目　录

绪论……（1）
第一章　地球化学变量……（7）
第一节　地球化学变量的特点及分类……（7）
第二节　变量的选择……（9）
第三节　变量的变换……（13）
第四节　数据集合的两种空间表示……（17）
第二章　聚类分析……（19）
第一节　一般概念……（19）
第二节　相似性度量……（19）
第三节　点群的谱系簇分……（23）
第四节　计算实例……（24）
第五节　逐步计算成图法……（28）
第六节　聚类分析在地质上的应用……（29）
第七节　聚类分析中一些问题的讨论……（34）
第八节　其它几种聚类分析方法……（36）
第三章　多元回归分析……（47）
第一节　基本概念……（47）
第二节　多元线性回归方程的建立……（47）
第三节　各自变量对回归效果的贡献……（52）
第四节　回归分析的一般步骤……（54）
第五节　最优回归方程的选择……（56）
第六节　几个问题讨论……（65）
第七节　回归分析在地球化学中的应用……（67）
第四章　趋势分析……（70）
第一节　概述……（70）
第二节　多项式趋势分析的数学方法……（70）
第三节　趋势方程的拟合度及显著性检验……（75）
第四节　几个问题的讨论……（76）
第五节　规则测网趋势面分析的数学方法……（78）
第六节　趋势分析在地球化学中的应用……（81）
第五章　判别分析……（84）
第一节　概述……（84）
第二节　费歇准则下的两类线性判别……（85）
第三节　两类判别计算实例……（91）

第四节　贝叶斯准则下的多类线性判别……(95)
第五节　逐渐判别分析……(103)
第六节　判别分析在地质上的应用……(110)
第七节　判别分析中应注意的问题……(112)
第六章　因子分析……(114)
第一节　概述……(114)
第二节　因子分析的数学模型……(115)
第三节　主成分分析……(121)
第四节　正交旋转——正交因子解……(123)
第五节　斜交旋转——斜交因子解……(126)
第六节　因子计量……(131)
第七节　因子分析的计算步骤与实例……(133)
第八节　Q型因子分析……(137)
第九节　因子分析在地球化学工作中的应用……(139)
第十节　因子分析的若干问题和讨论……(141)
第七章　对应分析和非线性映射……(143)
第一节　什么是对应分析……(143)
第二节　对应分析的数据变换方法……(143)
第三节　对偶定理……(144)
第四节　对应分析因子解……(145)
第五节　对应分析计算步骤……(146)
第六节　应用实例*……(147)
第七节　非线性映射简介……(150)
第八章　典型相关分析……(155)
第一节　概述……(155)
第二节　总体中的典型相关和典型变量……(155)
第三节　典型相关系数的显著性检验……(159)
第四节　用相关矩阵代替协方差矩阵的情形……(160)
第五节　应用实例*……(160)
第六节　问题与讨论……(163)
第九章　马尔科夫概型分析……(165)
第一节　基本概念……(165)
第二节　马尔科夫链的转移概率……(166)
第三节　连续参数非平稳马尔科夫过程……(169)
第四节　多元时空序列转移矩阵U……(173)
第五节　转移矩阵U的谱分解……(177)
第六节　过程的演化趋势……(179)
第十章　地质统计学(简介)……(184)
第一节　概述……(184)

第二节　区域化变量理论……(185)
第三节　变异函数及结构分析……(188)
第四节　线性克里格法……(194)
第十一章　稳健统计学* ……(198)
第一节　概述……(198)
第二节　稳健统计学中的几个概念……(198)
第三节　位置和尺度的稳健估计……(200)
第四节　方差的稳健估计……(206)
附录Ⅰ　线性代数初步知识……(209)
附录Ⅱ　数理统计常用表……(224)
主要参考文献……(234)

绪　论

近三十年来，在电子计算机应用的促进下，多元分析在地质地球化学领域中得到了迅速发展，对地球科学的定量化、地质数据的自动化处理、地质过程的计算机模拟以及地质思维的智能化等方面起到很大的促进作用，有力地推动了古老的地球科学与现代科学的结合。可以说多元分析方法的应用已经远远超出数据处理的范畴，已发展成为地质和化探工作中的一种富有生命力的新的研究手段，从而使地球科学在理论上和方法上产生了重大的变革。

一、地球化学中的多元分析的涵义

地球化学中的多元分析属于数学地质的一个分支。它以多元统计分析为主要手段，以电子计算机为工具来研究地球化学及化探问题，将数学与地球化学密切结合，使地球化学数学化、定量化，以便用计算机进行研究。

这里说的多元分析即多变量统计分析，它研究的对象是多变量的数据集合。这种数据集合一般由n个个体，每一个个体都有p个变量的观测值构成。形式上，可以把多元分析作为数理统计学在研究客观事物中多种因素间相互依赖的统计规律性的一个分支。多元分析主要讨论以下几方面问题：

1．讨论数据集合中变量与变量、样品与样品、变量与样品之间的相互关系或依赖关系。

2．对变量或样品进行分组、归类或判别。

3．研究数据集合的空间展布和时间序列变化规律。

4．简化数据结构。把具有内在联系的几个变量变换为一个新的独立变量或者约简数据集合的空间维数，用较简单的形式来表示所研究的复杂的多元数据集合。

5．提出假设和检验假设。主要是探讨以上四个方面所得结果的有效性、可靠性或可信性。

较常用的多元分析方法有：相关分析与回归分析、聚类分析、判别分析、趋势面分析、因子分析、对应分析、典型相关分析、马尔科夫概型分析和地质统计学。其中，回归分析、因子分析、对应分析、典型相关分析是研究变量和样品属性及依赖关系；聚类分析、判别分析等是对样品进行分组和归类；而趋势面分析、地质统计学及马尔科夫概型分析是探讨变量的时空展布。

地球化学中的多元分析的任务除了介绍这些多元分析的数学方法、原理以外，还要讨论它们在地球化学工作中的应用条件，所能解决地球化学问题的范围以及所计算得到数学结果的地球化学解释等问题。

二、迅速发展的原因

地球化学中的多元分析是现代地球化学、数学和电子计算机技术迅速发展的必然结果。从现代地球化学和地球化学勘查的发展趋势来看，主要表现为以下3个方面：

1．从经验到理论

在地球化学特别是地球化学探矿工作中，首先得到大量的事实规律。这些从实践中归纳出来的事实规律属于经验，需要上升为理论，用理论来解释它们，并反过来用于指导地球化学找矿的实践，目前很多经验性的事实规律已经被地球化学的理论所解释，但也有不少问题还有待于理论上的深化。可以说现在正面临着地球科学飞跃的前夕。在这个飞跃中，数学也将起到它应有的作用。

2．由定性走向定量

地球科学与其他自然科学相比，更多地使用定性的描述。研究人员根据自己所观察到的地质现象作人工的归纳、演译和推理。长期以来地质工作者较多地使用文字和各种图件来记录、反映地质现象。这样使得地质工作的野外工作量大，而且人为性强。对同一地质现象在不同研究者的文字描述中不容易作数量上的对比，因而也不利于地质工作的自动化。随着科学的发展，人们愈来愈强烈地认识到“一种科学仅当它成功地运用数学时，才算达到了完善的程度”。数学这门研究现实世界数量关系与空间形式的科学近年来很快地渗透到地质科学中，人们愈来愈渴望用简明的数学关系式或图表来表征各种地质现象和地质过程。正如S·C·鲁宾逊所说：“地质数据若用数学方法整理、核对、综合、计算和检验，则对地质解释、分类、概念检验、问题解答以及概率预报，能提供较之只限于任何个人的经验和记忆更为全面和客观的基础。而且，这样的基础可以作为许多地质人员的共同基础，他们每个人再加上自己的经验和知识去完成他们自己的解释工作。在推进地质学这门科学向前发展的过程中，这种共同基础正好可以成为一个主要因素。”由此看来，数学和地球科学的结合已成为地质科学向前发展的一个必然方向。

3．从单变量到多变量发展

客观世界中任何事情都是与其周围的事物或者与其所处的环境存在着千丝万缕的联系，在地球化学工作中也是一样，仅仅根据单变量的观测数据来分析问题已经远远不能满足研究工作的要求，而必须同时对多个变量进行观测、汇集与研究与问题全部有关的原始资料，较全面地研究它们之间的相互关系和相互影响，去粗取精、去伪存真，才能得出正确的结论。

地球化学和地球化学探矿不同于其他地质类学科的另一特点是它具有大量可测定的数据。仅从测试化学元素方面来看，从40、50年代只分析几个主要成矿元素，到60、70年代发展为既分析成矿元素又分析控矿元素和运矿元素。80年代以来，则向测试更多元素的方面发展。如地质矿产部1：20万化探扫面工作规定每个样品必须分析40多个元素，占元素周期表中能分析元素的大部分。

数学地质发展十分迅速的原因，从数理统计方面来看，各种统计方法越来越成熟，新的统计分析方法不断涌现，特别是多元统计分析发展很快，稳健统计分析和成分数据的统计分析方法的提出，使多元统计分析无论在理论上还是在应用方面都提高到一个新的水平。

50年代以来，电子计算机技术得到迅速发展，这为多元统计从数学方法的理论研究推

广到实际应用提供了必要的条件。如果没有电子计算机，要对一些数量较大的数据进行某些多元统计分析，如趋势分析、因子分析、判别分析等几乎是不现实的。电子计算机的使用对于多元分析在地质上的应用提供了一个必要的条件。

三、发展史及研究现状

早在1840年，英国地质学家莱尹尔（C.Lyell）运用统计的方法，根据近代海洋生物的相对含量，把第三纪地层中的生物种属与现代海生生物种属对比，成功地对第三纪地层作了进一步划分。

1890年，皮尔逊（K.Pearson）编写了《数学进化论贡献》一书，其中有古生物化石的统计分析。

1914—1934年间，俄国学者列文森——列辛格（Лёвцн сон—Лёссцнг）通过研究岩石的岩浆酸度系数的频率分布，探讨了安山岩、玄武岩、流纹岩的分类问题。

1936—1945年数学方法的应用由地质学的个别问题逐渐扩展到地质学的一些分支。在岩石学中引入了双变量或三变量分析，出现了三角岩比图、百分率图等。在矿物学和结晶学领域中，还运用了简单回归分析和方差分析，以了解矿物性质与化学成分之间的关系。1944年，苏联学者维斯捷列乌斯（Á.Б.Вцстёлцус）在苏联科学院报告集上发表了《分析地质学》一文，提出用定量方法研究地质问题的初步思想。他从事了三十多年的数学地质工作，成为苏联数学地质的创始人和国际数学地质学会的第一任主席。

1945—1960年间，数学地质迅速发展，电子计算机也开始在地质工作中应用。数学统计学的方法普遍应用于地质学的所有分支。1946年，苏联学者研究金属矿床元素统计分布特点。1954年，绍（D.M.Shaw）等人应用统计方法研究地球化学问题。1956年，切叶思（F.Chayes）应用均值、方差于岩石学研究。1956年，克伦宾(Graybill)在研究岩石的矿物和化学成分时，应用了多元统计分析方法。1958年，美国的克鲁拜因（W.C.Krumbein）从事区域地质统计分析方面的工作。同年，克鲁拜因首次在地质杂志上公布电子计算机地质计算程序。从此，地质学开始进入了应用电子计算机的新时代。

1961—1970年，数学方法在地质学中应用更为广泛，数学地质进入了发展昌盛时期，多元统计分析在地质学中大量应用，数学地质发展成为一门独立学科。自1961年开始，美国亚利桑那大学召开了一系列“电子计算机在矿产工业中的应用”讨论会。1966年，美国堪萨斯地质调查所召开了“电子计算机在地球科学中的应用”讨论会。1967年，国际地质科学联合会成立了“地质数据存储、自动处理和索取委员会”（COGEODATA）。1968年在巴黎召开的国际地质会议上成立了国际数学地质协会（IAMG），并开始出版国际数学地质协会杂志和地质计算机程序公报。美国地质调查所首次公布其电子计算机贡献文集——《电子计算机在地球科学中应用》一书出版。

近年来，数学地质在国内外发展很快。尤其在地质数据的存储、索取、自动处理和显示方面取得了长足的进步。据报导，到1980年止，世界上已建成了约500个大型地质数据库，涉及到地质学的各个重要领域。例如，瑞典设计了一个野外数据的存储索取系统（GEOMAP），用于野外地质填图工作。芬兰设计一种野外填图用的数据存储、索取和处理系统（GEOKU），具有自动查错和处理功能。加拿大戴维（P.David）和列比依斯（J.Lebuis）设计了一套全部地层信息标准记录系统（LEDA），可用于区域地质填图

和地层对比。此外，还设计出Geolog、Geosystem和野外数据处理系统（DASCH），用于地质、勘探、地球化学和地球物理野外调查等方面。DASCH系统可用于回归分析、因子分析、聚类分析、趋势面分析及自动绘图等。丹麦设计了一种称为System-Ⅱ的实验性数据处理系统，可以处理野外数据和实验室数据。

除上述记录野外数据的系统外，近年来在数据处理系统和计算机绘图方面又有新的进展，许多国家已经或正在建立一些更高级的地质数据的综合处理系统。其中，比较重要的有：

法国高等矿业学校发展了SIGMI系统，配备有较完整的软件，可直接用于地质数据的存储、索取和处理。

英国地球科学研究所用FORTRAN-Ⅳ语言编写的综合数据处理系统（G-EXEC系统），能在大多数计算机上应用，系统地处理任何类型的地质数据。英国“矿产矿山数据库”（MINTMINES）、“不列颠采矿数据库”（BRITPITS）和“矿产统计数据库”（MINSTATS）等，大都使用G-EXEC系统。此外，英国工程计算服务公司（ECS）的GEDNET地质应用软件系统，也可以处理各种地质数据。

加拿大西安略大学地质系建立了SAPRAS地质数据系统。这是一个以COBOL语言编写的系统，它可以存储、编辑、索取和处理各种地质数据。

美国地质调查所设计一套数据存储和索取系统（GRASP），最近又扩充为PACER和GANET两个子系统。它可用来直接输入地质数据，并能进行更新、检索和各种数据处理。美国俄克拉何马大学的通用信息处理系统（GIPSY），及由该系统所支持的美国“计算机化矿产资源信息库”(CRIB)，是世界上最大的一个矿产资源数据库。库内存有美国40 000个矿床和矿产地、国外6000个矿床和矿产记录。文件内容包括矿床位置、地质特征、储量、产量等数据。用户可以通过电子计算机网络在世界500多个城市用电话查阅或索取CRIB数据库中的数据。除上述综合系统外，还有一些专门数据系统。如“北美石油数据系统”（PDS）、“矿产可得性系统”（MAS）、“全国水文处理数据存储和检索系统”、“全国铀资源数据库”、“地球化学数据库”等。

捷克和斯洛伐克设计了一个数据处理系统，它包括10个数据文件。这些文件用统一的计算机系统联系起来，可处理环境地质数据，如处理矿床、野外地质填图、地下水、矿泉水、滑坡、水文学和工程地质的数据等。

加拿大地质调查所与主管地形测量等单位合作，应用电子计算机自动显示，实现了绘制地形图自动化和地质图半自动化。最近，以激光为基础的光学数据处理系统，将逐渐用于地质学中。

综上所述，国外数学地质方法在地球科学各个领域中已得到广泛的运用，并推动着地质学研究进一步走向定量化。地质过程的数学抽象、数学模型的建立及数学模拟等数学地质理论研究更加深入。在电子计算机两大功能（科学计算和信息处理）基础上建立起来的各种地质数据自动处理系统，已经开始从地质数据库向地质知识库，进而向具有人工智能的专家系统发展。

在我国数学地质大致经历了以下三个发展阶段。

1969—1978年为酝酿和引进阶段。在此期间，各科研单位和地质院校纷纷成立数学地质小组或数学地质研究室，举办各种数学地质短训班。在内容上，主要是引进国外比较成

熟的多元统计方法。1978年10月27日，在杭州召开了第一届全国数学地质学术讨论会，与会代表115人，提交了105篇论文资料。这是对该阶段工作的总结。

1979—1981年，为数学地质创立阶段。第一届全国数学地质学术讨论会之后，我国数学地质工作犹如雨后春笋，蓬勃发展。从事数学地质的专业人员不断增加，一支由老、中、青原数学工作者和地质工作者组成的数学地质队伍基本形成。1981年4月22日，在长沙召开了第二届全国数学地质学术讨论会，与会代表269人，提交论文203篇。会议期间成立了"全国数学地质专业委员会"，它标志着我国数学地质作为一门独立的学科正式创立。

1982—至今为稳定发展阶段。我国数学地质经过一段蓬勃发展之后，人们发现许多数学地质方法在使用时要受一定条件的限制。数学地质决不是简单的数学加上地质，数学地质的过程是一个研究和实践的过程。它要求数学地质专业人员既懂得数学的理论、方法以及应用条件，又要熟悉地质思维和各方面的专业知识，要在数学和地质的结合上下功夫。为了使数学地质更好地与生产实践相结合，迫使数学地质工作者对数学地质的方法内容、使用条件和解决实际问题的程度进行深入研究。这个阶段在数学理论方法上引进消化吸收了稳健多元统计分析和成分数据的统计方法，使对数学理论了解较少的地质工作者在数学的应用条件上降低了要求，便利了数学地质的推广，使数学分析方法的理论水平也提高了一步。在应用方面，对已有的数学地质方法的实用性进行了深入研究，新的数学地质模型的创立，适合我国国情的地质数据系统的建立，使数学地质在生产实践中逐渐形成生产能力。地质矿产部计算中心筹建了"全国化探扫面数据库"。中国地质大学（武汉）开始筹建"区域地球化学数据库"。

数学地质，尤其是多元统计分析发展很快，应用也十分广泛，但也有它的局限性。多元统计分析的结果直接依赖于原始数据，更明确地说多元统计分析只能起到对原始数据的综合、加工和提练的作用。因而在讨论多元统计的成果时，应该考虑原始数据中是否提供了与研究问题有关的信息，如果原始数据没有包含某种信息，那么在解释其结果时不能牵强附会。因此，数学地质工作者对于样品的采集和变量的选择是一个必须认真研究的问题。

四、多元分析方法的工作步骤

地球化学中的多元分析是以数学为工具，电子计算机为手段，解决地质地球化学问题为目的的，它解决问题的一般模式是：地质地球化学问题转化为数学问题，然后对数学问题的解进行地质解释。它是以地质地球化学为基础和出发点，经过变地质模型为数学模型，最后达到揭示地质问题的数学规律，或者明地质体和地质过程的数学特征，建立地质体和地质过程的数学模型，从而达到认识地质事件的客观特征，了解地质事件的运动规律为目的。

地球化学中的多元分析方法的工作步骤应当是：

1．了解地质背景，收集并分析前人工作资料；

2．制定地球化学样品的采集方案，研究对样品的测试项目（包括测试仪器、分析方案和分析灵敏度、精确度的确定），并对样品进行分析测试；

3．收集地球化学样品各种资料（包括样品的采集位置、地质产状、各项分析测试的原始数据等），建立地球化学数据库；

4．根据工作目的选择适当的多元统计分析方法，构成相应的数学模型；

5．确定合理的计算步骤和方法，并将计算步骤和方法编写出计算机程序或调用数据

处理系统中标准计算程序，选择地球化学数据库中适当样品的适当变量进行计算；

6．对计算的结果进行地质解释；

7．若对计算结果的地质解释不满意，可以调整计算方法、样品和变量的选择，然后再作计算和结果解释，通过反复实践，以获取较满意的结论。

第一章　地球化学变量

第一节　地球化学变量的特点及分类

一、地球化学变量的特点

各种地球化学特征在不同的研究对象（样品或测点）上进行测试，可以得到各自的数值。这种在不同对象上可以取值的地球化学特征称为地球化学变量。例如，元素含量、矿石的相对密度、矿体的厚度、薄片中某种副矿物的颗数、溶液的pH、Eh值、成矿成岩的温度和压力、热液包体溶液的盐度等等，都是地球化学变量。

地球化学变量具有3个特点。即它具有随机变量的性质，又具有确定性变量的特征，还具有区域化变量的结构性。

地球化学变量的随机性表现在以下几方面：

1．地球化学样品的采取具有抽样的性质。地质工作所研究的对象是天然物质。因此，不可能把整个大自然或一个区域的地质体都搬到实验室里来，而只能抽取其中一部分样品，这部分样品与所研究的对象相比，只占很少很少的数量。这样的工作方法本身就决定了地球化学变量具随机变量的性质。

2．从成因上看，地球化学变量的取值受多因素的控制。如某样品中某种元素的当今含量，它取决于漫长地质历史中的各种地质作用，在地质历史长河中，作用到某个研究对象上的各种因素，人们往往不能全部准确地推算到。常常遇到这样的情况，对于两个不同的研究对象，虽然它们有十分相似的地质地球化学环境，或者甚至是同一地点采取的，但它们的某些地球化学特征的测试结果却可以是不相同的，表现为随机变量的特征。

3．地球化学变量的随机性还表现在测试过程中，由于测试者的主观因素，测试仪器的精度不同，测试分析方案不同，测试时周围环境的影响，测试时间不同等等，对同一样品的同一变量的不同次测试结果就不一定完全相同。这可以归结为测试误差，误差就带有随机性特点。实际上，地球化学变量在某个个体上的真实值是不知道的，每一次测试只是对其真实值的一次估计。

地球化学变量又具有确定性的一面。这是因为每一个研究对象在指定的时刻，某种地球化学特征的真实数值是客观存在的，是有确定数值的。这个值不依赖于人们是否已经测试到了。

地球化学变量还具有结构性。所谓结构性是指地球化学变量由于受地质因素的控制，在不同方向上变化率的不同，具有不同的自相关的特性。当一个变量分布在空间内时（即赋于其空间位置时），称为区域化变量。区域化变量是一个函数$f(x)$，它在三维空间内的

每一点x[其坐标为(x_1, x_2, x_3)]取一个确定的值。但此函数的空间变异往往极不规则，具有异向性。

对地球化学变量的随机性、确定性和结构性，不同学者存在着不同的看法，有人较强调随机性的一面，而另外一些人则否定地球化学变量是随机变量。我们认为它具有随机性，又有确定性，还有结构性，只是由于选择数学方法的不同，有时把它当作随机变量，有时把它作为确定性变量。有时认为它是区域化变量。

二、地球化学变量的分类

地球化学变量按其在取值范围内的可能取值分为3类：

1．连续型变量

连续型变量在取值范围内有不可数无限多个值。这种变量可以取在某一个数值区间内的任何数值。如矿石中某种金属的相对含量可以在0.00%到100.00%之间变化。

2．离散型变量

离散型变量在取值范围内取有限多个或可数无限个值。如重砂样中某种副矿物的颗数。

3．二元变量

二元变量也称二态变量或有无变量，它是一种只有两个取值的离散变量。如某地区有矿或无矿，某种矿物有发光性或没有发光性，电性的正或负等等。二元变量在地质上并不少见。

另一种地球化学变量的分类方法是根据给予某个变量具体数值的方式来划分变量的类型。它把变量分为4类：

为叙述方便，先设变量x在两个观测对象A与B中的取值分别为x_A和x_B。

1．名义型变量

名义型变量把观测值分成互不相交的组别，这些组别可以用它们的名字、特征或符号来标识。例如岩石的不同蚀变类型，可以用其名称来区分，有钾化、硅化、绢云母化和绿泥石化等。又如矿物的颜色，有黑色、绿色、黄色、赭色等等。它们也可以用不同的字母符号来表示，还可以用事先约定的数字符号来标识，如约定钾化为1，硅化为2，绢云母化为3，绿泥石化为4等等。这些数字仅作为一种标识符，没有数值大小的涵义。即对样品A和B来讲，这种名义型变量只能判断是$x_A = x_B$还是$x_A \neq x_B$。如上例中，若有$x_A = x_B = 2$则说明A和B两者间为硅化样品；若$x_A = 2$，$x_B = 4$则说明A样品为硅化，而B样品为绿泥石化，A和B不具有相同的蚀变。除此之外不包含其他信息，更不能进行代数运算。

2．有序型变量

有序型变量对观测特征可以列出一个等级大小的顺序。对于A和B两个对象，它不仅能区别出$x_A = x_B$或$x_A \neq x_B$，还能进一步区分出$x_A > x_B$或$x_A < x_B$，即有了大小顺序的差别。但它们的级差不是等距的。如莫氏矿物硬度计，把矿物的硬度分成1，2，…，10共10个级别，随着级别的升高表示硬度增大，但相邻两级之间的硬度的差别却大不相同。又如在描述岩石的风化程度时，可以划分为新鲜的、弱风化的、风化的和强风化的4个级别，相应地亦可分别用数字0、1、2、3来标记。这里的数字只给出了一个风化程度强弱的顺序，而不是真正表示强风化是弱风化的两倍。再如变质岩研究中，按变质程度可以分成不同的变质相，不同变质相也可用不同数字来标识，但是各种变质相之间温度、压力的差别不一

定就是等间隔的。

3．间隔型变量

间隔型变量不但能区分它们之间的值是相等、不等、谁大谁小，而且还能表示谁比谁大多少。若$x_A=6$，$x_B=2$，则可以得到以下概念：x_A不等于x_B，x_A大于x_B，且x_A比x_B大4个单位。间隔型变量可以定量地表示它们之间的差异，但间隔型变量没有绝对零值。

4．比例型变量

比例型变量比间隔型变量更进一步。它具有间隔型变量的标识功能，而且还有真正的零点。若在样品A和B之间设$x_A=6$，$x_B=2$，则可以获得如下概念：①$x_A\neq x_B$；②$x_A>x_B$；③$x_A-x_B=4$；④$x_A/x_B=3$。比例型变量是最常见的变量，元素的含量，物体的体积，矿体的厚度等等都属于这类变量。

以上4类变量中，名义型变量和有序型变量属于定性变量，间隔型变量和比例型变量为定量变量。一般来说，定性变量只能参加布尔运算，定量变量可作代数运算。

地球化学变量按其预处理取值方法的不同还可以分为：

1．实测变量

它又叫观察变量或原始变量。它是用仪器或特定的分析手段对研究对象直接测定所得。

2．综合变量

综合变量是将两个或多个原始观察值加以综合，组成一个新的变量，具有特定的地质意义。这类变量是地质人员对原始变量的地球化学性质和地质含义进行累乘、累加或特殊的计算而得。如线金属量、因子得分、判别得分、特征分析中的关联指数等属于这类变量。它起到减少变量个数、简化数学模型和增强地质含义的作用。

3．人为变量

它是根据地质人员对地质现象和成矿理论人为确定的变量。如研究区内单位面积中出现的断裂的性质和发育程度，产出的岩性及其时代，距离岩体接触带的位置，甚至单元所处的地质构造环境等等都可以由研究人员人为赋给一定的数值。

第二节　变量的选择

地球化学工作中所遇到的变量是非常多的。为了解决某个地球化学问题而采用某种多元统计分析时，要取得理想的效果，必须首先对变量进行筛选。不可想象，不包含某种与研究问题有关信息的变量参加统计分析，就能够获得理想的结果。选择变量的原则可以从下述两方面来考虑，即与研究问题有关的变量应尽可能选进来；而与研究问题无关的变量都不要选入。也就是说，所选择的变量应能最有效地解决问题，变量的数目要尽可能少，而对研究问题有干扰的变量应去掉。在实际工作中，变量的选择也受分析测试条件和经费成本等因素的限制。

一、选择变量的方法

1．理论分析和经验法

(1) 针对所研究的问题，根据已有的地球化学和化探的理论知识选择变量。例如，为了研究某斑岩型Cu-Mo矿床的成矿晕地球化学机制和元素共生组合规律，采用因子分析。

根据矿床地球化学理论，可以选择以下几类元素：①成矿元素：Cu、Mo、Ag、S；②控矿元素：K、Na、Rb、Sr、Ca、Mg、Fe、Si、Al等；③运矿元素：F、Cl等。又如，对一个变质超基性铜镍矿床，为研究其成矿规律，可选择Ni、Co、S、Cr、Ti、B、Sn、As、Mo、Pb、Zn、V等12种化学元素。这是根据变质超基性岩类铜镍硫化物矿床的一般地球化学特征和该矿区岩体的具体地质特征确定的。其中，Ni、Co、Cr、S可以作为镍、钴、铜矿化的主要指示元素；Cr、Ti、V可以指示原始富含辉石的岩石；Ni、Co、Sn、As、B组合可以指示原始富含橄榄石的岩相，其中B还可以有效指示富橄榄石超基性岩的蛇纹石化；As可以指示晚期的热液矿化；Mo、Pb、Zn可以指示酸性岩脉与矿化的关系。从以上两个例子可以看出，理论分析法选择变量是对研究区的特定地质地球化学环境作具体分析的基础上选择变量的。在作法上，可以以不同矿种、不同类型的矿床中元素的共生组合规律的地球化学理论作为一般指导。

（2）采取经验的处理方法选择变量。是根据研究工作者本人或前人的工作结果，以相类似的或邻近矿床已经选用，并被实践证明是有效的变量作为工作地区选择变量的依据。

2．计算法

（1）相关系数

用研究的主要对象（如铜矿床的主要成矿元素Cu）与可供挑选的变量计算其相关系数。若它们之间存在明显的相关关系，则可选进。

对n个样品，变量x和变量y之间的简单相关系数$r_{(xy)}$可用下式求得：

$$r_{(xy)}=\frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2\cdot\sum_{i=1}^{n}(y_i-\bar{y})^2}}\qquad(f=1,\ 2,\ \cdots,\ n)\ ,\qquad(1\text{-}1)$$

其中

$$\bar{x}=\frac{1}{n}\sum_{i=1}^{n}x_i\ ,\quad \bar{y}=\frac{1}{n}\sum_{i=1}^{n}y_i\ .$$

（2）逐步回归分析

取被研究的变量为因变量，其余待选择的变量作为自变量，经过逐步回归分析，选择出最优的变量组合（详细计算步骤见第三章）。

（3）因子分析

对所有被挑选的变量作因子分析，并解释各主因子的地质意义，能代表研究问题的主因子所包含的变量组合，即是挑选出来的变量。

另外，如偏相关系数、秩相关系数、信息量估计等也可以作为变量选择的计算方法。

3．作图法

图形具有直观和形象的特点。它可表示出一般用数字和文字不容易表达清楚的一些现象和关系。因而也可用于变量的选择。

（1）单变量

从单个变量考虑，以该变量为坐标轴，把不同类型样品的该变量值投到坐标轴上，看这些点是否能够按不同类型分开。如其分辨率达到70%以上，说明该变量具有区分不同类

型样品的能力，可以将这个变量选进（见图1-1）。

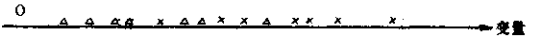


图1-1　单变量的分辨能力

（2）双变量

在直角坐标系上作两两变量的点聚图，既可区分不同类别（母体）的分离情况，又可用来直观地估计两两变量之间的相关关系。

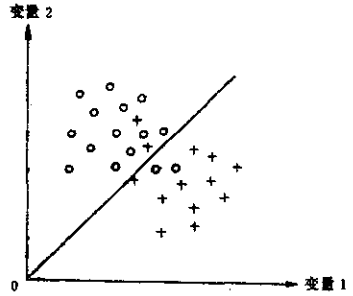


图1-2　双变量的分辨能力

当不同类型（母体）的样品点投到由这两个变量组成的直角坐标图上后，看这些用不同符号表示的样品点是否能分离成不同的点群，或者能否找到一条区分不同类别样品的分界线，使得它们能基本分开（区分率大于70%）。如果能分开，则说明这两个变量联合的判别效果是好的（见图1-2）。

（3）多变量

3个和3个以上变量可以用雷达图的办法来选择变量。假设有有矿及无矿两类样品，每个样品分别已测试了 p 个变量。雷达图的作法是，以 o 点为圆心，取适当长度为半径画圆，将圆周 p 等分，分别连接 o 和各等分点，得到 p 条成辐射状的半径。以这 p 条半径作为 p 个变量的坐标轴，根据每个变量数值波动大小（最好将原始数据作正规化变换），按 p 条坐标轴上刻度，将每个样品各变量值标于坐标轴上，按样品矿化类型不同连成多边形（p 个变量为 p 边形），分析对比各类总体所构成多边形的形状，以选择有判别能力的变量（见图1-3）。如变量SiO_2、CaO是判别能力强的变量，而K_2O则无判别能力。

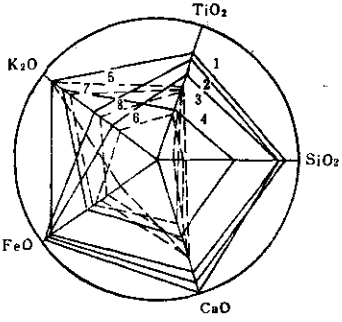


图1-3　多变量的分辨能力

二、选入变量数目与判别能力的关系

前面已经提到，选择变量的目的是确定一个最佳变量组合。所谓最佳变量组合是指该组合的判别能力最好，且变量的个数较少；从经济的角度上衡量，该组合的变量经济效益最好。不同的变量分析测试费用是不同的，同等分辨能力的变量组合，其经济成本也不一样。单从分辨能力来讲，随着变量个数的增加，分辨能力也不是直线上升的。郝瓦士（R. J. Howarth），1972年所作的变量组合与判别能力关系研究是一个很好的例子。例中所用的原始资料是某地复杂基岩的一组区域性水系沉积物的元素含量数据。构成判别函数的标准样本由153个样品组成。根据单变量的频率分布用目视估计，认为Pb、Sn、V、Cu、

Zn、Ti、Ni、Co、Mn、Cr和As这11种元素对于页岩夹砂岩、砂岩夹页岩、花岗岩、基性岩浆岩、铜矿和铅锌矿开采地区具有一定的判别能力。利用判别函数和逐次消去的方法对这些元素的所有可能的2.047种组合进行了计算，并算出正确分类的百分数。

按单个元素考虑，“最佳”元素是V、Co和Cr，其正确分类百分数分别为71.9、67.3和63.7(见表1-1)。由此看来似乎V-Co或V-Cr应是最优双变量组合。但表1-2指示最优双变量组合却是Ni-V，其正确分类的百分数为86.9，比V-Co(74.5)有很大的提高。根据上述情况，如果用“直观”单变量经验方法来估计最佳三元素组合或多元素组合，得到的结果可能更差。一般说来，新增加一个变量所取得的效果不仅依赖于其与已经被选进的变量之间的相关关系的性质，而且还取决于各类均值之间的可分程度和各变量之间的交互作用。

表1-1 某地水系沉积物根据单变量的判别函数所得正确分类的百分数

变量编号	1	2	3	4	5	6	7	8	9	10	11
变量名称	Pb	Sn	V	Cu	Zn	Ti	Ni	Co	Mn	Cr	As
正确分类%	52.9	49.7	71.9	60.8	61.4	54.9	65.4	67.3	47.7	67.3	53.6

表1-2 某地水系沉积物双变量的判别函数所得正确分类的百分数

%	Pb	Sn	V	Cu	Zn	Ti	Ni	Co	Mn	Cr	As
Pb		66.1	86.3	67.3	73.9	71.2	69.9	74.5	58.2	79.7	64.1
Sn			84.3	71.9	77.8	71.2	67.3	70.6	64.1	81.7	72.5
V				81.7	84.3	77.1	86.9	74.5	73.2	78.4	74.5
Cu					68.6	66.4	81.0	75.8	68.0	86.9	67.3
Zn						70.6	83.0	75.2	77.1	83.0	68.0
Ti							75.3	69.3	65.4	77.8	68.0
Ni								71.9	71.9	76.5	79.7
Co									66.0	75.2	77.1
Mn										68.0	67.3
Cr											75.2
As											

变量的C_n^k种所有可能组合正确分类的百分数的频率分布(见图1-4)反应出正确分类的百分数随k增加而迅速增大，直到某一点后，变为相对稳定。这时再增加一个新变量不会进一步提高正确分类的百分比，有时反而下降(如表1-3所示，当选进8个变量时，正确分类为98.0%，变量增加到11个时，正确分类的百分率反而下降到91.5%)。这说明了正确选择变量组合的重要性。

从这个例子至少可以得出以下几点结论：

(1) 多变量的判别分类能力比单变量的判别分类能力强。尤其在开始时，随着变量的增加，其分类能力很快增强。

(2) 变量个数增加到一定程度后，其分类能力不再随变量的增加而增加，有时反而有

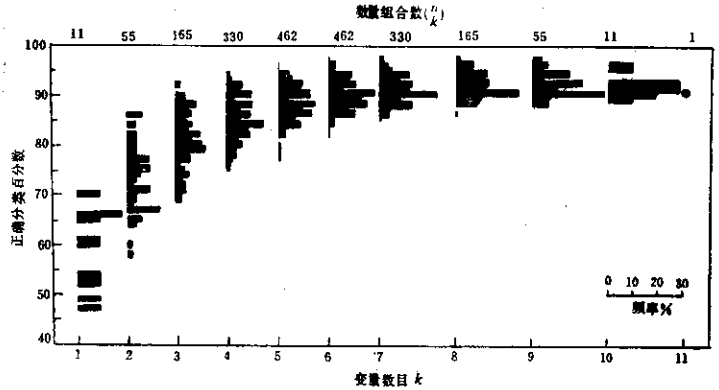


图1-4　利用逐次消去法作正确分类百分数与变量数 k 的频率分布图

表1-3　根据最佳元素组合的判别函数对某地水系沉积物正确分类百分数

变量数k	组合数 $\binom{n}{k}$	最高正确分类%	变　量　组　合*
1	11	71.9	3
2	55	86.9	3, 7
3	165	92.2	1, 3, 6
4	330	94.1	3, 4, 6, 7
5	462	96.1	3, 4, 6, 8, 10 3, 4, 6, 7, 8
6	462	97.4	1, 3, 4, 6, 7, 8 1, 3, 4, 6, 7, 9
7	330	97.4	1, 3, 4, 6, 7, 8, 9 1, 3, 4, 6, 8, 9, 10 1, 3, 5, 6, 8, 9, 10
8	165	98.0	1, 3, 4, 5, 6, 7, 8, 9
9	55	97.4	1, 3, 4, 5, 6, 7, 8, 9, 10
10	11	96.1	1, 3, 4, 5, 6, 7, 8, 9, 10, 11
11	1	91.6	1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11

* 1 = Pb, 2 = Sn, 3 = V, 4 = Cu, 5 = Zn, 6 = Ti, 7 = Ni, 8 = Co, 9 = Mo, 10 = Cr, 11 = As。

下降的可能。

（3）在同一个判别分类的水平上，可以有多种变量组合。

第三节　变量的变换

一、变量的均匀化变换

地球化学工作中得到的原始数据，各变量的单位（量纲）、数量级以及其代表的意义之间差异很大。这样直接用原始数据进行某些多元统计分析时，会出现如下问题：①岩石化

学分析或化探样品的元素含量中，主量元素和微量元素之间数量级相差很大，每一元素不同样品间的含量变动范围也有相当大的差异。主量元素Si、K、Al、Na、Ca的含量从百分之几到百分之几十，而Cu、Pb、As等元素通常只有几百ppm到几十ppm，地壳中丰度最大的元素（O=47%）比丰度最小的元素（Rn=7×10^{-16}）在含量上可大10^{17}倍。这样，在进行某些数学运算时，对于那些绝对值大的元素则可能扩大它的作用，而对于绝对值小的元素则可能缩小其实际作用，甚至因为计算时误差精度所限将会对某些绝对值小的元素的贡献舍去。②如果一组观测数据中包括性质不同的变量，象元素的含量、矿石的相对密度、矿体的埋深等，这些不同意义的变量一起参与某种统计计算时，就涉及到单位的统一问题。

以上两个方面的问题都可以归结为统一量纲的问题。变量的均匀化变换的基本目的是设法消除不同变量之间量纲的影响，将参与多元统计分析的所有变量统一成某种共同的数值特征。

变量均匀化变换所采用的方法是将同一变量的每个观测值除以与此变量有相同单位的某个均匀化因子，使原始变量变换为无量纲的相对数，从而消除单位（或数量级）不同所造成的影响。变量均匀化后的含义将取决于选择哪一种均匀化因子。下面分别介绍三种常用的均匀化变换方法。

设有n个样品，每个样品作p种地球化学变量的测定，则原始数据可表为矩阵

$$\mathbf{X}=\begin{pmatrix} x_{11} & x_{12}\cdots x_{1p} \\ x_{21} & x_{22}\cdots x_{2p} \\ \vdots & \vdots \quad \vdots \\ x_{n1} & x_{n2}\cdots x_{np} \end{pmatrix},$$

矩阵X中元素x_{ij}($i=1,\ 2,\ \cdots,\ n;\ j=1,\ 2,\ \cdots,\ p$)表示第$i$样品的第$j$个变量的观测值，其变换后的值记为$z_{ij}$($i=1,\ 2,\ \cdots,\ n;\ j=1,\ 2,\ \cdots,\ p$)。

1．均值计量变换

均值计量变换即用某变量的平均值作为均匀化因子去除该变量每个样品的原始观测数据。其数学表达式为：

$$z_{ij}=x_{ij}/\bar{x}_j \qquad (i=1,\ 2,\ \cdots,\ n;\ j=1,\ 2,\ \cdots,\ p) \quad , \qquad (1\text{-}2)$$

其中

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

显然，$\bar{z}_j=1$，$\sum_{i=1}^{n}(z_{ij}-\bar{z}_j)=0$，($j=1,\ 2,\ \cdots,\ p$)。

均值计量变换的数学意义是将坐标轴的单位化为$1/\bar{x}_j$。严格说来，该变换只适用于比例型变量。

2．极差变换（又称正规化或规格化变换）

变量的极差是指该变量中不同样品最大值与最小值之差。若以$x_{j\max}$表示变量j的最大值，以$x_{j\min}$表示变量j的最小值，则变量j的极差为$x_{j\max}-x_{j\min}$。极差变换就是将原始数据减去该变量的最小值然后以极差作为均匀化因子去除，其数学表达式为：

$$z_{ij}=\frac{x_{ij}-x_{j\min}}{x_{j\max}-x_{j\min}} \qquad (i=1,\ 2,\ \cdots,\ n;\ j=1,\ 2,\ \cdots,\ p)\ 。 \tag{1-3}$$

极差变换后$0\leqslant z_{ij}\leqslant 1$。极差变换的几何意义相当于把坐标原点移到最小值的位置，且把变量的极差作为该变量轴的单位长度。

3。标准化变换

标准化变换是将每一变量的原始观测值减去平均值，并以该变量的标准差作为均匀化因子去除，其数学表达式为：

$$z_{ij}=\frac{x_{ij}-\bar{x}_j}{S_j} \qquad (i=1,\ 2,\ \cdots,\ n;\ j=1,\ 2,\ \cdots,\ p)\ , \tag{1-4}$$

其中

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

标准化后变量的平均值为0，方差为1。其几何意义相当于将坐标原点移至重心（平均值）位置，且各变量按照各自的均方差比例化。

经标准化变换后的两个变量j和k之间的相关系数r_{jk}的计算公式可简化为：

$$r_{jk}=\frac{1}{n}\sum_{i=1}^{n}z_{ij}\cdot z_{ik} \qquad (i=1,\ 2,\ \cdots,\ n;\ j,\ k=1,\ 2,\ \cdots,\ p)\ 。 \tag{1-5}$$

标准化是最常用的一种均匀化变换。

二、变量的函数变换

大部分多元统计分析方法要求数据尽可能地服从正态分布，要求各变量之间的关系为线性关系。为了达到以上目的可以采用函数变换的办法来满足。

1。对数变换

地球化学样品中微量元素的含量常常呈对数正态分布。要使对数正态分布的数据转换为正态分布，只需要将原始数据作对数变换即可，其数学表达式为：

$$z_{ij}=\log(x_{ij}+c) \qquad (i=1,\ 2,\ \cdots,\ n;\ j=1,\ 2,\ \cdots,\ p)\ , \tag{1-6}$$

其中c为常数，用c可以调整某些变量的原始数值过小或过大的情况。特别是原始数据值很小导致其对数值出现负数，为避免上述缺点，在对数变换时可选择适当的c值。而一般情况下，c可取零值。

2。双曲函数、幂函数、指数函数变换

(1) 双曲函数变换

$$\frac{1}{z_{ij}}=a+\frac{b}{x_{ij}} \qquad (i=1,\ 2,\ \cdots,\ n;\ j=1,\ 2,\ \cdots,\ p)\ 。 \tag{1-7}$$

变换时常令 $z'_{ij}=\frac{1}{z_{ij}}$，$x'_{ij}=\frac{1}{x_{ij}}$，则

$$z'_{ij}=a+b\cdot x'_{ij}\ 。$$

(2) 幂函数变换

$$z_{ij}=dx_{ij}^{b} \qquad (i=1,\ 2,\ \cdots,\ n;\ j=1,\ 2,\ \cdots,\ p)\ , \tag{1-8}$$

变换时，两边常用对数，即令$z'_{ij}=\lg z_{ij}$，$x'_{ij}=\lg x_{ij}$，$a=\lg d$，则有$z'_{ij}=a+bx'_{ij}$。

（3）指数函数变换

$$z_{ij}=de^{bx_{ij}} \qquad (i=1,\ 2,\ \cdots,\ n;\ j=1,\ 2,\ \cdots,\ p)\ 。 \tag{1-9}$$

变换时两边取自然对数，令 $z'_{ij}=\ln z_{ij}$，$a=\ln d$，得 $z'_{ij}=a+bx_{ij}$。

3．三角函数和反三角函数变换

对于那些具有周期性变化的变量，化为线性关系时可用三角函数和反三角函数变换。其数学表达式为：

$$z_{ij}=\sin x_{ij}\ ,\quad z_{ij}=\cos x_{ij}\ ,$$
$$z_{ij}=\sin^{-1}x_{ij}\ ,\quad z_{ij}=\cos^{-1}x_{ij}\ 。 \qquad (i=1,\ 2,\ \cdots,\ n;\ j=1,\ 2,\ \cdots,\ p) \tag{1-10}$$

三、变量的组合

为了突出某种地球化学作用，加强变量的地球化学成因意义以及简化变量数目，可以把两个或两个以上的原始变量组合成一个新的变量。

1．累加变换

$$z_i=a_1x_{i1}+a_2x_{i2}+\cdots \qquad (i=1,\ 2,\ \cdots,\ n)\ , \tag{1-11}$$

其中a_1，a_2，…分别为第一个变量，第二个变量，……的权系数，其数值可根据变量的贡献大小、地球化学意义或数量级来取值。

2．累乘变换

$$z_i=x_{i1}^{a_1}\ x_{i2}^{a_2}\cdots \qquad (i=1,\ 2,\ \cdots)\ , \tag{1-12}$$

其中a_1，a_2，…为系数，在一般情况下可取1。当原始变量只有两个，且a_1与a_2分别为1和－1时，即为$z_i=x_{i1}/x_{i2}$，它是x_1与x_2两个元素对的比值。元素对的比值在地球化学和化探工作中特别有用。如在研究基性-超基性岩以及与其有关的矿床时，常用FeO/MgO；在硫化物矿床的研究中，常用Ni/Co或S/Se；在化探中常常用头晕元素与尾晕元素的比值来指示矿体的剥蚀深度等等。

3．多元统计分析法

多元统计分析法就是将原始数据先进行一次多元统计分析，对原始的变量进行组合或变换，得到符合一定要求的新变量。然后用新变量的数据再去做某种多元统计分析，所以这种方法也可以称为多重多元统计分析。

用来组合和约简变量最常用的多元统计分析方法是主成分分析和因子分析。为了得到有地球化学成因意义、互相独立、变量个数较原始变量少的新的变量组合，可以将原始数据先做一次主成分分析或因子分析，求得各样品新变量(主成分或因子)的因子计量数据，再用新的数据去做趋势分析、判别分析等等。有时这种多重多元分析方法效果很好。

除了因子分析外，其他的多元统计方法如：回归分析、判别分析、对应分析、马尔科夫概型分析等也可以用来变换原始变量。

多元统计分析的数学方法包括以后各章中的主要内容，详细的计算步骤请看以后相应章节。

第四节　数据集合的两种空间表示

对于一个个体（如一块标本）的 p 种地球化学特征（变量）进行测试，可以得到一个数据集合，可用向量：

$$\mathbf{x}=\begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_p \end{pmatrix}$$

来表示，向量$\mathbf{x}$的转置$\mathbf{x}'$表示为

$$\mathbf{x}'=[x_1,\ x_2,\ \cdots,\ x_p]\ 。$$

对于 n 个个体，每个个体测试 p 种地球化学特征（变量），所得的原始数据可以表示为一个 $n\times p$ 阶的矩阵：

$$\mathbf{X}=(x_{ij})=\begin{pmatrix} x_{11} & x_{12}\cdots x_{1p} \\ x_{21} & x_{22}\cdots x_{2p} \\ \vdots & \vdots \\ x_{n1} & x_{n2}\cdots x_{np} \end{pmatrix}\ 。$$

在这原始观测数据矩阵中，x_{ij}表示第 i 个个体的第 j 个变量的观测值。

对于观测矩阵有两种不同的几何表示方法。

一、数据集合的第一种空间表示

这种空间表示是把数据集合看作 p 维空间中的 n 个点。想像有 p 条正交轴，确定一个 p 维空间。集合中的任一个体对应于 p 维空间中的一个点，第 i 点的第 j 轴的坐标是 x_{ij}。因而 n 个个体就对应于 p 维空间中的 n 个点，而我们感兴趣的是这一群点的分布模式：它们究竟是坍塌成一个较低维数的扁平体，还是或多或少呈球状分布，或者呈伸长的形状；它们是否可以清楚地分成若干组，等等。当然，为了便于表示，只要能画出二维的图形就可以了，但就其性质来说，这个问题却具有一般性。

值得指出的是，这种几何的表示方法毫不依赖于三维以上空间的“存在”与否。我们可以把它看作是用一种较方便的语言来表示可以用代数术语表达的数学事实。比如，对于同一个数学事实，我们分别用如下的几何语言和代数语言表示，则显然几何语言的表达方式比较方便：

1．一个（$p-1$）维的超平面和一个 p 维的超球面相交于一个（$p-1$）维的超球面(类似于一个平面和一个球面相交于一个圆）。

2．$\sum_{i=1}^{p}(x_i-a)^2=b$ 和 $\sum_{j=1}^{p}c_jx_j=d$

的轨迹可以经过一个适当的坐标变换表示成如下的形式

$$\sum_{i=1}^{p-1}(y_i-k)^2=l\ 。$$

二、数据集合的第二种空间表示

这种空间表示是把数据集合看作 n 维空间中的 p 个点。

取 n 条正交轴，其中每一条轴对应于 n 个个体集合中的一个个体；并确定 p 个点 p_1，p_j 的坐标是 n 个值：$x_{1j}-x_{\cdot j}$，$x_{2j}-x_{\cdot j}$，…，$x_{nj}-x_{\cdot j}$。因而，对于每一个变量就有一个对应的点。p_j 与原点的距离由下式给出：

$$OP_j^2=\sum_{i=1}^{n}(x_{ij}-x_{\cdot j})^2 \quad (j=1,\ 2,\ \cdots,\ p)\ , \tag{1-13}$$

即 x_j 的方差的 n 倍。

如果将变量按方差为 1 进行标准化，则 p 个向量 OP_j 都具有单位长度，并且它们的端点都落在一个单位半径的超球面上。此外，OP_i 和 OP_j 之间所成夹角 θ_{ij} 的余弦

$$\cos\theta_{ij}=\sum_{k=1}^{n}(x_{ki}-x_{\cdot i})(x_{kj}-x_{\cdot j})\Big/\left[\sum_{k=1}^{n}(x_{ki}-x_{\cdot i})^2\sum_{k=1}^{n}(x_{kj}-x_{\cdot j})^2\right]^{1/2}$$

$$(i,\ j=1,\ 2,\ \cdots,\ p)\ , \tag{1-14}$$

是 x_i 和 x_j 之间的相关系数。

p 个向量 OP_1，OP_2，…，OP_p 确定一个嵌入在 n 维空间内的 p 维空间。如果两个变量高度相关，则 θ_{ij} 接近于零。如果两个变量为零相关，则 θ_{ij} 为 $\frac{\pi}{2}$，该两向量相互正交。也就是说，这些向量“集拢”的程度是它们所代表的变量之间相关系数大小的度量。如果一个向量位于其他向量所确定的（$p-1$）维空间内，则它所代表的变量是其他一些变量的线性函数。这里，向量的模式揭示变量之间相互关系的性质，如同第一种空间表示中 n 个点揭示个体之间相互关系的性质那样。

在多元统计分析中所谓“R 式分析”（研究变量之间的相互关系），就是用第二种空间表示；所谓“Q 式分析”，就是用第一种空间表示来研究样品（个体）之间的相互关系。对于同一批数据在上述两种空间表示法中都可得到两个点群之间存在着数学上的内在联系。

第二章　聚类分析

第一节　一般概念

分类问题是地球化学工作最常遇到的问题之一。如化探异常的分类、铁帽类型的划分、元素组合的分析等等。传统的方法中地质人员常常用定性的标志或单个指标进行分类。那些做法有时可能抓住一个或两个明显的特征对所研究的对象作一粗略的划分，但是，在许多情况下它不能揭示客观事物的本质和内在联系，分类时常常带有主观性和片面性，因此，不能反映客观实际情况。地球化学工作者为了克服以上的缺点，采用从事物的各个角度的不同特征去研究对象，进行分类，即综合多个指标的信息来划分研究对象的类别。然而，在多指标的情况下，用定性的、直观的经验方法有时就显得顾此失彼，甚至无法确定某对象的归属，这时，数学上的聚类分析在地球化学工作中的分类方面就显示了它的优越性，因而得到了迅速的推广。

聚类分析也叫点群分析、簇丛分析、群分析等等。它是根据研究对象（样品或变量）的多种特征在数值上可能存在着的相似性程度，将它们聚合为不同的点群的一种多元统计分析方法。

聚类分析中，从研究对象本身多个特征的数据作为出发点，考查它们每个研究对象之间的相似性程度，这种相似性度量是综合了多个指标的结果，从而把多变量转化为单变量，并按照这个相似性程度由亲到疏把所研究的个体逐次聚合成群。一般最后画出一个树枝状的谱系图，一目了然地表示出在不同相似水平上个体间的分群关系。

聚类分析按其研究对象的不同可分为两类：一类是对样品分类，称Q型分析；另一类是对变量分类，称R型分析。

第二节　相似性度量

在进行聚类分析时首先要对原始数据进行均匀化变换（见第一章第二节）。接着是把研究对象的多种特征综合成为可供比较研究对象之间亲疏关系的一个相似性统计量。常用的相似性统计量有以下几种：

一、距离系数d_{ih}

把n个研究对象（个体）表示为m维坐标空间中的n个点，很自然这些点之间的欧氏距离可用来度量两两点之间的亲疏关系，即两点之间的空间距离愈近，它们就愈相似。

设在m维空间中，第i点的坐标为x_{ij}，第h点的坐标为x_{hj}。那么，i点与h点间的欧

氏距离d_{ih}^{*}就可根据勾股定理推得：

$$d_{ih}^{*}=\left[\sum_{j=1}^{m}(x_{ij}-x_{hj})^{2}\right]^{1/2}\qquad(i,\ h=1,\ 2,\ \cdots,\ n)\ 。$$

为了使距离系d_{ih}在一定范围内变化，我们定义它为：

$$d_{ih}=\left[\frac{1}{m}\sum_{j=1}^{m}(x_{ij}-x_{hj})^{2}\right]^{1/2}\qquad(i,\ h=1,\ 2,\ \cdots,\ n)\ 。\qquad(2\text{-}1)$$

这里与欧氏距离不同的是根号里乘了一个$\frac{1}{m}$。这是因为欧氏距离d_{ih}^{*}是一个与空间坐标轴的维数有关的函数。除以m后即可使之限定在某个范围内。

经极差变换后的正规化数据，距离系数的取值范围为（0—1）之间。d_{ih}的值越小，表示点i与点h之间越相似。距离系数常用于样品点间的相似性度量，即用于Q型聚类分析中。

二、相关系数r_{ij}

相关系数r_{ij}是随机变量i和j之间相关程度的一种度量。因此，可以用于聚类分析中变量间的相似性度量。

设对n块标本，每块标本测试m个特征（变量），则任意两个变量i，j之间的相关系数r_{ij}可由下式计算：

$$r_{ij}=\frac{\sum_{i=1}^{n}(x_{ij}-\bar{x}_{i})(x_{ij}-\bar{x}_{j})}{\left[\sum_{i=1}^{n}(x_{ij}-\bar{x}_{i})^{2}\sum_{i=1}^{n}(x_{ij}-\bar{x}_{j})^{2}\right]^{1/2}}$$

$$=\frac{\sum_{i=1}^{n}x_{ij}x_{ij}-n\bar{x}_{i}\bar{x}_{j}}{\left[(\sum_{i=1}^{n}x_{ij}^{2}-n\bar{x}_{i}^{2})^{2}(\sum_{i=1}^{n}x_{ij}^{2}-n\bar{x}_{j}^{2})\right]^{1/2}}\qquad(i,j=1,2,\cdots,m)\ 。\qquad(2\text{-}2)$$

对于标准化变量，计算公式可简化为：

$$r_{ij}=\frac{1}{n}\sum_{i=1}^{n}x_{ij}x_{ij}\qquad(i,\ j=1,\ 2,\ \cdots,\ m)\ 。\qquad(2\text{-}3)$$

相关系数的取值范围为（-1—1）。相关系数越大，则说明变量越相似。

三、夹角余弦（或称相似系数）$\cos\theta_{ih}$

如果把任意两个研究对象表示为m维坐标空间中的两个向量i和h，那么这两个向量夹角的余弦$\cos\theta_{ih}$也可以用来表示这两个向量的相似性程度。

向量i和向量h之间的夹角余弦$\cos\theta_{ih}$可用下式计算：

$$\cos\theta_{ih}=\frac{\sum_{j=1}^{n}x_{ji}x_{jh}}{\left[\sum_{j=1}^{n}x_{ji}^2\sum_{j=1}^{n}x_{jh}^2\right]^{1/2}}\qquad(i,\ h=1,\ 2,\ \cdots,\ m)\ 。\qquad(2\text{-}4)$$

两个向量夹角余弦的值域为（$-1\leqslant\cos\theta\leqslant1$）。其物理意义可以理解为，向量 i 和向量 h 中对应元素成比例的程度。显然两个向量的夹角越小，$\cos\theta$越接近 1，则说明这两个向量越相似。几个较相似的向量，在坐标系中集合成一束或一群。

四、离差平方和增量ΔE_{pq}

这是聚类分析中常用的一个相似性统计量。离差平方和法也称误差平方和法，其思路是：点群在逐次聚合过程中，每一次挑选出两个点群，使得当二者合并为一新群后，其组内离差平方和 E 的增加值ΔE 比其它任何两个点群合并时增加的值都小，说明原来这两个点群最相似。

设在m维坐标空间中有 n 个个体，它们分为 g 个点群，任意第 k 个点群中包含有n_k个个体。记 x_{ijk} 是 m维空间中第 k 个点群第 i 个个体第 j 个变量的坐标值，则第 k 个点群的离差平方和E_k可用下式计算：

$$E_k=\sum_{j=1}^{m}\sum_{i=1}^{n_k}(x_{ijk}-\bar{x}_{jk})^2=\sum_{j=1}^{m}\sum_{i=1}^{n_k}x_{ijk}^2-n_k\sum_{j=1}^{m}\bar{x}_{jk}^2$$
$$(k=1,\ 2,\ \cdots,\ g)\ ,\qquad(2\text{-}5)$$

其中

$$\bar{x}_{jk}=\frac{1}{n_k}\sum_{i=1}^{n_k}x_{ijk}\ 。$$

整个坐标系中全部点群的离差平方和的总和E为：

$$E=\sum_{k=1}^{g}E_k\ 。\qquad(2\text{-}6)$$

在逐次聚合过程中，若将任意两个点群 p 和 q 加以合并形成新的点群 t，那么合并后比合并前离差平方和的增加ΔE_{pq}为：

$$\Delta E_{pq}=E_t-(E_p+E_q)\ 。\qquad(2\text{-}7)$$

将（2-5）式代入（2-7）式得：

$$\Delta E_{pq}=\left(\sum_{j=1}^{m}\sum_{i=1}^{n_t}x_{ijt}^2-n_t\sum_{j=1}^{m}\bar{x}_{jt}^2\right)-\left[\left(\sum_{j=1}^{m}\sum_{i=1}^{n_p}x_{ijp}^2-n_p\sum_{j=1}^{m}\bar{x}_{jp}^2\right)\right.$$
$$\left.+\left(\sum_{j=1}^{m}\sum_{i=1}^{n_q}x_{ijq}-n_q\sum_{j=1}^{m}\bar{x}_{jq}^2\right)\right]\ ,$$

消去x_{ijt}^2的项，化简后得：

$$\Delta E_{pq}=n_p\sum_{j=1}^{m}\bar{x}_{jp}^2+n_q\sum_{j=1}^{m}\bar{x}_{jq}^2-n_t\sum_{j=1}^{m}\bar{x}_{jt}^2\ ,\qquad(2\text{-}8)$$

其中

$$n_t = n_p + n_q \quad , \tag{2-9}$$

$$\bar{x}_{jt} = \frac{n_p\bar{x}_{jp} + n_q\bar{x}_{jq}}{n_p + n_q} \quad , \tag{2-10}$$

$$n_t\bar{x}_{jt} = n_p\bar{x}_{jp} + n_q\bar{x}_{jq} \quad 。 \tag{2-11}$$

为了化简（2-8）式，将（2-11）式两边平方得：

$$\begin{aligned} n_t^2\,\bar{x}_{jt}^{\,2} &= n_p^2\,\bar{x}_{jp}^{\,2} + n_q^2\,\bar{x}_{jq}^{\,2} + 2n_pn_q\bar{x}_{jp}\bar{x}_{jq} \\ &= n_p^2\,\bar{x}_{jp}^{\,2} + n_q^2\,\bar{x}_{jq}^{\,2} + \left[\bar{x}_{jp}^{\,2} + \bar{x}_{jq}^{\,2} - (\bar{x}_{jp} - \bar{x}_{jq})^2\right]\cdot n_pn_q \\ &= n_p(n_p + n_q)\bar{x}_{jp}^{\,2} + n_q(n_p + n_q)\bar{x}_{jq}^{\,2} - n_pn_q(\bar{x}_{jp} - \bar{x}_{jq})^2 \\ &= n_pn_t\,\bar{x}_{jp}^{\,2} + n_qn_t\bar{x}_{jq}^{\,2} - n_pn_q(\bar{x}_{jp} - \bar{x}_{jq})^2 \quad 。 \end{aligned}$$

将上式两边除以n_t^2得：

$$\bar{x}_{jt}^{\,2} = \frac{n_p}{n_t}\bar{x}_{jp}^{\,2} + \frac{n_q}{n_t}\bar{x}_{jq}^{\,2} - \frac{n_pn_q}{n_t^2}(\bar{x}_{jp} - \bar{x}_{jq})^2 \quad 。 \tag{2-12}$$

将（2-12）式代入（2-8）式得：

$$\Delta E_{pq} = \frac{n_pn_q}{n_p + n_q}\sum_{j=1}^{m}(\bar{x}_{jp} - \bar{x}_{jq})^2 \quad 。 \tag{2-13}$$

（2-13）式是计算离差平方和增量的最后公式。该式说明，p、q两点群合并后离差平方和的增量ΔE_{pq}与两点群中的各变量的重心（平均值）的距离成正比，并以$\frac{n_pn_q}{n_p + n_q}$为比例系数。离差平方和增量ΔE随着点群的逐次合并而增大，是个增函数。

五、其他一些相似统计量

以上4种是较常用的相似性统计量，其他还有一些不太常用的统计量：

1．绝对值距离系数（点群i和h间的绝对距离d_{ih}）

$$d_{ih} = \frac{1}{m}\sum_{j=1}^{m}|x_{ij} - x_{hj}| \quad 。 \tag{2-14}$$

2．斜交距离系数

$$d_{ik} = \left[\frac{1}{m^2}\sum_{j=1}^{m}\sum_{l=1}^{m}(x_{ij} - x_{kj})\cdot(x_{il} - x_{kl})\cdot r_{jl}\right]^{1/2} \quad 。 \tag{2-15}$$

3．切比雪夫距离

$$d_{ih} = \max_{1\leqslant k\leqslant m}|x_{ik} - x_{hk}| \quad 。 \tag{2-16}$$

4．兰氏距离

$$d_{ik}^2 = \left(\sum_{k=1}^{m}|x_{ik} - x_{hk}|\right)/(x_{ik} + x_{hk}) \quad 。 \tag{2-17}$$

5．马氏距离

$$d_{ik}^2=(x_i-x_k)'S^{-1}(x_i-x_k)\ ,\qquad (2\text{-}18)$$

其中，S^{-1}为样本协方差阵的逆矩阵，

$$S_{ik}=\frac{1}{n-1}\sum_{k=1}^{n}(x_{ki}-\bar{x_i})(x_{kk}-\bar{x_k})\ 。$$

6．指数相似系数

$$r_{ik}=\frac{1}{m}\sum_{k=1}^{m}e^{-\frac{3}{4}\frac{(x_{ik}-x_{jk})^2}{S^2}}\ 。\qquad (2\text{-}19)$$

以上各种相似性统计量中，一般说来距离类统计量用于Q型聚类分析，且其数值越小说明两个点群越相似；而相关系数类的统计量一般用于R型聚类分析，其数值越大，表示两组变量越相似。

第三节　点群的谱系簇分

有了相似性统计量就可以对点群进行谱系簇分或者说对点群聚合归类形成谱系图。

表2-1　相似形矩阵

s_{ij}	1	2	3	…	n
1		s_{12}	s_{13}	…	s_{1n}
2			s_{23}	…	s_{2n}
⋮				…	…
$n-1$					$s_{n-1\ n}$
n					

第二节中我们讨论了点群间的相似性度量，按照某种相似性度量，任何两两点之间都可以计算出其相似性统计量，记为s_{ij}，它构成一个相似性矩阵。由于相似性矩阵具有对称性，即$s_{ij}=s_{ji}$，（表2-1）。

此矩阵显示了能对比出最相似点群的信息。据此，可以按照某种原则作点群的聚合归类。点群的聚合归类过程称为谱系簇分。谱系簇分的方法很多，但基本上可以归为两大类：

一、一次计算形成法

一次计算形成法是根据原始数据一次计算出来的相似性矩阵S直接进行各个相似性水平分群成图。首先从相似性矩阵S中选择最相似的一对点群p、q（若$p<q$）连结成群，并随时把有关连结顺序、被连结的点群号和相似性水平记入连结顺序表中，连结完一对点群后，再选择相似性次大的一对，依次进行，直到把所有的点群都聚合为一群，并得到一个连结顺序表。根据该表，按照以下原则作谱系图：

1．若选出的两个点群，在已形成的群中都未出现过，则把它们形成一个独立的新群。

2．若选出的两个点群中，有一个是在已经分好的群中出现过，则把另一个点加入到该群中。

3．若选出的两个点群都分别出现在已分好的两个群中，则把这两个群合并为一个大群。

4．若选出的两个点都出现在同一群中，则此两个点就不再分了。

这样反复进行，直到所有的点群都聚合为一个群，形成一个分类系统。

一次计算成图法只需要计算一次相似性矩阵，计算比较简便。但是，群与群之间的联系只凭其中的两个点的相似程度，没有照顾点群的总体结构，因此当点群之间关系较复杂时，往往出现多解现象。此时，应运用逐步计算成群法。

二、逐步计算成群法

逐步计算成群法是在分群过程中，每聚合一次，都要重新计算新点群的代表性坐标和新点群与其它点群的相似性统计量，逐步计算，逐次形成新的点群。其具体计算步骤如下：

1. 计算相似性系数矩阵（初始矩阵 $S^{(0)}$），挑选出最相似的点对，把它们合并成群。

2. 把选出的点群原先的坐标值加权平均，计算得新群的坐标值。设第 k 次将第 i 点群与第 j 点群合并成 t 点群，t 点群的代表性坐标（重心）可用(2-20)式计算。

$$x_t^{(k)} = (n_i x_i^{(k-1)} + n_j x_j^{(k-1)})/(n_i + n_j) \quad , \qquad (2\text{-}20)$$

其中 n_i、n_j 分别为 i、j 点群中包含的个数。

3. 把数据矩阵中，第 $k-1$ 次的 i 点群和 j 点群的坐标值删去，代之以 t 点群的坐标值；

4. 根据 t 点群的坐标值，重新计算它与其它点群间的相似性统计量 $s_t^{(k)}$，刷新相似性矩阵，形成 $S^{(k)}$；

5. 重复上述过程，直至所有的点都聚合为一大群。

第四节　计算实例

本节将通过一个实际例子，说明聚类分析的计算步骤。

例子中的原始数据（表2-2），取自沈镛立、胡以铿等对陕西某超基性铜镍矿床的地球化学研究。

表2-2　陕西某超基性铜镍矿床钻孔中元素含量

组合样品号	样品组成	岩性及矿化	元素含量 (ppm)					
			Ni	Co	Cu	Cr	S	As
1	A孔10个样平均	蛇纹岩 A组矿化	1903	273	160	1178	8183	4
2	B孔11个样平均	蛇纹岩 无矿	2328	79	6	3175	536	14
3	C孔11个样平均	蛇纹岩 无矿	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	1775	94	13	3140	54	1
6	F孔12个样平均	滑镁岩 无矿	1046	44	6	2093	104	4

由于表中所列元素在岩石中呈对数正态分布，因此先将原始数据变换为对数值（表2-3）。

表2-3 陕西某超基性铜镍矿床钻孔中元素含量的对数值

元素 / 组合样号	Ni	Co	Cu	Cr	S	As
1	3.2794	2.4362	2.2041	3.0711	3.9119	0.6021
2	3.3670	1.8976	0.7782	3.5018	2.7679	1.1461
3	2.8716	1.4150	0.0000	2.9248	2.6284	0.4771
4	3.4443	2.4362	2.1761	3.3802	3.9156	1.5682
5	3.2492	1.9731	1.1139	3.4969	1.7324	0.0000
6	3.0195	1.6435	0.7782	3.3207	2.0170	0.6021

1. 用夹角余弦作为相似性统计量的一次计算形成法作Q型聚类分析

(1) 按照公式

$$\cos\theta_{ih}=\sum_{j=1}^{n}x_{ji}x_{jh}\Big/\Big[\sum_{j=1}^{n}x_{ji}^{2}\sum_{j=1}^{n}x_{jh}^{2}\Big]^{1/2},$$

计算得相似性统计量矩阵（表2-4）。

表2-4 陕西某超基性铜镍矿床组合样的余弦夹角

$\cos\theta_{ih}$	1	2	3	4	5	6
1	1.00					
2	0.9624	1.00				
3	0.9406	0.9848	1.00			
4	0.9907	0.9715	0.9352	1.00		
5	0.9357	0.9664	0.9409	0.9256	1.00	
6	0.9472	0.9806	0.9623	0.9516	0.9886	1.00

(2) 记下表中的最大值和相应点群对的号，即$\cos\theta_{h}^{(1)}=0.9907$，$i^{(1)}=4$，$h^{(1)}=1$。

(3) 挑选表中次大值和相应的点群号，$\cos\theta_{h}^{(2)}=0.9886$，$i^{(2)}=6$，$h^{(2)}=5$，并且检查这对点号是否已在前面分好的同一群中。现不在同一群中，故把它们记下。

(4) 再在表中挑选下一个次大值，$\cos\theta_{h}^{(3)}=0.9848$，$i^{(3)}=3$，$h^{(3)}=2$，同理检查其点群对是否在已分好的同一群中。

(5) 再在表中挑选下一个次大值，$\cos\theta_{h}^{(4)}=0.9806$，$i^{(4)}=6$，$h^{(4)}=2$，并检查其点群对是否在已分好的群中，但经检查 6 号点已与 5 号点合成一群，习惯上我们用群中号数最小的点代替，因此记$i^{(4)}=5$；2号点已与 3 号点合成一群。

(6) 再在表中挑选剩余的最大值，该值为0.9664，但经检查，它们的点群对 2 与 5 号已经于上一次合并成同一群，这时该值不用。

(7) 往下挑选剩余的最大值，得 0.9624，检查其点群对，$i^{(5)}=2$，且 2 号点已与 3、5、6号点合并成一群。检查$h^{(5)}=1$，1号点已与 4 号点合并成群，这时可以将该两群合并成一个大群，记下$\cos\theta_{h}^{(5)}=0.9624$，$i^{(5)}=2$，$h^{(5)}=1$。

到此为止已把所有的 6 个点聚合成一大群。总结以上步骤，可以列出一个点群连结顺序表，(表2-5)：

表2-5 点群连结顺序表

连结顺序	被连结的两群点号		相似性水平
1	1	4	0.9907
2	5	6	0.9886
3	2	3	0.9848
4	2、3	5、6	0.9806
5	1、4	2、3、5、6	0.9624

（8）据表2-5可作出对于6个组合样品的Q型聚类分析谱系图(图2-1)：

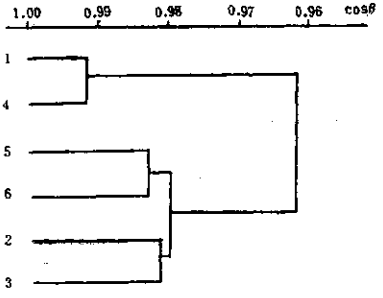


图2-1 Q型谱系图

（9）结果解释：由谱系图2-1与表2-2可以看出，6个组合样品间关系都比较密切(相似系数在0.90以上)。它们相互的聚合关系能客观地反映各组合样品的地质地球化学特征，它把有矿化1号与4号首先聚合为一群，把无矿的2、3、5、6号聚合为另一群。进一步又按岩性特征把蛇纹岩（2号和3号）和滑镁岩(5号、6号)分开成为两小群，从谱系图上可以直观地看到各组合样品的亲疏关系。

2．用相关系数作相似性统计量的逐步计算形成法做R型聚类分析

下面用上例的数据介绍逐步计算形成法的计算步骤。

（1）将表2-3的数据对各元素进行标准化得表2-6。

表2-6 陕西某超基性铜镍矿床钻孔组合样标准化数据

样号	Ni	Co	Cu	Cr	S	As
1	0.3738	1.2438	1.2999	0.9828	1.3007	-0.2805
2	0.8151	-0.1837	-0.5014	1.0185	-0.0733	0.3255
3	-1.6806	-1.4628	-1.4845	-1.6628	-0.2408	-0.5101
4	1.2045	1.2438	1.2645	-0.4535	1.3052	1.3682
5	0.2217	0.0164	-0.0773	0.9953	-1.3170	-1.4626
6	-0.9355	-0.8571	-0.5014	0.1770	-0.0751	-0.2605

（2）按标准化数据的相关系数计算公式计算得初始相关矩阵$R^{(0)}$：

$$R^{(0)}=\begin{array}{c|cccccc} & Ni & Co & Cu & Cr & S & As \\ Ni & 1 & & & & & \\ Co & 0.8462 & 1 & & & & \\ Cu & 0.7579 & 0.9802 & 1 & & & \\ Cr & 0.6431 & 0.2419 & 0.1811 & 1 & & \\ S & 0.5039 & 0.7370 & 0.7210 & -0.3075 & 1 & \\ As & 0.5603 & 0.4241 & 0.3930 & 0.1998 & 0.6802 & 1 \end{array}$$

（3）从$R^{(0)}$中选出相关系数最大的一对点群，记下$r_{Ni,Cu}=0.9802$。

(4) 用表2-6中数据计算$Co^{(1)}=\frac{Co+Cu}{2}$去替换Co的值，舍去Cu的值。即：$Co^{(1)}$=(1.2719、-0.3426、-1.4737、1.2542、-0.0305、-0.6793)。

(5) 重新计算$Co^{(1)}$与$Ni^{(0)}$、$Cr^{(0)}$、$S^{(0)}$、$As^{(0)}$间的相关系数，得

$$\mathbf{R}^{(1)}=\begin{array}{c} \\ Ni \\ Co^{(1)} \\ Cr \\ S \\ As \end{array}\begin{array}{c} \begin{array}{ccccc} Ni & Co^{(1)} & Cr & S & As \end{array} \\ \begin{bmatrix} 1 & & & & \\ 0.8019 & 1 & & & \\ 0.6431 & 0.2116 & 1 & & \\ 0.5039 & 0.7290 & -0.3075 & 1 & \\ 0.5603 & 0.4086 & 0.1998 & 0.6802 & 1 \end{bmatrix} \end{array}。$$

(6) 在$\mathbf{R}^{(1)}$中挑选相关系数最大的点群对，并记下$r_{Ni,Co}=0.8019$。

(7) 用$Ni^{(2)}=\frac{2\times Co^{(1)}+Ni}{3}$替换表2-6中Ni的数据，舍去$Co^{(1)}$，即：$Ni^{(2)}$=(0.9725、0.0433、-1.5427、1.2376、0.0536、-0.7647)。

(8) 计算$Ni^{(2)}$与剩余元素的Cr、S、As的相关系数，得

$$\mathbf{R}^{(2)}=\begin{array}{c} \\ Ni^{(2)} \\ Cr \\ S \\ As \end{array}\begin{array}{c} \begin{array}{cccc} Ni^{(2)} & Cr & S & As \end{array} \\ \begin{bmatrix} 1 & & & \\ 0.3554 & 1 & & \\ 0.6539 & -0.3075 & 1 & \\ 0.4591 & 0.1998 & 0.6802 & 1 \end{bmatrix} \end{array}。$$

(9) 从$\mathbf{R}^{(2)}$中选出相关系数最大的点群对，并记下$r_{S,As}^{(3)}=0.6802$。

(10) 用$S^{(3)}=\frac{S+As}{2}$替换表2-6中S的值，即：$S^{(3)}$=(1.5201、0.3761、-0.3755、1.4867、-1.3898、-0.6178)，并舍去As的值。

(11) 计算$S^{(3)}$与$Ni^{(2)}$、Cr间的相关系数，得

$$\mathbf{R}^{(3)}=\begin{array}{c} \\ Ni^{(2)} \\ Cr \\ S^{(3)} \end{array}\begin{array}{c} \begin{array}{ccc} Ni^{(2)} & Cr & S^{(3)} \end{array} \\ \begin{bmatrix} 1 & & \\ 0.3554 & 1 & \\ 0.5565 & -0.0538 & 1 \end{bmatrix} \end{array}$$

(12) 从$\mathbf{R}^{(3)}$中挑选相关系数最大者，并记下$r_{Ni^{(2)},S}^{(3)}=0.5565$。

(13) 用$Ni^{(4)}=\frac{3\times Ni^{(2)}+2\times S^{(3)}}{3+2}$替换表2-6中的$Ni^{(2)}$，即，$Ni^{(4)}$=(0.7915、0.1764、-1.0758、1.3372、-0.5234、-0.7039)，并舍去$S^{(3)}$的数据。

(14) 最后计算得$Ni^{(4)}$与Cr的相关系数$r^{(5)}=0.1918$。

总结以上步骤，列出连结顺序表2-7：

表2-7 R型聚类连结顺序表

连结顺序	被连结点群的元素		相似性水平
1	Co	Cu	0.9802
2	Ni	Co、Cu	0.8019
3	S	As	0.6802
4	S、As	Ni、Co、Cu	0.5565
5	Cr	Ni、Co、Cu、As、S	0.1918

(15) 根据表2-7作出R型聚类分析谱系图2-2：

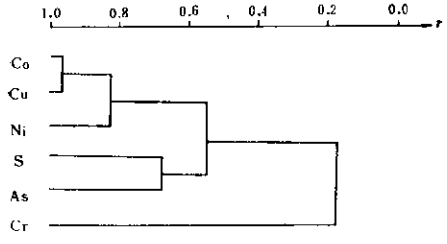


图2-2 R型聚类谱系图

(16) 结果的地质地球化学解释

由谱系图可以看出，在相关系数为0.4的水平上，可将元素分为两群：一群是亲硫的元素Co、Cu、Ni、S、As；另一群是亲氧元素Cr。在相关系数为0.6的相似性水平上又可将亲硫元素组合分为两组：一组是亲硫的阳离子Co、Cu、Ni；另一组是阴离子元素S和As。同时还可以看出，在Co、Cu、Ni三元素中，Co-Cu之间的关系更为密切，结合矿物学研究，Co一般以类质同象的形式存在于含Co的黄铜矿和黄铁矿中，而Ni可以单独形成针镍矿。从以上的结果可看出，对元素的R型聚类分析很好地揭示了这些元素在超基性铜镍矿床中的相互关系。

第五节 逐步计算成图法

上面已经介绍了在m维空间中的n个点可以用某种相似性统计量S_{ij}来度量两两点之间的相似性程度。在逐步计算形成法中，每合并一对点群后，都要计算新点群的代表性坐标。前面介绍的是用重心法，即以新点群的重心坐标作为它们的代表。实际上根据聚类时考虑问题的侧重点不同，新点群的代表性坐标的计算是多种多样的，其主要的算法有：①最短距离法，②最长距离法，③中间距离法，④重心法，⑤类平均法，⑥可变法，⑦可变平均法，⑧离差平方和法等。

在计算新的点群与其它点群的相似性统计量时，一般采用迭次推算成群法，即直接由合并成新点群的原来两个点群与其它点群之间的相似性统计量来推算新点群与其它点群之间的相似性统计量。

设由点群p和点群q合并为新点群t，则t点群与其它点群i之间的相似性统计量可用如下统一的递推公式求得：

$$S_{it}^2=\alpha_p S_{ip}^2+\alpha_q S_{iq}^2+\beta S_{pq}^2+\gamma\left|S_{ip}^2-S_{iq}^2\right|, \qquad (2\text{-}21)$$

其中α_p、α_q、β、γ的系数见表2-8：

表2-8 系统聚类算法参数表

方法	α_p	α_q	β	γ	附注
最短距离法	1/2	1/2	0	-1/2	
最长距离法	1/2	1/2	0	1/2	
中间距离法	1/2	1/2	$-1/4\leq\beta\leq 0$	0	
重心法	n_p/n_t	n_q/n_t	$-n_p n_q/n_t^2$	0	必须用欧氏距离
类平均法	n_p/n_t	n_q/n_t	0	0	
可变法	$(1-\beta)/2$	$(1-\beta)/2$	<1	0	
可变类平均法	$(1-\beta)\cdot n_p/n_t$	$(1-\beta)\cdot n_q/n_t$	<1	0	
离差平方和法	$\frac{n_k+n_p}{n_k+n_t}$	$\frac{n_k+n_q}{n_k+n_t}$	$-\frac{n_k}{n_k+n_t}$	0	必须用欧氏距离

其中：n_p、n_q、n_t分别为p点群、q点群和t点群中样品点的个数，且$n_t = n_p + n_q$。

第六节　聚类分析在地质上的应用

一、在地球化学工作中研究元素的组合关系或亲和能力

1．华南花岗岩微量元素间的共生组合关系

中国科学院贵阳地化所对华南花岗岩采集了103个标本（其中四堡-雪峰有16个，加里东期22个，海西期12个，燕山期53个）。每个标本测定了20个变量（即Nb、Ta、TR、Zr、Ba、Be、Co、La、Cr、Cu、Ga、Pb、Sn、Ni、Sr、V、Y、Yb、Ti、F）。相似性统计量采用相关系数，按相关系数从大到小排列的各元素对的顺序列于表2-9。

表2-9　华南花岗岩元素对相关系数表

元素对	相关系数	元素对	相关系数
Y, Yb	0.79	Nb, Ta	0.45
Ba, Sr	0.77	Co, Cr	0.45
V, Ti	0.68	TR, La	0.43
Ba, V	0.63	Nb, F	0.38
Ca, Ni	0.60	Zr, Co	0.32
Ga, Sn	0.58	Be, Ga	0.24
Ga, Pb	0.57	Nb, Be	0.07
Ga, Y	0.51	Nb, TR	0.02
Zr, Ba	0.49	Zr, Cu	0.01

对华南不同时代花岗岩中的微量元素进行R型聚类分析，其结果作成谱系图2-3。由图可以清楚地看出，20种微量元素明显地分成两大组，细分为7个亚组，12个次亚组（表2-10）。

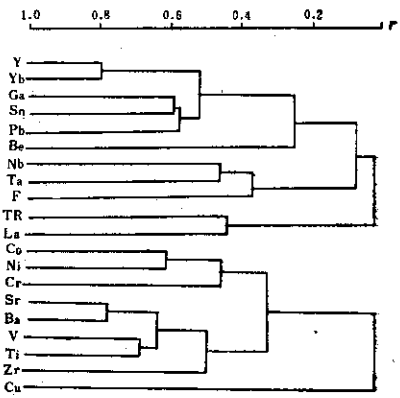


图2-3　华南花岗岩类微量元素聚类分析谱系图

第一大组所包括的元素有Y、Yb、Ga、Sn、Pb、Nb、Ta、F、Be、TR、La等11种。这一大组主要是稀有元素、稀土元素、有色金属和挥发性元素组合，这些元素是酸性岩浆所具有的典型元素。在华南花岗岩类中，从四堡-雪峰期到燕山期，其含量呈有规律地递增。

第二大组所包括的元素为Sr、Ba、V、Ti、Zr、Co、Cr、Cu、Ni等9种。这一大组以铁族元素为主，它们是基性岩浆所具有的典型元素，与第一大组恰恰相反，在华南花岗岩类中，从四堡-雪峰到燕山期，其含量逐渐减少。

表2-10 华南花岗岩类微量元素聚类分析分组

组	元 素	亚 组	元 素	次 亚 组	元 素
I	Y, Yb, Ga, Sn, Pb, Nb, Ta, F, Be, TR, La	1	Y, Yb, Ga, Sn, Pb	1	Y, Yb
				2	Ga, Sn, Pb
		2	Be	3	Be
		3	Nb, Ta, F	4	Nb, Ta
				5	F
		4	TR, La	6	TR, La
I	Co, Ni, Cr, Sr, Ba, V, Ti, Zr, Cu	5	Co, Cr, Ni	7	Co, Ni
				8	Cr
		6	Sr, Ba, V, Ti, Zr	9	Sr, Ba
				10	V, Ti
				11	Zr
		7	Cu	12	Cu

在0.35相似性水平上，第一大组又可分为4个亚组，第二大组可分为3个亚组。

第一亚组为重稀土元素Y、Yb和Pb、Sn、Ga。

第二亚组为稀有元素Be。

第三亚组为稀有元素Nb、Ta与挥发性元素F的组合。这种组合反映了在华南花岗岩类中Nb、Ta与F相关的显著特征。

第四亚组为TR与La的组合，即为轻稀土元素的组合。从对华南花岗岩中稀土配分来看，稀土总量以富轻稀土为特征。在聚类分析的结果中分出一个轻稀土元素组合。

第五亚组为铁族元素Co、Ni、Cr的组合。

第六亚组主要为铁族元素V、Ti和造岩元素Sr、Ba的组合。

第七亚组为Cu，是典型的亲硫元素，在花岗岩中主要呈硫化物存在。

华南花岗岩类微量元素的聚类分析结果把岩浆演化过程中地球化学行为截然相反的元素明显地分成两大组，并进一步分为7个亚组，12个次亚组。这就从数学的角度进一步揭示了在岩浆演化过程中微量元素间的内在联系。元素的原子核及核外电子排布的特征是决定元素地球化学行为的根本因素。实验证明，重放射性核受高能粒子轰击时，裂变产物就是稀土元素及其邻序元素。重放射性核衰变最终产物也主要是铅。而稀土、铌、钽、铍等元素，在地球化学迁移过程中与挥发性元素(如氟)组成易溶络合物已被大量实验所证实。因此，在第一大组中，稀土、铅、铌、钽、铍、氟出现在同一组合中并非偶然。

侯德封在研究地质历史各成矿期及内生矿床的成矿顺序时指出，成矿过程存在着两个基本系统，一是与过渡元素（Fe、Ti、Cr、Co、Ni等）有关的元素及Au、Cu元素系统；二是铀-钍-稀土及其外延的铌、钽系统。这两个基本系统在时间上从老到新逐渐发展着，贯穿整个地质发展史及每个成矿期。应用聚类分析对华南花岗岩微量元素的研究和这一结论十分吻合，这说明聚类分析可以从统计的角度为元素的共生组合研究提供有意义的信息。

二、利用聚类分析对研究对象进行分类

在地球化学工作中常常要对地球化学异常、岩体、铁帽、样品等进行分类。这种分类工作可采用聚类分析中的Q型分析。

这里举一个对铁帽分类的例子。我国南方某区地处亚热带，铁帽分布广泛，种类很多。从已掌握的资料来看，这些铁帽主要由矽卡岩型磁铁矿、多金属硫化物、黄铁矿、菱铁矿、赤铁矿等矿物集中的地质体风化而成。由于强烈的表生改造作用，用一般的地质工作方法对铁帽进行检查评价有一定的困难。

