多元数据分析方法

——纯代数处理

胡国定 张润楚 著

南开大学出版社

DUOYUANSHUJUFENXIFANGFA

多元数据分析方法

——纯代数处理

胡国定　张润楚　著

内 容 简 介

本书用纯代数分析观点系统地介绍了回归分析、逐步回归、相关分析、主成分分析、因子分析、典型分析、相应分析、判别分析、聚类分析、多维标度法和周期分析等一些常用的多元分析方法以及正交试验设计的一些基本理论。其内容深入浅出，方法多样实用，还配有一定例题。

本书主要以从事应用统计理论和实际工作的科技人员为阅读对象，也可以作为大专院校数理统计、经济管理和其它有关专业师生的教学参考书。

多元数据分析方法

——纯代数处理

胡国定　张润楚　著

南开大学出版社出版

（天津八里台南开大学校内）

邮政编码 300071　电话 34.9318

新华书店天津发行所发行

河北省邮电印刷厂印刷

1990年6月第1版　1990年6月第2次印刷

开本：850×1168 1/32 印张：13.625 插页：4

字数：338千　印数：501-2,500

ISBN 7-310-00239-3/O·42　定价：7.00元

前　言

多元分析是数理统计中最重要的分支之一，它的理论和方法在近半个世纪获得了飞速的发展，尤其是最近十几年来，由于电子计算机的普及，它在其他许多学科领域里，例如生物、医学、地质、农业、工程技术、气象、社会经济等，得到日益广泛的应用，使得这个分支的发展更加活跃和深入。

传统的多元统计分析是建立在概率论的基础之上的，对于给定的观测数据，一般是首先提出一种统计模型假设（通常要求满足独立和正态性条件），然后在这一假设条件下，导出它的一些理论和方法，作出统计推断（如参数估计和假设检验等）。诚然这是十分必要的，而且这些理论和方法已得到广泛的应用，成为数据分析的基础．但是在实际中还大量地存在另一类问题，即我们获得的许多数据资料，由于种种原因，可能不符合现有的某个统计模型的假设或者根本不知道它应该符合哪一种统计模型，这时如果仍然套用某个统计模型，运用传统分析方法，则往往达不到预期的效果．这类问题早已引起许多统计学家和实际工作者的注意和研究，随之另一类分析数据的方法即所谓探索性数据分析法便应运而生．此方法的基本思想是，一般不对数据作模型假设，而直接从实际数据出发，考察和分析大量数据的结构和特征，从中提取主要而准确的信息，必要时进一步归纳出反映数据的较切实的模型为统计推断作准备。在这类分析方法中，尤其当涉及多元数据分析时，代数方法和数值计算为其重要特征之一．迄今为止，这种分析数据的观点和方法已为统计方面的广大理论和实际工作者所接受，并在实际应用中发挥了显著的效用。

本书的目的是试图从纯代数分析的角度较系统地介绍一些常用的多元分析方法。在70年代末，本书作者编写了一本“数据分析”

讲义，曾在南开大学数学系有关专业作为多元分析课的部分内容讲授过多次，本书是在原来讲义的基础上补充修改而写成的。它的主要阅读对象是从事统计方面的理论和实际工作者，也可作为大专院校数理统计、经济管理和其它有关专业多元分析课的教学参考书。本书的内容比较浅显易懂，只要具有高等数学主要是线性代数基础知识的读者即可阅读。另外，它介绍的方法较多而且实用。因此我们也希望本书对多元分析方法的普及发挥一定的作用。

本书的第1章，介绍了向量和矩阵代数的一些基本知识，以作为以后各章节的数学工具。从第2章到第12章分别介绍了回归分析、逐步回归、相关分析、主成分分析、因子分析、典型分析、相应分析、判别分析、聚类分析、多维标度法和周期分析等一些常用多元分析方法。在这些章节中给出了各种技术的基本数学理论和方法，力求做到在理论上叙述严格，在方法介绍上多样实用，为此在书中基本上对每种方法都给出了一些例子说明它们如何应用。第13章介绍的是正交试验设计，以一种函数逼近和插补论观点介绍了这一题目的基本理论和思想，由于篇幅所限，这章举例不多，读者可以从有关专著或实际中找到相应例子。本书没有从概率论的观点出发介绍这些方法，所以不能说它们是这些方法的全部内容，尤其是没有涉及一些与概率论有关的理论问题。本书末尾列出了不少有关多元分析和试验设计方面的参考书目和文章，其中一些为本书的写作提供了素材，我们谨向这些著作的作者表示感谢，读者也可以从这些参考文献中找到许多自己感兴趣的问题.

