地球化学变量或类地球化学变量。

1.1　地球化学变量的特点

在地球化学研究中，无论是对某一岩体或矿体的地球化学特征的分析，抑或是进行区域性的金属和油气地球化学勘查、区域环境评价等，总是要确定若干个能反映研究对象(岩体、矿体、区域等)的基本地球化学特征的量。元素含量和有机组分的含量、成岩成矿的温度和压力、溶液的 pH、Eh、孔隙度、矿体厚度等都是常见的地球化学变量。通过采样(采集岩石、土壤或水样)并进行各种测试分析，可以获得地球化学变量的样本值。

地球化学变量具有下列基本特征：

(1)地球化学变量具有随机性，是随机变量。地球化学变量的随机性表现为：地球化学样品的采取具有抽样的性质；从成因上看，地球化学变量的取值受多种因素的控制；测试分析误差普遍存在。

(2)地球化学变量具有统计规律性。例如，在许多矿区元素含量具有明显的分带性。有学者把这种地球化学变量的时空结构性称为地球化学场(於崇文，1980)。

1.2　数据矩阵及基本统计量

记 p 个地球化学变量为 x_1，x_2，…，x_p，构成 p 维向量 $\boldsymbol{x}=(x_1, x_2, \cdots, x_p)$。采 n 个样品，分析每个样品的这 p 个变量的值，可以得到如下数据表格(表 1.1)。

表 1.1 样品的变量值

样品	x_1	x_2	…	x_p
样品 1	x_{11}	x_{12}	…	x_{1p}
样品 2	x_{21}	x_{22}	…	x_{2p}
⋮	⋮	⋮	⋮	⋮
样品 n	x_{n1}	x_{n2}	…	x_{np}

这些数据亦经常表示成数据矩阵的形式：

$$\boldsymbol{X}=\begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \vdots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix}=(x_{ij})_{n\times p}。$$

这里 x_{ij} 为第 i 个样品的第 j 个变量的测试数据。

在数据矩阵 $\boldsymbol{X}$ 中，每一行都是一个 p 维行向量，通常记为 $\boldsymbol{x}_{(1)}$，$\boldsymbol{x}_{(2)}$，…，$\boldsymbol{x}_{(n)}$，对应于各个样品的 p 个地球化学指标(变量)值；每一列则为一个 n 维列向量，对应于某个变量的 n 个测试分析值，通常仍以变量名 x_1，x_2，…，x_p 记之。

地球化学数据矩阵可以有两种空间表示方法：

(1)以变量名为坐标轴构成 p 维空间，每个样品为空间中的一个点(即样本点)，即数据矩阵 $\boldsymbol{X}$ 中各行向量表示的点。p 维变量空间中的 n 个样本点之间的距离远近反应了各样本点之间的亲疏关系，可以据此进行样品之间的相关性分析和分类等，称为 Q 型分析。

(2)以样品名为坐标轴构成 n 维空间，每个变量为空间中的一个点(即变量点)。n 维样品空间中的 p 个变量点之间的距离远近反应了各变量点之间的亲疏关系，可以据此进行变量之间的相关性分析和分类等，称为 R 型分析。

各变量的均值和方差是重要的统计特征，分别记为 $\bar{x}_j$ 和 s_j^2 $(j=1,\ 2,\ \cdots,\ p)$。均值即样本分析值的平均值，是该变量值的“中心”；方差是变量相对于该“中心”偏离程度的一种度量。均值和方差的计算公式分别为：

$$\bar{x}_j=\frac{1}{n}\sum_{i=1}^{n}x_{ij},$$

$$s_j^2=\frac{1}{n}\sum_{i=1}^{n}(x_{ij}-\bar{x}_j)^2=\frac{1}{n}\sum_{i=1}^{n}x_{ij}^2-n\bar{x}_j^2。$$

方差的平方根称为均方差。

变量 j 与变量 k 之间的协方差记为 s_{jk}，是变量 j 与变量 k 分别相对于其均

值偏差的交叉乘积之平均值：

$$s_{jk} = \frac{1}{n}\sum_{i=1}^{n}\left[(x_{ij} - \bar{x}_j)(x_{ik} - \bar{x}_k)\right] = \frac{1}{n}\sum_{i=1}^{n} x_{ij}x_{ik} - n\bar{x}_j\bar{x}_k。$$

可见，如果在统计的意义上，当同一样品中变量 j 的值相对于其平均值较大时，变量 k 的值相对于其平均值也较大，变量 j 的值相对于其平均值较小时，变量 k 的值相对于其平均值也较小，即变量 j 和变量 k 有相同或相近的变化规律，上式中交叉乘积的两项或都为正或都为负，结果为正值，称变量 j 与变量 k 正相关；反之，当同一样品中变量 j 的值相对于其平均值较大时，变量 k 的值相对于其平均值较小，变量 j 的值相对于其平均值较小时，变量 k 的值相对于其平均值较大，则上式中交叉乘积多为负，称变量 j 与变量 k 负相关。两变量完全相同时，其协方差最大，即为同一变量的方差。