利用该区铁帽样品所收集到的 Cu、Pb、Zn、Sn、As、Co、Ni 等7种元素的光谱半定量分析资料，取每个铁帽的平均值并进行对数变换，见表2-11。

表2-11　某地铁帽的光谱分析数据对数值

铁帽编号	Cu	Pb	Zn	Sn	As	Co	Ni
1	1.9798	2.3883	3.1098	3.4771	3.2813	0.5543	0.9361
2	1.9769	2.2878	2.9486	2.4279	2.6778	0.7720	1.0391
3	3.0540	3.2492	2.4355	2.8961	3.4771	0.3010	1.3127
4	2.0487	2.4631	3.0481	0.6990	3.0104	0.6990	2.3160
5	2.1991	0.3070	1.7385	0.3010	1.1391	0.3010	0.8891
6	2.1885	1.0338	1.8851	0.3010	3.1191	0.3010	1.1169
7	1.6848	1.1414	1.6019	0.3452	2.8654	0.8101	2.000
8	1.0003	0.7007	0.8287	0.3673	2.1875	0.3010	0.3835
9	1.1363	0.7850	1.3798	0.3010	2.4724	0.3010	1.1363
10	1.3316	0.8291	1.7914	0.4186	0.3010	0.3010	0.1176
11	1.5749	1.8170	2.4575	0.3360	2.8213	0.5248	1.2170
12	1.8130	2.9558	1.8632	0.3010	3.2659	0.3010	0.4591
13	1.0000	0.5000	1.3266	1.0000	0.8138	0.3010	1.8011
14	1.0000	0.5683	1.9260	0.3010	0.6930	0.3010	1.5511
15	1.0000	0.3598	1.7468	0.3010	2.2953	0.3010	1.5139
16	1.0000	0.7262	2.4834	0.5103	1.8369	0.6564	1.3266
17	2.9910	1.4785	1.8623	0.8770	3.2688	0.6990	1.8493
18	1.8968	0.7315	2.6499	0.6990	1.2088	0.6990	2.1993
19	3.0798	0.8356	1.8640	0.8224	3.1878	0.7457	0.9221

对上述数据采用标准化后，以离差平方和为相似性统计量做Q型聚类分析，其结果见谱系图2-4。

从谱系图中可以看出：大约在离差平方和为15的水平上，可将19个铁帽分成5组。A组有17、19号铁帽，它们是多金属硫化物型铁帽；B组的6、11、12号铁帽为黄铁矿型；C组由16、18、7、4号铁帽，它们都是菱铁矿型；D组的1、2、3号铁帽为矽卡岩磁铁矿-多金属硫化物型，其余的大致为赤铁矿-黄铁矿型铁帽。

从以上结果看，聚类分析可以对该区铁帽的分类和研究提供线索。若收集铁帽中的常量元素和特征资料，作进一步分析，可望得到更为满意的结果。

三、聚类分析在化探工作中的应用

在英格兰达比郡南部灰岩隆起区的区域分散流测量中，采用了多元素分析方法。该地

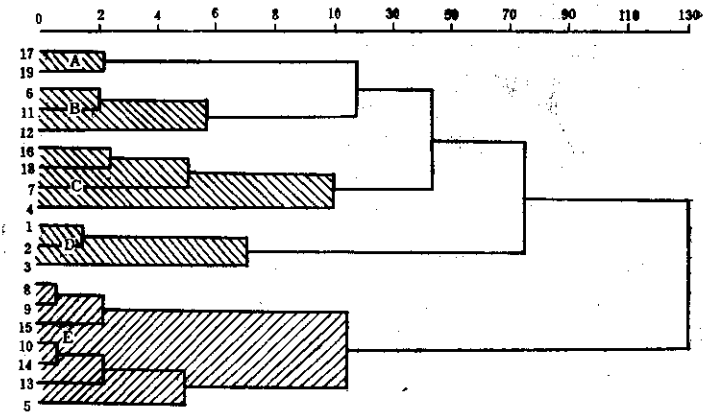


图2-4 某地铁帽Q型聚类谱系图

区有远景的矿化广泛分布在石炭系灰岩的铅-锌矿脉中，但被时代更新的厚层沉积物所覆盖。该区的地球化学工作者通过对该区河流沉积物样品中铜、铅及锌含量的分析，圈出上覆的艾达尔页岩中可能存在的次生异常，并根据次生异常进一步指导找矿。

开展这种区域性多元素测量工作，需要综合分析大量数据，用主观的肉眼观察方法处理数据，很繁琐，费时间，结果又常常不理想。本实例应用聚类分析方法和电子计算机处理化探数据取得了良好效果。

1．区域地质条件

研究区位于英格兰达比郡南部。石炭系灰岩出露于测区的最北部，向南为时代较新的石炭系地层，再向南地层便尖灭在三叠系地层和冰碛层之下。北部的灰岩发生了白云岩化，这种白云岩化灰岩是昆达型矿床的成矿母岩。测区内的矿化有两种类型，第一种类型出现在测区北部的白云岩化灰岩中(图2-5)，矿石由方铅矿和重晶石及少量萤石与微量的闪锌矿、黄铜矿、辉铜矿组成；第二种矿化沿第三系及其下伏地层间的不整合面出现，矿石由方铅矿、白铅矿、重晶石、辉铜矿、铜的碳酸盐类矿物、钒的化合物和某些含铀的碳氢化合物组成。

在测区100mile²*的面积上以每mile²1.7个样品的平均采样密度共采集了170个样品。80网目的全部样品用直读光谱分析了Al、Ca、Mg、K、Fe、Ti、Mn、Ba、Co、Cr、Cu、Ga、Li、Ni、Pb、Sr、Zn、Zr，还用比色分析测定了冷提取的Zn。在进行聚类分析时对原始数据进行了对数变换和标准化处理。

2．R型聚类分析

用相关系数为相似性统计量所作的R型聚类分析结果见谱系图2-6。在$r=0.45$的相似性水平上，出现6个清楚的点群：①Al_2O_3、Ga、K_2O、TiO_2、MgO、Cr、Zr；②Fe_2O_3、Li、V、Mn、Co、Ni、Cu；③CxZn、Zn；④CaO；⑤Ba、Sr；⑥Pb。点群①反映粘

* $1mile^2=2.59\times10^6m^2$

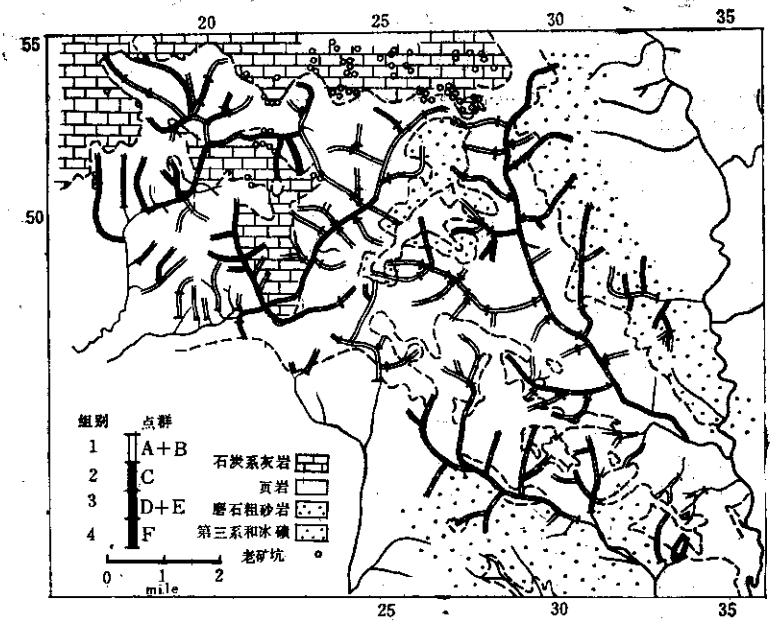


图2-5　矿区地质和Q式聚类分析结果

土矿物和风化残余矿物的元素组合；点群②是与Mn和Fe的氧化物有关的元素组合；点群③则是反映和矿化有关的元素组合；④、⑤和⑥是3个相关性比较小的独立组合。

CaO主要来自石灰岩，而Ba、Sr和Pb则经常出现于相对较贫CaO的岩性中，二者的表现显然不同。此外，Pb、Ba和Sr是与矿化有关的最重要元素。

Mn和Fe的点群中元素的分布模式密切相似，值得注意的是它们出现在某些异常中，根据已有资料，Mn和白云岩化灰岩相伴，并且以往曾进行过开采。因而验证所圈定的异常究竟是和矿化有关还是与矿化无关的局部富集所引起是一个重要问题。

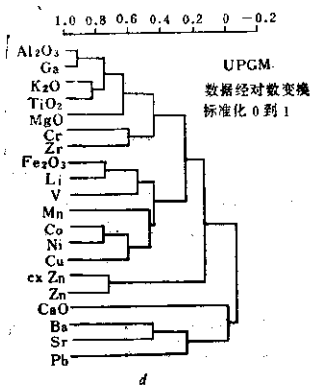


图2-6　R型分析谱系图

本区已知矿田中Pb和Zn之间的紧密组合关系在次生环境中已不再保持，这时Zn已与Pb以外的其他元素密切共生，这是由于风化环境中Zn的活动性高于Pb所致，另一原因则是Zn易于富集在Mn的氧化物中。Zn虽然可能富集于

Mn的氧化物中，但仍不失为矿致异常的指示元素。

R型谱系图的一个重要特征是化学元素组合和母体的分布特征之间存在某种联系。例如，第一个点群的元素，其频率分布基本上是单峰的，Cr是其中的唯一例外。第二个点群除Mn以外，其频率分布大多是双峰的。第三个点群的元素反映更复杂的多峰母体。

3．Q型聚类分析

通过Q型聚类，将170个样品分成4大组（图2-5）。第一组包括A和B两个点群；第二组为C点群；第三组包括D、E两个点群；第四组由F点群组成。由聚类分析得出的样品地区性分组反映了该区的基本岩性。重叠在这种地区性分组之上的，是由F点群所表示的强异常，（A+B）圈出了主要的页岩区，C点群圈出灰岩区，（D+E）表示米斯顿-格里特（磨石粗砂岩）及三叠系与德里费特（冰碛层）地区，它们都是砂质的。F点群所表示的强异常与已知的老矿点有关。

4．结论

通过对英格兰达比郡南部分散流数据进行聚类分析可以得以下结论：

（1）聚类分析对于研究元素的共生组合很有效，能用于揭示受到岩性、外生环境和矿化作用影响的复杂的元素分布模型。如本例中，用R式聚类分析将元素分成三大类，第一类为与粘土矿物和耐风化物有关的元素（Al、Ga、K、Ti、Cr、Zr、Mg、Ca）；第二类为与Mn、Fe伴生的元素（Fe、Mn、Ni、Co、V、Li、Cu），这些元素和岩性之间有着明显的关系；第三类是与主要矿化类型有关的元素（Ba、Sr、Pb、Zn），这些元素很分散，彼此之间的相关程度很低，但在灰岩区附近含量都显著增高。

（2）异常样品数量少，而且和其余样品不相似，因而通过Q型聚类分析可以将低相似性的划分出来。

第七节　聚类分析中一些问题的讨论

一、适用范围

聚类分析总的来说是对研究对象进行分组归类的一种多元统计分析分法。它在应用时不受工作区现在研究程度深浅的限制，可以用于研究程度不高的或研究对象的归属暂时不明的情况，也可以用于研究程度高、被分类对象的归属已知的情况，还可以用于部分样品分类已知而其余部分未知的情况，后一种情况有利于聚类分析结果的地质解释，即可以根据已知归属的样品来判断未知样品的归属。

聚类分析可用定量数据，也可以用半定量或定性数据，它不要求数据服从某种已知的频率分布或正态分布。这大大扩大了可能解决问题的范围。实际工作时，可以将R型和Q型聚类分析联合使用，往往先用R型聚类分析挑选出不相关的变量（正交变量），然后用这些正交变量作Q型聚类分析。

二、对变量的讨论

聚类分析中把各个样品的不同地球化学性质（变量）看作是同等重要的，即具有等价性。例如，pH、Eh、温度、元素的平均含量等等，一旦把它们选为一个变量，在聚类分

析中都平等对待，都看成是一条独立于其它变量的坐标轴。如元素含量，不论是微量元素还是主量元素(它们的绝对含量可以相差数千倍)，也不管其地质意义和经济价值的差别，在聚类分析中都一视同仁，它们对分类的贡献只是从变量本身的离差和判别能力上表现出来。这种等价性有时是与变量的地球化学意义的重要程度不一致。为了克服这个缺点，有人提出将变量按其重要程度的不同加权。但权系数的确定有很大的人为性。

从纯数学的意义上说，聚类分析要求各变量必须相互独立，互不相关。在用距离系数作为相似性度量时，欧氏距离的基本假定是坐标系统为正交坐标系。在实际问题中，多个变量之间往往彼此存在着不同程度的相关关系。在这种情况下，用欧氏距离来计算样品点群的空间距离，就会使空间中样品的距离发生畸变，从而使点群簇分时谱系的分群结构发生变化。为了避免这种现象，在数学上一般采用两种途径处理：一是在进行聚类分析之前，先把原始变量作主成分分析，将有相关关系的变量综合成主成分，经过主成分分析后，各主成分之间就相互独立了；另一种方法是用斜交距离系数来代替欧氏距离系数。

在处理实际问题时，地质工作者并不强调变量之间必须严格遵循相互独立的条件。相反，有时用几个地质意义比较清楚的、对所研究的问题有较强分辨力的相关变量参加聚类分析，反而能得到较好的分类效果。

三、聚类分析结果的合理利用

对于聚类分析所得的结果，不应只满足于通过按步就班的运算得到一个谱系簇分的结果去牵强附会地进行解释，而应当结合具体的研究对象，通过对聚类分析结果的仔细研究去发现存在于原始数据中过去未被注意的规律性，进一步对已知事实或现象提出新的解释，甚至可以根据新发现的规律提出新的假设。聚类分析所得的分类模型本身只是一种可能的结果，而并非一种无可怀疑的定论。聚类分析的结果可以给地质工作者一种"启发"，不同方法的分类结果只是存在于原始数据之中的真实分类的某种程度的反映。各种分类结果的比较和相互补充对所研究对象间的亲疏关系可以得到更全面和更深入的认识。

聚类分析的效果好坏首先依赖于原始数据，即样品的代表性和变量的选择，其次也受所采用的数学方法、步骤和某些假设是否正确所影响。当我们以聚类分析作为研究实际问题的工具时，往往事先很难估计究竟应选择什么变量，什么样品，什么组合变量，采取何种分析方法和假设才能得到所需要的正确分类结果。因此，在做了一轮聚类分析后不要轻易地马上下某种结论。当初次的分析结果与实际情况有一定出入，甚至有严重矛盾时，可以对原设计方案进行某些修改，再进行第二轮分析，使得结果更符合所研究问题的实际情况。如此循环多次，不断改善分类结果，直至获得满意的结果为止。

对同一批原始数据采用不同方法所得的分类结果差异较大，这很可能是原始数据中并无点群存在所造成的。如对样品进行聚类分析时，若无强分辨力的变量，样品点在变量空间中则呈相对均匀分布。在对变量进行分类时，各变量相互独立就会导致这一结果。有时虽然样品采自不同地质单元，但由于实验室中采用半定量的测试方法，使原始数据的分析误差掩盖了不同样品之间所存在的差异，将会导致聚类分析结果与实际研究问题之间的矛盾和解释上的困难。

第八节　其它几种聚类分析方法

上述介绍的聚类分析的基本思想是将n个个体各自成为一类，依据它们之间的相似程度，每一次合并一对点群，直至所有的点都合为一大群，形成一个由小到大的分类系统。这种方法称为系统聚类分析。作为系统聚类的改进，根据变量性质的不同，研究目的的不同，数学地质工作者还设计了动态聚类、模糊聚类和有序样品的聚类分析等方法。

一、动态聚类

由于系统聚类分析计算工作量大，在电算中要求储存相似性矩阵，当样品个数n很大时占用机器内存太多，甚至超过电子计算机的容量而无法计算。另外，从计算过程可以看到，系统聚类法一旦把某个样品划到一个点群以后，就一成不变了。为了改进聚类分析，一个想法是先粗略地进行分类，然后按照某种分群原则进行修正，直至分类比较合理为止。这就提出了动态聚类法，这种方法首先选择一批凝聚点，让样品按某种原则向这些凝聚点聚集形成初始分类，经过一轮初始分类后，修改凝聚点的坐标和个数，重新对全部样品点按新的凝聚中心聚合分类，反复进行，直至前后两次分类结果基本一致为止。

下面介绍麦奎因（Macqueen）所提出的方法，有人称为“K-means”，其步骤如下：

1．首先人为地定出一个分类数目K（当然可以根据实际情况预计），并任取K个样品作为凝聚点。

2．将其余（$N-K$）个样品分别归入到最近凝聚点的一类中，形成初始分类。

3．计算各类的重心，以重心代替凝点，再将N个样品重新归入最近凝聚点的一类。

4．再计算各类重心，重新分类，直至前后两次分类的结果相同为止。

这种方法计算简单，适合于对样品数目较大的作业。但是分类结果的好坏以及计算时收敛速度与初始凝聚点的选择有很大关系。如果第一次选取的K个凝聚点比较集中，则分类效果不会很好。一种改进的办法是在修改分类过程中，类的数目可以根据实际情况有所变化，太近的类可以合并，太远的样品可以分离出来产生新的类。

为了克服K-meams的缺点，提出了等混合法（Isomix法）。其基本步骤如下：

1．选取控制参数

（1）初始分类数m_0：可以根据实际预计选取n个。

（2）组内最少样品数N_S：一般根据要求分类的详细程度和样品总数N的大小来确定。

（3）组内各变量的最大标准差S_T：S_T越小分类越细。一般取变量最大标准差作为控制某组是否需要进行分裂的参数。

（4）组与组间的最短距离D_S：作为控制任意二组是否需要合并的参数。D_S越小，分组越细。

2．确定初始分组中心（即凝聚点），公式为：

$$g_{kj}=\mu_j+\left[\frac{2k-1}{m_0-1}-1\right]\sigma_j\,, \tag{2-22}$$

式中：g_{kj}—第k组j变量的坐标，$j=1,2,\cdots,p$；$k=1,2,\cdots,m_0$，

μ_j—j变量的平均值，

σ_j—j变量的标准差，

m_0—初始分组数。

3．初始分组

初始分组中心确定后，依次计算样品点距各分组中的距离d_{igk}，公式为：

$$d_{igk}=\sqrt{\sum_{j=1}^{p}(x_{ij}-g_{kj})^2}\quad , \tag{2-23}$$

式中：i—样品号，$i=1, 2, \cdots, n$，

j—变量号，$j=1, 2, \cdots, p$，

k—分组号，$k=1, 2\cdots, m_{g0}$

比较各样品距各分组中心的距离，根据择近原则进行归类。

4．修改分组

初始分组后得m_0组，这是相当粗糙的分类，因此需要调整，调整的方式有两种：

（1）分裂：计算每组各变量的平均值和标准离差，若任意组（k）任一变量（j）的标准差大于给定的参数S_t，这就说明本组内的样品过于分散，需要分裂。重复上述步骤，计算进一步分组后的中心。

（2）合并：计算组与组间的距离D_{pq}。用两组重心之间的距离来度量。计算公式如下：

$$D_{pq}=\sqrt{\frac{\sum_{j=1}^{p}(u_{pj}-u_{qj})^2}{\sigma_{pj}\cdot\sigma_{qj}}}\quad 。 \tag{2-24}$$

若任意两组间的距离$D_{pq}<D_t$（给定参数），则将该两组合并为一组。

根据上述原则不断分裂与合并，直到每一类都不需要再分裂或合并为止。最后得到分类树枝图，并进行地质解释。

二、模糊聚类

地质变量中存在有大量的模糊变量，对模糊变量的处理方法，已成为一门新的数学分支——模糊数学。美国控制论专家L.A.莱登（Zadan）于1965年首次提出模糊集合。因模糊的原文为“Fuzzy”，故又称“费晰数学”。

模糊数学的理论基础是模糊集合。模糊集合中用隶属函数来代替普通集合中绝对划分归属的关系。

隶属函数：给定论域U，其元素为u，A为U中的一个子集，对任意$u\in U$，都有一个表示归属程度的μ作为模糊集A的隶属函数，记作$\mu=A(u)$。隶属函数为闭区间〔0，1〕中任意实数值。这样，隶属函数把普通集合论中只取0、1两值推广到〔0，1〕闭区间。

模糊集合的运算：设A、B为域U的两个模糊子集，则它们的并集、交集和A的余集（A^c）都是模糊子集，它们的隶属函数分别定义为：

$$(A\vee B)(u)=\max[A(u)\cdot B(u)]\quad , \tag{2-25}$$

$$(A\wedge B)(u)=\min[A(u)\cdot B(u)]\quad , \tag{2-26}$$

$$A^c(u)=1-A(u)\quad 。 \tag{2-27}$$

模糊集没有明确的外延，它的边界是模糊的。然而在普通的数学中，描述一个概念应

该是一个明确的值。将模糊概念转化为明确的数学语言的办法是根据不同的置信水平 λ 值来确定其隶属关系$0\leqslant\lambda\leqslant1$。如当$A(u)\geqslant\lambda$时，就可以把研究对象 u 算作$u\in A$，否则$u\tilde{\in}A$。故 λ 值从 1 降到0，则A逐渐扩大。因此，模糊子集A是一个具有游移边界的集合，它随 λ 值的变小而增大。

在模糊数学中，对有限集合X的分类就是将X的内容分为T个子集。从集的概念出发，这种分类关系都具有下述三条性质。

(1) 反身性：$(x_i, x_i)\in R$，即集合中每个元素，自己与自己同属一类。

(2) 对称性：若 $(x_i, x_g)\in R$，则 $(x_g, x_i)\in R$，即集合中 i 元素与 g 元素同属于一类 (R) 时，g元素也必然与 i 元素同属于一类 (R) 。

(3) 传递性：若$(x_i, x_k)\in R$，$(x_g, x_k)\in R$，则$(x_i, x_g)\in R$，即集合中，若 i 元素与 k 元素同属于R 类，g 元素也与 k 元素同属R类，则 i 元素与 g 元素也同属于R类。

这三条性质称为等价关系。

模糊聚类方法的具体步骤如下：

1．模糊相似性标定

模糊相似性统计量常用的有模糊相似系数r_{ig}，模糊距离d和贴近度 ρ 三种。

(1) 模糊相似系数r_{ig}

它由取值在〔－1，1〕区间的普通相似系数$\cos\theta$的值“压缩”到〔0，1〕区间内的数而得，即

$$r_{ig}=\frac{1+\cos\theta}{2}\quad。\tag{2-28}$$

计算的 r_{ig} 可列成矩阵R，称为模糊相似系数矩阵。模糊相似系数矩阵R 除应具有反身性和对称性外，还一定要具有传递性，所以要对R作有限次褶积，使得

$$\mathbf{R}^n\circ\mathbf{R}^n=\mathbf{R}^n\quad。\tag{2-29}$$

如用 6 个样品 5 个变量作Q型聚类，经计算得$\cos\theta$矩阵如下：

$$\cos\theta=\begin{pmatrix}1 & 0.06 & -0.25 & -0.09 & 0.28 & 0.20\\ & 1 & -0.25 & -0.09 & 0.06 & -0.09\\ & & 1 & -0.25 & -0.25 & 0.06\\ & & & 1 & -0.09 & 0.10\\ & & & & 1 & 0.18\\ & & & & & 1\end{pmatrix}$$

由 (2-28) 式，得

$$\mathbf{R}=\begin{pmatrix}1 & 0.53 & 0.38 & 0.45 & 0.64 & 0.60\\ & 1 & 0.38 & 0.45 & 0.53 & 0.45\\ & & 1 & 0.38 & 0.38 & 0.53\\ & & & 1 & 0.45 & 0.55\\ & & & & 1 & 0.59\\ & & & & & 1\end{pmatrix}$$

将R 进行褶积，得$\mathbf{R}^2$。$\mathbf{R}^2$中元素是R 中元素用公式 (2-25) 、(2-26) 逻辑运算而得。如：

$$r_{11}^2=(1\wedge 1)\vee(0.53\wedge 0.53)\vee(0.38\wedge 0.38)\vee(0.45\wedge 0.45)\vee(0.64\wedge 0.64)\vee(0.60\wedge 0.60)$$
$$=1\vee 0.53\vee 0.38\vee 0.45\vee 0.64\vee 0.60=1,$$
$$r_{12}^2=(1\wedge 0.53)\vee(0.53\wedge 1)\vee(0.38\wedge 0.38)\vee(0.45\wedge 0.45)\vee(0.64\wedge 0.53)\vee(0.60\wedge 0.45)$$
$$=0.53\vee 0.53\vee 0.38\vee 0.45\vee 0.53\vee 0.45=0.53\ .$$

$$\mathbf{R}^2=\mathbf{R}\circ\mathbf{R}=\begin{pmatrix} 1 & 0.53 & 0.53 & 0.55 & 0.64 & 0.60\\ & 1 & 0.53 & 0.45 & 0.53 & 0.53\\ & & 1 & 0.53 & 0.53 & 0.53\\ & & & 1 & 0.55 & 0.53\\ & & & & 1 & 0.59\\ & & & & & 1 \end{pmatrix}\circ$$

再进行褶积

$$\mathbf{R}^4=\mathbf{R}^2\circ\mathbf{R}^2=\begin{pmatrix} 1 & 0.53 & 0.53 & 0.55 & 0.64 & 0.60\\ & 1 & 0.53 & 0.53 & 0.53 & 0.53\\ & & 1 & 0.53 & 0.53 & 0.53\\ & & & 1 & 0.55 & 0.55\\ & & & & 1 & 0.59\\ & & & & & 1 \end{pmatrix},$$

$$\mathbf{R}^8=\mathbf{R}^4\circ\mathbf{R}^4=\begin{pmatrix} 1 & 0.53 & 0.53 & 0.55 & 0.64 & 0.60\\ & 1 & 0.53 & 0.53 & 0.53 & 0.53\\ & & 1 & 0.53 & 0.53 & 0.53\\ & & & 1 & 0.55 & 0.55\\ & & & & 1 & 0.59\\ & & & & & 1 \end{pmatrix}\circ$$

此时$\mathbf{R}^8=\mathbf{R}^4$，说明$\mathbf{R}^4$已具有传递性，停止褶积。利用$\mathbf{R}^4$就可以对样品进行谱系簇分。

(2) 模糊距离d

两个模糊子集A、B间的距离称为模糊距离$d(A,B)$，其计算公式为：

$$d(A,B)=\|A\cup B\|-\|A\cap B\|\ ,\qquad (2\text{-}30)$$

即A与B之间距离为A，B并的模与A，B交的模之差。

例如，利用一个矿区测井数据（3个参数）解释钻孔内的岩性（6种）。先利用一些钻孔的数据作3个参数的回归分析，计算6种岩性的回归估值$\hat{y}$分组区间(表2-12)

表2-12 某矿区6种岩回归估值范围

岩性	1	2	3	4	5	6
$\hat{y}$	1.0—0.71	0.95—0.58	0.85—0.55	0.78—0.32	0.75—0.24	0.6—0

应用公式（2-30）可计算不同岩性之间的模糊距离d，如

$$d_{12}=\|(1-0.58)\|-\|(0.95-0.71)\|=0.42-0.24=0.18\ ,$$
$$d_{13}=\|(1-0.55)\|-\|(0.85-0.71)\|=0.45-0.14=0.31\ .$$

$$D=\begin{Bmatrix} 0 & 0.18 & 0.31 & 0.61 & 0.71 & 0.89 \\ & 0 & 0.10 & 0.43 & 0.54 & 0.93 \\ & & 0 & 0.30 & 0.34 & 0.57 \\ & & & 0 & 0.11 & 0.50 \\ & & & & 0 & 0.39 \\ & & & & & 0 \end{Bmatrix}。$$

(3) 贴近度ρ

模糊子集A与B的贴近度$\rho(A,\ B)$定义为：

$$\rho(A,\ B)=\begin{cases} \|A\cap B\|/\|A\cup B\| & 当\|A\cup B\|\neq 0， \\ 0 & 当\|A\cup B\|=0。 \end{cases} \tag{2-31}$$

即模糊子集A与B的贴近度为A与B交的模与它们并的模之比，当它们并的模为0时，取值0。

例如上例数据的贴近度矩阵为

$$\rho=\begin{Bmatrix} 1 & 0.571 & 0.311 & 0.103 & 0.053 & 0.00 \\ & 1 & 0.675 & 0.317 & 0.230 & 0.021 \\ & & 1 & 0.434 & 0.328 & 0.059 \\ & & & 1 & 0.837 & 0.359 \\ & & & & 1 & 0.480 \\ & & & & & 1 \end{Bmatrix}$$

在模糊相似系数、模糊距离和贴近度三个相似性统计量中，模糊相似系数和贴近度的数愈接近1，说明两个子集愈相似。而模糊距离则反之，其数值愈接近0。说明两个子集愈相似。

有了模糊子集的相似性矩阵，可仿照系统聚类进行谱系簇分。

三、有序样品的聚类——最优分割

许多地质数据是有序的，这些按一定顺序排列的地质变量，叫做有序地质变量。如，沿地层剖面采集的样品；钻孔中取出的岩芯样品；蚀变带的研究中，从岩体中心到围岩的蚀变剖面的样品等等，它们都是有序地质变量。

这类数据的特点是样品的前后次序不能变更。所以，一些不考虑样品排列顺序的数学处理方法对此不适用。有序样品的聚类分析就是对有序样品进行分段的统计方法。对n个有序样品进行分割，可能有

$$C_{n-1}^{1}+C_{n-1}^{2}+\cdots+C_{n-1}^{n-1}=2^{n-1}-1$$

种划分方法，每一种分法称为一种分割。在所有这些分割中，存在这样一种分割，它使得各段内部样品之间差异性最小，而段与段之间的差异性最大。这种对n个样品分段并使组内离差平方和最小的分割方法，称为最优分割法。

最优分割的计算步骤如下：

1．数据正规化

设原始数据矩阵为

$$X=\begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix}。$$

将X中的元素x_{ij}变换为

$$z_{ij}=(x_{ij}-\min\{x_{ij}\})/(\max\{x_{ij}\}-\min\{x_{ij}\})。$$

2．计算段直径矩阵D

$$d_{(i,j)}=\sum_{\alpha=i}^{j}\sum_{\beta=1}^{p}\left[\overline{z_{\alpha\beta}}-\overline{z_{\beta}}(i,j)\right]^2, \qquad (2\text{-}32)$$

$$1\leq i\leq j\leq n$$

其中

$$\overline{z_{\beta}}(i,j)=\frac{1}{j-i+1}\sum_{\alpha=i}^{j}z_{\alpha\beta}。 \qquad (2\text{-}33)$$

3．计算全部分割的组内离差平方和(或段直径之和)及各种分段的最优分割

记$S_m(k;\ a_1(m),\ a_2(m),\ \cdots,\ a_k(m))$为对$m$个样品的最优$k$段分割，其分割点为第$a_1(m),\ a_2(m),\ \cdots,\ a_k(m)$的组内离差平方和。

(1) 最优二段分割

由D矩阵对每一个$m=n,\ n-1,\ \cdots,\ 2$，计算相应的组内离差平方和$S_m(2;\ a_1(m))$：

$$S_m(2;j)=d(1,j)+d(j+1,m) \qquad \begin{pmatrix} j=1,\ 2,\ \cdots,\ m-1 \\ m=n,\ n-1,\ \cdots,\ 2 \end{pmatrix}。$$

找出最小值，确定相应的最优二段分割点

$$S_m(2;a_1(m))=\min_{1\leq j\leq m-1}\{S_m(2;j)\}。$$

从而得到n个样品$(m-n)$的最优二段分割点的$a_1(n)$。因此，将n个样品分割为$\{x_1,\ \cdots,\ x_{a_1(n)}\}\{x_{a_1(n)+1},\ \cdots,\ x_n\}$。

(2) 最优三段分割

根据D矩阵及最优二段分割的结果，对每一个$m=n,\ n-1,\ \cdots,\ 3$计算相应的三段分割的组内离差平方和：

$$S_m(3;a_1(j),j)=S_j(2;a_1(j))+d(j+1,m) \qquad \begin{pmatrix} j=2,3,\cdots,m-1 \\ m=n,n-1,\cdots,3 \end{pmatrix},$$

然后求出最小值，并确定相应的最优三段分割

$$S_m(3;a_1(m),a_2(m))=\min_{2\leq j\leq m-1}\{S_m(3;a_1(j),j)\},$$

从而得到n个样品$(m=n)$的最优三段分割为

$$\{x_1,\ \cdots,\ x_{a_1(n)}\}\{x_{a_1(n)+1},\ \cdots,\ x_{a_2(n)}\}\{x_{a_2(n)+1},\ \cdots,\ x_n\}$$

其中$a_1(n)$和$a_2(n)$为最优三段分割点。

(3) 最优k段分割

根据矩阵D及最优$k-1$段分割计算结果，对于每一个$m=n,\ n-1,\ \cdots,\ k$分别计算相对的k段分割的组内离差平方和：

$$S_m(k;\ a_1(j),\ a_2(j),\ \cdots,\ a_{k-2},\ j)=S_j(k-1;\ a_1(j),\ a_2(j),\ \cdots,\ a_{k-2}(j))+d(j+1,\ n)\quad\begin{pmatrix}j=k-1,\ k,\ \cdots,\ m-1\\ m=n,\ n-1,\ \cdots,\ k\end{pmatrix},\tag{2-34}$$

找出最小值，并确定相应的最优分割点，即

$$S_m(k,\ a_1(m),\ a_2(m),\ \cdots,\ a_{k-1}(m))=\min\{S_m(k;\ a_1(j),\ a_2(j),\ \cdots,\ a_{k-2}(j),\ j)\}\ 。$$

从而得到n个样品$(m=n)$的最优k段分割为：$\{x_1,\ \cdots,\ x_{c_1(n)}\}\{x_{c_1(n)+1},\ \cdots,\ x_{c_2(n)}\}\cdots\{x_{c(k-2)(n)+1},\ \cdots,\ x_{c(k-1)(n)}\}\{x_{c(k-1)(n)+1},\ \cdots,\ x_n\}$，其中$a_1(n)$，$a_2(n)$，$\cdots$，$a_{k-1}(n)$为最优$k$段分割点

4．计算举例

现有6个有序样品一个变量的正规化数据为

$$\mathbf{z}=\{1.00,\ 0.8571,\ 0.5143,\ 0.20,\ 0.0286,\ 0\}$$

(1) 计算段直径矩阵D

$$\mathbf{D}=\begin{Bmatrix} d(1,1) & d(1,2) & d(1,3) & d(1,4) & d(1,5) & d(1,6)\\ & d(2,2) & d(2,3) & d(2,4) & d(2,5) & d(2,6)\\ & & d(3,3) & d(3,4) & d(3,5) & d(3,6)\\ & & & d(4,4) & d(4,5) & d(4,6)\\ & & & & d(5,5) & d(5,6)\\ & & & & & d(6,6)\end{Bmatrix}$$

$$=\begin{Bmatrix} 0 & 0.0102 & 0.1246 & 0.3861 & 0.6880 & 0.9133\\ & 0 & 0.0588 & 0.2161 & 0.4000 & 0.5280\\ & & 0 & 0.0494 & 0.1214 & 0.1673\\ & & & 0 & 0.0147 & 0.0234\\ & & & & 0 & 0.0004\\ & & & & & 0\end{Bmatrix}。$$

(2) 最优二段分割。由D对于$m=6,\ 5,\ 4,\ 3,\ 2$时计算

$$S_m(2;\ j)=d(1,\ j)+d(j+1,\ n)\ ,\quad (j=1,\ 2,\ \cdots,\ m-1)\ 。$$

当$m=6$时，则

$$S_6(2;\ 1)=d(1,\ 1)+d(2,\ 6)=0.5280\ ,$$
$$S_6(2;\ 2)=d(1,\ 2)+d(3,\ 6)=0.1776\ ,$$
$$S_6(2;\ 3)=d(1,\ 3)+d(4,\ 6)=0.1480\ ,$$
$$S_6(2;\ 4)=d(1,\ 4)+d(5,\ 6)=0.3865\ ,$$
$$S_6(2,\ 5)=d(1,\ 5)+d(6,\ 6)=0.6880\ ,$$

其中

$$S_6(2;\ a_1(6))=\min\{S_6(2;\ j)\}=S_6(2;\ 3)=0.1480\ ,$$
$$a_1(6)=3\ 。$$

当$m=5$时，则

$$S_5(2;\ 1)=d(1,\ 1)+d(2,\ 5)=0.4000\ ,$$
$$S_5(2;\ 2)=d(1,\ 2)+d(3,\ 5)=0.1316\ ,$$

$$S_5(2;\ 3)=d(1,\ 3)+d(4,\ 5)=0.1393\ ,$$
$$S_5(2;\ 4)=d(1,\ 4)+d(5,\ 5)=0.3861\ ,$$

其中

$$S_5(2;\ a_1(5))=\min\{S_5(2;\ j)\}=S_5(2;\ 2)=0.1316\ 。$$

当$m=4$时，则

$$S_4(2;\ 1)=d(1,\ 1)+d(2,\ 4)=0.2161\ ,$$
$$S_4(2;\ 2)=d(1,\ 2)+d(3,\ 4)=0.0596\ ,$$
$$S_4(2;\ 3)=d(1,\ 3)+d(4,\ 4)=0.1246\ ,$$

其中

$$S_4(2;\ a_1(4))=\min\{S_4(2;\ j)\}=S_4(2;\ 2)=0.0596\ 。$$

当$m=3$时，则

$$S_3(2;\ 1)=d(1;\ 1)+d(2,\ 3)=0.0588\ ,$$
$$S_3(2;\ 2)=d(1,\ 2)+d(3,\ 3)=0.0102\ ,$$

其中

$$S_3(2;\ a_1(3))=\min\{S_3(2;\ j)\}=S_3(2;\ 2)=0.0588\ 。$$

当$m=2$时，则

$$S_2(2;\ a_1(2))=S_2(2;\ 1)=d(1,\ 1)+d(2,\ 2)=0\ 。$$

请注意：这里对$m=5,\ 4\cdots,\ 2$时，计算的$S_m(2;\ a_1(m))$的值是为最优三段分割作准备，而对于6个样品最优二段分割的分割点为：

$$S_6(2;\ a_1(6))=\min\{S_6(2;\ j)\}=S_6(2;\ 3)\ ,$$

从而得到6个样品的最优二段分割为：

$$\{x_1,\ x_2,\ x_3\}\{x_4,\ x_5,\ x_6\}\ ,$$

其中，x_3为分割点。

(3) 最优三段分割

对于$m=6,\ 5,\ 4,\ 3$时计算

$$S_m(3;\ a_1(j),\ j)=S_1(2;\ a(j))+d(j+1,\ n)\qquad (j=2,\ 3,\ \cdots,\ m-1)\ 。$$

当$m=6$时，则

$$S_6(3;\ a_1(2),\ 2)=S_2(2;\ a_1(2))+d(3,\ 6)=0.1673\ ,$$
$$S_6(3;\ a_1(3),\ 3)=S_3(2;\ a_1(3))+d(4,\ 6)=0.0336\ ,$$
$$S_6(3;\ a_1(4),\ 4)=S_4(2;\ a_1(4))+d(5,\ 6)=0.0060\ ,$$
$$S_6(3;\ a_1(5),\ 5)=S_5(2;\ a_1(5))+d(6,\ 6)=0.1316\ ,$$

其中

$$S_6(3;\ a_1(6),\ a_2(6))=\min_{2\le j\le 5}\{S_5(3;\ a_1(j),\ j)\}$$
$$=S_6(3;\ a_1(3)),\ 3)=0.0336\ 。$$

对于6个样品最优三段分割的分割点为：

$$a_1(6)=a_1(3)=2;\ a_2(6)=3\ 。$$

最优三段分割为：$\{x_1,\ x_2\}\{x_3\}\{x_4,\ x_5,\ x_6\}$。

继续算出$m=5,\ 4,\ 3$时3段分割

当$m=5$时，则

$$S_5(3;\ a_1(2),\ 2)=S_2(2;\ a_1(2))+d(3,\ 5)=0.1214\ ,$$

$$S_5(3;\ a_1(3),\ 3)=S_3(2;\ a_1(3))+d(4,\ 5)=0.0249\ ,$$

$$S_5(3;\ a_1(4),\ 4)=S_4(2;\ a_1(4))+d(5,\ 5)=0.0596\ ,$$

其中

$$S_5(3;\ a_1(5),\ a_2(5))=\min\{S_5(3;\ a_1(j),\ j)\}=0.0249\ 。$$

当$m=4$时，则

$$S_4(3;\ a_1(2),\ 2)=S_2(2;\ a_1(2))+d(3,\ 4)=0.0494\ ,$$

$$S_4(3;\ a_1(3),\ 3)=S_3(2;\ a_1(3))+d(4,\ 4)=0.0102\ ,$$

其中

$$S_4(3;\ a_1(4),\ a_2(4))=\min\{S_4(3;\ a_1(j),\ j)\}=0.0102\ 。$$

当$m=3$时，则

$$S_3(3;\ a_1(2),\ 2)=S_2(2;\ a_1(2))+d(3,\ 3)=0\ ,$$

即

$$S_3(3;\ a_1(3),\ a_2(3))=0\ 。$$

(4) 最优四段分割

即对$m=6,\ 5,\ 4$时计算

当$m=6$时，则

$$S_6(4;\ a_1(5),\ a_2(5),\ 5)=S_5(3;\ a_1(5),\ a_2(5))+d(6,\ 6)=0.0249\ ,$$

$$S_6(4;\ a_1(4),\ a_2(4),\ 4)=S_4(3;\ a_1(4),\ a_2(4))+d(5,\ 6)=0.0106\ ,$$

$$S_6(4;\ a_1(3),\ a_2(3),\ 3)=S_3(3;\ a_1(3),\ a_2(3))+d(4,\ 6)=0.0234\ ,$$

其中

$$S_6(4;\ a_1(6),\ a_2(6),\ a_3(6))=\min\{S_6(4;\ a_1(j),\ a_2(j),\ j)\}$$

$$=S_6(4;\ a_1(4),\ a_2(4),\ 4)=0.0106\ 。$$

所以，最优四段分割的分割点为

$$a_1(6)=a_1(4)=2;\ a_2(6)=a_2(4)=3;\ a_3(5)=4\ 。$$

继续计算

当$m=5$时，则

$$S_5(4;\ a_1(4),\ a_2(4),\ 4)=S_4(3;\ a_1(4),\ a_2(4))+d(5,\ 5)=0.0102\ ,$$

$$S_5(4;\ a_1(3),\ a_2(3),\ 3)=S_4(3;\ a_1(3),\ a_2(3))+d(4,\ 5)=0.0147\ ,$$

其中

$$S_5(4;\ a_1(5),\ a_2(5),\ a_3(5))=S_5(4;\ a_1(4),\ a_2(4),\ 4)=0.0102\ 。$$

当$m=4$时，则

$$S_4(4;\ a_1(3),\ a_2(3),\ 3)=S_3(3;\ a_1(3),\ a_2(3))+d(4,\ 4)=0\ ,$$

即

$$S_4(4;\ a_1(4),\ a_2(4),\ a_3(4))=S_4(4;\ a_1(3),\ a_2(3),\ 3)=0\ 。$$

故得最优四段分割为

$$\{x_1,\ x_2\}\{x_3\}\{x_4\}\{x_5,\ x_6\}\ 。$$

(5) 最优五段分割

当$m=6$时，则

$$S_6(5;\ a_1(5),\ a_2(5),\ a_3(5),\ 5)=S_5(4;\ a_1(5),\ a_2(5),\ a_3(5))+d(6,\ 6)$$
$$=0.0102\ ,$$
$$S_5(5;\ a_1(4),\ a_2(4),\ a_3(4),\ 4)=S_4(4;\ a_1(4),\ a_2(4),\ a_3(4))+d(5,\ 6)$$
$$=0.0004\ ,$$

其中

$$S_5(5;\ a_1(6),\ a_2(6),\ a_3(6),\ a_4(6))=S_5(5;\ a_1(4),\ a_2(4),\ a_3(4),\ 4)$$
$$=0.0004\ 。$$

故其分割点为：$a_1(6)=a_1(4)=1$；$a_2(6)=a_2(4)=2$；$a_3(6)=a_3(4)=3$；$a_4(6)=4$。从而得到6个样品的5段最优分割为：$\{x_1\}\{x_2\}\{x_3\}\{x_4\}\{x_5,\ x_6\}$ 。

因为6个样品的6段分割只有一种分割法，所以不必进行计算。到此，我们计算了6个样品的最优2段、3段、4段、5段和6段分割。

5．应用实例

天津地矿所于1978年对燕山地区震旦亚界用最优分割法进行地层划分，取得良好效果。现选其中利用光谱分析微量元素作为变量对蓟县标准剖面进行最优分割，并将其划分的地层与地质划分地层进行对比（表2-13）。从对比结果看，最优分割与地质划分的结果基本一致，而最优分割进一步又提供了一些信息：

表2-12　蓟县标准剖面最优分割与地质划分对比表

地质划分意见		最优分割结果
寒武系 t	府君山组（E_{1f}）	
青白口系 Z_q	景儿峪组（Z_{qj}）	
	下马岭组（Z_{qx}）	
蓟县系 Z_j	铁岭组（Z_{jt}）	
	洪水庄组（Z_{jh}）	
	雾迷山组（Z_{jw}）	
	杨庄组（Z_{jy}）	
	高于庄组（Z_{jg}）	
长城系 Z_c	大红峪组（Z_{cd}）	
	团山子组（Z_{ct}）	
	串岭沟组（Z_{cch}）	
	常州村组（Z_{cc}）	

（据秦正永等，1978年）

（1）高于庄组与杨庄组是一个系的分界线，在最优分割中未出现分割点。原因是该标准剖面通过盆地中心。但在燕山沉降盆地边缘，如昌平十三陵，滦县青龙山，桃园等剖面上分割点十分清楚。这说明在盆地中心沉积连续，地层完整，岩性变化小，而盆地边缘间断明显，为划分二个系的界线提供依据。

（2）团山子组和大红峪组界线，地质上有三种意见：以厚层石英岩为界；以薄层石英岩为界；以薄层石英岩中的紫红色凝灰岩为界。电算结果趋向于以薄层石英岩为界。

（3）铁岭组地质上分为二段，电算结果认为可以划分为两个独立的组。无论是标准剖

或外围剖面均有相当于分组的分割点。根据这个信息，最近在蓟县剖面中已发现铁岭组面、二段之间沉积间断面，且迭层石在一、二段亦有明显的变化，为建组提供了信息。

（4）对巨厚、岩性单一的高于庄组、雾迷山组，原地质上分别划分为7段和9段，而电算结果将它们各自分为4段和5段。

第三章　多元回归分析

第一节　基本概念

如果相关分析是研究随机变量之间相互关系的一种统计分析方法，那么回归分析就是研究随机变量对其它变量（可以是随机变量，也可以是确定性变量）的依赖关系的一种统计分析方法。虽然回归分析与相关分析之间有微小的区别，人们常常把它们统称为回归分析或相关分析。

在回归分析中，把被研究的变量称为因变量，把它所依赖的变量称为自变量。多元回归分析就是有多个自变量的回归分析。

因变量y对p个自变量x_1，x_2，…，x_p的依赖关系可以用一个数学模型来表示：

$$y=f(x_1,\ x_2,\ \cdots,\ x_p)+\varepsilon\ .$$

上式表示变量y可分成两部分，一部分是由与y有相关关系的p个自变量的函数来刻划，另一部分是由随机误差ε来刻划。当数学模型表示为一次式时，即

$$y=\beta_0+\beta_1x_1+\beta_2x_2+\cdots+\beta_px_p+\varepsilon\ , \tag{3-1}$$

（其中β_0，β_1，β_2，…，β_p为待估参数）

称这个数学模型为p元线性回归模型。当自变量为确定性变量时，称上述数学模型为p元线性模型。p元线性回归模型和p元线性模型的数学研究方法是相同的，统称为线性回归分析。

回归分析主要解决以下几方面的问题：

1．建立回归方程

$$\hat{y}=b_0+b_1x_1+b_2x_2+\cdots+b_px_p\ ,$$

并估计回归效果，检验回归的显著性。

2．讨论回归方程中各自变量的作用，或者说检验每个自变量对回归的贡献大小。

3．最优回归方程的选择。

4．对因变量y的预测或控制。

第二节　多元线性回归方程的建立

一元线性回归问题在数理统计课程中已有介绍，这里直接介绍多元线性回归分析。

一、多元线性回归方程的建立

假设随机变量y依赖于p个自变量x_1，x_2，…，x_p而存在，且y服从正态分布。则所要

建立的线性回归方程为

$$\hat{y}=b_0+b_1x_1+b_2x_2+\cdots+b_px_p\ , \tag{3-2}$$

这里b_i（$i=1,\ 2,\ \cdots,\ p$）称为y对x_i的偏回归系数，b_0为常数项，$\hat{y}$称为自变量x_i对因变量y的回归估值。

所谓建立回归方程就是要确定偏回归系数b_i和常数项b_0的具体数值。

假如随机变量y与另外p个变量$x_1,\ x_2,\ \cdots,\ x_p$的内在联系是线性的，它的第i次试验数据是

$$y_i,\ x_{i1},\ x_{i2},\ \cdots,\ x_{ip}\qquad (i=1,\ 2,\ \cdots,\ n)$$

，那么这n次试验的数据可以假设有如下结构式：

$$\begin{cases} y_1=b_0+b_1x_{11}+b_2x_{12}+\cdots+b_px_{1p}+\varepsilon_1 \\ y_2=b_0+b_1x_{21}+b_2x_{22}+\cdots+b_px_{2p}+\varepsilon_2 \\ \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \\ y_n=b_0+b_1x_{n1}+b_2x_{n2}+\cdots+b_px_{np}+\varepsilon_n \end{cases}\ , \tag{3-3}$$

其中，$b_0,\ b_1,\ b_2,\ \cdots,\ b_p$是$p+1$个待估参数；$x_1,\ x_2,\ \cdots,\ x_p$和$y$是$n$次试验的数据；$\varepsilon_1,\ \varepsilon_2,\ \cdots,\ \varepsilon_n$是$n$个相互独立且服从同一正态分布$N(0,\ \sigma)$的随机变量。

根据矩阵运算，如果令

$$\mathbf{Y}=\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix},\qquad \mathbf{X}=\begin{pmatrix} 1 & x_{11} & x_{12}\cdots x_{1p} \\ 1 & x_{21} & x_{22}\cdots x_{2p} \\ \vdots & \vdots & \vdots\ \ \vdots \\ 1 & x_{n1} & x_{n2}\cdots x_{np} \end{pmatrix},$$

$$\mathbf{b}=\begin{pmatrix} b_0 \\ b_1 \\ \vdots \\ b_p \end{pmatrix},\qquad e=\begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix},$$

那么（3-3）式就可以写成$\mathbf{Y}=\mathbf{X}\mathbf{b}+\boldsymbol{\varepsilon}$的形式。

人们常常用最小二乘法来求得偏回归系数$b_0,\ b_1,\ b_2,\ \cdots,\ b_p$的具体数值。

对照（3-3）和（3-2）式，y_i是实际观测值，$\hat{y}_i$是回归估值，它们之间的差$y_i-\hat{y}_i=\varepsilon_i$（$i=1,\ 2,\ \cdots,\ n$）称为残差（或剩余），残差刻划了观测值与回归估值之间的偏离程度。

最小二乘法的原则是选择这样一组偏回归系数b_i，使得残差平方和Q达到最小，即使

$$Q=\sum_{i=1}^{n}(y_i-\hat{y}_i)^2 \tag{3-4}$$

为最小。将（3-2）式代入得

$$Q=\sum_{i=1}^{n}\left[y_i-\left(b_0+b_1x_{i1}+b_2x_{i2}+\cdots+b_px_{ip}\right)\right]^2\ 。 \tag{3-5}$$

上式是b_0，b_1，…，b_p的二次函数，它是非负的，故有极小值存在。根据微积分中的极值原理可以求出b_0，b_1。…b_p。

在（3-5）式中分别求Q对b_j（$j=0$，1，2，…，p）的偏导数，并令其为0，可得到由$p+1$个方程组成的正规方程组：

$$\begin{cases} \dfrac{\partial Q}{\partial b_0} = -2\sum\limits_{i=1}^{n}\left(y_i - b_0 - b_1x_{i1} - b_2x_{i2} - \cdots - b_px_{ip}\right) = 0 \\ \dfrac{\partial Q}{\partial b_1} = -2\sum\limits_{i=1}^{n}\left(y_i - b_0 - b_1x_{i1} - b_2x_{i2} - \cdots - b_px_{ip}\right)\cdot x_{i1} = 0 \\ \cdots\cdots\cdots\cdots\cdots\cdots \\ \dfrac{\partial Q}{\partial b_p} = -2\sum\limits_{i=1}^{n}\left(y_i - b_0 - b_1x_{i1} - b_2x_{i2} - \cdots - b_px_{ip}\right)\cdot x_{ip} = 0 \end{cases}$$

它可进一步化简为（以$\sum\limits_{i}$代$\sum\limits_{i=1}^{n}$）

$$\begin{cases} nb_0 + \left(\sum\limits_i x_{i1}\right)b_1 + \left(\sum\limits_i x_{i2}\right)b_2 + \cdots + \left(\sum\limits_i x_{ip}\right)b_p = \sum\limits_i y_i \\ \left(\sum\limits_i x_{i1}\right)b_0 + \left(\sum\limits_i x_{i1}^2\right)b_1 + \left(\sum\limits_i x_{i1}x_{i2}\right)b_2 + \cdots + \left(\sum\limits_i x_{i1}x_{ip}\right)b_p = \sum\limits_i x_{i1}y_i \\ \cdots\cdots\cdots\cdots\cdots\cdots \\ \left(\sum\limits_i x_{ip}\right)b_0 + \left(\sum\limits_i x_{ip}x_{i1}\right)b_1 + \left(\sum\limits_i x_{ip}x_{i2}\right)b_2 + \cdots + \left(\sum\limits_i x_{ip}^2\right)b_p = \sum\limits_i x_{ip}y_i \end{cases} \quad (3\text{-}6)$$

正规方程组可以写成如下矩阵形式：

$$\mathbf{X'Xb = X'Y} \quad (3\text{-}7)$$

在$\mathbf{X'X}$为满秩矩阵时，即可求得偏回归系数

$$\mathbf{b} = (\mathbf{X'X})^{-1}\mathbf{X'Y}$$

有了偏回归系数b_j（$j=0$，1，…，p），就可以直接写出回归方程

$$\hat{y} = b_0 + b_1x_1 + b_2x_2 + \cdots + b_px_p$$

至此，所需要的回归方程就建立起来了。但在实际计算时，为了方便，要尽量地减低正规方程组的阶数。从（3-6）方程组的第一式中得到

$$b_0 = \bar{y} - (b_1\bar{x}_1 + b_2\bar{x}_2 + \cdots + b_p\bar{x}_p)$$

其中

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i \qquad \bar{x}_j = \frac{1}{n}\sum_{i=1}^{n}\bar{x}_{ij}$$

这时可把（3-5）式改写为：

$$Q = \sum_{i=1}^{n}\left[\left(y_i - \bar{y}\right) - b_1\left(x_{i1} - \bar{x}_1\right) - b_2\left(x_{i2} - \bar{x}_2\right) - \cdots - b_p\left(x_{ip} - \bar{x}_p\right)\right]^2 \quad (3\text{-}8)$$

跟前面的做法一样，利用（3-8）式分别求Q对b_j（$j=1$，2，…，p）的偏导数，并令其为0，整理后可得：

$$\begin{cases} S_{11}b_1 + S_{12}b_2 + \cdots + S_{1p}b_p = S_{1y}, \\ S_{21}b_1 + S_{22}b_2 + \cdots + S_{2p}b_p = S_{2y}, \\ \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \\ S_{p1}b_1 + S_{p2}b_2 + \cdots + S_{pp}b_p = S_{py} \end{cases} \tag{3-9}$$

其中

$$S_{ik} = S_{ki} = \sum_{i=1}^{n}(x_{ij} - \overline{x}_j)(x_{ik} - \overline{x}_k)$$

$$= \sum_{i=1}^{n}(x_{ij}x_{jk}) - n\overline{x}_j\overline{x}_k , \tag{3-10}$$

$$S_{iy} = \sum_{i=1}^{n}(x_{ij} - \overline{x}_j)(y_i - \overline{y}) = \sum_{i=1}^{n}x_{ij}y_i - n\overline{x}_i\overline{y} \quad 。 \tag{3-11}$$

若记

$$\mathbf{C} = \begin{pmatrix} x_{11} - \overline{x}_1, & x_{12} - \overline{x}_2, & \cdots, & x_{1p} - \overline{x}_p \\ x_{21} - \overline{x}_1, & x_{22} - \overline{x}_2, & \cdots, & x_{2p} - \overline{x}_p \\ \vdots & \vdots & & \vdots \\ x_{n1} - \overline{x}_1, & x_{n2} - \overline{x}_2, & \cdots, & x_{np} - \overline{x}_p \end{pmatrix},$$

$$\mathbf{b} = \begin{pmatrix} b_1 \\ b_2 \\ \vdots \\ b_p \end{pmatrix}, \qquad \mathbf{y} = \begin{pmatrix} y_1 - \overline{y} \\ y_2 - \overline{y} \\ \vdots \\ y_n - \overline{y} \end{pmatrix},$$

则（3-9）式可写成矩阵形式

$$\mathbf{C'Cb} = \mathbf{C'y} \quad 。$$

求解上式，得

$$\mathbf{b} = (\mathbf{C'C})^{-1}\mathbf{C'y} \quad 。$$

从上式中解得的向量$\mathbf{b}$，b_0可以从下式中求得

$$b_0 = \overline{y} - b_1\overline{x}_1 - b_2\overline{x}_2 - \cdots - b_p\overline{x}_p \quad 。 \tag{3-12}$$

这是用各变量离差平方和-离差叉积和组成的正规方程组系数矩阵$\mathbf{C'C}$中求得偏回归系数，并建立回归方程。

二、回归方程的显著性检验

在实际问题中，事先并不能断定随机变量y与一般变量x_1，x_2，…，x_p之间是否存在线性关系。当线性回归方程建立以后，就要讨论该回归方程的显著性如何。

多元线性回归效果的好坏，可以从因变量y的总离差平方和S_{yy}被回归估值$\hat{y}$说明了多少来衡量。

y的总离差平方和S_{yy}可以分解为两部分

$$S_{yy}=\sum_{i=1}^{n}(y_i-\bar{y})^2=\sum_{i=1}^{n}[(y_i-\hat{y}_i)+(\hat{y}_i-\bar{y})]^2$$

$$=\sum_{i=1}^{n}(y_i-\hat{y}_i)^2+2\sum_{i=1}^{n}(y_i-\hat{y}_i)(\hat{y}_i-\bar{y})+\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2$$

$$=\sum_{i=1}^{n}(y_i-\hat{y}_i)^2+\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2\ ,$$

其中，交叉项乘积为零，事实上从正规方程组可得

$$\sum_{i=1}^{n}(y_i-\hat{y}_i)=0\ ,\quad \sum_{i=1}^{n}(\hat{y}_i-\bar{y}_{i1})\cdot x_{il}=0\ 。$$

而

$$\bar{y}=b_0+b_1\bar{x}_1+b_2\bar{x}_2+\cdots+b_p\bar{x}_p\ ,$$

$$\hat{y}_i-\bar{y}=b_1(x_{i1}-\bar{x}_1)+b_2(x_{i2}-\bar{x}_2)+\cdots+b_p(x_{ip}-\bar{x}_p)$$

$$=\sum_{j=1}^{p}b_j(x_{ij}-\bar{x}_j)\ 。$$

从而

$$\sum_{i=1}^{n}(y_i-\hat{y}_i)(\hat{y}_i-\bar{y})=\sum_{i=1}^{n}(y_i-\hat{y}_i)\cdot\sum_{j=1}^{p}b_j(x_{ij}-\bar{x}_j)$$

$$=\sum_{j=1}^{p}b_j\sum_{i=1}^{n}(y_i-\hat{y}_i)(x_{ij}-\bar{x}_j)$$

$$=\sum_{j=1}^{p}b_j[\sum_{i=1}^{n}(y_i-\hat{y}_i)x_{ij}-\bar{x}_j\Sigma(y_i-\hat{y}_i)]$$

$$=0\ 。$$

令

$$S_{yy}=U+Q \tag{3-13}$$

其中

$$U=\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2$$

$$=\sum_{i=1}^{n}[b_0+b_1x_{i1}+\cdots+b_px_{ip}-\bar{y}]^2$$

$$=\sum_{i=1}^{n}[b_1(x_{i1}-\bar{x})+b_2(x_{i2}-\bar{x}_2)+\cdots+b_p(x_{ip}-\bar{x}_p)]^2\ 。$$

它反映了自变量x_1，x_2，…，x_p的变化对y的贡献，称为回归平方和，其自由度为p。

$$Q=\sum_{i=1}^{n}(y_i-\hat{y}_i)^2=\sum_{i=1}^{n}[y_i-(b_0+b\ x_{i1}+\cdots+b_px_{ip})]^2\ ,$$

称为残差平方和，其自由度为$n-p-1$。

显然，U值越大，也就是说Q值越小，回归效果就越显著。

利用方差分析，可以对y与x_j（$j=1,\ 2,\cdots,\ p$）之间的线性关系是否显著进行检验。

假设

$$H_0:\ b_1=b_2=b_3=\cdots=b_p=0\ ,$$

即假设回归效果是不显著的，则统计量

$$F=\frac{S_{回}^{2}}{S_{残}^{2}}=\frac{U/p}{Q/(n-p-1)}\ ,$$

服从第一自由度$f_1=p$，第二自由度$f_2=n-p-1$的F分布。列出方差分析表如下：

表3-1　多元线性回归方差分析表

方差来源	平　方　和	自由度	均　方	F
回归	$U=\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2=\sum_{j=1}^{p}b_jS_{jy}$	p	$S_{回}^{2}=\frac{U}{p}$	$F=\frac{S_{回}^{2}}{S_{残}^{2}}$
残差	$Q=\sum_{i=1}^{n}(y_i-\hat{y}_i)^2=S_{yy}-U$	$n-p-1$	$S_{残}^{2}=\frac{Q}{n-p-1}$	
总和	$S_{yy}=\sum_{i=1}^{n}(y_i-\bar{y})^2$	$n-1$		

对于指定的显著性水平α，可由F分布表查$F_\alpha(p,\ n-p-1)$，当计算所得的F值大于查表值F_α时，拒绝假设H_0，认为回归效果显著。反之，当$F<F_\alpha$时，接受假设H_0，认为回归方程没有显著意义。

把回归平方和在因变量y的总离差平方和S_{yy}中所占的比例称为拟合度，以c表示

$$c=U/S_{yy}\ 。\qquad(3\text{-}14)$$

将其开平方取正值，称为因变量y与所有自变量x_j（$j=1,\ 2,\ \cdots,\ p$）之间的复相关系数，记为R

$$R=\sqrt{U/S_{yy}}=\sqrt{1-Q/S_{yy}}\ 。\qquad(3\text{-}15)$$

第三节　各自变量对回归效果的贡献

在多元回归中，我们并不满足于整个回归方程是否显著这个结论。因为回归方程显著并不意味着方程中每个自变量对因变量y的作用都是显著的。我们讨论了回归方程总的效果后还要讨论各个自变量各自贡献的大小，后一种讨论有时在地球化学工作中更有意义。

一、偏相关系数

在双变量的数理统计中，用相关系r_{ij}来表示x_i与x_j两个变量相关关系的密切程度。这种相关系数称为简单相关系数。但在多变量的情况下，变量间的相关关系较复杂，有时两个变量本来不相关，由于其它变量分别与它们有一定的相关关系，致使这两个变量的简单相关系数增大。或者两个变量本来是正相关，由于其它变量的影响，计算出负相关等等。因此，在多变量的情况下，简单相关系数就不能真实反映两变量间相关关系。为了得到两个变量间真正的相关系数，必须除去其它变量的影响，这种消除了其它变量影响后得出的两变量的相关系数称为偏相关系数，记为$r_{ij\cdot c}$（其中c为其余变量子集合）。

在多元线性回归分析中，计算因变量y与各自变量x_i间的偏相关系数，可以反映该自变量x_i对因变量y的回归贡献大小。

在最简单的情况下，设有l，h，d三个变量，其偏相关系数可从它们两两间的简单相关系数r_{ij}（i，$j=l$，h，d）推算得到

$$r_{lh\cdot d}=\frac{r_{lh}-r_{ld}\cdot r_{hd}}{[(1-r_{ld}^2)(1-r_{hd}^2)]^{1/2}}。\tag{3-16}$$

当变量超过3个时，可用下列递推公式：

$$r_{lh\cdot cd}=\frac{r_{lh\cdot c}-r_{ld\cdot c}-r_{hd\cdot c}}{[(1-r_{ld\cdot c}^2)(1-r_{hd\cdot c}^2)]^{1/2}}。\tag{3-17}$$

由于上述递推公式计算麻烦，实际工作中常用下列公式计算偏相关系数：

$$r_{ih\cdot c}=-\frac{d_{lh}}{(d_{ll}\cdot d_{hh})^{1/2}},\tag{3-18}$$

其中d为相关矩阵逆矩阵中的元素，$l\neq h$。

偏相关系数的显著性，可用t统计量来检验。假设　H_0：$r_{lh\cdot c}=0$，这时统计量

$$t=\frac{r_{lh\cdot c}\sqrt{n-k-2}}{\sqrt{1-r_{lh\cdot c}^2}}\tag{3-19}$$

服从自由度$f=n-k-2$的t分布，其中k为子集c中的变量个数。按照指定的显著性水平α，查$t_\alpha(n-k-2)$分布表，当计算值$t>t_\alpha$查表值时否定原假设认为偏相关显著；否则$t<t_\alpha$，接受原假设，认为偏相关不显著。

二、偏回归平方和及其检验

各自变量对因变量贡献的大小更常见的是用偏回归平方和来衡量。上一节中讲到，回归平方和是所有自变量对y变量总贡献的度量。回归方程中，自变量个数增加，回归平方和随之增大；减少自变量，回归平方和也会相应减小。取消（或增加）某一自变量后，回归平方和的减小（或增大）的数值，称为该自变量的偏回归平方和。偏回归平方和可以衡量每个变量在回归中所起作用的大小。以P_i记自变量i的偏回归平方和，其计算公式为：

$$P_i=b_i^2/c_{ii},\tag{3-20}$$

式中，b_i为变量i的偏回归系数，c_{ii}为正规方程组系数矩阵的逆矩阵主对角线上第i个元素。

对偏回归平方和也可以作显著性检验。

$$H_0:\ P_i=0\ ,$$

假设，统计量$F_i=\dfrac{P_i}{S^2_{残}}=\dfrac{b_i^2}{c_{ii}S^2_{残}}$ (3-21)

服从第一自由度$f_1=1$，第二自由度$f_2=n-p-1$的F分布。当在某显著性水平α下，若F_i的计算值大于$F_\alpha(1,\ n-p-1)$查表值时，则拒绝原假设，认为该自变量的回归效果是显著的。当$F_i<F_\alpha$时，接受假设，认为该自变量的回归效果不显著。若经检验后，有几个自变量回归效果不显著，有时不能断言它们都不是影响因变量y的重要变量。但可以肯定，偏回归平方和最小的那个变量对y的贡献是不显著的。此时可以将这个变量从回归方程中剔除。剔除一个变量后重新建立回归方程，偏回归平方和也要重新检验，但一般新的偏回归系数b_j^*与原来的回归系数b_i之间有下列关系：

$$b_j^*=b_j-\frac{c_{ij}}{c_{ii}}b_i\qquad j\neq 1\ , \tag{3-22}$$

式中b_i为剔除变量的偏回归系数，c为正规方程组系数矩阵逆矩阵中的相应元素。

第四节　回归分析的一般步骤

这里通过一个实例来说明回归分析的一般步骤。

为利用元素分带性研究原生晕与矿体空间位置之间的关系，选择因变量y为矿体部位。当位于矿体尾部时，y取值为0；当位于矿体头部时取值为1，$0\leqslant y\leqslant 1$。在矿体的5个不同部位测得Sn、Pb、Zn的含量，所得数据经对数变换后列于表3-2中。

表3-2　原始数据对数值表

编　号	x_1(Sn)	x_2(Pb)	x_3(Zn)	y
1	1.7052	2.7733	2.8090	0.10
2	2.3976	1.7753	3.9253	0.41
3	2.3960	2.2503	3.9445	0.53
4	2.1084	1.3928	1.7474	0.84
5	1.5307	1.5618	2.5787	1.00
均　值	2.0275	1.9508	2.9992	0.5760

1．按公式$S_{ij}=\sum_{k=1}^{n}(x_{ki}-\bar{x}_i)(x_{kj}-\bar{x}_j)=\sum_{k=1}^{n}(x_{ki}x_{kj})-\frac{1}{n}\sum_{k=1}^{n}x_{ki}\sum_{k=1}^{n}x_{kj}$，求得正规方程组系数矩阵，见表3-3。

表3-3　正规方程组系数矩阵

S_{ij}	x_1	x_2	x_3	y
x_1	0.6305	-0.0712	0.8635	-0.1143
x_2	-0.0712	1.2601	0.8193	-0.6884
x_3	0.8635	0.8193	3.5348	-0.6111
y	-0.1143	-0.6884	-0.6111	0.5057

得正规方程组为

$$0.6305b_1 - 0.0712b_2 + 0.8635b_3 = -0.1143\ ,$$
$$-0.0712b_1 + 1.2601b_2 + 0.8193b_3 = -0.6881\ ,$$
$$0.8635b_1 + 0.8193b_2 + 3.5348b_3 = -0.6111\ 。$$

2．用求解求逆并行方案解上述正规方程组，具体步骤见表3-4。求得：

表3-4　正规方程组求解求逆计算表

方程	计算步骤	正规方程组				逆矩阵			检验
		x_1	x_2	x_3	y	c_1	c_2	c_3	Σ
①		0.6305	−0.0712	0.8635	−0.1143	1	0	0	2.3085
②		−0.0712	1.2601	0.8193	−0.6883	0	1	0	2.3198
③		0.8635	−0.8193	3.5348	−0.6111	0	0	1	5.0065
④	①+0.6305	1	−0.1129	1.3696	−0.1813	1.5860	0	0	3.6614
⑤	④×0.0712+②	0	1.2521	0.9168	−0.7013	0.1129	1	0	2.5805
⑥	④×(−0.8635)+③	0	0.9168	2.3521	−0.4545	−1.3695	0	1	2.4449
⑦	⑤×0.1129+④	1	0	1.4552	−0.2445	1.5962	0.0901	0	3.8941
⑧	⑤+1.2521	0	1	0.7322	−0.5601	0.0902	0.7986	0	2.0609
⑨	⑧×(−0.9168)+⑥	0	0	1.6808	−0.0590	−1.4532	−0.7321	1	0.5555
⑩	⑫×(−1.4552)+⑦	1	0	0	−0.2955	2.8509	0.7227	−0.8641	3.4140
⑪	⑫×(−0.7322)+⑧	0	1	0	−0.5858	0.7228	1.1176	−0.4357	1.8189
⑫	⑨+1.6808	0	0	1	0.0351	−0.8640	−0.4356	0.5950	0.3305

$b_1 = -0.2955$，$b_2 = -0.5858$，$b_3 = 0.0351$，

$$\begin{aligned} b_0 &= \bar{y} - [(-0.2955)\times\bar{x}_1 + (-0.5858)\times\bar{x}_2 + 0.0351\times\bar{x}_3] \\ &= 0.5760 - [(-0.2955)\times 2.0276 - 0.5858\times 1.9508 + 0.0351\times 2.9992 \\ &= 2.2129\ 。 \end{aligned}$$

写出回归方程

$$\hat{y} = 2.2127 - 0.2955x_1 - 0.5858x_2 + 0.0351x_3\ 。$$

且有正规方程组系数矩阵的逆矩阵为

$$\mathbf{C} = \begin{bmatrix} 2.8509 & 0.7227 & -0.8641 \\ 0.7228 & 1.1176 & -0.4357 \\ -0.8640 & -0.4356 & 0.5950 \end{bmatrix}\ 。$$

3．回归方程的显著性检验，见表3-5。

表3-5　回归方程的方差分析表

来源	平方和	自由度	均方	F
回归	$U = \sum_i b_i S_{iy} = 0.4157$	$p = 3$	$S^2_{回} = 0.1386$	
残差	$Q = S_{yy} - U = 0.0900$	$n - p - 1 = 1$	$S^2_{残} = 0.0900$	$F = 1.54$
总和	$S_{yy} = 0.5057$	$n - 1 = 4$		

检验结果不论α取0.01或0.05，F都小于F_α表明回归效果很差。这可能是以下原因引起：首先，某些自变量与因变量之相关性差，在回归中不起作用。其次，与原始的观测数

据太少等原因有关。对于第一种原因可以把回归作用不显著的变量剔除出回归方程。

4．求各自变量的偏回归平方和及其显著性检验：

$$p_1=\frac{b_1^2}{c_{11}}=\frac{(-0.2955)^2}{2.8509}=0.0306\ ,\qquad F_1=\frac{p_1}{S_{残}^2}=\frac{0.0306}{0.0900}=0.34\ ,$$

$$p_2=\frac{b_2^2}{c_{22}}=\frac{(-0.5858)^2}{1.1176}=0.3071\ ,\qquad F_2=\frac{p_2}{S_{残}^2}=\frac{0.3071}{0.0900}=3.41\ ,$$

$$p_3=\frac{b_3^2}{c_{33}}=\frac{0.0351^2}{0.5950}=0.0020\ ,\qquad F_3=\frac{p_3}{S_{残}^2}=\frac{0.0020}{0.0900}=0.022\ 。$$

根据检验结果，3个自变量的作用都不显著。因此将最不显著的x_3从回归方程中剔除。

5．从回归方程中剔除一个自变量x_3重新计算回归系数。

$$b_1^{*}=b_1-\frac{c_{31}}{c_{33}}b_3=-0.2955-\frac{-0.8640}{0.5950}\times 0.0351=-0.2445\ ,$$

$$b_2^{*}=b_2-\frac{c_{32}}{c_{33}}b_3=-0.5858-\frac{-0.4350}{0.5950}\times 0.0351=-0.5601\ ,$$

$$b_0^{*}=\bar{y}-(b_1^{*}\bar{x}_1+b_2^{*}\bar{x}_2)=0.5760+(0.2445\times 2.0275+0.5601\times 1.9508)$$
$$=2.1644\ 。$$

从而得回归方程：

$$\hat{y}=2.1644-0.2445x_1-0.5601x_2\ 。$$

重复前面方差分析步骤，得：

$$U=0.4153\qquad Q=0.0904\qquad S_{残}^2=0.0282$$

$$F=\frac{U/2}{Q/2}=4.6$$

经F检验结果，回归效果仍不显著，应进一步剔除不显著的变量。为此，计算其偏回归平方和，先求这两个变量的正规方程组系数的逆矩阵，然后计算偏回归平方和及F_i值，得$F_1=1.33, F_2=13.95$，重复上述步骤作剔除x_1的计算，建立与x_2的一元线性回归方程。

$$\hat{y}=1.6417-0.5463x_2$$

经方差分析得：$U=0.3761$，$Q=0.1296$，$F=8.7$。按自由度$f_1=1$，$f_2=3$，查F-分布表得$F_{0.05}=10.13$，$F_{0.1}=5.54$。可见在$\alpha=0.1$的水平上回归显著，$\hat{y}=1.6417-0.5463x_2$为最终的回归方程。

第五节　最优回归方程的选择

在有p个自变量的情况下，可能建立回归方程的自变量的不同组合数为(2^p-1)个。即假如有10个自变量，那么就可以建立1023个不同的回归方程。当然这些回归方程的效果有好有坏，而人们所希望的是回归效果最好的，也就是称为“最优”的回归方程。所谓最优回归方程包括以下含义：首先是其回归效果最佳；其次是自变量的个数最少。换句话说，一方面对因变量起显著作用的自变量都选进回归方程，另一方面对因变量作用不显著的自变量都剔出回归方程，选择一个最佳的变量组合。

选择最优回归方程有4种不同的方法：

1．从所有可能的变量组合中，选择一个最优的回归方程。这种方法当然是万无一失，一定能在所有的p个自变量中选出一个最优组合。但是其工作量特别大。当自变量数比较多时，一般的小型电子计算机可能也难以胜任。

2．从包含全部p个自变量组合的回归方程中逐一剔除对因变量作用不显著的自变量，直到保留在回归方程中自变量的作用都显著为止。

3．从自变量中一个一个挑选对因变量作用显著的自变量，直到没有被选入的变量对因变量的作用都不显著为止。

4．逐步回归分析方法。逐步回归按照自变量对因变量所起作用的显著程度，从大到小逐个地引入回归方程。当每一变量引入以后，若先前已经引入的变量由于后来变量的引入而使其作用变得不显著时，就及时从回归方程中剔除出去，直到作用显著的变量都引入到回归方程，而作用不显著的变量都剔出回归方程，得到一个最佳的变量组合为止。

以上第一种方法无疑能选择到最优的回归方程，而且在地球化学工作中能从不同的回归方程来分析与因变量有关的地球化学变量的不同作用。但是，这种方法计算工作量太大。第四种方法则综合了其它方法的优点，也克服了其它方法的缺点，既能选择最优回归方程，计算工作量又不太大。下面介绍逐步回归分析。

一、建立标准正规方程组

为了分辨p个自变量对因变量y所起影响或作用的大小，一个自然的想法是比较各自变量回归系数b_j（$j=1, 2, \cdots, p$）的绝对值的大小。因为根据回归系数的含义，x_j的回归系数b_j是在其余$p-1$个自变量保持不变的条件下，x_j改变一个单位所引起y平均变化的大小。因而它的绝对值的大小反映了它所代表的因素的重量程度。但是回归系数和自变量所取的单位（或数量级）有关，而各个自变量取不同的量纲的情况是常见的，因而不能将回归系数直接进行比较。为了消除这个影响，对自变量和因变量都要加以标准化。经过标准化的变量，其均值为0，标准离差为1。由标准化数据建立起来的正规方程组系数矩阵即为变量间的相关系数矩阵，称为标准化正规方程组。由它所解的b'_j称为标准回归系数，其常数项$b'_0=0$，因变量的总离差平方和$S_{yy}=1$。列出标准化正规方程组为：

$$\begin{cases} r_{11}b'_1 + r_{12}b'_2 + \cdots + r_{1p}b'_p = r_{1y} \\ r_{21}b'_1 + r_{22}b'_2 + \cdots + r_{2p}b'_p = r_{2y} \\ \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \\ r_{p1}b'_1 + r_{p2}b'_2 + \cdots + r_{pp}b'_p = r_{py} \end{cases} \quad 。 \tag{3-23}$$

二、变量的引入、剔除与消去法的关系

求解标准化正规方程组，建立标准回归方程，还需要解决以下两个问题：①引入变量和剔除变量的标准；②引入变量与剔除变量的方法。

假定已有l个自变量引入到回归方程，即

$$\hat{y} = b_0 + b_1x_1 + b_2x_2 + \cdots + b_lx_l \quad 。$$

相应的平方和分解公式是

$$S_{yy}=U+Q\ 。$$

为了表明U和Q与引入的自变量是有关的，分别用符号$U(x_1,\ \cdots,\ x_l)$和$Q(x_1,\ \cdots,\ x_l)$表示。当增加一个自变量x_i（$i=l+1,\ \cdots,\ p$）后，有了新的回归方程，相应的平方和分解公式是

$$S_{yy}=U(x_1,\cdots,x_l,x_i)+Q(x_1,\cdots,x_l,x_i)\ 。$$

原来的是

$$S_{yy}=U(x_1,\cdots,x_l)+Q(x_1,\cdots,x_l)\ 。$$

注意到上两式左端S_{yy}是一样的，当x_i引入后，回归平方和从$U(x_1,\cdots,x_l)$增加到$U(x_1,\ \cdots,\ x_l,\ x_i)$，而残差平方和从$Q(x_1,\ \cdots,\ x_l)$降到$Q(x_1,\ \cdots,\ x_l,\ x_i)$，并有

$$U(x_1,\cdots,x_l,x_i)-U(x_1,\cdots,x_l)=Q(x_1,\cdots,x_l)-Q(x_1,\cdots,x_l,x_i)\ 。$$

记

$$p_i=U(x_1,\cdots,x_l,x_i)-U(x_1,\cdots,x_l)\ ,$$

p_i就是回归方程中引入x_i后对回归平方和的贡献，即上一节中讨论的偏回归平方和，因此有

$$p_i=\frac{b_i^2}{c_{ii}}\ 。$$

并且统计量

$$F_3=\frac{p_i/1}{Q/(n-l-1)}=\frac{p_i}{S_{残}^2} \tag{3-24}$$

服从第一自由度$f_1=1$，第二自由度$f_2=n-l-1$的F-分布。经F检验，当x_i作用显著时，可将其引入。

同理，如果x_i原来已经在回归方程中，若检验后其作用不显著，可及时从回归方程中剔除出去。因此，取剔除和引入变量x_i的标准相同$F_4=F_3=p_i/S_{残}^2$。

有了引入和剔除变量的标准，就可以进行具体计算，求得回归系数。为了得到引入变量与剔除变量的具体方法，先回顾一下用消去法求解正规方程组的过程。

当消去正规方程组系数矩阵的第一列时，常数项列的第一个数就是只有x_1这一个自变量情况下所建立的回归方程的回归系数b_1。

第二次消去了正规方程组系数矩阵的第一、二两列时，常数项列中的第一、二两个数即为只有x_1，x_2两个自变量情况下所建立回归方程的回归系数b_1和b_2。依次类推，得到以后各个自变量的回归系数。

由此可得出结论：系数矩阵中每消去一列，等价于回归方程中引入一个新的变量，而且与变量排列的顺序无关。

在线性代数中为了节省空间（或计算机内存），推导出求解求逆紧凑格式。这种紧凑格式在逐步回归中还另有好处，如果在紧凑格式中对同一列进行偶数次消去运算，那么等同于没有消去该列。也就是说，对该矩阵中的某一列进行奇数次消去变换，则该变量被引入回归方程；若对该列进行了偶数次消去变法，则说明该变量被剔除出回归方程。引入一个变量与剔除一个变量的变换公式是相同的。在第s次对第k列消去的变换公式是：

$$r_{ij}^{(s)}=\begin{cases}1/r_{kk}^{(s-1)} & (i=k,\ j=k)\\ r_{kj}^{(s-1)}/r_{kk}^{(s-1)} & (i=k,\ j\neq k)\\ -r_{ik}^{(s-1)}/r_{kk}^{(s-1)} & (i\neq k,\ j=k)\\ r_{ij}^{(s-1)}-r_{ik}^{(s-1)}\cdot r_{kj}^{(s-1)}/r_{kk}^{(s-1)} & (i\neq k,\ j\neq k)\end{cases}。\quad (3\text{-}25)$$

进行上述变换后，回归分析中的剩余平方和 Q 的值即为系数矩阵中 r_{yy} 位置所得的结果。即有，$Q^{(0)}=r_{yy}^{(0)}=1$，$Q^{(s)}=r_{yy}^{(s)}$，$U^{(s)}=1-Q^{(s)}$，而且由相关矩阵构成的系数矩阵中，第 i 个变量的偏回归平方和 $p_i^{(s)}$ 为：

$$p_i^{(s)}=\left(r_{iy}^{(s-1)}\right)^2/r_{ii}^{(s-1)}。\quad (3\text{-}26)$$

所以 $p_i^{(s)}$ 可以作为下一步引进变量的指标。每一步引入可挑选不在回归方程变量中的最大者。

$$F_i^{(s)}=\frac{p_i^{(s)}/1}{Q^{(s)}/(n-l-1)},\quad (3\text{-}27)$$

式中，l 为已经引入到回归方程中的变量个数，F_i 服从 $f_1=1$，$f_2=n-l-1$ 的 F-分布。如果，经检验，已引进的变量中有不显著的，则选其最不显著者作剔除变换，然后再检验。再在未引入的变量中检验有无回归显著的变量，若有则挑选最显著的作引入的消去变换，然后再检验。反复进行，直到没有变量可以引进，也没有变量可以从方程中剔除为止。

最后，由于相关系数矩阵得到的是标准回归系数 b_i^*，应该把它化为一般回归系数 b_i。其关系为：

$$b_i=b_i^*\sqrt{S_{yy}/S_{ii}},\quad (3\text{-}28)$$

其中 S_{ii} 和 S_{yy} 为方差协方差矩阵中对应元素，即 i 变量和 y 因变量的方差。

$$b^{\circ}=\bar{y}-(b_1\bar{x}_1-b_2\bar{x}_2-\cdots-b_l\bar{x}_l)。$$

三、实例*

某种水泥在凝固时放出热量（cal/g）与水泥的4种化学成分有关。用逐步回归分析研究放热与水泥成分之间的关系。

设水泥凝固时放热为因变量 y，水泥的4种化学成分为自变量，x_1($3CaO\cdot Al_2O_3$)，x_2($3CaO\cdot SiO_2$)，x_3($4CaO\cdot Al_2O_3\cdot Fe_2O_3$)，$x_4$($2CaO\cdot SiO_2$)。原始数据见3-6表。

表3-6 原始数据表

样 号	x_1	x_2	x_3	x_4	y
1	7	26	6	60	78.5
2	1	29	15	52	74.3
3	11	56	8	20	104.3
4	11	31	8	47	87.6
5	7	52	6	33	95.9
6	11	55	9	22	109.2
7	3	71	17	6	102.7
8	1	31	22	44	72.5

* 选自中国科学院数学所数理统计组编的“回归分析方法”

续表3-6

样　号	x_1	x_2	x_3	x_4	y
9	2	54	18	22	93.1
10	21	47	4	26	115.9
11	1	40	23	34	83.8
12	11	66	9	12	113.3
13	10	68	8	12	109.4
均　值	7.46	48.15	11.77	30.00	95.42

先求方差-协方差矩阵（不除n），见表3-7。

$$S_{ij}=\sum_{k}(x_{ki}-\bar{x}_i)(x_{kj}-\bar{x}_j)。$$

表3-7　方差-协方差矩阵

S_{ij}	x_1	x_2	x_3	x_4	y
x_1	415.23	251.08	−372.62	−290.00	775.96
x_2		2905.69	−160.54	−3041.00	2292.95
x_3			492.31	38.00	−616.23
x_4				5362.00	−2481.70
y					2715.76

计算相关系数增广矩阵 $[r_{ij}^{(0)}]$

$$\mathbf{R}^{(0)}=\begin{pmatrix} 1 & 0.2286 & -0.8241 & -0.2454 & 0.7307 \\ 0.2286 & 1 & -0.1392 & -0.9730 & 0.8163 \\ -0.8241 & -0.1392 & 1 & 0.0295 & -0.5347 \\ -0.2454 & -0.9730 & 0.0295 & 1 & -0.8213 \\ 0.7307 & 0.8163 & -0.5347 & -0.8213 & 1 \end{pmatrix}$$

在进行逐步回归之前，先确定检验自变量显著性的F临界值。此例中，$n=13$，在全部4个变量中估计有2-3个变量可能被选入回归方程，因此第一自由度为1，第二自由度为13−2−1=10，取置信度$\alpha=0.10$，查F-分布表得$F=3.28$。下面进行逐步回归。

第Ⅰ步，考虑引入第一个变量

1．根据偏回归平方和的计算公式

$$p_i^{(s)}=\left(r_{iy}^{(s-1)}\right)^2/r_{ii}^{(s-1)},$$

求得：

$$p_1^{(1)}=\left(r_{1y}^{(0)}\right)^2/r_{11}^{(0)}=(0.7307)^2/1=0.5339\ ,$$

$$p_2^{(1)}=\left(r_{2y}^{(0)}\right)^2/r_{22}^{(0)}=(0.8163)^2/1=0.6663\ ,$$

$$p_3^{(1)}=\left(r_{3y}^{(0)}\right)^2/r_{33}^{(0)}=(-0.5347)^2/1=0.2859\ ,$$

$$p_4^{(1)}=\left(r_{4y}^{(0)}\right)^2/r_{44}^{(0)}=(-0.8213)^2/1=0.6745\ 。$$

选择最大者 $p_4^{(1)}=0.6745$作F-检验：

$$F_4^{(1)}=\frac{p_4^{(1)}/1}{Q^{(1)}/(n-1-1)}=\frac{(n-2)\cdot p_4^{(1)}}{1-p_4^{(1)}}=\frac{11\times 0.6745}{0.3255}=22.80\ 。$$

因为 $F_4^{(1)}=22.8>F_{(0.10)}=3.28$，所以$x_4$显著，应被引入。

2．作引入x_4的计算，利用公式（3-25）对$\mathbf{R}^{(0)}$的第4列进行消去变换，当$k=4$时，计算结果得到$\mathbf{R}^{(1)}$：

$$\mathbf{R}^{(1)}=\begin{bmatrix} 0.9398 & -0.0102 & -0.8169 & 0.2454 & 0.5291 \\ -0.0102 & 0.0534 & -0.1105 & 0.9730 & 0.0172 \\ -0.8169 & -0.1105 & 0.9991 & -0.0295 & -0.5104 \\ -0.2454 & -0.9730 & 0.0295 & 1 & -0.8213 \\ 0.5291 & 0.0172 & -0.5104 & 0.8213 & 0.3255 \end{bmatrix}。$$

3．引入x_4后的主要结果

方程中的自变量x_4，其标准回归系数 $b_4^{*(1)}=r_{4y}^{(1)}=-0.8213$ 。

剩余平方和

$$Q^{(1)}=r_{yy}^{(1)}=0.3255\ 。$$

一般回归系数为：

$$b_4^{(1)}=b_4^{*(1)}\sqrt{S_{yy}/S_{44}}=-0.8213\sqrt{2715.76/3362.00}=-0.7382\ 。$$

常数项为：

$$b_0^{(1)}=\bar{y}-b_4^{(1)}\bar{x}_4=95.42-(-0.7380)\times 30=117.57\ 。$$

因此，第一个回归方程为：

$$\hat{y}=117.57-0.7382x_4\ 。$$

因x_4是刚引入的第一个自变量，省去剔除变量的检验。

第Ⅰ步，考虑引入第二个变量

1．由$\mathbf{R}^{(1)}$计算未引入变量的偏回归平方和

$$p_1^{(2)}=\left(r_{1y}^{(1)}\right)^2/r_{11}^{(1)}=(0.5291)^2/0.9398=0.2980\ ,$$

$$p_2^{(2)}=\left(r_{2y}^{(1)}\right)^2/r_{22}^{(1)}=(0.0172)^2/0.0534=0.0055\ ,$$

$$p_3^{(2)}=\left(r_{3y}^{(1)}\right)^2/r_{33}^{(1)}=(-0.5104)^2/0.9991=0.2601\ 。$$