南开大学出版社的同志为本书的出版做了大量的工作，特别是王家骅副教授认真仔细地校阅了全部书稿，并提出了许多好的意见和建议，在此我们也向他们表示衷心的感谢。

由于我们水平所限，而且对这一新课题的研究还有待进一步完善和深入，因此书中难免会有一些缺点甚至错误。我们希望读者给予批评指正。

作　者

目　录

第1章　线性代数基本知识……（1）

§1.1　向量、矩阵、行列式……（1）

1.1.1　向量的概念……（1）

1.1.2　向量的运算……（2）

1.1.3　向量组的性质……（3）

1.1.4　矩阵的概念……（6）

1.1.5　矩阵的运算……（8）

1.1.6　几种特殊矩阵……（11）

1.1.7　初等方阵和矩阵的简化……（13）

1.1.8　行列式……（14）

§1.2　线性方程组……（16）

1.2.1　解存在的条件……（16）

1.2.2　相容方程的消去解法……（18）

1.2.3　不相容方程的组似近解法——最小二乘法……（27）

§1.3　向量空间、欧氏空间……（31）

1.3.1　向量空间的定义……（31）

1.3.2　向量空间的基底、维数……（32）

1.3.3　子空间、子空间的交与和……（34）

1.3.4　一般线性空间的定义……（34）

1.3.5　欧氏空间的概念……（35）

1.3.6　向量的长度、距离和夹角……（37）

1.3.7　向量的正交、向量与子空间的正交，向量到子空间的投影、距离……（38）

1.3.8　标准正交基及一般基的标准正交化……（40）

§1.4 二次型与格兰姆矩阵 ……（ 43 ）
1.4.1 二次型的分类和变换 ……（ 43 ）
1.4.2 矩阵的特征值和特征向量 ……（ 46 ）
1.4.3 格兰姆矩阵及其行列式 ……（ 50 ）
1.4.4 二次型的极值 ……（ 53 ）
第2章 回归分析 ……（ 56 ）
§2.1 问题的提出 ……（ 56 ）
2.1.1 从数据到方程 ……（ 56 ）
2.1.2 函数线性逼近和向量线性逼近 ……（ 58 ）
§2.2 回归分析的数学方法 ……（ 59 ）
2.2.1 线性逼近的求解和正规方程组 ……（ 59 ）
2.2.2 误差分析，回归效果的衡量 ……（ 61 ）
2.2.3 因子多少对于回归效果的影响 ……（ 65 ）
2.2.4 因子标度变换与回归效果的关系 ……（ 66 ）
2.2.5 对因子数据中心化的回归 ……（ 67 ）
2.2.6 降秩情形的处理 ……（ 69 ）
2.2.7 正规方程组的求解 ……（ 70 ）
2.2.8 例题 ……（ 71 ）
§2.3 回归分析的应用 ……（ 74 ）
2.3.1 用于复合函数的回归 ……（ 74 ）
2.3.2 复合回归的一些具体应用方法 ……（ 78 ）
2.3.3 时间函数和预报 ……（ 88 ）
2.3.4 趋势面分析 ……（ 92 ）
2.3.5 多个函数的回归 ……（ 95 ）
2.3.6 正交回归 ……（ 96 ）
第3章 逐步回归 ……（ 99 ）
§3.1 基本思想 ……（ 99 ）
§3.2 逐步回归的实施方法 ……（ 102 ）
§3.3 相关方阵 ……（ 103 ）