协方差矩阵为各变量之间协方差组成的矩阵：

$$\boldsymbol{S} = \begin{bmatrix} s_{11} & s_{12} & \cdots & s_{1p} \\ s_{21} & s_{22} & \cdots & s_{2p} \\ \vdots & \vdots & & \vdots \\ s_{p1} & s_{p2} & \cdots & s_{pp} \end{bmatrix} = (x_{jk})_{p\times p}。$$

显然协方差矩阵 $\boldsymbol{S}$ 为对称矩阵，即有 $s_{jk} = s_{kj}$，因为变量 j 与变量 k 之间的协方差当然等于变量 k 与变量 j 之间的协方差。还可以证明协方差矩阵 $\boldsymbol{S}$ 为正定矩阵。

虽然可以根据变量间的协方差的正负来判断两变量之间是正相关还是负相关，但协方差的大小显然与变量 xj 和 xk 的单位或量纲有关，不能据此判断两变量之间相关性的大小。为此可以用两变量的均方差进行“标准化”，类似于物理中的无量纲化，称为变量 k 与变量 j 的相关系数，记为 r_{jk}：

$$r_{jk} = \frac{s_{jk}}{s_j s_k} = \frac{\frac{1}{n}\sum_{i=1}^{n} x_{ij}x_{ik} - n\bar{x}_j\bar{x}_k}{\sqrt{\frac{1}{n}\sum_{i=1}^{n} x_{ij}^2 - n\bar{x}_j^2}\sqrt{\frac{1}{n}\sum_{i=1}^{n} x_{ik}^2 - n\bar{x}_k^2}}。$$

相关系数 r_{jk} 的值在 -1 与 $+1$ 之间，愈接近于 1，则变量 k 与变量 j 之间的正相关性愈好，高则均高，低则均低，同步协调；愈接近于 -1，则变量 k 与变量 j 之间的负相关性愈好，此高彼低；愈接近于 0，则变量 k 与变量 j 之间的相关性愈弱，彼此的高低取值互不影响，大抵是“独立行事”的。

假设各变量的均值为零。如果变量的均值非零，可以将样本值都减去该变量的平均值，即作变换：

$$x'_{ij} = x_{ij} - \bar{x}_j。$$

于是有：

$$s_{jk} = \frac{1}{n}\sum_{i=1}^{n} x_{ij}x_{ik} \quad 或 \quad S = X'X。$$

1.3 数据预处理

许多统计方法是针对变量服从某种分布(如正态分布)而言的。数据预处理就是通过变量变换，使变量服从某种分布。此外，地球化学观测数据中经常存在个别特异样本点，特别是异常高值点，称为离群样品(outlier)。数据预处理的另一作用是为了消除这些离群样品(outlier)的过强影响。

介质校正、批次差校正等也是数据变换的方法，其目的是尽可能消除数据误差，使统计的规律性更为清晰。

最常用的数据预处理方法是变量规一化。

地球化学原始数据中，各变量的量纲不同，其值的量级差异很大，如地质体中的 Al、Si、Fe 等常量元素的含量为 $10^{-2}\sim 10^{-1}$(百分含量)，Cu、Pb、Zn 的含量为 $10^{-6}\sim 10^{-4}$，Au 的含量则为 10^{-9}，这会给后面所要讨论的各种统计方法带来困难。事实上，以 Q 型分析为例，我们已经知道，Q 型分析的基础是在变量坐标构成的空间内考察各样本点之间的亲疏关系，显然要求各坐标的“刻度”，即各变量的量纲是一致的。大多数变量变换的主要目的是消除各变量数量级的差异，使各变量在统计分析中大致起相同的作用。

极差变换、标准化变换和对数变换是三种最常用的变量规一化方法。

1.3.1 极差变换

变换后的每个变量 $x_j(j=1, 2, \cdots, n)$ 的各样本值 x'_{ij} 为原值 x_{ij} 减去该变量的最小值后除以该变量最大值与最小值之差：

$$x'_{ij} = \frac{x_{ij} - x_{j\min}}{x_{j\max} - x_{j\min}}。$$

变换后 $0 \leqslant x'_{ij} \leqslant 1$，实现了变量数量级的规一化。

1.3.2 标准化变换

变换后的每个变量 $x_j(j=1, 2, \cdots, n)$ 的各样本值 x'_{ij} 为原值 x_{ij} 减去该变量的均值后除以该变量的标准差：

$$x'_{ij} = \frac{x_{ij} - \bar{x}_j}{s_j}。$$

变量标准化使变量的均值为 0，方差为 1。特别是当可以假定原各变量服从正态分布时，变换后的新变量均服从 $N(0, 1)$ 分布，不但能很好地起到变量规一化的作用，也便于理论分析。