选其中最大者作F-检验。

$$F_1^{(2)}=\frac{p_1^{(2)}/1}{\left(1-p_4^{(1)}-p_1^{(2)}\right)/(n-2-1)}=\frac{(n-3)\cdot p_1^{(2)}}{r_{yy}^{(1)}-p_1^{(2)}}$$

$$=\frac{10\times 0.2980}{0.3255-0.2980}=108.22\ 。$$

因为 $F_1^{(2)}=108.22>3.28$，因此，x_1显著，应将x_1引入回归方程。

2．对$\mathbf{R}^{(1)}$的第1列进行消去变换，$k=1$，计算结果为$\mathbf{R}^{(2)}$：

$$\mathbf{R}^{(2)}=\begin{bmatrix} 1.0641 & -0.0109 & -0.8693 & 0.2612 & 0.5631 \\ 0.0109 & 0.0532 & -0.1194 & 0.9756 & 0.0229 \\ 0.8693 & -0.1194 & 0.2891 & 0.1838 & -0.0505 \\ -0.2612 & -0.9756 & -0.1838 & 1.0641 & -0.6831 \\ -0.5631 & 0.0229 & -0.0505 & 0.6831 & 0.0275 \end{bmatrix} 。$$

3．求引入x_4，x_1后的回归方程，并作显著性检验。

标准回归系数为$b_1^{*(2)}=0.5631$，$b_4^{*(2)}=-0.6831$，代入公式（3-28）计算得一般回归系数：

$$b_1^{(2)}=1.4401,\quad b_4^{(2)}=-0.6140,\quad b_0^{(2)}=103.10 。$$

因此得第二个回归方程为

$$\hat{y}=103.10+1.4401x_1-0.6140x_4 。$$

其剩余平方和为

$$Q^{(2)}=r_{yy}^{(2)}=0.0275 。$$

第Ⅲ步，考虑剔除或引入新变量

1．现在方程中有x_1，x_4两个自变量，分别作显著性检验。

$$p_1^{(2)}=\frac{\left(r_{1y}^{(2)}\right)^2}{r_{11}^{(2)}},\qquad F_1^{(2)}=\frac{(n-2-1)\cdot p_1^{(2)}}{r_{yy}^{(2)}}=108.22 ,$$

$$p_4^{(2)}=\frac{\left(r_{4y}^{(2)}\right)^2}{r_{44}^{(2)}},\qquad F_4^{(2)}=\frac{(n-2-1)\cdot p_4^{(2)}}{r_{yy}^{(2)}}=159.30 。$$

计算结果都大于临界值3.28。因此，没有自变量可以剔除。

2．考虑引入新变量

计算不在方程中的变量x_2，x_3的偏回归平方和。

$$p_2^{(2)}=\left(r_{2y}^{(2)}\right)^2/r_{22}^{(2)}=(0.0229)^2/0.0532=0.0099 ,$$

$$p_3^{(2)}=\left(r_{3y}^{(2)}\right)^2/r_{33}^{(2)}=(-0.0505)^2/0.2891=0.0088 。$$

选其最大者$p_2^{(2)}$作F-检验：

$$F_2^{(2)}=\frac{(n-3-1)\cdot p_2^{(2)}}{\left(1-p^{(1)}-p_1^{(2)}\right)-p_2^{(2)}}=\frac{(n-4)\cdot p_2^{(2)}}{r_{yy}^{(2)}-p_2^{(2)}}$$

$$=\frac{9\times 0.0099}{0.0275-0.0099}=5.03 。$$

由于$F_2^{(2)}>F_{(0.10)}$，故x_2应引入回归方程。

3．对$\mathbf{R}^{(2)}$的第2列作消去变换，得$\mathbf{R}^{(3)}$：

$$\mathbf{R}^{(3)}=\begin{bmatrix} 1.0663 & 0.2044 & -0.8937 & 0.4606 & 0.5677 \\ 0.2044 & 18.7804 & -2.2423 & 18.3226 & 0.4304 \\ 0.8937 & 2.2423 & 0.0213 & 2.3714 & 0.0009 \\ 0.4606 & 18.3226 & -2.3714 & 18.9401 & -0.2632 \\ -0.5677 & -0.4304 & 0.0009 & 0.2632 & 0.0177 \end{bmatrix} 。$$

4．求引入x_4，x_1，x_2后的回归方程

标准回归系数为

$$b_1^{*(3)}=0.5677\ ,\quad b_2^{*(3)}=0.4304\ ,\quad b_4^{*(3)}=-0.2632\ ,$$

代入公式（3-28）计算得一般回归系数为：

$$b_1^{(3)}=1.4518\ ,\qquad b_2^{(3)}=0.4161\ ,$$

$$b_4=-0.2366\ ,\qquad b_0=71.65\ 。$$

因此，第三个回归方程为：

$$\hat{y}=71.65+1.4518x_1+0.4161x_2-0.2366x_4\ 。$$

其剩余回归平方和

$$Q^{(3)}=0.0177\ 。$$

第Ⅳ步，考虑第三个方程中因为引入x_2后，x_4，x_1中有没有被剔除的可能。

1．计算x_1，x_4在第三个回归方程中的偏回归平方和及相应的F值。

$$p_1^{(3)}=\left(r_{1y}^{(3)}\right)^2/r_{11}^{(3)}=(0.5677)^2/1.0663=0.3023\ ,$$

$$p_4^{(3)}=\left(r_{4y}^{(3)}\right)^2/r_{44}^{(3)}=(-0.2632)^2/18.9401=0.0037\ ,$$

$$F_1^{(3)}=\frac{(n-3-1)\cdot p_1^{(3)}}{r_{yy}^{(3)}}=\frac{9\times 0.3022}{0.0177}=154.01\ ,$$

$$F_4^{(3)}=\frac{(n-3-1)\cdot p_4^{(3)}}{r_{yy}^{(3)}}=\frac{9\times 0.0037}{0.0177}=1.87\ 。$$

因为$F_1^{(3)}=154.01>F_{(0.10)}=3.28$，回归显著，而$F_4^{(3)}=1.87<F_{(0.10)}=3.28$，$x_4$在第三个方程中由于变量$x_2$的引入变得不再显著，应将$x_4$从方程中剔除。

2．对$\mathbf{R}^{(3)}$的第4列再作一次消去变换，得$\mathbf{R}^{(4)}$

$$\mathbf{R}^{(4)}=\begin{bmatrix}1.0551 & -0.2412 & -0.8360 & -0.0243 & 0.5741\\ -0.2412 & 1.0551 & 0.0518 & -0.9674 & 0.6850\\ 0.8360 & -0.0518 & 0.3183 & -0.1252 & 0.0339\\ 0.0243 & 0.9674 & -0.1252 & 0.0528 & -0.0139\\ -0.5741 & -0.6850 & 0.0339 & -0.0139 & 0.6213\end{bmatrix}。$$

剔除x_4后有如下主要结果：

回归方程中包含的自变量为x_1和x_2，其标准回归系数为

$$b_1^{*(4)}=0.5741\ ,\quad b_2^{*(4)}=0.6850\ ,$$

代入（3-28）式求得一般回归系数：

$$b_1^{(4)}=1.4682\ ,\quad b_2^{(4)}=0.6622\ ,\quad b_0^{(4)}=52.58\ 。$$

因此，第四个回归方程为：

$$\hat{y}=52.58+1.4682x_1+0.6622x_2\ 。$$

其剩余平方和

$$Q^{(4)}=0.0213\ 。$$

第Ⅴ步，考虑有无变量可再剔除和有无变量可再引入。

1．对方程中的变量x_1，x_2进行显著性检验

$$p_1^{(4)}=\left(r_{1y}^{(4)}\right)^2/r_{11}^{(4)}=(0.5741)^2/1.0551=0.3124\ ,$$

$$p_2^{(4)}=\left(r_{2y}^{(4)}\right)^2/r_{22}^{(4)}=(0.6850)^2/1.0551=0.4447\ ,$$

$$F_1^{(4)}=\frac{p_1^{(4)}/1}{Q^{(4)}/(n-l-1)}=\frac{10\times0.3124}{0.0213}=146.52\ ,$$

$$F_2^{(4)}=\frac{p_2^{(4)}/1}{Q^{(4)}/(n-l-1)}=\frac{10\times0.4447}{0.0213}=208.58\ 。$$

它们都大于$F_a=3.28$临界值，所以x_1，x_2的回归作用显著。

2．对不在方程中的变量x_3，x_4进行检验，看是否还有可被引入的变量

$$p_3^{(4)}=\left(r_{3y}^{(4)}\right)^2/r_{33}=(0.0339)^2/0.3183=0.0036\ ,$$

$$p_4^{(4)}=\left(r_{4y}^{(4)}\right)^2/r_{44}=(-0.0139)^2/0.0528=0.0037\ ,$$

$$F_3^{(4)}=\frac{p_3^{(4)}/1}{Q^{(4)}/(n-l-1)}=\frac{10\times0.0036}{0.0213}=1.69\ ,$$

$$F_4^{(4)}=\frac{p_4^{(4)}/1}{Q^{(4)}/(n-l-1)}=\frac{10\times0.0037}{0.0213}=1.71\ 。$$

因为$F_3^{(4)}$和$F_4^{(4)}$都小于$F_a=3.28$临界值，说明变量x_3，x_4不需要再引入到回归方程中去。

3．最后得到的最优回归方程为

$$\hat{y}=52.58+1.4682x_1+0.66222x_2\ 。$$

4．对整个回归方程进行显著性检验

$$F=\frac{U/l}{Q/(n-l-1)}=\frac{(1-0.0213)/2}{0.0213/(13-2-1)}=229.5\ 。$$

在$\alpha=0.05$的水平上，查表$F_{(0.05)}(2,\ 10)=4.10$，

在$\alpha=0.01$的水平上，查表$F_{(0.01)}(2,\ 10)=7.56$。

所以，不论在$\alpha=0.05$或$\alpha=0.01$水平上，回归方程均高度显著。

5．计算拟合度c和复相关系数R

$$c=U/S_{yy}=(1-Q^{(4)})/S_{yy}=(1-0.0213)/1=97.87\%\ ,$$

$$R=\sqrt{c}=0.9893\ 。$$

从该例的逐步回归过程中可看出，变量x_4是最初引入方程的，但由于变量x_1和x_2的逐步引入，使得x_4在方程中不再起作用，最后被剔除出方程。变量x_3因始终被认为作用不显著而未被引入方程。所以，在多元回归分析中，自变量对因变量贡献的大小，不但要看某个自变量本身，而且要看自变量间的交互作用。逐步回归分析是从p个自变量中选择一个最佳的变量组合。在这个变量组合中任何减少或增加一个变量，都会影响对因变量的回归效果。

第六节　几个问题讨论

一、估计的标准误差

回归分析可将因变量y分成两部分，一部分是可以用p个自变量x_1，x_2，…，x_p的函数关系来表征的回归平方和；另一部分是剩余部分。如果说回归方程是正确的，则剩余平方和Q基本上可以看作是随机因素引起的。它可以用来计算估计误差S_y的大小：

$$S_y=\sqrt{Q/(n-p-1)}\ , \tag{3-29}$$

其中Q为剩余平方和，n为样品数，p为回归方程中自变量的个数。

S_y提供了一个回归估值准确度的量度，因此，可以在回归方程中加上$\pm S_y$项，即

$$\hat{y}=b_0+b_1x_1+b_2x_2+\cdots+b_px_p\pm S_y\ 。$$

在回归分析中，假设y_i呈正态分布，经过整个回归分析过程，得到它的两个参数，其数学期望$\hat{y}_i$，标准离差S_y。即y_i服从正态分布$N(\hat{y}_i,\ S_y)$。因此，在用回归值$\hat{y}$估计预测值y时就有

落在$(\hat{y}-S_y,\ \hat{y}+S_y)$的概率是68.3%；

落在$(\hat{y}-2S_y,\ \hat{y}+2S_y)$的概率是95.6%；

落在$(\hat{y}-3S_y,\ \hat{y}+3S_y)$的概率是99.7%；

落在$(\hat{y}-0.67S_y,\ \hat{y}+0.67S_y)$的概率是50%；

落在$(\hat{y}-1.96S_y,\ \hat{y}+1.96S_y)$的概率是95%。

二、回归方程的稳定性

回归方程是从作为随机变量的实际观测数据中计算出来的，而一批观测数据只是母体的一个子样，当取不同的子样时所建立起来的回归方程是否一样？即考查方程中回归系数b_i（$i=0$，1，…，p）是否相同？这种回归系数b_i随不同子样波动的性质称回归方程的稳定性，回归方程越稳定越好。

一般来说回归方程的稳定性与以下两个因素有关：

1．子样的大小，即样品数n的多少。n越大，建立的回归方程越稳定。有人提出一个经验性的标准，$n>10\times p$，即当样品数大于自变量个数的10倍时，所建立的回归方程才有意义。一般来说，样品数起码也要达到一个小子样（$n=30$）的容量才有统计规律。

2．假设变量的测试值是准确的前提下，回归方程的稳定性与变量的取值范围有关。变量取值范围愈大方程的稳定性愈好。

从第2点可得到推论，运用回归方程来预测时，预测值在y平均值附近时较可靠，远离y平均值时，预测的效果可能差一些，当超出建立回归方程的y观测值范围时，预测的结果是不可信的。

三、非线性回归

如果自变量与因变量之间为非线性关系时，通常的处理方法是将非线性关系化为线性关系，然后再作回归分析。当因变量与自变量为某种已知函数关系时，可以先对变量直接进行函数变换，当它们之间的关系不清楚时，可先作出散点图，大致估计它们之间的函数关系。常用的有以下几种：

1．双曲线 （图3-1）

$$\frac{1}{y}=a+\frac{b}{x}\ 。$$

变换时，令$y'=\frac{1}{y}$，$x'=\frac{1}{x}$，得 $y'=a+bx'$，即将双曲线变为直线。

2．幂函数 （图3-2）

$$y=dx^b\ 。$$

变换前，两边取常用对数，得 $\lg y=\lg d+b\lg x$，然后令$y'=\lg y$，$x'=\lg x$，$a=\lg d$，得直线方程 $y'=a+bx'$ 。

3．指数函数 （图3-3）

$$y=de^{bx}\ 。$$

变换时，先两边取自然对数，得 $\ln y=\ln d+bx$，令 $y'=\ln y$，$a=\ln d$，得 $y'=a+bx$ 。

4．对数函数 （图3-4）

$$y=a+b\lg x\ 。$$

令$x'=\lg x$，得 $y=a+bx'$ 。

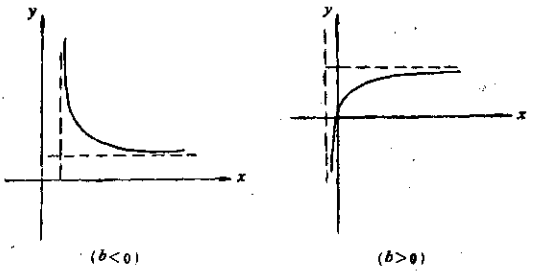


图3-1 双曲函数图象

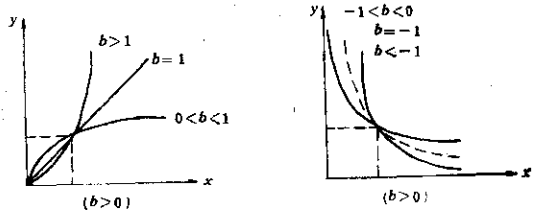


图3-2 幂函数图象

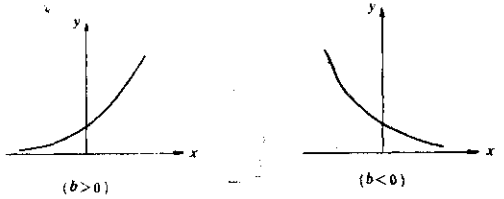


图3-3 指数函数图象

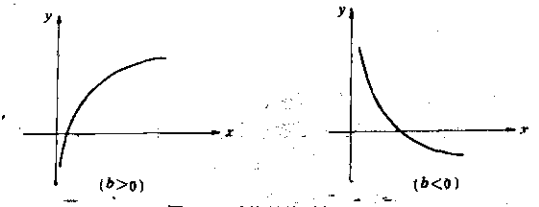


图3-4 对数函数图象

上面列举了几种常用的变换方法，虽然还有别的函数变换，但在实际工作中，对于数学模型所要求的地质变量性质，并非所有地质数据都能通过变换而使其得到满足的，研究者需要对具体问题具体分析，有时数据可能是多重母体混合分布，这就不能单纯用变换来解决问题，甚至会碰到变换不当使效果适得其反。

第七节 回归分析在地球化学中的应用

回归分析作为一种数学方法已经研究得比较深入了，但是它在地球化学工作中的应用却是比较灵活的。有的人能够熟练地应用并取得很好的效果；也有的人虽然能运用，但得不到好的效果。这主要取决于人们对回归分析数学方法的理解和地球化学知识多少两个方面。这里介绍回归分析在某些方面的应用。

一、用来研究某种地质作用与其它因素之间的关系

河北省某接触交代型铁矿，铁矿与燕山期中性岩浆侵入体有成因联系。矿体多产出于接触带，在内接触带的中性岩浆杂岩中普遍发育着钠长石化现象。

该区中性岩浆杂岩根据蚀变程度可分为4类：①未蚀变岩；②钠化岩石；③强烈钠化岩石；④叠加蚀变岩石。各类钠质交代岩石有以下基本特征：钠化岩石中钠长石含量占(70—80)%；中长石、钾长石被钠长石交代，角闪石被低铁阳起石、透闪石和次透辉石交代；磁铁矿基本消失，钠长石进而交代低铁阳起石、透闪石和次透辉石。叠加蚀变岩是在钠化以后一系列蚀变作用对钠质交代岩石进一步改造的产物。

钠质交代围岩蚀变是铁矿的一种找矿标志。为研究钠质交代与其它化学成分之间的关系，刘承祚等（1976）在本区选取了二长岩及钠质交代岩标本41块，以Na_2O为因变量，SiO_2、TiO_2、Al_2O_3、Fe_2O_3、FeO、MgO、CaO、K_2O为自变量，应用逐步回归分析

方法对其进行研究。

对于强烈钠化二长岩得到下列回归方程：

$$Na_2O = 8.99 - 1.01K_2O - 0.44MgO + 0.56FeO \text{ 。}$$

逐步回归第一步选进的变量为MgO，使剩余平方和Q由2.59降至1.85。说明Na_2O与MgO线性关系密切。第二步引进的变量为K_2O，使Q由1.85降至0.96，说明钠化过程中钾长石很快被钠长石交代（偏回归系数为－1.01），随着钠化作用的增强，K_2O有减少趋势。第三步选进的变量是FeO，偏回归系数为＋0.56，即随着钠的增加，FeO也增加。从这个例子可看出，逐步回归过程在一定程度上反映了钠化与铁矿形成之间的关系。

二、预测

回归分析应用最广的是预测。

上例利用所得回归方程，除进行铁矿成因解释以外，还能相当精确地预测Na_2O的含量值。在叠加蚀变二长岩中Na_2O的实测值与回归估值比较见表3-8。

表3-8　回归估值与Na_2O实测值比较表

标　本　号	北2-43	中760-15	中760-16	北2-42	北8-1	北2-23	北2-42
Na_2O实测值(%)	5.60	1.12	1.70	4.90	5.85	4.89	6.34
Na_2O回归估值(%)	5.61	1.12	1.70	4.90	5.85	4.89	6.34
误　　差	0.01	0.00	0.00	0.00	0.00	0.00	0.00

某矿区钒钛磁铁矿矿石的基本分析项目有TFe、TiO_2、V_2O_5。矿山每年的化验工作量很大，需要化验样品数万件。能否建立TFe与TiO_2及V_2O_5的回归方程，在工作中只化验TFe一项，而用TiO_2、和V_2O_5的回归估值分别代替各自的化验值，以减少TiO_2和V_2O_5的化验工作量？

这里先用回归分析研究矿石中V_2O_5与TFe的相关关系。从该矿的三个矿区各自抽取30个样品的化验资料作为样本，来研究V_2O_5与TFe的相关关系，得到以下的结果：

$\hat{y}_1 = -0.0323 + 0.0115x_1$，$Q_1 = 0.0010$，$r_1 = 0.988$；

$\hat{y}_2 = -0.0227 + 0.0089x_2$，$Q_2 = 0.0099$，$r_2 = 0.987$；

$\hat{y}_3 = -0.0362 + 0.0094x_3$，$Q_3 = 0.0166$，$r_3 = 0.942$。

（其中y为V_2O_5，x为TFe，1、2、3为三个矿区）

可见，三个矿区的相关系r均大于0.9，说明钒与铁高度相关；三个矿区的剩余标准差Q均小于0.02。说明用样品TFe含量代入回归方程计算得到的V_2O_5含量与实际化验值的差别很小，一般不超过0.02，在化验误差允许范围之内。据此，在勘探工作中和矿山生产中完全可以用V_2O_5的回归分析预测值代替化验值。TiO_2与V_2O_5的情况相似，也与TFe高度相关，建立的回归方程有效。因此，矿山在生产中每年只需要取少量的组合样品化验TFe、TiO_2和V_2O_5含量，以不断修正回归方程的系数，保证对V_2O_5和TiO_2估值的精确性，在大量的基本分析样品中只需化验TFe一项指标即可，这样每年可节约数十万元化验费用。

应当注意的是，在预测时要求地质条件相同。也就是说，不能某地建立的回归方程，搬到地质条件不同的另一地区去用。而且，一般来说，预测值的范围只能限制在建立回归

方程时所用的观测值的范围以内，否则都会导致不正确的结论。

三、在分散流背景值与异常划分中的应用

化探的基本问题是要把河流沉积物的金属量值分为二组，一组称为异常，另一组称为背景。传统的做法，是把高值者看作异常。然而，并非所有的高值都表示有矿，一些较低的数值也可能反映有矿。矿体仅仅是影响河流沉积物金属含量的多种地质因素之一。所以，在划定背景值上限时，应该根据样品地质条件的变化而变化。W.Rose (1970) 提出通过用回归分析估计总的地质效应来评价不同背景值上限的问题。一旦估计出总的地质效应，背景效应就可以被消除。金属含量的观测值与回归估值间的偏差值，可视为地质条件异常的效应，其中一种可能是矿致异常。

W.Rose在解释宾州东南部（7500mile^{2}区域内）约3000个分散流样品数据时，就遇到了这类问题。该区出露的岩石种类包括前寒武系的片麻岩、花岗岩、辉长岩、蛇纹岩，下古生代的海相灰岩、页岩和砂岩，以及上古生代三迭系的碎屑沉积岩，其中部分是非海相地层，有辉绿岩侵入。这次研究中，原来划分“异常”和“背景”样品，是用每种元素全测区的背景值划分的，但是出现了由于岩石种类不同而引起的变化，因此认识到需要用更精确的方法。回归分析被认为最有用且便于解释的方法。一旦在样品中定量地确定金属含量与这些金属环境变量之间的关系时，就能计算出每个样品中未知金属的预期含量，并将此预期值与观测值进行比较。这个过程相当于在较简单的解释方法中用“校正背景值”来替换固定背景值的计算方法。

引起河流沉积物中微量元素背景值变化的原因，除源地岩性外，还有 Fe-Mn 氧化物的吸附作用，离子交换能力强的粘土含量和含金属有机物的含量等。在回归分析时，将2800个样品按水系块段划分成 460 个分区，每一个分区中取包括 Fe、Mn、岩类所占面积百分数和地形高差等29种自变量。这样，Zn、Cu、Ni、Co、Cr 和V各元素和这些自变量的函数关系就可以用下面的回归方程式表示：

$$x_m = b_0 + b_1\mathrm{Mn} + b_2\mathrm{Fe} + b_3\mathrm{R}_1 + b_4\mathrm{R}_{k-2} + b_{k+1}\mathrm{H}$$

式中，x_m为微量金属m的对数含量，b为回归系数，Mn、Fe为锰和铁的对数含量，R为岩类所占面积百分比的对数，H为地形高差估计对数。

根据回归方程式，可以计算出每一分区的“校正背景值”。超过校正背景值下限的才被看作异常。用这样的方法圈定的异常与传统方法比较，有明显的差别。在该地区的东部，以前根据原始数据圈为异常的一些分区，现在根据新的下限值，则都划为背景分区。在新发现的异常的特点表明，前寒武系岩类分布地区锌的区域性高含量可能掩盖了灰岩分布区中有价值的异常，而灰岩分布区锌的背景值很低，可能产有如宾州弗里登斯维尔 (Friedensville) 的阿伯拉契亚型的锌矿床。特别有意义的是分布在宾州兰卡斯正西北的异常分区。这个分区有帮堡老铅锌矿和几个其它的矿山和矿点。根据原始数据，用传统方法并不能将帮堡分区从背景分区中划为异常而被漏掉，但是根据回归分析方法，它明显是一个异常分区。把这个分区识别为异常这一事实表明，用回归方法圈出的其它异常（尤其在灰岩地区）可能是铅锌矿床远景区。

第四章　趋势分析

第一节　概　述

趋势分析是一种研究随机变量在空间位置上变化规律的数理统计方法。它用某种数学模型，去拟合实测模型。在地质工作中经常需要研究某种地质特征的空间分布特征与变化规律，例如，为了查明某一地区的地质构造，需要研究某些地层单位的厚度或某一标准层的高程在该地区内的系统变化；为了了解某一岩浆侵入体物质成分空间变化特征，需要研究其矿物成分或化学成分在该岩体内的变化规律；又如在化探工作中为了发现矿化异常，需要研究化学元素在测区内的“区域趋势”和“局部异常”。地质上的这些变量常常随空间位置的变化而改变，可以说它们是空间位置的函数。这种随机变量可借用地质统计学中的一个名词——区域化变量来称呼它们，或者简单地理解为空间变量。

趋势分析根据数学手段和方法的不同，可以分为滑动平均（或叫移动平均）、多项式拟合趋势分析和调和趋势分析。本章主要介绍在地球化学工作中较常用的多项式拟合趋势分析。

趋势分析依空间的维数，又可分为一维，二维和三维。对于多项式趋势分析，不同维中按自变量的最高次数，又可分为一次，二次，…，六次等等。

由于地球化学变量在空间上表现为既有随机性，又有结构性(受周围点的含量控制)，因此可以采用趋势分析来进行研究。

趋势分析的任务有：

1．确定测区中地球化学变量空间分布的数学模型。

2．区分测区中地球化学变量的“区域变化趋势”和“局部异常”。

第二节　多项式趋势分析的数学方法

本节着重介绍二维多项式拟合趋势面的数学方法。

设以 $z(x, y)$ 表示所研究的地质特征，其中 (x, y) 是平面上点的坐标。第一节中已经指出，趋势分析的一个任务是将z的观测值的变化分解为两部分：

$$z(x, y)=\tau(x, y)+e\ , \tag{4-1}$$

其中 $\tau(x, y)$ 描述了区域趋势，称为“趋势值”。而e则表达了局部变化，称为“剩余值”，见图4-1。

在研究区内的n个测点，对于已知的观测数据

$$z_i: x_i,\ y_i\ ,\qquad i=1,\ 2,\ \cdots,\ n\ ,$$

通常用回归分析来求得它们的趋势值和剩余值，即根据已知n个点的坐标和观测值求出一个以z为因变量，坐标(x, y)为自变量的回归方程

$$\hat{z} = f(x, y)\ ,$$

使得

$$Q = \sum_{i=1}^{n}\left[z_i - f(x_i, y_i)\right]^2$$

达到最小。这就是在最小二乘意义下的曲面拟合问题。

通常选择多项式作为回归方程，是因为任何函数在一个适当的范围内总可以用多项式来逼近，而且调整多项式的次数可以使所求的回归方程适合问题的需要。

一次：$f(x, y) = a_0 + a_1x + a_2y$ ，

二次：$f(x, y) = a_0 + a_1x + a_2y + a_3x^2 + a_4xy + a_5y^2$ ，

三次：$f(x, y) = a_0 + a_1x + a_2y + a_3x^2 + a_4xy + a_5y^2$ ，

$\qquad + a_6x^3 + a_7x^2y + a_8xy^2 + a_9y^3$ ，

… … …

多项式的次数愈高则趋势值愈接近观测值，而剩余值就愈小。当然，具体工作中必须根据实际需要恰当地选择趋势面的次数，一般很少超过五次或六次，见图4-2。

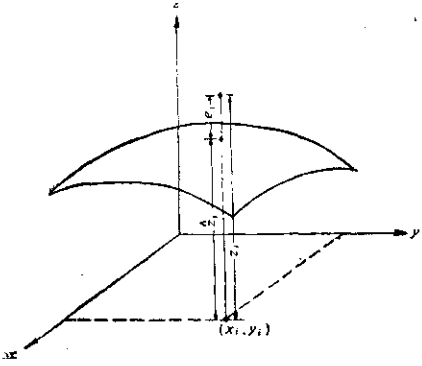


图4-1 趋势值和剩余值示意图

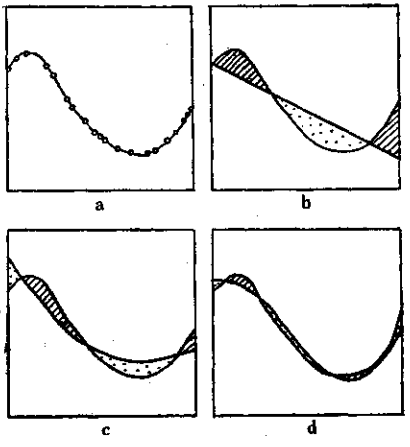


图4-2 趋势面的次数和观测值及剩余值的关系
a—原始观测点及其连线；b—以直线拟合；c—抛物线趋势；d—三次趋势；斜线阴影部分为正剩余，小点阴影部分为负剩余。(据Davis 1973)

如果记：

$$x = x_1,\ y = x_2,\ x^2 = x_3,\ xy = x_4,\ y^2 = x_5,\ \cdots$$

则趋势面方程可以写为：

$$\hat{z} = a_0 + a_1x_1 + a_2x_2 + a_3x_3 + \cdots + a_px_p\ . \qquad (4\text{-}2)$$

残差平方和为

$$Q=\sum_{i=1}^{n}(z_i-\hat{z}_i)^2=\sum_{i=1}^{n}[z_i-(a_0+a_1x_{1i}+a_2x_{2i}+\cdots+a_px_{pi})]^2\ , \qquad (4\text{-}3)$$

这里Q是a_0，a_1，…，a_p的函数。

根据最小二乘法的原则，要选择这样的系数a_0，a_1，…，a_p，($p<n$)，以使得Q达到极小。为此，求Q对a_0，a_1，…，a_p的偏导数：

$$\begin{cases}\dfrac{\partial Q}{\partial a_0}=-2\sum_{i=1}^{n}[z_i-(a_0+a_1x_{1i}+\cdots+a_px_{pi})]\cdot 1\\ \dfrac{\partial Q}{\partial a_1}=-2\sum_{i=1}^{n}[z_i-(a_0+a_1x_{1i}+\cdots+a_px_{pi})]\cdot x_{1i}\\ \cdots\cdots\\ \dfrac{\partial Q}{\partial a_p}=-2\sum_{i=1}^{n}[z_i-(a_0+a_1x_{1i}+\cdots+a_px_{pi})]\cdot x_{pi}\end{cases}\ . \qquad (4\text{-}4)$$

令$\dfrac{\partial Q}{\partial a_i}=0$，　($i=0$，1，2，…，$p$)，即得

$$\begin{cases}n\cdot a_0+\Sigma x_{1i}\cdot a_1+\cdots+\Sigma x_{pi}\cdot a_p=\Sigma z_i\\ \Sigma x_{1i}\cdot a_0+\Sigma x_{1i}^2\cdot a_1+\cdots+\Sigma x_{1i}x_{pi}\cdot a_p=\Sigma x_{1i}z_i\\ \cdots\cdots\\ \Sigma x_{pi}\cdot a_0+\Sigma x_{1i}x_{pi}\cdot a_1+\cdots+\Sigma x_{pi}^2\cdot a_p=\Sigma x_{pi}z_i\end{cases}\ . \qquad (4\text{-}5)$$

(4-5)式包括$p+1$个线性代数方程，解此线性方程组即得$p+1$个系数a_0，a_1，…，a_p。

若用$x=x_1$，$y=x_2$，$x^2=x_3$，$xy=x_4$，$y^2=x_5$，…回代到线性方程组(4-5)中，则实际上求解的线性方程组（以二维三次为例）是：（见下页）

此类矩阵是对称、正定的，解方程组可求出系数a_0、a_1、a_2、a_3、a_4、a_5、a_6、a_7、a_8、a_9，从而得到趋势面方程为

$$\hat{z}_i=a_0+a_1x_i+a_2y_i+a_3x_i^2+a_4x_iy_i+a_5y_i^2+a_6x_i^3+a_7x_i^2y_i+a_8x_iy_i^2+a_9y_i^3\ . \qquad (4\text{-}7)$$

趋势面方程建立后，将第i点的坐标(x_i，y_i)　($i=1$，2，…，n)代入方程(4-7)中，即可求出这些点的趋势值($\hat{z}_i$)。从理论上，在图幅范围内，只要给予任一坐标(x_i，y_i)，都可求得该点的趋势值，用这些值按一定间隔画出趋势等值线图，从而得出某一地质变量的变化趋势。这个趋势面方程和趋势等值线图，可以看作该地质变量在该地区的空间分布特征。

若从观测值(z_i)中减去其趋势值($\hat{z}_i$)，可得到诸点的剩余值($e_i=z_i-\hat{z}_i$)。根据这些剩余值，（也可进一步作某些数学处理）编制剩余等值线图，从而了解该地质变量的局部变化特征。

用同样的方法，可以得到更高次的趋势面方程。不过，次数越高未知数越多，求解越复杂。一般若配一个p次趋势面，其方程为

$$
\begin{array}{|ccc|ccc|ccc|c|c|}
\hline
N & \Sigma x & \Sigma y & \Sigma x^2 & \Sigma xy & \Sigma y^2 & \Sigma x^3 & \Sigma x^2y & \Sigma xy^2 & \Sigma y^3 & a_0 & \Sigma z \\
\hline
\Sigma x & \Sigma x^2 & \Sigma xy & \Sigma x^3 & \Sigma x^2y & \Sigma xy^2 & \Sigma x^4 & \Sigma x^3y & \Sigma x^2y^2 & \Sigma xy^3 & a_1 & \Sigma xz \\
\Sigma y & \Sigma xy & \Sigma y^2 & \Sigma x^2y & \Sigma xy^2 & \Sigma y^3 & \Sigma x^3y & \Sigma x^2y^2 & \Sigma xy^3 & \Sigma y^4 & a_2 & \Sigma yz \\
\hline
\Sigma x^2 & \Sigma x^3 & \Sigma x^2y & \Sigma x^4 & \Sigma x^3y & \Sigma x^2y^2 & \Sigma x^5 & \Sigma x^4y & \Sigma x^3y^2 & \Sigma x^2y^3 & a_3 & \Sigma x^2z \\
\Sigma xy & \Sigma x^2y & \Sigma xy^2 & \Sigma x^3y & \Sigma x^2y^2 & \Sigma xy^3 & \Sigma x^4y & \Sigma x^3y^2 & \Sigma x^2y^3 & \Sigma xy^4 & a_4 & \Sigma xyz \\
\Sigma y^2 & \Sigma xy^2 & \Sigma y^3 & \Sigma x^2y^2 & \Sigma xy^3 & \Sigma y^4 & \Sigma x^3y^2 & \Sigma x^2y^3 & \Sigma xy^4 & \Sigma y^5 & a_5 & \Sigma y^2z \\
\hline
\Sigma x^3 & \Sigma x^4 & \Sigma x^3y & \Sigma x^5 & \Sigma x^4y & \Sigma x^3y^2 & \Sigma x^6 & \Sigma x^5y & \Sigma x^4y^2 & \Sigma x^3y^3 & a_6 & \Sigma x^3z \\
\Sigma x^2y & \Sigma x^3y & \Sigma x^2y^2 & \Sigma x^4y & \Sigma x^3y^2 & \Sigma x^2y^3 & \Sigma x^5y & \Sigma x^4y^2 & \Sigma x^3y^3 & \Sigma x^2y^4 & a_7 & \Sigma x^2yz \\
\Sigma xy^2 & \Sigma x^2y^2 & \Sigma xy^3 & \Sigma x^3y^2 & \Sigma x^2y^3 & \Sigma xy^4 & \Sigma x^4y^2 & \Sigma x^3y^3 & \Sigma x^2y^4 & \Sigma xy^5 & a_8 & \Sigma xy^2z \\
\Sigma y^3 & \Sigma xy^3 & \Sigma y^4 & \Sigma x^2y^3 & \Sigma xy^4 & \Sigma y^5 & \Sigma x^3y^3 & \Sigma x^2y^4 & \Sigma xy^5 & \Sigma y^6 & a_9 & \Sigma y^3z \\
\hline
\end{array}
\quad = \quad \Sigma z \qquad (4\text{-}6)
$$

$$\hat{z}_i = a_0 + a_1 x_i + a_2 y_i + a_3 x_i^2 + a_4 x_i y_i + a_5 y_i^2 + \cdots + a_k y_i^k \quad , \tag{4-8}$$

其中

$$k = p(p+3)/2 \quad , \tag{4-9}$$

$k+1$则是多项式的项数，即多项式待定系数的个数。

显然，二次趋势面包含了一次趋势面的成分，三次趋势面包含了一、二次趋势面的成分，p次趋势面包含了所有低于p次趋势面的成分。

上述只是一个地质变量在二维平面上的情况，即为二维趋势面。对于一维剖面和三维立体的趋势分析与二维类似。图4-3给出了一维、二维和三维的一次、二次和三次趋势方程

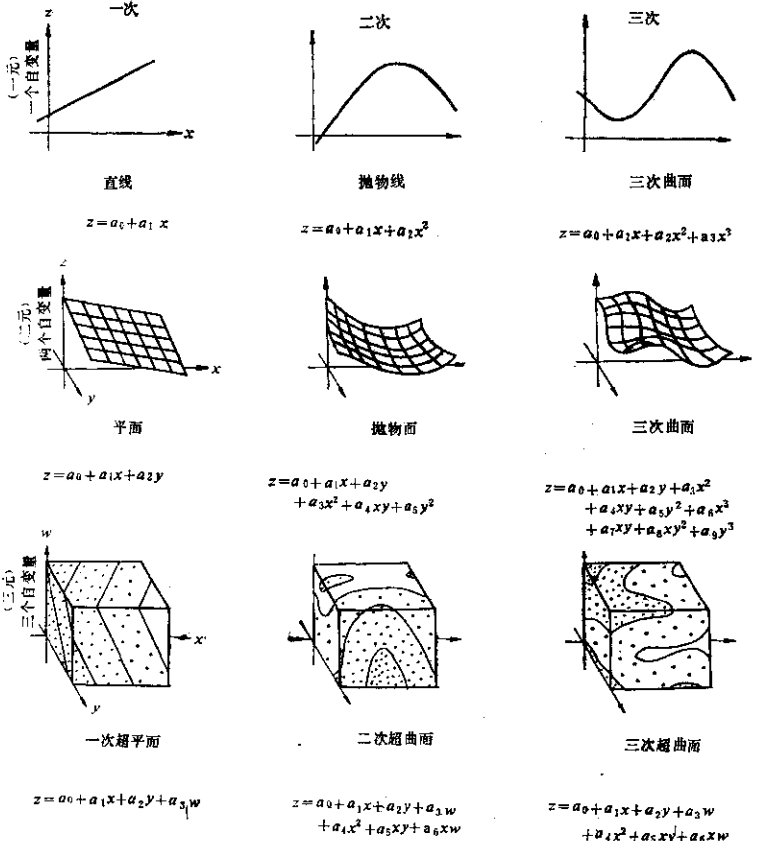


图4-3　趋势方程与几何图形之间的关系

式和图形之间的关系，更高次的趋势分析即可依此类推。

第三节　趋势方程的拟合度及显著性检验

一、趋势面的拟合度

得到趋势面方程后，可以计算其拟合度c。趋势面的拟合度，是指计算出的趋势面方程所代表的数学面与实际观测值（z_i）构成的曲面的逼近程度。

设有n个数据（z_i，$i=1$，2，…，n），其总的方差（即总离差平方和）

$$S_{总}=\sum_{i=1}^{n}(z_i-\bar{z})^2 \tag{4-10}$$

可以分解为两部分，即

$$S_{总}=\sum_{i=1}^{n}(\hat{z}_i-\bar{z})^2+\sum_{i=1}^{n}(z_i-\hat{z}_i)^2$$

令

$$S_{回}=\sum_{i=1}^{n}(\hat{z}_i-\bar{z}_i)^2\ , \tag{4-11}$$

$$S_{剩}=\sum_{i=1}^{n}(z_i-\hat{z}_i)^2\ , \tag{4-12}$$

则

$$S_{总}=S_{回}+S_{剩}\ 。 \tag{4-13}$$

$S_{回}$为回归离差平方和，代表由于趋势面方程所刻划的变差，其自由度为趋势面方程中自变量的个数p，$S_{剩}$为剩余离差平方和，代表局部异常和未加控制的因素所引起的误差，其自由度为$n-p-1$。趋势面的拟合度c，可以用回归平方和占总平方和的比例来衡量。也可以用剩余离差平方和的大小来衡量。一般来说拟合度都用百分数表示

$$\begin{aligned}c&=S_{回}/S_{总}\times 100\%\\&=(1-S_{剩}/S_{总})\times 100\%\ 。\end{aligned} \tag{4-14}$$

二、趋势面方程的显著性检验

与回归分析一样，趋势面方程建立之后，对其是否能反映因变量与自变量之间的客观规律，是否能够运用趋势面方程，尚须对其进行显著性检验。

1．趋势面方程的显著性检验

假设因变量与自变量及其高次幂和交叉乘积幂之间不相关，则偏回归系数均为零，即

$$H_0：a_1=a_2=\cdots=a_p=0\ ,$$

统计量

$$F=\frac{S_{回}/P}{S_{剩}/(n-p-1)}\ 。$$

在给定显著性水平α的情况下，服从第一自由度$f_1=p$，第二自由度$f_2=n-p-1$的F-分

布。查F分布表，得临界值$F_{\alpha}(p, n-p-1)$。如果$F \leqslant F_{\alpha}$，接受原假设H_0，则认为趋势面方程不显著；如果$F > F_{\alpha}$，拒绝原假设H_0，认为因变量与自变量之间相关密切，趋势面方程显著。

2．趋势面方程的逐次检验

趋势分析中，用增加自变量的次数，来提高因变量的拟合度。那么，趋势分析究竟做到哪一次好？用什么来衡量不同次数趋势方程的显著性？这就要对趋势方程进行逐次检验。

趋势方程的逐次检验，是检验$p+1$次趋势方程回归效果是否显著。也就是说，p次趋势面以后的趋势面方程，对1至p次回归有无新的贡献。仍然用F-检验法。

若p次趋势面方程，除常数项（a_0）外，包含q个系数（a_1，a_2，…，a_q）。$p+1$次趋势面方程除常数项外，包含k个系数（a_1，a_2，…，a_q，a_{q+1}，a_{q+2}，…，a_{k-1}，a_k）。假设趋势面方程增加一次，回归效果增加不显著，即增加项的偏回归系数均为零。

$$H_0 : a_{q+1} = a_{q+2} = \cdots = a_{k-1} = a_k = 0 \ 。$$

统计量

$$F = \frac{(S_{回}^{(p+1)} - S_{回}^{p})/(k-q)}{S_{剩}^{(p+1)}/k} \tag{4-15}$$

服从自由度为$f_1 = k - q$，$f_2 = k$的F-分布。

在给定显著性水平α情况下，查F-分布表，得临界值$F_{\alpha}(k-q, k)$。如果$F > F_{\alpha}$，否定原假设H_0，表明趋势面由p次增加到$p+1$次效果是显著的；如果$F \leqslant F_{\alpha}$，则说明趋势面方程增加一次其回归效果无显著增加，趋势面只要做到第p次就可以了。

第四节　几个问题的讨论

一、影响趋势面分析的主要因素

许多因素对趋势面分析产生重要影响，轻则使趋势面变形，重则使趋势面分析失效。这些因素主要有：因变量是否是空间位置自变量的函数关系；制图面积内控制点的数目和控制点的分布；正规方程组的病态问题。

在趋势分析中，假设因变量是位置坐标的函数$\hat{z}_i = f(x_i, y_i)$。这个假设在地质构造（褶曲）区研究中可以接受，但其它情况，有时就不一定。特别是，进一步用增加自变量的次数来提高拟合度的做法。因此，趋势分析并不是可以随意套用的。

下面讨论制图面积内控制点的数目和控制点的分布对趋势分析的影响。

显然制图面积内的控制点数或控制点的密度直接影响到趋势面和剩余面的形态。从数理统计的角度看，控制点数的下限必须超过多项式方程中系数的数目，否则将使回归分析的结果无效。在进行统计假设检验时，控制点数n决定着检验统计量F的自由度，而自由度必须足够大，以使F-检验具有统计学意义。当多项式系数的数目趋近控制点数时，F的自由度愈来愈小，只有极大的相关系数才可能是显著的，假设检验的有效性随着样本的减小而急剧下降。

在进行趋势面分析时，通常不考虑图幅边界以外的点。在图幅边界上控制点很少，甚

至根本没有控制点的情况下，几乎无法控制边界附近趋势面的形态。这样就产生了通常所说的“边界效应”。不论控制区内趋势面的坡度如何，将趋势值外推到控制区范围以外，将会偏离正确的趋势值，并且趋势面的次数愈高，外推结果的偏离愈大。因此，为了得到一个图幅完整的趋势面结果，最好控制点的范围稍大于图幅，这样就在图幅边界的边界效应集中地段形成一个“缓冲区”，在此区内的控制点对制图面积内的趋势面形态起约束、控制作用。

图幅内控制点的分布对于趋势面的形态也有显著的影响。图4-4表示控制点分布不均匀性对多项式趋势分析的影响。图4-4a表示控制点在一个盆状测试面上呈随机分布时所得到的二次趋势面，其拟合度超过95%，趋势面与原测试面基本相同。在图4-4b中，控制点在同一测试面上沿着一个狭长带分布，所得二次趋势面的拟合程度为93%，与图4-4a相比，趋势面的形状平行狭长带的延长方向发生畸变。上述事实清楚地说明控制点的分布对于趋势面的形状可产生严重的影响。还应注意，如果测试面有局部的“噪音”存在，则由于控制点不均匀分布所产生趋势面的畸变还要严重。

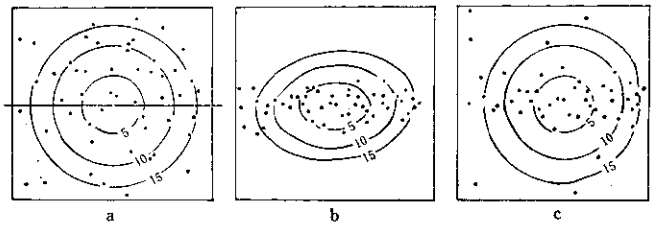


图4-4　控制点分布对趋势面的影响

因此，在设计控制点的分布时，要考虑到它的均匀性和随机性。所谓均匀性，指任一小单位面积中点的密度等于同样大小任一其它小面积中点的密度。所谓随机性是指采样时，一个采样点落在任一小面积内和落在同样大小的任一其它面积内的概率相等。

另外，制图面积内控制点的密度直接影响着趋势面形态的结构。密度大，能够更精细地刻划趋势面的内部结构；密度小，将忽略那些细微的变化。

二、正规方程组系数矩阵的病态问题

一般来说，趋势面方程的次数愈高，趋势面的拟合度也愈大，但事实并非完全如此。这是因为趋势分析中正规方程组系数矩阵具有下列两种值得注意的性质：

1．矩阵元素的大小级次相差悬殊；

2．高次趋势方程的系数矩阵接近于奇异矩阵。

因此，当趋势面方程次数增高时，正规方程组系数矩阵很快接近病态，即系数矩阵的行列式接近于0。病态矩阵在求逆求解时会引起很大的舍入误差。病态矩阵是影响趋势分析的重要因素。

解决的办法：

1．调整坐标比例尺，使自变量取值的数量尽可能在一位整数；

2．选择适当的求解求逆数学方法，对于矩形规则测网可采用正交多项式的方法。

第五节　规则测网趋势面分析的数学方法

当制图面积内控制点呈等间距矩形规则网格分布时，自变量x_i，y_i可以自动生成，利用正交多项式可拟合趋势面。这种方法的优点是：利用正交多项式拟合趋势面可以使正规方程组的系数矩阵变为对角阵，从而避免求解病态方程，提高方程解的精度；同时正规方程组变为p个一元方程，十分容易求解。

首先改变正规方程组（4-5）的写法，使其系数的概率意义更为明显。

记

$$\overline{z}=\frac{1}{n}\sum_{i=1}^{n}x_i\ ,\qquad \overline{x_j}=\frac{1}{n}\sum_{i=1}^{n}x_{ji}\ ,$$

并引进符号：

$$S_{jk}=\sum_{i=1}^{n}(x_{ji}-\overline{x_j})(x_{ki}-\overline{x_k})\ ,$$

$$S_{jz}=\sum_{i=1}^{n}(x_{ji}-\overline{x_j})(z_i-\overline{z})\ ,$$

$$S_{zz}=\sum_{i=1}^{n}(z_i-\overline{z})^2\ 。$$

在方程组（4-5）中，由第一个方程可以得到

$$a_0=\frac{1}{n}\sum_{i=1}^{n}z_i-\frac{1}{n}\left[\sum_{i=1}^{n}x_{1i}a_1+\cdots+\sum_{i=1}^{n}x_{pi}a_p\right]\ 。$$

用上面的符号，可以写成

$$a_0=\overline{z}-(\overline{x_1}a_1+\overline{x_2}a_2+\cdots+\overline{x_p}a_p)\ 。\tag{4-16}$$

剩余平方和的公式（4-3）可写成

$$Q=\sum_{i=1}^{n}\left[(z_i-\overline{z})-a_1(x_{1i}-\overline{x_1})-\cdots-a_p(x_{pi}-\overline{x_p})\right]^2\ 。\tag{4-17}$$

根据最小二乘法，由（4-17）式出发，求Q对a_j的偏导数，并使$\dfrac{\partial Q}{\partial a_j}=0$　$(j=1,2,\cdots,p)$，可得方程组

$$\begin{cases}S_{11}a_1+S_{12}a_2+\cdots+S_{1p}a_p=S_{1z}\\ S_{21}a_1+S_{22}a_2+\cdots+S_{2p}a_p=S_{2z}\\ \cdots\cdots\cdots\cdots\cdots\cdots\\ S_{p1}a_1+S_{p2}a_2+\cdots+S_{pp}a_p=S_{pz}\end{cases}\ 。\tag{4-18}$$

如果利用正交多项式就可以使正规方程组（4-18）的系数矩阵变为对角阵，即

$$\begin{Bmatrix}S_{11} & & & 0\\ & S_{22} & & \\ & & \ddots & \\ 0 & & & S_{pp}\end{Bmatrix},\qquad S_{jk}=\begin{cases}0 & \text{当}j\neq k\ ,\\ \neq 0 & \text{当}j=k\ 。\end{cases}$$

下面着重讨论一维的情形。设自变量x取等间距的值，不妨设$x_1=1$，$x_2=2$，…，$x_i=i$，…，$x_n=n$。如果$x_1=a+h$，$x_2=a+2h$，…，$x_n=a+nh$，则可作如下变换：

令 $x'=\frac{x-a}{h}$，这时就有$x'_1=1$，$x'_2=2$，…，$x'_n=n$。对应于x_i的观测值仍记为z_i。要配一个p次的多项式回归曲线：

$$\hat{z}=a_0+a_1x+a_2x^2+\cdots+a_px^p \text{ 。}$$

把x，x^2，…，x^p作适当变换，而要求$\hat{z}$形如

$$\hat{z}=b_0+b_1\psi(x)+b_2\psi(x)+\cdots+b_p\psi_p(x)=\sum_{j=0}^{p}b_j\psi_j(x)$$

其中$\psi_0(x)=1$，$\psi_j(x)$是j次多项式。现在研究应如何选择这些多项式，以使得正规方程组的系数矩阵成为对角矩阵。现将$\psi_j(x)$看作新的变量，其回归系数b_j应由下面正规方程组来决定：

$$\begin{cases} S_{11}b_1+S_{12}b_2+\cdots+S_{1p}b_p=S_{1z} \\ S_{21}b_1+S_{22}b_2+\cdots+S_{2p}b_p=S_{2z} \\ \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \\ S_{p1}b_1+S_{p2}b_2+\cdots+S_{pp}b_p=S_{pz} \end{cases} \text{ 。}$$

而

$$b_0=\bar{z}-[\bar{\psi}_1(x)b_1+\bar{\psi}_2(x)b_2+\cdots+\bar{\psi}_p(x)b_p] \text{ ，}$$

其中

$$S_{jk}=\sum_{i=1}^{n}[\psi_j(x_i)-\bar{\psi}(x)][\psi_k(x_i)-\bar{\psi}(x)] \text{ ，}$$

$$S_{jz}=\sum_{i=1}^{n}[\psi_j(x_i)-\bar{\psi}(x)](z_i-\bar{z}) \text{ 。}$$

再选择这样的$\psi_j(x)$，使得

$$\begin{cases} \sum_{i=1}^{n}\psi_j(x_i)=0 \qquad (j=1,\ 2,\ \cdots,\ p) \\ \sum_{i=1}^{n}\psi_j(x_i)\psi_k(x_i)=0 \qquad (j\neq k) \end{cases} \tag{4-19}$$

于是

$$\bar{\psi}_j(x)=\frac{1}{n}\sum_{i=1}^{n}\psi_j(x_i)=0 \text{ ，}$$

$$S_{jk}=\begin{cases} 0 & (j\neq k) \text{ ，} \\ \sum_{i=1}^{n}\psi_j^2(x_i) & (j=k) \text{ ，} \end{cases}$$

$$S_{jz}=\sum_{i=1}^{n}\psi_j(x_i)z_i \text{ 。}$$

这时正规方程组简化为

$$\begin{cases} \sum_{i=1}^{n}\psi_1^2(x_i)\cdot b_1 = \sum_{i=1}^{n}\psi_1(x_i)z_i \\ \sum_{i=1}^{n}\psi_2^2(x_i)\cdot b_2 = \sum_{i=1}^{n}\psi_2(x_i)z_i \\ \cdots\qquad\qquad \cdots \\ \sum_{i=1}^{n}\psi_p^2(x_i)\cdot b_p = \sum_{i=1}^{n}\psi_p(x_i)z_i \end{cases}, \tag{4-20}$$

即

$$\begin{bmatrix} \sum_{i=1}^{n}\psi_1^2(x_i) & & & 0 \\ & \sum_{i=1}^{n}\psi_2^2(x_i) & & \\ & & \ddots & \\ 0 & & & \sum_{i=1}^{n}\psi_p^2(x_i) \end{bmatrix} \cdot \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_p \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^{n}\psi_1(x_i)z_i \\ \sum_{i=1}^{n}\psi_2(x_i)z_i \\ \vdots \\ \sum_{i=1}^{n}\psi_p(x_i)z_i \end{bmatrix} .$$

同时，因为正规方程组已变为p个一元方程，因而可以立即求出回归系数

$$b_j = \frac{\sum_{i=1}^{n}\psi_j(x_i)z_i}{\sum_{i=1}^{n}\psi_j^2(x_i)} \qquad (j=1,\ 2,\ \cdots,\ p)\ , \tag{4-21}$$

而b_0也简化为

$$b_0 = \bar{z} = \frac{1}{n}\sum_{i=1}^{n}z_i\ 。 \tag{4-22}$$

条件（4-19）称为“正交性”。满足正交性条件（4-19）的这种多项式是存在的。费歇（R·A·Fisher）推导出了这样的一组多项式，以满足条件（4-19）：

$$\begin{cases} \psi_1(x) = x - \bar{x} \\ \psi_2(x) = (x-\bar{x})^2 - \dfrac{n^2-1}{12} \\ \psi_3(x) = (x-\bar{x})^3 - \dfrac{3n^2-7}{20}(x-\bar{x}) \\ \psi_4(x) = (x-\bar{x})^4 - \dfrac{3n^2-13}{14}(x-\bar{x}) + \dfrac{3(n^2-1)(n^2-9)}{560} \\ \cdots\cdots \\ \psi_{j+1}(x) = \psi_1(x)\psi_j(x) - \dfrac{j^2(n^2-j^2)}{4(4j^2-1)}\psi_{j-1}(x) \end{cases} \tag{4-23}$$

对于二维的情形，也可以用正交多项式来配多元多项式回归方程。同理，要求各自变

量取等间隔的值。

设$x_i=i$ $(i=1, 2, \cdots, n)$，$y_j=j$ $(j=1, 2, \cdots, m)$，以z_{ij}表示对应于x_i，y_j的观测值。现在要求用正交多项式对z_{ij}配一个二维二次回归方程：

$$\hat{z}=a_{00}+a_{10}x+a_{01}y+x_{20}x^2+a_{02}y^2+a_{11}xy \ 。$$

利用正交多项式

$$\psi_{10}(x)=x-\bar{x}\ ,\qquad \psi_{20}(x)=(x-\bar{x})^2-\frac{n^2-1}{12}\ ,$$

$$\psi_{01}(y)=y-\bar{y}\ ,\qquad \psi_{02}(y)=(y-\bar{y})^2-\frac{m^2-1}{12}\ 。$$

将上式改写为回归方程的如下形式：

$$\hat{z}=b_{00}+b_{10}\psi_{10}(x)+b_{01}\psi_{01}(y)+b_{20}\psi_{20}(x)+b_{02}\psi_{02}(y)+b_{11}\psi_{10}(x)\psi_{01}(y)\ 。 \tag{4-24}$$

这时正规方程的解是

$$\left\{\begin{aligned}
b_{00}&=\bar{z}=\frac{\sum\limits_i\sum\limits_j z_{ij}}{nm}\\
b_{10}&=\frac{\sum\limits_i\sum\limits_j \psi_{10}(x_i)z_{ij}}{\sum\limits_i\sum\limits_j \psi_{10}^2(x_i)}=\frac{\sum\limits_i\sum\limits_j \psi_{10}(x_i)z_{ij}}{m\sum\limits_i \psi_{10}^2(x_i)}\\
b_{20}&=\frac{\sum\limits_i\sum\limits_j \psi_{20}(x_i)z_{ij}}{m\psi_{20}^2(x_i)}\\
b_{01}&=\frac{\sum\limits_i\sum\limits_j \psi_{01}(y_j)z_{ij}}{n\sum\limits_j \psi_{01}^2(y_j)}\\
b_{02}&=\frac{\sum\limits_i\sum\limits_j \psi_{02}(y_j)z_{ij}}{n\sum\limits_j \psi_{02}^2(y_j)}\\
b_{11}&=\frac{\sum\limits_i\sum\limits_j \psi_{10}(x_i)\psi_{01}(y_j)z_{ij}}{\sum\limits_i\sum\limits_j[\psi_{10}(x_i)\psi_{01}(y_j)]^2}
\end{aligned}\right. 。 \tag{4-25}$$

对于更高次的正交多项式，可以用（4-23）递推公式进行计算。

解得回归方程系数后，同以前的方法一样即建立趋势面方程，并可进一步计算趋势值和剩余值等。

第六节　趋势分析在地球化学中的应用

一、两种趋势面模型

趋势面分析通常用来研究和解决两种不同类型的问题，一类问题是发现系统趋势，另

一类问题是区分局部异常。这两种应用的目的不同，所依据的假设也不同，但由于用的是同一种方法，就很容易将这一差别掩盖起来，因此必须作进一步的分析，将二者区别开来。首先我们注意到和这两类问题相应的有两种趋势面模型。

第一类问题的趋势面模型是

$$z_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + \varepsilon_i \quad 。 \tag{4-26}$$

这种模型的剩余部分只包括ε_i，ε_i是观测误差（包括取样误差，仪器误差，分析误差等）。通常误差ε_i在回归面附近是随机、正态分布，其均值为零，方差为常数，而这又说明各种误差相互独立。

此时，z_i的观测值落在回归面附近比较均匀的误差分布带之内（图4-5），指出剩余部分仅只反映误差项的大小，而并不反映局部异常。对于这种情形，可以用一个趋势面来恰当地拟合因变量z_i的观测值。

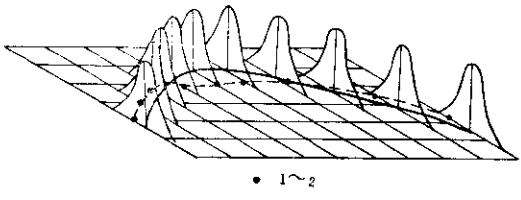


图4-5　第一种趋势面模型

1—实际观测点，2—多项回归曲线。

第二类问题的趋势面模型是

$$z_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + (\gamma_i + \varepsilon_i) \quad 。 \tag{4-27}$$

这种模型的剩余部分除误差项外，还包括局部异常，通常二者混合在一起，并且误差项ε_i往往比局部异常γ_i小得多。因此可以认为几乎全部剩余方差均是用趋势面所不能拟合的（图4-6）。尤其重要的是随机误差ε_i的均值为零，且对所有的z_i值都是相互独立的，而$(\gamma_i + \varepsilon_i)$虽然均值亦为零，但通常对于$z_i$值并不相互独立。正是在这一点上两种趋势面模型有着根本的区别，这一区别集中表现在剩余值之间是否存在自相关关系。

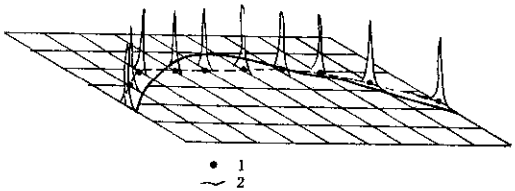


图4-6　第二种趋势面模型

1—实际观测点，2—多项回归曲线。

在第一种模型里，剩余值（随机误差）之间相互独立，并不存在自相关关系；但是在第二种模型里，相邻剩余值（主要是局部异常）之间往往存在一定程度的自相关。正是这

种自相关剩余指出了有地质意义的现象存在，特别是正的自相关剩余指示可能的矿致异常。因此趋势面分析的目的就是要圈定出一定大小的区域，在此区域内各项（$\gamma_i+\varepsilon_i$）相互相关，换句话说，趋势面分析的主要目的就是要寻找自相关剩余地段，局部异常有较强烈的影响。顺便指出，由于剩余不呈随机分布，因此不能用F比来检验回归效果的显著性，因为这时高的F比不是由于回归效果显著，而是由于局部剩余地段上的自相关所致。

二、建立地球化学变量的空间分布模型

对于第一类趋势面模型，可以用它作为某种地球化学变量在某一地区空间分布的数学模型（趋势方程——代数模型；趋势面等值线图——几何模型）。这种数学模型直观地表示了改变量空间分布的几何形态，并且在控制区范围内，只要给定任一空间坐标，就可以从趋势方程中求得估计值。

三、异常下限的确定

在第二类问题的趋势面模型中，把观测值的整个变差分解为区域趋势、局部异常和噪声三个组成部分。在实际中，正确区分这三个组成部分是一项极其困难的工作。下面只对已经确定趋势值的前提下如何从剩余值中区分出局部异常，也就是如何确定异常下限的问题作一初步分析。

确定异常下限有好几种方法，这里只介绍一种经验方法。设R_i代表正剩余值，M代表正剩余的总个数，N代表观测点的总数，则正剩余为局部异常A_i与噪声L_i之和，即

$$R_i = A_i + L_i \qquad (i=1,\ 2,\ \cdots,\ M)。$$

以经验值L_e来代替L_i：

$$L_e = \frac{\sum_{i=1}^{M} R_i}{K}, \tag{4-28}$$

其中$K=\frac{1}{3}N$（也可根据实际情况作N或$\frac{1}{2}N$），为一经验系数。

求出异常下限后，可按化探中的方法圈定异常。

四、其它应用

在计算机制图中，趋势面的方法被用来作为等值线图的插值计算。

在实际工作中，趋势分析的应用是多方面的。如用于平缓褶皱构造研究；元素或热液浓集中心、浓集趋势的研究；甚至有人（Degeoffroy，1968）在对威斯康辛州锌矿区地球化学数据处理中利用趋势分析来选择钻孔勘等网。

第五章　判别分析

第一节　概　　述

一、判别分析的概念

判别分析是判断样品归属的一种多元统计分析方法。它根据已知的两个或多个类别母体的多种特征建立起一个或多个判别函数，用它来判断未知样品可能属于已知类别中的哪一个母体。它是一种分类的数学方法，也是一种将某个样品的多个特征同它可能归属的各种类型进行对比，然后确定将这个样品分归可能性最大的一类母体中的方法。

和聚类分析不同，判别分析旨在把类型未知的样品根据其多种地球化学特征分别归入预先确定的各个组（例如矿化组或无矿组），在所有的方法中，不管所使用的判别标准（判别函数）的性质如何，首先必须有一个代表各个组的样品的“训练组”，每一类的训练样品应当尽可能的多，以便保证足够的统计代表性。理想的情况是，一旦确定了判别标准，就应当根据已确定的判别规则，对一个独立的样品“试验组”进行分类。

判别分析具有如下特点：

（1）它必须有已知类型的样本作为“训练组”。这是与聚类分析最重要的差别。

（2）已知类别（母体）与未知的样品共同有多种特征，即多个指标。

（3）已知类别母体之间的分布一般有相互重叠的部分。特别就某单个特征（指标）来看时，总有相互重叠的部分（图5-1）。因此，有一部分个体在划分类别时，可能出现错划现象。

（4）判别分析将一个多维的向量空间（多种特征）约简为一维的判别空间，从而大大地简化了空间维数，而不致损失划分不同母体的更多信息。

（5）根据中心极限定理，多个变量的线性函数比单变量有较大可能服从正态分布规律，因而判别函数比原始变量能更好地满足多元正态分布这一重要假设。

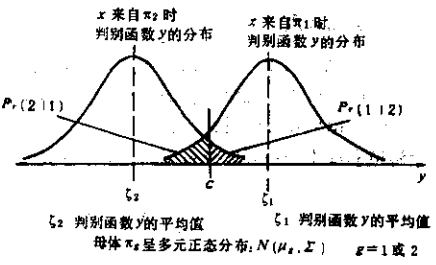


图5-1　母体分布与“错分损失”

二、判别分析的分类

判别分析按已知类别的数目，可以分为两类（或叫两级）判别和多类（或叫多级）判别；按建立判别函数时所用的准则分类，有费歇准则下的判别分析和贝叶斯准则下的判别分析等；按区分不同母体所用的数学模型，分为线性判别和非线性判别。

三、判别函数

若有两个母体，它们的p种特征（变量）分布状态可以用p维空间中两个椭球体的等密度轨迹表示。对于任一单个特征来说，判别或区分这两类母体的效果都不是最佳的。按某种准则，把这p种特征组合成一个综合指标，对于已知类别样品所建立的关系式，称为判别函数。如

$$y=f(x_1,\ x_2,\ \cdots,\ x_p)\ 。$$

这个判别函数所得的值y，称为判别计量（或判别得分）。

判别函数中最简单的形式是线性函数，即

$$y=c_1x_1+c_2x_2+\cdots+c_px_p=\sum_{j=1}^{p}c_j\cdot x_j\ 。 \tag{5-1}$$

通过将p个指标综合成判别函数后，两类母体的区分率达到最好（图5-2）。

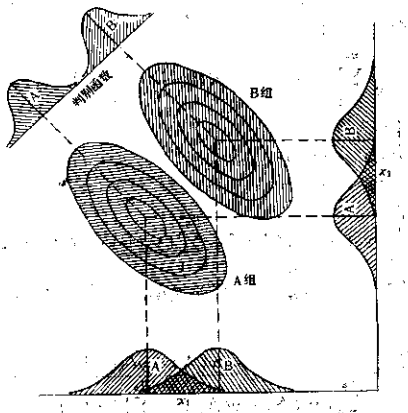


图5-2 判别分析几何解释

第二节 费歇准则下的两类线性判别

费歇准则是由英国数学家费歇（R. A. Fisher）在1936年首先提出来的。

属于不同母体的两类样品，每个样品都可以测定p种特征，它们在p维空间中可以看成是两个点群。这两个点群可能有重叠的部分，但不能完全重合。在一般情况下，两点群各自的离差也不一定相等。怎样来区分上述两个点群呢？费歇提出了如下准则：通过坐标原

点，在两个点群之间寻找一个最优分割面（其数学表达式就是判别函数），使得两个母体的样品点在该分割面上投影达到最大限度的分离，而且每类样品内部的离散最小(图5-3)。满足类间距离最大，而类内离差最小的准则，在统计学上称为费歇准则。

一、线性判别函数的建立

在p维空间，通过原点的最优分割面写成函数形式为

$$y = c_1x_1 + c_2x_2 + \cdots + c_px_p \ 。$$

所谓建立判别函数就是要确定上式中的系数c_j（$j=1, 2, \cdots, p$）。

设在已知的两类母体A、B中，分别抽取n_a和n_b个样品。每个样品都测试p个特征。以x_{ki}(A)表示A类母体中第k个样品的第i个变量的观测值；以x_{hi}（B）表示B母体中第h个样品第i个变量的观测值（$k=1, 2, \cdots, n_a$；$h=1, 2, \cdots, n_b$；$i=1, 2, \cdots, p$；$n=n_a+n_b$）。这n个样品点在分割面上有相应n个投影点，这些投影点的y坐标值分别用y_{ki}（A）和y_{hi}（B）来表示，它们的重心分别记为$\bar{y}$（A）和$\bar{y}$（B），则

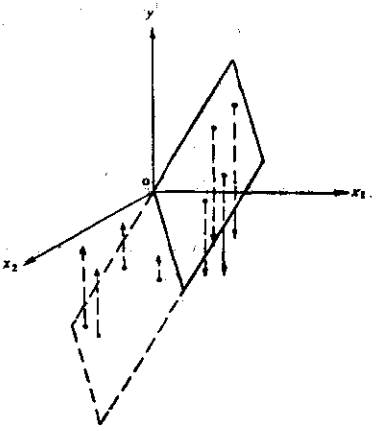


图5-3 费歇判别平面示意图

$$\bar{y}(A) = \frac{1}{n_a}\sum_{k=1}^{n_a} y_k \ (A) \ , \ \bar{y}(B) = \frac{1}{n_b}\sum_{h=1}^{n_b} y_h(B) \ 。$$

A和B两类母体的类间离差Q为

$$Q = [\bar{y}(A) - \bar{y}(B)]^2 \ 。$$

A和B类样品y的类内离差F为

$$F = \sum_{k=1}^{n_a}[y_k(A) - \bar{y}(A)]^2 + \sum_{h=1}^{n_b}[y_h(B) - \bar{y}(B)]^2 \ 。$$

按费歇准则，要使类间离差最大，类内离差最小，也就是它们之比$Q/F = I$为最大，写成数学表达式为

$$I = \frac{Q}{F} = \frac{[\bar{y}(A) - \bar{y}(B)]^2}{\sum_{k=1}^{n_a}[y_k(A) - \bar{y}(A)]^2 + \sum_{h=1}^{n_b}[y_h(B) - \bar{y}(B)]^2} \ 。 \qquad (5\text{-}2)$$

根据方程（5-1）式，y是c_i的函数，要使I达到极大，根据微积分的知识，分别求I对c_i（$j=1, 2, \cdots, p$）的偏导数，并令其为0，即

$$\frac{\partial I}{\partial c_i} = 0 \ 。 \qquad (5\text{-}3)$$

由（5-2）式，先两边取对数

$$\ln I = \ln Q - \ln F \ ,$$

然后求c_i的偏导数，并令其为0，

$$\frac{\partial \ln I}{\partial c_j} = \frac{\partial \ln Q}{\partial c_j} - \frac{\partial \ln F}{\partial c_j} = 0 \text{ 。} \tag{5-4}$$

由此可得：

$$\frac{1}{Q}\frac{\partial Q}{\partial c_j} = \frac{1}{F}\frac{\partial F}{\partial c_j} \text{ 。} \tag{5-5}$$

因为

$$\begin{aligned} Q &= [\bar{y}(\mathrm{A}) - \bar{y}(\mathrm{B})]^2 \\ &= \left[\sum_{j=1}^{p} c_j \bar{x}_j(\mathrm{A}) - \sum_{j=1}^{p} c_j \bar{x}_j(\mathrm{B})\right]^2 \\ &= \left[\sum_{j=1}^{p} c_j \left(\bar{x}_j(\mathrm{A}) - \bar{x}_j(\mathrm{B})\right)\right]^2 \text{ 。} \end{aligned}$$

记

$$d_j = \bar{x}_j(\mathrm{A}) - \bar{x}_j(\mathrm{B}) \text{ ,} \tag{5-6}$$

则

$$Q = \left[\sum_{j=1}^{p} c_j d_j\right]^2 \text{ 。} \tag{5-7}$$

而

$$\begin{aligned} F &= \sum_{k=1}^{n_a}\left[y_k(\mathrm{A}) - \bar{y}(\mathrm{A})\right]^2 + \sum_{k=1}^{n_b}\left[y_k(\mathrm{B}) - \bar{y}(\mathrm{B})\right]^2 \\ &= \sum_{k=1}^{n_a}\left[\sum_{j=1}^{p} c_j\left(x_{kj}(\mathrm{A}) - \bar{x}_j(\mathrm{A})\right)\right]^2 + \sum_{k=1}^{n_b}\left[\sum_{j=1}^{p} c_j\left(x_{kj}(\mathrm{B}) - \bar{x}_j(\mathrm{B})\right)\right]^2 \\ &= \sum_{k=1}^{n_a}\left[\sum_{j=1}^{p} c_j\left(x_{kj}(\mathrm{A}) - \bar{x}_j(\mathrm{A})\right) \cdot \sum_{i=1}^{p} c_i\left(x_{ki}(\mathrm{A}) - \bar{x}_i(\mathrm{A})\right)\right] \\ &\quad + \sum_{k=1}^{n_b}\left[\sum_{j=1}^{p} c_j\left(x_{kj}(\mathrm{B}) - \bar{x}_j(\mathrm{B})\right) \cdot \sum_{i=1}^{p} c_i\left(x_{ki}(\mathrm{B}) - \bar{x}_i(\mathrm{B})\right)\right] \\ &= \sum_{j=1}^{p}\sum_{i=1}^{p} c_j c_i\left[\sum_{k=1}^{n_a}\left(x_{kj}(\mathrm{A}) - \bar{x}_j(\mathrm{A})\right) \cdot \left(x_{ki}(\mathrm{A}) - \bar{x}_i(\mathrm{A})\right)\right. \\ &\quad \left. + \sum_{k=1}^{n_b}\left(x_{kj}(\mathrm{B}) - \bar{x}_j(\mathrm{B})\right) \cdot \left(x_{ki}(\mathrm{B}) - \bar{x}_i(\mathrm{B})\right)\right] \text{ ,} \end{aligned}$$

记

$$\begin{aligned} S_{ij} &= \sum_{k=1}^{n_a}\left(x_{ki}(\mathrm{A}) - \bar{x}_i(\mathrm{A})\right) \cdot \left(x_{kj}(\mathrm{A}) - \bar{x}_j(\mathrm{A})\right) \\ &\quad + \sum_{k=1}^{n_b}\left(x_{ki}(\mathrm{B}) - \bar{x}_i(\mathrm{B})\right) \cdot \left(x_{kj}(\mathrm{B}) - \bar{x}_j(\mathrm{B})\right) \\ &\qquad (i, j = 1, 2, \cdots, p) \text{ ,} \end{aligned} \tag{5-8}$$

则：

$$F = \sum_{i=1}^{p}\sum_{j=1}^{p} c_i c_j S_{ij} \text{ 。} \tag{5-9}$$

由（5-7）式得

$$\frac{\partial Q}{\partial c_j}=2(c_1d_1+c_2d_2+\cdots+c_pd_p)\cdot d_j=2\left(\sum_{i=1}^{p}c_id_i\right)\cdot d_j\ ,$$

由（5-9）式得

$$\frac{\partial F}{\partial c_j}=2(c_1S_{1j}+c_2S_{2j}+\cdots+c_pS_{pj})=2\sum_{i=1}^{p}c_iS_{ij}\ ,$$

则（5-5）式可写为

$$\frac{F}{Q}\left(\sum_{i=1}^{p}c_id_i\right)\cdot d_j=\sum_{i=1}^{p}c_iS_{ij}\ 。\qquad (5\text{-}10)$$

令

$$\frac{F}{Q}\left(\sum_{i=1}^{p}c_i\cdot d_i\right)=\beta\ 。\qquad (5\text{-}11)$$

（5-10）式可写为

$$\sum_{i=1}^{p}c_iS_{ij}=\beta d_j\ ,$$

即

$$c_1S_{1j}+c_2S_{2j}+\cdots+c_pS_{pj}=\beta\cdot d_j\qquad (j=1,\ 2,\ \cdots,\ p)\ 。\qquad (5\text{-}12)$$

（5-12）式是包含有c_1，c_2，…，c_p的p个未知数的线性方程组，式中β是常数因子，对线性方程组的解不起作用，因此可不考虑。于是上述线性方程组简化为：

$$\begin{cases}S_{11}c_1+S_{12}c_2+\cdots+S_{1p}c_p=\bar{x}_1(\mathrm{A})-\bar{x}_1(\mathrm{B})\\ S_{21}c_1+S_{22}c_2+\cdots+S_{2p}c_p=\bar{x}_2(\mathrm{A})-\bar{x}_2(\mathrm{B})\\ \cdots\cdots\cdots\cdots\\ S_{p1}c_1+S_{p2}c_2+\cdots+S_{pp}c_p=\bar{x}_p(\mathrm{A})-\bar{x}_p(\mathrm{B})\end{cases}\qquad (5\text{-}13)$$

其中

$$S_{ij}=\sum_{h=1}^{n_a}\left[x_{hi}(\mathrm{A})-\bar{x}_i(\mathrm{A})\right]\cdot\left[x_{hj}(\mathrm{A})-\bar{x}_j(\mathrm{A})\right]+\sum_{h=1}^{n_b}\left[x_{hi}(\mathrm{B})-\bar{x}_i(\mathrm{B})\right]\cdot\left[x_{hj}(\mathrm{B})-\bar{x}_j(\mathrm{B})\right]\ 。$$

解线性方程组（5-13）即可求出各判别系数c_1，c_2，…c_p，得到判别函数：

$$y=c_1x_1+c_2x_2+\cdots+c_px_p$$

二、未知样品的判别

用所建立的判别函数，分别求出A母体和B母体已知样品的判别计量平均值$\bar{y}(\mathrm{A})$和$\bar{y}(\mathrm{B})$：

$$\bar{y}(\mathrm{A})=\sum_{j=1}^{p}c_j\bar{x}_j(\mathrm{A})=c_1\bar{x}_1(\mathrm{A})+c_2\bar{x}_2(\mathrm{A})+\cdots+c_p\bar{x}_p(\mathrm{A})\ ,$$

$$\bar{y}(\mathrm{B})=\sum_{j=1}^{p}c_j\bar{x}_j(\mathrm{B})=c_1\bar{x}_1(\mathrm{B})+c_2\bar{x}_2(\mathrm{B})+\cdots+c_p\bar{x}_p(\mathrm{B})\ 。$$

由于$\bar{y}(\mathrm{A})$是n_a个样品的平均值，$\bar{y}(\mathrm{B})$是n_b个样品的平均值，取$\bar{y}(\mathrm{A})$和$\bar{y}(\mathrm{B})$的加权平均值作为判别A、B两类母体样品的临界值y。

$$y_0 = \frac{n_a \cdot \bar{y}(A) + n_b \cdot \bar{y}(B)}{n_a + n_b} \quad (5\text{-}14)$$

假定有$\bar{y}(A) > y_0 > \bar{y}(B)$，则把未知归属样品的$p$个变量测试值代入判别方程，求得其判别值$y_i$，凡是$y_i > y_0$的样品，判它属于A母体；凡是$y_i < y_0$的样品，判它属于B母体。

三、二类判别函数的显著性检验及判别正确率

根据p个变量所建立的判别函数，判别是否有效，判别结果好坏程度如何？第一个问题是要解决判别函数对区分两个母体是否显著；第二个问题是以正确判别率来衡量判别好坏程度。

1．判别函数的显著性检验

判别函数的显著性检验，可采F-检验法和马氏距离检验法。

（1）F-检验　判别分析的假设前提是假设两组样品取自两个不同的母体。两个母体差异性是否显著，可从类平均差异情况进行考查。在类平均值相等条件下，即检验假设

$$H_0: \bar{y}(A) = \bar{y}(B) ,$$

统计量

$$F = \frac{S_{回}/f_{回}}{S_{剩}/f_{剩}} \quad (5\text{-}15)$$

服从由自度$f_{回}$和$f_{剩}$的F-分布。式中

$$S_{回} = \sum_{k=1}^{n_a} \left(\bar{y}(A) - \bar{y}\right)^2 + \sum_{h=1}^{n_b} \left(\bar{y}(B) - \bar{y}\right)^2$$

$$= n_a \left(\bar{y}(A) - \bar{y}\right)^2 + n_b \left(\bar{y}(B) - \bar{y}\right)^2 , \quad (5\text{-}16)$$

$$\bar{y} = \frac{n_a \bar{y}(A) + n_b \bar{y}(B)}{n_a + n_b} ,$$

$$f_{回} = 2 - 1 = 1 ,$$

$$S_{剩} = \sum_{k=1}^{n_a} \left(y_k(A) - \bar{y}(A)\right)^2 + \sum_{h=1}^{n_b} \left(y_h(B) - \bar{y}(B)\right)^2 , \quad (5\text{-}17)$$

$$f_{剩} = n_a + n_b - 1 - 1 .$$

在给定显著性水平α情况下，查F分布表，得$F_\alpha(1, n-2)$。如果$F > F_\alpha$，否定原假设H_0，说明两组间差别显著，判别有效；如果$F \leqslant F_\alpha$，则接受原假设H_0，说明两组间差异不显著，判别函数是无效的。

（2）马氏距离检验，检验两个母体间差异性是否显著，还可通过检验母体中各个变量的类平均值之差

$$d_j = \bar{x}_j(A) - \bar{x}_j(B) \qquad (j = 1, 2, \cdots, p)$$

是否足够大，即两类多元平均值在统计上是否存在显著差异。这种显著性检验，一般采用马氏距离（D^2）为基础构成统计量

$$F = \frac{n_a \cdot n_b \cdot (n_a + n_b - p - 1)}{(n_a + n_b) \cdot (n_a + n_b - 2) \cdot p} \cdot D^2 , \quad (5\text{-}18)$$

其中

$$D^2=(n_a+n_b-2)\cdot\sum_{j=1}^{p}c_jd_j\ 。\qquad(5\text{-}19)$$

在假设$\bar{x}_i(A)=\bar{x}_i(B)$的情况下，$F$服从自由度$f_1=p$，$f_2=(n_a+n_b-p-1)$的$F$分布。同理，在给定显著性水平$\alpha_F$，查$F$-分布表，得临界值$F_\alpha(p,\ n_a+n_b-p-1)$。若$F>F_\alpha$，说明两个母体多元均值间差异显著，判别方程有效；否则，判别方程无效。

2．判别正确率

判别函数经过检验为有效之后，再用已知样品的变量数据代入判别函数，回判每个已知样品的判别值，重新判别它的归类。根据判别正确率（或错判率），来衡量判别效果的好坏程度。

判别正确率是指原观测属于A类或B类的样品，根据判别函数仍判为A类或B类的样品占全部参加判别分析计算样品总数的百分比。若原来从A母体中抽取n_a个样品，经过判别，有m_a个判为A类；原来从B母体中抽取n_b个样品，经过判别有m_b个判为B类，则判别正确率为

$$r=\frac{m_a+m_b}{n_a+n_b}\times100\%\ ,\qquad(5\text{-}20)$$

错判率为

$$e=\frac{n_a+n_b-m_a-m_b}{n_a+n_b}\times100\%=1-r\ 。\qquad(5\text{-}21)$$

四、变量的选择

在判别分析中，如果选取变量太少，会由于信息量不足而导致判别效果差；但选取变量太多，还会由于变量之间的相互干扰而降低判别效果。为了获得最佳判别函数，应设法在不影响判别效果的前提下，挑选那些分辨母体能力较强的变量参加建立判别函数。

1．根据马氏距离的贡献挑选变量

每个变量两母体间的距离占总距离的百分比，作为各变量的贡献Dx_i（$j=1,\ 2,\ \cdots,\ p$），

$$\begin{aligned}Dx_j&=c_jd_j\Big/\sum_{j=1}^{p}c_jd_j\times100\%\\&=c_j\cdot\Big(\bar{x}_j(A)-\bar{x}_j(B)\Big)\Big/\Big[\sum_{j=1}^{p}c_j\Big(\bar{x}_j(A)-\bar{x}_j(B)\Big)\Big]\times100\%\ 。\end{aligned}\qquad(5\text{-}22)$$

在计算过程中，可斟酌取舍，挑取对马氏距离贡献大的变量参加建立判别函数。

2．I值最大法挑选变量

由费歇准则，得

$$I=\frac{Q}{F}=\frac{\Big[\bar{y}(A)-\bar{y}(B)\Big]^2}{\sum_{k=1}^{n_a}\Big[y_k(A)-\bar{y}(A)\Big]^2+\sum_{k=1}^{n_b}\Big[y_k(B)-\bar{y}(B)\Big]^2}$$

为最大。若仅考虑一个变量x_i时，相应地亦可计算出每个变量的I_i值（$j=1,\ 2,\ \cdots,\ p$）。

$$I_j=\frac{\left[\bar{x}_i(\mathrm{A})-\bar{x}_j(\mathrm{B})\right]^2}{\sum_{k=1}^{n_a}\left[x_{kj}(\mathrm{A})-\bar{x}_j(\mathrm{A})\right]^2+\sum_{k=1}^{n_b}\left[x_{kj}(\mathrm{B})-\bar{x}_j(\mathrm{B})\right]^2}。\qquad(5\text{-}23)$$

显然，I_j越大的变量，分辨母体的能力越强。在具体工作中，可根据具体情况选择I_j大于某个数值的变量参加建立判别函数。

第三节　两类判别计算实例

例一

某地区经化探工作圈出了一批异常，其中7个为已知见矿异常，另7个为已知无矿异常。分别测定了Cu、Ag、Bi三元素的平均值（其平均含量对数值见表5-1），试作判别分析。

表5-1　某地14个异常含量对数值

编号	异常性质	x_1 ln (Cu)	x_2 ln (100×Ag)	x_3 ln (Bi)
1		2.58	0.90	0.95
2	见	2.90	1.23	1.00
3	矿	3.55	1.15	1.00
4	异	2.35	1.15	0.79
5	常	3.54	1.85	0.79
6	(A)	2.70	2.23	1.30
7		2.70	1.70	0.48
8		2.25	1.98	1.06
9	无	2.16	1.80	1.06
10	矿	2.33	1.74	1.10
11	异	1.96	1.48	1.04
12	常	1.94	1.40	1.00
13	(B)	3.00	1.30	1.00
14		2.78	1.20	1.48

1．计算两类样品各变量的平均值

$$\bar{x}_1(\mathrm{A})=\frac{1}{n_a}\sum_{k=1}^{n_a}x_{k1}(\mathrm{A})=2.902,$$

$\bar{x}_2(\mathrm{A})=1.458$，$\bar{x}_3(\mathrm{A})=0.9014$，

$\bar{x}_1(\mathrm{B})=2.346$，$\bar{x}_2(\mathrm{B})=1.628$，$\bar{x}_3(\mathrm{B})=1.1057$。

2．计算d_j

$$d_j=\bar{x}_j(\mathrm{A})-\bar{x}_j(\mathrm{B}),$$

$d_1=0.556$，$d_2=-0.170$，$d_3=-0.205$。

3．计算平方和，交叉乘积和矩阵S

$$S_{ij}=S_{ji}=\sum_{k=1}^{n_a}\left[x_{ki}(\mathrm{A})-\bar{x}_i(\mathrm{A})\right]\left[x_{kj}(\mathrm{A})-\bar{x}_j(\mathrm{A})\right]+\sum_{k=1}^{n_b}\left[x_{ki}(\mathrm{B})-\bar{x}_i(\mathrm{B})\right]\left[x_{kj}(\mathrm{B})-\bar{x}_j(\mathrm{B})\right]$$

得

$$S_{11}=2.28 \quad S_{12}=0.11 \quad S_{13}=0.24$$
$$S_{22}=1.69 \quad S_{23}=0.20$$
$$S_{33}=0.54$$

列出线性方程组

$$\begin{cases} 2.28c_1+0.11c_2+0.24c_3=0.556 \\ 0.11c_1+1.69c_2+0.20c_3=-0.17 \\ 0.24c_1+0.20c_2+0.54c_3=-0.205 \end{cases}。$$

4．解上述方程组得

$$c_1=0.298,\ c_2=-0.062,\ c_3=-0.489。$$

5．列出判别函数

$$y=0.298x_1-0.062x_2-0.489x_3。$$

6．计算判别临界值y_0

$$\bar{y}(A)=c_1\bar{x}_1(A)+c_2\bar{x}_2(A)+c_3\bar{x}_3(A)=0.33,$$
$$\bar{y}(B)=c_1\bar{x}_1(B)+c_2\bar{x}_2(B)+c_3\bar{x}_3(B)=0.06,$$
$$y_0=\frac{n_a\cdot\bar{y}(A)+n_b\cdot\bar{y}(B)}{n_a+n_b}=0.2。$$
$$\bar{y}(A)>y_0>\bar{y}(B)。$$