§3.4 具体计算步骤 ……（114）
§3.5 例题 ……（117）
第4章 相关分析……（125）
§4.1 两个变量的相关 ……（125）
4.1.1 引言……（125）
4.1.2 相关系数 ……（125）
4.1.3 相关系数的几何意义……（128）
4.1.4 最大相关系数 ……（130）
§4.2 多个变量的相关 ……（134）
4.2.1 广义相关系数和广义方差……（134）
4.2.2 广义相关系数的性质……（137）
§4.3 多重相关和偏相关 ……（139）
4.3.1 多重相关系数 ……（139）
4.3.2 偏相关系数 ……（141）
§4.4 变量组的相关筛选 ……（144）
4.4.1 多重相关法 ……（145）
4.4.2 内相关性法 ……（145）
4.4.3 相对无关系数法……（146）
4.4.4 主成分分析法 ……（147）
第5章 主成分分析……（149）
§5.1 引言 ……（149）
§5.2 数据拟合观点 ……（150）
5.2.1 在R^p中拟合数据点 ……（150）
5.2.2 主轴、主坐标和主成分……（152）
5.2.3 用主成分恢复原始数据……（154）
5.2.4 在R^n中拟合以及与在R^p中拟合的关系……（155）
§5.3 最优依次逼近观点 ……（157）
5.3.1 问题的提出 ……（157）
5.3.2 最优组合因子的导出……（158）

§5.4 主成分的性质和计算步骤 ……（162）
5.4.1 主成分的性质 ……（162）
5.4.2 算法步骤 ……（165）
§5.5 主成分分析的应用 ……（166）
5.5.1 在R^p中的分析 ……（166）
5.5.2 在R^n中的分析 ……（169）
5.5.3 补充变量和补充个体 ……（171）
5.5.4 结果的图表示和解释 ……（173）
5.5.5 例题 ……（175）
§5.6 主成分回归分析 ……（179）
第6章 因子分析 ……（182）
§6.1 基本因子模型 ……（182）
6.1.1 引言 ……（182）
6.1.2 基本因子分析模型 ……（183）
§6.2 因子模型的基本性质 ……（186）
6.2.1 共性变差和个性变差 ……（186）
6.2.2 标度变换不变性 ……（187）
6.2.3 因子载荷的不唯一性 ……（188）
§6.3 因子模型的求解 ……（189）
6.3.1 主因子法 ……（189）
6.3.2 主成分分析法 ……（197）
6.3.3 重心法 ……（201）
§6.4 因子得分 ……（204）
§6.5 方差最大正交旋转 ……（206）
第7章 典型分析 ……（213）
§7.1 问题的阐述和记号 ……（213）
7.1.1 引言 ……（213）
7.1.2 数学描述和记号 ……（214）
§7.2 求解方法和典型变量的性质 ……（215）

7.2.1 数学解法 …… (215)
7.2.2 典型变量的性质 …… (217)
7.2.3 计算步骤 …… (219)
§7.3 典型分析的几何解释 …… (222)
7.3.1 典型分析视为几何变换 …… (222)
7.3.2 和多重回归的关系 …… (225)
7.3.3 降秩情形的处理 …… (226)
§7.4 典型得分和预测 …… (226)
§7.5 定性数据的典型分析 …… (227)
第 8 章 相应分析 …… (231)
§8.1 引言 …… (231)
§8.2 相应分析的一般提法 …… (233)
8.2.1 一般 2 维列联表 …… (233)
8.2.2 分布轮廓及有关记号 …… (235)
8.2.3 距离的选择 …… (237)
8.2.4 拟合优度准则的选择 …… (239)
8.2.5 问题的一般提法 …… (241)
§8.3 相应分析的求解 …… (242)
8.3.1 关于任意距离和任意准则的一般分析 …… (242)
8.3.2 相应分析的求解 …… (243)
8.3.3 计算步骤 …… (247)
§8.4 结果的图示和解释 …… (250)
8.4.1 联立表示的图示方法 …… (250)
8.4.2 关于主轴和因素关系的解释 …… (252)
8.4.3 在重心情形下的相应分析 …… (254)
§8.5 与典型分析的关系 …… (255)
第 9 章 判别分析 …… (259)
§9.1 问题的提出 …… (259)
§9.2 以直线划分的判别法 …… (259)