1.3.3　对数变换

通常变量取以 10 为底的对数：

$$x'_{ij} = \ln x_{ij}。$$

取对数后大大减小了数量级的差异，如原相差 10 倍的两个变量，经对数变换后只相差 1。另外，许多研究者认为，许多地球化学变量，特别是微量元素服从对数正态分布，即取对数后能使各变量服从正态分布。

对数变换属于非线性变换。根据变量分布的特点，还可以有其他各种非线性变换。

除变量规一化方法外，对离群样本点的数据预处理方法还有非线性映射和局域化数据预处理等。

第 2 章　聚类分析

聚类分析(cluster analysis)的功能是建立一种分类方法，将一批样品或变量，按照它们在性质上的相似程度进行分类，把相似程度大的并成一类，而把相似程度小的分为不同的类。

在前面一章中已经指出，地球化学数据矩阵可以有两种空间表示方法：p 维变量空间中的 n 个样本点和 n 维样本空间中的 p 个变量点。相应地，在样本空间中是对变量进行分类，称 R 型聚类分析；在变量空间中是对样本进行分类，称 Q 型聚类分析。

2.1　距离与其他相似性系数

2.1.1　距离

两点间的距离(distance)是表征两空间点之间“亲疏”关系的最直接、最自然的度量。在三维空间中，两点间的距离的平方为各坐标值差的平方和，称为欧几里得距离(Euclidean distance)。在高维的抽象空间中，点 i 和点 j 之间的距离 d_{ij} 可以有各种不同的定义，只要其满足距离公理：

(1)对一切的 i，j，$d_{ij} \geqslant 0$；

(2)$d_{ij}=0$ 等价于点 i 和点 j 为同一点，即 $x_{(i)}=x_{(j)}$；

(3)对一切的 i，j，$d_{ij}=d_{ji}$；

(4)三角不等式成立，即对一切的 i，j，k，有 $d_{ij} \leqslant d_{ik}+d_{kj}$。

据此，可以定义下列距离。

2.1.1.1　绝对值距离

绝对值距离的公式为：

$$d_{ij}(1)=\sum_{k=1}^{p}|x_{ik}-x_{jk}|。\tag{2.1}$$

根据在第 1 章中对数据矩阵的约定，这里 d_{ij} 显然是变量空间中样本 i 与样本 j 之间的距离，适用于样本分类，即 Q 型聚类分析。事实上聚类分析主要是

Q 型分析。如果欲进行 R 型分析，则相应的公式为：

$$d_{ij}(1) = \sum_{k=1}^{n} |x_{ki} - x_{kj}|。 \tag{2.2}$$

这在后面也类似，不再一一说明。

2.1.1.2 欧氏距离

欧氏距离的公式为：

$$d_{ij}(2) = \left[\sum_{k=1}^{p} (x_{ik} - x_{jk})^2 \right]^{1/2}。 \tag{2.3}$$

事实上，上述两种距离属于明氏(Minkowski)距离：

$$d_{ij}(q) = \left[\sum_{k=1}^{p} (x_{ik} - x_{jk})^q \right]^{1/q}。 \tag{2.4}$$

当 $q=1$，2 时的特例；当 q 趋于无穷大时，则为切比雪夫距离：

$$d_{ij}(\infty) = \max_{1 \leqslant k \leqslant p} |x_{ik} - x_{jk}|。 \tag{2.5}$$

2.1.1.3 马氏距离

欧氏距离没有考虑变量之间的相关性。马氏(Mahalanobis)距离对此进行改进，其定义为：

$$d_{ij}(M) = (x_{(i)} - x_{(j)})' \boldsymbol{S}^{-1} (x_{(i)} - x_{(j)})。 \tag{2.6}$$