7．判别显著性检验

$$D^2=(n_a+n_b-2)\cdot\sum_{j=1}^{p}c_jd_j=3.312。$$
$$F=\frac{n_an_b}{(n_a+n_b)(n_a+n_b-2)}\cdot\frac{n_a+n_b-p-1}{p}\cdot D^2=3.22,$$
$$f_1=p=3,\ f_2=n_a+n_b-p-1=10,$$

查F-分布表（取$\alpha=0.10$和$\alpha=0.05$），

$$F_{0.10}=2.73,\ F_{0.05}=3.71。$$

由于$2.73<3.22<3.71$，认为只在$\alpha=0.10$的水平上判别效果显著。

8．对未知异常含矿性判别

有一未知异常，其三个变量平均值对数为：

$$\ln(Cu)=2.945,\ \ln(100\times Ag)=2.147,\ \ln(Bi)=1.537。$$

将这组数据代入判别函数，求得判别值

$$y=0.296\times2.945-0.062\times2.147-0.487\times1.537=-0.07。$$

因为计算的y值（-0.07）小于临界值（0.2），所以认为该异常属于无矿异常。

例二

该例是福建省某地选用两类判别模型进行判别分析的典型实例（据张宋业，1981），原始数据见表5-2。

1．变量的挑选

（1）计算各变量类内平均值和类内离差平方和，由公式

表5-2 福建某地两类判别原始数据表

类别	变量 / 样品	Al_2O_3 (x_1)	MnO (x_2)	TiO_2 (x_3)	SO_3 (x_4)	P_2O_5 (x_5)	Ag (x_6)	Fe_2O_3 (x_7)	CaO (x_8)	MgO (x_9)
A类	1	22.46	0.12	0.66	3.70	0.93	18.40	10.24	10.11	2.52
	2	19.37	0.15	0.80	0.09	0.10	12.35	24.38	0.11	6.69
	3	25.01	0.13	0.98	0.17	1.70	10.91	15.31	2.86	4.28
	4	26.03	0.12	0.74	1.79	1.28	17.10	12.28	9.50	3.67
	5	25.00	0.12	1.00	0.29	1.44	10.37	15.42	3.86	3.65
	6	28.94	0.07	1.08	0.05	1.46	31.62	11.59	2.06	3.86
	7	21.06	0.07	0.66	0.02	2.74	17.70	19.18	2.86	4.04
	8	18.88	0.09	0.56	1.60	1.16	15.93	12.03	12.03	1.45
	9	17.84	0.06	0.86	0.17	0.17	24.14	12.64	0.45	4.68
	10	22.09	0.16	0.87	2.67	0.33	17.18	12.38	4.06	3.27
	11	21.07	0.32	0.41	1.59	0.08	24.26	8.44	7.71	2.55
	12	31.23	0.04	1.92	0.03	0.05	20.54	8.92	0.12	3.38
	13	27.76	0.06	0.61	0.88	0.20	14.08	6.03	1.55	2.23
B类	1	6.83	0.18	0.32	9.10	0.13	34.15	34.72	8.83	2.80
	2	12.00	0.17	1.38	2.96	0.22	12.21	6.28	4.35	2.22
	3	13.56	0.28	0.56	1.74	0.04	12.91	12.48	2.15	5.06
	4	8.04	0.43	0.53	3.77	0.13	15.12	10.24	5.74	3.55
	5	8.30	1.21	0.50	5.82	0.10	13.61	7.30	32.72	0.56
	6	9.02	0.23	0.46	0.14	0.13	28.15	10.00	0.17	2.73
	7	10.76	0.52	0.73	4.48	0.16	15.74	16.07	13.27	3.28
	8	12.45	0.38	0.72	4.04	0.28	8.69	19.56	4.89	5.19
	9	9.82	0.42	0.94	3.76	0.21	11.25	13.07	4.43	3.13
	10	18.42	0.20	0.65	0.06	0.38	22.98	16.97	0.81	4.11
	11	9.15	0.22	0.44	0.84	0.13	18.14	23.08	0.97	3.77
	12	6.73	0.20	2.40	1.30	0.16	17.04	8.75	2.10	1.38
	13	12.55	0.23	0.46	0.15	0.24	18.84	12.19	0.34	4.36
	14	9.25	0.26	0.65	1.82	0.21	12.02	20.86	2.54	4.37
	15	10.16	0.16	0.80	0.18	0.24	10.98	12.48	0.48	2.77
	16	10.44	0.44	0.53	2.45	0.11	15.24	17.50	2.82	3.79

注：表中数据均为百分含量。

$$\bar{x}_i(\mathrm{A}) = \frac{1}{n_a}\sum_{k=1}^{n_a} x_{ki}(\mathrm{A}) \ , \quad \bar{x}_i(\mathrm{B}) = \frac{1}{n_b}\sum_{k=1}^{n_b} x_{ki}(\mathrm{B}) \ ,$$

$$F_{ii} = \sum_{k=1}^{n_a}\left(x_{ki}(\mathrm{A}) - \bar{x}_i(\mathrm{A})\right)^2 + \sum_{k=1}^{n_b}\left(x_{ki}(\mathrm{B}) - \bar{x}_i(\mathrm{B})\right)^2$$

计算得表5-3的结果。

表5-3 类内平均值与类内离差平方和

变量	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9
$\bar{x}_i$ (A)	23.5954	0.1162	0.8500	1.0038	0.8954	18.0446	13.0524	4.4062	3.5592
$\bar{x}_i$ (B)	10.4675	0.3450	0.6294	2.6856	0.1813	16.6888	15.1025	5.4431	3.3481
F_{ii}	330.7982	1.0436	2.6241	109.8975	8.3251	1010.9985	1943.2930	1174.8752	206.3089

由
$$I_i=\frac{\left(\bar{x}_i(\mathrm{A})-\bar{x}_i(\mathrm{B})\right)^2}{\sum_{k=1}^{n_a}\left(x_{ki}(\mathrm{A})-\bar{x}_i(\mathrm{A})\right)^2+\sum_{k=1}^{n_b}\left(x_{ki}(\mathrm{B})-\bar{x}_i(\mathrm{B})\right)^2}$$
计算得各变量I值：

$I_1=0.5299$，$I_2=0.0502$，$I_3=0.0186$，

$I_4=0.0257$，$I_5=0.0612$，$I_6=0.0017$，

$I_7=0.0047$，$I_8=0.0009$，$I_9=0.0002$。

（2）选取变量，选取$I_i>0.01$的变量，即I_1，I_2，I_3，I_4，I_5所代表的变量参加判别函数，该5个变量已占总I值的99.32%，其余4个变量只占0.68%，故可忽略不计。

2．建立判别函数

（1）计算所选取5个变量的组内方差与协方差，得线性方程组：
$$\begin{pmatrix}12.2518&-0.1325&0.5878&-2.5160&0.1164\\-0.1325&0.0387&-0.0094&0.1661&-0.0104\\0.5878&-0.0094&0.0972&-0.0981&-0.0155\\-2.5160&0.1661&-0.0981&4.0702&-0.1109\\0.1164&-0.0104&-0.0155&-0.1109&0.3083\end{pmatrix}\cdot\begin{pmatrix}c_1\\c_2\\c_3\\c_4\\c_5\end{pmatrix}=\begin{pmatrix}13.1279\\-0.2288\\0.2206\\-1.6818\\0.7141\end{pmatrix}。$$
（2）解上述方程组得

$c_1=1.3986$，$c_2=-4.4145$，$c_3=-5.8293$，

$c_4=0.5319$，$c_5=1.5390$。

则判别函数为

$$y=1.3986x_1-4.4145x_2-5.8293x_3+0.5319x_4+1.5390x_5。$$

3．判别函数显著性检验

（1）计算马氏距离，由公式得

$D^2=18.2893$

（2）计算F统计量
$$F=\frac{13\times16\times(13+16-1-5)}{(13+16)\times(13+13-2)\times5}\times18.2893=22.35。$$
给定$\alpha=0.05$，查F-分布表得临界值$F_{0.05(5,23)}=2.64$，由于$F=22.35>2.64$（$F_{0.05(5,23)}$），故认为判别函数用以分辨母体是显著的。

4．判别归类

（1）计算类平均值和判别临界值
$$\bar{y}(\mathrm{A})=\sum_{j=1}^{p}c_j\cdot\bar{x}_j(\mathrm{A})=29.4446，$$
$$\bar{y}(\mathrm{B})=\sum_{j=1}^{p}c_j\cdot\bar{x}_j(\mathrm{B})=11.1554，$$
$$y_0=\frac{n_a\cdot\bar{y}(\mathrm{A})+n_b\cdot\bar{y}(\mathrm{B})}{n_a+n_b}=19.3540。$$
（2）对已知样品进行回判验证

将已知分类的29个样品的5个变量代入判别函数，求得判别值y_i，若$y_i>y_0$(19.3540,)则判为A类；若$y_i<y_0$（19.3540），则判为B类。其结果见表5-4。

表5-4 已知样品判别归类表

类别	样品	判别值 (y)	判归	类别	样品	判别值 (y)	判归
A	1	30.4344	A	B	1	11.9340	B
	2	21.9645	A		2	9.0009	B
	3	31.3983	A		3	15.4494	B
	4	34.4844	A		4	8.4618	B
	5	35.5482	A		5	6.8015	B
	6	36.1449	A		6	3.1935	B
	7	29.5245	A		7	11.3184	B
	8	25.3800	A		8	14.2074	B
	9	20.0259	A		9	8.7237	B
	10	27.0450	A		10	21.8592	A
	11	26.4690	A		11	9.6099	B
	12	32.4080	A		12	17.0424	B
	13	35.7804	A		13	14.3046	B
					14	9.2907	B
					15	9.3042	B
					16	11.0889	B

注：分界值$y_0=19.3540$。

由表5-4中可以看出，29个样品中有28个与已知分类一致，只有一个样品错分，计算其判别正确率为

$$r=\frac{28}{29}\times 100\%=96.55\%\ 。$$

因此，判别效果是好的。

（3）对未知样品判别分类

有4个未知类别的待判样品，将其5个变量值代入判别函数，判得其中20，21号样品属于B类，而30，24号样品属于A类。上述结论也得到了其它地质方法的验证。

表5-5 未知样品的原始数据及判别结果

样号 \ 变量	Al_2O_3 (x_1)	MnO (x_2)	TiO_2 (x_3)	SO_3 (x_4)	P_2O_5 (x_5)	判别值 (y)	判别归类
20	9.71	0.25	0.50	3.16	0.16	11.4385	B
30	25.64	0.05	1.54	0.05	0.19	26.2315	A
24	18.41	0.06	0.84	0.02	0.14	20.8116	A
21	14.64	0.20	0.98	3.28	0.19	15.9381	B

注：分界值$y_0=19.3540$。

第四节 贝叶斯准则下的多类线性判别

两类线性判别分析是利用一个判别函数将样品划分为两类，从而确定未知样品的归属。有时研究的个体可能来自多个母体，如G个母体（$G>2$），这时确定未知样品归属的数学

方法称为多类判别分析。

多类判别问题，可按费歇准则的思路，用逐步两级判别分析方法来解决，即在每两类间建立一个判别函数，若是三类判别，须建立3个判别函数；若是四类判别，就需要建立6个判别函数；若是n类，则须建立C_n^2个判别函数。这样在实际应用中，计算工作量太大，很不方便。因此，对多类判别常采用贝叶斯（Bayes）准则的判别模型。这种方法在进行G类判别时，只须建立G个判别函数，然后计算未知样品出现在各类的概率，以最大概率进行归类。费歇准则是为p维空间两个点群寻找最优分割面，而贝叶斯准则是为p维空间G个点群寻找最优空间划分方法。

一、贝叶斯准则

若从G个母体（A_1，A_2，…，A_G）中取得n个样品，每个样品必然属于这G个母体中的一个母体（A_g）。如果每个样品测得p个变量（x_1，x_2，…，x_p），可把每个样品看作p维空间｛R｝中的一个点。n个样品组成一个p维样本空间｛R｝。

如果设法将样品空间划分为G个两两互斥的子空间：｛R_1｝，｛R_2｝，…，｛R_G｝，并分别与G个母体（A_1，A_2，…，A_G）对应。那么，对于一个未知样品x（x_1，x_2，…，x_p）同样也作为p维空间中的一点，看它落在哪个子空间的概率最大，就可把它划归那个母体。

对空间｛R｝，是根据已知G类样品划分的，由于G个子样的分布往往有重叠，所以对空间｛R｝划分时，任何一种划分方法都可能存在错分现象，错分就会造成损失。若把原属于母体A_g的样品错分到A_h母体中去，则造成的损失记为：$L(h|g)$，并约定

当$h=g$时，$L(h|g)=0$；

当$h\neq g$时，$L(h|g)>0$。

若把原属于Ag的样品错分到A_h的概率记为：$p\{h|g\}$，则当G个母体的概率分布密度$f_g(x)$已知时，有

$$P(h|g)=\int_{R_h} f_g(\mathbf{x})d\mathbf{x}\ 。 \tag{5-24}$$

把原属于A_g的样品错分到A_h母体中去所造成的平均损失为

$$W_h=\sum_{\substack{g=1\\g\neq h}}^{G} L(h|g)P\{h|g\}=\sum_{\substack{g=1\\g\neq h}}^{G} L(h|g)\int_{R_h} f_g(\mathbf{x})d\mathbf{x}\ 。 \tag{5-25}$$

在各母体的先验概率q_h已知时，G类母体总的错分平均损失为

$$W_R=\sum_{h=1}^{G} q_h W_h=\sum_{h=1}^{G} q_h\sum_{\substack{g=1\\g\neq h}}^{G} L(h|g)P(h|g)\ 。 \tag{5-26}$$

如果把原来不属于A_g的样品错分到A_g类中所造成的损失记为：$L(g|h)$，错分概率记为：$P(g|h)$。当母体的概率分布密度已知时，有

$$P(g|h)=\int_{R_g} f_h(\mathbf{x})d\mathbf{x}\ , \tag{5-27}$$

则把原不属A_g的样品错划入A_g类所造成的平均损失为

$$W_g=\sum_{\substack{h=1\\h\neq g}}^{G} L(g|h)P(g|h)=\sum_{\substack{h=1\\h\neq g}}^{G} L(g|h)\int_{R_g} f_h(\mathbf{x})d\mathbf{x}\ 。 \tag{5-28}$$

同样，当母体的先验概率q_g已知时，G类母体错分的总平均损失为

$$\dot{W}_R=\sum_{g=1}^{G}q_g\sum_{\substack{h=1\\h\neq g}}^{G}L(g|h)\int_{R_g}f_h(\mathbf{x})d\mathbf{x}\ 。 \tag{5-29}$$

显然，人们希望划分空间，使总的错分损失（W_R）最小。

如图5-4所示，面积Ⅰ表示原属A_g的个体错误地划入A_h母体的错分概率$P(h|g)$，面积Ⅱ表示原来不属于A_g的个体错误地划入A_g母体的错分概率$P(g|h)$。面积Ⅰ、Ⅱ又表示在x发生条件下，实际上属于A_g和A_h的后验概率，它们分别为$P(A_g|\mathbf{x})$和$P(A_h|\mathbf{x})$。由贝叶斯概率公式可知：

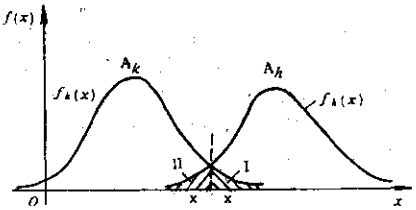


图5-4 错分概率示意图

$$P(h|g)=P(A_g|\mathbf{x})=\frac{q_gP(\mathbf{x}|A_g)}{\sum_{i=1}^{G}q_iP(\mathbf{x}|A_i)}\ , \tag{5-30}$$

$$P(g|h)=P(A_h|\mathbf{x})=\frac{q_hP(\mathbf{x}|A_h)}{\sum_{i=1}^{G}q_iP(\mathbf{x}|A_i)}\ 。 \tag{5-31}$$

由（5-30）和（5-31）式说明，无论x的分布如何，如果$P(\mathbf{x}|A_g)$和$P(\mathbf{x}|A_h)$条件概率已知，则平均错分的概率即可求得。当x为多元正态分布时，即$N(\mathbf{a}_i,\ \Sigma)$（其中，$\mathbf{a}_i$为母体的数学期望或平均值向量；Σ为协方差矩阵），上式中的条件概率$P(\mathbf{x}|A_i)$可用母体的概率密度$f_i(\mathbf{x})$来代替，因而（5-30）、（5-31）式可改写为

$$P(A_g|\mathbf{x})=\frac{q_gf_g(\mathbf{x})}{\sum_{i=1}^{G}q_if_i(\mathbf{x})} \tag{5-32}$$

$$P(A_h|\mathbf{x})=\frac{q_hf_h(\mathbf{x})}{\sum_{i=1}^{G}q_if_i(\mathbf{x})} \tag{5-33}$$

（5-25）式和（5-28）式可改写为

$$W_h=\sum_{\substack{h=1\\h\neq g}}^{G}L(g|h)\frac{q_hf_h(\mathbf{x})}{\sum_{i=1}^{G}q_if_i(\mathbf{x})}\ , \tag{5-34}$$

$$W_g=\sum_{\substack{h=1\\h\neq g}}^{G}L(h|g)\frac{q_gf_g(\mathbf{x})}{\sum_{i=1}^{G}q_if_i(\mathbf{x})}\ 。 \tag{5-35}$$

贝叶斯已证明了，要使总错分平均损失（W_R）为最小，则划分空间{R}的方法应满足

以下不等式：

$$\{R_g\}:\ \sum_{\substack{g=1\\ g\neq h}}^{G} L(h|g)\,q_g f_g(x) > \sum_{\substack{h=1\\ h\neq g}}^{G} L(g|h)\,q_h f_h(x)\ , \qquad (5\text{-}36)$$

即对任一研究个体$x=(x_1, x_2, \cdots, x_p)$，若把它划归$A_g$母体所造成的平均损失小于把它划归其它母体所造成的平均损失时，则将它划归A_g母体。因此，将在给定母体先验概率q_g时，划分空间$\{R\}$使错分平均损失最小的原则，称为贝叶斯准则，（贝叶斯准则的另一种说法是，划分空间$\{R\}$，使得观测样品中，原属于A_g母体的样品x_g，落在$\{R_g\}$子空间的概率$P(A_g|x)$最大）。按照这一准则划分空间$\{R\}$的结果，称为关于q_g的贝叶斯解。

在实际问题中，错分损失不易准确计量，要精确给出能反映客观实际的$L(h|g)$值是困难的。此时，往往假定

$$L(h|g)=L(g|h) \qquad (h\neq g)\ ,$$

则贝叶斯准则（5-36）可写成

$$\{R_g\}:\ \sum_{\substack{g=1\\ g\neq h}}^{G} q_g f_g(x) > \sum_{\substack{h=1\\ h\neq g}}^{G} q_h f_h(x)\ ,$$

即在给定母体先验概率q_g时，划分空间$\{R\}$使原来为A_g母体的样品，落在$\{R_g\}$子空间的概率比落在任何其它子空间$\{R_h\}$（$h\neq g$）的概率都大。

将上式两边同时减$\sum_{i=1}^{G} q_i f_i(x)$，可得

$$\{R\}:\ q_h f_h(x) < q_g f_g(x)\ ,$$

将上式两边再同除以$\sum_{i=1}^{G} q_i f_i(x)$，可得

$$\{R\}:\ \frac{q_h f_h(x)}{\sum_{i=1}^{G} q_i f_i(x)} < \frac{q_g f_g(x)}{\sum_{i=1}^{G} q_i f_i(x)}\ 。 \qquad (5\text{-}37)$$

（5-37）式两端就是个体$x=(x_1, x_2, \cdots, x_p)$落在母体$A_h$和$A_g$的后验概率。所以，贝叶斯准则又称为以指定具有较高后验概率的母体作为判别个体x归属的准则。也就是说，把属于A_g母体的后验概率最大的个体，归属A_g母体。由（5-37）式可知，后验概率最大与$q_g f_g(x)$最大是等价的。因此

$$q_g f_g(x) \qquad (g=1, 2, \cdots, G)\ , \qquad (5\text{-}38)$$

称为赖以对任一个体x进行判别归属的判别函数。

二、多类判别函数的建立

设有G个母体A_g（$g=1, 2, \cdots, G$），服从正态分布$N(a_g, \Sigma)(g=1, 2, \cdots, G)$，其概率密度函数为

$$f_g(x)=(2\pi)^{-\frac{p}{2}}\left|\Sigma^{-1}\right|^{\frac{1}{2}}\exp\left[-\frac{1}{2}(x-a_g)'\Sigma^{-1}(x-a_g)\right] \qquad (5\text{-}39)$$

其中，$\mathbf{x}=[x_1, x_2, \cdots, x_p]'$；参数$\mathbf{a}_g$和$\Sigma$分别为母体$A_g$的均值向量和协方差矩阵，即

$$\mathbf{a}_g=[a_{1g}, a_{2g}, \cdots, a_{pg}]',$$

$$\Sigma=(\sigma_{ij})=\begin{pmatrix}\sigma_{11} & \sigma_{12} & \cdots & \sigma_{1p}\\ \sigma_{21} & \sigma_{22} & \cdots & \sigma_{2p}\\ \vdots & \vdots & & \vdots\\ \sigma_{p1} & \sigma_{p2} & \cdots & \sigma_{pp}\end{pmatrix}.$$

Σ是对称阵；Σ^{-1}是Σ的逆矩阵；$|\Sigma^{-1}|$为Σ^{-1}的行列式。

当$L(h|g)=L(g|h)$ $(h\neq g)$，且母体先验概率q_g和母体参数已知，母体协方差矩阵Σ无显著差异（统计相等）时，将（5-39）式代入（5-38）式，可得G个判别函数，为

$$q_g f_g(\mathbf{x})=q_g(2\pi)^{-\frac{p}{2}}\left|\Sigma^{-1}\right|^{\frac{1}{2}}\exp\left[-\frac{1}{2}(\mathbf{x}-\mathbf{a}_g)'\Sigma^{-1}(\mathbf{x}-\mathbf{a}_g)\right]$$

$$(g=1, 2, \cdots, G)。 \qquad (5\text{-}40)$$

两边取对数，经整理得：

$$\ln q_g f_g(\mathbf{x})=\ln\left[(2\pi)^{-\frac{p}{2}}\left|\Sigma^{-1}\right|^{\frac{1}{2}}\right]-\frac{1}{2}\mathbf{x}'\Sigma^{-1}\mathbf{x}+\ln q_g$$

$$-\frac{1}{2}\mathbf{a}_g'\Sigma^{-1}\mathbf{a}_g+\mathbf{a}_g'\Sigma^{-1}\mathbf{x}, \qquad (5\text{-}41)$$

式中第一、二项与g无关，可视为常数，故令

$$b=\ln\left[(2\pi)^{-\frac{p}{2}}\left|\Sigma^{-1}\right|^{\frac{1}{2}}\right]-\frac{1}{2}\mathbf{x}'\Sigma^{-1}\mathbf{x},$$

$$c_g=\mathbf{a}_g'\Sigma^{-1}=[c_{1g}, c_{2g}, \cdots, c_{pg}],$$

$$c_{0g}=-\frac{1}{2}a_g'\Sigma^{-1}\mathbf{a}_g,$$

则

$$\ln q_g f_g(\mathbf{x})=b+\ln q_g+c_{0g}+c_g\mathbf{x} \quad (g=1, 2, \cdots, G)。 \qquad (5\text{-}42)$$

由于b与g无关，不影响G个函数之间的比较，故可将b略去，令

$$y_g(\mathbf{x})=\ln q_g+c_{0g}+c_g\mathbf{x},$$

则（5-42）式可改写为

$$\ln q_g f_g(\mathbf{x})=b+y_g(\mathbf{x}) \qquad (g=1, 2, \cdots, G)。 \qquad (5\text{-}43)$$

显然，对$g=1, 2, \cdots, G$来说，当$y_g(\mathbf{x})$最大时，$\ln q_g f_g(\mathbf{x})$亦达到最大，故称

$$y_g(\mathbf{x})=\ln q_g+c_{0g}+c_g\mathbf{x}$$

$$=\ln q_g+c_{0g}+c_{1g}x_1+c_{2g}x_2+\cdots+c_{pg}x_p,$$

$$(g=1, 2, \cdots, G) \qquad (5\text{-}44)$$

为贝叶斯准则下G个正态母体的多类线性判别函数。其中$\ln q_g+c_{0g}$为x的零次项系数（常数项）；c_{ig} $(j=1, 2, \cdots, p)$为x的非零次项系数。

由$c_g=\mathbf{a}_g'\Sigma^{-1}$展开得

$$c^j{}_g=\sum_{k=1}^{p}\mathbf{a}_{kg}G^{kj} \qquad (5\text{-}45)$$

$$c_{0g}=-\frac{1}{2}\mathbf{a}_g'\Sigma^{-1}\mathbf{a}_g=-\frac{1}{2}c_g\mathbf{a}_g=-\frac{1}{2}\sum_{j=1}^{p}c_{jg}a_{jg}$$

$$(g=1, 2, \cdots, G), \qquad (5\text{-}46)$$

式中，σ^{kj}为逆矩阵Σ^{-1}中第k行第j列元素；a_{kj}，a_{ig}为A_g母体中第k、j个变量的平均值。当母体参数确定后，c_{ig}和c_{0g}就完全确定；而如果母体参数未知，则须用容量足够大的样本对母体参数进行估计。用估计值代替理论值，求得近似解。

假设收集了G组（类）样品（每组样品个数可不相同），若第g组（类）有n_g个样品，则共有n（$n=\sum_{g=1}^{G}n_g$）个样品，每个样品取p个变量，那末第g类，第i个样品，第j个变量的观测值为x_{gij}（$g=1, 2, \cdots, G$；$i=1, 2, \cdots, n_g$；$j=1, 2, \cdots, p$）。此外，还有一些未知类别的待判样品（第h类）。现将所有样品观测值排成如下形式：

组（类）1　$x_{111}, x_{112}, \cdots, x_{11p}$

$x_{121}, x_{122}, \cdots, x_{12p}$

$\cdots \quad \cdots \quad \cdots$

$x_{1n_11}, x_{1n_12}, \cdots, x_{1n_1p}$

组（类）2　$x_{211}, x_{212}, \cdots, x_{21p}$

$x_{221}, x_{222}, \cdots, x_{22p}$

$\cdots \quad \cdots \quad \cdots$

$x_{2n_21}, x_{2n_22}, \cdots, x_{2n_2p}$

$\cdots \quad \cdots \quad \cdots$

组（类）G　$x_{G11}, x_{G12}, \cdots, x_{G1p}$

$x_{G21}, x_{G22}, \cdots, x_{G2p}$

$\cdots \quad \cdots \quad \cdots$

$x_{Gn_G1}, x_{Gn_G2}, \cdots, x_{Gn_Gp}$

组（类）h　$x_{h11}, x_{h12}, \cdots, x_{h1p}$

$x_{h21}, x_{h22}, \cdots, x_{h2p}$

$\cdots \quad \cdots \quad \cdots$

$x_{hn_h1}, x_{hn_h2}, \cdots, x_{hn_hp}$

首先根据已知的G个母体的样本观测值，对母体参数（a_g和Σ）进行估计，以便建立判别函数。

估计均值向量（$\bar{x}_g$）

$$\bar{x}_g=[\bar{x}_{g1}, \bar{x}_{g2}, \cdots, \bar{x}_{gp}]',$$

其中

$$\bar{x}_{gj}=\frac{1}{n_g}\sum_{i=1}^{n_g}x_{gij} \quad (j=1, 2, \cdots, p;\ g=1, 2, \cdots, G)。 \tag{5-47}$$

计算样本协方差矩阵S，是用样本变量离差交叉乘积和除以其自由度作为母体方差、协方差的无偏估计。令组内离差交叉和矩阵为W，

$$\mathbf{W}=\begin{pmatrix} w_{11} & w_{12} & \cdots & w_{1p} \\ w_{21} & w_{22} & \cdots & w_{2p} \\ \vdots & \vdots & & \vdots \\ w_{p1} & w_{p2} & \cdots & w_{pp} \end{pmatrix},$$

$$\mathbf{S}=\begin{pmatrix} \mathbf{S}_{11} & \mathbf{S}_{12} & \cdots & \mathbf{S}_{1p} \\ \mathbf{S}_{21} & \mathbf{S}_{22} & \cdots & \mathbf{S}_{2p} \\ \vdots & \vdots & & \vdots \\ \mathbf{S}_{p1} & \mathbf{S}_{p2} & \cdots & \mathbf{S}_{pp} \end{pmatrix}.$$

其中

$$w_{hj}=\sum_{g=1}^{G}\sum_{i=1}^{n_g}(x_{gih}-\bar{x}_{gh})(x_{gij}-\bar{x}_{gj}),\quad (h,j=1,2,\cdots,p) \tag{5-48}$$

$$s_{hj}=\frac{w_{hj}}{n-G},\quad n=\sum_{g=1}^{G}n_g,$$

s_{hj}是母体理论方差与协方差σ_{hj}的无偏估计。S是母体理论方差协方差矩阵Σ的无偏估计。

为使用方便，把S的逆矩阵$[\mathbf{S}^{-1}]$的元素记为：s^{hj} $(h, j=1, 2, \cdots, p)$。G个母体的先验概率可用各母体的样品出现的频率近似地代替，即

$$q_g=\frac{n_g}{n}\qquad (g=1, 2, \cdots, G)。 \tag{5-49}$$

求出上述参数的估计值后，判别系数和常数项则可用公式（5-45），（5-46）求得

$$c_{jg}=\sum_{h=1}^{p}\bar{x}_{gh}s^{hj}\qquad (j=1, 2, \cdots, p;\ g=1, 2, \cdots, G), \tag{5-50}$$

$$c_{0g}=-\frac{1}{2}\sum_{j=1}^{p}c_{jg}\bar{x}_{gj}\qquad (g=1, 2, \cdots, G)。 \tag{5-51}$$

G类母体的判别函数即可建立

$$y_g(\mathbf{x})=\ln q_g+c_{0g}+c_{1g}x_1+c_{2g}x_2+\cdots+G_{pg}x_p$$
$$(g=1, 2, \cdots, G)。 \tag{5-52}$$

为了对某一个体进行分类，就可将该个体的一组x值代入（5-52）式的G个判别函数式中，得到该个体的G个线性判别值$y_g(\mathbf{x})$。如果

$$y_g^*(\mathbf{x})=\max_{1\leqslant g\leqslant G}\{y_g(\mathbf{x})\},$$

则把该个体划归A_g*母体。

三、后验概率计算

贝叶斯准则是后验概率最大准则。任一未知个体x属于多元正态母体A_g的后验概率为

$$P(\mathrm{A}_g|\mathbf{x})=\frac{q_g f_g(\mathbf{x})}{\sum_{i=1}^{G}q_i f_i(\mathbf{x})} \tag{5-53}$$

当线性判别函数$y_g(\mathbf{x})$算出之后，后验概率的计算公式可进一步简化。由于

$$\ln q_g f_g(\mathbf{x}) = b + y_g(\mathbf{x}) ,$$

则

$$q_g f_g(\mathbf{x}) = e^{y_g(\mathbf{x})+b} = e^b e^{y_g(\mathbf{x})} ,$$

故

$$P(A_g|\mathbf{x}) = \frac{e^b e^{y_g(\mathbf{x})}}{e^b \sum_{i=1}^{G} e^{y_i(\mathbf{x})}} = \frac{e^{y_g(\mathbf{x})}}{\sum_{i=1}^{G} e^{y_i(\mathbf{x})}} 。 \tag{5-54}$$

式（5-54）较式（5-53）简单，但计算时，会由于$y_g(\mathbf{x})$值过大或过小而产生溢出。因此，进行极差变换，改写为如下形式

$$P(A_g|\mathbf{x}) = \frac{e^{y'_g(\mathbf{x})}}{\sum_{i=1}^{G} e^{y'_i(\mathbf{x})}} , \tag{5-55}$$

式中

$$y'_g(\mathbf{x}) = y_g(\mathbf{x}) - \max_{i \le g \le G} \{y_g(\mathbf{x})\} 。$$

在判别了未知样品的归属后，可用（5-55）式计算某样品属于A_g母体的后验概率。后验概率指出了样品归属于某母体可能性的大小。在实际工作中，还可以把次大的后验概率作为判别样品的参考，特别是在最大后验概率接近0.5的时候。

四、检验对多个母体的判别效果

判别分析的前提是母体间客观上存在着差别，但判别函数是根据样本建立的，如果母体间不存在差别（实为一个母体），提出判别就毫无意义；如果母体之间客观上存在着差别，但从母体中抽取出的样品变量能否反映G个母体之间的差别，事先并不完全了解，因此判别函数建立起来之后，它能否区分母体，即用它判别某个个体归属是否有效，仍须进行统计假设检验。对两个母体的差异性假设检验，曾介绍了马氏距离检验法。对多个母体的判别效果的检验，多运用威尔克斯（Wilks）准则（这是单变量方差分析在多元情况下的推广）。

设样品数据为x_{gij}（$g=1, 2, \cdots G$; $i=1, 2, \cdots, n_g$; $j=1, 2, \cdots, p$），它们来自G个具有相同协方差矩阵的p维正态母体N（a_g, Σ），并定义：组内离差矩阵$\mathbf{W}=[w_{kj}]$；组间离差矩阵$\mathbf{B}=[b_{kj}]$；总的离差矩阵$\mathbf{T}=\mathbf{W}+\mathbf{B}=[t_{kj}]$。其中

$$w_{kj} = \sum_{g=1}^{G} \sum_{i=1}^{n_g} (x_{gik} - \bar{x}_{gk})(x_{gij} - \bar{x}_{gj}) , \tag{5-56}$$

$$b_{kj} = \sum_{g=1}^{G} n_g (\bar{x}_{gk} - \bar{x}_k)(\bar{x}_{gj} - \bar{x}_j) , \tag{5-57}$$

$$t_{kj} = w_{kj} + b_{kj} = \sum_{g=1}^{G} \sum_{i=1}^{n_g} (x_{gik} - \bar{x}_k)(x_{gij} - \bar{x}_j) \tag{5-58}$$

$$(k, j = 1, 2, \cdots, p) 。$$

式中

$$\bar{x}_{gj}=\frac{1}{n_g}\sum_{i=1}^{n_g}x_{gij}\qquad (j=1,\ 2,\ \cdots,\ p)\ ,$$

$$\bar{x}_j=\frac{1}{n}\sum_{g=1}^{G}\sum_{i=1}^{n_g}x_{gij}\ ,\ n=\sum_{g=1}^{G}n_g\ 。$$

用威尔克斯准则

$$U=\frac{|\mathbf{W}|}{|\mathbf{T}|}=\frac{|\mathbf{W}|}{|\mathbf{W}+\mathbf{B}|}\tag{5-59}$$

来检验假设H_0：$a_1=a_2=\cdots=a_p$，即检验p个变量对于区分G个母体的能力。显然，U（两个行列式之比）的值愈小，说明母体间差别愈大，愈有利于G个母体的区分；U值愈大，母体间差别愈小，则母体不易区分，因此，可把U作为判别能力的一种度量。由于随机变量U的分布函数计算困难，在实际计算中，常用其它分布函数来近似地代替，如C.R.Rao的近似式，即

$$F\cong\left(\frac{1-U^{\frac{1}{a}}}{U^{\frac{1}{a}}}\right)\left(\frac{la+2\lambda}{2z}\right)\tag{5-60}$$

近似地服从自由度为$2z$，$la+2\lambda$的F分布，式中

$$a=\begin{cases}\sqrt{(p^2\cdot(G-1)^2-4)/(p^2+(G-1)^2-5)} & \text{当}p^2+(G-1)^2\neq 5\ ,\\ 1 & \text{当}p^2+(G-1)^2=5\ ,\end{cases}$$

$$l=n-1-\frac{1}{2}(p+G)\ ,$$

$$\lambda=\frac{1}{4}\left[2-p\cdot(G-1)\right]\ ,$$

$$z=\frac{1}{2}p\cdot(G-1)\ 。$$

假设在单变量x_i情况下的两类判别，则

$$U=U_i=\frac{w_{ii}}{t_{ii}}\ 。$$

（5-60）式变为

$$F=\left(\frac{t_{ii}-w_{ii}}{w_{ii}}\right)\left(\frac{n-G}{G-1}\right)\ ,$$

即组间均方与组内均方之比，这就是单因素方差分析中的F统计量。

当$p=1$，2（G任意）或$G=2$，3（p任意）时，式（5-60）就成为一个精确表达式。

第四节　逐步判别分析

一、逐步判别分析的基本思想

在判别分析中，若表征母体的变量很多，且每个变量对母体的分辨能力不同，如果这些变量都拿来参加建立判别函数，不但计算工作量大，而且还可能由于变量的不独立性而

影响判别效果。因此，在建立判别函数之前，挑选那些对分辨母体能力较强的变量参加建立判别函数就显得十分必要。逐步判别分析是完成这一工作的好方法。在两类判别分析中，曾介绍过用I值法和用对马氏距离贡献大小来挑选变量，这两种方法都是孤立地对每个变量进行考察，而忽略了变量之间的相互关系。

逐步判别分析和逐步回归分析的基本思想相似，都是根据每一个变量在各类（组）判别式中所起判别作用的重要性来挑选判别效果最好的变量（即分辨母体能力最强的变量）进入判别式；同时从判别式中剔除那些由于新变量的引入而失去判别作用的变量，使最后的判别式中，只保留对母体判别能力较强的变量。

逐步判别分析的作法是：

（1）首先从一个变量开始，比较各个变量的判别能力，把其中分辨母体能力最强的一个变量挑选出来，并检验其对母体分辨能力是否显著。如果显著，就将其引入判别函数。

（2）再把未选入的每一个变量同已选入变量结合在一起，比较每两个变量结合之后分辨母体的能力，从中选择分辨能力最大的变量，并检验其对母体分辨能力是否显著。如果显著，再把这个变量选入判别式。

（3）当后一个变量选入后，考察原先已引入到判别式中的变量，是否由于新变量的引入而判别能力下降。如果由于新变量的引入，使得原先已在判别函数中的某变量的判别效果不再显著，就将其从判别式中剔除；若因新变量的引入，其判别效果仍显著，则可考虑引进另一个变量。如此反复进行，直至既没有变量可引入，又没有变量可剔为止。最后，利用选入的变量建立判别函数。

这种通过逐步计算变量的判别能力，并逐步进行显著性检验，自动选择变量，最后获得最优判别函数的方法，称为逐步判别分析。其判别的数学模型，仍是在贝叶斯准则下的多类线性判别模型。

二、变量的判别能力及变量取舍的标准

1．变量的综合判别能力

在讨论判别效果检验时，曾用威尔克斯统计量

$$U=\frac{|\mathbf{W}|}{|\mathbf{T}|}$$

来检验多个变量结合在一起对G个母体的分辨能力是否显著的问题。其中，U反映了多个变量的综合判别能力。

假设收集了G类样品，第g类有n_g个样品，当只有一个变量时，其组内离差平方和矩阵与总离差平方和矩阵均为一阶矩阵，只有一个元素，即

$$\mathbf{W}_1=w_{11}=\sum_{g=1}^{G}\sum_{i=1}^{n_g}(x_{gi}-\bar{x}_g)^2\ ,$$

$$\mathbf{T}_1=t_{11}=\sum_{g=1}^{G}\sum_{i=1}^{n_g}(x_{gi}-\bar{x})^2\ ,$$

则

$$U_1=\frac{|\mathbf{W}_1|}{|\mathbf{T}_1|} \tag{5-61}$$

可表示某变量x的判别能力。

当有两个变量（如x_1，x_2）时，其组内离差矩阵与总离差矩阵均为二阶矩阵，即

$$\mathbf{W}_2=\begin{pmatrix} w_{11} & w_{12} \\ w_{21} & w_{22} \end{pmatrix},\quad \mathbf{T}_2=\begin{pmatrix} t_{11} & t_{12} \\ t_{21} & t_{22} \end{pmatrix},$$

其中

$$w_{kj}=\sum_{g=1}^{G}\sum_{i=1}^{n_g}(x_{gik}-\bar{x}_{gk})(x_{gij}-\bar{x}_{gj}),$$

$$t_{kj}=\sum_{g=1}^{G}\sum_{i=1}^{n_g}(x_{gik}-\bar{x}_{k})(x_{gij}-\bar{x}_{k}),$$

则

$$U_{1,2}=\frac{|\mathbf{W}_2|}{|\mathbf{T}_2|} \tag{5-62}$$

反映了两个变量x_1，x_2组合在一起分辨母体的能力，即$U_{1,2}$值越小，则x_1，x_2联合的判别能力越强。

当有L个变量（如x_1，x_2，…，x_L）时，其组内离差矩阵与总离差矩阵为L阶矩阵，即

$$\mathbf{W}_L=\begin{pmatrix} w_{11} & w_{12} & \cdots & w_{1L} \\ w_{21} & w_{22} & \cdots & w_{2L} \\ \vdots & \vdots & & \vdots \\ w_{L1} & w_{L2} & \cdots & w_{LL} \end{pmatrix},$$

$$\mathbf{T}_L=\begin{pmatrix} t_{11} & t_{12} & \cdots & t_{1L} \\ t_{21} & t_{22} & \cdots & t_{2L} \\ \vdots & \vdots & & \vdots \\ t_{L1} & t_{L2} & \cdots & t_{LL} \end{pmatrix},$$

其中

$$w_{kj}=\sum_{g=1}^{G}\sum_{i=1}^{n_g}(x_{gik}-\bar{x}_{gk})(x_{gij}-\bar{x}_{gj}),$$

$$t_{kj}=\sum_{g=1}^{G}\sum_{i=1}^{n_g}(x_{gik}-\bar{x}_{k})(x_{gij}-\bar{x}_{j}),\qquad (k,\ j=1,\ 2,\ \cdots,\ L),$$

则

$$U_{1,2,\cdots,L}=\frac{|\mathbf{W}_L|}{|\mathbf{T}_L|} \tag{5-63}$$

反映L个变量组合在一起分辨母体的能力。同样$U_{1,2,\cdots,L}$值越小，其判别能力越强。

2．未选入变量的判别能力及引进变量的标准

（1）未被选进变量的判别能力

假设变量x_1，x_2，…x_L已经给定，然后再添加一个新变量x_r（$r>L$），现讨论变量x_r的判别能力。

把（$L+1$）个变量分为两组：第一组是前 L 个已给定的变量；第二组仅有 x_r 一个变量，那 $L+1$ 个变量的组内离差矩阵为

$$\mathbf{W}_{L+1}=\left(\begin{array}{cccc:c} w_{11} & w_{12} & \cdots & w_{1L} & w_{1(L+1)} \\ w_{21} & w_{22} & \cdots & w_{2L} & w_{2(L+1)} \\ \vdots & \vdots & & \vdots & \vdots \\ w_{L1} & w_{L2} & \cdots & w_{LL} & w_{L(L+1)} \\ \hdashline w_{(L+1)1} & w_{(L+1)2} & \cdots & w_{(L+1)L} & w_{(L+1)(L+1)} \end{array}\right),$$

写成分块矩阵形式

$$\mathbf{W}_{L+1}=\begin{pmatrix} \mathbf{W}_{11} & \mathbf{W}_{12} \\ \mathbf{W}_{21} & \mathbf{W}_{rr} \end{pmatrix},$$

其中

$$\mathbf{W}_{11}=\begin{pmatrix} w_{11} & w_{12} & \cdots & w_{1L} \\ w_{21} & w_{22} & \cdots & w_{2L} \\ \vdots & \vdots & & \vdots \\ w_{L1} & w_{L2} & \cdots & w_{LL} \end{pmatrix},$$

$$\mathbf{W}'_{12}=\mathbf{W}_{21}=(w_{1(L+1)}w_{2(L+1)}\cdots w_{L(L+1)}),$$

$$\mathbf{W}_{rr}=w_{rr}=w_{(L+1)(L+1)}。$$

将行列式进行初等变换，可得

$$|\mathbf{W}_{L+1}|=|\mathbf{W}_{11}|\begin{vmatrix} \mathbf{I} & \mathbf{W}_{11}^{-1}\mathbf{W}_{12} \\ 0 & \mathbf{W}_{rr}-\mathbf{W}_{21}\mathbf{W}_{11}^{-1}\mathbf{W}_{12} \end{vmatrix}$$

$$=|\mathbf{W}_{11}|\,|\mathbf{W}_{rr}-\mathbf{W}_{21}\mathbf{W}_{11}^{-1}\mathbf{W}_{12}|。$$

令

$$\mathbf{W}^{(l)}=|\mathbf{W}_{rr}-\mathbf{W}_{21}\mathbf{W}_{11}^{-1}\mathbf{W}_{12}|，\text{则}$$

$$|\mathbf{W}_{L+1}|=|\mathbf{W}_{11}|\cdot w^{(l)}。 \tag{5-64}$$

同理

$$|\mathbf{T}_{L+1}|=|\mathbf{T}_{11}|\cdot t^{(l)}。 \tag{5-65}$$

于是

$$\frac{|\mathbf{W}_{L+1}|}{|\mathbf{T}_{L+1}|}=\frac{|\mathbf{W}_{11}|}{|\mathbf{T}_{11}|}\cdot\frac{w_{rr}^{(l)}}{t_{rr}^{(l)}}， \tag{5-66}$$

记为

$$U_{1,2,\cdots,L,r}=U_{1,2,\cdots,L}\cdot U_r|_{1,2,\cdots,L}， \tag{5-67}$$

其中，$U_{1,2,\cdots,L,r}$表示（$L+1$）个变量$x_1,x_2,\cdots,x_L$和x_r的联合判别能力；$U_{1,2,\cdots,L}$表示前L个变量x_1，x_2，…，x_L的综合判别能力；$U_r|_{1,2,\cdots,L}$表示在给定前L个变量前提下，变量x_r的判别能力。显然，在给定$U_{1,2,\cdots,L}$时，$U_{1,2,\cdots,L,r}$值的大小与$U_r|_{1,2,\cdots,L}$有关。如果x_1，x_2，…，x_L代表已引入变量，则$U_r|_{1,2,\cdots,L}$表示尚未引入变量x_r在与x_1，x_2，…，x_L结合在一起时的判别能力。即尚未引入变量x_r的判别能力，简记为：$U_r|_{(L)}$。

$U_r|_{(L)}$也是一个威尔克斯统计量，它可以检验假设H_0：$a_1|_{(L)}=a_2|_{(L)}=\cdots=a_G|_{(L)}$，其中，$a_g|_{(L)}$（$g=1,2,\cdots,G$）是在给定$x_1,x_2,\cdots,x_L$条件第$g$个母体（$L+1$）元的条件分布均值向量。为避开$U$分布函数计算，可建立与$U_r|_{(L)}$等价的$F$统计量计算式

$$F_{(G-1,n-G-L)}=\frac{1-U_r|_{(L)}}{U_r|_{(L)}}\cdot\frac{n-G-L}{G-1}=\frac{t_{rr}^{(l)}-w_{rr}^{(l)}}{w_{rr}^{(l)}}\cdot\frac{n-G-L}{G-1}\quad。\tag{5-68}$$

来检验变量x_r的判别能力的显著性。

（2）引进变量的标准

假设已经计算了l步，并引入了某L个变量（包括：$l=0$；$L=0$），尚有$x_{L+1},x_{L+2},\cdots,x_{L+m}$个变量尚未引入，现要确立第$l+1$步再引入一个新变量$x_r$（$r=L+1,L+2,\cdots,L+m$）的标准。

在给定$x_1,x_2,\cdots,x_L$的条件下，每一个未选入的变量x_r的判别能力均可由

$$U_r|_{(L)}=\frac{w_{rr}^{(l)}}{t_{rr}^{(l)}}\tag{5-69}$$

给出。其中，必有一个最小者

$$U_{r^*|(L)}^{*}=\min_{L+1<r<L+m}\{U_r|_{(L)}\}\ ,$$

说明变量x^*的判别能力最强。用$U_{r^*|(L)}^{*}$等价的F计算式

$$F=\frac{1-U_{r^*|(L)}^{*}}{U_{r^*|(L)}^{*}}\cdot\frac{n-G-L}{G-1}=\frac{t_{rr}^{(l)*}-w_{rr}^{(l)*}}{w_{rr}^{(l)*}}\cdot\frac{n-G-L}{G-1}\tag{5-70}$$

进行显著性检验。当计算值$F>F_\alpha(G-1,n-G-L)$时，则认为变量x^*的判别能力显著，应该将它引入判别函数。（5-70）式中的统计量F，可称为“引入F”，记为F_3，作为引入变量的标准。

3．已选变量的判别能力及剔除变量的标准

（1）已选变量的判别能力

在逐步引入变量时，已经引入变量的判别能力可能会由于后来引入了新变量而下降。因此，每引入一个新变量后，须对已经引入的其它变量的判别能力重新估价，对某个由于新变量的引入而判别能力下降到判别效果不显著的变量则应及时从判别函数中剔除。

假设已经计算了l步，并引入了包括x_r在内的L个变量，现在要确定第$l+1$步剔除变量x_r的标准。为了方便，假设x_r是第l步引入的，即前$l-1$步引入了不包括x_r在内的$L-1$个变量。因此，该问题就转化为在给定前$L-1$个变量的条件下，检验第l步引入的变量x_r的判别能力。由（5-69）式可知：

$$U_{r|(L-1)}=\frac{w_{rr}^{(L-1)}}{t_{rr}^{(L-1)}}\ ,\tag{5-71}$$

对于任意的r（$r=1,2,\cdots,L$），$U_{r|(L-1)}$，表示已选入的变量x_r的判别能力。显然，$U_{r|(L-1)}$值愈大，x_r的判别能力愈弱。

（2）剔除变量的标准

若对第l步引入的L个变量，计算其判别能力$U_r|_{(L-1)}$（$1\leqslant r\leqslant L$）。其中必有一个

$$U^*_{r|(L-1)}=\max_{1\leqslant r\leqslant L}\{U_r|_{(L-1)}\}\ ,$$

说明变量x_r的判别能力最弱。运用与其等价的F统计量

$$F=\frac{1-U^*_{r|(L-1)}}{U^*_{r|(L-1)}}\cdot\frac{n-G-(L-1)}{G-1}=\frac{w^{(l)*}_{rr}-t^{(l)*}_{rr}}{t^{(l)*}_{rr}}\cdot\frac{n-G-L+1}{G-1}$$

(5-72)

检验其判别效果。如果计算值$F\leqslant F_{\alpha(G-1,\,n-G-(L-1))}$时，则表明在给定显著性水平下$x_r$的判别能力不显著，应将其从判别式中剔除。反之，表明x_r仍显著，x_r留在判别式中。这里的F可称为“剔除F”，用F_4记之。

由上述看出，在挑选变量过程中，每一步都要计算许多行列式的值；当无变量可剔除又无新变量可引入时，根据最终选入判别式的变量建立判别函数时，还须计算协方差矩阵（S）的逆矩阵（S^{-1}）。如果用一般方法，计算工作量很大，占用内存很多。为了逐步求得行列式值及逆矩阵，可用紧凑方案。

三、逐步判别的计算步骤

设原始数据x_{gij}（$g=1,\ 2,\ \cdots,\ G;\ i=1,\ 2,\ \cdots,\ n_g;\ j=1,\ 2,\ \cdots,\ p$），为第$g$类中第$i$个样品第$j$个变量的观测值。

1．计算类平均值$\bar{x}_{gj}$和总平均值$\bar{x}_j$

$$\bar{x}_{gj}=\frac{1}{n_g}\sum_{i=1}^{n_g}x_{gij}\qquad(g=1,\ 2,\ \cdots G;\ j=1,\ 2,\ \cdots,\ p)\ ,$$

$$\bar{x}_j=\frac{1}{n}\sum_{g=1}^{G}\sum_{i=1}^{n_g}x_{gij}\qquad(j=1,\ 2,\ \cdots,\ p)\ .$$

2．计算组内离差矩阵（W）和总离差矩阵（T）

$$\mathbf{W}=[w_{kj}],$$

$$w_{kj}=\sum_{g=1}^{G}\sum_{i=1}^{n_g}(x_{gik}-\bar{x}_{gk})(x_{gij}-\bar{x}_{gj})\ ,$$

$$\mathbf{T}=[t_{kj}]\ ,$$

$$t_{kj}=\sum_{g=1}^{G}\sum_{i=1}^{n_g}(x_{gik}-\bar{x}_k)(x_{gij}-\bar{x}_j)\ ,$$

$$(k,\ j=1,\ 2,\ \cdots,\ p)\ .$$

3．逐步计算，挑选变量

设已知计算了l步（包括$l=0$），判别函数中已引入了某L个变量，则第$l+1$的计算内容是：

（1）计算全部变量的判别能力

若x_r是未选变量，则

$$U_i=\frac{w_{ii}^{(l)}}{t_{ii}^{(l)}} \qquad (i=L+1,\ L+2,\ \cdots,\ p)\ ;$$

若x_i是已选变量，则

$$U_i=-\frac{t_{ij}^{(l)}}{w_{jj}^{(l)}} \qquad (j=1,\ 2,\ \cdots,\ L)\ 。$$

（2）挑选变量并作显著性检验

首先从已选变量中检查有无需要剔除的变量，寻找U_j最大者

$$U_k^*=\max_{j\in L}\{U_j\} \qquad (j\in L,\ \text{表示}x_j\text{是已选变量})$$

进行F检验

$$F_4=\frac{1-U_k^*}{U_k^*}\cdot\frac{n-G-L+1}{G-1}=\frac{w_{kk}^{(l)*}-t_{kk}^{(l)*}}{t_{kk}^{(l)*}}\cdot\frac{n-G-L+1}{G-1}\ 。$$

若$F_4\leqslant F_{\alpha(G-1,\,n-G-L+1)}$则把$x_j$从判别函数中剔除；若$F_4>F_{\alpha(G-1,\,n-G-L+1)}$，即没有变量可剔除，则考虑从未选变量中挑选判别能力显著的变量，假设有x_k满足

$$U_k^*=\min_{i\notin L}\{U_i\} \qquad (i\notin L,\ \text{表示}x_i\text{是未选变量})\ ,$$

计算F_3的值：

$$F_3=\frac{1-U_k^*}{U_k^*}\cdot\frac{n-G-L}{G-1}=\frac{t_{kk}^{(l)*}-w_{kk}^{(l)*}}{w_{kk}^{(l)*}}\cdot\frac{n-G-L}{G-1}\ ,$$

若$F_3>F_{\alpha(G-1,\,n-G-L)}$，则把变量$x_k$引入判别函数；若$F_3\leqslant F_{\alpha(G-1,\,n-G-L)}$，结束变量的引入。

（3）不论变量x_k是引入还是剔除，都要用求解求逆紧凑方案对**W**和**T**进行消去变换，计算出新的**W**和**T**，从中求得判别系数。第$l+1$步对第k列进行消去变换的公式是：

$$w_{ij}^{(l+1)}=\begin{cases}w_{kj}^{(l)}/w_{kk}^{(l)} & (i=k,\ j\neq k)\\ w_{ij}^{(l)}-w_{ik}^{(l)}\cdot w_{kj}^{(l)}/w_{kk}^{(l)} & (i\neq k,\ j\neq k)\\ 1/w_{kk}^{(l)} & (i=k,\ j=k)\\ -w_{ik}^{(l)}/w_{kk}^{(l)} & (i\neq k,\ j=k)\end{cases}\ 。 \tag{5-73}$$

$$t_{ij}^{(l+1)}=\begin{cases}t_{kj}^{(l)}/t_{kk}^{(l)} & (i=k,\ j\neq k)\\ t_{ij}^{(l)}-t_{ik}^{(l)}\cdot t_{kj}^{(l)}/t_{kk}^{(l)} & (i\neq k,\ j\neq k)\\ 1/t_{kk}^{(l)} & (i=k,\ j=k)\\ -t_{ik}^{(l)}/t_{kk}^{(l)} & (i\neq k,\ j=k)\end{cases}\ 。 \tag{5-74}$$

重复（1），（2），（3）步，直至被引入的变量，经检验都是显著的，未被引入的变量都是判别能力不显著的为止。并记下被引入的变量号（凡经奇数次变换的变量x_k）。

4．建立判别函数

假设共计算了$l+1$步，最终选入L个变量

（1）计算判别系数

$$c_{ig}=(n-G)\sum_{k=1}^{L}w_{ki}^{(l+1)}\cdot\bar{x}_{gk}\ ,$$

$$c_{0g}=-\frac{1}{2}\sum_{i=1}^{L}c_{ig}\cdot\bar{x}_{gi}\ .$$

（2）构成判别函数

$$y_g(x)=\ln q_g+c_{0g}+\sum_{i=1}^{L}c_{ig}\cdot x_i$$

$$=\ln q_g+c_{0g}+c_{1g}x_1+c_{2g}x_2+\cdots+c_{Lg}x_L\qquad(g=1,\ 2,\ \cdots,\ G)\ .$$

5．判别分类

（1）将每一个体的p的变量观测值代入判别函数，求得G个判别值$y_g(x)$，若

$$y_g^*(x)=\max_{1\leqslant g\leqslant G}\{y_g(x)\}\ ,$$

则将个体划归第g^*个母体。

（2）根据判别值计算每一个体属于各母体的后验概率，即

$$P(A_g|x)=\frac{e^{y_g(x)}}{\sum_{h=1}^{G}e^{y_h(x)}}=\frac{e^{y'_g(x)}}{\sum_{h=1}^{G}e^{y'_h(x)}}\ ,$$

式中

$$y'_h(x)=y_h(x)-y_g^*(x)\qquad(h=1,\ 2,\ \cdots,\ G)\ .$$

6．检验判别函数的判别效果

用（5-60）式计算F近似值，若$F>F_\alpha(2z,\ l\alpha+2\lambda)$，则判别效果显著；反之，判别效果不显著。

第六节　判别分析在地质上的应用

目前，判别分析在地质学各领域中均有应用。主要分以下几类。第一类是判断研究对象的归属。如判断化探异常的性质；判别岩体的类型；判断地体的构造环境；判断铁帽的含矿性等等。这一类用途最广。凡是在工作过程确定了某些已知类别的母体，来判别未知类别的个体属于已知母体中哪一个的问题，都可以用判别分析去解决。第二类问题是研究或选择对某些类别有分辨能力的变量。如选择区分不同成矿类型的指标；确定不同剥蚀深度的评价指标；研究不同构造环境的指示元素等。第三类是用于矿床统计预测。在工作区内选择一些训练单元，建立判别含矿性的数学模型，然后用于判断未知单元的含矿性。

下面举一实例来说明判别分析的应用。此例是我们对我国某变质超基性岩成矿规律地球化学研究的一部分。

一、矿区地质概况

矿区位于秦岭东西向构造带的南亚带与龙门山东北向构造带复合部位。矿区及其外围广布下古生界变质基性火山岩，构成本区的基底。其上零星覆盖着上古生代的泥盆、石炭系灰岩、炭质板岩和千枚岩等。区内岩浆活动十分频繁，超基性岩、基性岩和中、酸性岩浆岩均有出露，其分布明显受褶皱轴部及断裂的控制。北西西向、近东西向和北东向三组主要断裂的相互交接处即为含镍超基性岩体出露部位。

超基性岩体东西长约5km，宽0.4—1.4km，出露面积近6km²。岩体总的走向为285—290°，向北东倾斜，倾角一般50°左右。岩体被剥蚀的深度不大。目前认为，主岩体形成于海西早期或加里东期，而分枝岩体形成较晚。

矿区内超基性岩经多次强烈变质，原岩几无保存。现在主岩体北部主要由纤胶蛇纹岩组成，伴有滑石菱镁岩，中部为滑石菱镁岩和叶蛇纹岩过渡带，而南部则为滑镁岩和石英菱镁岩带。

根据变质岩石中残余结构以及岩石化学和岩石中微量元素的地球化学研究，推断原始岩相有纯橄榄岩、斜方辉橄岩、橄榄岩和橄榄辉石岩类。岩体经受了自变质作用和它变质作用。前者表现为蛇纹石化、透闪石化和绿泥石化；后者则为滑石化和滑石碳酸盐化。当酸性侵入体侵入时，在接触带上发生了硅化，铁白云石化等。

据勘探已发现主要镍矿体产于岩体中西部南侧。矿石主要分为两种类型：A组矿石组合为磁黄铁矿-镍黄铁矿型；B组矿石为黄铁矿-针镍矿-辉镍矿型。

逐步多类判别的目的是以岩体西部已知地段为基础，评价中部地区的含矿性。

二、工作方法

这次工作采取逐步三类判别，即对待判样品进行“矿”、“矿化”、“无矿”三类判别。

1．已知样品的选择

（1）样品的分配尽量照顾到空间分布的合理性，即尽量照顾到不同测线的不同钻孔。在一个钻孔中又尽量考虑到不同深度的间隔。以保证在已知母体中的随机性。

（2）按已知成矿地段和待判地段内各种主要岩性的样品比例数抽样，以保证由已知地段构成的判别函数适用于待判地段。

（3）每类样品数确定为50个，属大子样，以尽量使样本接近正态分布。

（4）在每个样品做25种元素检查的基础上，挑选了具有能指示成矿作用、成岩作用和变质作用的12个变量，它们是Ni、Co、Cu、Cr、Ti、B、Sn、Mo、Pb、Zn、S、As。

2．判别函数的计算

（1）F值的确定

变量判别能力的显著性与显著性水平α有关。为了使较多的变量参加到判别函数中去，进行F检验的显著性水平α不宜太小，取$\alpha=0.10$。相应F值为2.30。

（2）变量的筛选过程

在计算过程中，具体步骤如下：

第一步引进的变量是Ni，其F值为1189.75，因而是高度显著的。

第二步引进了Cu，其F值为13.42，仍然是显著的。

第三步引进的变量是S，其F值为4.53，还是显著的。

第四步引进的变量是Co，其F值为2.87，它在$\alpha=0.10$水平下显著。

其余的变量已不再显著，因而最终判别函数中只包括了这4个元素。从地球化学的角度看，我们的目的是要判别铜-镍硫化物矿床样品的含矿性，在数学上选择Ni、Cu、S、Co 4元素作为判别矿、矿化和无矿的指标无疑是正确的。

最后所建立的判别函数如下：

$$y_1(\text{矿}) = -77.05 + 0.027\text{Ni} + 0.016\text{Cu} + 0.004\text{S} + 0.001\text{Co}\ ,$$

$$y_2(\text{矿化}) = -25.01 + 0.016\text{Ni} + 0.006\text{Cu} - 0.006\text{S} + 0.001\text{Co}\ ,$$

$$y_3(\text{无矿}) = -1.69 + 0.004\text{Ni} + 0.003\text{Cu} - 0.003\text{S} + 0.003\text{Co}\ 。$$

3．判别结果

（1）将150个已知类别样品代入判别函数，进行判别归类，结果见表5-6

表5-6　已知样品的判别分类

	判为矿类样品数	判为矿化样品数	判为无矿样品数
50个已知矿类样品中	50	0	0
50个已知矿化样品中	2	48	0
50个已知无矿样品中			50

由表5-6可见，参加建立判别函数的样品，典型性是好的，所得判别函数的分辨能力也是强的，分辨正确率达为98.7%。

（2）为了便于对中部待判地段与西部已知成矿地段进行评价对比，我们利用所得判别函数，搜集了两批新的样品进行分类。一批来自西部已知成矿地段693个样品；一批来自中部评价地段的750个样品。判别分类结果见表5-7。

表5-7　西部和中部地段样品判别结果

划入类别 / 采样地段	矿类		矿化类		矿和矿化类		非矿类	
	样品数	百分比	样品数	百分比	样品数	百分比	样品数	百分比
西部地段（693个）	228	33%	123	18%	351	51%	343	49%
中部地段（750个）	107	14%	233	31%	340	45%	410	55%

（3）判别结果分析

将西部已知成矿地段与中部评价地段样品的判别结果进行对比，两个地段划归矿和矿化类的百分比相差不多（分别为51%和45%），但分别划入矿类和矿化类的百分比却有明显的差异。已知成矿地段样品划归矿类占33%，划归矿化类的仅占18%；而评价地段样品划归矿类则仅有14%，划归矿化类的则占31%。以上事实指出评价地段的矿化特征与已知成矿的西部地段相差不多，但矿化的富集程度有显著的差异，即已知成矿地段矿化的富集程度明显高于评价地段。

第七节　判别分析中应注意的问题

数学家们常常从数学方法本身来讨论某种方法的效果。在实际应用的时候却必须十分

重视应用条件。

判别分析在应用时须注意以下问题：

1．建立判别函数时，要求所解决问题的地质条件要清楚，所利用的地质资料要详细、准确、可靠。这是应用判别分析方法时要注意的一个前提条件。

2．要求构成判别函数的标准样本尽量接近待判母体。具体地说应考虑到标准类别数是否齐全、符合实际？如果标准类别数和待判母体数不符，就不可避免地会出现错分现象。标准类别数确定以后，还要看各个标准类别是否典型，是否具有充分的代表性？显然，如果由实际样品所构成的各个标准类别不够典型，不能充分代表待判母体，也不能得到良好的判别效果。

3．有时有人希望用一次判别分析来区分很多的类别。但应注意，类别数增大，往往会降低正确分类的平均概率。因为类别愈多，则对于类别之间差异的分辨要求愈高，对于变量的判别能力要求愈严。如果不能满足这些条件，就会出现错误分类。

4．贝叶斯准则下的判别分析要求已知的标准样本服从多维正态分布，各组方差、协方差相等。为此，在条件允许时，应采取足够数目的样品，以满足母体分布的正态性前提。一般，标准样品的个数以50个左右（大子样）为宜。

5．恰当选取地质变量是应用效果好坏的关键问题之一。不同性质的变量可以混合使用，如将元素含量、颗粒大小、Eh值、pH值以及其它地质的、地球物理的、遥感的数据混合使用。一般，变量增多，可以提高正确分类的概率。但也要注意，所选的变量应与所研究的问题有较密切的联系，并且有较强的分辨力。最好在确定变量之前，先进行必要的探索性试验，并认真检查供应选入变量不被遗漏。当然为了节省人力、物力，也应尽量避免选变量的不必要重复，避免无判别能力的变量参加计算。

第六章　因子分析

第一节　概　述

因子分析方法最早是1904年由德国心理学家斯卑尔曼（C.Spearman)提出的，1957年由美国卡伦宾(W.C.Krumbein)引进地质学，当时主要用以研究沉积学，1962年由爱布尔(J.Imbrie)和珀迪(E.G.Purdy)加以完善和发展。近年来,因子分析在地质学中被广泛应用。

因子分析是用来研究一组变量的相关性，或用来研究相关矩阵（或协方差矩阵）内部结构的一种多元统计分析方法。它将多个变量综合成为少数的“因子”，也就是在较少损失原始数据信息的前提下，用少量的因子去代替原始的变量，从而达到对原始变量的分类，揭露原始变量之间的内在联系。

因子分析将从以下三个方面为地质工作中的成因推理提供重大帮助。

1．压缩原始数据　地质人员在研究每一个地质问题时都愿意获取尽可能多的数据，而在最终综合这些数据以形成地质成因概念时又为面对这大量复杂的、通常又是相互矛盾的数据而深感苦恼，简单地说，地质人员在搜集数据时总希望尽可能多，而在分析、综合数据时又希望尽可能少。因子分析恰恰提供了一条科学的、逻辑的途径，能把大量的原始数据大大精简，以利于地质人员进行综合分析。这种精简是以不影响主要地质结论的精确性为前提，或者说是在不损失地质成因信息的前提下进行的。

2．指示成因推理的方向　在形成成因结论的过程中，人们的思维和推理是最重要的一环。从大量复杂的地质数据中理出一个成因上的头绪并不是一件容易的事。不同的人对同一组地质数据，往往导出不同的成因结论，其原因是人们在推理过程中渗杂了主观、片面的意见。因子分析有可能把庞杂的原始数据按成因上的联系进行归纳、整理、精炼和分类，理出几条比较客观的成因线索，为地质人员提供逻辑推理的方向，帮助他们导出正确的成因结论。

3．分解叠加的地质过程　现在所看到的地质现象往往是多种成因过程叠加的产物，既有时间上不同过程的叠加，又有空间上不同过程的叠加，各个过程互相干扰，互相掩盖，造成了地质成因研究的复杂化。因子分析能提供从复合过程中弄清每个单一过程的性质和特征的途径。

因子分析在地质成因研究中，潜在的解决地质问题的能力是很大的，但是并非使用因子分析就能完全克服成因研究中的各种困难。

因子分析（广义）可以分为R型因子分析、Q型因子分析和R-Q型因子分析。R型因子分析用来研究变量之间的相关关系；Q型因子分析用来研究样品之间的相关关系；R-Q型因子分析用来研究变量与样品间的对应关系。

第二节　因子分析的数学模型

一、基本统计量

在对某一地质问题进行统计研究时，所研究的对象往往是一组具有某些共同属性的个体。约定以N表示个体的数目，以n表示对每一个体所测得的变量数，记号x_{ji}表示第j个变量在第i个个体上的观测值，则原始数据矩阵表示为：

$$(x_{ji})=\begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1N} \\ x_{21} & x_{22} & \cdots & x_{2N} \\ \vdots & \vdots & & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nN} \end{pmatrix}。$$

约定变量都是标准化的（下同），即

$$z_{ji}=(x_{ji}-\bar{x}_j)/s_j，\tag{6-1}$$

其中

$$\bar{x}_j=\frac{1}{N-1}\sum_{i=1}^{N}x_{ji}，$$

$$S_j^2=\frac{1}{N-1}\sum_{i=1}^{N}(x_{ji}-\bar{x}_j)^2。$$

对任何一个确定的z，$z_{ji}(i=1,2,\cdots,N)$所有值的集合称为标准化的统计变量z_{ji}。显然，标准化变量的均值为零$\left(\mathrm{E}(z_j)=0\right)$，方差为1$\left(\mathrm{var}(z_j)=1\right)$。

任意两个标准化变量的相关系数定义为

$$r_{jk}=\frac{1}{N-1}\sum_{i=1}^{N}z_{ji}z_{ki}=\frac{1}{N-1}z_jz_k'。\tag{6-2}$$

n个变量两两之间的相关系数构成一个方阵

$$=\begin{pmatrix} r_{11} & r_{12} & \cdots & r_{1n} \\ r_{21} & r_{22} & \cdots & r_{2n} \\ \vdots & \vdots & & \vdots \\ r_{n1} & r_{n2} & \cdots & r_{nn} \end{pmatrix}$$

称为相关矩阵。显然它的对角线元素$r_{jj}=1$。相关矩阵是一个实对称矩阵。

二、因子分析的数学模型

现用一实例来导出因子模型。

假若一沉积盆地，有三个剥蚀区供给沉积物（图6-1）。现从盆地沉积物中取N个样品，每个样品测得n个变量。若这些变量是各种重矿物的百分比含量，显然每个样品中不同重矿物的百分比含量与三个剥蚀区重矿物组成有关。由于剥蚀区母岩成分不同，造成其重矿物共生组合也有一定差别。若以F_1、F_2、F_3分别代表三个剥蚀区重矿物的共生组

合，则盆地中每个样品的各种重矿物百分比含量主要受F_1、F_2、F_3的控制，其次可能与其它因素（如风成作用带来的个别矿物，记为U_i）有关。因此，每个变量可以看作是F_1、F_2、F_3及U_i的函数，即

$$z_j = f(F_1, F_2, F_3, U_j)$$
$$= a_{j1}F_1 + a_{j2}F_2 + a_{j3}F_3 + b_jU_j \quad (j = 1,2,\cdots,n)。$$

由此可见，原始数据的信息通过各个变量表现出来，而这些变量之间存在着十分复杂的关系，有时数个有成因联系的变量共同反映某种信息，有时候某种特殊的信息只在某个变量中显示。因此可以设想将每个变量的信息分解开来，表示为各个因素的线性组合。这种表示变量和因素之间关系的数学模型称为因子模型，



图6-1 沉积盆地与剥蚀区示意图

$$z_j = a_{j1}F_1 + a_{j2}F_2 + \cdots + a_{jm}F_m + b_jU_j。 \tag{6-3}$$

将$j = 1,2,\cdots,n$展开，即为

$$\begin{cases} z_1 = a_{11}F_1 + a_{12}F_2 + \cdots + a_{1m}F_m + b_1U_1 \\ z_2 = a_{21}F_1 + a_{22}F_2 + \cdots + a_{2m}F_m + b_2U_2 \\ \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \\ z_n = a_{n1}F_1 + a_{n2}F_2 + \cdots + a_{nm}F_m + b_nU_n \end{cases}。 \tag{6-4}$$

因子模型中F_1，F_2，…，F_m（$m<n$）称为公因子，它们出现在一个以上的变量中；U_i称为单一因子（或特殊因子、唯一因子）；a_{j1}，a_{j2}，…，a_{im}称为公因子负荷（载荷），同一个公因子中几个较大因子负荷的变量形成一个变量组合，同一个变量组合具有成因联系，该公因子的物理意义由因子负荷较大的变量指示；b_i称为单一因子负荷，它表示该变量的某种特殊性。

为了将因子模型表示为矩阵形式，记

$$\mathbf{Z} = \begin{pmatrix} z_{11} & z_{12} & \cdots & z_{1N} \\ z_{21} & z_{22} & \cdots & z_{2N} \\ \vdots & \vdots & & \vdots \\ z_{n1} & z_{n2} & \cdots & z_{nN} \end{pmatrix}, \quad \mathbf{F} = \begin{pmatrix} F_{11} & F_{12} & \cdots & F_{1N} \\ F_{21} & F_{22} & \cdots & F_{2N} \\ \vdots & \vdots & & \vdots \\ F_{m1} & F_{m2} & \cdots & F_{mN} \end{pmatrix},$$

$$\mathbf{U} = \begin{pmatrix} u_{11} & u_{12} & \cdots & u_{1N} \\ u_{21} & u_{22} & \cdots & u_{2N} \\ \vdots & \vdots & & \vdots \\ u_{n1} & u_{n2} & \cdots & u_{nN} \end{pmatrix}, \quad \mathbf{A} = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1m} \\ a_{21} & a_{22} & \cdots & a_{2m} \\ \vdots & \vdots & & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nm} \end{pmatrix},$$

$$\mathbf{b} = \begin{pmatrix} b_1 & & & 0 \\ & b_2 & & \\ & & \ddots & \\ 0 & & & b_n \end{pmatrix},$$

则因子模型可以写成如下矩阵形式：

$$Z = AF + bU \quad 。 \tag{6-5}$$

在因子模型中，各公因子$F_i(i=1, 2, \cdots, m)$之间可以是互相独立的，也可以是互相相关的，但单一因子$U_j(j=1, 2, \cdots, n)$与公因子F_i之间则永远是相互独立的。

三、因子分析的基本定理

因子分析首先要解决的问题是根据变量的相关关系确定因子模型中A矩阵的元素，即确定各变量在各公因子上的负荷。

由于变量都已标准化，所以变量z_j与z_k的相关系数为

$$r_{jk} = \frac{1}{N-1}\sum_{i=1}^{N} x_{ji}x_{ki} \qquad (j, k = 1, 2, \cdots, n) \quad 。$$

相关系数矩阵为

$$R = \frac{1}{N-1}ZZ' \quad 。$$

将因子模型（6-5）式代入，得

$$R = \frac{1}{N-1}[AF + bU][AF + bU]'$$

$$= \frac{1}{N-1}[AFF'A' + bUF'A' + AFU'b' + bUU'b'] \quad 。$$

根据公因子及单一因子的独立性及标准化假设，即假设无论是公因子之间，公因子与单一因子之间，单一因子之间都是互相独立的，则有

$$\frac{1}{N-1}UF' = 0 \quad , \qquad \frac{1}{N-1}FU' = 0 \quad ,$$

$$\frac{1}{N-1}FF' = I \quad , \qquad \frac{1}{N-1}UU' = I \quad ,$$

式中I为单位矩阵。这里$\frac{1}{N-1}FF'$和$\frac{1}{N-1}UU'$实际上是公因子间相关矩阵和单一因子间相关矩阵。其中的元素

$$r_{F_jF_k} = \frac{1}{N-1}\sum_{i=1}^{N} F_{ji}F_{ki} = \begin{cases} 1 & \text{当} j = k \\ 0 & \text{当} j \neq k \end{cases} ,$$

$$(j, k = 1, 2, \cdots, m) \tag{6-6}$$

$$r_{u_ju_k} = \frac{1}{N-1}\sum_{i=1}^{N} u_{ji}u_{ki} = \begin{cases} 1 & \text{当} j = k \\ 0 & \text{当} j \neq k \end{cases}$$

$$(j, k = 1, 2, \cdots, n) \tag{6-7}$$

所以，当公因子间相互独立，单一因子间也相互独立，公因子和单一因子都已标准化的条件下，它们的相关矩阵才是单位阵。

由上可得：

$$R = AA' + bb' \quad 。 \tag{6-8}$$

令$R^* = R - bb'$，并称之为“约相关矩阵”，则据（6-8），约相关矩阵R^*与因子负荷矩阵

A有如下关系：

$$\mathbf{R}^* = \mathbf{AA}' \tag{6-9}$$

(6-9) 式指出，因子分析要求将n个变量之间的相关关系转化为这n个变量和m个公因子之间的相关关系。换言之，所谓因子分析就是在已知约相关矩阵$\mathbf{R}^*$的条件下求解因子负荷矩阵，使$\mathbf{AA}' = \mathbf{R}^*$。因此可以将关系式 (6-9) 看作因子分析的基本定理。

四、变量方差的组成

一个标准化变量j的总方差为1，即相关矩阵中主对角线元素r_{jj}由 (6-8) 式可得：

$$r_{jj} = a_{j1}^2 + a_{j2}^2 + \cdots + a_{jm}^2 + b_j^2 = 1 \ 。$$

令

$$h_j^2 = \sum_{i=1}^{n} a_{ji}^2 = a_{j1}^2 + a_{j2}^2 + \cdots + a_{jm}^2 \ , \tag{6-10}$$

称h_j^2为变量 j 的公因子方差。它表示全部公因子对变量 j 的总方差所作的贡献。b_j^2称为单一因子方差，是由于变量 j 的特殊因素对总方差的贡献。并且有

$$1 = h_j^2 + b_j^2 \ 。 \tag{6-11}$$

顺便指出，约相关矩阵$\mathbf{R}^*$与相关矩阵R的差别仅在于R的对角线元素为1，而$\mathbf{R}^*$的对角线元素是由诸变量的公因子方差组成的，即

$$\mathbf{R}^* = \begin{pmatrix} 1-b_1^2 & r_{12} & \cdots & r_{1n} \\ r_{21} & 1-b_2^2 & \cdots & r_{2n} \\ \vdots & \vdots & & \vdots \\ r_{n1} & r_{n2} & \cdots & 1-b_n^2 \end{pmatrix} = \begin{pmatrix} h_1^2 & r_{12} & \cdots & r_{1n} \\ r_{21} & h_2^2 & \cdots & r_{2n} \\ \vdots & \vdots & & \vdots \\ r_{n1} & r_{n2} & \cdots & h_n^2 \end{pmatrix} 。$$

变量 j 的公因子方差h_j^2是因子负荷矩阵中第j行元素的平方和。现讨论某个公因子F_k的方差贡献S_k^2，把它定义为因子负荷矩阵中第k列元素的平方和，即

$$S_k^2 = \sum_{j=1}^{n} a_{jk}^2 = a_{1k}^2 + a_{2k}^2 + \cdots + a_{nk}^2 \qquad (k = 1, 2, \cdots, m) \ , \tag{6-12}$$

它是同一个公因子F_k对各变量所提供的方差总和。S_k^2愈大，公因子 F_k 愈重要，它是衡量因子相对重要性的指标。

n个标准化变量的总方差为n，公因子F_k的方差贡献S_k^2表示了在n个变量方差总量中的一部分，所以它有时也可用方差百分比来表示$\left(\frac{S_k^2}{n} \times 100\%\right)$。公因子$F_k$的方差百分比愈大，说明该公因子愈重要。$m$个公因子的累计方差百分比$\left(\left(\sum_{k=1}^{m} S_k^2\right) \Big/ n \times 100\%\right)$表示了$n$个变量用$m$个公因子来表示时，所占方差比例。

因子分析的目的之一是为了把握事物的主要矛盾，研究事物主要因素的顺序，因此提

取的各公因子F_k，可以从约相关矩阵R*中按照它们方差贡献S^2_k的大小来逐一提取。或者说，将提取的各公因子F_k，按照它们的方差贡献S^2_k的大小排列，即$S^2_1 \geqslant S^2_2 \geqslant \cdots \geqslant S^2_m$。这样，把对应的$F_1$，$F_2$，…，$F_m$分别称为第1主因子，第2主因子，…，第$m$主因子。

五、因子负荷矩阵A的导出

由因子分析的基本定理

$$R^* = AA' ,$$

用r^*_{jk}来表示R*中的元素，则有

$$r^*_{jk} = \sum_{i=1}^{m} a_{ji}a_{ki} \qquad (j,\ k=1,\ 2,\ \cdots,\ n) 。 \tag{6-13}$$

选取第1主因子F_1，即确定F_1的因子负荷a_{j1}，使得其方差贡献S^2_1为最大，也就是说要求$S_{21} = \sum_{j=1}^{n} a^2_{j1}$达到极大值的解。这是一个条件极值问题，令

$$2T = S^2_1 - \sum_{j,k=1}^{n} \mu_{jk} r^*_{jk} = S^2_1 - \sum_{j,k=1}^{n} \sum_{i=1}^{m} \mu_{jk} a_{ji} a_{ki} , \tag{6-14}$$

式中T为一新的函数；μ_{jk}为拉格朗日乘子。根据数学分析中求极值的原理，函数T极大的充分必要条件是求T对a_{j1}的偏导数，并令其等于零，其中T对每个a_{j1} $(j=1,2,\cdots,n)$的偏导数为

$$\frac{\partial T}{\partial a_{j1}} = a_{j1} - \sum_{k=1}^{n} \mu_{jk} a_{k1} = 0 。 \tag{6-15}$$

同样，求T对其余每一个变量$a_{jp}$$(p \neq 1)$的偏导数，且令其等于零，即

$$\frac{\partial T}{\partial a_{jp}} = - \sum_{k=1}^{n} \mu_{jk} a_{kp} = 0 。 \tag{6-16}$$

可以将(6-15)，(6-16)两组方程式结合起来，写成

$$\frac{\partial T}{\partial a_{jp}} = \delta_{1p} a_{j1} - \sum_{k=1}^{n} \mu_{jk} a_{kp} = 0 \qquad (p=1,2,\cdots,n) \tag{6-17}$$

式中克罗内克符号δ_{1p}取值为：

$\delta_{1p}=1$，若$p=1$，

$\delta_{1p}=0$，若$p \neq 1$。

将(6-17)式两边乘a_{j1}，并对j求和，得

$$\delta_{1p} \sum_{j=1}^{n} a^2_{j1} - \sum_{j=1}^{n} \sum_{k=1}^{n} \mu_{jk} a_{j1} a_{kp} = 0 。 \tag{6-18}$$

据(6-15)式$\sum_{j=1}^{n} \mu_{jk} a_{j1} = a_{k1}$；令$\sum_{j=1}^{n} a^2_{j1} = \lambda_1$则(6-18)式可写成

$$\delta_{1p}\lambda_1-\sum_{k=1}^{n}a_{k1}a_{kp}=0\ 。 \tag{6-19}$$

将（6-19）式两边乘以a_{jp}，并对p求和

$$a_{j1}\lambda_1-\sum_{k=1}^{n}\Big(\sum_{p=1}^{m}a_{jp}a_{kp}\Big)=0\ 。 \tag{6-20}$$

将（6-13）代入上式

$$a_{j1}\lambda_1-\sum_{k=1}^{n}r_{jk}^{*}a_{k1}=0\ 。 \tag{6-21}$$

对（6-21）乘（-1），并按j（$j=1,2,\cdots,n$）展开

$$\begin{cases}(h_1^2-\lambda)a_{11}+r_{12}a_{21}+\cdots+r_{1n}a_{n1}=0\\ r_{21}a_{11}+(h_2^2-\lambda)a_{21}+\cdots+r_{2n}a_{n1}=0\\ \cdots\cdots\cdots\cdots\cdots\cdots\\ r_{n1}a_{11}+r_{n2}a_{21}+\cdots+(h_n^2-\lambda)a_{n1}=0\end{cases}\ 。 \tag{6-22}$$

（6-22）式可写成下列矩阵形式：

$$\begin{pmatrix}(h_1^2-\lambda)&r_{12}&\cdots&r_{1n}\\ r_{21}&(h_2^2-\lambda)\cdots&&r_{2n}\\ \vdots&\vdots&&\vdots\\ r_{n1}&r_{n2}&\cdots&(h_n^2-\lambda)\end{pmatrix}\cdot\begin{pmatrix}a_{11}\\ a_{21}\\ \vdots\\ a_{n1}\end{pmatrix}=\begin{pmatrix}0\\ 0\\ \vdots\\ 0\end{pmatrix}, \tag{6-23}$$

或写成

$$[\mathbf{R}^*-\lambda\mathbf{I}]\cdot\mathbf{a}=\mathbf{0}\ , \tag{6-24}$$

式中$\mathbf{R}^*=(r_{jk}^{*})$ $(j,k=1,2,\cdots,n)$ 为约相关矩阵；$\mathbf{I}$为单位矩阵；a_{j1}为第1主因子的因子负荷。

根据因子分析的要求和目的，$a_{j1}=(a_{11},a_{21},\cdots,a_{n1})$不应同时为零，因此上述齐次方程组有非显然解。存在非显然解的充分必要条件是因子负荷a_{j1}的系数行列式等于0，即

$$\begin{vmatrix}(h_1^2-\lambda)&r_{12}&\cdots&r_{1n}\\ r_{21}&(h_2^2-\lambda)\cdots&&r_{2n}\\ \vdots&\vdots&&\vdots\\ r_{n1}&r_{n2}&&(h_n^2-\lambda)\end{vmatrix}=0\ , \tag{6-25}$$

或写成

$$|\mathbf{R}^*-\lambda\mathbf{I}|=0\ 。 \tag{6-26}$$

式（6-25）是矩阵$\mathbf{R}^*$的特征方程，其中λ为特征方程的根，即$\mathbf{R}^*$的特征值。由于$\mathbf{R}^*$是实对称矩阵，且$|\mathbf{R}^*|\geqslant 0$，因此它的所有特征值均为非负实数。

在建立(6-23)方程组的过程中，由于令$\lambda_j=\sum_{j=1}^{n}a_{j1}^2$，而$\sum_{j=1}^{n}a_{j1}^2$又是$F_j$的方差贡献$S_j^2$，

即

$$\lambda_i = \sum_{j=1}^{n} a_{ji}^2 = S_i^2 \quad (6\text{-}27)$$

因此，可以选择最大特征根作为λ_1，即

$$\lambda_1 = \max_{i} \{\lambda_i\}$$

将λ_1代入（6-22）式，并解这一齐次方程组，可得一组解a_{11}，a_{21}，…，a_{n1}。

同理可求得第2公因子，第3公因子，…，第m公因子，并求得它们的解。

六、公因子个数m的确定

至今，还没有一个统计学的基本检验能普遍地用来确定公因子的恰当个数。有用的准则是：

1．利用m个公因子总方差的累积百分数，

$$\left(\sum_{k=1}^{m} S_k^2\right)/n \times 100\% = \sum_{k=1}^{m} \lambda_k / n \times 100\% ,$$

当它大于某个值（如85%—95%）时的m数。

2．特征值超过1.0的因子可以认为是“有意义”的。

3．在特征值与因子数目相应的曲线图上，曲线的斜率表现出一个“有意义”的拐点（Cattell，1966）。

4．保留那些与一个以上变量有重大关系的因子。

5．保留那些从地球化学解释的观点来看有意义的因子。

第三节　主成分分析

一、主成分分析的数学模型

在第二节中介绍了因子负荷矩阵的求法，但实际工作中并不一定要那样做。因为，实际上约相关矩阵中公因子方差的确定是很麻烦的，逐个提取公因子的办法可以用其它办法来代替。

在因子模型中，可把变量表示为公因子和单一因子两部分，即

$$\mathbf{Z} = \mathbf{AF} + \mathbf{bU} \ 。$$

当因子出现在一个以上变量中时称为公因子；当它只在一个变量中出现时称为单一因子，因此，公因子和单一因子的区别只是它出现的变量个数是否大于1，也就是说，它们的因子负荷是不是只在一个变量中有较大的值，而其它并没有本质的区别。所以我们可以把公因子和单一因子统称为因子，这时变量z_i可以表示为n个因子的函数式，即

$$z_j = a_{j1}F_1 + a_{j2}F_2 + \cdots + a_{jn}F_n \ 。 \quad (6\text{-}28)$$

写成矩阵形式

$$\mathbf{Z} = \mathbf{AF} \ 。 \quad (6\text{-}29)$$

这就是主成分分析的数学模型，其中：

$$\mathbf{Z}=\begin{pmatrix} z_{11} & z_{12} & \cdots & z_{1N} \\ z_{21} & z_{22} & \cdots & z_{2N} \\ \vdots & \vdots & & \vdots \\ z_{n1} & z_{n2} & \cdots & z_{nN} \end{pmatrix},\quad \mathbf{A}=\begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{pmatrix},$$

$$\mathbf{F}=\begin{pmatrix} F_{11} & F_{12} & \cdots & F_{1N} \\ F_{21} & F_{22} & \cdots & F_{2N} \\ \vdots & \vdots & & \vdots \\ F_{n1} & F_{n2} & \cdots & F_{nN} \end{pmatrix}。$$