9.2.1 基本思想 ……（259）
9.2.2 两个类别情形的判别 ……（260）
9.2.3 三个类别情形的判别 ……（265）
§9.3 以曲线划分的二类判别 ……（269）
9.3.1 引言 ……（269）
9.3.2 以马氏距离为尺度导致的曲线判别 ……（270）
9.3.3 一般的多项式曲线划分的判别 ……（272）
§9.4 费歇判别法 ……（274）
9.4.1 一般情形 ……（274）
9.4.2 两个群体的情形 ……（277）
§9.5 与回归分析的关系 ……（278）
§9.6 与典型分析的关系 ……（279）
§9.7 判别准则好坏的衡量 ……（282）
9.7.1 回代法 ……（282）
9.7.2 刀切法 ……（284）
第10章 聚类分析 ……（286）
§10.1 相似和距离 ……（286）
10.1.1 相似系数 ……（287）
10.1.2 距离 ……（289）
§10.2 系统聚类法 ……（291）
§10.3 一次形成分类法 ……（300）
§10.4 K水准逐步形成分类法 ……（303）
§10.5 有序样品的聚类方法 ……（308）
10.5.1 精确最优解方法——费歇算法 ……（308）
10.5.2 例题 ……（312）
§10.6 移动中心聚类法 ……（317）
10.6.1 基本算法 ……（317）
10.6.2 算法的基本原理 ……（319）
第11章 多维标度法 ……（322）

§11.1 引言……（322）
§11.2 距离阵和经典解……（323）
11.2.1 距离阵 ……（323）
11.2.2 欧氏距离阵及其判定定理……（325）
11.2.3 多维标度的经典解 ……（329）
11.2.4 相似阵情形……（333）
§11.3 经典解的优良性质……（338）
11.3.1 经典解和主成分分析的关系……（338）
11.3.2 经典解的最优性质和拟合优度……（339）
§11.4 非度量方法 ……（343）
11.4.1 引言 ……（343）
11.4.2 舍帕得-克卢斯卡算法 ……（346）
第12章 周期分析……（349）
§12.1 问题的提出……（349）
§12.2 周期分析方法……（349）
§12.3 计算步骤……（354）
第13章 正交试验设计……（356）
§13.1 问题的提出……（356）
§13.2 正交表的定义和性质……（357）
13.2.1 正交表的定义……（357）
13.2.2 正交表的性质……（360）
§13.3 正交表的构造 ……（364）
13.3.1 几个概念……（364）
13.3.2 第1种 r 级正交表的构造 ……（366）
13.3.3 第2种 r 级正交表的构造 ……（370）
13.3.4 任意第 l 种 r 级正交表的构造 ……（375）
§13.4 函数的逼近与函数的插补……（377）
§13.5 可加数学模型……（379）
§13.6 自变量的变换……（384）

13.6.1 在r级正交表中安排取$q(<r)$个不同值的变量 …（384）
13.6.2 在r级正交表中安排取$q(>r)$个不同值的变量 …（388）
§13.7 交互作用……（389）
13.7.1 两个变量的交互作用 ……（389）
13.7.2 两个以上变量的交互作用 ……（393）
§13.8 部分正交表的使用……（397）
§13.9 可加函数的拓广……（402）
附录：一些常用正交表 ……（404）
参考文献……（419）

第 1 章　线性代数基本知识

我们利用这一章向读者介绍本书要用到的线性代数方面的一些基本概念和结果，熟悉这部分内容的读者可略过这一章从第 2 章读起。

§1.1　向量、矩阵、行列式

1.1.1　向量的概念

客观世界中，有许许多多事物必须要同时用多个数量来刻画和描述，例如：

一个质点在空间的位置，需要在选定坐标系下用三个坐标值来描述；

空间中的一个力有大小和方向，也需要在选定坐标系下用三个分量值来描述；

一个人的健康水平，需要用反映健康情况的身体各器官的功能指标来衡量；

在地形变测量中，某地两点间十天内的高差变化情况，需要用十天内测量到的一系列数据才能反映出来；

……

在数学上，称一组有序的数

$$(x_1, x_2, \cdots, x_n) \tag{1.1.1}$$

为**向量**。为方便计，我们用希腊字母α，β，γ，…或小写英文正体字母x，y，…（必要时带有下标）表示向量，例如记

$$x=(x_1,\ x_2,\ \cdots,\ x_n).$$

而用小写斜体英文字母（或同时带有下标）表示向量的分量，如上式中的x_i（$i=1,\ 2,\ \cdots,\ n$）表示x的分量。分量的个数称为向量的**维数**。例如上述向量x的维数为n。