式中：$\boldsymbol{S}$ 为数据矩阵的协方差矩阵，$\boldsymbol{S}^{-1}$ 为 $\boldsymbol{S}$ 的逆矩阵。

理论上讲，距离公理保证不同定义的距离都能表征空间点之间的相对“远近”关系，也就是说都能用于空间点群的划分。在实际应用中一般多采用欧氏距离。

距离均为正值，以距离越小表征两空间点愈相近，应归为同一类。

在计算空间点之间的距离前，必须对变量进行量纲的规一化(见第 1 章)，否则数量级小的变量在距离公式中基本不起作用。

2.1.2 其他相似性度量

除距离外，还有其他相似性度量可以表征空间点群之间的“亲疏”性。

2.1.2.1 相关系数

距离系数主要用于 Q 型分析，相关系数主要用于 R 型分析。从第 1 章知道，变量 x_j 与变量 x_k 之间的“亲疏”性的一个自然的度量是两变量的相关系数：

$$r_{jk}=\frac{\frac{1}{n}\sum_{i=1}^{n}x_{ij}x_{ik}-n\bar{x}_j\bar{x}_k}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}x_{ij}^2-n\bar{x}_j^2}\sqrt{\frac{1}{n}\sum_{i=1}^{n}x_{ik}^2-n\bar{x}_k^2}}。\tag{2.7}$$

相关系数的值域为(-1, 1)，其值越大，即越接近于1，则相关性愈好，被认为两空间点愈相似，应归为同一类。

2.1.2.2 夹角余弦

两空间点的"亲疏"程度除用距离表征外，还可以用两空间点所成的矢量间夹角的大小来反映。在样本空间中两变量向量 $\boldsymbol{x}_j$ 和 $\boldsymbol{x}_k$ 的夹角余弦为两向量的内积并为向量长度所标定：

$$\cos\theta_{jk}=\frac{\boldsymbol{x}_j\boldsymbol{x}_k}{|\boldsymbol{x}_j||\boldsymbol{x}_k|}=\frac{\boldsymbol{x}_j'\boldsymbol{x}_k}{|\boldsymbol{x}_j||\boldsymbol{x}_k|}=\frac{\sum_{i=1}^{n}x_{ij}x_{ik}}{\sqrt{\sum_{i=1}^{n}x_{ij}^2}\sqrt{\sum_{i=1}^{n}x_{ik}^2}}。\tag{2.8}$$

与相关系数比较可以发现，如果两变量的均值为0，则两变量的夹角余弦等于两变量的相关系数。

在变量空间中两样本向量之间的夹角余弦可以类似给出。夹角余弦的值域为(-1, 1)，其值越大，即越接近于1，则夹角愈小，被认为两空间点愈相似，应归为同一类。

2.2 系统聚类法

依据前面定义的能表征空间点之间亲疏关系的相似性度量，人们可以进行空间点群的分类。为简便起见，设定用前面介绍的某种距离，如欧氏距离作为相似性度量，则系统聚类的步骤为：

(1)将每个样看成一类，此时共有 n 类；

(2)计算类与类之间的距离，合并距离最近的两个类；

(3) 重复步骤(2)，直至所有样品归为一类。

上述过程结果可以通过绘制系统聚类谱系图来直观表征。

由于类与类之间的距离可以有不同的定义，就产生了不同的系统聚类法。

2.2.1 最短距离法

定义类 G_q 与类 G_r 之间的距离为所有 G_q 中的点与所有 G_r 中的点最近的点对的距离，其数学表述为：

$$D_{qr} = \min_{x_{(i)} \in G_q, x_{(j)} \in G_r} d_{ij}。 \tag{2.9}$$

当采用相关系数或夹角余弦等作为相似性度量时，上式中的 min 应为 max。

例 2.1 设有如下数据：

表 2.1 某地超基性岩 Cu－Ni 硫化物矿床地球化学数据

样号	样品组成	岩性及矿化	元素含量					
			Ni	Co	Cu	Cr	S	As
1	A 孔 16 个样平均	蛇纹岩，A 组矿化	1903	273	160	1178	8163	4
2	B 孔 11 个样平均	蛇纹岩，无矿化	2328	79	6	3175	586	14
3	C 孔 11 个样平均	蛇纹岩，无矿化	744	26	1	841	425	3
4	D 孔 12 个样平均	滑镁岩，B 组矿化	2782	273	150	2400	8231	37
5	E 孔 10 个样平均	滑镁岩，无矿化	1775	94	13	3140	54	1
6	F 孔 12 个样平均	滑镁岩，无矿化	1046	44	6	2093	104	4

对表 2.1 的数据进行变量分类。为消除各变量量级上的差异，对原始数据取以 10 为底的对数，结果如表 2.2 所示。

表 2.2 表 2.1 中数据对数变换结果

样号	Ni	Co	Cu	Cr	S	As
1	3.2794	2.4362	2.2041	3.0711	3.9118	0.6021
2	3.3670	1.8976	0.7782	3.5017	2.7679	1.1461
3	2.8716	1.4150	0	2.9248	2.6284	0.4771
4	3.4444	2.4362	2.1761	3.3802	3.9155	1.5682
5	3.2492	1.9731	1.1139	3.4969	1.7324	0
6	3.0195	1.6435	0.7782	3.3208	2.0170	0.6021

用相关系数作为相似性统计量，得如下相关系数矩阵：