写成展开的形式

$$\begin{cases} z_1 = a_{11}F_1 + a_{12}F_2 + \cdots + a_{1m}F_m + \cdots + a_{1n}F_n \\ z_2 = a_{21}F_1 + a_{22}F_2 + \cdots + a_{2m}F_m + \cdots + a_{2n}F_n \\ \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \\ z_n = a_{n1}F_1 + a_{n2}F_2 + \cdots + a_{nm}F_m + \cdots + a_{nn}F_n \end{cases} \tag{6-30}$$

在主成分分析的数学模型中，假设各因子的顺序也是按其方差贡献的大小排列，即

$$S_1^2 \geqslant S_2^2 \geqslant \cdots \geqslant S_m^2 \geqslant \cdots \geqslant S_n^2,$$

那么，主成分分析解中的前m列与因子负荷矩阵是相同的。

二、主因子解的求得

由主成分分析的数学模型$\mathbf{Z}=\mathbf{AF}$两边同乘其转置矩阵，得：

$$\mathbf{ZZ'}=\mathbf{AFF'A'}。 \tag{6-31}$$

同因子分析基本定理推导过程一样，在各变量和各因子是标准化的，且各公因子之间相互独立的假设条件下有

$$\mathbf{R}=\mathbf{AA'}。 \tag{6-31}$$

(6-31)式表示，主成分的因子负荷矩阵A可以从相关矩阵分解得来。而主成分负荷矩阵的前m列与因子负荷矩阵相同。

由矩阵的特征值和特征向量性质可知，对任何非奇异矩阵，经过有限次的正交变换，都可以使它成为对角阵，其对角线元素为特征根，相应正交变换的向量为特征向量。相关矩阵R是实对称矩阵，当然可以进行足够多次正交变换后成为对角阵。

记Λ为对角阵，其对角线元素为λ_i；P为正交阵；$\mathbf{M}_k=\mathbf{P}_1\mathbf{P}_2\cdots\mathbf{P}_k$；$\mathbf{M}'$为M的转置，则

$$\mathbf{M}_k'\mathbf{RM}=\Lambda, \tag{6-32}$$

即对相关矩阵R进行足够多次（k次）正交变换后，R成为具有对角线元素λ_i的对角阵Λ，而$\mathbf{M}_k$则为R的特征向量所构成的矩阵。(6-32)式两边左乘$\mathbf{M}_k$，右乘$\mathbf{M}_k'$得

$$\mathbf{R}=\mathbf{M}_k\Lambda\mathbf{M}_k'。$$

据(6-31)式，则

$$\mathbf{AA'}=\mathbf{M}_k\Lambda^{1/2}\Lambda^{1/2}\mathbf{M}_k',$$

其中$\Lambda^{1/2}$为具有对角线元素$\sqrt{\lambda_i}$的对角阵，从而

$$\mathbf{A}=\mathbf{M}\Lambda^{1/2}, \tag{6-33}$$

其中

$$M=\begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{pmatrix}, \quad \Lambda^{1/2}=\begin{pmatrix} \sqrt{\lambda_1} & & & 0 \\ & \sqrt{\lambda_2} & & \\ & & \ddots & \\ 0 & & & \sqrt{\lambda_n} \end{pmatrix},$$

且有$\lambda_1 \geqslant \lambda_2 \geqslant \cdots \geqslant \lambda_n$。

将（6-33）式写成展开形式

$$\begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{pmatrix}=\begin{pmatrix} a_{11}\sqrt{\lambda_1} & a_{12}\sqrt{\lambda_2} & \cdots & a_{1n}\sqrt{\lambda_n} \\ a_{21}\sqrt{\lambda_1} & a_{22}\sqrt{\lambda_2} & \cdots & a_{2n}\sqrt{\lambda_n} \\ \vdots & \vdots & & \vdots \\ a_{n1}\sqrt{\lambda_1} & a_{n2}\sqrt{\lambda_2} & \cdots & a_{nn}\sqrt{\lambda_n} \end{pmatrix},$$

即

$$a_{ij}=a_{ij}\sqrt{\lambda_i} \quad 。 \tag{6-34}$$

由第二节中用拉格朗日乘数法求条件极值，同样得出主因子解的实质就是求相关矩阵的特征值和特征向量。对于主成分分析可以从相关矩阵出发，但更多的是从方差-协方差矩阵出发，求得主因子解，这时的主因子称为主成分。由于主成分分析在这种情况下，变量没有标准化，所求得的特征值相对来说是不均匀的，特征向量也没有正规化。

求矩阵特征根和特征向量的方法很多，目前电算中常用的是雅可比法和迭代法。

求得因子负荷矩阵后，同样按上节中的原则，选取前m个作为主成分。之后可计算公因子方差h_i^2，方差贡献S_i^2和累积方差百分比。

第四节　正交旋转——正交因子解

求得初始因子矩阵A之后，可以按照公因子进行各种地质解释工作。但主因子解不是唯一的。在理论上存在着无穷多组彼此等价的因子解，这些解都和已知的实测相关矩阵相对应。从几何的意义上说，公因子空间一旦被确定以后就有无穷多个因子轴系可以相互旋转而同样确切地再现实测相关矩阵。事实上，对于任一正交阵T，令$B=AT$，则有

$$BB'=(AT)(AT)'=ATT'A'=AA'=R \quad ,$$

即AT仍是一个因子解。

因此，在初始因子解还没有与自然模型相吻合时，可以对初始因子解作旋转，使得因子模型尽量和自然模型相吻合。也就是说，旋转后的因子模型更便于地质解释。

图6-2表示主因子解中初始因子轴与变量之间的关系；图6-3表示初始因子轴绕原点并保持因子轴相互正交的条件下旋转一定角度后与变量的关系。显然，经旋转后每一因子只和较少的变量有相关关系（即在每一因子轴上有相当多的a_{ij}趋近于零）。也就是说，旋转后对某因子轴，有一部分变量与其夹角趋近于零，而另外几个变量与其之夹角趋近于90°。从代数意义上说，即使得因子的方差达到最大。

正交因子旋转有多种方法，最常用的是方差极大法（Varimax method）。在数学上，是使各因子负荷的方差达到极大作为因子负荷矩阵简单化的准则。

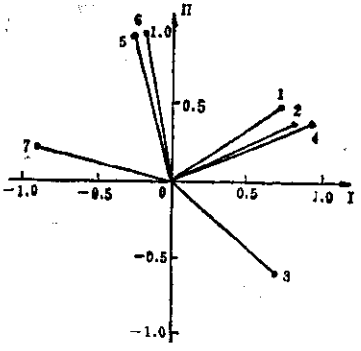


图6-2 初始因子轴与变量之间的关系

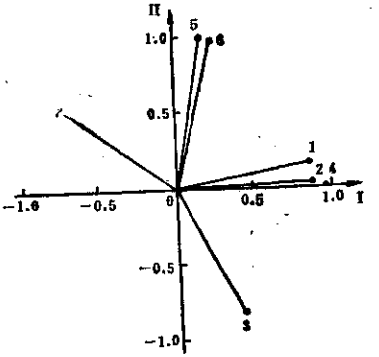


图6-3 正交旋转后因子轴与变量之间的关系

对于初始因子解的负荷矩阵

$$\mathbf{A}=\begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1m} \\ a_{21} & a_{22} & \cdots & a_{2m} \\ \vdots & \vdots & & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nm} \end{pmatrix},$$

令旋转后的因子负荷矩阵为

$$\mathbf{B}=\begin{pmatrix} b_{11} & b_{12} & \cdots & b_{1m} \\ b_{21} & b_{22} & \cdots & b_{2m} \\ \vdots & \vdots & & \vdots \\ b_{n1} & b_{n2} & \cdots & b_{nm} \end{pmatrix},$$

则

$$\mathbf{B}=\mathbf{AT}\ , \tag{6-35}$$

其中 T为正交阵。

先来讨论取p、q两列的情况，(6-35)式可写为

$$\begin{pmatrix} b_{1p} & b_{1q} \\ b_{2p} & b_{2q} \\ \vdots & \vdots \\ b_{np} & b_{nq} \end{pmatrix}=\begin{pmatrix} a_{1p} & a_{1q} \\ a_{2p} & a_{2q} \\ \vdots & \vdots \\ a_{np} & a_{nq} \end{pmatrix}\begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix},$$

写成一般公式为

$$\begin{cases} b_{jp}=a_{jp}\cos\theta+a_{jq}\sin\theta \\ b_{jq}=-a_{jp}\sin\theta+a_{jq}\cos\theta \end{cases} \quad \begin{pmatrix} j=1,2,\cdots,n; \\ p,q=1,2,\cdots,m \end{pmatrix} \tag{6-36}$$

方差极大法是将某一因子p的简单性定义为因子p的各因子负荷平方(b_{jp}^2)的方差：

$$S_p^2=\frac{1}{n}\sum_{j=1}^{n}(b_{jp}^2-\overline{b_i^2})^2=\frac{1}{n}\sum_{j=1}^{n}(b_{jp}^2)^2-\frac{1}{n^2}\Big(\sum_{j=1}^{n}b_{jp}^2\Big)^2 \tag{6-37}$$

为极大时，因子p则具有最大的简单性。即F,中各因子负荷b_{ip}有的趋近于1，有的则趋近于零，向两极分化，使得不同变量对同一因子的负荷有尽可能大的差异。该公因子的物理意义可以更明确地由因子负荷接近于1的变量组合来表征。对所有m个公因子而言，就应使

$$S^2=\sum_{p=1}^{m}S_p^2=\frac{1}{n}\sum_{p=1}^{m}\sum_{j=1}^{n}\left(b_{jp}^4\right)-\frac{1}{n^2}\sum_{r=1}^{m}\left(\sum_{j=1}^{n}b_{jp}^2\right)^2 \tag{6-38}$$

为极大，整个因子矩阵可获得最简单结构。这一原则，称为原始方差极大准则。但由于各变量公因子方差(h_j^2)之间的差异使其对（6-38）式的作用造成不平衡，因此在实际应用时，常常消除各变量公因子方差的影响，将第j个变量的各因子负荷的平方(b_{jp}^2)除以h_j^2，即

$$S^2=\frac{1}{n}\sum_{p=1}^{m}\sum_{j=1}^{n}\left(b_{jp}^2/h_j^2\right)^2-\frac{1}{n^2}\sum_{p=1}^{m}\left(\sum_{j=1}^{n}b_{jp}^2/h_j^2\right)^2 。 \tag{6-39}$$

又因用常量n^2乘上式后极值点不变，于是得

$$V=n\sum_{p=1}^{m}\sum_{j=1}^{n}\left(b_{jp}^2/h_j^2\right)^2-\sum_{p=1}^{m}\left(\sum_{j=1}^{n}b_{jp}^2/h_j^2\right)^2 。 \tag{6-40}$$

以V取得极大为准则来求得最后因子负矩阵B，这一准则又称正规化方差极大准则。

将（6-36）式代入上式，对θ求一阶导数，并令其等于零，解出方程，得

$$\tan 4\theta=\frac{D-2AB/n}{C-(A^2-B^2)/n}=\frac{\nu}{\delta} \tag{6-41}$$

其中，令

$$u_j=(a_{jp}/h_j)^2-(a_{jq}/h_j)^2 ,$$

$$v_j=2(a_{jp}/h_j)(a_{jq}/h_j) ,$$

$$(p,q=1,2,\cdots,m;\ p\neq q)$$

则

$$A=\sum_{j=1}^{n}u_j \qquad B=\sum_{j=1}^{n}v_j ,$$

$$C=\sum_{j=1}^{n}\left(u_j^2-v_j^2\right) , \quad D=2\sum_{j=1}^{n}u_jv_j .$$

这样，使得V为极大值的θ角可从（6-41）式解得。其计算步骤为：

（1）先确定一对因子$p,q(p,q=1,2,\cdots,m;\ p\neq q)$求得$u_j$、$v_j$，进而求得$A$、$B$、$C$、$D$,代入（6-41）式求出$\tan 4\theta$。

（2）根据（6-41）式中ν和δ的代数符号查表6-1确定θ角。

（3）因子初始矩阵A（或前一次旋转的矩阵）中的第p、q列右乘正交阵T_{pq}（即按式（6-36）计算），得一对因子轴旋转后的结果。

（4）经过所有两两不同的因子对$(p,q=1,2,\cdots,m;\ p\neq q)$的组合，它们共有$C_m^2=\frac{m(m-1)}{2}$次正交变换，求得一轮循环的旋转因子矩阵$B_1$，

$$B_1 = AT_{12}T_{13}\cdots T(m-1)m \ 。$$

经过若干轮循环，

$$B = B_1B_2\cdots B_s \ 。$$

直到V值的增加量小于某预定精度为止，然后经正规化还原，从而得到最终正交多因子解

$$B = A\ T \ 。$$

表6-1 旋 转 角

式（6-41）的代数符号			$\cos 4\theta$ 的符号	4θ 所在象限	θ 的变化范围
分子（γ）（和$\sin 4\theta$）	分 母（δ）	$\tan 4\theta$			
+	+	+	+	I：$0° < 4\theta < 90°$	$0 \sim 22.5°$
+	−	−	−	II：$90° < 4\theta < 180°$	$22.5° \sim 45°$
−	−	+	−	III：$-180 < 4\theta < -90°$	$-45° \sim -22.5°$
−	+	−	+	IV：$-90° < 4\theta < 0°$	$-22.5° \sim 0°$

第五节 斜交旋转——斜交因子解

正交旋转是在旋转过程中因子互相正交，始终保持初始解中因子间不相关的特点。但在实际上公因子之间常常具有一定相关性。这时就须用斜交旋转。斜交旋转得到斜交因子解。在斜交因子解中因子模型的公因子系数就不是变量与因子间的相关系数。因此，一个完全的斜交因子解，除了因子模型外，还要另有反映变量与因子间相关系数的因子结构和因子间相关关系的相关矩阵。

一、斜交因子模型

如果n个变量z_1，z_2，…，z_n可以用m个相关的斜交公因子T_1，T_2，…，T_m来表示，若不考虑特殊因子，其斜交因子模型可以写成：

$$\begin{cases} z_1 = b_{11}T_1 + b_{12}T_2 + \cdots + b_{1m}T_m \\ z_2 = b_{21}T_1 + b_{22}T_2 + \cdots + b_{2m}T_m \\ \cdots \quad \cdots \quad \cdots \quad \cdots \\ z_n = b_{n1}T_1 + b_{n2}T_2 + \cdots + b_{nm}T_m \end{cases}, \tag{6-42}$$

式中，T_1，T_2，…，T_m是斜交公因子，可以看成斜坐标的单位向量；b_{ij}表示变量z_i的向量p_i在斜因子轴T_j上的坐标，称为斜交因子负荷，可以看成因子空间中表示变量z_i的点p_i的第j个斜坐标。如果把（6-42）式写成矩阵形式，即

$$\begin{bmatrix} z_1 \\ z_2 \\ \vdots \\ z_n \end{bmatrix} = \begin{bmatrix} b_{11} & b_{12} & \cdots & b_{1m} \\ b_{21} & b_{22} & \cdots & b_{2m} \\ \vdots & \vdots & & \vdots \\ b_{n1} & b_{n2} & \cdots & b_{nm} \end{bmatrix} \begin{bmatrix} T_1 \\ T_2 \\ \vdots \\ T_m \end{bmatrix},$$

或

$$Z = BT$$

矩阵B叫做变量Z（关于斜交因子T）的因子模型（因子负荷）矩阵。

对于正交因子，因子模型就是因子解，这时因子模型和因子结构是一致的。但对于斜交因子，两者是有区别的。

如图6-4所示，T_1、T_2是斜因子轴，OP表示变量z_j的向量，其长度为1，它在T_1、T_2的投影是OQ和OR，它们就是斜交因子负荷（b_1，b_2）。OM、ON分别表示变量z_j与因子T_1、T_2的相关系数（$OM=r_{jT_1}$，$ON=r_{jT_2}$）。即因子结构中的元素，它们和因子负荷是不同的。$\cos\theta_{12}°$表示了因子T_1和T_2的相关程度。

斜交因子解是由正交因子解变换而来。

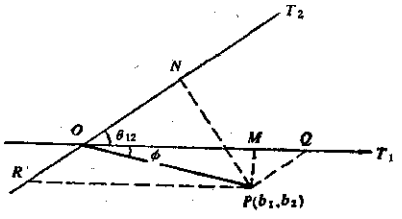


图6-4 斜交因子轴系中因子负荷与相关系数的关系

(1) 因子变换矩阵

所谓因子变换矩阵，就是从正交因子F变换（旋转）成斜交因子T的变换矩阵。现从正交因子解出发，研究正交与斜交的变换关系。如果把T_1、T_2、…、T_m看成斜坐标轴系的单位向量，则T_j在m个正交因子F_1、F_2、…、F_m方向（正交坐标轴系方向）上的投影：t_{1j}、t_{2j}、…、t_{mj}（即T_j端点的正交坐标）就是斜坐标轴T_j相对于正交坐标因子轴F_1、F_2、…、F_m的夹角余弦，其平方和等于1，即

$$t_{ij}=|T_j|\cos(T_j,\ F_i)\qquad (i=1,2,\cdots,m)\ ,$$

$$T_j=[t_{1j},\ t_{2j},\ \cdots,\ t_{mj}]'\ ,$$

$$\sum_{i=1}^{m}t_{ij}^2=1\qquad (j=1,2,\cdots,m)\ 。$$

$$\mathbf{T}=[T_1,\ T_2,\ \cdots,\ T_m]=\begin{bmatrix}t_{11} & t_{12} & \cdots & t_{1m}\\ t_{21} & t_{22} & \cdots & t_{2m}\\ \vdots & \vdots & & \vdots\\ t_{m1} & t_{m2} & \cdots & t_{mm}\end{bmatrix}。$$

(2) 斜因子相关矩阵

m个斜因子轴既互不独立，必有相关，就要确定刻划m个斜交因子之间相关性的相关矩阵，$\mathbf{L}=(l_{ij})$ $(i,j,=1,2,\cdots,m)$，其中元素l_{ij}表示斜因子T_i与T_j的相关系数，

$$l_{ij}=T_iT_j'=t_{1i}t_{1j}+t_{2i}t_{2j}+\cdots+t_{mi}t_{mj}\ 。$$

把它写成矩阵的形式

$$\mathbf{L}=\mathbf{T}'\mathbf{T}\qquad (6\text{-}43)$$

(3) 因子结构矩阵

$$\mathbf{S}=(s_{ij})\ ,$$

其中元素s_{ij}（$i=1,2,\cdots,n$；$j=1,\ 2,\cdots,\ m$）表示第i个变量z_i的向量p_i在斜因子轴T_j上的投影。以二维情况为例，图6-5中，s_{i1}表示向量p_i在T_1上的投影，即向量OM；s_{i2}表示向量P_i在T_2上的投影，即向量ON。因为OM就是向量P_i和斜因子轴T_1的内积，故OM是变

量 i 与斜因子 T_1 的相关系数。一般 s_{ij} $(i=1,2,\cdots,n;\ j=1,2,\cdots,m)$ 是变量 i 与斜因子 j 的相关系数。

因子结构矩阵S可以通过正交因子负荷矩阵A和相对应的因子变换矩阵T得到，它们三者的关系可用二维图示说明（图6-5）。

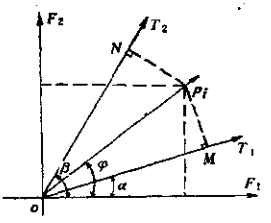


图6-5 变量、斜因子、相关系数之间的关系

向量P_i在斜因子轴T_1及T_2上的投影分别为：

$$OM = |P_i|\cos(\varphi-\alpha)$$

$$ON = |P_i|\cos(\beta-\varphi)$$

式中OM、ON分别为变量 i 与斜因子轴 T_1 及 T_2 的相关系数r_{iT_1}、r_{iT_2}；向量长度 $|P_i| = h_i$，即变量i的公因子方差h_i^2的平方根，于是有

$$r_{iT_1} = h_i(\cos\varphi\cos\alpha + \sin\varphi\sin\alpha)$$

$$= (h_i\cos\varphi)\cos\alpha + (h_i\sin\varphi)\sin\alpha = a_{i1}\cos\alpha + a_i2\sin\alpha\ ,$$

$$r_{iT_2} = h_i(\cos\beta\cos\varphi + \sin\beta\sin\varphi)$$

$$= (h_i\cos\varphi)\cos\beta + (h_i\sin\varphi)\sin\beta = a_{i1}\cos\beta + a_{i2}\sin\beta\ ,$$

即

$$[r_{iT_1}, r_{iT_2}] = [a_{N1}\ \ a_{i2}]\begin{bmatrix}\cos\alpha & \cos\beta\\ \sin\alpha & \sin\beta\end{bmatrix}$$

$$= [a_{i1}\ \ a_{i2}]\begin{bmatrix}t_{11} & t_{12}\\ t_{21} & t_{22}\end{bmatrix}。$$

一般，对于m个因子的情况，有

$$\begin{bmatrix}r_{1T_1} & r_{1T_2} & \cdots & r_{1T_m}\\ r_{2T_1} & r_{2T_2} & \cdots & r_{2T_m}\\ \vdots & \vdots & & \vdots\\ r_{nT_1} & r_{nT_2} & \cdots & r_{nT_m}\end{bmatrix} = \begin{bmatrix}a_{11} & a_{12} & \cdots & a_{1m}\\ a_{21} & a_{22} & \cdots & a_{2m}\\ \vdots & \vdots & & \vdots\\ a_{n1} & a_{n2} & \cdots & a_{nm}\end{bmatrix} \cdot \begin{bmatrix}t_{11} & t_{12} & \cdots & t_{1m}\\ t_{21} & t_{22} & \cdots & t_{2m}\\ \vdots & \vdots & & \vdots\\ t_{m1} & t_{m2} & \cdots & t_{mm}\end{bmatrix},$$

即

$$\mathbf{S = AT}\ 。 \qquad (6\text{-}44)$$

（4）因子模型矩阵

由（6-42）式斜交因子模型

$$z_i = b_{i1}T_1 + b_{i2}T_2 + \cdots + b_{im}T_m \qquad (i=1,2,\cdots,n)\ 。$$

于是有

$$r_{iT_j} = E(z_iT_j) = b_{i1}r_{T_1T_j} + b_{i2}r_{T_2T_j} + \cdots + b_{ij}r_{T_jT_j} + \cdots + b_{im}r_{T_mT_j}$$

$$(i=1,2,\cdots,n;\ j=1,2,\cdots,m)\ ,$$

即

$$r_{iT_1} = b_{i1} + b_{i2}r_{T_2T_1} + \cdots + b_{im}r_{T_mT_1}\ ,$$

$$r_{iT_2}=b_{i1}r_{T_1T_2}+b_{i2}+\cdots+b_{im}r_{T_mT_2}\ ,$$

$$\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots$$

$$r_{iT_m}=b_{i1}r_{T_1T_m}+b_{i2}r_{T_2T_m}+\cdots+b_{im}\ ,$$

$$(i=1,2,\cdots,n)\ .$$

由（6-43）式，则有

$$\begin{bmatrix} r_{1T_1} & r_{1T_2} & \cdots & r_{1T_m} \\ r_{2T_1} & r_{2T_2} & \cdots & r_{2T_m} \\ \vdots & \vdots & & \vdots \\ r_{nT_1} & r_{nT_2} & \cdots & r_{nT_m} \end{bmatrix}=\begin{bmatrix} b_{11} & b_{12} & \cdots & b_{1m} \\ b_{21} & b_{22} & \cdots & b_{2m} \\ \vdots & \vdots & & \vdots \\ b_{n1} & b_{n2} & \cdots & b_{nm} \end{bmatrix}\cdot\begin{bmatrix} l_{11} & l_{12} & \cdots & l_{1m} \\ l_{21} & l_{22} & \cdots & l_{2m} \\ \vdots & \vdots & & \vdots \\ l_{m1} & l_{m2} & \cdots & l_{mm} \end{bmatrix},$$

写成矩阵

$$\mathbf{S}=\mathbf{B}\,\mathbf{L}\ . \qquad (6\text{-}45)$$

把（6-43）和（6-44）式代入（6-45）式得

$$\mathbf{A}\mathbf{T}=\mathbf{B}[\mathbf{T}'\mathbf{T}]\ ,$$

从而有

$$\mathbf{B}=\mathbf{A}[\mathbf{T}']^{-1} \qquad (6\text{-}46)$$

（6-46）式表示，当已求得正交因子负荷矩阵（已知A）时，只要知道T′就能求出斜交因子模型B。但由于T不是对称矩阵，所以也可以用（6-45）式求B，即

$$\mathbf{B}=\mathbf{S}\,\mathbf{L}^{-1}\ .$$

二、斜交参考解

为了满足简单结构准则，使因子解中出现一定数目的零，可引进一组斜交参考轴$\{\Lambda_j\}$它们正好是斜主因子坐标超平面的垂直线。以两维为例，如图6-6所示，斜因子轴T_1，T_2通过变量点群的中心，坐标平面是斜因子轴T_1和T_2组成，作新的斜交参考轴Λ_2和Λ_1，它们分别垂直于T_1和T_2，1—4四个点在Λ_2上的投影以及5—8四个点在Λ_1上投影都非常接近0。

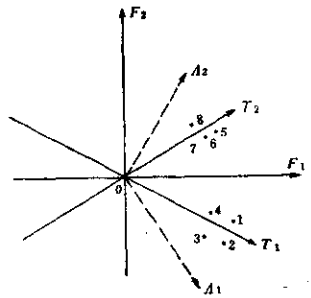


图6-6 Λ，T，F之间的关系

现在来看T与Λ的关系，知道其中一个，如何求另一个？

令

$$\mathbf{D}=\mathbf{T}'\Lambda=\begin{bmatrix} T_1' \\ T_2' \\ \vdots \\ T_m' \end{bmatrix}[\Lambda_1\ \Lambda_2\ \cdots\ \Lambda_m]$$

$$= \begin{bmatrix} T_1'\Lambda_1 & T_1'\Lambda_2 & \cdots & T_1'\Lambda_m \\ T_2'\Lambda_1 & T_2'\Lambda_2 & \cdots & T_2'\Lambda_m \\ \cdot & \cdot & & \cdot \\ T_m'\Lambda_1 & T_m'\Lambda_2 & \cdots & T_m'\Lambda_m \end{bmatrix} .$$

则D是由斜主因子T_1，T_2，…，T_m与斜参因子$\Lambda_1,\Lambda_2,\cdots,\Lambda_m$的相关系数所组成的矩阵。

设 $\Lambda^{-1}=(\mu_{ij})$ 则

$$T'=T'[\Lambda\Lambda^{-1}]=[T'\Lambda]\Lambda^{-1}$$

$$\cdot\cdot \begin{bmatrix} T_1'\Lambda_1\mu_{11} & \cdots & T_1'\Lambda_1\mu_{1m} \\ T_2'\Lambda_2\mu_{21} & \cdots & T_2'\Lambda_2\mu_{2m} \\ \vdots & & \vdots \\ T_m'\Lambda_m\mu_{m1} & \cdots & T_m'\Lambda_m\mu_{mm} \end{bmatrix}$$

因T'每一行元素平方和应为1，即

$$[T_i'\Lambda_i]^2[\mu_{i1}^2+\mu_{i2}^2+\cdots+\mu_{im}^2]=1 ,$$

故

$$T_i'\Lambda_i=\frac{1}{\sqrt{\mu_{i1}^2+\mu_{i2}^2+\cdots+\mu_{im}^2}}=\frac{1}{(\sum_{l=1}^{m}\mu_{il}^2)^{\frac{1}{2}}} .$$

由于所作斜交参考轴Λ_i垂直于斜因子轴Tl（$i\neq l$），所以D矩阵是一个对角阵，对角线上的元素为 $T_i'\Lambda_i=1\Big/(\sum_{l=1}^{m}\mu_{il}^2)^{\frac{1}{2}}$，即

$$D=T'\Lambda=\begin{bmatrix} 1/(\sum_{l=1}^{m}\mu_{1l}^2)^{\frac{1}{2}} & & & 0 \\ & 1/(\sum_{l=1}^{m}\mu_{2l}^2)^{\frac{1}{2}} & & \\ & & \ddots & \\ 0 & & & 1/(\sum_{l=1}^{m}\mu_{ml}^2)^{\frac{1}{2}} \end{bmatrix}$$

$$T'=D\Lambda^{-1}=\begin{bmatrix} 1/(\sum_{l=1}^{m}\mu_{1l}^2)^{\frac{1}{2}} & & & 0 \\ & 1/(\sum_{l=1}^{m}\mu_{2l}^2)^{\frac{1}{2}} & & \\ & & \ddots & \\ 0 & & & 1/(\sum_{l=1}^{m}\mu_{ml}^2)^{\frac{1}{2}} \end{bmatrix}\cdot\begin{bmatrix} \mu_{11} & \mu_{12} & \cdots & \mu_{1m} \\ \mu_{21} & \mu_{22} & \cdots & \mu_{2m} \\ \vdots & \vdots & & \vdots \\ \mu_{m1} & \mu_{m2} & \cdots & \mu_{mm} \end{bmatrix}$$

$$= \begin{bmatrix} \mu_{11}\Big/\left(\sum_{l=1}^{m}\mu_{1l}^2\right)^{\frac{1}{2}} & \mu_{12}\Big/\left(\sum_{l=1}^{m}\mu_{1l}^2\right)^{\frac{1}{2}} & \cdots & \mu_{1m}\Big/\left(\sum_{l=1}^{m}\mu_{1l}^2\right)^{\frac{1}{2}} \\ \mu_{21}\Big/\left(\sum_{l=1}^{m}\mu_{2l}^2\right)^{\frac{1}{2}} & \mu_{22}\Big/\left(\sum_{l=1}^{m}\mu_{2l}^2\right)^{\frac{1}{2}} & \cdots & \mu_{2m}\Big/\left(\sum_{l=1}^{m}\mu_{2l}^2\right)^{\frac{1}{2}} \\ \vdots & \vdots & & \vdots \\ \mu_{m1}\Big/\left(\sum_{l=1}^{m}\mu_{ml}^2\right)^{\frac{1}{2}} & \mu_{m2}\Big/\left(\sum_{l=1}^{m}\mu_{ml}^2\right)^{\frac{1}{2}} & \cdots & \mu_{mm}\Big/\left(\sum_{l=1}^{m}\mu_{ml}^2\right)^{\frac{1}{2}} \end{bmatrix} 。$$

T′就等于A的逆（A^{-1}）再按行正规化（长度为1）

三、斜交旋转

斜交旋转是从某一正交因子解出发，经过变换（旋转）最终求出斜主因子解，这个解应包括斜主因子模型，因子结构矩阵和斜交因子相关矩阵。其计算步骤如下：

设经方差极大旋转后正交因子模型为A

（1）将A按行正规化（长度为1）得矩阵A*。

（2）将A*的各元素绝对值k次幂（k为＞2的适当正整数），并保留其原来符号，得矩阵H。

（3）建立A*对H的最小二乘拟合，即令

$$A^*C = H \ , \tag{6-47}$$

其中C是m阶方阵，用$A^{*\prime}$去左乘（6-47）式两边，然后再用 $[A^{*\prime}A^*]^{-1}$ 左乘所得方程的两边，则可解得

$$C = [A^{*\prime}A^*]^{-1}A^{*\prime}H \ 。$$

（4）将C按列正规化，得矩阵A。

（5）将A^{-1}按行正规化得T′。

（6）据（6-43），（6-45），（6-46）式分别求得斜交因子解的相关、结构、模型矩阵：

$$L = T'T \ ,$$

$$S = AT \ ,$$

$$B = A[T']^{-1} \text{或} B = SL^{-1} \ 。$$

为了使所得的解更“理想”，可让k为2、3、4、…，依次进行旋转，比较各次的结果，k值增大到因子相关矩阵相对稳定下来时止。根据实践经验，较为适当的k值通常在2—4。一般来讲，给定的一组变量之间相关关系越复杂，k值也就越高，k值应取多大，则具体问题要具体分析。

第六节　因子计量

因子计量又称因子得分。在n个变量中提取m个有成因意义的公因子后，可以将每个公因子看成一个新的组合变量。那么这些新变量（因子）在每个样品上取值就是该样品对某个因子的因子计量，记作 $f_{jk}(j=1,2,\cdots,m;\ k=1,2,\cdots,N)$。$R$型因子分析是通过对每

个样品因子计量的计算，考察该样品主要是由何种因子确定的，在此基础上再对样品进行成因解释和分类。

在介绍因子分析的数学模型时，把变量表示为m个因子的线性组合：

$$z_j = a_{j1}F_1 + a_{j2}F_2 + \cdots + a_{jm}F_m + b_jU_j \ ,$$

从而得出主因子解、正交因子解或斜交因子解。这是因子分析一个方面的问题。此外，由于公因子能充分反映包含在原始变量中的相关信息，所以用因子取代原始变量，将更有益于描述研究对象的特征。因此，因子分析的另一个问题是将因子表示为变量的线性组合，如

$$F_k = \beta_{k1}z_1 + \beta_{k2}z_2 + \cdots + \beta_{kn}z_n \qquad (k = 1,2,\cdots,m) \ 。$$

这样对每一个样品，利用其变量的标准化值计算相应的因子F_k值，就称为样品在该因子上的因子计量（得分）。因子计量的计算公式应视因子模型而定。

一、因子计量的直接解

当运用主成分因子模型

$$\mathbf{Z} = \mathbf{A}\ \mathbf{F} \ ,$$

对包含n个变量的数据所提取的因子个数等于变量数时，可获得因子得分的直接解。

此时，由于主因子解A是一个$n \times n$阶非奇异方阵，它的逆矩阵存在，所以因子得分为

$$\mathbf{F} = \mathbf{A}^{-1}\mathbf{z} \ 。 \qquad (6\text{-}48)$$

这个解是准确的、唯一的，不包含估计成分。

二、因子得分的估计式

当因子模型包括公因子和单一因子

$$\mathbf{Z} = \mathbf{AF} + \mathbf{bU}$$

时，主因子解A矩阵的列数等于公因子数，它是一个$n \times m$阶矩阵，由于$m < n$，A矩阵不能直接求得逆矩阵，或者说方程组

$$F_k = \beta_{k1}z_1 + \beta_{k2}z_2 + \cdots + \beta_{kn}z_n \qquad (k = 1,2,\cdots,m)$$

是有n个未知数，而只有m个方程的线性方程组，它或者无解或者有无穷多组解。对于这种情况，一般采用最小二乘法意义下最佳拟合得出因子得分函数估计式，即求得因子对n个变量的回归方程

$$\begin{aligned}\hat{F}_k &= \beta_{k1}z_1 + \beta_{k2}z_2 + \cdots + \beta_{kn}z_n \\ &= \boldsymbol{\beta}_k'\mathbf{z} \qquad (k = 1,2,\cdots,m)\end{aligned} \qquad (6\text{-}49)$$

式中，$\beta_{kj}(j = 1,2,\cdots,n)$为回归系数，由于变量和主因子都是标准化了的，故没有常数项。

根据多元回归分析的理论可知，确定回归系数β_{kj}的标准化正规方程组为：

$$\begin{cases} r_{11}\beta_{k1} + r_{12}\beta_{k2} + \cdots + r_{1n}\beta_{kn} = r_{1F_k} \\ r_{21}\beta_{k1} + r_{22}\beta_{k2} + \cdots + r_{2n}\beta_{kn} = r_{2F_k} \\ \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \\ r_{n1}\beta_{k1} + r_{n2}\beta_{k2} + \cdots + r_{nn}\beta_{kn} = r_{nF_k} \end{cases} \qquad (k = 1,2,\cdots,m) \ , \qquad (6\text{-}50)$$

其中r_{il}为第 l，第 j 变量间的相关系数；r_{jF_k}为变量 j 和公因子F_k间的相关系数，即

$$r_{jF_k}=a_{jk}\ 。$$

(6-49) 式写成矩阵形式为

$$\begin{bmatrix} r_{11} & r_{12} & \cdots & r_{1n} \\ r_{21} & r_{22} & \cdots & r_{2n} \\ \vdots & \vdots & & \vdots \\ r_{n1} & r_{n2} & \cdots & r_{nn} \end{bmatrix} \begin{bmatrix} \beta_{k1} \\ \beta_{k2} \\ \vdots \\ \beta_{kn} \end{bmatrix} = \begin{bmatrix} a_{1k} \\ a_{2k} \\ \vdots \\ a_{nk} \end{bmatrix},$$

即

$$\mathbf{R}\beta_k=\mathbf{a}_k\ 。 \tag{6-51}$$

解正规方程组，得

$$\beta_k=\mathbf{R}^{-1}\mathbf{a}_k\ 。 \tag{6-52}$$

把求得的回归系数代入 (6-49) 式，便可获得在第 k 个因子上因子计量的估计式，为

$$\hat{\mathbf{F}}_k=[\mathbf{R}^{-1}\mathbf{a}_k]'\mathbf{Z}=\mathbf{a}_k'\mathbf{R}^{-1}\mathbf{Z}\ , \tag{6-53}$$

式中，$\mathbf{a}_k'$为$\mathbf{a}_k$（即因子负荷矩阵A中第 k 列元素所组成的列向量）的转置；$\mathbf{R}^{-1}$为变量相关矩阵的逆矩阵；$\mathbf{Z}$为原始数据的标准化矩阵。

依次让$k=1,2,\cdots,m$，则

$$\hat{\mathbf{F}}=\begin{bmatrix} \hat{\mathbf{F}}_1 \\ \hat{\mathbf{F}}_2 \\ \vdots \\ \hat{\mathbf{F}}_m \end{bmatrix}=\begin{bmatrix} a_1' \\ a_2' \\ \vdots \\ a_m' \end{bmatrix}\mathbf{R}^{-1}\mathbf{Z}=\mathbf{A}'\mathbf{R}^{-1}\mathbf{Z}\ , \tag{6-54}$$

这就是因子计量的估计式。

$\mathbf{A}'=[a_{ij}]_{m\times n}$，在正交旋转中，它就是旋转后的因子负荷矩阵的转置；当因子斜交时，它就是因子结构矩阵的转置。

第七节　因子分析的计算步骤与实例

一、计算步骤

将前面所讨论的计算过程归纳为如下步骤：

1．列出原始数据表；

2．将原始数据标准化；

3．求变量间的相关系数，列出相关矩阵R；

4．求R的全部特征值和特征向量；

5．确定公因子个数m；

6．计算初始因子矩阵A，并可作初步地质解释；

7．对A作正交旋转；

8．求正交因子解各因子计量；

9．对A作斜交旋转；

10．求斜交因子解各因子计量。

二、实例

在广西某地，为了研究该地区的成矿机理，在钻孔中采集14个矿石样品，每个样分析Cu、Pb、Zn、Sn、As、Mo、Co、Ni、Ag、Cr、Bi等11个元素，作因子分析。

1．原始数据，见表6-2。

2．求变量间的相关矩阵R，得表6-3。

3．求出R的11个特征值，按其大小顺序为：4.118，2.672，1.395，1.101，0.835，0.433，0.218，0.127，0.078，0.017，0.006。

表6-2 广西某矿原始数据

样号	Cu	Pb	Zn	Sn	As	Mo	Co	Ni	Ag	Cr	Bi
1	40	10	40	5	400	5	10	65	4	100	10
2	40	15	40	5	85	250	10	15	8	30	10
3	600	300	120	5	800	15	10	500	4	60	50
4	3000	650	120	350	400	5	10	500	18	30	50
5	100	450	40	150	300	5	10	400	2	30	300
6	4000	4000	160	100	1000	5	10	450	30	200	3000
7	3000	500	200	450	700	5	10	400	10	60	600
8	60	400	40	250	200	5	10	150	1	100	100
9	60	3000	3000	60	1500	5	5	150	2	80	3000
10	200	3000	3000	40	1500	5	40	150	2	60	1000
11	3000	150	60	40	300	5	40	250	2	100	10
12	300	150	60	20	100	15	10	120	2	30	10
13	600	1000	100	10	100	25	10	70	6	100	300
14	40	20	20	10	30	10	10	25	1	15	10

表6-3 相关矩阵

	Cu	Pb	Zn	Sn	As	Mo	Co	Ni	Ag	Cr	Bi
Cu	1	0.27	-0.24	0.50	0.13	-0.23	0.17	0.65	0.77	0.48	0.28
Pb		1	0.67	-0.05	0.81	-0.23	0.15	0.19	0.48	0.59	0.91
Zn			1	-0.14	0.84	-0.15	0.32	-0.16	-0.20	0.01	0.58
Sn				1	0.03	-0.24	-0.18	0.54	0.33	-0.06	-0.01
As					1	-0.29	0.24	0.28	0.17	0.32	0.75
Mo						1	-0.12	-0.37	0.03	-0.25	-0.19
Co							1	-0.06	-0.21	0.07	-0.12
Ni								1	0.53	0.17	0.18
Ag									1	0.54	0.46
Cr										1	0.59
Bi											1

4．$\lambda \geqslant 1$的4个特征值及其对应的特征向量，列于表6-4。

表6-4 特征值与特征向量

	I	II	III	IV
λ	4.12	2.67	1.39	1.10
Cu	0.29	0.41	-0.07	0.23
Pb	0.44	-0.21	0.14	-0.02
Zn	0.23	-0.48	-0.20	-0.18
Sn	0.11	0.38	-0.36	-0.37
As	0.39	-0.29	-0.13	-0.16
Mo	-0.19	-0.04	0.51	-0.02
Co	0.06	-0.19	-0.40	0.72
Ni	0.25	0.37	-0.30	-0.13
Ag	0.32	0.35	0.33	0.03
Cr	0.33	0.07	0.32	0.38
Bi	0.43	-0.16	0.24	-0.19

5．求初始矩阵A，得表6-5。

表6-5中因子负荷反映了变量与公因子之间的相关程度，公因子的地质意义是由负荷较大的元素表征的。矩阵A中各行元素平方和为公因子方差 h_j^2，它的大小表示变量的信息被公因子所提取的方差。它的值较小时（如Mo的$h_6^2=0.51$）说明Mo的变差还有近一半不能被公因子说明。A的各列平方和等于特征值。某公因子的特征值除以变量数（$n=11$）的百分比为该公因子占总方差的百分比。4个公因子的因子负荷平方和等于9.286，占总方差的84.4%，这4个公因子已经能说明全部变量方差的大部分。

表6-5 初始因子矩阵A（主因子解）

	F_1	F_2	F_3	F_4	h_j^2
Cu	0.59	0.67	-0.08	0.31	0.80
Pb	0.90	-0.34	0.16	-0.02	0.95
Zn	0.48	-0.79	-0.24	-0.19	0.95
Sn	0.23	0.62	-0.42	-0.39	0.77
As	0.79	-0.47	-0.22	-0.16	0.92
Mo	-0.39	-0.07	0.60	-0.02	0.51
Co	0.11	-0.31	-0.47	0.76	0.90
Ni	0.51	0.61	-0.36	-0.14	0.77
Ag	0.64	0.57	0.39	0.03	0.90
Cr	0.68	0.11	0.37	0.40	0.77
Bi	0.87	-0.27	0.28	-0.20	0.94
$S_i^2=\lambda_i$	4.12	2.67	1.39	1.11	9.28
方差百分比	37.45%	24.27%	12.64%	10.09%	
累积方差	37.45%	61.72%	74.36%	84.45%	

6．对初始因子矩阵A作正交旋转得表6-6。正交旋转后各行公因子差不变，但各列的平方和不再等于特征值。正交旋转后，各公因子的负荷向两极分化，有的变量的负荷接近于1，而另一些变量的负荷接近0。公因子的地质意义（变量组合）更加明确。

表6-6　正交旋转因子解

	F_1	F_2	F_3	F_4	h_j^2
Cu	-0.07	0.74	-0.56	0.17	0.90
Pb	0.86	0.46	-0.01	0.03	0.95
Sn	-0.08	0.03	-0.84	-0.22	0.77
As	0.93	0.09	-0.19	0.15	0.92
Mo	-0.27	0.04	0.56	-0.35	0.51
Co	0.09	0.01	0.06	0.95	0.90
Ni	0.04	0.33	-0.81	-0.03	0.77
Ag	0.08	0.85	-0.29	-0.29	0.90
Cr	0.29	0.82	0.07	0.14	0.77
Bi	0.84	0.46	-0.01	-0.19	0.94
Σa_{ij}^2	3.30	2.56	2.14	1.29	9.28

正交因子解说明：F_1为Pb、Zn、As、Bi的组合，该因子反映了受裂隙控制的热液矿化。F_2为Cu、Ag、Cr元素组合，根据Cr的地球化学性质在成矿过程中活动性很弱，应属典型的内矽卡岩产物，因此F_2(Cu、Ag、Cr组合)反映了本区矽卡岩矿化。F_3为Sn、Ni组合，另外也有部分Cu，反映了成岩作用中金属元素的富集。从Cu还有Pb、Zn等元素它们除出现一个因子外，还在其它因子上有较高的因子负荷，这说明本区的矿化不只一次，而是多期次多阶段的成矿作用。F_4是Co的因子，根据单矿物分析结果，Co在黄铁矿及磁黄铁矿中一般达到300ppm。故这一因子可能反映本区广泛发育的黄铁矿和磁黄铁矿化。至于Mo，注意到其公因子方差只有0.51，应该尚有F_5属于Mo的因子。

顺便指出，初始因子解与选取公因子个数m无关，但正交因子解和斜交因子解与m的大小有关。取不同的公因子数目，旋转后结果是不一样的。旋转后，公因子的方差贡献可以不作为公因子重要性的依据。

表6-7　正交因子计量

样号	1	2	3	4	5	6	7	8	9	10	11	12	13	14
F_1	-0.51	-0.71	-0.20	-0.48	-0.30	0.90	-0.16	-0.43	2.51	1.84	-0.86	-0.57	-0.41	-0.62
F_2	-0.21	-0.05	-0.50	0.23	-0.64	3.13	-0.08	-0.42	-0.63	-0.69	0.57	-0.45	0.36	-0.56
F_3	0.73	1.27	-0.53	-1.85	-0.37	-0.02	-1.95	0.09	0.29	0.18	-0.09	0.51	1.00	0.73
F_4	-0.49	-0.61	-0.13	-0.25	-0.43	-0.54	-0.23	-0.31	-0.84	2.10	2.42	-0.29	-0.25	-0.34

7．计算正交因子解的因子计量，结果见表6-7。从中看出F_1因子计量较高的样品是9号和10号样品，该两样品矿化好。F_2得分高的是6号样品。而F_3得分高的应是7号和4号样品，因为代表该因子元素组合的因子负荷是负号。当同一公因子中若同时有变量的负荷高正值和高负值存在时，则应将它们分为两个元素组合看待，但这两个元素组合间存在着负相关关系。F_4的因子得分高的样品有10号和11号

8．对初始因子矩阵A作斜交旋转，得表6-8、6-9、6-10。分别为斜交因子结构，斜交因子模型和斜交因子相关矩阵。

表6-8 斜交因子结构

	GⅠ	GⅠ	GⅡ	GⅣ
Cu	0.503	0.277	−0.316	0.264
Pb	0.296	−0.637	0.093	−0.047
Zn	−0.326	−0.876	−0.024	0.030
Sn	−0.293	−0.049	−0.795	−0.177
As	−0.108	−0.816	−0.185	0.024
Mo	0.235	0.251	0.556	−0.315
Co	0.127	0.109	0.059	0.917
Ni	−0.003	−0.030	−0.681	0.016
Ag	0.628	0.119	−0.031	−0.200
Cr	0.744	0.034	0.297	0.172
Bi	0.270	−0.858	0.095	−0.265

表6-9 斜交因子模型

	Ⅰ	Ⅱ	Ⅲ	Ⅳ
Cu	1.182	−0.074	−1.022	0.056
Pb	0.956	−1.100	−0.239	0.138
Zn	−0.049	−1.054	0.248	0.408
Sn	0.353	0.120	−1.039	−0.278
As	0.545	−1.131	−0.262	0.309
Mo	−0.253	0.367	0.633	−0.400
Co	−0.047	−0.256	0.133	1.015
Ni	0.779	−0.112	−1.137	−0.094
Ag	1.298	−0.194	−0.786	−0.398
Cr	1.158	−0.499	−0.306	0.104
Bi	0.960	−1.033	−0.257	−0.104

表6-10 斜交因子相关矩阵

	Ⅰ	Ⅱ	Ⅲ	Ⅳ
Ⅰ	1	0.492	0.621	0.214
Ⅱ		1	0.236	0.352
Ⅲ			1	0.015
Ⅳ				1

第八节 Q 型因子分析

Q 型因子分析和R型因子分析的不同点在于：R 型因子分析是研究变量之间的相互关系，而Q型因子分析方法则是研究样品之间的相互关系。因此，在进行Q型因子分析时需要把R型因子分析中的原始数据矩阵进行转置。

正如把相关系数作为变量之间线性关系的相似性度量一样，在进行Q型因子分析时，应当选择样品之间合适的相似性度量。一般除相关系数外，还有以下二种选择，即：一种由英布里(Imbrie)和珀迪(Purdy，1962)提出的比例相似系数，即夹角余弦；另一种是各种距离系数。样品和样品间的相似性矩阵记号为Q。

在Q型因子分析时，如选用相似系数或距离系数作为样品间的相似性度量时，变量的均匀化一般也选用相应的单位化和正规化（或规格化）。

在获得样品与样品之间的相似性矩阵Q后，Q型因子分析可按R型因子分析的数学方法，逐个地提取特征根，求得相应的特征向量，进而求得因子负荷矩阵。在此基础上再进行正交旋转或斜交旋转，求得正交解或斜交解。也可以分别求得它们的因子得分。

下面以一假想例子说明Q型因子分析的过程及结果。这个例子代表一个因果过程为已知的成分数据集合。假设给定10个重砂样品的矿物成分数据。样品1、2、3分别为三个端元样品，它们各自的10种变量的百分含量事先给定，然后按表6-11中给定的其它样品占端元样品组分的百分比计算每个样品的百分含量。如第7号样品变量1的值$x_{17}=5.0\times20\%+10.0\times70\%+3.0\times10\%=8.30$。结果见表6-12。

表6-11　各样品相对端元样品的百分比

端元组分样品	1	2	3	4	5	6	7	8	9	10
1	100	0	0	50	80	40	20	50	20	10
2	0	100	0	50	0	0	70	30	60	10
3	0	0	100	0	20	60	10	20	20	80

表6-12　假设数据表

变量	样品									
	1	2	3	4	5	6	7	8	9	10
x_1	5.0	10.0	3.0	7.5	4.6	3.8	8.3	6.1	7.6	3.9
x_2	25.0	30.0	6.0	27.5	21.2	13.6	26.6	22.7	24.2	10.3
x_3	15.0	17.0	10.0	16.0	14.0	12.0	15.9	14.6	15.2	11.2
x_4	5.0	17.0	13.0	11.0	6.6	9.8	14.2	10.2	13.8	12.6
x_5	5.0	8.0	25.0	6.5	9.0	17.0	9.1	9.9	10.8	21.3
x_6	20.0	8.0	15.0	14.0	19.0	17.0	11.1	15.4	11.8	14.8
x_7	10.0	5.0	13.0	7.5	10.6	11.8	6.8	9.1	7.6	11.9
x_8	5.0	4.0	8.0	4.5	5.6	6.8	4.6	5.3	5.0	7.3
x_9	5.0	1.0	5.0	3.0	5.0	5.0	2.2	3.8	2.6	4.6
x_{10}	5.0	0.0	2.0	2.5	4.4	3.2	1.2	2.9	1.4	2.1

根据表6-11，10个样品成分的三角图见图6-7。下面的计算是企图只要给出原始数据，通过Q型因子分析就能揭示上述给定的（假设已知）因果关系，即验证假设。

根据6-11的原始数据，按Q型因子分析所得结果分别列于表6-13、6-14、6-15和6-16中。

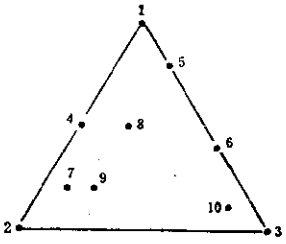


图6-7　样品成分三角图

表6-13 $\cos\theta$矩阵[$\cos\theta_{(P)}$]

样品	1	2	3
1	1.000	0.874	0.691
2	0.874	1.000	0.648
3	0.691	0.648	1.000
4	0.965	0.970	0.691
5	0.989	0.873	0.791
6	0.885	0.802	0.948
7	0.924	0.990	0.723
8	0.971	0.939	0.794
9	0.928	0.978	0.775
10	0.788	0.753	0.988

表6-14 初始因子矩阵

样品	因子1	因子2	因子3
1	0.95056	0.18806	−0.26113
2	0.93094	0.28623	0.22666
3	0.84184	−0.53438	0.07586
4	0.97147	0.23700	−0.00817
5	0.97818	0.03201	−0.20536
6	0.96057	−0.27009	−0.06612
7	0.97224	0.19711	0.12800
8	0.99604	0.07678	−0.04485
9	0.98655	0.11832	0.11272
10	0.91587	−0.39670	0.06183

表6-15 方差极大旋转矩阵

样品	因子1	因子2	因子3
1	0.66049	0.38982	0.64172
2	0.91337	0.34894	0.20867
3	0.30957	0.93134	0.19178
4	0.81794	0.38079	0.43124
5	0.62287	0.52208	0.58268
6	0.49013	0.77133	0.40602
7	0.84948	0.43093	0.30441
8	0.72961	0.51820	0.45112
9	0.80805	0.50045	0.31007
10	0.43646	0.86311	0.25406

表6-16 斜交因子结构矩阵

样品	端元组分		
	1	2	3
1	1.000	0.000	0.000
2	0.000	1.000	0.000
3	0.000	0.000	1.000
4	0.497	0.537	0.001
5	0.347	0.000	0.207
6	0.411	0.000	0.644
7	0.200	0.753	0.098
8	0.530	0.343	0.207
9	0.206	0.668	0.210
10	0.108	0.116	0.839

由最后结果（表6-16）（斜交因子结构矩阵）可以看出，各端元组分对每个样品的贡献已十分接近表6-11中给定假设百分比，但还不是完全相同，这是因为斜交因子矩阵将样品分解为标准化端元组分的贡献，而表6-11是用非标准化组分百分比。

由以上结果可以看出，因子分析能够由果及因地揭示各种成因关系。但是选用的数学模型要符合自然模型，如本例中三个端元组分存在一种负相关，用斜交因子解能较好的表征，若采用各因子间互相独立的正交因子模型，就很难解释其成因关系。

第九节 因子分析在地球化学工作中的应用

一、研究元素的共生组合

元素共生组合的研究是地球化学工作中最基本的问题之一。因子分析中，一般人认为R型分析是一种研究元素共生组合的有效手段和方法。近年来有的地球化学工作者认为因子分析不单单是研究元素共生组合的有效手段和方法，而且是一种研究元素共生组合的新的地球化学理论，正象原子的内部结构对元素结合规律所起的影响作用一样，因为在因子分析中每一个因子所包含的主要元素，不仅仅表示它们的一种组合关系，而且反映了一种内

在的成因联系。

在研究元素共生组合的基础上，因子分析还为进一步研究元素的赋存状态或存在形式提供了信息。在一般的地球化学工作中，对那些稀有分散元素、超微量元素的赋存状态和存在形式的研究是比较困难的。恰当地运用因子分析，可以为这方面的研究提供信息。

二、解决分类问题

因子分析与聚类分析一样，可用于研究对象的分类。因子分析的分类是建立在具有成因意义基础之上的，因此更有意义。

例如在化探扫面中发现了一批异常，这些异常按其成因可分成m类。我们可以在这些异常区内采集N个样品，每个样品测试n个元素所组成的原始数据进行因子分析。R型因子分析可以综合出反映m种成因的m个因子。根据某个样品在某因子上得分的高低可以将样品分成m类。Q型因子分析则直接按照因子负荷对样品进行分类。

三、划分地球化学过程，区别成矿作用

现在所见到的地质体都经历了很长的地质历史，或者经历了多个地质阶段，它们都可以看作是一些不同的地质地球化学过程的现时产物。例如某矿床的形成，它经历了岩浆作用，岩浆期后热液作用，后来的区域变质作用等不同的地质地球化学过程。如何从数学的角度分解这些地质过程，并找出其中哪个作用对于成矿处于主导地位，解决这些问题因子分析是一种十分有用的方法。

例如对陕西某超基性铜镍矿床的成矿机理研究中，在成矿地段采取了693个岩石样品，每个样品分析了Ni、Co、Cu、Cr、Ti、B、Sn、Mn、Pb、Zn、S、As等12个元素，通过R型因子分析，得斜交因子结构矩阵（表6-17）。

表6-17 陕西某矿床斜交因子结构矩阵

元 素	F_1	F_2	F_3	F_4
Ni	0.6029	-0.2808	-0.0675	-0.0047
Co	0.0221	-0.7350	0.1227	0.1822
Cu	-0.0322	-0.7822	-0.0888	-0.0228
Cr	0.7688	0.1654	-0.0018	-0.0201
Ti	-0.6248	0.1410	-0.2128	0.1015
B	0.0252	-0.0207	-0.0543	-0.7173
Sn	0.6284	0.1713	-0.2187	0.0573
Mo	-0.3338	-0.0472	-0.2040	0.4210
Pb	0.1082	0.0841	-0.0078	-0.7243
Zn	-0.0175	0.2576	0.7048	0.0950
S	-0.0471	-0.7978	-0.0682	-0.0673
As	-0.0173	-0.1005	0.7064	-0.0622

由表6-17可以看出：第一主因子（F_1）由Ni、Cr、Sn组成，代表岩浆作用时期的元素，表示一种基性、超基性的元素组合特点，而Ni为本矿区的主要成矿元素，所以说Ni矿化主要与超基性岩浆作用有关，从地质上看这期矿化称为A组矿化。第二公因子（F_2）由Co、Cu、S组成，是后期硫化物阶段的产物，从地质上看Cu、Co的矿化属于硫化物化，即B组矿化。第三公因子（F_3）由As和Zn两元素组成，是典型的热液作用产物，反映了热

液对矿化的改造。第四公因子（F_4）由B、Pb元素组成，为蛇纹岩化的产物，与矿化关系不大。

第十节　因子分析的若干问题和讨论

一、应用因子分析的前提

因子分析的应用必须以下列假设为前提：

1．变量和因子之间存在线性关系。变量和因子之间的线性关系又必须以原始变量之间存在不同程度的相关关系为基础，因子分析是研究相关矩阵内部结构的一种多元统计分析方法，如果原始变量之间不存在相关关系，就不能把n个变量的信息压缩成$m(m<n)$个公因子，也就没有必要作因子分析了。

2．对于任意一已知变量、各个因子的作用应具有加和性、因子之间无交互作用（但可以有相关关系）。另一种说法是，某一因子仅当其它因子的作用达到某种程度后才被"允许"起作用。在因子分析中，因子模型为

$$z_i = a_{i1}F_1 + a_{i2}F_2 + \cdots + a_{im}F_m + b_iU_i \ ,$$

它是一个线性的数学模型，各个因子解都是由它推导出来的。该模型决定了各因子之间应服从加和性的条件。

应当指出，应用因子分析来解决实际问题时，有的自然模型具有加和性，而有些问题并不满足加和性。在这种情况下，对因子分析的应用和解释都必须慎重。

二、因子解的地质解释

因子分析在研究地质成因问题上是一种非常有用的数学方法，它由果及因来反推地质过程，讨论地质成因问题，为地质学家运用自己广博的地质地球化学知识和丰富的经验的机会要比别的统计方法多。但是正由于在因子分析中存在着客观的计算结果同实际工作者的主观思维之间相互影响这一特点，在使用因子分析方法时必须特别慎重。

其它多元统计分析（如判别分析，回归分析）的计算结果基本上是一个最终结果，可以直接予以应用，而因子分析的计算结果（因子解）只能看作是一个中间结果，剩下的部分需要人们用自己的思维来完成，这就涉及地质地球化学知识、经验，甚至于思维方式和哲学思想，所以，因子分析是一个客观计算同主观思维相结合的过程。

因子分析所采用的主要是一套坐标旋转、空间变换的数学技术，其目的在于选择若干个观察地质数据的"最好方向"，以便在这几个方向上最清楚地识别地质数据所展示的成因意义。因子分析中初始因子（或主成分）、正交因子解、斜交因子解都可以用于地质解释，但究竟哪个因子解用作地质解最好，需要具体分析，也就是说要看哪个解更符合自然模型，例如，第七节中的一个利用假设数据来研究沉积盆地沉积物来源的例子。该例子是一个已知成因模型的Q型因子分析，从所得的计算结果看，其初始因子解和正交因子解都不理想，斜交因子解却非常好地解释了各个因子，这是因为初始因子解和正交因子解的数学模型不符合自然模型，它们都是以各因子间互相独立为条件的，但实际上，该例假设沉积盆地中每个样品的物质来源是由三个剥蚀区供给的，组成每样品的原始数据是一种成分数

据（定和数据），成分数据各因素之间存在一种负相关（某一成分增加，其主成分必然减少），因此斜交因子模型（它允许因子间有相关关系）才符合自然模型，斜交因子解的解释才是正确的。

在实际工作中，自然模型往往是未知的。这给因子分析结果的地质解释带来困难。

因子分析中另一个需要注意的问题是，参加分析的原始数据（包括样品的代表性和变量的选取）是否包含指示成因的信息，如果没有包含有关的信息，无论采用何种模型都不会得到满意的结果。

总的来看，因子分析的结果将启发我们得出许多有用的地质和成因方面的设想，具体有以下三方面：

1．形成科学的成因假设。当对象的地质研究程度很差、对于成因尚不清楚的情况下，因子分析的结果将活跃研究者的思维，启发他们形成地质成因概念，但这种概念有待进一步工作的验证。

2．检验已有的成因观点。当对象的地质研究程度较高，已有了一种或若干种地质成因模式的情况下，因子分析的结果可以旁证研究者的成因观点是否合理，并帮助他们修订出更合理的成因模式。修正后的成因模式仍需进一步工作予以确证。

3．确立地质成因模式。当由因子分析所提出的地质成因模式已被其它更可靠的直接手段所确证时，则可将该模式确立为一个适合于使用的成因模式。这时因子分析的结果已不是地质假设，而是地质结论，可以用这个因子模型来解决其他地质问题。

三、关于公因子的重要程度

公因子的方差贡献（或者说特征值）的大小是衡量某公因子重要程度的数学标准。在进行实际工作的时候，这种数学标准和真实情况可能会不一致。原因有以下几个方面：

1．因子分析的数学模型中，变量或因子的重要程度都是以其方差大小来衡量的。方差大小与变量平均值的大小是两回事。也就是说，因子分析只强调变量的离差（变化量）而不强调变量在样品中的比重（百分含量）。

2．在实际工作中，常常遇到这样的情况，有的变量虽然方差很小，但却有很强的成因指示意义，而另一些变量虽然方差很大却很少成因指示意义。

3．无论是方差还是均值的大小不但与地质意义有一定的差别，与经济价值之间也有不同。地质找矿是寻找有经济意义的地质体，很多变量的方差不能指示经济价值。

4．正交旋转和斜交旋转后，公因子的方差与初始因子的方差是不同的。正交旋转和斜交旋转的结果与所选定公因子的个数 m 有关。为了避免取不同的 m 值使旋转结果不同，建议用 n 代替 m，即选取公因子的个数与变量数相同，先作旋转，然后在地质解释时只取前 m 个公因子。

这样看来，某些方差贡献小的因子只要它具有明确的地质意义或重要的经济意义，就应当把它列入因子解中。

第七章　对应分析和非线性映射

第一节　什么是对应分析

对应分析又称R-Q型因子分析，是由法国Benzeci于1970年首次提出，在因子分析的基础上发展起来的，因此，也可以说它属于广义因子分析中的一个分支。

在地质和化探数据的统计分析中，经常要处理三种关系，即变量之间的关系，样品之间的关系以及样品和变量之间的关系。第六章因子分析中提出了R型分析和Q型分析，它们分别可以用来研究变量之间和样品之间的关系，两种因子分析通常是分别进行，甚至只作R型因子分析，不作Q型因子分析，或者只作Q型因子分析，不作R型因子分析，这样人为地把R型和Q型分析割裂开来，结果漏掉了许多有用的信息。地质成因问题是通过不同特征的样品表现出来的，因此，成因与样品是有联系的。R型分析和Q型分析之间不可分割，有一种对应关系。

另外，一般来说原始数据矩阵中，样品的数目远远超过变量数目，这样就给Q型分析的计算带来极大的困难。如果有100个样品，每个样品测定10个指标，这样R型分析只要计算一个（10×10）阶相关矩阵的特征值和特征向量，而进行Q型分析就要计算阶数大得多的矩阵的特征值和特征向量，这给计算带来极大的困难。于是人们就设法从R型分析的结果推导Q型分析的特征值和特征向量。

对应分析就是把R型分析与Q型分析统一起来，把变量和样品同时反映到相同坐标轴（因子轴）的一张图上，这样就便于地质解释与推断。对应分析点聚图中，变量点群表示了具有同一地质作用的元素组合，样品点群表示了相同类型的样品，而且样品点群的地质成因是由它们邻近的变量所表征。这就有助于对样品类型的地质解释，同时通过样品在空间的分布可以了解地质过程的空间关系。

第二节　对应分析的数据变换方法

假定有N个样品，每个样品有n个变量，它们组成了一个$n\times N$的原始数据矩阵。

首先对原始数据进行概率化标度，令

$$p_{ij}=x_{ij}/T \qquad (i=1,\ 2,\ \cdots,\ n;\ j=1,\ 2,\ \cdots,\ N), \tag{7-1}$$

即：

$$\mathbf{P}=\frac{1}{T}\mathbf{X}=\begin{pmatrix} p_{11} & p_{12} & \cdots & p_{1N} \\ p_{21} & p_{22} & \cdots & p_{2N} \\ \cdot & \cdot & \cdots & \cdot \\ p_{n1} & p_{n2} & \cdots & p_{nN} \end{pmatrix}.$$

其中

$$T=\sum_{i=1}^{n}\sum_{j=1}^{N}x_{ij}\ 。$$

为了使R型分析结果与Q型分析结果能够表示在同一张点聚图上，记

$$Z_{ij}=\frac{p_{ij}-p_{i\cdot}p_{\cdot j}}{\sqrt{p_{i\cdot}p_{\cdot j}}}\ ,\qquad (7\text{-}2)$$

其中

$$p_{i\cdot}=\sum_{j=1}^{N}p_{ij}\ ,$$

$$p_{\cdot j}=\sum_{i=1}^{n}p_{ij}\ ,\qquad (i=1,\ 2,\ \cdots,\ n;\ j=1,\ 2,\ \cdots,\ N)\ 。$$

从$[Z_{ij}]_n x_N$矩阵出发可以分别计算各变量间的协方差矩阵 **A** 和各样品间的协方差矩阵 **B**：

$$\mathbf{A}=\mathbf{ZZ}'\ ,\qquad (7\text{-}3)$$

$$\mathbf{B}=\mathbf{Z}'\mathbf{Z}\ 。\qquad (7\text{-}4)$$

第三节　对偶定理

设λ_k是矩阵$\mathbf{A}=\mathbf{ZZ}'$的第k个非0特征值，$\mathbf{U}_k$是相应的特征向量，则有

$$\mathbf{ZZ}'\mathbf{U}_k=\lambda_k\mathbf{U}_k\ 。\qquad (7\text{-}5)$$

两边同乘$\mathbf{Z}'$，得

$$\mathbf{Z}'\mathbf{ZZ}'\mathbf{U}_k=\lambda_k\mathbf{Z}'\mathbf{U}_k\ 。\qquad (7\text{-}6)$$

由（7-4）式，$\mathbf{B}=\mathbf{Z}'\mathbf{Z}$，则

$$\mathbf{BZ}'\mathbf{U}_k=\lambda_k\mathbf{Z}'\mathbf{U}_k\ ,$$

记 $\mathbf{V}_k=\mathbf{Z}'\mathbf{U}_k$，则：

$$\mathbf{BV}_k=\lambda_k\mathbf{V}_k\ 。\qquad (7\text{-}8)$$

这就是说，λ_k也是$\mathbf{B}=\mathbf{Z}'\mathbf{Z}$矩阵的第$k$个非0特征值，且其对应的特征向量$\mathbf{V}_k=\mathbf{Z}'\mathbf{U}_k$。

反之，设λ_l是矩阵$\mathbf{B}=\mathbf{Z}'\mathbf{Z}$的第$l$个特征值，$\mathbf{V}_k$是相应的特征向量。同样可以得到，$\lambda_l$是矩阵$\mathbf{A}=\mathbf{ZZ}'$的非0特征值，并且其相应的特征向量为$\mathbf{U}_k=\mathbf{ZV}_k$。

由此证明矩阵$\mathbf{A}=\mathbf{ZZ}'$与矩阵$\mathbf{B}=\mathbf{Z}'\mathbf{Z}$具有相同的特征值。若**U**是**A**的特征向量，则$\mathbf{Z}'\mathbf{U}$是**B**的特征向量；如**V**是**B**的特征向量，则**ZU**是**A**的特征向量。这就是对偶性定理。

由于**A**与**B**具有相同的特征值，这就建立了R型分析与Q型分析之间的联系。借助这个关系可从R型因子分析出发，直接获得Q型因子分析的结果，从而克服了由于样品容量大而造成进行Q型因子分析时计算上的困难。

如果$\mathbf{U}_k$是矩阵$\mathbf{A}=\mathbf{ZZ}'$的单位化特征向量，那么$\mathbf{V}_k=\mathbf{Z}'\mathbf{U}_k$也就不会是单位向量；反之，若$\mathbf{V}_k$是单位特征向量，则$\mathbf{U}_k=\mathbf{ZV}_k$不再是单位向量。在因子分析中求因子负荷矩阵时，特征向量是被单位化的。为此，令

$$\mathbf{V}_k^* = \frac{1}{\sqrt{\lambda_k}}\mathbf{Z}'\mathbf{U}_k \ , \tag{7-9}$$

由于$\mathbf{V}_k = \mathbf{Z}'\mathbf{U}_k$是B的特征向量，所以$\mathbf{V}_k^*$也是B的特征向量，而且

$$\begin{aligned}\mathbf{V}_k^{*\prime}\,\mathbf{V}_k^* &= \frac{1}{\sqrt{\lambda_k}}\mathbf{U}_k'\mathbf{Z}\cdot\frac{1}{\sqrt{\lambda_k}}\mathbf{Z}'\mathbf{U}_k \\ &= \frac{1}{\lambda_k}\mathbf{U}_k'\mathbf{Z}\mathbf{Z}'\mathbf{U}_k \\ &= \mathbf{U}_k'\mathbf{U}_k = 1 \ 。\end{aligned}$$

（由7-5式$\mathbf{Z}\mathbf{Z}'\mathbf{U}_k = \lambda_k\mathbf{U}_k$，所以$\mathbf{Z}\mathbf{Z}' = \lambda_k$）这就是说，$\mathbf{V}_k^*$才是B的单位特征向量。

同理，如果$\mathbf{V}_k$是B的单位特征向量，则

$$\mathbf{U}_k^* = \frac{1}{\sqrt{\lambda_k}}\mathbf{Z}\mathbf{V}_k \tag{7-10}$$

是A的单位特征向量。

综上所述，由于A与B具有相同的特征值，关系式

$$\mathbf{V}_k^* = \frac{1}{\sqrt{\lambda_k}}\mathbf{Z}'\mathbf{U}_k \text{ 与 } \mathbf{U}_k^* = \frac{1}{\sqrt{\lambda_k}}\mathbf{Z}\mathbf{V}_k$$

称为变量空间与样品空间的对偶性（或二重性）。

第四节　对应分析因子解

一、因子负荷的计算

在R型因子分析中，设$\mathbf{U}_k$为$\mathbf{A}-\mathbf{Z}\mathbf{Z}'$矩阵第$k$个特征值相应的单位特征向量，于是第$j$个变量在第$k$个因子轴($F_k$)上的因子负荷$f_{jk}$为

$$f_{jk} = u_{jk}\sqrt{\lambda_k} \qquad (j=1,\ 2,\ \cdots,\ n;\ k=1,\ 2,\ \cdots,\ m)\ , \tag{7-11}$$

从而对应于$\mathbf{B}=\mathbf{Z}'\mathbf{Z}$矩阵的第$k$个单位特征向量$\mathbf{V}_k^*$，于是第$i$个样品在第$k$个因子轴($G_k$)上的因子负荷（$g_{ik}$）为

$$\begin{aligned}\mathbf{g}_k &= \mathbf{V}_k^*\sqrt{\lambda_k} = \left(\frac{1}{\sqrt{\lambda_k}}\mathbf{Z}'\mathbf{U}_k\right)\sqrt{\lambda_k} = \mathbf{Z}'\mathbf{U}_k \\ &= [\mathbf{Z}_1\mathbf{U}_k,\ \mathbf{Z}_2\mathbf{U}_k,\ \cdots,\ \mathbf{Z}_N\mathbf{U}_k] \qquad (j=1,\ 2,\ \cdots,\ n)\ ,\end{aligned} \tag{7-12}$$

即

$$g_{ik} = \sum_{j=1}^{n} z_{ij}u_{jk} \qquad (i=1,\ 2,\ \cdots,\ N;\ k=1,\ 2,\ \cdots,\ m)\ 。 \tag{7-13}$$

这样，由（7-11），（7-13）式可以求得公因子F_k对于变量z_j的因子负荷f_{jk}和公因子G_k对于样品z_i的因子负荷g_{ik}。

二、对应分析因子负荷点聚图

由因子分析可知，每个特征值表示与其对应因子的方差贡献。又因为A与B具有相同

的特征值，因此在变量空间中第k个因子与样品空间第k个因子（$k=1, 2, \cdots, m$）提供了相同的方差贡献。从几何意义上看，在变量空间和样品空间中的因子轴具有相同单位长度。所以，当求出各变量及各样品的因子负荷以后，可用同一因子轴表示变量和样品，即把R型与Q型因子分析的结果同时反映在同一因子平面上。

如λ_1和λ_2分别表示$\mathbf{A}=\mathbf{ZZ}'$的最大和次大特征值，也是$\mathbf{B}=\mathbf{Z}'\mathbf{Z}$的最大和次大特征值，那么$F_1$和$F_2$表示R型因子分析的两个公因子轴，而且也是Q型因子分析的两个公因子轴G_1，G_2。相应地R型因子负荷向量f_{i1}和f_{i2}（$j=1, 2, \cdots n,$）可以作为j变量的坐标值，在F_1和F_2坐标系中得到变量j的投影点；g_{i1}和g_{i2}（$j=1, 2, \cdots, N$）可以作为i样品的坐标值，在同一个坐标系中得样品i的投影点。这样把R型分析和Q型分析的结果表示在同一张点聚图上。

三、对应分析结果的解释

对于上述点聚图上的样品点群和变量点群进行直观的分析和研究之后，可以得到下列信息：

1．同一样品点群可以解释为同一地质过程所产生的结果或归属于同一类型的样品。

2．同一变量点群表示变量之间具有相同成因的相关关系。

3．一个样品点群可以用靠近这个点群的变量点群来表征，有助于鉴定类别及成因。

第五节　对应分析计算步骤

设有N个样品，每个样品测得n个变量，其原始数据矩阵为

$$\mathbf{X}=\begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1N} \\ x_{21} & x_{22} & \cdots & x_{2N} \\ \vdots & \vdots & & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nN} \end{bmatrix}。$$

1．进行数据变换

（1）将$\mathbf{X}$按行、按列分别求和（表7-1）。

表7-1　对应分析数据变换

变量 \ 样品	x_1	x_2	$\cdots$	x_N	$x_{i\cdot}=\sum_{j=1}^{N}x_{ij}$
1	x_{11}	x_{12}	$\cdots$	x_{1N}	$x_{1\cdot}$
2	x_{21}	x_{22}	$\cdots$	x_{2N}	$x_{2\cdot}$
$\vdots$	$\vdots$	$\vdots$		$\vdots$	$\vdots$
n	x_{n1}	x_{n2}	$\cdots$	x_{nN}	$x_{n\cdot}$
$\sum_{i=1}^{n}x_{ij}=x_{\cdot j}$	$x_{\cdot 1}$	$x_{\cdot 2}$	$\cdots$	$x_{\cdot N}$	$T=\sum_{i=1}^{n}\sum_{j=1}^{N}x_{ij}$

（2）计算矩阵

$$\mathbf{Z}=[z_{ij}]\ ,$$

其中

$$z_{ij}=\frac{x_{ij}-x_{i\cdot}x_{\cdot j}/T}{\sqrt{x_{i\cdot}x_{\cdot j}}}\ 。\qquad (7\text{-}14)$$

2．R型因子分析

计算矩阵$\mathbf{A}=\mathbf{ZZ}'$的特征值$\lambda_1\geqslant\lambda_2\geqslant\cdots\geqslant\lambda_m$及相应的特征向量$u_1$，$u_2$，…，$u_m$，进而求得R型因子负荷矩阵（表7-2）。

表7-2　R型因子负荷矩阵

变量 \ 因子	F_1	F_2	…	F_m
x_1	$u_{11}\sqrt{\lambda_1}$	$u_{12}\sqrt{\lambda_2}$	…	$u_{1m}\sqrt{\lambda_m}$
x_2	$u_{21}\sqrt{\lambda_1}$	$u_{22}\sqrt{\lambda_2}$	…	$u_{2m}\sqrt{\lambda_m}$
⋮	⋮	⋮	…	⋮
x_m	$u_{m1}\sqrt{\lambda_1}$	$u_{m2}\sqrt{\lambda_2}$	…	$u_{mm}\sqrt{\lambda_m}$

3．Q型因子分析

根据R型因子分析时所得的m个特征值$\lambda_1\geqslant\lambda_2\geqslant\cdots\geqslant\lambda_m$和相应的特征向量$u_1$，$u_2$，…，$u_m$求得Q型因子负荷矩阵（表7-3）。

表7-3　Q型因子负荷矩阵

样品 \ 因子	G_1	G_2	…	G_m
1	$\sum_{j=1}^{n}z_{j1}u_{j1}$	$\sum_{j=1}^{n}z_{j1}u_{j2}$	…	$\sum_{j=1}^{n}z_{j1}u_{jm}$
2	$\sum_{j=1}^{n}z_{j2}u_{j1}$	$\sum_{j=1}^{n}z_{j2}u_{j2}$	…	$\sum_{j=1}^{n}z_{j2}u_{jm}$
⋮	⋮	⋮	…	…
N	$\sum_{j=1}^{n}z_{jN}u_{j1}$	$\sum_{j=1}^{n}z_{jN}u_{j2}$	…	$\sum_{j=1}^{n}z_{jN}u_{jm}$

4．作点聚图进行地质解释和推断

根据R型和Q型因子负荷矩阵，在两个因子轴组成的平面上作变量和样品的点聚图，并进行成因解释和分类。

第六节　应用实例

本例是应用对应分析对我国主要岩浆岩的分类特点和主要的化学成分作一些验证和探讨。数据引自《简明地球化学手册》（表7-4）。

由电子计算机计算得对应分析的结果列于表7-5和表7-6中。

从对应分析结果可以看出，第一主因子方差贡献达到56%，前两个主因子的累积贡献

表7-4 中国主要岩浆岩种类的平均化学成分

岩类	SiO_2	TiO_2	Al_2O_3	Fe_2O_3	FeO	MnO	MgO	CaO	Na_2O	K_2O	H_2O^+	P_2O_5
酸性岩	71.27	0.25	14.25	1.24	1.62	0.08	0.80	1.62	3.79	4.03	0.56	0.16
	72.06	0.29	13.49	1.81	2.11	0.08	0.28	0.75	3.74	4.55	0.63	0.10
	72.38	0.32	14.15	0.84	1.43	0.05	0.69	1.66	3.77	2.61	0.82	0.10
	77.08	0.15	11.98	1.26	0.68	0.05	0.63	0.77	5.42	1.15	0.73	0.10
	64.98	0.52	16.33	1.89	2.49	0.09	1.94	3.70	3.67	2.95	0.83	0.32
	65.70	0.65	15.24	2.88	1.56	0.10	1.57	4.00	3.13	2.83	2.18	0.16
	65.74	0.75	15.89	1.87	2.52	0.13	1.64	3.27	3.29	3.67	0.69	0.20
中性岩	57.39	0.89	16.42	3.10	4.13	0.18	3.77	5.58	4.26	2.57	0.89	0.37
	60.51	0.73	16.70	2.84	3.49	0.14	2.54	4.63	3.88	2.65	0.88	0.46
	56.75	0.76	18.09	3.58	3.26	0.15	3.42	6.97	3.07	2.01	0.79	0.49
碱性岩	63.30	0.57	17.52	2.79	1.15	0.29	0.86	1.43	4.07	7.48	0.41	0.13
	71.41	0.22	14.37	1.22	1.75	0.07	0.96	1.04	3.90	4.34	0.66	0.06
	57.92	0.86	17.50	3.17	2.57	0.09	1.76	5.51	5.51	4.49	0.32	0.31
	64.33	0.51	18.03	2.49	3.17	0.16	0.31	1.30	5.43	5.35	0.29	0.09
基性岩	47.62	1.67	14.52	4.09	9.37	0.22	6.47	8.75	2.07	1.18	2.02	0.46
	48.28	2.21	14.99	4.18	6.95	0.20	7.00	8.07	3.40	2.51	1.26	0.60
超基性岩	45.05	1.00	4.79	4.28	8.14	0.31	21.61	10.78	0.97	0.87	2.28	0.16
	40.38	1.15	6.21	5.76	8.64	0.20	25.53	6.06	0.69	0.34	4.97	0.07
	39.39	0.06	0.47	2.96	5.42	0.20	46.49	0.22	0.64	0.09	3.66	0.05

表7-5 R型分析因子负荷矩阵

变量	F_1	F_2	F_3	F_4
SiO_2	−4.0451	−0.3145	−0.5394	−0.4265
TiO_2	2.4243	3.1701	−0.0823	−0.7270
Al_2O_3	−2.7956	2.9834	0.1292	−0.2716
Fe_2O_3	3.6881	1.3417	0.6740	−0.9271
FeO	3.9303	0.8629	−0.0923	−0.3315
MnO	3.0025	0.8080	2.5612	0.8406
MgO	3.3307	−2.2962	0.1967	0.8868
CaO	2.8524	2.6719	−0.5881	−0.0544
Na_2O	−3.3175	1.5713	0.0282	−0.2171
K_2O	−3.0116	0.9303	2.6607	−0.3572
H_2O^+	3.5481	−1.7420	−0.3590	−1.1990
P_2O_5	0.8661	3.0426	−1.1835	1.2622
方差贡献%	56.54	25.14	7.56	2.99
累积方差贡献%	56.54	81.68	89.24	92.23

百分比达到81.68%，可认为这两个主因子就能说明原始数据大部信息。由这两个主因子所组成的平面作出点聚图（见图7-1）可以用来解释变量和样品、变量和变量、样品和样品之间的成因联系和分类。由于图中邻近的变量点表示它们之间紧密关联和相同的地球化学意义，邻近的样品点具有相似的属性，属于相同的岩石类型；同一区域内的变量与样品点之间，变量指示着邻近样品的地质地球化学意义，而样品点则反过来可以看成某种地质作用下的产物。

表7-6 Q型分析因子负荷矩阵

样 品	G_1	G_2	G_3	G_4
1	−2.3137	−0.8255	−0.1964	0.1279
2	−2.2219	−1.0136	0.1748	−0.3190
3	−2.2829	−1.1918	−0.8504	−0.0654
4	−2.6527	−1.4473	−1.3802	−0.0732
5	−1.2613	0.2719	−0.7205	0.3120
6	−0.3665	−0.3699	−0.4038	−0.8925
7	−1.1783	0.1019	0.0744	0.0355
8	0.1531	1.3506	−0.0466	0.3455
9	−0.2793	1.2106	−0.5276	0.5752
10	0.3894	1.7403	−0.8662	0.6305
11	−1.6292	0.4382	3.1485	0.1975
12	−2.4944	−1.1945	0.0167	−0.2547
13	−1.2432	1.5171	0.0386	−0.2348
14	−1.9604	0.0852	1.3284	−0.3417
15	3.1417	2.3358	−0.4181	−0.1809
16	2.4835	3.2985	−0.2635	0.1061
17	4.7285	−0.3522	0.5305	0.4746
18	5.3584	−1.4946	0.0425	−1.6251
19	3.8295	−4.4613	0.1190	−1.1825

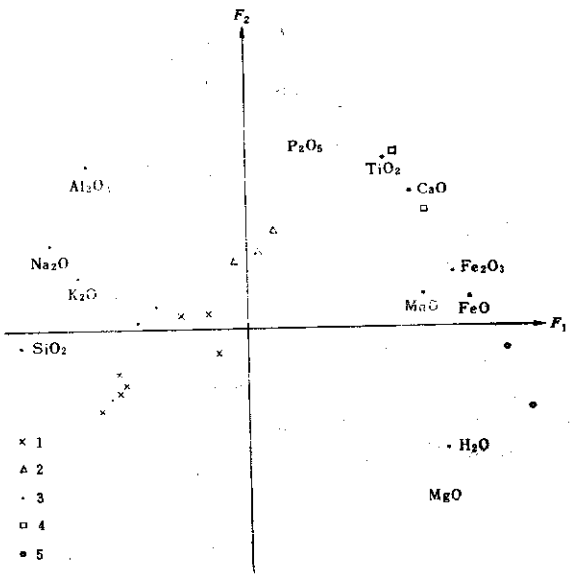


图7-1 对应分析（因子轴1—2）点聚图

1—酸性岩；2—中性岩；3—碱性岩；4—基性岩；5—超基性岩

从以上一般原则来解释图7-1的结果，可得到如下几点认识：

1．酸性岩区　全部7个酸性岩样品都落在该区内，它们的化学成分以SiO_2，Na_2O，K_2O和Al_2O_3为特征，特别是SiO_2的高绝对值有效地反映了样品的这一性质。

2．中性岩区　3个中性岩样品点落在一个较小的范围内，以第一主因子轴Q型因子负荷接近0为特征。

3．基性岩区　2个基性岩样品集中在第一象限，以第二因子的正因子负荷为特征。

4．超基性岩区　3个超基性岩样品落在最右侧，以第一主因子的高正负荷和第二主因子负的负荷为特征。

5．碱性岩区　4个碱性岩样品点与酸性岩样品有一定的重叠，但碱性岩样品与K_2O的关系更为密切。在第二主因子轴与第三主因子轴所组成的平面内（或其它因子轴平面内）可以把它们清楚地分开。

6．总结以上可以看出，第一主因子轴是岩石基性程度的综合指示，它的化学成分从负值到正值就依次排列顺序为：SiO_2，Na_2O，K_2O，Al_2O_3，P_2O_5，TiO_2，CaO，MnO，MgO，H_2O^+，Fe_2O_3，FeO。

7．第二主因子轴能进一步区分基性岩与超基性岩。基性岩以CaO，P_2O_5，TiO_2，Al_2O_3的正值为特征，超基性岩以MgO，H_2O^+的高绝对值为特征。

8．第一主因子和第二主因子不能说明的部分，可以用第三主因子加以说明，它将酸性岩和碱性岩进一步加以区分。

以上的总结，不能说概括了对应分析的全部结果。一个好的对应分析解，就象一张地质图一样，它所包含的地球化学内容十分丰富，一个地球化学工作者可以根据自己的知识、经验和对工作区研究、认识程度得到不同的有用信息。

第七节　非线性映射简介

一、什么是非线性映射

非线性映射是一种几何图象降维的数学方法，前面所讲的因子分析也是一种降维方法，而那是将高维变量综合为少数几个综合因子（原始变量的线性组合），使这些因子能再现原始变量间的相关关系，最大限度地表征原来多个指标的信息。非线性映射是通过某种非线性变换，把高维空间的几何图象变换成低维（一维、二维或三维）空间中的图象，要求变换后仍能近似地保持原来的几何关系。这种方法直观形象，使我们能够在低维空间中看到一些高维样品点相互关系的近似图象。

设有N个样品，每个样品包含p个变量，则每一个样品点就相当于R^p空间中的一个p维点：

$$X_i=\{x_{i1},\ x_{i2},\ \cdots,\ x_{ip}\}\qquad(i=1,\ 2,\ \cdots,\ N)\ 。$$

现在要将R^p空间中的N个$X_i(i=1,\ 2,\ \cdots,\ N)$映射到低维空间$R^l(l<p)$中，即通过非线性映射后，把$R^p$中$N$个样品点变为$R^l$空间中的$N$个点：

$$Y_i=\{y_{i1},\ y_{i2},\ \cdots,\ y_{il}\}\qquad(i=1,\ 2,\ \cdots,\ N)\ 。$$

要求经过映射后R^l空间中N个点Y_i间的距离仍然近似于R^p空间中N个样品点X_i间的距

离。

要达到上述目的，引入非线性映射判据（由高维变换到低维的约束条件）：

$$K=\frac{1}{\sum_{i<j} d_{ij}^{*}}\sum_{i<j}\frac{(d_{ij}^{*}-d_{ij})^{2}}{d_{ij}^{*}}=\frac{1}{NF}\sum_{i<j} w_{ij}(d_{ij}^{*}-d_{ij})^{2}\ , \qquad (7\text{-}15)$$

其中

$$NF=\sum_{i<j} d_{ij}^{*}=\sum_{i=1}^{N-1}\sum_{j=i+1}^{N} d_{ij}^{*}\ ,\ \text{称为标准化因子；}$$

$$w_{ij}=\frac{1}{d_{ij}^{*}}\ ,\ \text{为权系数；}$$

d_{ij}^{*}为原空间R^{p}中点x_i与x_j的距离；

d_{ij}为新空间R^{l}中点Y_i与Y_j的距离。

K的含义是使原空间与新空间距离之差平方和达到极小时求得新空间的几何构形。

二、计算方法

1．给出原始数据矩阵

$$\mathbf{X}=\begin{bmatrix} X_1 & X_2 & \cdots & X_N \\ x_{11} & x_{12} & \cdots & x_{1N} \\ x_{21} & x_{22} & \cdots & x_{2N} \\ \vdots & \vdots & & \vdots \\ x_{p1} & x_{p2} & \cdots & x_{pN} \end{bmatrix}$$

2．计算R^{p}空间的距离系数矩阵$\mathbf{D}^{*}$

$$d_{ij}^{*}=\sqrt{\sum_{k=1}^{p}(x_{ki}-x_{kj})^{2}}\qquad (i,\ j=1,\ 2,\ \cdots,\ p)\ , \qquad (7\text{-}16)$$

得

$$\mathbf{D}^{*}=\begin{bmatrix} d_{12}^{*} & d_{13}^{*}\cdots & d_{1N}^{*} \\ & d_{23}^{*}\cdots & d_{2N}^{*} \\ & \ddots & \vdots \\ & & d_{N-1N} \end{bmatrix}。$$

3．求R^{l}空间中N个点的坐标$Y_i(i=1,\ 2,\ \cdots,\ N)$

任取R^{l}空间中的N个初值点：

$$Y_1=\{y_{11},\ y_{21},\ \cdots,\ y_{l1}\}\ ,$$

$$Y_2=\{y_{12},\ y_{22},\ \cdots,\ y_{l2}\}\ ,$$

$$\cdots\cdots\cdots\cdots\cdots\cdots$$

$$Y_N=\{y_{1N},\ y_{2N},\ \cdots,\ y_{lN}\}\ 。$$

将它们代入K中，则K为$l\times N$个变量y_{ij}的函数，用迭代法求出当使K达到极小值时新空间R^{l}中的Y_i值，如果用欧氏距离则

$$d_{ij}(m)=\sqrt{\sum_{k=1}^{l}(y_{ki}(m)-y_{kj}(m))^{2}},$$

m为迭代次数，其迭代公式为

$$y_{ij}(m+1)=y_{ij}(m)-MF\Delta_{ij}(m)\ , \tag{7-17}$$

其中：MF称为魔力因子，可经验地取$MF=0.3$或0.4.

$$\Delta_{ij}(m)=\frac{\partial K(m)}{\partial y_{ij}(m)}\Big/\left|\frac{\partial^2 K(m)}{\partial (y_{ij}(m))^2}\right|\ , \tag{7-18}$$

$$\frac{\partial K(m)}{\partial y_{ij}(m)}=\frac{-2}{NF}\sum_{\substack{\alpha=1\\ \alpha\neq i}}^{N}\left[\frac{d_{i\alpha}^{*}-d_{i\alpha}}{d_{i\alpha}d_{i\alpha}^{*}}\right](y_{ij}-y_{\alpha j}) \tag{7-19}$$

$$\frac{\partial^2 K(m)}{\partial y_{ij}^2(m)}=\frac{-2}{NF}\sum_{\substack{\alpha=1\\ \alpha\neq i}}^{N}\frac{1}{d_{i\alpha}d_{i\alpha}^{*}}\left[(d_{i\alpha}^{*}-d_{i\alpha})-\frac{(y_{ij}-y_{\alpha j})^2}{d_{i\alpha}}\left(1+\frac{d_{i\alpha}^{*}-d_{i\alpha}}{d_{i\alpha}}\right)\right]\ . \tag{7-20}$$

为了便于作图，常取$l=2$，有时为减少计算时间，先对原始数据$\mathbf{X}=[x_{ij}]$进行R型因子分析，取第一、二主成分构成一个因子平面，然后把N个样品点X_1，X_2，…，X_N在此因子面上的因子得分（计量）点Y_1，Y_2，…，Y_N作为初始值进行迭代。

三、实例*

例1　原始数据为6维空间的正三角形，由9个点组成，它们均匀分布在三条边上，把它们映射到二维平面上。原始数据见表7-7。

取二维平面上的正方形为初始构形，二维平面上8个初始点均匀分布在4条边上，一个点在中心，进行迭代。迭代结果$K=1.0\times10^{-22}$时，$Y_i(i=1,\ 2,\ \cdots,\ 9)$的二维映射如表7-7和图7-2所示。

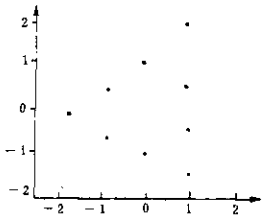


图7-2　映射图

由表7-7可作出二维空间中映射构形图(图7-2)，它仍然是二维空间中的一个正三角形，通过映射可以直观地在二维空间中看到高维空间的近似图象，以便研究点与点之间的关系。

例2　原数据为一批岩浆岩岩石化学成分，由32个样品、12个变量组成。把12维样品点映射到二维平面上，初始$Y_i(i=1,\ 2,\cdots,32)$值选取12个变量经主成分分析后32个样品点在前两个主因子平面上的因子得分点，进行迭代。迭代前用初始值计算的$K=0.397$，经迭代后的K值降为0.0255，其原始资料和映射结果见表7-8和图7-3。

从以上结果看，非线性映射将32个岩浆岩样品的12维数据点大致划分成5个簇群。Ⅰ和Ⅱ簇群为酸性岩区，Ⅰ中的酸性岩含SiO_2高、而MgO低，Ⅱ中的酸性岩相对来说含SiO_2低、而MgO高。Ⅲ簇群为中性岩区，其中喷出岩靠右而侵入岩靠左。Ⅳ簇群为基性岩区，其中喷出岩靠左而侵入岩靠右。Ⅴ簇群为超基性岩区。在5个簇群以外，碱性岩

*据中国地质科学院地球物理勘探研究所，1987。

表7-7 六维空间的正三角形和非线性映射结果

样号 \ 坐标	六维样品点						二维映射点	
	1	2	3	4	5	6	1	2
1	0.000	0.000	0.000	0.000	0.000	0.000	0.770	1.542
2	0.370	0.640	−0.429	−0.061	−0.496	0.140	−0.067	0.996
3	0.739	1.280	−0.857	−0.122	−0.993	0.280	−0.905	0.450
4	1.109	1.920	−1.286	−0.184	−1.489	0.420	−1.743	−0.095
5	1.172	1.811	−1.225	−0.612	−0.697	0.030	−0.852	−0.548
6	1.236	1.701	−1.164	−1.041	−0.116	−0.360	0.040	−1.001
7	1.299	1.591	−1.102	−1.470	0.919	−0.750	0.932	−1.454
8	0.866	1.061	−0.735	−0.980	0.612	−0.500	0.878	−0.455
9	0.433	0.530	−0.367	−0.490	0.306	−0.250	0.824	0.543

表7-8 32块火成岩标本的化学成分和非线性映象的二维结果

样品号	分析指标 / 岩性	十维												二维	
		SiO_2	TiO_2	Al_2O_3	Fe_2O_3	FeO	MnO	MgO	CaO	Na_2O	K_2O	H_2O	P_2O_5	1	2
1	花岗岩	71.27	0.25	14.25	1.24	1.62	0.08	0.80	1.61	3.79	4.03	0.56	0.16	19.47	18.90
2	黑云母花岗岩	71.99	0.21	13.81	1.37	1.72	0.12	0.81	1.55	3.42	3.81	0.64	0.20	19.98	18.83
3	二云母花岗岩	74.32	0.12	13.60	0.90	1.24	0.09	0.55	0.79	3.12	4.52	0.65	0.10	19.37	18.04
4	角闪黑云母花岗岩	67.25	0.58	14.78	1.44	3.07	0.09	1.96	4.65	3.02	2.50	0.36	0.30	20.86	19.75
5	流纹岩	72.06	0.29	13.40	1.81	2.11	0.38	0.28	0.75	3.74	4.55	0.63	0.10	19.69	18.47
6	碱流岩	70.02	0.39	12.24	2.61	3.35	0.11	0.09	0.71	5.36	4.42	0.26	0.05	18.12	19.77
7	霏细岩	70.21	0.33	15.88	1.80	1.30	0.00	0.78	1.03	2.77	4.16	1.32	0.19	20.66	17.84
8	石英斑岩	72.38	0.32	14.15	0.84	1.43	0.06	0.69	1.66	3.77	2.61	0.82	0.10	18.87	18.92
9	石英角斑岩	77.08	0.15	11.98	1.26	0.68	0.05	0.63	0.77	5.42	1.15	0.73	0.10	18.91	17.02
10	花岗闪长岩	64.98	0.52	16.33	1.89	2.49	0.09	1.64	3.70	3.67	2.95	0.83	0.32	20.19	20.14
11	英安岩	65.70	0.65	15.24	2.88	1.56	0.10	1.57	4.00	3.13	2.83	2.18	0.16	21.75	19.25
12	石英二长岩	65.74	0.73	15.89	1.87	2.52	0.13	1.64	3.27	3.29	3.67	0.69	0.20	20.42	19.71
13	闪长岩	57.39	0.89	16.42	3.10	4.15	0.18	3.77	5.58	4.25	2.57	0.89	0.37	20.82	21.66
14	石英闪长岩	60.51	0.73	16.70	2.84	3.49	0.14	2.54	4.63	3.68	2.65	0.88	0.46	20.90	21.04
15	安山岩	56.75	0.76	18.60	3.58	3.26	0.15	3.42	6.97	3.07	2.01	0.79	0.49	21.76	21186
16	黑云母角闪石安山岩	58.19	0.42	19.51	2.20	3.02	0.23	3.25	7.80	3.08	1.65	0.30	0.35	22.89	20.24
17	角闪石安山岩	59.01	1.23	17.43	3.95	3.10	0.11	4.51	6.03	2.88	1.33	0.23	0.19	22.27	21.00
18	辉石安山岩	55.25	0.52	20.22	3.76	3.28	0.03	3.51	8.12	2.57	1.34	0.80	0.59	22.63	21.85
19	正长岩	63.30	0.57	17.52	2.79	1.15	0.29	0.86	1.43	4.07	7.48	0.41	0.13	22.36	17.21
20	石英正长岩	71.41	0.22	14.37	1.22	1.76	0.07	0.96	1.04	3.90	4.34	0.66	0.06	19.21	18.57
21	辉石正长岩	57	0.85	17.50	3.17	2.67	0.09	1.76	5.51	5.51	4.49	0.32	0.31	19.12	21.35
22	粗面岩	64.33	0.51	15.03	2.49	3.17	0.16	0.31	1.30	5.43	5.35	0.29	0.09	18.33	20.51
23	辉长岩	47.62	1.67	14.52	4.09	9.37	0.22	6.47	8.75	2.97	1.18	2.02	0.46	23.37	24.02
24	角闪石辉长岩	46.84	0.32	19.82	4.67	6.87	0.26	6.83	9.47	2.45	0.29	2.04	0.14	24.99	21.56
25	苏长辉长岩	48.79	1.01	12.90	3.63	10.57	0.30	8.34	9.43	2.21	0.81	1.54	0.23	24.69	23.51
26	玄武岩	48.28	2.21	14.99	4.18	6.95	0.20	7.00	8.07	3.40	2.51	1.26	0.60	22.02	24.18
27	白榴石玄武岩	46.83	2.48	13.16	4.85	6.37	0.11	8.22	8.07	3.53	4.18	1.25	0.95	20.71	25.43
28	白榴石碧玄岩	50.10	2.39	14.15	3.80	5.59	0.12	6.01	6.84	3.54	4.98	0.54	1.04	19.64	24.86
29	粗面玄武岩	49.16	2.19	14.76	3.44	6.99	0.13	7.06	7.70	3.53	2.88	0.91	0.74	21.01	24.12
30	辉石岩	45.05	1.03	4.76	4.28	8.14	0.31	21.61	10.78	0.97	0.37	2.28	0.18	23.67	23.47
31	角闪橄榄岩	40.38	1.15	6.21	5.76	8.64	0.20	25.53	6.06	0.69	0.34	4.97	0.07	23.32	21.62
32	纯橄榄岩	39.39	0.06	0.47	2.92	5.42	0.20	46.49	0.22	0.64	0.09	3.66	0.05	28.41	17.81

（19、21、22）以及酸性岩中含碱高的碱流岩（6）和石英角斑岩（9）环绕在酸性岩区的周围，这些点以富含钠和钾为特点。另外，簇群Ⅰ中“混进”了一个石英正长岩（20），经检查这个样品的各项化学成分与花岗岩非常一致，对该样品可能岩石定名有误。

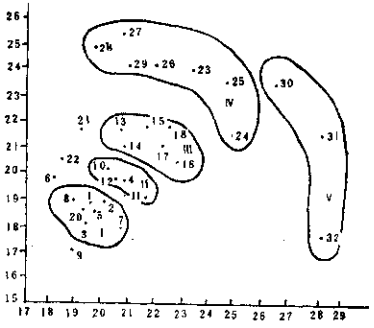


图7-3　32个岩浆岩化学成分非线性映射

从以上实例可以看出，非线性映射方法对研究多变量数据的分类和样品间的相对关系是有效的。同聚类分析相比，非线性映射可以分析较为复杂的结构，它不象聚类分析那样需要人们主观地选择某种相似性系数，从而得出不同的结果，而对这种不同的聚类结果的统一评价是困难的。另外，非线性映射得出的散点图，特别方便于人们的直观考查和分类。

当所处理的结构较为复杂时，可将非线性映射的维数增加。对于结构非常复杂、维数相当高的原始数据，当K太大时（例如$K\gg0.1$），所得出的二维或三维散点图就可能不足以刻划原始数据的固有结构。这是方法本身的一个限制。另外，由于本方法计算中需要将d_{ij}^{*}和d_{ij}保存起来，故占用机器的内存较多，对于内存为32k的计算机，大约可以处理150个样品（变量数≤30）。当样品数较大时，必须另作处理。

第八章　典型相关分析

第一节　概　述

典型相关分析是研究两组变量之间相关关系的一种统计分析方法。

假定考虑两组具有联合分布的变量x_1，x_2，…，x_k及y_1，y_2，…，y_s，要研究这两组变量之间的相关关系，一种方法是分别对每个变量进行研究，求出第一组的k个变量和第二组的s个变量之间的$k\times s$个相关系数。这样做既烦琐也不容易抓住问题的本质。另一种方法是采用典型相关分析，它类似主成分分析，在每一组变量中，都选择若干个有代表性的综合指标，通过研究两组综合指标间的关系来反映两组变量间的相关关系。说得更确切些，在第一组变量中找出一个变量线性组合，在第二组变量中也找出一个线性组合，使其具有最大的相关，然后又在每一组中找出第二个线性组合，使得在与第一个线性组合不相关的条件下，这两个线性组合之间最相关。将此程序继续进行。直到两组变量间相关被提取完毕为止。

在地球化学工作中，研究两组元素之间关系的问题是很多的。例如，研究成矿的机制时，成矿元素与其它元素之间的关系，确定控矿因素。研究超基性岩时，把铬尖晶石的化学成分作为一组变量，把铬尖晶石所在的围岩作为另一组变量，运用典型相关分析研究两组变量间的相关关系，以了解铬尖晶石形成的岩浆环境和它最终固定下来时的围岩环境。又如研究地层，可将两条区域剖面的微量元素含量作为两组变量，用典型相关分析可得出两个地区剖面的对比结果。

因为典型相关分析揭露了两个因素“集团”之间的内部联系，而两个因素“集团”的内容和变量数目又可以不同，所以它在地质上应用较广。

第二节　总体中的典型相关和典型变量

设有两个具有联合正态分布的随机向量

$$\mathbf{x}=(x_1,\ x_2,\ \cdots,\ x_k)',\tag{8-1}$$

$$\mathbf{y}=(y_1,\ y_2,\ \cdots,\ y_s)',\tag{8-2}$$

它们具有协方差矩阵Σ，把它分块写成

$$\Sigma=\begin{bmatrix}\Sigma_{xx} & \Sigma_{xy}\\ \Sigma_{yx} & \Sigma_{yy}\end{bmatrix},\tag{8-3}$$

或更简单地写成

$$\Sigma=\begin{pmatrix}\Sigma_{11} & \Sigma_{12}\\ \Sigma_{21} & \Sigma_{22}\end{pmatrix}$$

其中 Σ_{11}为第一组变量x之间的协方差矩阵；Σ_{22}为第二组变量y之间的协方差矩阵；$\Sigma_{12}=\Sigma_{21}$为第一组变量x和第二组变量y之间的协方差矩阵。假设 Σ 是正定的，从而 Σ_{11}，Σ_{12}，Σ_{22}也是正定的。

为了研究两组随机变量$\mathbf{x}=(x_1, x_2, \cdots, x_k)'$与$\mathbf{y}=(y_1, y_2, \cdots y_s)'$之间的相关关系，现考虑它们的线性组合之间的相关关系。

设有

$$u=a_1x_1+a_2x_2+\cdots+a_kx_k\ ,$$
$$v=b_1y_1+b_2y_2+\cdots+b_sy_s\ ;$$

或

$$u=(a_1a_2\cdots a_k)\begin{pmatrix}x_1\\ x_2\\ \vdots\\ x_k\end{pmatrix}=\mathbf{a}'\mathbf{x}\ , \tag{8 4}$$

$$v=(b_1,\ b_2,\ \cdots,\ b_s)\begin{pmatrix}y_1\\ y_2\\ \vdots\\ y_s\end{pmatrix}=\mathbf{b}'\mathbf{y}\ , \tag{8 5}$$

其中 a_1，a_2，$\cdots$，a_k和b_1，b_2，$\cdots$，b_s为待定系数。我们确定这两个系数向量的原则是使u和v之间的相关系数达到最大。

$$r=\frac{\operatorname{cov}(\mathbf{a}'\mathbf{x},\mathbf{b}'\mathbf{y})}{[\operatorname{var}(\mathbf{a}'\mathbf{x})\operatorname{var}(\mathbf{b}'\mathbf{y})]^{\frac{1}{2}}} \tag{8-6}$$

在所有线性组合u和v中选出相关系数最大的那一对，记为$u_1=\mathbf{a}^{(1)\prime}\mathbf{x}$和$v_1=\mathbf{b}^{(1)\prime}\mathbf{y}$，然后又在所有分别同$u_1$、$v_1$不相关的线性组合中选出相关系数最大的一对，记为$u_2=\mathbf{a}^{(2)\prime}\mathbf{x}$和$v_2=\mathbf{b}^{(2)\prime}\mathbf{y}$；继续下去，直到分别同$u_1$，$u_2$，$\cdots$，$u_m$和$v_1$，$v_2$，$\cdots$，$v_m$都不相关的线性组合$u$与$v$不相关为止，这个$m$正好等于协方差矩阵 Σ_{12}的秩。

因为u与v乘上一个常数后保持相关系数不变，所以对于a和b可施行规格化。假定a、b经过规格化后，使得u、v的均值为0，方差为1，则有

$$1=\mathrm{E}u^2-(\mathrm{E}u)^2=\mathrm{E}u^2=\mathrm{E}(\mathbf{a}'\mathbf{x}\mathbf{a}'\mathbf{x})$$
$$=\mathrm{E}(\mathbf{a}'\mathbf{x}\mathbf{x}'\mathbf{a})=\mathbf{a}'\ \Sigma_{11}\mathbf{a}\ , \tag{8-7}$$
$$1=\mathrm{E}v^2-(\mathrm{E}v)^2\ =\mathrm{E}v^2\ =\mathrm{E}(\mathbf{b}'\mathbf{y}\mathbf{b}'\mathbf{y})$$
$$=\mathrm{E}(\mathbf{b}'\mathbf{y}\mathbf{y}'\mathbf{b})=\mathbf{b}'\ \Sigma_{22}\mathbf{b}\ 。 \tag{8-8}$$

u与v之间的相关系数为

$$r=\operatorname{cov}(\mathbf{a}'\mathbf{x},\ \mathbf{b}'\mathbf{y})=\mathrm{E}(uv)=\mathrm{E}(\mathbf{a}'\mathbf{x}\mathbf{b}'\mathbf{y})$$
$$=\mathrm{E}(\mathbf{a}'\mathbf{x}\mathbf{y}'\mathbf{b})=\mathbf{a}'\mathrm{E}(\mathbf{x}\mathbf{y}')\mathbf{b}=\mathbf{a}'\ \Sigma_{12}\mathbf{b}\ 。 \tag{8-9}$$

于是，问题转化为在$\mathbf{a}'\ \Sigma_{11}\mathbf{a}-1=0$和$\mathbf{b}'\ \Sigma_{22}\mathbf{b}-1=0$的附加条件下，使$r=\mathbf{a}'\ \Sigma_{12}\mathbf{b}$为最大去确定a和b。为此，求

$$\theta = \mathbf{a}'\boldsymbol{\Sigma}_{12}\mathbf{b} - \frac{1}{2}\lambda_1(\mathbf{a}'\boldsymbol{\Sigma}_{11}\mathbf{a} - 1) - \frac{1}{2}\lambda_2(\mathbf{b}'\boldsymbol{\Sigma}_{22}\mathbf{b} - 1) \tag{8-10}$$

关于a和b的极值，其中$\frac{1}{2}\lambda_1$和$\frac{1}{2}\lambda_2$为拉格郎日乘数。

根据求极值原理，使θ最大的必要条件是θ对向量a和b的一阶偏导数为0，即

$$\frac{\partial\theta}{\partial\mathbf{a}} = \boldsymbol{\Sigma}_{12}\mathbf{b} - \frac{1}{2}\lambda_1(2\boldsymbol{\Sigma}_{11}\mathbf{a}) = \boldsymbol{\Sigma}_{12}\mathbf{b} - \lambda_1\boldsymbol{\Sigma}_{11}\mathbf{a} = 0 \ , \tag{8-11}$$

$$\frac{\partial\theta}{\partial\mathbf{b}} = \boldsymbol{\Sigma}_{21}\mathbf{a} - \frac{1}{2}\lambda_1(2\boldsymbol{\Sigma}_{22}\mathbf{b}) = \boldsymbol{\Sigma}_{21}\mathbf{a} - \lambda_2\boldsymbol{\Sigma}_{22}\mathbf{b} = 0 \ 。 \tag{8-12}$$

将（8-11）、（8-12）两式两边分别左乘$\mathbf{a}'$和$\mathbf{b}'$得

$$\mathbf{a}'\boldsymbol{\Sigma}_{12}\mathbf{b} - \lambda_1\mathbf{a}'\boldsymbol{\Sigma}_{11}\mathbf{a} = 0 \ ,$$

$$\mathbf{b}'\boldsymbol{\Sigma}_{21}\mathbf{a} - \lambda_2\mathbf{b}'\boldsymbol{\Sigma}_{22}\mathbf{b} = 0 \ 。$$

根据附加条件（8-7）、（8-8）得

$$\mathbf{a}'\boldsymbol{\Sigma}_{12}\mathbf{b} = \lambda_1 \ , \quad \mathbf{b}'\boldsymbol{\Sigma}_{21}\mathbf{a} = \lambda_2\mathbf{a}。$$

又因

$$(\mathbf{a}'\boldsymbol{\Sigma}_{12}\mathbf{b})' = \mathbf{b}'\boldsymbol{\Sigma}_{12}'(\mathbf{a}')' = \mathbf{b}'\boldsymbol{\Sigma}_{21}\mathbf{a} \ ,$$

即

$$\lambda_1' = \lambda_2$$

而实数λ_1的转置为λ_1本身，所以

$$\lambda_1 = \lambda_2 = \mathbf{a}'\boldsymbol{\Sigma}_{12}\mathbf{b} \ 。 \tag{8-13}$$

由（8-9）和（8-13）式，知$\lambda_1 = \lambda_2 = r$，即$\lambda_1$为$u$向量与$v$向量之间的相关系数。但a和b是待求向量，为了求得$\lambda$值尚须进一步变换。为此，将（8-11）和（8-12）式写为

$$\boldsymbol{\Sigma}_{12}\mathbf{b} - \lambda\boldsymbol{\Sigma}_{11}\mathbf{a} = 0 \ , \tag{8-14}$$

$$\boldsymbol{\Sigma}_{21}\mathbf{a} - \lambda\boldsymbol{\Sigma}_{22}\mathbf{b} = 0 \ 。 \tag{8-15}$$

将（8-14）式左乘$\boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}$得

$$\boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12}\mathbf{b} - \lambda\boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{11}\mathbf{a} = 0 \ ,$$

即

$$\boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12}\mathbf{b} - \lambda\boldsymbol{\Sigma}_{21}\mathbf{a} = 0 \ 。$$

将（8-15）式代入，得

$$\boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12}\mathbf{b} - \lambda^2\boldsymbol{\Sigma}_{22}\mathbf{b} = 0 \ 。$$

等式两边再左乘$\boldsymbol{\Sigma}_{22}^{-\frac{1}{2}}$，得

$$\boldsymbol{\Sigma}_{22}^{-\frac{1}{2}}\boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12}\mathbf{b} - \lambda^2\mathbf{b} = 0 \ ,$$

或

$$(\boldsymbol{\Sigma}_{22}^{-\frac{1}{2}}\boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12} - \lambda^2)\mathbf{b} = 0 \ 。 \tag{8-16}$$

同理，将（8-15）式两边左乘$\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}$，然后将(8-14)式代入，等式两边再左乘$\boldsymbol{\Sigma}_{11}^{-1}$，则得

$$(\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21} - \lambda^2)\mathbf{a} = 0 \ 。 \tag{8-17}$$

(8-16)、(8-17)式中协方差矩阵均为已知，

记

$$\mathbf{A}=\boldsymbol{\Sigma}_{11}^{-\frac{1}{2}}\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}\ , \tag{8-18}$$

$$\mathbf{B}=\boldsymbol{\Sigma}_{22}^{-\frac{1}{2}}\boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12}\ 。 \tag{8-19}$$

得

$$(\mathbf{A}-\lambda^2)\mathbf{a}=0\ , \tag{8-20}$$

$$(\mathbf{B}-\lambda^2)\mathbf{b}=0\ 。 \tag{8-21}$$

可见所要求的典型相关系数r就是矩阵**A**或**B**的特征值的平方根（λ^2是矩阵**A**，也是矩阵**B**的特征值），而待定向量**a**、**b**分别是**A**、**B**的相应特征向量。

由于假定协方差矩阵（8-3）是正定的，所以**A**和**B**也是正定的。**A**(或**B**)的秩等于$\boldsymbol{\Sigma}_{12}$的秩，设为m，则**A**、**B**的非零特征值的个数相等，都等于m。

设矩阵**A**的特征值为

$$\lambda_1^2\geqslant\lambda_2^2\geqslant\cdots\geqslant\lambda_k^2\geqslant0\ ,$$

矩阵**B**的特征值为

$$\lambda_1^2\geqslant\lambda_2^2\geqslant\cdots\geqslant\lambda_t^2\geqslant0\ ,$$

因为**A**和**B**的全部非零特征值是相同的，可用相同的符号写为

$$\lambda_1^2\geqslant\lambda_2^2\geqslant\cdots\geqslant\lambda_m^2\geqslant0\ 。$$

（m等于矩阵$\boldsymbol{\Sigma}_{12}$的秩），此时全部非零特征值的正平方根（因为λ是选取所有线性组合u和v之间的最大相关系数，故取λ^2的正平方根）为

$$\lambda_1\geqslant\lambda_2\geqslant\cdots\geqslant\lambda_m\geqslant0\ , \tag{8-22}$$

称为典型相关系数。它们分别表示了关于$\mathbf{x}=(x_1,\ x_2,\ \cdots,\ x_k)$和$\mathbf{y}=(y_1,\ y_2,\ \cdots,\ y_s)$的以下$m$对典型变量之间的相关系数：

$$\begin{cases}u_1=a_1^{(1)}x_1+a_2^{(1)}x_2+\cdots+a_k^{(1)}x_k\\ v_1=b_1^{(1)}y_1+b_2^{(1)}y_2+\cdots+b_s^{(1)}y_s\end{cases},$$

$$\begin{cases}u_2=a_1^{(2)}x_1+a_2^{(2)}x_2+\cdots+a_k^{(2)}x_k\\ v_2=b_1^{(2)}y_1+b_2^{(2)}y_2+\cdots+b_s^{(2)}y_s\end{cases}, \tag{8-23}$$

$$\vdots\qquad\qquad\vdots\qquad\qquad\vdots$$

$$\begin{cases}u_m=a_1^{(m)}x_1+a_2^{(m)}x_2+\cdots+a_k^{(m)}x_k\\ v_m=b_1^{(m)}y_1+b_2^{(m)}y_2+\cdots+b_s^{(m)}y_s\end{cases}。$$

典型变量有如下性质：

1. 由$\mathbf{x}=(x_1,\ x_2,\ \cdots,\ x_k)$组成的典型变量$u_1,\ u_2,\ \cdots,\ u_m$都互不相关，并且其方差为1。由$\mathbf{y}=(y_1,\ y_2,\ \cdots,\ y_s)$组成的典型变量$v_1,\ v_2,\ \cdots,\ v_m$亦然。

2. 由于$\mathbf{x}=(x_1,\ x_2,\ \cdots,\ x_k)$组成的前$m$个典型变量$u_i(i=1,\ 2,\ \cdots,\ m)$，分别同由$\mathbf{y}=(y_1,\ y_2,\ \cdots,\ y_s)$组成的前$m$个典型变量$v_i(i=1,\ 2,\ \cdots,\ m)$之间具有相关系数$\lambda_i\neq0(i=1,\ 2,\ \cdots,\ m)$。除此之外的由**x**组成的任何一个典型变量，同由**y**组成的任何一个典型变量都不相关。

3. 当$k=s=1$时，λ就是熟知的两个随机变量间的简单相关系数。

第三节　典型相关系数的显著性检验

设对母体（一般是两个母体）抽取一个容量为n的子样，得到个数分别为k和s的两组变量数据为

$$\mathbf{X}=\begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1n} \\ x_{21} & x_{22} & \cdots & x_{2n} \\ \vdots & \vdots & & \vdots \\ x_{k1} & x_{k2} & \cdots & x_{kn} \end{pmatrix},\quad \mathbf{Y}=\begin{pmatrix} y_{11} & y_{12} & \cdots & y_{1n} \\ y_{21} & y_{22} & \cdots & y_{2n} \\ \vdots & \vdots & & \vdots \\ y_{s1} & y_{s2} & \cdots & y_{sn} \end{pmatrix}。$$

求得协方差矩阵的无偏估计 $\boldsymbol{\Sigma}$：

$$\boldsymbol{\Sigma}=\begin{pmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{pmatrix}=(s_{ij})\ ,$$

其中 $\boldsymbol{\Sigma}_{11}$，$\boldsymbol{\Sigma}_{22}$和 $\boldsymbol{\Sigma}_{12}$的元素分别为

$$s_{ij}=\frac{1}{n-1}\sum_{t=1}^{n}(x_{it}-\overline{x_i})(x_{jt}-\overline{x_j})\qquad (i,\ j=1,\ 2,\ \cdots,\ k)\qquad (8\text{-}24)$$

$$s_{ij}=\frac{1}{n-1}\sum_{t=1}^{n}(y_{it}-\overline{y_i})(y_{jt}-\overline{y_j})\qquad (i,\ j=1,\ 2,\ \cdots,\ s)\qquad (8\text{-}25)$$

$$s_{ij}=\frac{1}{n-1}\sum_{t=1}^{n}(x_{it}-\overline{x_i})(y_{jt}-\overline{y_j})\qquad (i=1,2,\cdots,k;j=1,2,\cdots,s)。(8\text{-}26)$$

利用（8-18）、（8-19）、（8-20）和（8-21）式计算子样的典型相关系数和相应的子样典型变量。此时，要作典型相关系数的显著性检验。

假设x与y互不相关，则 $\boldsymbol{\Sigma}_{12}$所有元素为零，于是典型相关系数 $\lambda_i=\mathbf{a}^{(i)\prime}\boldsymbol{\Sigma}_{12}\mathbf{b}^{(i)}=0(i=1,\ 2,\ \cdots,\ m)$。

设矩阵A（或B）的m个特征值为

$$\lambda_1^2\geqslant\lambda_2^2\geqslant\cdots\geqslant\lambda_m^2\ 。$$

求

$$\Lambda_1=(1-\lambda_1^2)(1-\lambda_2^2)\cdots(1-\lambda_m^2)=\prod_{i=1}^{m}(1-\lambda_i^2)\qquad (8\text{-}27)$$

在x与y不相关的假设下，对于大子样（n较大）的统计量

$$Q_1=-\left[n-1-\frac{1}{2}(k+s+1)\right]\ln\Lambda_1\qquad (8\text{-}28)$$

近似地服从自由度为$k\times s$的χ^2-分布。给定显著性水平α，当$Q_1\geqslant\chi_\alpha^2(k\times s)$时，则至少有一个典型相关系数$\lambda_1$为显著相关。

再检验其余$m-1$个λ值的显著性，此时计算$\Lambda_2=(1-\lambda_2^2)(1-\lambda_3^2)\cdots(1-\lambda_m^2)$，统计量

$$Q_2=-\left[n-2-(k+s+1)/2\right]\ln\Lambda_2\qquad (8\text{-}29)$$

近似地服从$(k-1)(s-1)$为自由度的χ^2-分布。当$Q_2\geqslant\chi_\alpha^2[(k-1)(s-1)]$时，则$\lambda_2$也显著。逐个检验，直至相关不显著为止。

第四节　用相关矩阵代替协方差矩阵的情形

在实际计算中也可从子样的相关矩阵R出发来计算典型相关系数和典型变量。将相关矩阵R分块为4个矩阵：

$$\mathbf{R}=\begin{pmatrix}\mathbf{R}_{11} & \mathbf{R}_{12}\\ \mathbf{R}_{21} & \mathbf{R}_{22}\end{pmatrix}=(r_{ij}),$$

其中

$$r_{ij}=\frac{S_{ij}}{\sqrt{S_{ii}S_{jj}}}。$$

如果用R代替Σ，按(8-18)，(8-19)，(8-20)和(8-21)式解出典型相关系数为λ_1，λ_2，…，λ_m，矩阵A的相应的特征向量为$\alpha^{(i)}(i=1,2,\cdots,m)$。矩阵B的相应特征向量为$\beta^{(i)}(i=1,2,\cdots,m)$。那么所要求的典型相关系数仍然是$\lambda_1$，$\lambda_2$，…，$\lambda_m$。但是所要求的典型变量的系数向量$\mathbf{a}^{(i)}$却与$\alpha^{(i)}$有关系

$$\alpha^{(i)}=\mathbf{S}_x\mathbf{a}^{(i)}。\tag{8-30}$$

$\mathbf{b}^{(i)}$与$\beta^{(i)}$有关系

$$\beta^{(i)}=\mathbf{S}_y\mathbf{b}^{(i)},\tag{8-31}$$

其中$\mathbf{S}_x$和$\mathbf{S}_y$分别为

$$\mathbf{S}_x=\begin{pmatrix}\sqrt{S_{11}} & & 0\\ & \sqrt{S_{22}} & \\ & & \ddots\\ 0 & & \sqrt{S_{kk}}\end{pmatrix},\quad \mathbf{S}_y=\begin{pmatrix}\sqrt{S_{11}} & & 0\\ & \sqrt{S_{22}} & \\ & & \ddots\\ 0 & & \sqrt{S_{ss}}\end{pmatrix},\tag{8-32}$$

这是因为$\Sigma_{11}=\mathbf{S}_x\mathbf{R}_{11}\mathbf{S}_x$，$\Sigma_{22}=\mathbf{S}_y\mathbf{R}_{22}\mathbf{S}_y$，$\Sigma_{12}=\mathbf{S}_x\mathbf{R}_{12}\mathbf{S}_y$的缘故。

第五节　应用实例*

为了研究某铅锌矿铅锌的矿源层，利用12种主量和微量元素对某武系、泥盆系和矿体三部分样品分别作对应分析，确定它们的指示元素为SiO_2、Al_2O_3、Fe_2O_3、K_2O、CaO、CO_2、MgO；Pb、Zn、S。然后利用矿体的13个样品分别对成矿元素Pb、Zn与SiO_2、Al_2O_3、Fe_2O_3、K_2O；成矿元素Pb、Zn与CaO、MgO、CO_2进行典型相关分析，以研究Pb、Zn来自哪一个地层。典型相关分析计算步骤如下：

1．原始数据，见表8-1。

表8-1　某铅锌矿原始数据表

样号	CaO	MgO	CO_2	SiO_2	Al_2O_3	Fe_2O_3	K_2O	Pb	Zn
1	31.66	20.57	42.23	0.09	0.26	0.072	0.075	0.025	0.085
2	48.77	5.51	42.90	0.17	0.40	0.16	0.125	0.025	0.21
3	29.96	21.50	43.12	0.17	0.26	0.072	0.075	0.025	1

*据有色金属公司矿产地质研究院邓国英，1987。

续表8-1

样号	CaO	MgO	CO_2	SiO_2	Al_2O_3	Fe_2O_3	K_2O	Pb	Zn
4	28.70	20.06	42.90	2.24	1.14	0.28	0.40	0.175	0.56
5	30.98	20.06	43.42	0.95	0.60	0.11	0.15	0.10	0.64
6	27.45	19.72	41.18	2.17	1.20	0.57	0.40	0.25	1.41
7	27.11	19.46	40.09	4.34	2.31	0.57	0.70	0.162	0.909
8	28.87	20.78	43.50	0.26	0.46	1.00	0.10	0.125	0.905
9	31.32	10.80	42.52	0.52	0.53	0.14	0.01	0.025	0.10
10	29.50	20.44	43.59	1.32	0.81	0.10	0.05	0.10	0.235
11	28.55	21.17	44.29	0.45	0.32	0.16	0.005	0.055	0.333
12	29.80	21.67	45.58	0.12	0.27	0.038	0.005	0.115	0.047
13	29.10	21.54	44.33	0.67	0.26	0.02	0.005	0.16	0.877
$\bar{x}$	30.9046	10.4662	45.0508	1.0362	0.6785	0.2581	0.1685	0.0886	0.4970
s_n	5.3217	4.0734	1.3436	1.1871	0.5624	0.2742	0.1987	0.0755	0.4083

2．计算泥盆系3个指示元素和Pb、Zn共5个变量的相关矩阵

$$R=\left[\begin{array}{ccc:cc} 1 & -0.9462 & 0.0512 & -0.4029 & -0.3676 \\ -0.9462 & 1 & 0.1727 & 0.1661 & 0.1324 \\ 0.0512 & 0.1727 & 1 & -0.4537 & -0.4389 \\ \hdashline -0.4029 & 0.1661 & -0.4537 & 1 & 0.9409 \\ -0.3676 & 0.1324 & -0.4389 & 0.9409 & 1 \end{array}\right]。$$

3．求逆矩阵R_{11}^{-1}和R_{22}^{-1}：

$$R_{11}^{-1}=\left[\begin{array}{ccc} 17.4723 & 17.1998 & -3.8650 \\ 17.1998 & 17.9622 & -3.9827 \\ -3.8650 & -3.9827 & 1.8857 \end{array}\right],$$

$$R_{22}^{-1}=\left[\begin{array}{cc} 8.7179 & -8.2026 \\ -8.2026 & 8.7179 \end{array}\right]。$$

4．计算$A=R_{11}^{-1}R_{12}R_{22}^{-1}R_{21}$和$B=R_{22}^{-1}R_{21}R_{11}^{-1}R_{12}$：

$$A=\left[\begin{array}{ccc} 0.9623 & -0.3691 & 1.1193 \\ 0.8435 & -0.3171 & 0.9898 \\ -0.0134 & 0.0008 & -0.0211 \end{array}\right],$$

$$B=\left[\begin{array}{cc} 0.4189 & 0.3901 \\ 0.1979 & 0.2051 \end{array}\right]。$$

5．计算典型相关系数和典型变量

求矩阵A与B的特征值。由于A与B有相同的非零特征值，故可以求阶数小的矩阵B的特征值：

$$|B-\lambda^2 I|=(\lambda^2)^2-0.6240\lambda^2+0.0069=0,$$

解之得：$\lambda_1^2=0.6127$，$\lambda_2^2=0.0113$。

可以验证，此λ_1^2和λ_2^2也满足A的特征方程：

$$|A-\lambda^2 I|=0 \ 。$$

故两个典型相关系数为$\lambda_1=0.7828$，$\lambda_2=0.1063$。

求矩阵A对应于λ_1^2和λ_2^2的特征向量

$$\alpha^{(1)}=S_x a^{(1)}=\begin{pmatrix}50.1188\\44.4023\\-1.0000\end{pmatrix}，\quad \alpha^{(2)}=S_x a^{(2)}=\begin{pmatrix}2.7056\\3.9353\\-1.0000\end{pmatrix}，$$

其中

$$S_x=\begin{pmatrix}\sqrt{S_{11}}&0&\\&\sqrt{S_{22}}&\\0&&\sqrt{S_{33}}\end{pmatrix}=\sqrt{13}\begin{pmatrix}5.3217&0&\\&4.0734&\\0&&1.3436\end{pmatrix}。$$

求得$a^{(1)}$、$a^{(2)}$，并单位化，得

$$a^{(1)}=\begin{pmatrix}0.6529\\0.7557\\-0.0521\end{pmatrix}，\quad a^{(2)}=\begin{pmatrix}0.3848\\0.7312\\-0.5632\end{pmatrix}。$$

同样，求得矩阵B对应于λ_1^2、λ_2^2的特征向量分别为

$$\beta^{(1)}=S_y b^{(1)}=\begin{pmatrix}-2.0593\\-1.0000\end{pmatrix}，$$

$$\beta^{(2)}=S_y b^{(2)}=\begin{pmatrix}0.9791\\-1.0000\end{pmatrix}，$$

其中

$$S_y=\begin{pmatrix}\sqrt{S_{11}}&0\\0&\sqrt{S_{22}}\end{pmatrix}=\begin{pmatrix}0.0756&0\\0&0.4083\end{pmatrix}。$$

求得$b^{(1)}$、$b^{(2)}$，并经单位化后为

$$b^{(1)}=\begin{pmatrix}-0.9960\\-0.0896\end{pmatrix}，\quad b^{(2)}=\begin{pmatrix}0.9826\\-0.1858\end{pmatrix}。$$

因此，两组变量的两个典型相关系数及其典型变量为

$$\lambda_1=0.7828，\ u_1=0.0529\mathrm{CaO}+0.7557\mathrm{MgO}-0.0521\mathrm{CO_2}，$$
$$v_1=-0.9960\mathrm{Pb}-0.0896\mathrm{Zn}\ ;$$
$$\lambda_2=0.1063，\ u_2=0.3848\mathrm{CaO}+0.7312\mathrm{MgO}-0.5632\mathrm{CO_2}，$$
$$v_2=0.9826\mathrm{Pb}-0.1858\mathrm{Zn}\ 。$$

6. 相关系数的显著性检验

因为$\lambda_1^2=0.6127$，$\lambda_2^2=0.0113$，

$$A_0=(1-\lambda_1^2)(1-\lambda_2^2)=0.3829，$$

$$Q_0=-\left[13-1-\frac{1}{2}\times(3+2+1)\right]\times\ln 0.3829=8.6398 。$$

而　　$\chi^2_{0.05}(3\times2)=12.592$，　　$Q_0<\chi^2_{(\alpha)}$ 。

故u_1与v_1在$\alpha=0.05$水平上不相关。

又

$$\Lambda_1 = (1-\lambda_2^2) = 0.9887\ ,$$

$$Q_1 = -\left[13-2-\frac{1}{2}\times(3+2+1)\right]\times\ln 0.9887 = 0.0909\ 。$$

而 $\chi^2_{0.05}(2\times1) = 5.991$，$Q_1 < \chi^2_\alpha$ 。

故u_2与v_2也不相关，可见成矿元素Pb、Zn与泥盆系无关。

7. 用同样方法求得寒武系变量SiO_2，Al_2O_3，Fe_2O_3，K_2O与成矿元素Pb，Zn的典型相关系数及其相应的典型变量。

$$\lambda_1 = 0.9647,\ u_1 = 4.0547SiO_2 - 4.9905Al_2O_3 + 0.5282Fe_2O_3 + 1.4115K_2O\ ,$$
$$v_1 = 1.6091Pb - 0.6756Zn\ ;$$

$$\lambda_2 = 0.6936,\ u_2 = 0.1927SiO_2 + 1.5512Al_2O_3 + 1.0076Fe_2O_3 - 2.1469K_2O\ ,$$
$$v_2 = -2.4742Pb + 2.8730Zn\ 。$$

因为

$$Q_0 = -\left[n-1-\frac{1}{2}(k+s+1)\right]\ln\left[(1-\lambda_1^2)(1-\lambda_2^2)\right] = 28.256,$$

而 $\chi^2_{0.05}(4\times2) = 15.507$，$Q_0 > \chi^2_\alpha$ ，

故u_1与v_1显著相关；

又

$$Q_1 = -\left[13-2-\frac{1}{2}\times(4+2+1)\right]\cdot\ln(1-0.6936^2) = 4.9203,$$

而 $\chi^2_{0.05}(3\times1) = 7.815$，$Q_1 < \chi^2_\alpha$ 。

故u_1与v_1不相关。

从以上结果可以看出，成矿元素Pb、Zn与泥盆系化学组分无关，没有成因联系。而Pb、Zn与寒武系关系密切（指第一对典型变量），有成因联系。主要是成矿元素Pb、Zn与碎屑粘土成分Al_2O_3、SiO_2、K_2O之间密切相关，是同生沉积的结果，即寒武系砂页岩就是铅锌矿的矿源层。

第六节　问题与讨论

1. 典型相关分析是对待研究的两组变量进行适当的线性变换，将整个相关结构简化概括为m个典型相关关系。典型相关的几何涵义是：在一个$p=k+s$多维空间内，一个对p个变量进行观测的样本确定了一个k维的超平面和一个s维的超平面，二者相交于坐标原点，并且分别包含着代表两组变量的点。对k个坐标轴和s个坐标轴作线性变换，使得这两个超平面在新的p维空间内尽可能互相平行。这两个超平面之间夹角的余弦定义为典型相关系数。

2. 典型相关分析和因子分析有某些相似之处，二者均研究变量组合之间的相关关系，也都探求因果关系，但二者也有重要区别：

（1）在典型相关分析中，变量组合是预先规定的，在某种意义上是任意的，从而这种

方法带有主观的成分；而在因子分析中，由因子所代表的变量组合则是通过分析后加以确定的，因而较为客观。

（2）在因果联系方面，典型相关分析是寻求变量组合之间产生最大相关的可能原因，而因子分析则探求诸因子的成因意义。

（3）因子分析可以通过因子旋转来提高因子模型与自然物理模型之间的吻合程度，而在典型相关分析中则不可能实现典型变量的旋转，因为典型变量的任何进一步的改变都将破坏典型相关，并将外加的相关关系重新引入所研究的体系。

由以上三点可见，两种方法的应用范围有所区别，因果分析的广度和深度也有所不同。

3．在主成分分析中先寻找不相关的变量，然后证明所找出的变量具有极值性质；而在典型相关分析中则相反，先寻求最大相关，然后证明所得的变量是不相关的。

在主成分分析中，主成分具有由特征值所表示的最大方差；在典型相关分析中，典型相关分量并没有这样的最大方差，而是通过将相关关系归约到一个最小集合来实现最优化。

第九章　马尔科夫概型分析

第一节　基本概念

一、随机过程

要了解随机过程，首先介绍随机变量和随机函数的概念。

1．随机变量：简单地说，具有一定概率分布的变量称为随机变量。也就是说，随机变量的取值不是一个确定的值，而是受随机因素w的影响，呈一定的概率分布，记作$Z=Z(w)$。当对Z进行一次观测（随机试验），其结果是一个确定的值，叫Z的一个现实（或实现），记作z（或Z）。

2．随机函数：除受随机因素w影响外，还受其它自变量（一个或多个）控制的随机变量，例如，设有两个自变量u、v，$u\in U$，$v\in V$，则称$Z(u,v,w)$为定义在$\{U,V\}$的随机函数。则

（1）当u，v固定时，$Z(u,v,w)$为一随机变量；

（2）每一次观测的结果为一确定的函数$Z(u,v)$称为$Z(u,v,w)$的一个现实。

故随机函数$Z(u,v,w)$既可以理解为具有两个参数u，v的随机变量族，也可以理解为它的所有现实$Z(u,v)$的集合。

3．随机过程：当随机函数中只有一个自变量t（一般t表示时间）时，称它为随机过程，记为$Z(t,w)$。或者说，任何一个受概率定律支配的过程称为随机过程。随机过程中每一个t都是一个随机变量。因此，随机过程中必须用一组随机变量的集合来刻划。

4．马尔科夫过程：如果一个过程发展的将来状态，只取决于现在的状态，而与过去的状态无关，这种性质叫做马尔科夫性。具有马尔科夫性的随机过程称为马尔科夫过程。若$t_1<t_2<\cdots<t_n$为时间参数值，且$1<j<n$，则对于给定的$x(t_j)$，随机变量的集合$\{x(t_1)\cdots,x(t_{j-1})\}$和$\{x(t_{j+1}),\cdots,x(t_n)\}$相互独立；或者等价地说，对于给定的$x(t_1),\cdots,x(t_{n-1})$；$x(t_n)$的条件概率分布只依赖于$x(t_{n-1})$的特定值，即事实上就是在给定$x(t_{n-1})$的条件下，$x(t_n)$的条件概率分布：

$$P\{x(t_n)\leqslant x_n\mid x(t_1)=x_1,\cdots,x(t_{n-1})=x_{n-1}\}$$
$$=P\{x(t_n)\leqslant x_n\mid x(t_{n-1})=x_{n-1}\}\quad。\qquad(9\text{-}1)$$

因此，一个马尔科夫过程对于最近瞬间之前是无“记忆”的，这个性质称为无后效性。故马尔科夫过程又称无后效随机过程。

状态空间间断的马尔科夫过程称为马尔科夫链。通常的马尔科夫链中，体系可能处的状态数目是有穷或可数无穷多个的。

马尔科夫过程的数学模型称为马尔科夫概型。

地球科学是一种历史的科学，地球上所发生的种种地质现象都曾经历过一个历史演变过程。例如成矿过程，从地球化学的观点看，是构成矿床的大量化学元素的原子在成矿作用的物理化学和热力学、动力学条件下的迁移沉淀的过程。这种过程具有随机过程的性质。因此，运用随机过程的理论，特别是马尔科夫过程的理论来研究有明显的优越性。马尔科夫概型在地质中常常用来划分沉积旋回，进行地层对比，查明火山岩系和侵入杂岩体的活动顺序，揭露成矿过程，划分成矿阶段等研究工作。

第二节　马尔科夫链的转移概率

一、转移概率与转移矩阵

在一个马尔科夫链中，系统状态的转移可以用一个转移概率矩阵$\mathbf{P}=(p_{ij})$来表达。转移概率矩阵中的元素p_{ij}称为转移概率。马尔科夫链的转移概率可以表为

$$P\{x(t_n)=j \mid x(t_{n-1})=i\}=p_{ij}\ , \tag{9-2}$$

p_{ij}表示已知在时刻t_{n-1}系统状态处于i的条件下，系统在时刻$t_n(t_n>t_{n-1})$处于状态j的概率，并称为转移概率。p_{ij}下标的次序对应于转移方向，即从状态i向状态j转移。转移矩阵$\mathbf{P}=(p_{ij})$的形式是

$$\mathbf{P}=\begin{pmatrix} p_{11} & p_{12} & p_{13} & \cdots \\ p_{21} & p_{22} & p_{23} & \cdots \\ p_{31} & p_{32} & p_{33} & \cdots \\ \cdots & \cdots & \cdots & \cdots \end{pmatrix}。$$

有时为了代数运算的方便，也用它的转置矩阵的形式$\mathbf{P}=(p_{ij})'$来表示。

转移矩阵有两个重要性质，即其所有元素都是非负的，且所有各行之和均为1：

$$1\geqslant p_{ij}\geqslant 0,$$

$$\sum_{j=1} p_{ij}=1,\qquad i=1,\ 2,\ \cdots\ 。 \tag{9-3}$$

转移概率描述了马尔科夫过程的概率统计特征，因此如何确定转移概率的问题就成为马尔科夫概型分析中所要解决的一个重要问题。

转移概率在理论上是绝对概率，实际应用中往往用频率来代替，得到“转移频率矩阵”。如，设有一地层剖面，包含4种岩性：砂岩（A），灰岩（B），页岩（C）和煤层（D）。由剖面底部到顶部，这4种岩性出现的顺序如下：

底部　ABACDCDABCBADCDCBACABDABCDBACDA　顶部。

计算它们从一种岩性转移到其它岩性的频率，可以写出一个4×4矩阵，用其中的元素（出现的次数）来表示从一种已知状态（岩性，以字母表示）转移到另一种状态的次数。

		到 A	B	C	D	每行之和
从	A	0	4	3	1	8
	B	4	0	2	1	7
	C	1	2	0	5	8
	D	3	1	3	0	7
每列之和		8	7	8	7	

将频数换算成频率所构成的矩阵，被称为“转移频率矩阵”，将上述矩阵中每一行的每个元素除以该行各元素之和求得：

		到 A	B	C	D	每行之和
从	A	0	0.500	0.375	0.125	1.000
	B	0.571	0	0.286	0.143	1.000
	C	0.125	0.250	0	0.625	1.000
	D	0.428	0.143	0.429	0	1.000

二、柴普曼-柯尔莫各洛夫（Chapman-Kolmgorov）方程

利用转移矩阵可以计算从状态i经过两步或三步转移到状态j的转移概率。

例如：对于一个包含砂岩，粉砂岩，页岩的系统，由图(9-1)可见，从粉砂岩经过两步转移到页岩有3条不同的途径，使得

$$p_{23}^{(2)}=\sum_{j=1}^{3}p_{2j}p_{j3}=p_{21}p_{13}+p_{22}p_{23}+p_{23}p_{33}$$
$$=0.4\times0.5+0.4\times0.2+0.2\times0.1$$
$$=0.30\ 。$$

这是从粉砂岩由3条不同途径经过两步转移到页岩的概率之和，j表示途径数。记号$p_{23}^{(2)}$读作“从状态2经过两步转移到状态3的概率”。

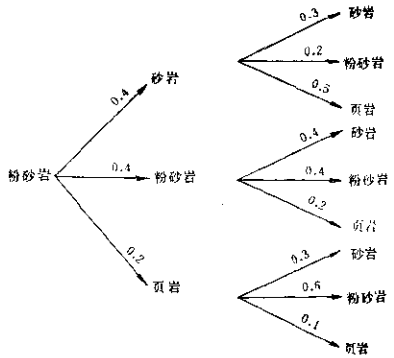


图9-1 粉砂岩、砂岩、页岩转移概率图

同样，可以计算

$$p_{22}^{(2)}=\sum_{j=1}^{3}p_{2j}p_{j2}=p_{21}p_{12}+p_{22}p_{22}+p_{23}p_{32}$$
$$=0.4\times0.2+0.4\times0.4+0.2\times0.6=0.36\ ,$$

$$p_{21}^{(2)}=\sum_{j=1}^{3}p_{2j}p_{j1}=p_{21}p_{11}+p_{22}p_{21}+p_{23}p_{31}$$
$$=0.4\times0.3+0.4\times0.4+0.2\times0.3=0.34\ 。$$

两步的一般公式是：

$$p_{ik}^{(2)} = \sum_j p_{ij} p_{jk} \quad 。$$

推广到两步以上，则一般公式为：

$$p_{ik}^{(n)} = \sum_j p_{ij}^{(m)} p_{jk}^{(n-m)} \qquad (1<m<n)\ , \tag{9-4}$$

其中n指步数。(9-4)式指出从状态i经过n步转移到状态k过程，可以看作先经过m步转移到状态j，再由状态j经过$(n-m)$步转移到状态k。(9-4)式称为“柴普曼-柯尔莫各洛夫”方程（离散参数平稳马尔科夫链）。它是研究马尔科夫过程的基础。

柴普曼-柯尔莫各洛夫方程可以写成另一种形式：

$$p_{ik}(t_2) = \sum_j p_{ij}(t_1) p_{jk}(t_2 - t_1) \quad 。 \tag{9-5}$$

这一方程可以这样来解释：设体系在时刻0处于状态i，在时刻t_2处于状态k。从状态i向状态k的转移过程必须经过在时刻t_1体系所处的中间状态$j(0<t_1<t_2)$。由于t_1和t_2之间的这一段过程独立于过去，因而体系在时刻t_1处于状态j和在时刻t_2处于状态k的概率是两个独立随机事件的概率之乘积：$p_{ij}(t_1)p_{jk}(t_2-t_1)$。

显然，在已知体系中当时刻0处于状态i，在时刻t_2处于状态k的概率应该对j的所有可能的取值求和，这就是（9-5）式的右边。

值得指出的是，对于离散参数齐次马尔科夫链情形下的柴普曼-柯尔莫各洛夫方程（9-5）或（9-4）式，有：①时间参数限于取整数值，即从体系在时刻t所处的状态只能直接转移到时刻$(t+1)$或$(t-1)$的相邻状态；②过程是时间齐次的，即所谓“齐次马尔科夫链”或称“平稳马尔科夫链”。转移概率不随时间而变（$p_{ik}(t)$对于长度t的所有时间区间均相同），而只依赖于时间之差，实际上可将（9-5）式写成：$p_{ik}(t_2-0) = \sum_j p_{ij}(t_1-0)p_{jk}(t_2-t_1)$。

三、绝对概率

马尔科夫概型分析很重要的一个问题是在已知系统初始状态和转移矩阵的情况下，如何计算经n步转移后，体系处于各种状态的概率。

设有下列转移矩阵P

$$\begin{array}{cc} & \text{从}i \\ \text{到}j & \begin{pmatrix} & \text{I} & \text{II} & \text{III} & \text{IV} \\ \text{I} & 0 & 0 & 0 & 0 \\ \text{II} & 0.5 & 0 & 0 & 0 \\ \text{III} & 0.5 & 0.5 & 1 & 0 \\ \text{IV} & 0 & 0.5 & 0 & 1 \end{pmatrix} \end{array}$$

令p_i为过程发生前体系处于状态i的概率，过程发生后以概率$p_j^{(n)}$处于状态j，转移一步即为$p_j^{(1)}$。在一个$p_j^{(1)}$依赖于过程发生前体系处于i的概率p_i（初始状态），记$p^{(0)}=p_i$，则当过程进行第一步（或在时刻1），体系处于状态$j(j=\text{I},\ \text{II},\ \text{III},\ \text{IV})$的绝对概率为

$$p_{\text{I}}^{(1)} = 0,$$

$$p_{\text{II}}^{(1)} = 0.5 \times p\text{I}\ ,$$

$$p_{\text{III}}^{(1)} = p_{\text{III}} + 0.5 \times (p_{\text{I}} + p_{\text{II}}),$$

$$p_{\mathrm{N}}^{(1)}=p_{\mathrm{N}}+0.5\times p_{\mathrm{I}}\ 。$$

写成矩阵形式：

$$\begin{bmatrix} p_{\mathrm{I}}^{(1)} \\ p_{\mathrm{II}}^{(1)} \\ p_{\mathrm{III}}^{(1)} \\ p_{\mathrm{N}}^{(1)} \end{bmatrix}=\begin{bmatrix} 0 & 0 & 0 & 0 \\ 0.5 & 0 & 0 & 0 \\ 0.5 & 0.5 & 1 & 0 \\ 0 & 0.5 & 0 & 1 \end{bmatrix}\cdot\begin{bmatrix} p_{\mathrm{I}} \\ p_{\mathrm{II}} \\ p_{\mathrm{III}} \\ p_{\mathrm{N}} \end{bmatrix},$$

即

$$\mathbf{P}^{(1)}=\mathbf{P}\cdot\mathbf{P}^{(0)}\ ,$$

式中$\mathbf{P}^{(1)}$为体系过程发生第一步后的绝对概率；$\mathbf{P}$为转移概率矩阵(简称转移矩阵)；$p^{(0)}$为体系在过程发生之前的概率分布（初始概率分布）。

如果过程继续进行到第二步，则在计算时刻2时体系的绝对概率为$\mathbf{P}^{(2)}$，可以把体系在时刻1时的绝对概率$\mathbf{P}^{(1)}$当作其初始状态，也就是说：

$$\mathbf{P}^{(2)}=\mathbf{P}\mathbf{P}^{(1)}=\mathbf{P}\mathbf{P}\mathbf{P}^{(0)}=\mathbf{P}^{2}\mathbf{P}^{(0)}\ ,$$

其中

$$\mathbf{P}^{2}=\begin{bmatrix} 0 & 0 & 0 & 0 \\ 0.5 & 0 & 0 & 0 \\ 0.5 & 0.5 & 1 & 0 \\ 0 & 0.5 & 0 & 1 \end{bmatrix}\begin{bmatrix} 0 & 0 & 0 & 0 \\ 0.5 & 0 & 0 & 0 \\ 0.5 & 0.5 & 1 & 0 \\ 0 & 0.5 & 0 & 1 \end{bmatrix}=\begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0.75 & 0.5 & 1 & 0 \\ 0.25 & 0.5 & 0 & 1 \end{bmatrix}。$$

同理可推得，当过程进行到第三步后（时刻3），体系处于各特定状态的绝对概率为

$$\mathbf{P}^{(3)}=\mathbf{P}^{3}\mathbf{P}^{(0)}\ 。$$

则第n步后的绝对概率分布为

$$\mathbf{P}^{(n)}=\mathbf{P}^{n}\mathbf{P}^{(0)}\ 。\qquad(9\text{-}6)$$

因此矩阵$\mathbf{P}^{n}$给出了所需要的n步转移概率$[p_{ij}^{(n)}]$。

可以证明，随着过程的发展，初始阶段的影响逐渐消失，因而当n甚大时，体系在时刻n处于状态j的概率将不依赖于在时刻0的初始状态，即绝对概率分布$\mathbf{P}^{(n)}$收敛于一个独立于初始分布$\mathbf{P}^{(0)}$的极限分布$\mathbf{P}$：

$$\lim_{n\to\infty}\mathbf{P}^{(n)}=\mathbf{P}\ 。$$

它表示当n甚大时，体系处于某特定状态的概率。

这是马尔科夫链的一个极重要的性质，称为“各态历经性质”（遍历性）。

第三节　连续参数非平稳马尔科夫过程

一、柴普曼-柯尔莫各洛夫方程

上一节中讨论了离散参数柴普曼-柯尔莫各洛夫方程。用$p_{ik}^{(n)}$表示从状态i到状态k的n步转移概率，现在用$p_{ij}(s,\ t)$表示已知在时间s体系处于i状态条件下，在时间t体系处于状

态j的条件概率，即

$$P\{x(t)=j \mid x(s)=i\}=P_{ij}(s,\ t)\ 。 \tag{9-7}$$

现在考虑3个时刻s，t和u，且$s<t<u$。设体系在时刻s处于状态i，在时刻u处于状态n，则所研究的过程必然经历在时刻t的某一状态j。对于一个马尔科夫过程，体系由状态i经状态j到状态n这样一个特定过程的概率是$p_{ij}(s,\ t)p_{jn}(t,\ u)$，因为据定义，这一特定过程的两个阶段是彼此独立的。因此，对j求和即得已知在时刻s体系处于状态i的条件下，在时刻u体系处于状态n的概率，即

$$p_{in}(s,\ u)=\sum_j p_{ij}(s,t)p_{jn}(t,\ n)\ 。 \tag{9-8}$$

(9-8)式就是连续参数条件下的柴普曼-柯尔莫各洛夫方程，它是和离散参数条件下的柴普曼-柯尔莫各洛夫方程(9-5)式相对应的，而(9-5)式的时间只能取整数值。

为了表达得更简洁，用矩阵记法

$$\mathbf{P}(t\,|s)\equiv[p_{ij}(s,\ t)]'\ , \tag{9-9}$$

这里，将转移矩阵用$[p_{ij}]$的转置来定义，以便矩阵$\mathbf{P}$中的元素的下标保持习惯的次序，而$\mathbf{P}$中的各列则为条件分布。

将(9-8)式用(9-9)式的矩阵形式表达，则可将柴普曼-柯尔莫各洛夫方程写成如下形式：

$$\mathbf{P}(u\,|\,s)=\mathbf{P}(u\,|t)\mathbf{P}(t\,|s)\ 。 \tag{9-10}$$

至此，我们一直未作时间齐次的假设。若过程对时间为齐次的，则转移概率$p_{ij}(s,t)$将不随初始时刻s而变，而只依赖于时刻t和s之差，即

$$p_{ij}(s,\ t)=p_{ij}(t-s)\ 。 \tag{9-11}$$

设σ和τ分别为相继两个时间区间的长度，则对于这样两个相继时间区间，可以用稍微不同的记号将(9-10)式写成

$$\mathbf{P}(\sigma+\tau)=\mathbf{P}(\sigma)\mathbf{P}(\tau)\ 。 \tag{9-12}$$

二、时间齐次条件下的柯尔莫各洛夫微分方程

令时刻s和t固定，而时刻$u=t+\Delta t$，则当$\Delta t\to 0$时可由连续参数条件下柴普曼-柯尔莫各洛夫方程(9-8)推导出两个基本微分方程：

1．柯尔莫各洛夫“前向微分方程”

$$\frac{\partial p_{in}(s,t)}{\partial t}=-\lambda_n(t)p_{in}(s,t)+\sum_{j\neq u}p_{ij}(s,\ t)\lambda_j(t)q_{jn}(t)\ , \tag{9-13}$$

式中$\lambda_n(t)\Delta t$为在时间区间Δt内发生某些变化的概率；

$q_{jn}(t)$为在时刻t从状态j转移到状态n的概率，它是以已发生的某种变化为条件的一种条件概率，称为“极限转移概率”。

若用对角阵$\mathbf{\Lambda}(t)\Delta t$中的第$n$个元素表示在时间区间$\Delta t$内由状态$n$发生某些变化的概率，用$\mathbf{Q}(t)$记极限条件转移矩阵，则可将前向方程写成如下矩阵形式：

$$\begin{aligned}\frac{\partial\mathbf{P}(t\,|s)}{\partial t}&=\{\mathbf{Q}(t)-\mathbf{I}\}\mathbf{\Lambda}(t)\mathbf{P}(t\,|s)\\&=\mathbf{R}(t)\mathbf{P}(t\,|s)\ ,\end{aligned} \tag{9-14}$$

其中

$$\mathbf{R}(t)=\{\mathbf{Q}(t)-\mathbf{I}\}\mathbf{\Lambda}(t)\ 。$$

2。柯尔莫各洛夫“后向微分方程”

$$\frac{\partial p_{in}(s,t)}{\partial s}=\lambda_i(s)p_{in}(s,\ t)-\lambda_i(s)\sum_{j\neq i}q_{ij}(s)p_{jn}(s,t)\ 。 \tag{9-15}$$

同样，可将后向方程写成如下矩阵形式：

$$\frac{\partial \mathbf{P}(t\,|s)}{\partial s}=-\mathbf{P}(t|\,s)\mathbf{R}(s)\ 。 \tag{9-16}$$

在时间齐次条件下，$p_{ij}(s,\ t)$只依赖于时刻t和s之差$t-s$，显然λ_n和q_{in}是和t无关的常数。因而前向方程（9-13）可写为

$$\frac{dp_{in}(t)}{dt}=-\lambda_n p_{in}(t)+\sum_{j\neq n}\lambda_j q_{jn}p_{ij}(t)\ , \tag{9-17}$$

而后向方程（9-15）则可写为

$$\frac{dp_{in}(t)}{dt}=-\lambda_i p_{in}(t)+\lambda_i\sum_{j\neq n}q_{ij}p_{jn}(t)\ 。 \tag{9-18}$$

必须注意，这里用的是对于t（而不是对于s）的微分方程系数。

如果用矩阵形式表达，则前向方程（9-17）可写为

$$\frac{d\mathbf{P}((t)}{dt}=\mathbf{RP}(t)\ , \tag{9-19}$$

而后向方程（9-18）可写为

$$\frac{d\mathbf{P}(t)}{dt}=\mathbf{P}(t)\mathbf{R}\ 。 \tag{9-20}$$

（9-19）和（9-20）式分别是（9-14）和（9-16）式的齐次形式。但现在$\mathbf{R}=[r_{ij}]$，$r_{ii}=-\lambda_i$，$r_{ij}=q_{ij}\lambda_i(i\neq j)$，均和$t$的值无关。当然（9-19）和（9-20）式的初始条件均为

$$\mathbf{P}(0)=\mathbf{I}\ 。$$

最后，在时刻t的绝对概率$\mathbf{P}(t)$必须满足

$$\frac{d\mathbf{P}(t)}{dt}=\mathbf{RP}(t)\ 。 \tag{9-21}$$

三、柯尔莫各洛夫微分方程的解

在非齐次的条件下，矩阵$\mathbf{R}(t)$随时间t而变，因而前向微分方程（9-14）和后向微分方程（9-16）有不同的解；而在齐次条件下，矩阵$\mathbf{R}$不随时间而变，因而前向微分方程（9-19）和后向微分方程（9-20）有一个满足初始条件和柴普曼-柯尔莫各洛夫方程共同形式解，即

$$\mathbf{P}(t)=\exp(\mathbf{R}(t))。 \tag{9-22}$$

如果体系只有有穷个状态，$\mathbf{R}$是一个有穷矩阵，则用$\mathbf{R}$的幂展开的指数展开式能够成立，因而可以利用谱分解的方法，即若

$$\mathbf{R}=\sum_{i=1}^{k}a_iA_i\ , \tag{9-23}$$

式中A_i为谱集合；a_i为$\mathbf{R}$的特征值，则

$$\exp(\mathbf{R}(t)) = \sum_{i=1}^{k} e^{a_i t} A_i$$

得柯尔莫各洛夫方程解（9-22）的显式

$$\mathbf{P}(t) = \sum_{i=1}^{k} e^{a_i t} A_i \quad 。 \tag{9-24}$$

四、结论

当$t=1$时，（9-22）式简化为

$$\mathbf{P}(1) = e^{\mathbf{R}} \quad 。$$

在实际工作中，使单位时间间隔等于取样间距，这样，若将t表为单位取样间距，并将$\mathbf{P}(1)$写作$\mathbf{P}$，则（9-22）式变为

$$\mathbf{P}(s) = \mathbf{P}^s \quad 。 \tag{9-25}$$

（9-25）式就是s步转移矩阵$[p_{ij}(s)]$。

利用谱分解法，由（9-25）式得

$$\mathbf{P}(s) = \sum_{i=1}^{k} \lambda_i^s A_i \quad 。 \tag{9-26}$$

由于矩阵$\mathbf{R}$和矩阵$\mathbf{P}$有相同的谱集合，而其特征值$a_i = \ln\lambda_i$，故将矩阵$\mathbf{R}$确定以后，即可由（9-22）式计算转移矩阵$\mathbf{P}(t)$。

但由（9-25）式可知，在求出前向和后向柯尔莫各洛夫微分方程的共同解之后，只要将t表为单位取样间距，就可以不必计算$\mathbf{R}$的显式而直接得到s步转移矩阵$\mathbf{P}(s)$。

既然在齐次条件下，柯尔莫各洛夫前向方程和后向方程有共同解，而过程总是向正的时间方向发展，故今后我们用前向微分方程进行分析。如用柯尔莫各洛夫前向微分方程则可直接得到前向转移矩阵（设用$\mathbf{U}$表示）。如果用$\mathbf{x}_k$表示在时刻k的观测向量，则连续过程就可以用

$$\hat{\mathbf{x}}_{k+s} = \mathbf{U}^s \mathbf{x}_k$$

进行描述，其中s为一非负连续变量。

当$s=0$，$\mathbf{U}^s$变为单位阵，上式就简化为

$$\hat{\mathbf{x}}_k = \mathbf{x}_k \quad ,$$

即利用柯尔莫各洛夫前向微分方程所构造的连续过程恰巧相当于$s=0$时时间序列的实际观测值。

当$s=1$，即对于单位时间间隔或单位取样间距，就有

$$\hat{\mathbf{x}}_{k+1} = \mathbf{U}\mathbf{x}_k \quad ,$$

这就是构成多元时空序列前向转移矩阵$\mathbf{U}$的基础。

总之，从以上分析可以得出下列两点主要结论：

1．通常任何一种地质过程大都是连续参数非平稳的马尔科夫过程，但是由于：

（1）一个过程之所以表现出非平稳性是因为其中存在着一种趋势（或者说随机函数$\{x(t),\ t\in T\}$的数学期望不是常数），如果将此趋势从原始数据中减去，则剩余部分将构

成一个平稳过程，或者说，表征一个地质过程的随机函数的数学期望不是常数，对于过程的研究而言是非本质的，因为，如果将原始随机函数$x(t)$中心化以后，它的数学期望将恒等于零，因而就满足数学期望等于常数的条件。所以，如果随机过程非平稳，只是由于存在一种趋势或者说只是由于随机函数的数学期望不是常数，那末这并不妨碍把它作为平稳过程来进行研究。

(2) 从本节的具体分析可见，且仅当时间齐次的假设成立时，才有可能实际解决转移矩阵的计算问题。

由于以上两方面的考虑，因而把地质上连续非平稳马尔科夫过程近似地当作离散参数平稳马尔科夫链来进行研究。

2．既然在齐次条件下，柯尔莫洛夫前向方程和后向方程有共同解，而过程总是向正的时间方向发展，所以用前向方程进行概率统计分析。首先将由前向方程直接得出前向转移矩阵的表达式。

第四节　多元时空序列转移矩阵U

一、序列相关与自相关

如何正确估计一个马尔科夫概型的转移概率是应用马尔科夫概型分析的一个重要问题，为了便于对这一问题进行讨论，先对序列相关和自相关作一初步介绍。

通常将序列相关定义为一个时间序列中的当前值和另一个时间序列中的滞后值之间的相关。

设有x_i和x_j两个变量（i，$j=1$，2，$\cdots$，p），各自构成一个时间序列：$x_{i,k}(k=1,2,\cdots,n)$和$x_{j,k}(k=1,2,\cdots,n)$，则滞后为s时，这两个序列之间的序列相关系数$r_{(ij)_s}$应为：

$$r_{(ij)_s}=\frac{\operatorname{cov}(x_{i,k},\ x_{j,k-s})}{(\operatorname{var}x_{i,k}\cdot\operatorname{var}x_{j,k-s})^{\frac{1}{2}}},\qquad(9\text{-}27)$$

或

$$r_{(ij)_s}=\frac{\sum_{k=s+1}^{n}(x_{i,k}-\overline{x_i})(x_{j,k-s}-\overline{x_j})}{\left[\sum_{k=s+1}^{n}(x_{i,k}-\overline{x_i})^2\cdot\sum_{k=s+1}^{n}(x_{j,k-s}-\overline{x_j})^2\right]^{\frac{1}{2}}}$$

$$=\frac{\sum_{k=s+1}^{n}x_{i,k}x_{j,k-s}-\frac{1}{n-s}\sum_{k=s+1}^{n}x_{i,k}\sum_{k=s+1}^{n}x_{j,k-s}}{\left[\left\{\sum_{k=s+1}^{n}x_{i,k}^2-\frac{1}{n-s}\left(\sum_{k=s+1}^{n}x_{i,k}\right)^2\right\}\left\{\sum_{k=s+1}^{n}x_{j,k-s}-\frac{1}{n-s}\left(\sum_{k=s+1}^{n}x_{j,k-s}\right)^2\right\}\right]^{\frac{1}{2}}}$$

(9-28)

显然，理论上序列相关系数应随滞后l的增大而逐渐减小（图9-2）。

自相关可以定义为一个时间序列中的每一个值与同一时间序列中滞后值之间的相关。由于是同一个序列，序列$x_{i,k}$和$x_{i,k-s}$中的数列是相同的，因而$\operatorname{var}x_{i,k}=\operatorname{var}x_{i,k-s}$，(9-27)

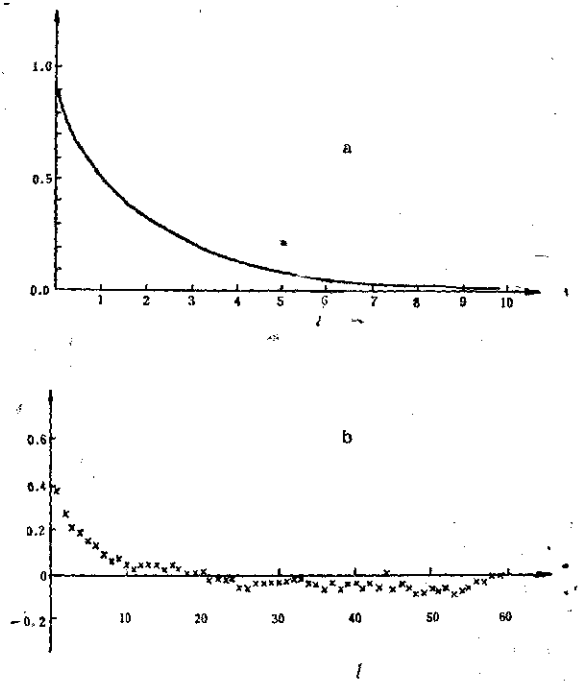


图9-2 a. r_1-l理论曲线 b. r_1-l关系的一个实例

式就变成

$$r_{(ii)s}=\frac{\mathrm{cov}(x_{i,t,k},\ x_{(i,t,k-s)})}{\mathrm{var}(x_{i,t,k})}\ 。 \tag{9-29}$$

故由（9-28）式可得自相关系数的公式

$$r_{(ii)s}=\frac{\sum_{k=s+1}^{n}(x_{i,t,k}-\overline{x_i})(x_{i,t,k-s}-\overline{x_i})}{\sum_{k=s+1}^{n}(x_{i,t,k}-\overline{x_i})^2}$$

$$=\frac{\sum_{k=s+1}^{n}x_{i,t,k}x_{i,t,k-s}-\frac{1}{n-s}\left(\sum_{k=s+1}^{n}x_{i,t,k}\right)\left(\sum_{k=s+1}^{n}x_{i,t,k-s}\right)}{\sum_{k=s+1}^{n}x_{i,t,k}^2-\frac{1}{n-s}\left(\sum_{k=s+1}^{n}x_{i,t,k}\right)^2}\ 。 \tag{9-30}$$

必须着重指出，自相关系数与滞后之间的关系可以作为检验一个时间序列是否具有马

尔科夫性质的重要准则：如果滞后为1时的自相关系数显著较大，并且随着滞后的增大而迅速减小，然后在其平均值附近波动，则所研究的序列具有马尔科夫性质。

二、转移矩阵U的构成

第二节中已经提到在一个马尔科夫链中，体系状态的转移可以用一个“转移矩阵”来表达。

设有一p变量时间序列，如果对每个变量相继地进行n次观测，则所得观测数据$\{x_{ik}: i=1, 2, \cdots, p; k=1, 2, \cdots, n\}$可以用一个观测矩阵X表示：

$$\mathbf{X}=\begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1n} \\ x_{21} & x_{22} & \cdots & x_{2n} \\ \vdots & \vdots & & \vdots \\ x_{p1} & x_{p2} & \cdots & x_{pn} \end{bmatrix}。 \tag{9-31}$$

此矩阵包含n个列向量$\mathbf{x}_k(k=1, 2, \cdots, n)$。

转移矩阵P中的任意一个转移概率p_{ij}指出时间序列中时刻k($k=1, 2, \cdots, n$)的观测值$x_{i,k}(i=1, 2, \cdots, p)$可以期望出现在下一时刻$k+1$的观测值$x_{j,k+1}(j=1, 2, \cdots, p)$中的分数；换言之，$x_{i,k}$的$p_{ij}$部分转移到了$x_{j,k+1}$，即$p_{ij}x_{i,k}$是$x_{j,k+1}$中的可预测部分。因而为了确定转移概率P，必须知道任意一个观测值$x_{i,k}$的百分之多少转移到了观测值$x_{j,k+1}$。在第二节中由观测矩阵直接计算转移频率得到转移频率矩阵，但是在地质上许多多元时间序列虽然时间参数是离散的，然而状态空间是连续的，即地质上的许多多元时间序列属连续型马尔科夫过程，所以不能用离散参数马尔科夫链中所采用的上述方法来构造转移矩阵。今设前向转移矩U为

$$\mathbf{U}=\begin{bmatrix} u_{11} & u_{12} & \cdots & u_{1p} \\ u_{21} & u_{22} & \cdots & u_{2p} \\ \vdots & \vdots & & \vdots \\ u_{p1} & u_{p2} & \cdots & u_{pp} \end{bmatrix}, \tag{9-32}$$

其中的元素$u_{ij}(i, j=1, 2, \cdots, p)$为转移概率。

根据前面所说转移概率矩阵中每一转移概率的意义，显然下列矩阵方程成立：

$$\begin{bmatrix} x_{1,k+1} \\ x_{2,k+1} \\ \vdots \\ x_{p,k+1} \end{bmatrix}=\begin{bmatrix} u_{11} & u_{12} & \cdots & u_{1p} \\ u_{21} & u_{22} & \cdots & u_{2p} \\ \vdots & \vdots & & \vdots \\ u_{p1} & u_{p2} & \cdots & u_{pp} \end{bmatrix}\cdot\begin{bmatrix} x_{1,k} \\ x_{2,k} \\ \vdots \\ x_{p,k} \end{bmatrix}+\begin{bmatrix} e_{1,k+1} \\ e_{2,k+1} \\ \\ e_{p,k+1} \end{bmatrix} \tag{9-33}$$

写成矩阵形式：

$$\mathbf{x}_{k+1}=\mathbf{U}\mathbf{x}_k+e_{k+1}。 \tag{9-34}$$

将(9-33)式展开，得

$$\begin{cases} x_{1,k+1}=u_{11}x_{1,k}+u_{12}x_{2,k}+\cdots+u_{1p}x_{p,k}+e_{1,k+1} \\ x_{2,k+1}=u_{21}x_{1,k}+u_{22}x_{2,k}+\cdots+u_{2p}x_{p,k}+e_{2,k+1} \\ \cdots\quad\cdots\quad\cdots\quad\cdots\quad\cdots \\ x_{p,k+1}=u_{p1}x_{1,k}+u_{p2}x_{2,k}+\cdots+u_{pp}x_{p,k}+e_{p,k+1} \end{cases}。 \tag{9-35}$$

方程组(9-35)描述了所研究体系的多元马尔科夫概型。$e_{i,k+1}(i=1, 2, \cdots, p)$为随机数，

它们统计独立于除$x_{i,k+1}$以外的所有观测值。

假定$x_{i,k}$和$e_{i,k}$的平均值为零。

(9-35)式中的每一个方程指出$x_{i,k}$的u_{ij}部分转移到了$x_{j,k+1}$，即$u_{ij}x_{i,k}$是$x_{j,k+1}$中的可预测部分。

矩阵U和矩阵P不同的是：对于矩阵U①u_{ij}可以是负数；②$\sum_j u_{ij}$不一定等于1，并且一般矩阵U是不对称的，即$u_{ij}\neq u_{ji}$。

三、转移概率的估计

时间序列分析的一个重要问题是如何对概型参数作出正确估计。对于马尔科夫概型来说，就是如何正确估计转移概率u_{ij}。

1．单变量马尔科夫概型的参数估计

设有一序列x_1，x_2，…，x_n满足下列单变量马尔科夫概型：

$$x_{k+1}=ux_k+e_{k+1}\ , \tag{9-36}$$

其中e_{k+1}为独立正态变量，其均值为零，方差为σ_e^2，则可用x_{k+1}对x_k的线性回归或用最大似然法得到u的估计：

$$\hat{u}=\left(\sum_{k=1}^{n}x_{k+1}x_k\right)\Big/\sum_{k=1}^{n}x_k^2\ 。 \tag{9-37}$$

可以证明$\hat{u}$就是第一自相关系数ρ_1(即滞后为1的自相关系数)的最佳估计。

2．多变量马尔科夫概型的参数估计

在第四节之二中已经得到矩阵方程

$$\mathbf{x}_{k+1}=\mathbf{U}\mathbf{x}_k+\mathbf{e}_{k+1}\ ,$$

上式两边各右乘以行向量x'_k，得

$$\mathbf{x}_{k+1}\mathbf{x}'_k=\mathbf{U}\mathbf{x}_k\mathbf{x}'_k+\mathbf{e}_{k+1}\mathbf{x}'_k\ ,$$

将上式两边对n个观测值求和，并且由于$\mathbf{e}_{k+1}$统计独立于$\mathbf{x}_k$，得

$$\sum_{k=1}^{n}\mathbf{x}_{k+1}\mathbf{x}'_k=\mathbf{U}\sum_{k=1}^{n}\mathbf{x}_k\mathbf{x}'_k\ , \tag{9-38}$$

令$\mathbf{x}_k$，$\mathbf{x}_{k+1}$，$\mathbf{x}'_k$各向量中的元素均中心化，即

$$\mathbf{x}_k=\begin{bmatrix}x_{1,k}-\bar{x}_1\\x_{2,k}-\bar{x}_2\\\vdots\\x_{p,k}-\bar{x}_p\end{bmatrix},\quad \mathbf{x}_{k+1}=\begin{bmatrix}x_{1,k+1}-\bar{x}_1\\x_{2,k+1}-\bar{x}_2\\\cdots\\x_{p,k+1}-\bar{x}_p\end{bmatrix},$$

$$x'_k=[x_{1,k}-\bar{x}_1,\ x_{2,k}-\bar{x}_2,\ \cdots,\ x_{pk}-\bar{x}_p]\ 。$$

末项校正为：对于$k+1>n$，$x_{i,n+k}=x_{i,k}(i=1,\ 2,\ \cdots,\ p;k=1,\ 2,\ \cdots,\ n)$。

观测值的方差-协方差矩阵为

$$\mathbf{C}_0=\frac{1}{n-1}\sum_{k=1}^{n}\mathbf{x}_k\mathbf{x}'_k\ 。 \tag{9-39}$$

观测值$x_{i,k}$和滞后为1的观测值$x_{i,k+1}$之间的方差-协方差矩阵为

$$C_1=\frac{1}{n-1}\sum_{k=1}^{n}x_{k+1}x_k' \quad , \tag{9-40}$$

故（9-38）式可写成

$$C_1=UC_0 \quad 。 \tag{9-41}$$

或

$$U=C_1C_0^{-1} \quad 。$$

第五节　转移矩阵U的谱分解

应用马尔科夫概型分析的最终目的是要阐明地质过程的演化趋势。由于地质上时空序列的观测值不可避免地包含着某种程度的随机起伏，因而用整个转移矩阵U推导而得的过程模型会偏离真实的地质过程。为了查明地质过程的演化趋势，可以利用谱分解的方法将转移矩阵U分解，提取优势部分，舍去次要组成或反映某种随时空区间的增大而迅速消逝的系统变化的那些组成。

为了分析问题的需要，先引进左、右特征向量的概念。

由于一个矩阵的行列式等于其转置矩阵的行列式，容易得出U'与U的特征值相等，因此有

$$U't=\lambda t \quad ,$$

其中t是U'对应于特征值λ的特征向量。因而

$$t'U=\lambda t' \quad , \tag{9-42}$$

其中t'称为U的对应于λ的“左特征向量”，而通常方程

$$Uv=\lambda v \tag{9-43}$$

中的特征向量v称为U对应于λ的“右特征向量”。

下面讨论谱分解方法。

设有一马尔科夫过程，具有p个状态，因而转移矩阵U为$p\times p$阶的矩阵。现在讨论谱分解的一种最简单的情形。假定特征方程

$$|U-\lambda I|=0$$

的所有各根全不相同。分别对每一个λ_i定义一个右向量v_i和左向量t_i'，即

$$Uv_i=\lambda_i v_i \quad , \quad t_i'U=\lambda_i t_i' \qquad (i=1,\ 2,\ \cdots,\ p) \quad 。 \tag{9-44}$$

或者，如果

$$V=[v_1,\ v_2,\ \cdots,\ v_p],\ T=[t_1,\ t_2,\ \cdots,\ t_p],\ T'=\begin{bmatrix} t_1' \\ t_2' \\ \vdots \\ t_p' \end{bmatrix} ,$$

则由（9-44）立即可得

$$UV=V\Lambda,\ T'U=\Lambda T' \quad , \tag{9-45}$$

其中Λ为特征值λ的对角阵

$$\Lambda=\begin{bmatrix}\lambda_1 & & 0\\ & \lambda_2 & \\ & & \ddots \\ 0 & & \lambda_p\end{bmatrix}。 \tag{9-46}$$

由（9-45）式可得U的另一种表达式：

$$U=V\Lambda V^{-1}=(T')^{-1}\Lambda T'。 \tag{9-47}$$

由方程（9-44）可见特征向量v_i和t_j'满足下列关系：

$$t_j'(Uv_i)=t_j'(\lambda_i v_i)=(t_j'\lambda_i)v_i \qquad (j=1,2,\cdots,p)。 \tag{9-48}$$

若诸特征值不同，则由（9-48）式得

$$t_j'v_i=0 \qquad (j\neq i,\ \lambda_j\neq\lambda_i)。 \tag{9-49}$$

因为$\lambda_i\neq\lambda_j$，因而使$t_j'(\lambda_i v_i)=(t_j'\lambda_j)v_i$成立，除非$t_j'v_i=0$。其次，可以选择$t_i'$和$v_i$的标度，使

$$t_i'v_i=1， \tag{9-50}$$

因而

$$T'V=I。 \tag{9-51}$$

从而（9-47）式变为

$$U=V\Lambda T'=\sum_{i=1}^{p}\lambda_i v_i t_i'，$$

或

$$U=\sum_{i=1}^{p}\lambda_i A_i， \tag{9-52}$$

其中 $A_i=v_i t_i'$。以3×3的矩阵U为例，（9-52）式中

$$U=V\Lambda T'=\begin{bmatrix}v_1 v_2 v_3\end{bmatrix}\begin{bmatrix}\lambda_1 & 0 & 0\\ 0 & \lambda_2 & 0\\ 0 & 0 & \lambda_3\end{bmatrix}\cdot\begin{bmatrix}t_1'\\ t_2'\\ t_3'\end{bmatrix}$$

$$=\begin{bmatrix}v_{11} & v_{21} & v_{31}\\ v_{12} & v_{22} & v_{32}\\ v_{13} & v_{23} & v_{33}\end{bmatrix}\cdot\begin{bmatrix}\lambda_1 & 0 & 0\\ 0 & \lambda_2 & 0\\ 0 & 0 & \lambda_3\end{bmatrix}\cdot\begin{bmatrix}t_{11} & t_{12} & t_{13}\\ t_{21} & t_{22} & t_{23}\\ t_{31} & t_{32} & t_{33}\end{bmatrix}。$$

由此得

$$U=\lambda_1\begin{bmatrix}v_{11}\\ v_{12}\\ v_{13}\end{bmatrix}\cdot\begin{bmatrix}t_{11}t_{12}t_{13}\end{bmatrix}+\lambda_2\begin{bmatrix}v_{21}\\ v_{22}\\ v_{23}\end{bmatrix}\cdot\begin{bmatrix}t_{21}t_{22}t_{23}\end{bmatrix}$$

$$+\lambda_3\begin{bmatrix}v_{31}\\ v_{32}\\ v_{33}\end{bmatrix}\cdot\begin{bmatrix}t_{31}t_{32}t_{33}\end{bmatrix}，$$

或 $$U=\lambda_1 v_1 t_1' + \lambda_2 v_2 t_2' + \lambda_3 v_3 t_3' \text{ 。}$$

推广到$p\times p$矩阵得

$$U=\sum_{i=1}^{p}\lambda_i v_i t_i' \quad 或 \quad U=\sum_{i=1}^{p}\lambda_i A_i \text{。}$$

(9-52) 式指出，可以将转移矩阵U分解成各部分$U_i=\lambda_i v_i t_i'$ $(i=1, 2, \cdots, p)$。

(9-52) 式中诸矩阵

$$A_i = v_i t_i' \tag{9-53}$$

构成一个“谱集合”，它具有三性质：

$$A_i A_j = \begin{cases} 0 & (i\neq j), \\ A_i & (i=j), \end{cases}$$
$$\sum_{i=1}^{p} A_i = I \text{。} \tag{9-54}$$

诸矩阵A_i的这个集合之所以称为一个“谱集合”是因为它将转移矩阵U分解成各个组成部分，就像一个棱镜或者一面光栅将多色光谱分解成不同波长的谱线一样。(9-54) 式的三个性质分别保证谱集合的元素的选择作用是不叠加、等效和无遗漏的。这就进一步指出，(9-52) 式不仅将U分解为p个组成部分$U=\lambda_i v_i t_i'$，并且这p个组成部分既不重复，又无遗漏，且属等效；因而可以对每一个组成部分分别独立地加以考虑。这是将从转移矩阵U中提取其优势组成U_1，用它来反映马尔科夫过程优势倾向的理论根据。

最后，还可以利用 (9-52) 式立即写出U的n次幂表达式：

$$U^n=\sum_{i=1}^{p}\lambda_i^n A_i^n \text{。}$$

据 (9-54) 式性质2

$$A_i^n = A_i \text{，}$$

因此

$$U^n=\sum_{i=1}^{p}\lambda_i^n A_i \text{。} \tag{9-55}$$

至此，可以利用 (9-55) 式，根据 (9-6) 式由初始分布和转移矩阵来计算任意一阶段的绝对概率。

第六节　过程的演化趋势

一、过程的优势倾向

第五节中已经指出，利用矩阵的谱分解，可以将转移矩阵U分解为各组成部分，并且可对每一部分分别加以考虑：

$$U=\sum_{i=1}^{p}\lambda_i v_i t_i' \text{。}$$

而U的优势成分则为

$$U_1 = \lambda_1 v_1 t_1' \quad , \tag{9-56}$$

其中λ_1为矩阵U的最大特征值，而v_1，t_1'分别为对应于λ_1的右特征向量和左特征向量。U的优势成分反映随机过程的优势倾向。现对此进行几何解释。

p个变量$x_i(i=1, 2, \cdots, p)$的n个观测值$x_{i,k}(i=1, 2, \cdots, p; k=1, 2, \cdots, n)$可以用$p$维空间中的$n$个点来表示。图9-3中点$x_k$，$x_{k+1}$，$x_{k+2}$是$x_{i,k}$的观测值所构成的马尔科夫链的一部分。现在作一转移矩阵U，使点x_k转移到点Ux_k，后者即为x_{k+1}的可预测部。同样，Ux_{k+1}为x_{k+2}的预测点的位置，而点U^2x_k则为由x_k预测x_{k+2}的预测点的位置。图中单箭头指出马尔科夫链的实测转移，双箭头代表借U推断、预测的转移。

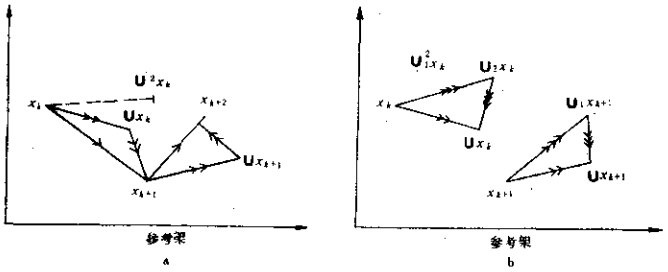


图9-3　转移矩阵u及其优势成分与观测向量关系

实测点与预测点之间的差异是由于随机向量所造成的。如果x值的链相对较短，则可假定由组成转移矩阵U还不能消除所有这些随机变化。此时可从U提取其中的优势成分U_1，以确定变化过程的优势倾向。图9-3b中的三重箭头指出如何由U提取U_1。

优势成分U_1具有下列性质：对于一个已知的点x_k，它给出序列中所有后继各点x_{k+s}（$s \geqslant 1$）的方向。例如U_1x_k指出x_{k+1}的方向；$U_1^2x_k$位于x_k和U_1x_k的直线上，它指出x_{k+2}的方向……$U_1^sx_k$指出x_{k+s}的方向。

一般，连结x_k与U^2x_k的向量和连结x_k与U_1x_k的向量之间的夹角小于连结x_k与U_1x_k的向量之间的夹角。当s增大时，由x_k至U^sx_k的向量和由x_k至U_1x_k的向量之间的夹角趋近于零。

二、过程演化的趋势因子

下面用两种方法来分析反映过程演化的趋势因子。

1．第一种方法

据（9-34）式，一个多元时间序列$x_k(k=1, 2, \cdots, n)$中相继的观测值之间有下列关系：

$$x_k = Ux_{k-1} + e_k \quad 。$$

令yx_k代表所研究诸变量的一个线性组合，其中x_k为列向量，而y为行向量。现在要求将y最优化，使得新变量$yx_k(k=1, 2, \cdots, n)$能够显示过程的“最大演化趋势”。

为此，首先必须对yx_k的这一性质赋予定量的意义。要找出具有最大自相关的yx_k，这

就是说要使下列马尔科夫概型方程中的第一自相关系数ρ_1（滞后为1）为极大：

$$\mathbf{y}x_k = \rho_1 \mathbf{y}x_{k-1} + e_k^* \quad , \tag{9-57}$$

其中$e_k^*(k=2, 3, \cdots, n)$代表随机剩余。

如果将（9-34）式两边左乘$\mathbf{y}$，则得

$$\mathbf{y}x_k = \mathbf{y}\mathbf{U}x_{k-1} + \mathbf{E}_k \quad , \tag{9-58}$$

其中 $\mathbf{E}_k = \mathbf{y}e_k$ 。

现在利用下述极小极大准则。要求向量$\mathbf{y}$中的元素满足以下条件，即（9-57）式中的ρ_1，$\mathbf{y}x_{k-1}$与（9-59）式中的$\mathbf{y}\mathbf{U}x_{k-1}$之差为极小，而$\rho_1$为极大。

显然，如果

$$\mathbf{y}\mathbf{U} = \rho_1 \mathbf{y} \quad , \tag{9-59}$$

则$\rho_1\mathbf{y}x_{k-1}$与$\mathbf{y}\mathbf{U}x_{k-1}$之差为零。此时，随机剩余$e_k^* = \mathbf{E}_k$。显然仅当$\rho_1$是$\mathbf{U}$的一个特征值，且$\mathbf{y}$是相应的特征向量时（9-59）式才能成立。

由于要使ρ_1为极大，所以ρ_1应等于$\mathbf{U}$的最大特征值λ_1，而$\mathbf{y}$则应是相应的左特征向量$\mathbf{t}_1'$。因此$\mathbf{t}_1'\mathbf{x}_k$就是所要求的能够显示过程“最大演化趋势”的新变量。

2．第二种方法

由（9-44）式

$$\mathbf{t}_i'\mathbf{U} = \lambda_i \mathbf{t}_i' \quad ,$$

上述关系对于每一个行向量$\mathbf{t}_i'$都成立，故有

$$\mathbf{t}_i'\mathbf{U}x_k = \lambda_i \mathbf{t}_i' x_k \quad 。 \tag{9-60}$$

在（9-34）式中将k换为$k+1$，得

$$\mathbf{x}_{k+1} = \mathbf{U}x_k + \mathbf{e}_{k+1} \quad 。$$

左乘$\mathbf{t}_i'$，并以（9-60）式代入，得

$$\mathbf{t}_i'\mathbf{x}_{k+1} = \lambda_i \mathbf{t}_i'\mathbf{x}_k + \mathbf{t}_i'\mathbf{e}_{k+1} \quad ,$$

其中$\mathbf{t}_i'\mathbf{e}_{k+1}$为一新随机剩余，记作$e_{k+1}^*$，于是上式可写作

$$\mathbf{t}_i'\mathbf{x}_{k+1} = \lambda_i(\mathbf{t}_i'\mathbf{x}_k) + e_{k+1}^* \quad 。 \tag{9-61}$$

由此可知：

（1）新变量$\mathbf{t}_i'\mathbf{x}_k$是$p$个变量的一个线性组合，可以称之为一个“典型变量”。

（2）每一个典型变量$\mathbf{t}_i'\mathbf{x}_k$都遵循一个参数为$\lambda_i$的单变量马尔科夫概型。

当$i=1$时，（9-61）式变为

$$\mathbf{t}_1'\mathbf{x}_{k+1} = \lambda_1 \mathbf{t}_1'\mathbf{x}_k + e_{k+1}^* \quad , \tag{9-62}$$

其中$\mathbf{t}_1'$即为矩阵$\mathbf{U}$的优势成分$\mathbf{U}_1$对应于最大特征值λ_1的左特征向量。

再将（9-62）式与一般单一变量马尔科夫概型

$$\mathbf{x}_{k+1} = \rho_1 \mathbf{x}_k + \mathbf{e}_{k+1}$$

进行比较，又可知此马尔科夫概型的参数λ_1即为第一自相关系数ρ_1。

一个观测值x_i转移矩阵$\mathbf{U}$从$\mathbf{x}_k$向$\mathbf{U}\mathbf{x}_k$的转移可以看作是沿着p个轴所发生的p个个别转移之和，这p个轴彼此正交，并且无论$\mathbf{x}_k$的位置如何，它们之间是平行的。$\mathbf{U}$的右特征向量

v_1给出了这p个轴的方向余弦，其中对应于最大特征值λ_1的右特征向量v_1则是代表x的最多数转移的那个方向。

根据以上分析可知典型变量t'_1x_t显示多元时间序列的最大序列相关，或者说，它反映演化过程的优势趋势，可以称之为“趋势因子”；而用方程表示趋势因子的p个变量的线性组合则可称之为“趋势因子方程”。

三、应用实例

为了研究我国某超基性岩铜镍矿床的成矿规律和划分成矿阶段及Ni、Co、Cu的空间分布，选择矿区内3403、3601、3803等10个钻孔全部的Ni、Co、Cu、Cr、Ti、B、Sn、Mo、Pb、Zn、S、As等12种元素含量的数据进行了马尔科夫概型分析。其主要结果如下：

1．所研究的10个钻孔的多元时空序列属于马尔科夫概型。证据为：

(1) 原始变量的自相关系数当滞后等于1时的取值显著大于滞后大于1时的取值。见图9-4。

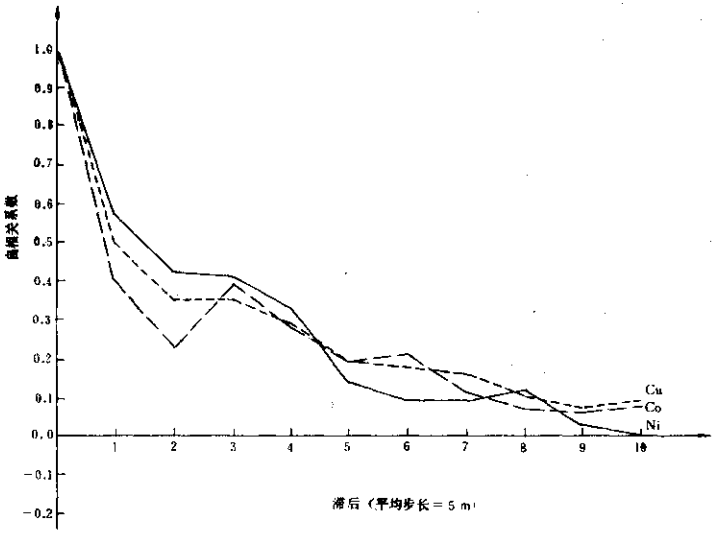


图9-4 Ni、Co、Cu的滞后-自相关关系图

(2) 各钻孔趋势因子t'_1x的第一自相关系数ρ_1与相应转移矩阵U的最大特征值λ_1高度符合。见表9-1。

表9-1 各钻孔中U的最大特征值λ_1与第一自相关系数ρ_1的对比

钻孔号	3103	3301	3803	4003	4401	101	6601	7401	8001
λ_1	0.371	0.853	0.814	0.682	0.907	0.895	0.646	0.861	0.862
ρ_1	0.376	0.862	0.815	0.684	0.925	0.909	0.653	0.864	0.864
λ_1的顺序	9	4	6	7	1	2	8	5	3

2．趋势因子序列的演化趋势反映出本区Ni、Co、Cu矿化的基本特征

从矿床A、B组矿化与趋势因子$t'_{1}x$值的关系来看：所研究的钻孔中A组矿化地段趋势因子$t'_{1}x$均为负值；B组矿化与无矿地段趋势因子$t'_{1}x$均为正值。因此，可以利用A、B组矿化与趋势因子$t'_{1}x$值的上述关系来判断和鉴别钻孔中所见到的矿化类型，它也反映了本矿区的成矿规律（图9-5）。

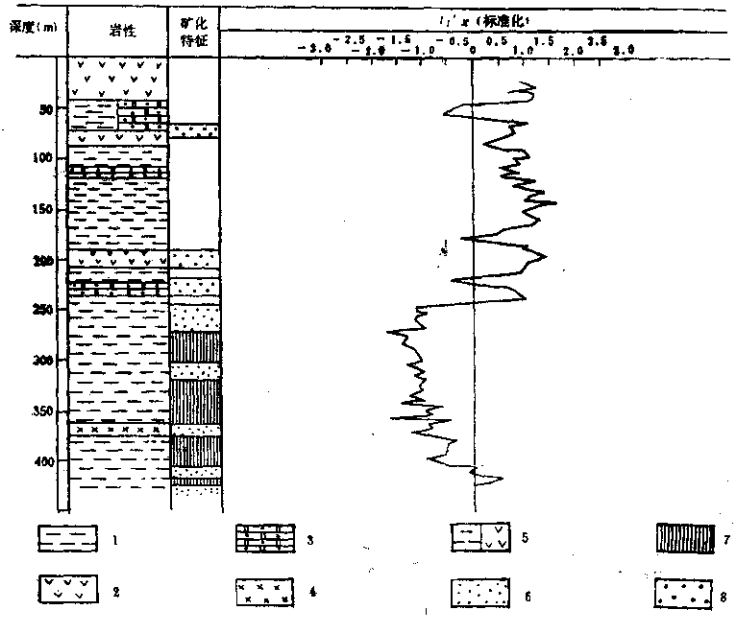


图9-5　3601孔马尔科夫概型趋势图

1．滑铁岩；2．菱铁岩；3．铁白云岩；4．钠长斑岩；5．滑铁岩＋菱铁岩；6．A组矿化；7．A组矿体；8．B组矿体。

第十章　地质统计学(简介)

第一节　概　述

地质统计学又称克里格法，1951年由克里格（D.G.Krige）提出，1962年由马特隆（G.Matheron）完善。

地质统计学原以矿石品位和矿床储量的精确估计为主要目的，近年来被引进到地球化学中，用以研究地球化学变量的空间展布，在环境污染，天气的变化趋势等方面也有应用。地质统计学对于区域化变量空间分布规律的研究是一种十分有效的方法。

地质统计学以变量的空间结构（空间相关）为基础，以区域化变量为核心，以变异函数为基本工具，是数学地质中一种较新的理论和方法。

地质统计学有以下几个优点：

1．对真值估计使用的方法是一种线性无偏最优估计方法。任何一种储量计算方法都是对真实值的一种估计，克里格法以最简单一次线性方程，对均值估计是无偏的，并给出最小估计方差，也就是说，它是一种线性无偏最优估计，是一种精确的估值方法。

2．克里格法在估计块段品位时随参予估值样品的空间位置不同，给样品以不同的权系数。如图10-1，用7个钻孔样品估计方块V的品位时，给各钻孔样品z_i以权系数λ_i。图中权系数λ_i有以下关系：

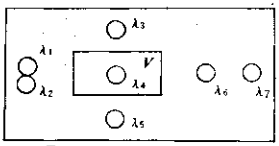


图10-1　块段V的克里格构形

（1）对称关系。$\lambda_3=\lambda_5$。

（2）距离估计区V的中心愈近其权系数愈大$\lambda_4>\lambda_i$（$i\neq4$）。

（3）z_6屏蔽了z_7，因而z_7的权系数λ_7很小。

（4）z_1部分屏蔽z_2，z_2部分屏蔽z_1，z_1+z_2组合在一起与z_6对称，有$\lambda_1+\lambda_2\approx\lambda_6$。

（5）权系数总和为1，$\sum_{i=1}^{7}\lambda_i=1$。

3．克里格法可以计算不同几何域的离散方差。

4．克里格法可以给出计算方法本身的估计方差，以了解估值的可靠程度。根据克里格法估计方差还可以评价勘探网度和储量块段划分的合理性等。

地质统计学根据基本假设不同，计算方法也不同，可分为：线性克里格法、随机克里格法、对数克里格法、泛克里法和析取克里格法等。使用最多的是线性克里格法，下面主要讨论线性克里格法。

第二节　区域化变量理论

一、区域化变量

当一个变量分布在空间内时，称为“区域化”变量。区域化变量是在空间上有数据的实函数，它在三维空间的每一个点x（x_u，x_v，x_w）都有一个值，如样品的品位、矿层的厚度、地形观测值、物化探观测值等。

区域化变量具有两重性：

（1）随机性　区域化变量具有局部的、随机的、不规则的一面，其中包含着随机变量的概念。

（2）结构性　区域化变量具有整体的(或平均的)结构化的一面。即存在着某种空间自相关，可以用某一函数来表示。结构性表现在不同的方向上具有不同的空间自相关。

由于区域化变量具有结构性的一面，因此研究时必须注意到样品的几何形态。区域化变量的几何形态可以用样品的体积、形状、规格和方法（统称“支撑”）来确定。在一个研究区内，如果支撑改变了，就会得出新的区域化变量。区域化变量的重要数字特征有：均值（数学期望）、方差、协方差和变异函数。变异函数和协方差是区域化变量的结构函数。随机变量的均值、方差和协方差的定义是众所周知的。变异函数定义为区域化变量z（x）增量平方的期望值：

$$2\gamma(x_1, x_2) = \mathrm{E}\{[z(x_1) - z(x_2)]^2\} 。$$

而

$$\gamma(x_1, x_2) = \frac{1}{2}\mathrm{E}\{[z(x_1) - z(x_2)]^2\} \tag{10-1}$$

称为半变异函数。

二、平稳性假设

区域化变量是一个随机变量。为了进行统计推断，必须有随机变量$z(x)$的若干现实。在实际地质工作中，同一地点只有一个这样的现实，即在同一空间一般只能采样一次。为了克服这个困难，可以假设：这些随机变量不依赖于它们的具体位置，对于不同位置x_i的实验数据$z(x_i)$，都可以看成是随机变量$z(x)$一次不同的现实，于是在一个矿化带中就有随机变量$z(x)$的多次现实，则统计推断就变为可能。为此，对区域化变量提出如下平稳性假设：

1．在研究领域内，数学期望是一个常数（m），它不依赖于位置x。即

$$\mathrm{E}\{z(x)\} = m 。$$

2．在研究领域内，空间协方差只依赖于分隔距离向量h。任意两点之间，只要h相同，就有相同的协方差。即

$$C(h) = \mathrm{E}\{z(x+h) \cdot z(x)\} - m^2 。 \tag{10-2}$$

协方差的平稳性，意味着方差和变异函数也具有平稳性：

$$\mathrm{var}\{z(x)\} = \mathrm{E}\{[z(x)-m]^2\} = C(0)\ , \tag{10-3}$$

$$\gamma(h) = \frac{1}{2}\mathrm{E}\{[z(x+h)-z(x)]^2\} = C(0) - C(h)\ 。 \tag{10-4}$$

地质统计学研究的是区域化变量的一、二阶矩，只要假设它满足二阶平稳，一般也就够了。还需指出，平稳与不平稳是相对的。在某一领域内是平稳的，超出这个领域也可能就是不平稳的。实际上，地质统计分析可以确定适当大小的移动领域，使随机变量的数学期望和协方差在该领域内是平稳的，因而有效数据可以进行统计推断。这种领域称为平稳域。

如果方差和协方差不具平稳性，只有变异函数和数学期望具平稳性时，称之为“内蕴假设”。

平稳假设有重要的地质意义，例如在同一个砂卡岩矿床中有一个较大的矿体跨越内、外接触带，内、外接触带的界线有可能将这个矿体划分为两个平稳域。一个矿体浅部和深部也可能是不同的平稳域。不同的平稳域应分别统计和计算。

三、平均值$\bar{\gamma}$和$\bar{C}$的计算

克里格法运算经常用到点半变异函数$\gamma(h)$的平均值$\bar{\gamma}$：

$$\bar{\gamma}(V,\ V') = \frac{1}{VV'}\int_V dx \int_{V'} \gamma(x-x')dx'\ ,$$

其中$\bar{\gamma}(V,\ V')$表示当向量$h=x-x'$的一端x描述域V，而另一端x独立地描述域V'时，函数$\gamma(h)$的平均值。域V，V'可以相等，也可以不相等；可以重合，也可以不重合。若x是三维空间的一个点$(x_u,\ x_v,\ x_w)$，记号$\int_V dx$代表三重积分$\int_{V_1}\int_{V_2}\int_{V_3} dx_u,\ dx_v,\ dx_w$，欲求$\bar{\gamma}(V,\ V')$需要六重积分。为克服这一困难，将域$V$和$V'$分别化成$n$和$n'$个离散点，用数值积分法逼近$\bar{\gamma}(V,\ V')$，即

$$\gamma(V,\ V') = \frac{1}{nn'}\sum_{i=1}^{n}\sum_{j=1}^{n'}\gamma(x_i - x_j')\ , \tag{10-5}$$

其中$x_i - x_j'$为域V离散点x_i与域V'离散点x_j'的距离向量。n与n'可以相等，也可以不等。

经计算，在保证克里格方差损失小于1%的条件下，克里格法所需要的离散点的密度：对于一维的域，10个点；对于二维的域，6×6个点；对于三维的域，4×4×4个点；根据经验，上述点数还可以减少。

仿照上法可计算

$$\bar{C}(V,\ V') = \frac{1}{nn'}\sum_{i=1}^{n}\sum_{j=1}^{n'}C(x_i - x_j')\ 。 \tag{10-6}$$

四、估计方差

区域化变量的估计量$\hat{z}(x)$与区域化变量$z(x)$本身总会有所不同，也就是说估计量一般均有估计误差：

$$R(x) = z(x) - \hat{z}(x)\ 。$$

在平稳假设下，以 $E\{R(x)\}=m_e$（常数）表征平均误差；以 $\mathrm{var}\{R(x)\}=E[z(x)-\hat{z}(x)]^2=\sigma_e^2$（常数）表征误差的离差程度，称为估计方差。

设待估计域为$V(x)$，估计域为$v(x')$，经推算其估计方差可写成下式：

$$\sigma_e^2=\bar{C}(V,V)+\bar{C}(v,v)-2\bar{C}(V,v)$$

或

$$\sigma_e^2=2\bar{\gamma}(V,v)-\bar{\gamma}(V,V)-\gamma(v,v)\ 。 \tag{10-7}$$

五、离散方差

在研究区内，需要了解区域化变量的波动情况，离散方差则可用作区域化变量波动情况的刻划。

设区域V范围内有N个单元v，则单元中区域化变量的平均值$z_v(x_i)$对于区域V的平均值$z_V(x)$的离散情况可用均离差来表征：

$$S^2(x)=\frac{1}{N}\sum_{j=1}^{N}[z_v(x_i)-z_V(x)]^2 \qquad V>v\ ,$$

$S^2(x)$又称实验离散方差。

在平稳假设条件下，区域V内N个单元v的离散方差定义为$S^2(x)$的平稳期望：

$$D^2(v/V)=E\{S^2(x)\}=E\left\{\frac{1}{N}\sum_{j=1}^{N}[z_V(x)-z_v(x)]^2\right\}$$

经推导有：

$$D^2(v/V)=\bar{C}(v,v)-\bar{C}(V,V)$$

或

$$D^2(v/V)=\bar{\gamma}(V,V)-\bar{\gamma}(v,v) \qquad V>v\ 。 \tag{10-8}$$

六、正则化

实际工作中，点数据$z(y)$是极罕见的，通常观测数据$z_V(x)$确定在中心位于点支撑$v(x)$上，此样品的品位$z_v(x)$是点品位$z(y)$在$v(x)$中的平均值：

$$z_v(x)=\frac{1}{N}\sum_{j=1}^{N}z(y_i)\ 。$$

称平均值$z_v(x)$为点变量$z(y)$在体积$v(x)$内的正则化。一个二阶平稳点随机变量$z(y)$的正则化随机函数$z_v(x)$也是二阶平稳的，它具有：

1. $$E\{z_v(x)\}=E\left\{\frac{1}{N}\sum_{i=1}^{N}z(y_i)\right\}$$
$$=\frac{1}{N}\sum_{i=1}^{N}E\{z(y_i)\}=\frac{1}{N}\sum_{i=1}^{N}m=m\ 。$$

2. $$\gamma_v(h)=\frac{1}{2}E\{[z_v(x+h)-z_v(x)]^2\}$$

$$= \bar{\gamma}(v, v_h) - \bar{\gamma}(v, v) \ , \tag{10-9}$$

$$\mathrm{var}\{z_v(x)\} = C_v(0) = \bar{C}(v, v) \ , \tag{10-10}$$

$$C_v(h) = C_v(0) - \gamma_v(h) = \bar{C}(v, v_h) \ , \tag{10-11}$$

其中v_h表示从v平移了向量h的支撑。

当距离$|h|$与支撑v相比甚大时，则得简化的近似公式：

$$\gamma_v(h) \approx \gamma(h) - \bar{\gamma}(v, v) \qquad |h| \gg v \ 。 \tag{10-12}$$

支撑$v = s \times l$中的横截面积s与长度l相比甚小时，则（10-9）和（10-12）式化为：

$$\gamma_l(h) = \bar{\gamma}(l, l_h) - \bar{\gamma}(l, l) \qquad l \gg s \ , \tag{10-13}$$

$$\gamma_l(h) \approx \gamma(h) - \bar{\gamma}(l, l) \qquad |h| \gg l \gg s \ 。 \tag{10-14}$$

$\bar{\gamma}(l, l)$只依赖于正则化支撑长度l而与距离h无关，是一个常数。（10-14）式中的半变异函数如果具有球状模型（见本章第二节）在满足变程$a \geqslant 3l$时，其正则化模型$\gamma_l(h)$可近似地将点模型$\gamma(h)$中的块金C_0减去常数$\bar{\gamma}(l, l)$，正则化的变程a_l近似地等于点变程a，其它参数不变。

实际上，由于现有数据经常确定于非点的支撑v上，因而往往需要构造一个与实测的正则化半变异函数$\gamma_v(h)$相一致的理论点模型$\gamma(h)$。从正则化模型$\gamma_v(h)$推演到点模型$\gamma(h)$等于是模型$\gamma_v(h)$的去正则化。

去正则化的步骤是：首先拟合实验所得的正则化曲线$\gamma^*(h)$得模型$\gamma_v(h)$。再初略确定一个点模型$\gamma'(h)$。然后计算点模型$\gamma'(h)$的正则化理论表达式$\gamma'_v(h)$，并与$\gamma_v(h)$相比较，再调整点模型$\gamma'(h)$的各种参数（如球状模型的变程、基台等），使$\gamma'_v(h)$与$\gamma_v(h)$一致，此时$\gamma'(h)$即为所求的点模型$\gamma(h)$。

钻探中，勘探样支撑v一般长2米，待估块支撑V一般不小于生产单元块，支撑$V \gg v$，在这种情况下可以把支撑v视为点支撑，由勘探样获得的变异函数模型可以作为点模型，由此产生对V的估计误差可以忽略不计。

当点模型确定后，实际需要的确定在某支撑V上的正则化理论表达式$\gamma_V(h)$才容易计算。

区域化变量若确定在点支撑上，可视为随机变量。区域化变量通常确定在非点支撑上，实际上是点随机变量的正则化随机函数。一个二阶平稳的点随机变量，它的正则化随机函数也是二阶平稳的。

第三节　变异函数及结构分析

一、变异函数的性质

根据变异函数的定义，实验半变异函数由下式确定：

$$\gamma^*(h) = \frac{1}{2N(h)} \sum_{j=1}^{N(h)} [z(x_i + h) - z(x_i)]^2 \ , \tag{10-15}$$

式中$N(h)$是以向量h相隔的实验数据对$[z(x), z(x_i + h)]$的个数。

变异函数具有以下性质：

1．可迁现象　半变异函数一般随矢量h的增大而增大。当$|h|\geqslant a$时，半变异函数不再增大，而是稳定在一个极限值附近。这一性质称为跃迁性，距离a称为变程，变程a对于特定研究区的特定方向是一个定数。半变异函数稳定的极限值称为“基台值”，它就是区域化变量的先验方差：

$$\gamma(\infty)=\mathrm{var}\{z(x)\}=C(0)\ 。$$

这种型式的变异函数可用基台值和变程来刻划，称为“跃迁型”。

2．影响范围一定　在跃迁型中，落在$|h|\leqslant a$内的任何数据$z(x_i)$和$z(x_i+h)$之间互相相关，当距离$|h|\geqslant a$，区域化变量$z(x)$和$z(x+h)$不再相关。

3．异向性和各向同性　当矿化现象在每一方向上呈现出不同的变异称为各向异性。跃迁型的各向异性表现为不同方向上有不同的变程或不同的基台值。当矿化现象在每一方向上的变异性，只与距离的模$|h|$有关时，称为各向同性。

4．矿化的连续性　区域化变量$z(x)$在空间中的连续性与$\gamma(h)$在原点附近的性状有关，按连续性递降次序可分为4种类型，见图10-2。

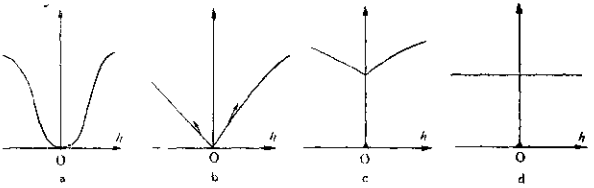


图10-2　原点附近变异函数的性状

a．抛物线性状；b．线性性状；c．块金效应；d．纯块金效应

二、变异函数的理论模型

据（10-15）式建立的变异曲线，由于数据有限，实际为非平滑的锯齿状曲线，称为“实验变异曲线”。在实际工作中要按照具体情况和观测数据选用一种数学模型对它进行模拟，以得出一条平滑的用确定的数学式可以表示的曲线，这就是变异函数的理论模型。为了使理论模型论述上的方便，这里暂时不考虑块金效应，而只考虑各向同性的和基台值为1的标准化随机变量。

三维空间内各向同性记为$\gamma(|h|)=\gamma(r)$。

其中

$$r=\sqrt{h_x^2+h_y^2+h_z^2}\ 。$$

而h_x，h_y，h_z是向量h的三个分量。下面分别叙述几种理论模型：

1．有基台的模型（跃迁型）

（1）球状模型

$$\gamma(r)=\begin{cases}\dfrac{3r}{2a}-\dfrac{r^3}{2a^3} & 0\leqslant r<a\ ,\\ 1=\text{基台} & r\geqslant a\ 。\end{cases}\qquad(10\text{-}16)$$

（2）指数模型

$$\gamma(r)=1-\exp\left(-\frac{r}{a}\right)\quad。\tag{10-17}$$

当实际变程为$a'=3a$时，$\gamma(a')=1-e^{-3}=0.95\approx1$。

（3）高斯模型

$$\gamma(r)=1-\exp\left(-\frac{r^2}{a^2}\right)\quad。\tag{10-18}$$

当实际变程为$a'=\sqrt{3}\,a$时，$\gamma(a')=0.95\approx1$。

有基台模型见图10-3。

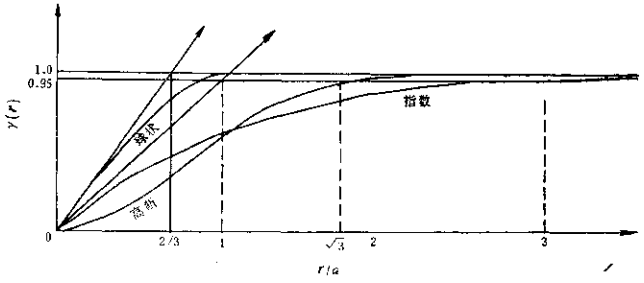


图10-3 有基台的变异函数理论模型

2．无基台的模型 这些模型对应于不存在有限方差和协方差的随机变量$z(x)$，其变异函数具有：

（1）r^θ型模型

$$\gamma(r)=\omega r^\theta\qquad 0<\theta<2\ ,$$

式中 ω、θ为常数。

（2）对数模型

$$\gamma(r)=\log r+b\qquad r\geqslant1\ ,$$

式中b为常数。

此外，还有空穴效应模型及其它模型，但都不常用，常用的是有基台的模型。至于用指数模型或球状模型只是一个习惯问题，并不影响计算结果。

三、结构分析

1．各向同性和各向异性 $\gamma(h)$是向量h的函数。当函数$\gamma(h)$只依赖于向量h的模$|h|$，则称为“各向同性”，当$\gamma(h)$随方向而异，则称为“各向异性”。

（1）几何异向性 当仅用坐标的线性变换将异向性约简为各向同性时，则说明$\gamma(h)$及$C(h)$具有几何异向性。

例如，h_1和h_2为沿空间两个特定方向上测得的距离，在该二方向上计算的变异函数$\gamma_1(h_1)$和$\gamma_2(h_2)$可以表示为两个球状模型，它们的基台值相同（均为1），但变程分别为a_1和a_2。将这种各向异性的$\gamma_1(h_1)$和$\gamma_2(h_2)$转换为各向同性，只要以比值a_2/a_1乘h_1，或以比值a_1/a_2乘h_2即可。如：

$$\gamma_1(h_1)=\frac{3}{2a_1}h_1-\frac{1}{2a_1^3}h_1^2 \qquad h_1\leqslant a_1\ ,$$

$$\gamma_2(h_2)=\frac{3}{2a_2}h_2-\frac{1}{2a_2^3}h_2^2 \qquad h_2\leqslant a_2\ 。$$

令$h_1'=h_1\frac{a_2}{a_1}$，可求得：

$$\gamma_1(h_1)=\frac{3}{2a_1}h_1-\frac{1}{2a_1^3}h_1^3=\frac{3}{2a_1}h_1\frac{a_2}{a_2}-\frac{1}{2a_1^3}h_1^3\frac{a_2^3}{a_2^3}$$

$$=\frac{3}{2a_2}h_1\frac{a_2}{a_1}-\frac{1}{2a_1^3}h_1^3\left(\frac{a_2}{a_1}\right)^3$$

$$=\frac{3}{2a_2}h_1'-\frac{1}{2a_2^3}(h_1')^3 \qquad h_1'\leqslant a_2\ ,$$

即$\gamma_1(h_1)=\gamma_1(h_1')$，两个方向上的变异函数可以用一个变程$a_2$的球状模型来表示。

若4个水平方向有基台值相同，而变程a_1、a_2、a_3、a_4各异的4个半变异函数，它们的变程值的终端可以用一个椭圆来近似，将坐标轴旋转角φ，使之平行于椭圆的主轴，则可以将4个方向的半变异函数约简为与椭圆的主轴同方向的两个半变异函数，进而可转换为各向同性模型。

比例效应：

考虑中心位于两个不同点x_0和x_0'的两个拟平稳域$v(x_0)$与$V(x_0')$，区域化变量的均值分别为$m(x_0)$与$m(x_0')$，令$\gamma(h,\ x_0)$与$\gamma(h,\ x_0')$为确定在这两个邻域的半变异函数，在采矿实践中乘上一个因子后往往使$\gamma(h,\ x_0)$与$\gamma(h,\ x_0')$重合。假定有一个独立于邻区的平稳模型$\gamma_0(h)$，使：

$$\gamma(h,\ x_0)/f[m(x_0)]=\gamma(h,\ x_0')/f[m(x_0')]=\gamma_0(h)\ , \qquad (10\text{-}19)$$

则称$\gamma(h,\ x_0)$与$\gamma(h,\ x_0')$存在比例效应，称（10-19）式为拟平稳模型。

比例效应存在时，可以根据$\gamma(h,\ x)$与$m(x)$的关系，给图解拟平稳的地质统计分析的计算带来方便。令$\lambda=f[m(x)]$，若δ是估计方差、离散方差和正则化中任一个运算，则有：

$$\delta(\lambda\gamma)=\lambda\delta(\gamma)\ ,$$

$$\delta[\gamma(h,\ x_0)]=f[m(x_0)]\delta[\gamma_0(h)]\ ,$$

$$\delta[\gamma(h,\ x_0)]=\frac{f[m(x_0)]}{f[m(x_0')]}\delta[\gamma(h,\ x_0')]\ 。$$

如果有套合结构$\gamma=\sum_i\gamma_i$存在，则每一个组成结构的比例效应可能不同。

（2）带状异向性　带状异向性模型应用最广，因为任何一个不能用坐标的简单线性变换约简的实测异向性就是“带状异向性”。

令$\gamma(h)$为一个表征三维空间内某种变异的套合模式：

$$\gamma(h)=\sum_i\gamma_i(h),$$

其中h是坐标为$(h_u,\ h_v,\ h_w)$的向量，该套合模型中的每一个组成结构$\gamma_i(h)$都可能对h为各向异性。如图10-4，沿垂直方向构成的半变异曲线$\gamma(h_w)$揭示了一种强变异性，而水平

方向为各向同性的半变异曲线$\gamma\left(\sqrt{h_u^2+h_v^2}\right)$则揭示了一种弱变异性。这样，可以认为：在$\gamma(h_w)$中除了包含与水平方向一致的各向同性结构部分之外，还包含有一个异性部分：

$$\gamma_2(h_w)=\gamma(h_w)-\gamma\left(\sqrt{h_u^2+h_v^2}\right)。$$

这样，就可以采用三维套合模型$\gamma(h)$来描绘所研究的区域化变量的变异性：

$$\gamma(h)=\gamma_1(|h|)+\gamma_2(h_w)。\tag{10-20}$$

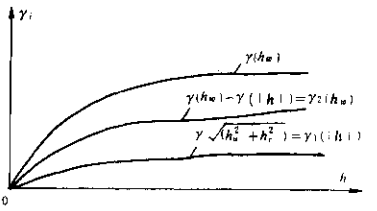


图10-4 带状异向性

2．套合结构 上面讨论了各个方向的变异，这里讨论同方向上由于不同尺度的各种原因引起的变异性。套合结构可以方便地表示为许多变异曲线之和，每一个变异曲线表示一种特定尺度上的变异性：

$$\gamma(h)=\gamma_0(h)+\gamma_1(h)+\cdots+\gamma_n(h)$$

$$=\sum_{i=1}^{n}\gamma_i(h)。\tag{10-21}$$

(10-21)式中的$\gamma_i(h)$均可以是跃迁型或其它型。图10-5中，当距离h稍大于数据支撑时，$\gamma_0(h)$将迅速到达基台值C_0。因此$\gamma_0(h)$兼有所有微观变异性（如岩石物性差异、观测差异）。$\gamma_1(h)$是另一种跃迁模式，例如具有一个表示扁豆状矿体的较大的变程(如a_1=10m)。而$\gamma_2(h)$是第三种跃迁模式，其变程a_2反映了地层交替或均匀矿化带的延伸（如a_2=200m）。

在较小距离上（$|h|<10$m）观测到的总变异性取决于$\gamma_0(h)+\gamma_1(h)$，而对于较大一些距离，总变异性就取决于(10-21)式中所有$\gamma_i(h)$。

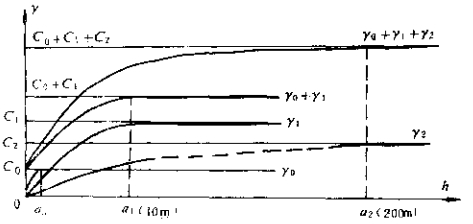


图10-5 套合结构与块金效应

3．块金效应 (10-21)式中，在$h>a_0$的观测标度上，微观结构$\gamma_0(h)$在总变异中以幅度C_0的表现不连续出现在原点处，这就叫块金效应。称C_0为块金常数。

块金效应的含义与观测尺度有着密切的联系。如图10-5所示，一个变程为a_1=10m，基台值为C_1的结构函数$\gamma_1(h)$在小于10m的距离内随h变化，但在30—50m的采样网格上，它就只是一种“块金效应”，在这样大的网格上凭经验无法区分块金常数中的两种参数C_0和C_1。因此在一个规则的网格上取样后，需要在一个更大的网格上取一些样，这就能利用两种不同观测尺度的资料去研究套合结构，并推断某种可能存在的块金效应。

4．结构分析的步骤

（1）由观测数据计算各个基本的实验半变异函数$\gamma_k^*(h)$以及实验平均值m_k^*与实验离散方差S_k^{2*}。

采用二维克里格法时，应在水平方向上计算N-S，E-W，NW-SE，NE-SW 4个方向的$\gamma^*(h)$，在采用三维克里格法时，除上述4个方向外，尚须计算垂直方向上的$\gamma^*(h)$值。当数据列线，但点距不等时，需确定距离h(步长)的容许误差限，而对于非列线之不规则数据还要给出角度误差限，方可计算某方向的$\gamma^*(h)$。此外，可靠性距离$|h|$应小于所考虑的区域L之一半，即$|h|<L/2$，其有效数据对的对数N_h要足够多，如$N_h>30$对，这样计算的实验半变异函数才有意义。

（2）用理论曲线拟合平均实验半变异曲线，建立变异函数的理论模型。

（3）去正则化，因观测数据通常确定在非点支撑上，上述步骤获得的变异函数，实际上是确定在该支撑上的正则化模型，所以必须由这个正则化模型转化为一般结构的点模型。

（4）构成一般结构模型。为了在计算机上运算的方便，将各方向上的半变异函数表为N个各向同性结构套合之和。

$$\gamma(h)=\sum_{i=1}^{N}\gamma_i(h)\ 。 \tag{10-22}$$

当组成结构$\gamma_i(h)$具有特定的异向性时，则可将向量$\mathbf{h}$的初始直角坐标(h_u，h_v，h_w)作线性变换：$[\mathbf{h}_i]=[\mathbf{A}_i][\mathbf{h}]$，然后把它们作为各向同性结构组合到一般结构模型中去。

图10-4中，$\gamma(h)=\gamma_1(|h|)+\gamma_2(h_w)$中的$\gamma_1(|h|)$代表一个三维空间各向同性的结构，可以把它看成是由矢量$\mathbf{h}$经坐标变换后得到的$[\mathbf{h}_1]=[\mathbf{A}_1][h]$，其中：

$$[\mathbf{A}_1]=\begin{pmatrix}1&0&0\\0&1&0\\0&0&1\end{pmatrix}。$$

变换后向量为：

$$h_{1u}=h_u+0\times h_v+0\times h_w=h_u\ ,$$
$$h_{1v}=0\times h_u+h_v+0\times h_w=h_v\ ,$$
$$h_{1w}=0\times h_u+0\times h_v+h_w=h_w\ 。$$

而$\gamma_2(h_w)$表示一个垂直方向上的异向性结构，可以把它看成对于变换后向量$\mathbf{h}_2$的各向同性，即$\gamma_2|h|=\gamma_2(h_w)$。$\mathbf{h}_2$是由矢量$\mathbf{h}$经坐标变换后得到的$[\mathbf{h}_2]=[\mathbf{A}_2][\mathbf{h}]$，其中：

$$\mathbf{A}_2=\begin{pmatrix}0&0&0\\0&0&0\\0&0&1\end{pmatrix}。$$

变换后的向量为：

$$h_{2u}=0\times h_u+0\times h_v+0\times h_w=0\ ,$$
$$h_{2v}=0\times h_u+0\times h_v+0\times h_w=0\ ,$$
$$h_{2w}=0\times h_u+0\times h_v+h_w=h_w\ 。$$

这样，上述一般结构模型可写成各向同性结构套合之和：

$$\gamma(h)=\gamma_1(|h_1|)+\gamma_2(|h_2|)\ 。$$

当存在块金效应时：

$$\gamma(h)=\gamma_0(|h_0|)+\gamma_1(|h_1|)+\gamma_2(|h_2|)\ 。 \tag{10-23}$$

说明一下，h变成为h_1实际上没有变化，但仍然有意义，因为它是由向量$\mathbf{h}$左乘一个矩

阵$[A_1]$得来的，与h左乘$[A_2]$得到h_2的形式相同，可用相同的方法计算$|h_1|$、$|h_2|$，在计算机上实现起来特别方便。

存在比例效应时，尚需计算拟平稳域的比例数应。

第四节　线性克里格法

一、线性克里格法的理论

一般，对整个矿床的总体估计是通过对其逐个块段的局部估计组合而得。局部估计是要找出位于一个平稳域或准平稳域内的一个块段的平均品位的最佳估计，总体估计则是在大于平稳的范围内对整个矿床进行估计。

今要对局部块段$V(x_0)$的平均值$z_V(x_0)$进行估计。所用的实验数据是一组离散的品位值$\{z_\alpha,\ \alpha=1,\ 2,\ \cdots,\ n\}$，被估域$V(x_0)$的克里格估值$\hat{z}_V(x_0)$是这$n$个数值的线性组合：

$$\hat{z}_V(x_0)=\sum_{\alpha=1}^{n}\lambda_\alpha z_\alpha\ 。\tag{10-24}$$

这里品位值z_α确定在支撑v_α上。支撑v_α（$\alpha=1,\ 2,\ \cdots,\ n$）可以是不同类型不同维数的域，如点、线长或体积。

若λ_α为已知，即可据（10-24）式，用z_α估计$\hat{z}_V(x_0)$。现在的问题是如何求解λ_α？

在满足无偏条件（$\sum_\alpha\lambda_\alpha=1$）下，使估计方差$\sigma_E^2=E[z_V-\hat{z}_V]^2$达到最小，可以用拉格朗日乘子求条件极值的方法来解决这个问题。为推导方便，令拉格朗日乘子$\mu'=-2\mu$，使

$$f=E[z_V-\sum_\alpha\lambda_\alpha z_\alpha]^2-2\mu(\sum_\alpha\lambda_\alpha-1)$$

达到极小，令其偏导数为零，即$\frac{\partial f}{\partial\lambda_\alpha}=0$，$\frac{\partial f}{\partial\mu}=0$，经推算得：

$$\begin{cases}\sum_{\beta=1}^{n}\lambda_\beta\bar{C}(v_\alpha,\ v_\beta)-\mu=\bar{C}(v_\alpha,\ V) & (\alpha=1,\ 2,\ \cdots,\ n)\ ,\\ \sum_{\beta=1}^{n}\lambda_\beta=1\ 。\end{cases}\tag{10-25}$$

$$\sigma_E^2=\bar{C}(V,\ V)+\mu-\sum_{\alpha=1}^{n}\lambda_\alpha\bar{C}(v_\alpha,\ V)\ ,\tag{10-26}$$

或

$$\begin{cases}\sum_{\beta=1}^{n}\lambda_\beta\bar{\gamma}(v_\alpha,\ v_\beta)+\mu=\bar{\gamma}(v_\alpha,\ V) & (\alpha=1,\ 2,\ \cdots,\ n)\ ,\\ \sum_{\beta=1}^{n}\lambda_\beta=1\ ,\end{cases}\tag{10-27}$$

$$\sigma_E^2=\sum_{\alpha=1}^{n}\lambda_\alpha\bar{\gamma}(v_\alpha,\ V)+\mu-\bar{\gamma}(V,\ V)\ 。\tag{10-28}$$

在内蕴假设下，协方差$C(h)$、$C(0)$不存在，仍可用（10-27）式计算。但（10-27）式用矩阵形式表示时，系数矩阵主对角线非主元，需用全主元素消去法求解，比较费事，因此，在内蕴假设下，可定义一个伪协方差$C(h)$：

$$C(h)=A-\gamma(h)\ 。$$

为使$C(h)>0$，常数A是满足条件$A>\gamma(h)$的一个任意正数，用$\gamma(h)=A-C(h)$代入（10-27）式，A可消去，（10-27）式就化为（10-25）式。（10-25）式的优点是系数矩阵主对角线为主元，用顺序消去法求解比较简便。

二、克里格方程的矩阵形式

（10-25）式是一组包括（$n+1$）个未知数的$n+1$个线性方程，称之为克里格方程组，它可以表示为矩阵形式：

$$\begin{pmatrix} \overline{C}(v_1,v_1)\cdots\overline{C}(v_1,v_\beta)\cdots\overline{C}(v_1,v_n) & 1 \\ \vdots \quad\quad \vdots \quad\quad \vdots & \vdots \\ \overline{C}(v_\beta,v_1)\cdots\overline{C}(v_\beta,v_\beta)\cdots\overline{C}(v_\beta,v_n) & 1 \\ \vdots \quad\quad \vdots \quad\quad \vdots & \vdots \\ \overline{C}(v_n,v_1)\cdots\overline{C}(v_n,v_\beta)\cdots\overline{C}(v_n,v_n) & 1 \\ 1 \quad\cdots\quad 1 \quad\cdots\quad 1 & 0 \end{pmatrix} \begin{pmatrix} \lambda_1 \\ \vdots \\ \lambda_\beta \\ \vdots \\ \lambda_n \\ -\mu \end{pmatrix} = \begin{pmatrix} \overline{C}(v_1,V) \\ \vdots \\ \overline{C}(v_\beta,V) \\ \vdots \\ \overline{C}(v_n,V) \\ 1 \end{pmatrix} 。 \tag{10-29}$$

简写为$\mathbf{K}\lambda=\mathbf{M}$

其中

$$\mathbf{K}=\begin{pmatrix} \overline{C}(v_1,v_1)\cdots\overline{C}(v_1,v_\beta)\cdots\overline{C}(v_1,v_n) & 1 \\ \vdots \quad\quad \vdots \quad\quad \vdots & \vdots \\ \overline{C}(v_\beta,v_1)\cdots\overline{C}(v_\beta,v_\beta)\cdots\overline{C}(v_\beta,v_n) & 1 \\ \vdots \quad\quad \vdots \quad\quad \vdots & \vdots \\ \overline{C}(v_n,v_1)\cdots\overline{C}(v_n,v_\beta)\cdots\overline{C}(v_n,v_n) & 1 \\ 1 \quad\cdots\quad 1 \quad\cdots\quad 1 & 0 \end{pmatrix},$$

$$\lambda=\begin{pmatrix} \lambda_1 \\ \vdots \\ \lambda_\beta \\ \vdots \\ \lambda_n \\ -\mu \end{pmatrix}, \qquad \mathbf{M}=\begin{pmatrix} \overline{C}(v_1,V) \\ \vdots \\ \overline{C}(v_\beta,V) \\ \vdots \\ \overline{C}(v_n,V) \\ 1 \end{pmatrix}。$$

由此可得：

$$\lambda=\mathbf{K}^{-1}\mathbf{M}\ 。$$

$$\sigma_E^2=\overline{C}(V,V)-\lambda'\mathbf{M}\ 。 \tag{10-30}$$

三、克里格方程组的几点性质

1. 解的唯一性，要求协方差矩阵$[\overline{C}(v_\alpha,v_\beta)]$是严格正定的。

2. 克里格估值是精确的插值，若V与v相重合时，给出①$\hat{z}_V(x_0)=z_\alpha$；②$\sigma_E^2=0$。

3. 克里格方程组和克里格方差只依赖于结构模型$C(h)$或$\gamma(h)$和支撑v_α，v_β，V的相对

几何特征，并不依赖于数据z_α的特定数值。所以，当数据构形已知，则在打钻之前即可解出克里格方程组，并预测相应的最小估计方差。这样，就可以用克里格方差来平衡新的钻探施工成本与实效。

4．$C(h)$或$\gamma(h)$的影响反映在连续性好的方向上求出的权系数大。

5．矩阵K只依赖于估计邻域数据支撑v_α、v_β的相对几何特征，并不依赖于待估域的支撑。因此，两个恒同的数据构形将给出同一个K，从而只需求一次逆矩阵$\mathbf{K}^{-1}$与M或M′之乘积：$\lambda=\mathbf{K}^{-1}\mathbf{M}$，$\lambda'=\mathbf{K}^{-1}\mathbf{M}'$。这样就可以提高程序编排的效率。为此，尽量采用规则网格取样，否则也应采用一种“平均的”数据构形，即“随机克里格法构形”，使数据构形规则化。

在规则数据构形中若有数据缺失的现象，必须填补缺失，恢复完整的数据构形。最简单的填补方法是用邻近数据取算术平均值替补。这种用邻近数据平均值补缺的方法会降低估值的精度，但是计算的估计方差σ_E^2并没有增大。

四、克里格法的实施步骤

1．选择区域化变量　根据要研究的问题，选择区域化变量。例如，金属含量、矿体厚度、含矿率等。

2．对有效数据的审查　审查的目的在于纠正各种差错，并对结果提出解释。审查的内容包括数据的可靠性、代表性、均一性等。

3．结构分析　这是地质统计分析重要的一步，其目的是要构置一个变异函数模型，对于全部有效结构信息作定量化概括，以表征所研究区域化变量的主要特征。

4．确定测区模型的形式和待估块尺寸，最好用三维固定方块模型　这时每一个待估块段应形态相同、大小相等，以便保证M阵恒同，有利于求解克里格方程组。

5．估计邻区的确定　确定估计邻区的主要目的是为了保证在估计中避免产生偏倚，理论上可以根据半变异函数的变程和屏蔽效应来确定。在实际工作中，若采用三维固定方块矿床模型时，估计邻区可取以待估方块为中心的三层共27个块段。

6．计算$\overline{C}$或$\overline{\gamma}$，组成克里格方程组，求解克里格方程组得λ_α，计算局部估计量$\hat{z}_{V_i}$：

$$\hat{z}_V=\sum_{\alpha=1}^{n}\lambda_\alpha z_\alpha \ 。$$

7．总体估计　对于克里格局部估计量大于边界品位的诸块段组合而得总体估计量。

设测区中n个待估块段的集合为$\{V_i;\ i=1,\ 2,\ \cdots,\ n\}$，其相应的局部克里格估计量的集合为$\{\hat{z}_{V_i};\ i=1,\ 2,\ \cdots,\ n\}$确定边界品位$z_C$。设有$N$个待估块满足$\hat{z}_{V_i}>z_C$，于是以上两集合的相应子集$\{V_k;\ k=1,\ 2,\ \cdots,\ N\}$和$\{\hat{z}_{V_k};\ k=1,\ 2,\ \cdots,\ N\}$给出总体估计量，即

$$D=\sum_{k=1}^{N}V_k\ , \tag{10-31}$$

$$\hat{z}_D=\frac{1}{D}\sum_{k=1}^{N}V_k\cdot z_{V_k}\ 。 \tag{10-32}$$

D为矿床中矿石量的总体积，$\hat{z}_D$为区域化变量在D中的平均值。在三维固定方块模型中，

N个支撑V_k均相等，则（10-32）式化为：

$$\hat{z}_D = \frac{1}{N}\sum_{k=1}^{N}\hat{z}_{Vk} \quad 。 \tag{10-33}$$

令W_k为待估块V_k相应的体重，则矿床中矿石总重量为

$$Q = \sum_{k=1}^{N} V_k W_k \quad 。 \tag{10-34}$$

8．总体估计方差的计算　由公式（10-30）计算局部估计量$\hat{z}_{Vk}$的估计方差σ_{Ek}^2。克里格估计方差的全矿平均值$\sigma_{KV}^2 = \frac{1}{N}\sum_{k=1}^{N}\sigma_{ok}$。

第十一章　稳健统计学*

第一节　概　述

统计学家们已经注意到当母体中混有离群数据时的统计方法的稳健性问题。稳健性可以理解为某种统计方法对于离群数据的抵抗能力，即方法对于偏离统计假设情况的适应能力。传统的统计方法都对被研究总体的特性，尤其是分布型式作出一些假设，如假定总体服从正态分布。实际应用中，当数据符合方法所规定的假设时，方法具有最佳性能，计算结果可靠。但是，当数据偏离假设的情况下，常规统计方法适应性差，计算的结果不正确，甚至会导致谬误，这就是说，常规的方法稳健性差。

40年代始，稳健统计学迅速发展，以J.W.Tukey 教授为首的美国普林斯顿大学统计学研究小组开始强调传统方法的不稳健性，并提出了若干可行的替代方法。1964年瑞士学者尤伯（P.J.Huber）发表了“位置参数的稳健估计”，首次回答了如何设计稳健方法的问题，并提出了M估计法。1971年美国学者汉培尔(F.R.Hampel)提出了影响曲线及崩溃点两个重要概念，为稳健性的数学定义奠定了基础。1970—1971年，稳健统计学的创始者们被邀请到普林斯顿大学进行了长达一年的学术讨论。1981年Huber出版了专著《稳健统计学》，该书从数学理论上对稳健统计学进行了系统的总结。目前稳健统计学还在不断地发展和完善。

第二节　稳健统计学的几个概念

一、模糊概念及其意义

在稳健统计学中提出了如位置、尺度、相关性等概念。这些概念是模糊概念，即本身没有明确的数学定义。位置(location)是指一个分布的中心位置或典型值；尺度（scale）又称离散度（spread）是指一个分布的宽窄；而相关性（association）则指两个或多个分布的相关关系。

模糊概念可以通过许多方式明确化，如位置可用算术平均值、中位数等许多统计量来估计；尺度用标准离差或几个分位数的均值来估计；相关性用各种相关系数（如皮尔逊相关系数、秩相关系数等）来估计。

模糊概念的引入实际上是对各种估计值进行了分类，使我们可以在此基础上对同一模

* 周蓉《稳健统计学及其地质应用》1987

糊概念的不同估计方法进行比较，讨论其稳健性，以便得到更可靠的估计值。

二、稳健性

统计方法的稳健性指的是某方法对于数据偏离假设情况的适应能力。在现阶段，主要指原始数据偏离分布模型时的适应能力。如果经验分布显著偏离假设时，统计方法仍能给出可靠的结果，则该方法是稳健的。

统计方法的稳健性取决于它所使用的统计量。例如，估计方法的好坏可用估计量的偏差和方差来衡量。在检验问题中，方法的好坏则用检验水平和功率来衡量。这些都是由估计统计量和检验统计量的分布推算出的特征参数。因此，经验分布偏离理论分布时，统计方法的稳健性就表现为统计量分布的变化。如果它变化小，则称统计方法是稳健的。

三、崩溃点

崩溃点是指某种统计方法中，允许经验分布偏离理论分布的最大界限。也就是说，经验分布不超过这个界限时，统计量的分布受偏离影响不大，结果基本可靠。一旦经验分布超过这个界限，统计量与理论真值的偏差很大，统计结果不可靠，即这种统计方法崩溃，这个界限就叫做崩溃点。粗略地讲，崩溃点就是在不造成破坏性影响的前提下，统计量所能容许的经验分布与理论分布之间的最大偏离。

经验分布一般来说可看作是由理论分布加离群数据而成的污染分布。如果 ε 是样本中污染的百分比。则 ε 可用来代表经验分布对模型分布的偏离程度。这时，渐近崩溃点即为这样的ε^*值，它使得，当$\varepsilon>\varepsilon^*$时，统计量的最大渐近偏差达到最大值。可以证明(Huber, 1981)，样本均值的 $\varepsilon^*=0$，而中位数的 $\varepsilon^*=0.5$。这说明样本均值的稳健性极差，不能容纳任何离群值，而中位数则十分稳健，离群值所占百分比可接近50%。

四、影响曲线

影响曲线也叫影响函数，是稳健统计学中的另一个重要概念，用于描述统计量受不同位置上离群值影响的细节。

设总体分布为F。当样本容量n很大时，样本的分布接近F。如果样本中，在数值为x处混入了比例为 ε 的离群值，则其经验分布为：

$$F_\varepsilon=(1-\varepsilon)F+\varepsilon\delta_x\ , \tag{11-1}$$

其中δ_x表示离群值的分布。这时，统计量T的影响曲线定义为T在x处变化的极限，即：

$$IC(x,F,T)=\lim_{\varepsilon\to 0}\frac{T(F_\varepsilon)-T(F)}{\varepsilon}$$

$$=\frac{d}{d_\varepsilon}T(F_\varepsilon)\Big|_{\varepsilon=0}\ 。 \tag{11-2}$$

这个定义表明，对于给定的分布F和统计量T，影响曲线IC是x的函数，反映了在x处的离群值引起统计量的相对于离群值比例的变化率。换句话说，影响函数表示当样本容量很大时，增加一个值为 x 的离群数时将会引起统计量的多大变化。

五、效率

既然中位数比算术平均值稳健，为什么不用中位数代替算术平均值？这是因为还需要考虑统计量的效率，即数据所含的信息被该统计量所萃取的比例。计算均值时每一个数据都对它有影响，而计算中位数时，只有一个或两个观测值对中位数的大小有影响。

对于给定的分布，某统计量的效率被定义为：

$$\text{效率}=\frac{\text{对该分布可获得的最小方差}}{\text{该统计量的渐近方差}}。$$

在正态分布条件下，样本均值是位置的最小二乘估计，因而具有最小方差。所以，均值在正态分布情况下的效率为1，其它位置估计量的效率都小于1。效率有时候用百分数来表示。

在研究有限样本时可用到效率的概念，这时效率被定义为：

$$\text{效率}=\frac{\text{模型的理论方差}}{\text{实际的实验方差}}。$$

在实际工作中，希望某统计量有较大的效率，但一般来说，效率与稳健性是矛盾的。因为要达到稳健就需要剔除某些被认为是离群的数据，或者给这些观测值较低的权，这就必然降低信息的萃取率。换句话说，稳健性的取得通常以牺牲效率作为代价。这意味着对于某一实际问题，必须选择能达到稳健性和效率之间最理想的平衡。

第三节　位置和尺度的稳健估计

位置和尺度的稳健估计方法可大致归纳为三大类：

1．M估计，即极大似然估计（maximum likelihood estimates）。

2．L估计，即顺序统计量线性组合（linear combinations of order statistics）。

3．R估计，即秩估计（rank estimates）。

一、位置的M估计

1．定义

设某一容量为n的样本$x_1,x_2,\cdots,x_n$的位置记为T，而误差即为$r_i=(x_i-T)$ $(i=1,2,\cdots,n)$。我们知道样本均值是位置的最小二乘估计值，即是使误差的平方和Σr_i^2极小化的解。它是一个很不稳健的统计量，其原因在于r_i值的二次方的形式出现在被极小化的函数中（Σr_i^2）有很大的r_i值的离群数据对Σr_i^2的影响比其它值的影响大得多，要使Σr_i^2极小化，必然要迁就这些离群数据。

那么如果用一个随r_i值的增长速度没有Σr_i^2快的函数$\Sigma\rho(r_i)$来进行极小化，是否可以得到比平均值更稳健的位置估计呢？M估计就是基于这样的想法提出来的。

对于样本$x_1,x_2,\cdots,x_n$，位置的M估计定义为极小化问题：

$$\Sigma\rho(x_i)=\min \tag{11-3}$$

的解；另一种定义是，它为方程

$$\Sigma\psi(r_i)=0 \tag{11-4}$$

的解。如果ψ是ρ的导数，即$\psi(r)=\rho'(r)$时，这两个定义是等价的。

2．计算方法

计算位置的M估计时，还要引入尺度参数s，并将r_i用$z_i=r_i/s$来代替。这是为了保持估计值尺度不变性，即当数据的尺度改变时估计值也发生相同的改变。此时，T为下述方程的解：

$$\sum_{i=1}^{n}\psi(z_i)=\sum_{i=1}^{n}\psi\left(\frac{x_i-T}{S}\right)=0 \quad 。 \tag{11-5}$$

位置的M估计有时被化成加权平均的形式来计算，定义权系数为

$$w_i=\psi(z_i)/z_i \qquad (i=1,2,\cdots,n) \tag{11-6}$$

将（11-6）式代入（11-5）式，得

$$\sum w_i(x_i-T)=0 \ ,$$

其解为：

$$T=\sum w_i x_i/\sum w_i \quad 。 \tag{11-7}$$

可见，位置的M估计可看作是数的某种加权平均值，其权的大小与ψ函数成正比，与z成反比。

具体计算时均用迭代法，首先由给定的初始值$T^{(0)}$和$s^{(0)}$（可用最小二乘或某种简单的稳健估计求出），当迭代收敛时，得到位置和尺度的M估计值。

3．几种常用的M估计

根据ψ函数的不同，常用M估计有：

（1）尤伯（Huber）估计

Huber(1964) 在首次引入M估计的概念时，以第二提案提出了一个后来被广泛采用的M估计，其ψ函数为：

$$\psi(z)=\begin{cases} z & |z|<C, \\ C\cdot\mathrm{sign}(z) & |z|\geqslant C, \end{cases} \tag{11-8}$$

式中$z_i=(x_i-T)/s$，$\mathrm{sign}(z)$指z的符号，而C为调节常数。

调节常数C可用于控制稳健性与效率之间的平衡。C越大，效率越高，但稳健性越差。在正常分布情形下，$C=0.35$时，效率为75%，$C=1.345$时效率为95%，而$C=2.0$时效率为99%。

设样本$x_1,x_2,\cdots,x_n$中，符合$z_i\leqslant -C(x_i\leqslant T-Cs)$的样品数为$n_1$；符合$z_i\geqslant C(x_i\geqslant T+Cs)$的样品数为$n_3$，其余，即$-C<z_i<C(T-Cs<x_i<T+Cs)$的样品数为$n_2=n-(n_1+n_3)$则由（11-5）和（11-8）式可得：

$$\sum_{i=1}^{n}\psi(z)=\sum_{i=n_1+1}^{n-n_3}(x_i-T)/s+(n_3-n_1)C=0 \ ,$$

解出T值为：

$$T=\frac{1}{n_2}\sum_{i=n_1+1}^{n-n_3}x_i+Cs\frac{n_3-n_1}{n_2} \quad . \tag{11-9}$$

（2）汉培尔（Hampel）三参数估计

在尤伯估计中，两端数值的影响被限制为某一常数。因而是有界的，但不是零。还有一些M估计使得极端数值的影响降为零，这类估计的ψ函数是零尾（即重降）函数。

汉培尔的三参数估计使中部数值的影响与数值大小成正比，向外一部分数值的影响限制为某一常数，而更远处极端数值的影响逐渐降为零。其ψ函数为

$$\psi(z)=\begin{cases} z & |z|<a\ , \\ a\times \mathrm{sign}(z) & a\leqslant |z|<b\ , \\ \dfrac{C-|z|}{C-b}a\times \mathrm{sign}(z) & b\leqslant |z|<c\ , \\ 0 & |z|\leqslant c\ 。 \end{cases} \tag{11-10}$$

汉培尔三参数估计的计算用到（11-5）式的一阶台劳展开式，第$m+1$次迭代的位置估计值为：

$$T^{(m+1)}=T^{(m)}-\sum_{i=1}^{n}\psi(z_i^{(m)})\Big/\sum_{i=1}^{n}\frac{\partial}{\partial T}\psi(z_i^{(m)})\ , \tag{11-11}$$

其中：

$$z_i^{(m)}=(x_i-T^{(m)})/s^{(m)}\ ,$$

$$\sum_{i=1}^{n}\varphi(z_i)=\sum^{n_1}z_i+\sum^{n_2}a\times\mathrm{sign}(z_i)+\sum^{n_3}\frac{C-|z_i|}{C-b}a\times\mathrm{sign}(z_i)\ ,$$

$$\sum_{i=1}^{n}\frac{\partial}{\partial T}\psi(z_i)=n_1+\frac{a}{b-C}n_3\ 。$$

（11-11）式中，n_1,n_2,n_3分别代表符合$|z|<a$，$a\leqslant|z|<b$，$b\leqslant|z|<C$三段的样品数，而$\sum^{n_1}$，$\sum^{n_2}$，$\sum^{n_3}$分别表示在这三部分内求和。

（3）正弦估计

正弦估计又称为安宙斯（Andrews）估计，它是由加拿大数学家Andrews提出的。其ψ函数的中部是一个正弦函数，而两端为零：

$$\psi(z)=\begin{cases} \sin(z/C) & |z|<C\pi\ , \\ 0 & |z|\geqslant C\pi\ 。 \end{cases} \tag{11-13}$$

正弦估计近似于汉培尔三参数估计，但具有简单得多的数学表达式，因而在计算时较为方便。其缺点是即使在中心部分影响曲线也只是近似地与z值大小成正比，而不是严格成正比。

调节常数C的大小确定了估计的稳健性与效率之间的平衡。当C趋近无穷大时，估计值趋近于最小二乘估计。Andrews(1974) 推荐取$C=1.5$—2。当$C=1.339$时，正弦估计在正态分布情况下的渐近效率为95%。

正弦估计的计算与汉培尔三参数的计算类似，也要用到一阶台劳展开式。当ψ函数由（11-13）式定义时，（11-5）式的一阶台劳展开式可近似表达为：

$$\mathrm{tg}\left(\frac{T^{(m+1)}-T^{(m)}}{Cs}\right)=\sum^{n_2}\sin\left(\frac{x_i-T^{(m)}}{Cs}\right)\Big/\sum^{n_2}\cos\left(\frac{x_i-T^{(m)}}{Cs}\right)\ 。$$

第$m+1$次迭代的估计值为：

$$T^{(m+1)}=T^{(m)}+Cs\operatorname{tg}^{-1}\Big(\sum^{n_2}\sin\big(\frac{x_i-T^{(m)}}{Cs}\big)\Big/\sum^{n_2}\cos\big(\frac{x_i-T^{(m)}}{Cs}\big)\Big)\ , \qquad (11\text{-}14)$$

式中的s可由中位绝对中位数差来估计。

二、位置的L估计

L估计是用顺序统计量的线性组合来进行估计的一类估计方法。如果将某一总体的一组观测值按从小到大的次序排列起来，即$x_1\leqslant x_2\leqslant\cdots\leqslant x_n$，则它们就是一组顺序统计量的现实值。而L估计则是它们的线性组合：

$$T=\sum_{i=1}^{n}a_ix_i \qquad (11\text{-}15)$$

其中 $\Sigma a_i=1$ 。

显然，位置的L估计也是数据的加权平均值。不同的L估计给不同的顺序统计量以不同的系数。

1．α切尾均值

α切尾均值（$0\leqslant\alpha<0.5$）是顺序统计量的这样一种线性组合

$$T=\frac{1}{(1-2\alpha)n}\sum_{i=1}^{n}a_ix_i, \qquad (11\text{-}16)$$

其中

$$a_i=\begin{cases}1 & [\alpha n]+1<i<n-[\alpha n]\ ,\\ p & i=[\alpha n]+1\text{或}i=n-[\alpha n]\ ,\\ 0 & i<[\alpha n]+1\text{或}i>n-[\alpha n]\ ,\end{cases}$$

式中$[\alpha n]$表αn的整数部分，$p=1+[\alpha n]-\alpha n$。

α切尾均值实际上是将数据的最大和最小的$[\alpha n]$个数去掉之后剩下部分的加权平均值。当$[\alpha n]\neq\alpha n$，即αn不为整数时，剩余部分的最小和最大的各一个数的权为p，其余中间的数的权为1；当αn为整数时，剩余的所有数的权均为1。

由定义（11-16）式可知，当$\alpha=0$时，T就是样本的算术平均值，当α趋近于0.5时，T是中位数。

2．α文瑟化均值

对一组数据进行文瑟化就是按下式用x_i^*代替x_i：

$$x_i^*=\begin{cases}x_i & A\leqslant x_i\leqslant B\ ,\\ A & x_i<A\ ,\\ B & x_i>B\ 。\end{cases} \qquad (11\text{-}17)$$

这实际上是将数据中的两端数值用某个常数来代替。

一般α文瑟化均值（$0\leqslant\alpha<0.5$）是对数据中最小的$[\alpha n]$个数用$x_{[\alpha n]+1}$来代替，最大的$[\alpha n]$个数用$x_{n-[\alpha n]}$来代替之后的算术平均值。

3．跳跃估计

记样本的第一个及第四个4分位数（即概率为25%和75%处的x值）为p_{25}及p_{75}，按

下式定义两点h_1及h_2：

$$h_1 = p_{25} - C(p_{75} - p_{25}) ,$$
$$h_2 = p_{75} + C(P_{75} - P_{25}) 。 \tag{11-18}$$

其中C是调节常数，通常取1.0—2.0之间的值。位置的单重跳跃估计是先将小于h_1及大于h_2的数据“跳跃过去”，再对剩余部分求算术平均值或某种稳健的位置估计值。多重跳跃法则对前一次的剩余部分重新求h_1和h_2，然后再进行跳跃，直到没有数据可被跳跃为止。

4．百分位数的线性组合

百分位数的线性组合是将样本数据顺序排列后，按概率的百分位数值组合而得位置的估计。记p_k为样本的第k个百分位数（概率为k处的观测值），被采用过的估计公式有

$$T = p_{50} \quad （即中位数），$$
$$T = (p_{25} + p_{75})/2 ,$$
$$T = (p_{15} + p_{84})/2 , \tag{11-19}$$
$$T = (p_{16} + p_{50} + p_{84})/3 ,$$
$$T = 0.3p_{33.3} + 0.4p_{50} + 0.3p_{66.7} ;$$

这些估计计算简便，但数据的利用率低，一般可作粗略的初始估计。

三、位置的R估计

位置的R估计即秩估计，是由双样本秩检验导出的一系列位置估计方法，用于对称或近似对称分布的情形。

设$\mathbf{x} = (x_1, x_2, \cdots, x_n)$是从某一对称分布的总体中抽出的样本，$T$是总体的位置估计值，则可由下式组成第二样本$\mathbf{y} = (y_1, y_2, \cdots, y_n)$

$$y_i = 2T - x_i \quad (i = 1, 2, \cdots, n) 。 \tag{11-20}$$

因为$y_i - T = -(x_i - T)$，可见$\mathbf{y}$是$\mathbf{x}$的以T为中心的镜象。如果T是位置的真值，则$\mathbf{x}$和$\mathbf{y}$所代表的分布应完全重合。

将两样本混合为一容量为$2n$的样本，记$x_1, x_2, \cdots, x_n$在混合样本中的序数为$R_1, R_2, \cdots, R_n$，定义秩统计量$S_{n,n}$为：

$$S_{n,n} = \sum_{i=1}^{n} \alpha(R_i) , \tag{11-21}$$

其中$d(R_i)$是R_i的函数，被称为“得分”。$d(R_i)$函数形式的选择应使得当总体分布为对称分布且T等于位置的真值时，$S_{n,n} = 0$。例如，函数

$$d(R_i) = \frac{R_i}{2n+1} - \frac{1}{2} \tag{11-22}$$

就满足这一要求。

可以想象，如果T显著偏离真值，必有较多的$\mathbf{x}$样品集中于混合样本的某一侧，即$R_1, R_2 \cdots, R_n$中的大部分将大于n或小于n，致使$S_{n,n}$不等于0。为了求得位置的R估计，可以调节T的值，直至$S_{n,n} = 0$或接近于0为止。

四、尺度的稳健估计

单纯的尺度估计问题在现实工作中是很少的。尺度常常是在位置估计中所需要的一个

参数。很多尺度的稳健估计方法与位置的稳健估计方法类似，只是在使用时要更加注意尾部信息的处理问题。由于尺度的估计在很大程度上取决于尾部而不是中部，抛弃过多的尾部信息有可能带来较大的估计偏差。

1．尺度的M估计

与位置的M估计定义相类似，尺度S的M估计为下列函数的解：

$$\Sigma\chi(r_i/S)=0\ , \tag{11-23}$$

其中$r_i=x_i-T$，而χ是一个适当选择的偶函数。它在正半轴上非降，且有一次变号，即$\chi(0)<0$，而$\chi(\pm\infty)>0$。

对于给定n和样本值$x_1,x_2,\cdots,x_n$，恒有$S\to\infty$时，$\Sigma\chi(r_i/S)\to\Sigma\chi(0)<0$及当$S\to 0$时，$\Sigma\chi(r_i/S)\to\Sigma\chi(\pm\infty)>0$，故方程（11-23）式一定有解。

尺度的M估计与位置的M估计往往是可以对应起来的，如果取χ函数满足下式：

$$\chi(z_i)=z_i\psi(z_i)-1 \qquad (i=1,2,\cdots,n) \tag{11-24}$$

则由（11-23）式定义的尺度估计与由（11-5）式定义的位置估计相对应。它们均是方程

$$\Sigma\rho(z_i)=\min$$

的解，即：

$$\psi(z_i)=-S\frac{\partial}{\partial T}\rho(z_i)\ ,$$

$$\chi(z_i)=-S\frac{\partial}{\partial S}\rho(z_i)\ 。$$

这种关系常用来同时进行位置和尺度的估计。

(1) 尤伯估计

在提出位置估计（11-8）式的同时，尤伯（1964）提出了相应的尺度估计，其χ函数为：

$$\chi(z)=\psi^2(z)-\beta\ ,$$

即

$$\chi(z)=\begin{cases}z^2-\beta & |z|<C\ ,\\ C^2-\beta & |z|\geqslant C\ ,\end{cases} \tag{11-25}$$

其中：

$$\beta=\int_{-\infty}^{+\infty}\psi^2(t)\psi(t)dt\ 。$$

而$\psi(t)$为标准正态分布的密度函数

$$\psi(t)=\frac{1}{\sqrt{2\pi}}\exp(-t^2/2)\ ,$$

β的取值使标准正态分布情况下尤伯尺度估计的值为1。

位置和尺度的尤伯估计可同时求解。此时每次迭代中依次使用以下公式：

$$\overline{x'}=\frac{1}{n_2}\sum_{i=n_1+1}^{n-n_3}x_i\ , \tag{11-26}$$

$$S^2=\sum_{i=n_1+1}^{n-n_3}(x_i-\overline{x'})^2\Big/\left[(n-1)\beta-\left(n_1+n_2+\frac{(n_3-n_1)^2}{n_2}\right)C^2\right]\ , \tag{11-27}$$

$$T = \overline{x'} + Cs(n_3 - n_1)/n_2 \ , \tag{11-28}$$

其中：

$$\beta = 2\left(\frac{C^2}{2} - \frac{C}{\sqrt{2\pi}}e^{-C^2/2} - \frac{(C^2-1)}{\sqrt{\pi}}\int_0^C e^{-t^2}dt\right) \ , \tag{11-29}$$

其余符号同前。

(2) 中位绝对中位差及中位数法

中位绝对中位差即相对于中位数的偏离绝对值的中位数：

$$MAD = \text{med}\{|x_i - \text{med}\{x_i\}|\}/0.6745 \ 。 \tag{11-30}$$

其尺度的M估计为：

$$\chi(z) = \text{sign}(|z| - 1) \ 。 \tag{11-31}$$

通常将中位数称为中位绝对中位差相对应的位置估计。实际上，中位数的定义函数为

$$\psi(z) = \text{sign}(z) \ 。$$

由于中位数及中位绝对中位差的计算简便，且都有高崩溃点 (1/2)，因而常被用作迭代计算的初始值。

2. 尺度的L估计

许多对位置的L估计方法可参照使用，形成对尺度的L估计，如切尾均方差、文瑟化均方差、尺度的跳跃估计等。在求这些尺度估计时都要用到位置估计。

以样本百分数为基础的尺度估计公式如下：

$$\begin{aligned} S &= (p_{75} - p_{25})/2 \ , \\ S &= (p_{84} - p_{16})/2 \ , \\ S &= (p_{84} - p_{16})/4 + (p_{95} - p_5)/6.6 \ 。 \end{aligned} \tag{11-32}$$

尺度的L估计存在一个标准化的问题，即使该估计在标准正态分布的情况下的数值为1。(11-32)式中的第一式需除以常数0.6745，α切尾均方差的标准化需乘以系数$[1-2\alpha-2\xi\cdot\psi(\xi)]^{-\frac{1}{2}}$，其中$\xi$为标准正态分布在$(1-\alpha)$处的分位数，而$\psi(\xi)$为$\xi$处的标准正态分布频率。

第四节　方差的稳健估计

一、概述

在单变量统计中，离群值的处理常常可以利用对每一变量的一元方法来检测。对多变量数据来说观测被认为是离群点的，常常是指当每一个变量的值被认为是与其它变量有关时，有些不能够保持那些在大部分观测中显然存在的变量间关系的模式。由于传统方法的性能严重受到异常值的影响，所以几乎不受异常数据影响的稳健方法就提供了一种具有吸引力的补救办法。Huber(1972)和Hampel(1973)的综述论文总结了早年一元稳健分析的重要方法和成果，后Huber(1977)和Hampel(1977)又发表了更新的成果。Hogg (1977) 的引导论文给出了基本的想法。

二、多元位置和尺度的稳健估计方法

在多元M估计中马氏距离起了一个重要的作用。第m次观测相对于观测均值的马氏平方距离定义为：

$$d_m^2 = (\mathbf{x}_m - \bar{\mathbf{x}})' \mathbf{V}^{-1} (\mathbf{x}_m - \bar{\mathbf{x}}) \ , \tag{11-33}$$

其中$\mathbf{V}$表示样品的方差矩阵。

M估计可以被认为是对传统估计算法的一种简单修正，它对被认为来自于数据主体的观测赋于全权，而对来自干扰分布的尾端的观测则减少其权重或影响。实际上，对具有过分大马氏距离的观测的权重进行了减少。

这里用来定义均值和方差稳健估计的方程如下：

$$\bar{\mathbf{x}} = \sum_{m=1}^{n} w_m \mathbf{x}_m \Big/ \sum_{m=1}^{n} w_m \ , \tag{11-34}$$

$$\mathbf{V} = \sum_{m=1}^{n} w_m^2 (\mathbf{x}_m - \bar{\mathbf{x}})(\mathbf{x}_m - \bar{\mathbf{x}})' \Big/ \Big(\sum_{m=1}^{n} w_m^2 - 1 \Big) \ . \tag{11-35}$$

其中：

$$w_m = w(d_m) = \omega(d_m)/d_m$$

$$d_m = \{(\mathbf{x}_m - \bar{\mathbf{x}})' \mathbf{V}^{-1} (\mathbf{x}_m - \bar{\mathbf{x}})\}^{\frac{1}{2}} \ ,$$

通过迭代法可求解$\bar{\mathbf{x}}$和$\mathbf{V}$。

这里，ω是一个有界的影响函数，在d_m的值域上与合理的数据对应是线性的，而在这个值域外有界。Hampel (1973) 提出的函数为

$$\omega(d) = \begin{cases} d & d \leqslant d_0 \ , \\ d_0 \exp\{-\frac{1}{2}(d-d_0)^2/b_2^2\} . & d > d_0 \ , \end{cases} \tag{11-36}$$

其中：

$d_0 = \sqrt{v} + b_1/\sqrt{2}$ ，

b_1是导于某个标准正态分布的百分点，

b_2控制了影响函数下降为零的速度。

从广泛的实践经验来看，考虑到（11-36）式的三种情况，它们是：

① $b_1 = \infty$所有观测具有单位权重，同常规计算结果相同。

② $b_1 = 2$，$b_2 = \infty$Huber(1964)提出的非降形式。

③ $b_1 = 2$，$b_2 = 1.25$Hampel(1973) 提出的加速下降形式。

三、稳健统计计算例子

这是一组假想的数据，共有13个样品，每个样品有3个变量，结果见表（11-1）。其中，变量x_1的数据近似正态分布，x_2是偏正态分布的数据，x_3为严重偏离正态分布。用该组数据进行了常规统计方法和稳健统计方法的计算，得出它们各自的均值、标准离差和相关系数，结果列于表（11-2）和（11-3）。

表11-1

样 号	x_1	x_2	x_3	样 号	x_1	x_2	x_3
1	9.00	90.00	7.00	8	35.00	65.00	36.00
2	11.00	91.00	8.00	9	38.00	65.00	37.00
3	17.00	82.00	15.00	10	45.00	41.00	43.00
4	24.00	89.00	22.00	11	51.00	39.00	51.00
5	28.00	76.00	26.00	12	50.00	110.00	500.00
6	30.00	72.00	31.00	13	52.00	300.00	1010.00
7	31.00	71.00	36.00				

表11-2

变 量	均 值			标 准 差		
	常规方法	稳健方法	改 变 率	常规方法	稳健方法	改 变 率
x_1	32.38	29.00	11.6%	14.65	13.25	10.5%
x_2	91.61	71.00	29.0%	65.61	17.97	265.1%
x_3	140.15	28.36	394.0%	292.35	14.21	1957.3%

表11-3

相关系数	常 规 方 法			稳 健 方 法		
	x_1	x_2	x_3	x_1	x_2	x_3
x_1	1.0000	0.2547	0.5724	1.0000	−0.9408	0.9885
x_2	0.2547	1.0000	0.9161	−0.9408	1.0000	−0.9182
x_3	0.5724	0.9161	1.0000	0.9885	−0.9182	1.0000

从表11-2可见，变量x_1的原始数据近似正态分布，用常规方法和稳健方法计算得到的均值与标准离差数比较接近；变量x_2为中等偏离正态分布，两种不同方法计算得的均值和标准差差别较大，均值和标准差的改变率分别为29%和265.1%；变量x_3是严重偏离正态分布的数据，两种统计方法计算的结果相差甚大，均值的改变率为394.0%，标准差的改变率则更大，为1957%。

从原始数据表11-1中可看出，变量x_1的数值是基本上按照从小到大的顺序排列的；变量x_2的数值除第12、13两个离群样品外，是按照从大到小下降排列的；变量x_3的排列与变量x_1相似，而与x_2的顺序相反。显而易见，x_1与x_2应该是负相关；x_1与x_3应该是正相关。由于离群样品的干扰，常规统计方法所得的结果得不到这种规律，$r_{1,2}$(常)$=0.2547$，$r_{2,3}$(常)$=0.9161$，$r_{1,3}$(常)$=0.5724$。稳健统计方法的结果正确地揭示了原始数据之间的相关关系：$r_{1,2}$(稳)$=-0.9408$，$r_{2,3}$(稳)$=-0.9182$，$r_{1,3}$(稳)$=0.9885$。因此，在实际工作中，如果不分析数据的分布情况，盲目地套用常规的方法进行各种统计计算，有时会导致错误的结论。而稳健统计方法有可能排除离群样品的干扰，保证样本特征参数及变量之间相关关系的正确结果。

附录 I　线性代数初步知识

矩阵的运算

一、有关矩阵的概念与定义

数域F上的$m\times n$个数a_{ij}（$i=1,\ 2,\ \cdots,\ m;\ j=1,\ 2,\ \cdots,\ n$）按确定的位置排成的矩形阵列（或者说表），称为$m\times n$矩阵。记作

$$\mathbf{A}=\begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{pmatrix},$$

其中横的一排叫行，竖的一排叫做列。第i行第j列相交处的数a_{ij}（$i=1,\ 2,\ \cdots,\ m;\ j=1,\ 2,\ \cdots,\ n$）称为矩阵$\mathbf{A}$的元（或叫元素）。矩阵$\mathbf{A}$简记为$(a_{ij})$或$(a_{ij})_{m\times n}$。

$n\times n$矩阵也称为n阶方阵，$a_{11},\ a_{22},\ \cdots,\ a_{nn}$（即下标$i=j$的元$a_{ii}$）称为矩阵$\mathbf{A}$的主对角线的元。

二、矩阵的相等、加、减、数乘、乘法与转置

1．矩阵$\mathbf{A}_{s\times n}$与矩阵$\mathbf{B}_{r\times m}$的相等定义，如果

① $s=r;\ n=m$，

② $a_{ij}=b_{ij}$（$i=1,\ 2,\ \cdots,\ s;\ j=1,\ 2,\ \cdots,\ n$），

则称 $\mathbf{A}=\mathbf{B}$。

也就是说相等矩阵必须具有相同行数与相同列数；而且，两矩阵相等，各对应位置的元素都分别相等。

2．矩阵的加法和减法

定义：

$$\begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \cdots & \cdots & \cdots & \cdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{pmatrix} \pm \begin{pmatrix} b_{11} & b_{12} & \cdots & b_{1n} \\ b_{21} & b_{22} & \cdots & b_{2n} \\ \cdots & \cdots & \cdots & \cdots \\ b_{m1} & b_{m2} & \cdots & b_{mn} \end{pmatrix}$$

$$= \begin{pmatrix} a_{11}\pm b_{11} & a_{12}\pm b_{12} & \cdots & a_{1n}\pm b_{1n} \\ a_{21}\pm b_{21} & a_{22}\pm b_{22} & \cdots & a_{2n}\pm b_{2n} \\ \cdots\cdots & \cdots\cdots & \cdots\cdots & \cdots\cdots \\ a_{m1}\pm b_{m1} & a_{m2}\pm b_{m2} & \cdots & a_{mn}\pm b_{mn} \end{pmatrix}$$

两个同类型的矩阵相加减，就是把它们各对应位置的元素相加减。

矩阵的加减法满足如下性质：

① 交换律：$\mathbf{A}+\mathbf{B}=\mathbf{B}+\mathbf{A}$ 。

② 结合律：$\mathbf{A}+(\mathbf{B}+\mathbf{C})=(\mathbf{A}+\mathbf{B})+\mathbf{C}$ 。

3。数乘

一个数k乘上一个矩阵$\mathbf{A}$定义为

$$k\begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \cdots\cdots & & & \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{pmatrix}=\begin{pmatrix} ka_{11} & ka_{12} & \cdots & ka_{1n} \\ ka_{21} & ka_{22} & \cdots & ka_{2n} \\ \cdots\cdots & & & \\ ka_{m1} & ka_{m2} & \cdots & ka_{mn} \end{pmatrix},$$

也就是说数乘矩阵时，将乘到矩阵的每个元素上。

显然数乘矩阵有如下性质

① $k\mathbf{A}=\mathbf{A}k$ 。　　（交换律）

② $k(\mathbf{A}+\mathbf{B})=k\mathbf{A}+k\mathbf{B}$ 。　（分配律）

其中：k为任意数。

4。矩阵的乘法

若 $\mathbf{A}=(a_{ik})$为$m\times n$矩阵，

$\mathbf{B}=(b_{kj})$为$n\times s$矩阵，则，

$$\mathbf{A}_{m\times n}\mathbf{B}_{n\times s}=\mathbf{C}_{m\times s}\ ,$$

式中$C_{ij}=\sum_{k=1}^{n}a_{ik}b_{kj}$　　$(i=1,\ 2,\ \cdots m,\ j=1,\ 2,\ \cdots,\ s)$ 。

即

$$\begin{pmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & & \vdots \\ a_{m1} & \cdots & c_{mn} \end{pmatrix}\begin{pmatrix} b_{11} & \cdots & b_{1s} \\ \vdots & & \vdots \\ b_{n1} & \cdots & b_{ns} \end{pmatrix}=\begin{pmatrix} a_{11}b_{11}+\cdots+a_{1n}b_{n1} & \cdots & a_{n}b_{1s}+\cdots+a_{1n}b_{ns} \\ \vdots & & \vdots \\ a_{m1}b_{11}+\cdots+a_{mn}b_{n1} & \cdots & a_{m1}b_{1s}+\cdots+a_{mn}b_{ns} \end{pmatrix}。$$

我们称左边的矩阵$\mathbf{A}$为左矩阵，右边的矩阵$\mathbf{B}$为右矩阵，两个矩阵相乘，左矩阵的列数必须等于右矩阵的行数。乘积的元素C_{ij}，等于左矩阵的第i行与右矩阵的第j列相应元素相乘之后相加。

例。

$$\begin{pmatrix} 1 & 2 \\ 4 & 5 \\ 3 & 6 \end{pmatrix}\begin{pmatrix} 3 & 1 \\ 2 & 4 \end{pmatrix}=\begin{pmatrix} 1\times3+2\times4 & 1\times1+2\times4 \\ 4\times3+5\times2 & 4\times1+5\times4 \\ 3\times3+6\times2 & 3\times1+6\times4 \end{pmatrix}=\begin{pmatrix} 7 & 9 \\ 22 & 24 \\ 21 & 27 \end{pmatrix}。$$

矩阵的乘法有下列性质：

① 满足结合律　$(\mathbf{AB})\mathbf{C}=\mathbf{A}(\mathbf{BC})$ 。

② 满足分配律 $(\mathbf{A}+\mathbf{B})\mathbf{C}=\mathbf{AC}+\mathbf{BC}$，

$\mathbf{C}(\mathbf{A}+\mathbf{B})=\mathbf{CA}+\mathbf{CB}$。

③ 一般情况下不满足交换律 $\mathbf{AB}\neq\mathbf{BA}$

5. 矩阵的转置

把$m\times n$矩阵$\mathbf{A}=(a_{ij})$的列与行互换后所得到的$n\times m$矩阵称为$\mathbf{A}$的换置矩阵，记作$\mathbf{A}'$，即

$$\mathbf{A}'=\begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{pmatrix}'=\begin{pmatrix} a_{11} & a_{21} & \cdots & a_{m1} \\ a_{12} & a_{22} & \cdots & a_{m2} \\ \vdots & \vdots & & \vdots \\ a_{1n} & a_{2n} & \cdots & a_{mn} \end{pmatrix},$$

也就是$a'_{ij}=a_{ji}$。

转置矩阵满足如下性质：

① $(\mathbf{A}+\mathbf{B})'=\mathbf{A}'+\mathbf{B}'$。

② $(\mathbf{AB})'=\mathbf{B}'\mathbf{A}'$。

第②条称为反序定律。

三、几个特殊的矩阵

1. 零矩阵

元素a_{ij}的值全为0的矩阵称为零矩阵，记作

$$\mathbf{0}=\begin{pmatrix} 0 & 0 & \cdots & 0 \\ 0 & 0 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \cdots & 0 \end{pmatrix}。$$

零矩阵具有性质：

① $\mathbf{0}+\mathbf{A}=\mathbf{A}+\mathbf{0}=\mathbf{A}$。

② $\mathbf{0A}=\mathbf{A0}=\mathbf{0}$。

③ $\mathbf{A}+(-\mathbf{A})=\mathbf{0}$ 称$-\mathbf{A}$为$\mathbf{A}$的负矩阵。

2. 对角矩阵

主对角线以外的元素都是零（$d_{ij}=0$，$i\neq j$）的方阵称为对角阵，记作

$$\mathbf{D}=\begin{pmatrix} d_{11} & & & \mathbf{0} \\ & d_{22} & & \\ & & \ddots & \\ \mathbf{0} & & & d_{nn} \end{pmatrix}。$$

3. 单位矩阵

主对角线上的元素全为1，（$d_{ij}=1$，$i=j$），且主对角线以外的元素全为0，（$d_{ij}=0$，

当$j \neq i$）称为单位矩阵，记作

$$\mathbf{1}=\begin{pmatrix} 1 & & 0 \\ & 1 & \\ & \ddots & \\ 0 & & 1 \end{pmatrix}。$$

显然 $|\mathbf{B}=\mathbf{B}|=\mathbf{B}$ 。

4．对称矩阵

满足条件$a_{ij}=a_{ji}$ （$i, j=1, 2, \cdots, n$）的方阵$\mathbf{A}=(a_{ij})$称为对称矩阵。例如

$$\mathbf{A}=\begin{pmatrix} 5 & 1 & 3 \\ 1 & 6 & -2 \\ 3 & -2 & 3 \end{pmatrix}。$$

显然对称矩阵$\mathbf{A}$的转置矩阵$\mathbf{A}'$仍为它本身，即$\mathbf{A}'=\mathbf{A}$ 。

5．逆矩阵

对n阶方阵$\mathbf{A}$，若存在一个n阶方阵$\mathbf{B}$，使得$\mathbf{AB}=\mathbf{BA}=\mathbf{1}$，称$\mathbf{B}$为$\mathbf{A}$的逆矩阵，记作$\mathbf{A}^{-1}$，它满足$\mathbf{AA}^{-1}=\mathbf{A}^{-1}\mathbf{A}=\mathbf{1}$。

只要方阵$\mathbf{A}$为非奇异（$|\mathbf{A}| \neq 0$）的矩阵，$\mathbf{A}$的逆矩阵总是存在。

6．正交矩阵

若有n阶方阵$\mathbf{A}$，它的转置矩阵$\mathbf{A}'$与其逆阵$\mathbf{A}^{-1}$相等（$\mathbf{A}'=\mathbf{A}^{-1}$）时，称方阵$\mathbf{A}$为正交矩阵。例如。

$$\mathbf{A}=\begin{bmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{bmatrix}。$$

正交矩阵每列（或行）的平方和等于1，不同的两列（或行）的乘积为零。

7．矩阵相似

如果有一非奇异矩阵$\mathbf{X}$（即$|\mathbf{X}| \neq 0$）使得

$$\mathbf{B}=\mathbf{X}^{-1}\mathbf{AX}，$$

那末矩阵$\mathbf{A}$与矩阵$\mathbf{B}$相似，也称$\mathbf{A}$经相似变换为$\mathbf{B}$，记作$\mathbf{A}\sim\mathbf{B}$。这种变换过程称为矩阵的相似变换。若$\mathbf{X}$为正交矩阵，则称正交变换。

四、向量

当矩阵的列数为一时，称为列向量，记作

$$\mathbf{a}=\begin{pmatrix} a_1 \\ a_2 \\ \vdots \\ a_n \end{pmatrix}_{n\times 1}$$

n为向量a的维数。向量a的转置为$a' = (a_1 a_2 \cdots a_n)$。

在几何学上，一个n维向量可以在n维空间中表示为一条具有方向的线段，也可以用一个点来表示，这个点就是向量的端点。

向量的长度和单位化

一个向量的长度即为坐标原点到向量端点的空间距离，记作$|a|$（或$\| a \|$）。由三角中的勾股定理，我们容易得到

$$|a| = \sqrt{\sum_{i=1}^{n} a_i^2} = (\mathbf{a}'\mathbf{a})^{\frac{1}{2}} \quad 。$$

我们把一个n维向量中每一个元素a_i，（$i = 1, 2, \cdots, n$）分别除以其长度$|a|$，称为向量的单位化。

即

$$a_i^* = \frac{a_i}{|\mathbf{a}|} \quad 。$$

经过单位化的向量$\mathbf{a}^*$的各元素之和等于1，即

$$\sum_{i=1}^{n} a_i^* = 1 \quad 。$$

在n维空间中有两个向量$\mathbf{a}$和$\mathbf{b}$。

$$\mathbf{a} = \begin{pmatrix} a_1 \\ a_2 \\ \vdots \\ a_n \end{pmatrix}, \qquad \mathbf{b} = \begin{pmatrix} b_1 \\ b_2 \\ \vdots \\ b_n \end{pmatrix} 。$$

这两个向量相乘，可以看为

$$\mathbf{a}'\mathbf{b} = \mathbf{b}'\mathbf{a} = \sum_{i=1}^{n} a_i b_i = a_1 b_1 + a_2 b_2 + \cdots + a_n b_n ,$$

所乘的结果是一个数，称这个数为这两个向量的内积。

若两个单位化了的向量，其内积为零，称这两个向量相互正交（或独立）。

两个向量端点之距离定义为这两个向量之距离，记作d，则

$$d = |\mathbf{a} - \mathbf{b}| = [(\mathbf{a} - \mathbf{b})'(\mathbf{a} - \mathbf{b})]^{\frac{1}{2}} \quad 。$$

两个向量夹角θ的余弦定义为：

$$\cos\theta = \frac{\mathbf{a}'\mathbf{b}}{(\mathbf{a}'\mathbf{a})^{\frac{1}{2}} (\mathbf{b}'\mathbf{b})^{\frac{1}{2}}} \quad 。$$

五、线性方程组

线性方程组的一般形式为

$$\begin{cases} a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n = y_1 \\ a_{21}x_1 + a_{22}x_2 + \cdots + a_{2n}x_n = y_2 \\ \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \\ a_{n1}x_1 + a_{n2}x_2 + \cdots + a_{nn}x_n = y_n \end{cases} 。$$

令

$$A = \begin{pmatrix} a_{11} & a_{12}\cdots a_{1n} \\ a_{21} & a_{22}\cdots a_{2n} \\ \vdots & \vdots \quad \vdots \\ a_{n1} & a_{n2}\cdots a_{nn} \end{pmatrix}, \quad x = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix}, \quad y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix},$$

则有

$$\mathbf{Ax = y} 。$$

从上面的方程组我们可以作y_1，$y_2\cdots y_n$能够用x_1，$x_2\cdots x_n$线性表示，称为向量x线性变换为向量y。

1．线性方程组的解

有如下一个线性方程组

$$\begin{cases} 12x_1 - 3x_2 + 2x_3 = 12 \\ -18x_1 + 3x_2 - x_3 = -15 \\ x_1 + x_2 + x_3 = 6 \end{cases},$$

由消元法，我们可以求得解向量x，步骤如下

由以上消去法可得解向量

$$x_1 = 1,\ x_2 = 2,\ x_3 = 3 。$$

令

$$A = \begin{pmatrix} 12 & -3 & 2 \\ -18 & 3 & -1 \\ 1 & 1 & 1 \end{pmatrix}, \quad x = \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix}, \quad b = \begin{pmatrix} 12 \\ -15 \\ 6 \end{pmatrix},$$

则

$$\mathbf{Ax = b} 。$$

从计算过程得，当系数矩阵A变换为单位矩阵I时，右端项就是方程组的解向量，即

$$\mathbf{A^{-1}Ax = A^{-1}b} 。$$

步骤方程	计算说明	系数			右端项	检查Σ
		a_1	a_2	a_3		
①		12	-3	2	12	23
②		-18	3	-1	-15	-31
③		1	1	1	6	9
④	①÷12	1	-0.25	0.1667	1	1.9167
⑤	②-④×(-18)	0	-1.5	2	3	3.5
⑥	③-④×(1)	0	1.25	0.8333	5	7.0833

续表

步骤方程	计　算　说　明	系　数			右端项	检查Σ
		a_1	a_2	a_3		
⑦	④-⑧×(-0.25)	1	0	-0.1667	0.5	1.3333
⑧	⑤+(-1.5)	0	1	-1.3333	-2	-2.3333
⑨	⑥-⑧×1.25	0	0	2.5	7.5	10
⑩	⑦-⑫×(-0.167)	1	0	0	1	2
⑪	⑧-⑫×(-1.33)	0	1	0	2	3
⑫	⑨÷2.5	0	0	1	3	4

$$\mathbf{x}=\mathbf{A}^{-1}\mathbf{b}\ .$$

2．矩阵求逆

从上面可知：当系数矩阵A经过一线性变换$\mathbf{A}^{-1}$即成为单位矩阵，反过来对一个与A同阶的单位矩阵也经过这样一个线性变换，就可以得到其逆矩阵。因此可以同样用消去法来求A的逆矩阵。

例：

有

$$\mathbf{A}=\begin{pmatrix}-1 & 8 & -2\\ -6 & 49 & -10\\ -4 & 34 & -5\end{pmatrix},$$

		A			I			检查
		a_1	a_2	a_3				Σ
①		-1	8	-2	1	0	0	6
②		-6	49	-10	0	1	0	34
③		-4	34	-5	0	0	1	28
④	①+(-1)	1	-8	2	-1	0	0	-6
⑤	④×(+6)+②	0	1	2	-6	1	0	-2
⑥	④×(+4)+③	0	2	3	-4	0	1	2
⑦	⑧×(+8)+④	1	0	18	-49	8	0	-22
⑧	⑤+1	0	1	2	-6	1	0	-2
⑨	⑧×(-2)+⑥	0	0	-1	8	-2	1	6
⑩	⑫×(-18)+⑦	1	0	0	95	-28	18	86
⑪	⑫×(-2)+⑧	0	1	0	10	-3	2	10
⑫	⑨×(-1)	0	0	1	-8	2	-1	-6

解得

$$\mathbf{A}^{-1}=\begin{pmatrix}95 & -28 & 18\\ 10 & -3 & 2\\ -8 & 2 & -1\end{pmatrix}.$$

3．线性方程组求解求逆并行方案

例．有方程组

$$\begin{cases} 10x_1 + 7x_2 + 4x_3 = 4 \\ 7x_1 + 7x_2 + 3x_3 = 4 \\ 4x_1 + 3x_2 + 4x_3 = 3 \end{cases},$$

	方程	计算说明	A			右端项	I			检查
	①		10	7	4	4	1	0	0	26
	②		7	7	3	4	0	1	0	22
	③		4	3	4	3	0	0	1	15
第一次消去第一列	④	①+10	1	$\frac{7}{10}$	$\frac{4}{10}$	$\frac{4}{10}$	$\frac{1}{10}$	0	0	$\frac{26}{10}$
	⑤	②−④×7	0	$\frac{21}{10}$	$\frac{2}{10}$	$\frac{12}{10}$	$-\frac{7}{10}$	1	0	$\frac{38}{10}$
	⑥	③−①×4	0	$\frac{2}{10}$	$\frac{24}{10}$	$\frac{14}{10}$	$-\frac{4}{10}$	0	1	$\frac{46}{10}$
第二次消去第二列	⑦	④−⑧×$\frac{7}{10}$	1	0	$\frac{1}{3}$	0	$\frac{1}{3}$	$-\frac{1}{3}$	0	$\frac{4}{3}$
	⑧	⑤÷$\frac{21}{10}$	0	1	$\frac{2}{21}$	$\frac{12}{21}$	$-\frac{7}{21}$	$\frac{10}{21}$	0	$\frac{38}{21}$
	⑨	⑥−⑧×$\frac{2}{10}$	0	0	$\frac{50}{21}$	$\frac{27}{21}$	$-\frac{1}{3}$	$-\frac{2}{21}$	1	$\frac{89}{21}$
第三次消去第三列	⑩	⑦−⑫×$\frac{1}{3}$	1	0	0	$-\frac{9}{50}$	$\frac{19}{50}$	$-\frac{16}{50}$	$-\frac{7}{50}$	$\frac{37}{50}$
	⑪	⑧−⑫×$\frac{2}{21}$	0	1	0	$\frac{26}{50}$	$-\frac{16}{50}$	$\frac{24}{50}$	$-\frac{2}{50}$	$\frac{82}{50}$
	⑫	⑨÷$\frac{50}{21}$	0	0	1	$\frac{27}{50}$	$-\frac{7}{50}$	$-\frac{2}{50}$	$\frac{21}{50}$	$\frac{89}{50}$

得 $\mathbf{x} = \begin{pmatrix} -\frac{9}{50} \\ \frac{26}{50} \\ \frac{27}{50} \end{pmatrix}$，$\mathbf{A}^{-1} = \begin{pmatrix} \frac{19}{50} & -\frac{16}{50} & -\frac{7}{50} \\ -\frac{16}{50} & \frac{24}{50} & -\frac{2}{50} \\ -\frac{7}{50} & -\frac{2}{50} & \frac{21}{50} \end{pmatrix}$。

由以上我们可以总结出第s次消去第k列的变换公式：

$$a_{ij}^{(s)} = \begin{cases} a_{ki}^{(s-1)} / a_{kk}^{(s-1)} & i = k, \\ a_{ij}^{(s-1)} - a_{ik}^{(s-1)} \cdot a_{ki}^{(s-1)} / a_{kk}^{(s-1)} & i \neq k, \end{cases}$$

这种方法称为高斯—约当求解和消去法求逆并行方案。

4. 求解求逆紧凑方案

从消去法中可以看出，每一步矩阵里都有这样三个列向量：〔1, 0, 0〕′，〔0, 1, 0〕′和〔0, 0, 1〕′。当第一次消去第一列后，原系数矩阵中的第一列变为〔1, 0, 0〕′，在以后的变换中该列总是不变，而右边单位阵中只有第一列有变化。消去其它各列也有同样情况。因此，完全可以不把这个单位矩阵写出来。将右边逆矩阵存放的列调到左边消去后出现的单位矩阵的各列位置上去。

以前例为例，第一次消去第一列后所得的变换矩阵为：

$$\begin{pmatrix} \frac{1}{10} & \frac{7}{10} & \frac{4}{10} & \frac{4}{10} \\ -\frac{7}{10} & \frac{21}{10} & \frac{2}{10} & \frac{12}{10} \\ -\frac{4}{10} & \frac{2}{10} & \frac{24}{10} & \frac{14}{10} \end{pmatrix}。$$

第二次消去第二列后得变换矩阵为：

$$\begin{pmatrix} \frac{1}{3} & -\frac{1}{3} & \frac{1}{3} & 0 \\ -\frac{7}{21} & \frac{10}{21} & \frac{2}{21} & \frac{12}{21} \\ -\frac{1}{3} & -\frac{2}{21} & \frac{50}{21} & \frac{27}{21} \end{pmatrix}。$$

第三次消去第三列后得变换矩阵为：

$$\begin{pmatrix} \frac{19}{50} & -\frac{16}{50} & -\frac{7}{50} & -\frac{9}{50} \\ -\frac{16}{50} & \frac{24}{50} & -\frac{2}{50} & \frac{26}{50} \\ -\frac{7}{50} & -\frac{2}{50} & \frac{21}{50} & \frac{27}{50} \end{pmatrix}。$$

化为小数表示

$$\begin{pmatrix} 0.38 & -0.32 & -0.14 & -0.18 \\ -0.32 & 0.48 & -0.04 & 0.52 \\ -0.14 & -0.04 & 0.42 & 0.54 \end{pmatrix}。$$

矩阵中前三列为逆矩阵，后一列为解向量。

紧凑方案中，第s次消去第k列的变换公式为：

$$a_{ij}^{(s)}=\begin{cases} 1/a_{kk}^{(s-1)} & i=k,\ j=k \\ -a_{ik}^{(s-1)}/a_{kk}^{(s-1)} & i\neq k,\ j=k \\ a_{kk}^{(s-1)}/a_{kk}^{(s-1)} & i=k,\ j\neq k \\ a_{ij}^{(s-1)}-a_{ik}^{(s-1)}\cdot a_{kj}^{(s-1)}/a_{kk}^{(s-1)} & i\neq k,\ j\neq k \end{cases}。$$

同时可以证明，所得结果与消去列的顺序无关。

六、矩阵的特征值与特征向量

对于一个线性变换$\mathbf{Ax}=\mathbf{b}$，及其相应的矩阵$\mathbf{A}$，若$\mathbf{x}$不是零向量，且

$$\mathbf{b}=\lambda\mathbf{x}\ ,$$

其中λ是一个数，也就是当

$$\mathbf{Ax}=\lambda\mathbf{x}$$

时，则称λ为矩阵A的“特征值”，称向量x是矩阵A的对应于λ的“特征向量”。

由上式可得

$$(\mathbf{A}-\lambda\mathbf{I})\mathbf{x}=0\ ,$$

这是一个齐次方程组，即

$$\begin{cases}(a_{11}-\lambda)x_1+a_{12}x_2+\cdots+a_{1n}x_n=0\\ a_{21}x_1+(a_{22}-\lambda)x_2+\cdots+a_{2n}x_n=0\\ \cdots\cdots\cdots\cdots\cdots\cdots\\ a_{n1}x_1+a_{n2}x_2+\cdots+(a_{nn}-\lambda)x_n=0\end{cases},$$

因x为非零向量，若要使x有非零解，则必须满足

$$|\mathbf{A}-\lambda\mathbf{I}|=\begin{vmatrix}a_{11}-\lambda & a_{12} & \cdots a_{1n}\\ a_{21} & a_{22}-\lambda & \cdots a_{2n}\\ \vdots & \vdots & \vdots\\ a_{n1} & a_{n2} & \cdots a_{nn}-\lambda\end{vmatrix}=0\ 。$$

将此行列式展开，其结果是λ的一个n次多项式，这个多项式称为矩阵A的特征方程，记作：$\varphi(\lambda)=0$，求此特征方程就得到矩阵A的n个特征根λ_i（$i=1,\ 2,\ \cdots,\ n$）；再把这n个特征根分别代进齐次方程组中，即可求得对应于每一个特征根λ_i的n个特征向量x_i（$i=1,\ 2,\ \cdots,\ n$）。

例。

$$\mathbf{A}=\begin{bmatrix}-3 & 5\\ 4 & -2\end{bmatrix}\ 。$$

写出特征方程式：

$$\begin{cases}(-3-\lambda_i)x_{1i}+5x_{2i}=0\\ 4x_{1i}+(-2-\lambda_i)x_{2i}=0\end{cases},$$

其特征多项式为

$$\begin{vmatrix}-3-\lambda_i & 5\\ 4 & -2-\lambda_i\end{vmatrix}=0\ 。$$

展开整理得 $\lambda_i^2+5\lambda_i-14=0$ 。

解之

$$\lambda_i=\frac{-b\pm\sqrt{b^2-4ac}}{2a}=\frac{-5\pm9}{2}\ ,$$

$$\begin{cases}\lambda_1=-7\ ,\\ \lambda_2=2\ 。\end{cases}$$

将$\lambda_1=-7$代进齐次方程组。

$$\begin{cases}(-3+7)x_{1i}+5x_{2i}=0\\4x_{1i}-(2-7)x_{2i}=0\end{cases}$$

$$\begin{cases}4x_{1i}+5x_{2i}=0\\4x_{1i}+5x_{2i}=0\end{cases},$$

令 $x_{11}=5$，

则 $x_{21}=-4$。

$$\therefore\quad X_1=\begin{pmatrix}5\\-4\end{pmatrix}。$$

x_1是矩阵A对应于$\lambda_1=-7$的特征向量。

同理将$\lambda_2=2$代入齐次方程组：

$$\begin{cases}-5x_{1i}+5x_{1i}=0\\4x_{1i}-(2+2)x_{2i}=0\end{cases},$$

令$x_{12}=1$，则$x_{22}=1$。

$$x_2=\begin{pmatrix}1\\1\end{pmatrix}$$是矩阵A对应于$\lambda_2=2$的特征向量。

特征方程$\varphi(\lambda)=0$的n个根λ_1，λ_2，…，λ_n称为矩阵A的谱，记作chA，求矩阵A的谱集合及其对应特征向量的过程称为矩阵A的谱分解。

特征值与特征向量有下列主要性质：

设λ_1，λ_2，…，λ_n为n阶方阵A的特征值，则

（1）A的逆矩阵A^{-1}的特征值为λ_1^{-1}，λ_2^{-1}，…，λ_n^{-1}。

（2）n阶方阵A的特征值之和等于A的迹，

即

$$\lambda_1+\lambda_2+\lambda_n\cdots=a_{11}+a_{22}+\cdots+a_{nn},$$

或写成

$$\sum_{i=1}^{n}\lambda_i=\sum_{h=1}^{n}a_{ii}=t_rA。$$

（3）矩阵A的n个特征值之积等于A的行列式，即

$$\lambda_1\cdot\lambda_2\cdots\lambda_n=|A|。$$

（4）实对称矩阵的特征都是实数，并且其对应的n个特征值向量互相正交（独立）。

（5）A矩阵与其转置A'有相同的特征值。更广泛一点说，矩阵的相似变换其特征值保持不变。

根据性质（5）可以设计多种求矩阵特征值的方法。如求实对称矩阵的雅可比法就是我们常用的一种方法。

七、求实对称矩阵特征值和特征向量的雅可比法

设n阶实对称矩阵$A=(a_{ij})_{n\times n}$的特征值是λ_1，λ_2，…，λ_n则必存在一正交矩阵Q，使

得：

$$\mathbf{Q}^{-1}\mathbf{A}\mathbf{Q}=\begin{pmatrix}\lambda_1 & & & 0\\ & \lambda_2 & & \\ & & \ddots & \\ 0 & & & \lambda_n\end{pmatrix}。$$

正交矩阵Q可用一系列旋转矩阵的积来逼近：

$$\mathbf{Q}=\mathbf{U}_1\mathbf{U}_2\cdots\mathbf{U}_m=\prod_{i=1}^{m}\mathbf{U}_i$$

式中

$$\mathbf{U}_i=\mathbf{U}_{pq}=\begin{array}{c}\begin{matrix}(p) & & (q)\end{matrix}\\ \begin{pmatrix}1 & & & & & 0\\ & \ddots & & & & \\ & & 1 & & & \\ & & & \cos\theta & \cdots & \sin\theta\\ & & & \vdots & \ddots & \vdots\\ & & & -\sin\theta & \cdots & \cos\theta\\ & & & & & & 1\\ & & & & & & & \ddots\\ 0 & & & & & & & & 1\end{pmatrix}\end{array}\begin{matrix}\\ \\ \\ (p)\\ \\ (q)\\ \\ \\ \end{matrix}$$

$(p<q,\ p=1,\ 2,\ \cdots,\ n-1;\ q=2,\ 3,\ \cdots,\ n)$。

取

$$\theta=\frac{1}{2}\operatorname{arcctg}\frac{a_{qq}-a_{pp}}{2a_{pq}}。$$

因为在这种旋转变换下，消去了矩阵中位于第p行第q列（$p\neq q$）交点上的元素，而矩阵所有元素的平方和保持不变，并且对角线上的元素的平方和增大，因而非对角线元素的平方和随之减小。因此，当旋转次数足够大时，可使非对角线元素的绝对值足够小，对于预先给定的精度$\varepsilon>0$，如果$|a_{ij}|<\varepsilon$（$i\neq j$）则可认为$a_{ij}\approx 0$。于是得到求矩阵A的特征值与特征向量的具体迭代方法。

1．按以下递推公式求特征值$\lambda_1,\ \lambda_2,\ \cdots,\ \lambda_n$：

$$\begin{cases}\zeta_k=\operatorname{ctg}2\theta=\dfrac{a_{qq}^{(k)}-a_{pp}^{(k)}}{2a_{pq}^{(k)}}\\ t_k=\operatorname{tg}\theta=\begin{cases}(\zeta_k+\sqrt{1-\zeta_k^2})^{-1} & (\zeta_k\geqslant 0)\\ -(|\zeta_k|+\sqrt{1-\zeta_k^2})^{-1} & (\zeta_k<0)\end{cases},\end{cases}$$

$$\begin{cases} v_k = \operatorname{tg}\dfrac{\theta}{2} = \begin{cases} \dfrac{\sqrt{1+t_k^2}-1}{t_k} & (t_k>0) \\ \dfrac{1-\sqrt{1+t_k^2}}{t_k} & (t_k<0) \end{cases} \\ s_k = \sin\theta = \dfrac{t_k}{\sqrt{1+t_k^2}} \end{cases}$$

$$\begin{cases} a_{pp}^{(k+1)} = a_{pp}^{(k)} - t_k a_{pq}^{(k)} \\ a_{qq}^{(k+1)} = a_{qq}^{(k)} + t_k a_{pq}^{(k)} \\ a_{pj}^{(k+1)} = a_{pj}^{(k)} - s_k(a_{qj}^{(k)} + v_k a_{pj}^{(k)}) & (j \neq p) \\ a_{qj}^{(k+1)} = a_{qj}^{(k)} + s_k(a_{pj}^{(k)} - v_k a_{qj}^{(k)}) & (j \neq q) \\ a_{ij}^{(k+1)} = a_{ij}^{(k)} & (i \neq p,\ q,\ j \neq p,\ q) \\ a_{ij}^{(1)} = a_{ij} & (i,\ j = 1,\ 2,\ \cdots)(k = 1,\ 2,\ \cdots) \end{cases}$$

假定当 $|a_{ij}^{(m)}| < \varepsilon$ $(i \neq j)$ 时，可以认为 $a_{ij}^{(m)} \approx 0$，则迭代到 $k = m-1$ 即可，而取 $a_{ij}^{(m)}$ 作为 λ_i 的近似值：

$$\lambda_i \approx a_{ii}^{(m)} \qquad (i = 1,\ 2,\ \cdots,\ n)\ 。$$

2．求特征向量

从前有

$$\mathbf{U}_m' \mathbf{U}_{m-1}' \cdots \mathbf{U}_1' \mathbf{A} \mathbf{U}_1 \cdots \mathbf{U}_{m-1} \mathbf{U}_m = \begin{pmatrix} \lambda_1 & & \mathbf{0} \\ & \lambda_2 & \\ \mathbf{0} & & \ddots \\ & & & \lambda_m \end{pmatrix}。$$

记

$$\mathbf{P}_m = \mathbf{U}_1 \cdots \mathbf{U}_{m-1} \mathbf{U}_m\ ,$$

则

$$\mathbf{A}\mathbf{P}_m = \mathbf{P}_m \begin{pmatrix} \lambda_1 & & \mathbf{0} \\ & \lambda_2 & \\ \mathbf{0} & & \ddots \\ & & & \lambda_m \end{pmatrix}。$$

根据特征向量的定义，$\mathbf{P}_m$ 为特征向量矩阵。

$\mathbf{P}_m$ 由下列递推公式算出

$$\begin{cases} u_{ip}^{(k+1)} = u_{ip}^{(k)} - s_k(u_{iq}^{(k)} + v_k u_{ip}^{(k)}) \\ u_{iq}^{(k+1)} = u_{iq}^{(k)} + s_k(u_{ip}^{(k)} - v_k u_{iq}^{(k)}) \\ u_{ij}^{(k+1)} = u_{ij}^{(k)} \qquad (j \neq p,\ q,\ i = 1,\ 2,\ \cdots,\ n) \\ u_{ij}^{(1)} = u_{ij} \qquad (i,\ j = 1,\ 2,\ \cdots,\ n)(k = 1,\ 2,\ \cdots,\ m-1) \end{cases}$$

最后得到

$$\rho_m = (u_{ij}^{(m)}) ,$$

即 $$u_j^{(m)} = [u_{1j}^{(m)},\ u_{2j}^{(m)},\ \cdots,\ u_{nj}^{(m)}]' 。$$

为对应于特征值λ_j的特征向量的近似值

八、用迭代法（幂法）求矩阵的特征值和特征向量

在多元分析经常遇到求实矩阵最大特征值的问题。这里我们简单介绍用迭代法来求绝对值最大的特征值及相应的特征向量。

假定一个n阶方阵$\mathbf{A}$的特征值和特征向量为λ_j和v_j（$j=1,\ 2,\ \cdots,\ n$）。不妨设想，特征向量v_j所对应的特征值λ_j是按照其绝对值的大小编号，即

$$|\lambda_1| > |\lambda_2| > |\lambda_3| > \cdots > |\lambda_n| 。$$

在n维空间中，对于任何一个向x都可以表示成向量v_i的线性组合，即

$$x = a_1 v_1 + a_2 v_2 + \cdots + a_n v_n 。$$

因此，我们任取n维向量x_0计算

$$\mathbf{x}_{k+1} = \mathbf{A}\mathbf{x}_k,\quad k = 0,\ 1,\ 2,\ \cdots 。$$

由于$\mathbf{x}_0$可表示成特征向量v_i的线性组合，即

$$\mathbf{x}_0 = a_1 v_1 + a_2 v_2 + \cdots + a_n v_n 。$$

根据特征值和特征向量的定义

$$\mathbf{A}\mathbf{v}_j = \lambda \mathbf{v}_j 。$$

所以有：

$$\mathbf{x}_1 = \mathbf{A}x_0 = a_1\lambda_1 v_1 + a_2\lambda_2 v_2 + \cdots + a_n\lambda_n v_n$$

$$= \lambda_1\left[a_1 v_1 + \left(\frac{\lambda_2}{\lambda_1}\right)a_2 v_2 + \cdots + \left(\frac{\lambda_n}{\lambda_1}\right)a_n v_n\right] 。$$

一般地

$$\mathbf{x}_k = \mathbf{A}\mathbf{x}_{k-1} = \mathbf{A}^k\mathbf{x}_0$$

$$= a_1\lambda_1^k v_1 + a_2\lambda_2^k v_2 + \cdots + a_n\lambda_n^k v_n$$

$$= \lambda_1^k\left[a_1 v_1 + \left(\frac{\lambda_2}{\lambda_1}\right)^k a_2 v_2 + \cdots + \left(\frac{\lambda_n}{\lambda_1}\right)^k a_n v\right] 。$$

当k足够大时，从

$$\left|\frac{\lambda_j}{\lambda_1}\right| < 1 \qquad (j = 2,\ 3,\ \cdots,\ n)$$

可知

$$\left(\frac{\lambda_j}{\lambda_1}\right)^k \approx 0 \qquad (j = 2,\ 3,\ \cdots,\ n) 。$$

从而（9-2）可为

$$\mathbf{x}_k \approx \lambda_1^k a_1 v_1,$$

$$\mathbf{x}_{k+1} \approx \lambda_1\lambda_1^k a_1 \mathbf{v}_1 \approx \lambda_1 \mathbf{x}_k 。$$

由此可见，当k足够大时（使$\frac{\lambda_j^k}{\lambda_1^k}\approx 0$时），$x_{k+1}$和$x_k$应当大体上只差一个常数倍，这个常数便是矩阵A的绝对值最大的特征值λ_1。这也可说成向量x_{k+1}和x_k当k足够大时大体上呈线性相关。

至于如何具体地确定λ_1，可以有几种处理方式。

1。取向量x_{k+1}的第j元与向量x_k中的第j元之比，即

$$\lambda_1=\frac{(x_{k+1})_j}{(x_k)_j}\qquad (j=1,\ 2,\ \cdots,\ n)\ 。$$

2。取向量(x_{k+1})与(x_k)各元素比的平均值，即

$$\lambda_1=\frac{1}{j}\sum_{j=1}^{n}\frac{(x_{k+1})_j}{(x_k)_j}\ 。$$

对于矩阵A相应λ_1的特征向量可以取向量x_{k+1}或v_1。因为由公式$x_k\approx\lambda_1^k\alpha_1 v_1$可知$v_1$与$x_k$之间只差$\lambda_1^k\alpha_1$倍，$x_k$与$x_{k+1}$之间只差$\lambda_1$倍，所以根据特征向量的性质$v_1$，$x_k$和$x_{k+1}$都是对应于$\lambda_1$的特征向量。通常我们可以将其单位化，即取

$$v_1=\frac{x_k}{||x_k||}\ 。$$

在求出了矩阵A的绝对值最大特征值及相应特征向量的基础上，可以将原矩阵A作适当的修改成一新矩阵A_1。再对新矩阵A_1用幂法求得其绝对值最大（相对A的次大）特征值和特征向量。倘若重复采用这一措施，在原则上可以将A的全部特征值和特征向量求出，不过由于前几个特征值都是近似的，加上计算中的舍入误差，后面计算出来的特征值是很不可靠的。

当A的最大特征值λ_1和相应的特征向量v_1求出后，如何修改原矩阵A呢？

1。当A的实对称矩阵时，将A修改成A_1的方法

我们可取

$$A_1=A-\lambda_1(v_1'v_1)^{-1}v_1v_1'\ 。$$

由于

$$A_1v_1=Av_1-\lambda_1(v_1'v_1)^{-1}v_1(v_1'v_1)=0\ ,$$

$$A_1v_j=Av_j-\lambda_1(v_1'v_1)^{-1}v_1v_1'v_j=\lambda_jv_j\qquad (j=2,\ 3,\ \cdots,\ n)\ 。$$

所以A_1的特征值为λ_2，λ_3，$\cdots$，λ_n和0，相应的特征向量为v_2，v_3，$\cdots$，v_n和v_1，也就是说矩阵A_1与A除第一特征值不同外其他的特征值和特征向量都相同，且A_1的最大特征值刚好为A的次大特征值。

对A_1应用幂法即可求得A的λ_2和v_2，重复上述，可求得前几个特征值和特征向量。

2。当A的非对称矩阵时

我们取

$$A_1=A-\lambda_1v_1u_1'\ ,$$

其中u_1为A'的特征值λ_1相应的特征向量。

附　录　Ⅱ

数理统计常用表

1. 在5%和1%信度下所需相关系数的最低值表

自由度 $n-2$	5%	1%	自由度 $n-2$	5%	1%	自由度 $n-2$	5%	1%
1	0.997	0.999	16	0.468	0.590	35	0.325	0.418
2	0.950	0.990	17	0.456	0.575	40	0.304	0.393
3	0.878	0.959	18	0.444	0.561	45	0.288	0.372
4	0.811	0.917	19	0.433	0.549	50	0.273	0.354
5	0.754	0.874	20	0.423	0.537	60	0.250	0.325
6	0.707	0.834	21	0.413	0.526	70	0.232	0.302
7	0.666	0.798	22	0.404	0.515	80	0.217	0.283
8	0.632	0.765	23	0.396	0.505	90	0.205	0.267
9	0.602	0.735	24	0.388	0.496	100	0.195	0.254
10	0.576	0.708	25	0.381	0.487	125	0.174	0.228
11	0.553	0.684	26	0.374	0.478	150	0.159	0.208
12	0.532	0.661	27	0.367	0.470	200	0.138	0.181
13	0.514	0.641	28	0.361	0.463	300	0.113	0.148
14	0.497	0.623	29	0.355	0.456	400	0.098	0.128
15	0.482	0.606	30	0.349	0.449	1000	0.062	0.081

2. 标准正态分布函数$F(u)$的数值表

$$F(u)=\int_{-\infty}^{u}\frac{1}{\sqrt{2\pi}}e^{-\frac{t^2}{2}}dt \quad (\text{当}u<0\text{时})$$

$F(u)$	0	1	2	3	4	5	6	7	8	9
-0.0	0.5000	0.4960	0.4920	0.4880	0.4840	0.4810	0.4761	0.4721	0.4681	0.4641
-0.1	0.4602	0.4562	0.4522	0.4438	0.4443	0.4404	0.4394	0.4325	0.4286	0.4247
-0.2	0.4207	0.4168	0.4129	0.4090	0.4052	0.4013	0.3974	0.3936	0.3897	0.3859
-0.3	0.3821	0.3783	0.3745	0.3707	0.3669	0.3632	0.3594	0.3557	0.3520	0.3483
-0.4	0.3446	0.3409	0.3372	0.3336	0.3300	0.3264	0.3228	0.3192	0.3156	0.3121
-0.5	0.3085	0.3050	0.3015	0.2981	0.2946	0.2912	0.2877	0.2843	0.2810	0.2776
-0.6	0.2743	0.2709	0.2676	0.2643	0.2611	0.2578	0.2546	0.2514	0.2483	0.2451
-0.7	0.2420	0.2389	0.2358	0.2327	0.2296	0.2266	0.2236	0.2206	0.2177	0.2148
-0.8	0.2119	0.2090	0.2061	0.2033	0.2005	0.1977	0.1949	0.1922	0.1894	0.1867
-0.9	0.1841	0.1814	0.1788	0.1762	0.1736	0.1711	0.1685	0.1660	0.1635	0.1611
-1.0	0.1587	0.1562	0.1539	0.1515	0.1492	0.1469	0.1446	0.1423	0.1401	0.1379
-1.1	0.1357	0.1335	0.1314	0.1292	0.1271	0.1251	0.1230	0.1210	0.1190	0.1170

续

$F(u)$	0	1	2	3	4	5	6	7	8	9
-1.2	0.1151	0.1131	0.1112	0.1093	0.1075	0.1056	0.1038	0.1020	0.1003	0.0985
-1.3	0.0968	0.0951	0.0934	0.0918	0.0901	0.0885	0.0869	0.0853	0.0838	0.0823
-1.4	0.0808	0.0793	0.0778	0.0764	0.0749	0.0735	0.0721	0.0708	0.0694	0.0681
-1.5	0.0668	0.0655	0.0643	0.0630	0.0618	0.0606	0.0594	0.0582	0.0571	0.0559
-1.6	0.0548	0.0537	0.0526	0.0516	0.0505	0.0495	0.0485	0.0475	0.0465	0.0455
-1.7	0.0446	0.0436	0.0427	0.0418	0.0409	0.0401	0.0392	0.0384	0.0375	0.0367
-1.8	0.0359	0.0351	0.0344	0.0336	0.0329	0.0322	0.0314	0.0307	0.0301	0.0294
-1.9	0.0287	0.0281	0.0274	0.0268	0.0262	0.0256	0.0250	0.0244	0.0239	0.0233
-2.0	0.0228	0.0222	0.0217	0.0212	0.0207	0.0202	0.0197	0.0192	0.0188	0.0183
-2.1	0.0179	0.0174	0.0170	0.0166	0.0162	0.0158	0.0154	0.0150	0.0146	0.0143
-2.2	0.0139	0.0136	0.0132	0.0129	0.0125	0.0122	0.0119	0.0116	0.0113	0.0110
-2.3	0.0107	0.0104	0.0102	0.0099	0.0096	0.0094	0.0091	0.0089	0.0087	0.0084
-2.4	0.0082	0.0080	0.0078	0.0075	0.0073	0.0071	0.0069	0.0068	0.0066	0.0064
-2.5	0.0062	0.0060	0.0059	0.0057	0.0055	0.0054	0.0052	0.0051	0.0049	0.0048
-2.6	0.0047	0.0045	0.0044	0.0043	0.0041	0.0040	0.0039	0.0038	0.0037	0.0036
-2.7	0.0035	0.0034	0.0033	0.0032	0.0031	0.0030	0.0029	0.0025	0.0027	0.0026
-2.8	0.0026	0.0025	0.0024	0.0023	0.0023	0.0022	0.0021	0.0021	0.0020	0.0019
-2.9	0.0019	0.0018	0.0018	0.0017	0.0016	0.0016	0.0015	0.0015	0.0014	0.0014
-3.0	0.0013	0.0013	0.0013	0.0012	0.0012	0.0011	0.0011	0.0011	0.0010	0.0010
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6843	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.6942	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9098	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767

续

$F(u)$	0	1	2	3	4	5	6	7	8	9
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9901	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986
3.0	0.9987	0.9987	0.9987	0.9988	0.9988	0.9989	0.9989	0.9989	0.9990	0.9990

3. 最大累积频率绝对差D_N的临界值$D_{N,\alpha}$表

$$P\{D_N > D_{N,\alpha}\} = \alpha$$

N \ α	0.10	0.05	0.01	N \ α	0.10	0.05	0.10
1	0.9500	0.9750	0.9950	31	0.2141	0.2379	0.2853
2	0.7764	0.8419	0.9293	32	0.2109	0.2342	0.2809
3	0.6360	0.7076	0.8290	33	0.2077	0.2308	0.2768
4	0.5652	0.6239	0.7342	34	0.2047	0.2274	0.2728
5	0.5095	0.5633	0.6685	35	0.2019	0.2243	0.2690
6	0.4680	0.5193	0.6166	36	0.1991	0.2212	0.2653
7	0.4361	0.4834	0.5758	37	0.1965	0.2188	0.2618
8	0.4096	0.4543	0.5418	38	0.1939	0.2154	0.2584
9	0.3875	0.4300	0.5133	39	0.1915	0.2127	0.2552
10	0.3867	0.4093	0.4889	40	0.1891	0.2101	0.2521
11	0.3524	0.3912	0.4677	41	0.1869	0.2076	0.2490
12	0.3382	0.3754	0.4491	42	0.1847	0.2052	0.2461
13	0.3255	0.3614	0.4325	43	0.1826	0.2028	0.2433
14	0.3142	0.3489	0.4176	44	0.1805	0.2006	0.2406
15	0.3040	0.3376	0.4042	45	0.1786	0.1984	0.2379
16	0.2947	0.3273	0.3920	46	0.1767	0.1963	0.2354
17	0.2863	0.3180	0.3809	47	0.1748	0.1942	0.2330
18	0.2785	0.3094	0.3706	48	0.1730	0.1922	0.2306
19	0.2714	0.3014	0.3612	49	0.1713	0.1903	0.2283
20	0.2647	0.2941	0.3524	50	0.1696	0.1884	0.2260
21	0.2586	0.2872	0.3443	55	0.1619	0.1798	0.2157
22	0.2528	0.2809	0.3367	60	0.1551	0.1723	0.2067
23	0.2475	0.2749	0.3295	65	0.1491	0.1657	0.1988
24	0.2424	0.2693	0.3229	70	0.1438	0.1598	0.1917
25	0.2377	0.2640	0.3166	75	0.1390	0.1544	0.1853
26	0.2332	0.2591	0.3106	80	0.1347	0.1496	0.1795
27	0.2260	0.2544	0.3050	85	0.1307	0.1452	0.1742
28	0.2250	0.2499	0.2997	90	0.1271	0.1412	0.1694
29	0.2212	0.2457	0.2947	95	0.1238	0.1375	0.1649
30	0.2176	0.2417	0.2899	100	0.1207	0.1340	0.1608

当$N>100$时，$D_{N,0.10}\approx\frac{1.23}{\sqrt{N}}$，$D_{N,0.05}\approx\frac{1.36}{\sqrt{N}}$，$D_{N,0.01}\approx\frac{1.63}{\sqrt{N}}$，按此近似公式算出$N$从120到300的几个临界值列于右表中，对右表中未列的120到300的其他N值，可查图。

N \ α	0.10	0.05	0.01
120	0.1123	0.1242	0.1489
140	0.1040	0.1151	0.1378
160	0.0972	0.1075	0.1289
180	0.0917	0.1013	0.1215
200	0.0870	0.0962	0.1153
220	0.0829	0.0917	0.1099
240	0.0794	0.0878	0.1052
260	0.0763	0.0844	0.1011
280	0.0735	0.0813	0.0974
300	0.0710	0.0785	0.0941

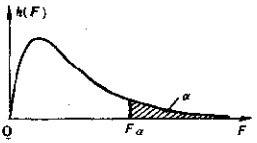
4. χ^2分布临界值χ_α^2表

$$P\{\chi^2\geqslant\chi_\alpha^2\}=\int_{\chi_\alpha^2}^{\infty}k(\chi^2)d\chi^2=\alpha$$

自由度ν	显著性水平α			自由度ν	显著性水平α		
	0.10	0.05	0.01		0.10	0.05	0.01
1	2.71	3.84	6.64	16	23.54	26.30	32.00
2	4.61	5.99	9.21	17	24.77	27.59	33.41
3	6.25	7.82	11.35	18	25.99	28.87	34.81
4	7.78	9.49	13.28	19	27.20	30.14	36.19
5	9.24	11.07	15.09	20	28.41	31.41	37.57
6	10.65	12.59	16.81	21	29.62	32.67	38.93
7	12.02	14.07	18.48	22	30.81	33.92	40.29
8	13.36	15.51	20.09	23	32.01	35.17	41.64
9	14.68	16.92	21.97	24	33.20	36.42	42.98
10	15.99	18.31	23.21	25	34.38	37.65	44.31
11	17.28	19.68	24.73	26	35.56	38.89	45.64
12	18.55	21.03	26.22	27	36.74	40.11	46.96
13	19.81	22.36	27.69	28	37.92	41.34	48.28
14	21.06	23.69	29.14	29	39.09	42.56	49.59
15	22.31	25.00	30.58	30	40.26	43.77	50.89

5. F分布临界值F_α表

$$P\{F \geqslant F_\alpha\} = \int_{F_\alpha} h(F)dF = \alpha$$



$\alpha = 0.10$

f_2 \ f_1	1	2	3	4	5	6	7	8	9	10	15	20	30	50	100	200	500	∞	f_1 / f_2
1	39.9	49.5	53.6	55.8	57.2	58.2	58.9	59.4	59.9	60.2	61.2	61.7	62.3	62.7	63.0	63.2	63.3	63.3	1
2	8.53	9.00	9.16	9.24	9.29	9.33	9.35	9.37	9.38	9.39	9.42	9.44	9.46	9.47	9.48	9.49	9.49	9.49	2
3	5.54	5.46	5.39	5.34	5.31	5.28	5.27	5.25	5.24	5.23	5.20	5.18	5.17	5.15	5.14	5.14	5.14	5.13	3
4	4.54	4.32	4.19	4.11	4.05	4.01	3.98	3.95	3.94	3.92	3.87	3.84	3.82	3.80	3.78	3.77	3.76	3.76	4
5	4.06	3.78	3.62	3.52	3.45	3.40	3.37	3.34	3.32	3.30	3.24	3.21	3.17	3.15	3.13	3.12	3.11	3.10	5
6	3.78	3.46	3.29	3.18	3.11	3.05	3.01	2.98	2.96	2.94	2.87	2.84	2.80	2.77	2.75	2.73	2.73	2.72	6
7	3.59	3.26	3.07	2.96	2.88	2.83	2.78	2.75	2.72	2.70	2.63	2.59	2.56	2.52	2.50	2.48	2.48	2.47	7
8	3.46	3.11	2.92	2.81	2.73	2.67	2.62	2.59	2.56	2.54	2.46	2.42	2.38	2.35	2.32	2.31	2.30	2.29	8
9	3.36	3.01	2.81	2.69	2.61	2.55	2.51	2.47	2.44	2.42	2.34	2.30	2.25	2.22	2.19	2.17	2.17	2.16	9
10	3.28	2.92	2.73	2.61	2.52	2.46	2.41	2.38	2.35	2.32	2.24	2.20	2.16	2.12	2.09	2.07	2.06	2.06	10
11	3.23	2.86	2.66	2.54	2.45	2.39	2.34	2.30	2.27	2.25	2.17	2.12	2.08	2.04	2.00	1.99	1.98	1.97	11
12	3.18	2.81	2.61	2.48	2.39	2.33	2.24	2.24	2.21	2.19	2.10	2.06	2.01	1.97	1.94	1.92	1.91	1.90	12
13	3.14	2.76	2.56	2.43	2.35	2.28	2.23	2.20	2.16	2.14	2.05	2.01	1.96	1.92	1.88	1.86	1.85	1.85	13
14	3.10	2.73	2.52	2.39	2.31	2.24	2.19	2.15	2.12	2.10	2.01	1.96	1.91	1.87	1.83	1.82	1.80	1.80	14
15	3.07	2.70	2.49	2.36	2.27	2.21	2.16	2.12	2.09	2.06	1.97	1.92	1.87	1.83	1.79	1.77	1.76	1.76	15
16	3.05	2.67	2.46	2.33	2.24	2.18	2.13	2.09	2.06	2.03	1.94	1.89	1.84	1.79	1.76	1.74	1.73	1.72	16
17	3.03	2.64	2.44	2.31	2.22	2.15	2.20	2.06	2.03	2.00	1.91	1.86	1.81	1.76	1.73	1.71	1.69	1.69	17
18	3.01	2.62	2.42	2.29	2.20	2.13	2.08	2.04	2.00	1.98	1.89	1.84	1.78	1.74	1.70	1.68	1.67	1.66	18
19	2.99	2.61	2.40	2.27	2.18	2.11	2.06	2.02	1.98	1.96	1.86	1.81	1.76	1.72	1.67	1.65	1.64	1.63	19
20	2.97	2.59	2.38	2.25	2.16	2.09	2.04	2.00	1.96	1.94	1.84	1.79	1.74	1.69	1.65	1.63	1.62	1.61	20
22	2.95	2.56	2.35	2.22	2.13	2.06	2.01	1.97	1.93	1.90	1.81	1.76	1.70	1.65	1.61	1.59	1.58	1.57	22
24	2.93	2.54	2.33	2.19	2.10	2.04	1.98	1.94	1.91	1.88	1.78	1.73	1.67	1.62	1.58	1.56	1.54	1.53	24
26	2.91	2.52	2.31	2.17	2.08	2.01	1.96	1.92	1.88	1.86	1.76	1.71	1.65	1.59	1.55	1.53	1.51	1.50	26
28	2.89	2.50	2.29	2.16	2.06	2.00	1.94	1.90	1.87	1.84	1.74	1.69	1.63	1.57	1.53	1.50	1.49	1.48	28
30	2.88	2.49	2.28	2.14	2.05	1.98	1.93	1.88	1.85	1.82	1.72	1.67	1.61	1.55	1.51	1.48	1.47	1.46	30
40	2.84	2.44	2.23	2.09	2.00	1.93	1.87	1.83	1.79	1.76	1.66	1.61	1.54	1.48	1.43	1.41	1.39	1.38	40
50	2.81	2.41	2.20	2.06	1.97	1.90	1.84	1.80	1.76	1.73	1.63	1.57	1.50	1.44	1.39	1.36	1.34	1.33	50
60	2.79	2.39	2.18	2.04	1.95	1.87	1.82	1.77	1.74	1.71	1.60	1.54	1.48	1.41	1.36	1.33	1.31	1.29	60
80	2.77	2.37	2.15	2.02	1.92	1.85	1.79	1.75	1.71	1.68	1.57	1.51	1.44	1.38	2.32	1.28	1.26	1.24	80
100	2.76	2.36	2.14	2.00	1.91	1.83	1.78	1.73	1.70	1.66	1.56	1.49	1.42	1.35	1.29	1.26	1.23	1.21	100
200	2.73	2.33	2.11	1.97	1.88	1.80	1.75	1.70	1.66	1.63	1.52	1.46	1.38	1.31	1.24	1.20	1.17	1.14	200
500	2.72	2.31	2.10	1.96	1.86	1.79	1.73	1.68	1.64	1.61	1.50	1.44	1.36	1.28	1.21	1.16	1.12	1.09	500
∞	2.71	2.30	2.08	1.94	1.85	1.77	1.72	1.67	1.63	1.60	1.49	1.42	1.34	1.26	1.18	1.13	1.08	1.00	∞

续

$\alpha = 0.05$

f_2 \ f_1	1	2	3	4	5	6	7	8	9	10	12	14	16	18	20	f_1 / f_2
1	161	200	216	225	230	234	237	239	241	242	244	245	246	247	248	1
2	18.5	19.0	19.2	19.2	19.3	19.3	19.4	19.4	19.4	19.4	19.4	19.4	19.4	19.4	19.4	2
3	10.1	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81	8.79	8.74	8.71	8.69	8.67	8.66	3
4	7.71	6.94	6.56	6.39	6.26	6.16	6.09	6.04	6.00	5.96	5.91	5.87	5.84	5.82	5.80	4
5	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77	4.74	4.68	4.64	4.60	4.58	4.56	5
6	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10	4.06	4.00	3.96	3.92	3.90	3.87	6
7	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68	3.64	3.57	3.53	3.49	3.47	3.44	7
8	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39	3.35	3.28	3.24	3.20	3.17	3.15	8
9	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18	3.14	3.07	3.03	2.99	2.96	2.94	9
10	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02	2.98	2.91	2.86	2.83	2.80	2.77	10
11	4.84	3.98	3.59	3.36	3.20	3.09	3.01	2.95	2.90	2.85	2.79	2.74	2.70	2.67	2.65	11
12	4.75	3.89	3.49	3.26	3.11	3.00	2.91	2.85	2.80	2.75	2.69	2.64	2.60	2.57	2.54	12
13	4.67	3.81	3.41	3.18	3.03	2.92	2.83	2.77	2.71	2.67	2.60	2.55	2.51	2.48	2.46	13
14	4.60	3.74	3.34	3.11	2.96	2.85	2.76	2.70	2.65	2.60	2.53	2.48	2.44	2.41	2.39	14
15	4.54	3.68	3.29	3.06	2.90	2.79	2.71	2.64	2.59	2.54	2.48	2.42	2.38	2.35	2.33	15
16	4.49	3.63	3.24	3.01	2.85	2.74	2.66	2.59	2.54	2.49	2.42	2.37	2.33	2.30	2.28	16
17	4.45	3.59	3.20	2.96	2.81	2.70	2.61	2.55	2.49	2.45	2.38	2.33	2.29	2.26	2.23	17
18	4.41	3.55	3.16	2.93	2.77	2.66	2.58	2.51	2.46	2.41	2.34	2.29	2.25	2.22	2.19	18
19	4.38	3.52	3.13	2.90	2.74	2.63	2.54	2.48	2.42	2.38	2.31	2.26	2.21	2.18	2.16	19
20	4.35	3.49	3.10	2.87	2.71	2.60	2.51	2.45	2.39	2.35	2.28	2.22	2.18	2.15	2.12	20
21	4.32	3.47	3.07	2.84	2.68	2.57	2.49	2.42	2.37	2.32	2.25	2.20	2.16	2.12	2.10	21
22	4.30	3.44	3.05	2.82	2.66	2.55	2.46	2.40	2.34	2.30	2.23	2.17	2.13	2.10	2.07	22
23	4.28	3.42	3.03	2.80	2.64	2.53	2.44	2.37	2.32	2.27	2.20	2.15	2.11	2.07	2.05	23
24	4.26	3.40	3.01	2.78	2.62	2.51	2.42	2.36	2.30	2.25	2.18	2.13	2.09	2.05	2.03	24
25	4.24	3.39	2.99	2.76	2.60	2.49	2.40	2.34	2.28	2.24	2.16	2.11	2.07	2.04	2.01	25
26	4.23	3.37	2.98	2.74	2.59	2.47	2.39	2.32	2.27	2.22	2.15	2.09	2.05	2.02	1.99	26
27	4.21	3.35	2.96	2.73	2.57	2.46	2.37	2.31	2.25	2.20	2.13	2.08	2.04	2.00	1.97	27
28	4.20	3.34	2.95	2.71	2.56	2.45	2.36	2.29	2.24	2.19	2.12	2.06	2.02	1.99	1.96	28
29	4.18	3.33	2.93	2.70	2.55	2.43	2.35	2.28	2.22	2.18	2.10	2.05	2.01	1.97	1.94	29
30	4.17	3.32	2.92	2.69	2.53	2.42	2.33	2.27	2.21	2.16	2.09	2.04	1.99	1.96	1.93	30
32	4.15	3.29	2.90	2.67	2.51	2.40	2.31	2.24	2.19	2.14	2.07	2.01	1.97	1.94	1.91	32
34	4.13	3.28	2.88	2.65	2.49	2.38	2.29	2.23	2.17	2.12	2.05	1.99	1.95	1.92	1.89	34
36	4.11	3.26	2.87	2.63	2.54	2.36	2.28	2.21	2.15	2.11	2.03	1.98	1.93	1.90	1.87	36
38	4.10	3.24	2.85	2.62	2.46	2.35	2.26	2.19	2.14	2.09	2.02	1.96	1.92	1.88	1.85	38
40	4.08	3.23	2.84	2.61	2.45	2.34	2.25	2.18	2.12	2.08	2.00	1.95	1.90	1.87	1.84	40
42	4.07	3.22	2.83	2.59	2.44	2.32	2.24	2.17	2.11	2.06	1.99	1.93	1.89	1.86	1.83	42
44	4.06	3.21	2.82	2.58	2.43	2.31	2.23	2.16	2.10	2.05	1.98	1.92	1.88	1.84	1.81	44
46	4.05	3.20	2.81	2.57	2.42	2.30	2.22	2.15	2.09	2.04	1.97	1.91	1.87	1.83	1.80	46
48	4.04	3.19	2.80	2.57	2.41	2.29	2.21	2.14	2.08	2.03	1.96	1.90	1.86	1.82	1.79	48
50	4.03	3.18	2.79	2.56	2.40	2.29	2.20	2.13	2.07	2.03	1.95	1.89	1.85	1.81	1.78	50

续

$\alpha=0.05$

f_2 \ f_1	1	2	3	4	5	6	7	8	9	10	12	14	16	18	20	f_1 / f_2
60	4.00	3.15	2.76	2.53	2.37	2.25	2.17	2.10	2.04	1.99	1.92	1.86	1.82	1.78	1.75	60
80	3.96	3.11	2.72	2.49	2.33	2.21	2.13	2.06	2.00	1.95	1.88	1.82	1.77	1.73	1.70	80
100	3.94	3.09	2.70	2.46	2.31	2.19	2.10	2.03	1.97	1.93	1.85	1.79	1.75	1.71	1.68	100
125	3.92	3.07	2.68	2.44	2.29	2.17	2.08	2.01	1.96	1.91	1.83	1.77	1.72	1.69	1.65	125
150	3.90	3.06	2.66	2.43	2.27	2.16	2.07	2.00	1.94	1.89	1.82	1.76	1.71	1.67	1.64	150
200	3.89	3.04	2.65	2.42	2.26	2.14	2.06	1.98	1.93	1.88	1.80	1.74	1.69	1.66	1.62	200
300	3.87	3.03	2.63	2.40	2.24	2.13	2.04	1.97	1.91	1.86	1.78	1.72	1.68	1.64	1.61	300
500	3.86	3.01	2.62	2.39	2.23	2.12	2.03	1.96	1.90	1.85	1.77	1.71	1.66	1.62	1.59	500
1000	3.85	3.00	2.61	2.38	2.22	2.11	2.02	1.95	1.89	1.84	1.76	1.70	1.65	1.61	1.58	1000
∞	3.84	3.00	2.60	2.37	2.21	2.10	2.01	1.94	1.88	1.83	1.75	1.69	1.64	1.60	1.57	∞

$\alpha=0.05$

f_2 \ f_1	22	24	26	28	30	35	40	45	50	60	80	100	200	500	∞	f_1 / f_2
1	249	249	249	250	250	251	251	251	252	252	252	253	254	254	254	1
2	19.5	19.5	19.5	19.5	19.5	19.5	19.5	19.5	19.5	19.5	19.5	19.5	19.5	19.5	19.5	2
3	8.65	8.64	8.63	8.62	8.62	8.60	8.59	8.59	8.58	8.57	8.56	8.55	8.54	8.53	8.53	3
4	5.79	5.77	5.76	5.75	5.75	5.73	5.72	5.71	5.70	5.69	5.67	5.66	5.65	5.64	5.63	4
5	4.54	4.53	4.52	4.50	4.50	4.48	4.46	4.45	4.44	4.43	4.41	4.41	4.39	4.37	4.37	5
6	3.86	3.84	3.83	3.82	3.81	3.79	3.77	3.76	3.75	3.74	3.72	3.71	3.69	3.68	3.67	6
7	3.43	3.41	3.40	3.39	3.38	3.36	3.34	3.33	3.32	3.30	3.29	3.27	3.25	3.24	3.23	7
8	3.13	3.12	3.10	3.09	3.08	3.06	3.04	3.03	3.02	3.01	2.99	2.97	2.95	2.94	2.93	8
9	2.92	2.90	2.89	2.87	2.86	2.84	2.83	2.81	2.80	2.79	2.77	2.76	2.73	2.72	2.71	9
10	2.75	2.74	2.72	2.71	2.70	2.68	2.66	2.65	2.64	2.62	2.60	2.59	2.56	2.55	2.54	10
11	2.63	2.61	2.59	2.58	2.57	2.55	2.53	2.52	2.51	2.49	2.47	2.46	2.43	2.42	2.40	11
12	2.52	2.51	2.49	2.48	2.47	2.44	2.43	2.41	2.40	2.38	2.36	2.35	2.32	2.31	2.30	12
13	2.44	2.42	2.41	2.39	2.38	2.36	2.34	2.33	2.31	2.30	2.27	2.26	2.23	2.22	2.21	13
14	2.37	2.35	2.33	2.32	2.31	2.28	2.27	2.25	2.24	2.22	2.20	2.19	2.16	2.14	2.13	14
15	2.31	2.29	2.27	2.26	2.25	2.22	2.20	2.19	2.18	2.16	2.14	2.12	2.10	2.08	2.07	15
16	2.25	2.24	2.22	2.21	2.19	2.17	2.15	2.14	2.12	2.11	2.08	2.07	2.04	2.02	2.01	16
17	2.21	2.19	2.17	2.16	2.15	2.12	2.10	2.09	2.08	2.06	2.03	2.02	1.99	1.97	1.96	17
18	2.17	2.15	2.13	2.12	2.11	2.08	2.06	2.05	2.04	2.02	1.99	1.98	1.95	1.93	1.92	18
19	2.13	2.11	2.10	2.08	2.07	2.05	2.03	2.01	2.00	1.98	1.96	1.94	1.91	1.89	1.88	19
20	2.10	2.08	2.07	2.05	2.04	2.01	1.99	1.98	1.97	1.95	1.92	1.91	1.88	1.86	1.84	20
21	2.07	2.05	2.04	2.02	2.01	1.98	1.96	1.95	1.94	1.92	1.89	1.88	1.84	1.82	1.81	21
22	2.05	2.03	2.01	2.00	1.98	1.96	1.94	1.92	1.91	1.89	1.86	1.85	1.82	1.80	1.78	22
23	2.02	2.00	1.99	1.97	1.96	1.93	1.91	1.90	1.88	1.86	1.84	1.82	1.79	1.77	1.76	23
24	2.00	1.98	1.97	1.96	1.94	1.91	1.89	1.88	1.86	1.84	1.82	1.80	1.77	1.75	1.73	24
25	1.98	1.96	1.95	1.93	1.92	1.89	1.87	1.86	1.84	1.82	1.80	1.78	1.75	1.73	1.71	25
26	1.97	1.95	1.93	1.91	1.90	1.87	1.85	1.84	1.82	1.80	1.78	1.76	1.73	1.71	1.69	26
27	1.95	1.93	1.91	1.90	1.88	1.86	1.84	1.82	1.81	1.79	1.76	1.74	1.71	1.69	1.67	27
28	1.93	1.91	1.90	1.88	1.87	1.84	1.82	1.80	1.79	1.77	1.74	1.73	1.69	1.67	1.65	28
29	1.92	1.90	1.88	1.87	1.85	1.83	1.81	1.79	1.77	1.75	1.73	1.71	1.67	1.65	1.64	29
30	1.91	1.89	1.87	1.85	1.84	1.81	1.79	1.77	1.76	1.74	1.71	1.70	1.66	1.64	1.62	30

续

$\alpha = 0.05$

f_2 \ f_1	22	24	26	28	30	35	40	45	50	60	80	100	200	500	∞	f_1 / f_2
32	1.88	1.86	1.85	1.83	1.82	1.79	1.77	1.75	1.74	1.71	1.69	1.67	1.63	1.61	1.59	32
34	1.86	1.84	1.82	1.80	1.80	1.77	1.75	1.73	1.71	1.69	1.66	1.65	1.61	1.59	1.57	34
36	1.85	1.82	1.81	1.79	1.78	1.75	1.73	1.71	1.69	1.67	1.64	1.62	1.59	1.56	1.55	36
38	1.83	1.81	1.79	1.77	1.76	1.73	1.71	1.69	1.68	1.65	1.62	1.61	1.57	1.54	1.53	38
40	1.81	1.79	1.77	1.76	1.74	1.72	1.69	1.67	1.66	1.64	1.61	1.59	1.55	1.53	1.51	40
42	1.80	1.78	1.76	1.74	1.73	1.70	1.68	1.66	1.65	1.62	1.59	1.57	1.53	1.51	1.49	42
44	1.79	1.77	1.75	1.73	1.72	1.69	1.67	1.65	1.63	1.61	1.58	1.56	1.52	1.49	1.48	44
46	1.78	1.76	1.74	1.72	1.71	1.68	1.65	1.64	1.62	1.60	1.57	1.55	1.51	1.48	1.46	46
48	1.77	1.75	1.73	1.71	1.70	1.67	1.64	1.62	1.61	1.59	1.56	1.54	1.49	1.47	1.45	48
50	1.76	1.74	1.72	1.70	1.69	1.66	1.63	1.61	1.60	1.58	1.54	1.52	1.48	1.46	1.44	50
60	1.72	1.70	1.68	1.66	1.65	1.62	1.59	1.57	1.56	1.53	1.50	1.48	1.44	1.41	1.39	60
80	1.68	1.65	1.63	1.62	1.60	1.57	1.54	1.52	1.51	1.48	1.45	1.43	1.38	1.35	1.32	80
100	1.65	1.63	1.61	1.59	1.57	1.54	1.52	1.49	1.48	1.45	1.41	1.39	1.34	1.31	1.28	100
125	1.63	1.60	1.58	1.57	1.55	1.52	1.49	1.47	1.45	1.42	1.39	1.36	1.31	1.27	1.25	125
150	1.61	1.59	1.57	1.55	1.53	1.50	1.48	1.45	1.44	1.41	1.37	1.34	1.29	1.25	1.22	150
200	1.60	1.57	1.55	1.53	1.52	1.48	1.46	1.43	1.41	1.39	1.35	1.32	1.26	1.22	1.19	200
300	1.58	1.55	1.53	1.51	1.50	1.46	1.43	1.41	1.39	1.36	1.32	1.30	1.23	1.19	1.15	300
500	1.56	1.54	1.52	1.50	1.48	1.45	1.42	1.40	1.38	1.34	1.30	1.28	1.21	1.16	1.11	500
1000	1.55	1.53	1.51	1.49	1.47	1.44	1.41	1.38	1.36	1.33	1.29	1.26	1.19	1.13	1.08	1000
∞	1.54	1.52	1.50	1.48	1.46	1.42	1.39	1.37	1.35	1.32	1.27	1.24	1.17	1.11	1.00	∞

$\alpha = 0.01$

f_2 \ f_1	1	2	3	4	5	6	7	8	9	10	12	14	16	18	20	f_1 / f
1	405	500	540	563	576	586	593	598	602	606	611	614	617	619	621	1
2	98.5	99.0	99.2	99.2	99.3	99.3	99.4	99.4	99.4	99.4	99.4	99.4	99.4	99.4	99.4	2
3	34.1	30.8	29.5	28.7	28.2	27.9	27.7	27.5	27.3	27.2	27.1	26.9	26.8	26.8	26.7	3
4	21.2	18.0	15.7	16.0	15.5	15.2	15.0	14.8	14.7	14.5	14.4	14.2	14.2	14.1	14.0	4
5	16.3	13.3	12.1	11.4	11.0	10.7	10.5	10.3	10.2	10.1	9.89	9.77	9.68	9.61	9.55	5
6	13.7	10.9	9.78	9.15	8.75	8.47	8.26	8.10	7.98	7.87	7.72	7.60	7.52	7.45	7.40	6
7	12.2	9.55	8.45	7.85	7.46	7.19	6.99	6.84	6.72	6.62	6.47	6.36	6.27	6.21	6.16	7
8	11.3	8.65	7.59	7.01	6.63	6.37	6.18	6.03	5.91	5.81	5.67	5.56	5.48	5.41	5.36	8
9	10.6	8.02	6.99	6.42	6.06	5.80	5.61	5.47	5.35	5.26	5.11	5.00	4.92	4.86	4.81	9
10	10.0	7.56	6.55	5.99	5.64	5.39	5.20	5.06	4.94	4.85	4.71	4.60	4.52	4.46	4.41	10
11	9.65	7.21	6.22	5.67	5.32	5.07	4.89	4.74	4.63	4.54	4.40	4.29	4.21	4.15	4.10	11
12	9.33	6.93	5.95	5.41	5.06	4.82	4.64	4.50	4.39	4.30	4.16	4.05	3.97	3.91	3.86	12
13	9.07	6.70	5.74	5.21	4.86	4.62	4.44	4.30	4.19	4.10	3.96	3.86	3.78	3.71	3.66	13
14	8.86	6.51	5.56	5.04	4.70	4.46	4.28	4.14	4.03	3.94	3.80	3.70	3.62	3.56	3.51	14
15	8.68	6.36	5.42	4.89	4.56	4.32	4.14	4.00	3.89	3.80	3.67	3.56	3.49	3.42	3.37	15
16	8.53	6.23	5.29	4.77	4.44	4.20	4.03	3.89	3.78	3.69	3.55	3.45	3.37	3.31	3.26	16
17	8.40	6.11	5.18	4.67	4.34	4.10	3.93	3.79	3.68	3.59	3.46	3.35	3.27	3.21	3.16	17
18	8.29	6.01	5.09	4.58	4.25	4.01	3.84	3.71	3.60	3.51	3.37	3.27	3.19	3.13	3.08	18
19	8.18	5.93	5.01	4.50	4.17	3.94	3.77	3.63	3.52	3.43	3.30	3.19	3.12	3.05	3.00	19
20	8.10	5.85	4.94	4.43	4.10	3.87	3.70	3.56	3.46	3.37	3.23	3.13	3.05	2.99	2.94	20

续 $\alpha = 0.01$

f_2 \ f_1	1	2	3	4	5	6	7	8	9	10	12	14	16	18	20	f_1 / f_2
21	8.02	5.78	4.87	4.37	4.04	3.81	3.64	3.51	3.40	3.31	3.17	3.07	2.99	2.93	2.88	21
22	7.95	5.72	4.82	4.31	3.99	3.76	3.59	3.45	3.35	3.26	3.12	3.02	2.94	2.88	2.83	22
23	7.88	5.66	4.76	4.26	3.94	3.71	3.54	3.41	3.30	3.21	3.07	2.97	2.89	2.83	2.78	23
24	7.82	5.61	4.72	4.22	3.90	3.67	3.50	3.36	3.26	3.17	3.03	2.93	2.85	2.79	2.74	24
25	7.77	5.57	4.68	4.18	3.86	3.63	3.46	3.32	3.22	3.13	2.99	2.89	2.81	2.75	2.70	25
26	7.72	5.53	4.64	4.14	3.82	3.59	3.42	3.29	3.18	3.09	2.96	2.86	2.78	2.72	2.66	26
27	7.68	5.49	4.60	4.11	3.78	3.56	3.39	3.26	3.15	3.06	2.93	2.82	2.75	2.68	2.63	27
28	7.64	5.45	4.57	4.07	3.75	3.53	3.36	3.23	3.12	3.03	2.90	2.79	2.72	2.65	2.60	28
29	7.60	5.42	4.54	4.04	3.73	3.50	3.33	3.20	3.09	3.00	2.87	2.77	2.69	2.63	2.57	29
30	7.56	5.39	4.51	4.02	3.70	3.47	3.30	3.17	3.07	2.98	2.84	2.74	2.66	2.60	2.55	30
32	7.50	5.34	4.46	3.97	3.65	3.43	3.26	3.13	3.02	2.93	2.80	2.70	2.62	2.55	2.50	32
34	7.44	5.29	4.42	3.93	3.61	3.39	3.22	3.09	2.98	2.89	2.76	2.66	2.58	2.51	2.46	34
36	7.40	5.25	4.38	3.89	3.57	3.35	3.18	3.05	2.95	2.86	2.72	2.62	2.54	2.48	2.43	36
38	7.35	5.21	4.34	3.86	3.54	3.32	3.15	3.02	2.92	2.83	2.69	2.59	2.51	2.45	2.40	38
40	7.31	5.18	4.31	3.83	3.51	3.29	3.12	2.99	2.89	2.80	2.66	2.56	2.48	2.42	2.37	40
42	7.28	5.15	4.29	3.80	3.49	3.27	3.10	2.97	2.86	2.78	2.64	2.54	2.46	2.40	2.34	42
44	7.25	5.12	4.26	3.78	3.47	3.24	3.08	2.95	2.84	2.75	2.62	2.52	2.44	2.37	2.32	44
46	7.22	5.10	4.24	3.76	3.44	3.22	3.06	2.93	2.82	2.73	2.60	2.50	2.42	2.35	2.30	46
48	7.20	5.08	4.22	3.74	3.43	3.20	3.04	2.91	2.80	2.72	2.58	2.49	2.40	2.33	2.28	48
50	7.17	5.06	4.20	3.72	3.41	3.19	3.02	2.89	2.79	2.70	2.56	2.46	2.38	2.32	2.27	50
60	7.08	4.98	4.13	3.65	3.34	3.12	2.95	2.82	2.72	2.63	2.50	2.39	2.31	2.25	2.20	60
80	6.96	4.88	4.04	3.56	3.26	3.04	2.87	2.74	2.64	2.55	2.42	2.31	2.23	2.17	2.12	80
100	6.90	4.82	3.98	3.51	3.21	2.99	2.82	2.69	2.59	2.50	2.37	2.26	2.19	2.12	2.07	100
125	6.84	4.78	3.94	3.47	3.17	2.95	2.79	2.66	2.55	2.47	2.33	2.23	2.15	2.08	2.03	125
150	6.81	4.75	3.92	3.45	3.14	2.92	2.76	2.63	2.53	2.44	2.31	2.20	2.12	2.06	2.00	150
200	6.76	4.71	3.88	3.41	3.11	2.89	2.73	2.60	2.50	2.41	2.27	2.17	2.09	2.02	1.97	200
300	6.72	4.68	3.85	3.38	3.08	2.86	2.70	2.57	2.47	2.38	2.24	2.14	2.06	1.99	1.94	300
500	6.69	4.65	3.82	3.36	3.05	2.84	2.68	2.55	2.44	2.36	2.22	2.12	2.04	1.97	1.92	500
1000	6.66	4.63	3.80	3.34	3.04	2.82	2.66	2.53	2.43	2.34	2.20	2.10	2.02	1.95	1.90	1000
∞	6.63	4.61	3.78	3.32	3.02	2.80	2.64	2.51	2.41	2.32	2.18	2.08	2.00	1.93	1.88	∞

$\alpha = 0.01$

f_2 \ f_1	22	24	26	28	30	35	40	45	50	60	80	100	200	500	∞	f_1 / f_2
1	622	623	624	625	626	628	629	630	630	631	633	633	635	636	637	1
2	99.5	99.5	99.5	99.5	99.5	99.5	99.5	99.5	99.5	99.5	99.5	99.5	99.5	99.5	99.5	2
3	26.6	26.6	26.6	26.5	26.5	26.5	26.4	26.4	26.4	26.3	26.3	26.2	26.2	26.1	26.1	3
4	14.0	13.9	13.9	13.9	13.8	13.8	13.7	13.7	13.7	13.7	13.6	13.6	13.5	13.5	13.5	4
5	9.51	9.47	9.43	9.40	9.38	9.33	9.29	9.26	9.24	9.20	9.16	9.13	9.08	9.04	9.02	5
6	7.35	7.31	7.28	7.25	7.23	7.18	7.14	7.11	7.09	7.06	7.01	6.99	6.93	6.90	6.88	6
7	6.11	6.07	6.04	6.02	5.99	5.94	5.91	5.88	5.86	5.82	5.78	5.75	5.70	5.67	5.65	7
8	5.32	5.28	5.25	5.22	5.20	5.15	5.12	5.09	5.07	5.03	4.99	4.96	4.91	4.88	4.86	8
9	4.77	4.73	4.70	4.67	4.65	4.60	4.57	4.54	4.52	4.48	4.44	4.42	4.36	4.33	3.31	9
10	4.36	4.33	4.30	4.27	4.25	4.20	4.17	4.14	4.12	4.08	4.04	4.01	3.96	3.93	3.91	10

续

$\alpha = 0.01$

f_2 \ f_1	22	24	26	28	30	35	40	45	50	60	80	100	200	500	∞	f_1 / f_2
11	4.06	4.02	3.99	3.96	3.94	3.89	3.86	3.83	3.81	3.78	3.73	3.71	3.66	3.62	3.60	11
12	3.82	3.78	3.75	3.72	3.70	3.65	3.62	3.59	3.57	3.54	3.49	3.47	3.41	3.38	3.36	12
13	3.62	3.59	3.56	3.53	3.51	3.46	3.43	3.40	3.38	3.34	3.30	3.27	3.22	3.19	3.17	13
14	3.46	3.43	3.40	3.37	3.35	3.30	3.27	3.24	3.22	3.18	3.14	3.11	3.06	3.03	3.00	14
15	3.33	3.29	3.26	3.24	3.21	3.17	3.13	3.10	3.08	3.05	3.00	2.98	2.92	2.89	2.87	15
16	3.22	3.18	3.15	3.12	3.10	3.05	3.02	2.99	2.97	2.93	2.89	2.86	2.81	2.78	2.75	16
17	3.12	3.08	3.05	3.03	3.00	2.96	2.92	2.89	2.87	2.83	2.79	2.76	2.71	2.68	2.65	17
18	3.03	3.00	2.97	2.94	2.92	2.87	2.84	2.81	2.78	2.75	2.70	2.68	2.62	2.54	2.57	18
19	2.96	2.92	2.89	2.87	2.84	2.80	2.76	2.73	2.71	2.67	2.63	2.60	2.55	2.51	2.49	19
20	2.90	2.86	2.83	2.80	2.78	2.73	2.69	2.67	2.64	2.61	2.56	2.54	2.48	2.44	2.42	20
21	2.84	2.80	2.77	2.74	2.72	2.67	2.64	2.61	2.58	2.55	2.50	2.48	2.42	2.38	2.36	21
22	2.78	2.75	2.72	2.69	2.67	2.62	2.58	2.55	2.53	2.50	2.45	2.42	2.36	2.33	2.31	22
23	2.74	2.70	2.67	2.64	2.62	2.57	2.54	2.51	2.48	2.45	2.40	2.37	2.32	2.28	2.26	23
24	2.70	2.66	2.63	2.60	2.58	2.53	2.49	2.46	2.44	2.40	2.36	2.33	2.27	2.24	2.21	24
25	2.66	2.62	2.59	2.56	2.54	2.49	2.45	2.42	2.40	2.36	2.32	2.29	2.23	2.19	2.17	25
26	2.62	2.58	2.55	2.53	2.50	2.45	2.42	2.39	2.36	2.33	2.28	2.25	2.19	2.16	2.13	26
27	2.59	2.55	2.52	2.49	2.47	2.42	2.38	2.35	2.33	2.29	2.25	2.22	2.16	2.12	2.10	27
28	2.56	2.52	2.49	2.46	2.44	2.39	2.35	2.32	2.30	2.26	2.22	2.19	2.13	2.09	2.06	28
29	2.53	2.49	2.46	2.44	2.41	2.36	2.33	2.30	2.27	2.23	2.19	2.16	2.10	2.06	2.03	29
30	2.51	2.47	2.44	2.41	2.39	2.34	2.30	2.27	2.25	2.21	2.16	2.13	2.07	2.03	2.01	30
32	2.46	2.42	2.39	2.36	2.34	2.29	2.25	2.22	2.20	2.16	2.11	2.08	2.02	1.98	1.96	32
34	2.42	2.38	2.35	2.32	2.30	2.25	2.21	2.18	2.16	2.12	2.07	2.04	1.98	1.94	1.91	34
36	2.38	2.35	2.32	2.29	2.26	2.21	2.17	2.14	2.12	2.08	2.03	2.00	1.94	1.90	1.87	36
38	2.35	2.32	2.28	2.26	2.23	2.18	2.14	2.11	2.09	2.05	2.00	1.97	1.90	1.86	1.84	38
40	2.33	2.29	2.26	2.23	2.20	2.15	2.11	2.08	2.06	2.02	1.97	1.94	1.87	1.83	1.80	40
42	2.30	2.26	2.23	2.20	2.18	2.13	2.09	2.06	2.03	1.99	1.94	1.91	1.85	1.80	1.78	42
44	2.28	2.24	2.21	2.18	2.15	2.10	2.06	2.03	2.01	1.97	1.92	1.89	1.82	1.78	1.75	44
46	2.26	2.22	2.19	2.16	2.13	2.08	2.04	2.01	1.99	1.95	1.90	1.86	1.80	1.75	1.73	46
48	2.24	2.20	2.17	2.14	2.12	2.06	2.02	1.99	1.97	1.93	1.88	1.84	1.78	1.73	1.70	48
50	2.22	2.18	2.15	2.12	2.10	2.05	2.01	1.97	1.95	1.91	1.86	1.82	1.76	1.71	1.68	50
60	2.15	2.12	2.03	2.05	2.03	1.98	1.94	1.90	1.88	1.84	1.78	1.75	1.68	1.63	1.60	60
80	2.07	2.03	2.00	1.97	1.94	1.89	1.85	1.81	1.79	1.75	1.69	1.66	1.58	1.53	1.49	80
100	2.02	1.98	1.94	1.92	1.89	1.84	1.80	1.76	1.73	1.69	1.63	1.60	1.52	1.47	1.43	100
125	1.98	1.94	1.91	1.88	1.85	1.80	1.76	1.72	1.69	1.65	1.59	1.55	1.47	1.41	1.37	125
150	1.96	1.92	1.88	1.85	1.83	1.77	1.73	1.69	1.66	1.62	1.56	1.52	1.43	1.38	1.33	150
200	1.93	1.89	1.85	1.82	1.79	1.74	1.69	1.66	1.63	1.58	1.52	1.48	1.39	1.33	1.28	200
300	1.89	1.85	1.82	1.79	1.76	1.71	1.66	1.62	1.59	1.55	1.48	1.44	1.35	1.28	1.22	300
500	1.87	1.83	1.79	1.76	1.74	1.68	1.63	1.60	1.56	1.52	1.45	1.41	1.31	1.23	1.16	500
1000	1.85	1.81	1.77	1.74	1.72	1.66	1.61	1.57	1.54	1.50	1.43	1.38	1.28	1.19	1.11	1000
∞	1.83	1.79	1.76	1.72	1.70	1.64	1.59	1.55	1.52	1.47	1.40	1.36	1.25	1.15	1.00	∞

主要参考文献

王仁铎、胡光道，1983，线性地质统计学，地质出版社。

王学仁，1982，地质数据的多变量统计分析，科学出版社。

中国科学院地质研究所，1977，数学地质引论，地质出版社。

刘承祚、孙惠文，1981，数学地质基本方法及应用，地质出版社。

余金生、李裕伟，1985，地质因子分析，地质出版社。

朱裕生，1984，矿产资源评价方法学导论，地质出版社。

於崇文等，1980，数学地质的方法与应用，冶金出版社。

赵鹏大、胡旺亮、李紫金，1983，矿床统计预测，地质出版社。

韩金炎，1987，数学地质，煤炭工业出版社。

植东升，1987，地质找矿勘探中的概率统计方法，中南工业大学出版社。

侯景儒、黄竞先，1981，地质统计学及其在矿产储量计算中的应用，地质出版社。

小乔治，S. K.，琢查德，F. L.，1978，地质数据统计分析，科学出版社。

阿格特伯格，F. P.，1980，地质数学，科学出版社。

John, C. Davis, 1983, Statistics and Data Analysis in Geology.

Hampel, F. R., 1968, Contribution to the theory of robust estimation, Ph. D. Dissertaion, Univ. of california, Berkeley.

Huber, P. J., 1964, Robust estimation of Location perameter, Am. Math. Statist. 35, 73-101.

Huber, P. J. 1981, Robust Statistics, Jahn Eiley & Sons, New York, 308P.

Tukey, J. W., 1960, Asurvey of sampling from contaminated distributions, in I, olkin, Contributions to probability Statistics, Stanford Univ. Press.

Tukey, J. W., 1962, The future of data analysis, Am. Math. Statist. 33, 1-67.

Tukey, J. W., 1977, Exploratory data analysis, Addison-Wesley Reading, Mass.