1.1.2 向量的运算

对于向量$\alpha=(a_1,\ a_2,\ \cdots,\ a_n)$和$\beta=(b_1,\ b_2,\ \cdots,\ b_n)$，如果$a_i=b_i$（$i=1,\ 2,\ \cdots,\ n$），则称向量$\alpha$与$\beta$**相等**，记$\alpha=\beta$。

若α的每个分量为0，则称α为**零向量**，记为θ。

给出向量的如下运算。

(i)*加法* 设$\alpha=(a_1,\ a_2,\ \cdots,\ a_n)$，$\beta=(b_1,\ b_2,\ \cdots,\ b_n)$，则称向量$(a_1+b_1,\ a_2+b_2,\ \cdots,\ a_n+b_n)$为$\alpha$与$\beta$的和，记之为$\alpha+\beta$。

(ii)*数乘* 设c为一数，$\alpha=(a_1,\ a_2,\ \cdots,\ a_n)$，则称向量$(ca_1,\ ca_2,\ \cdots,\ ca_n)$为数$c$与向量$\alpha$的乘积，简称数乘，记为$c\alpha$。

向量以及向量运算的几何意义可以二维情形为例来说明。

在选定的坐标系下，任意向量

$$\alpha=(x,\ y)$$

可视为联结原点O和点$M(x,\ y)$的一个有向线段$\overrightarrow{OM}$（见图1.1.1）。容易看出，不同的向量有从原点出发的不同有向线段与之对应，反之，从原点出发的不同有向线段有不同的向量与之对应，这样，向量与有向线段间建立了一一对应关系。于是，两向量$\alpha=(x_1,\ y_1)$与$\beta=(x_2,\ y_2)$之和（记为γ）

$$\gamma=\alpha+\beta=(x_1+x_2,\ y_1+y_2)$$

就是以α，β为边的平行四边形的对角线所表示的向量，而数c与$\alpha=(x_1,\ y_1)$的乘积就是有向线段α伸长c倍所表示的向量。特

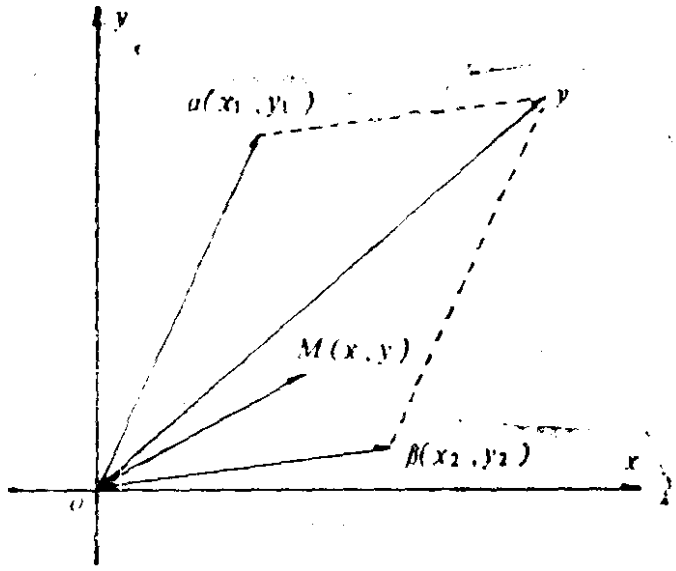

图 1.1.1

别地，当 $c=-1$ 时，$-\alpha=(-1)\cdot\alpha=(-x_1,\ -y_1)$ 表示与α长度相等但指向相反的向量，称为α的**负向量**.

1.1.3 向量组的性质

为了研究向量间的关系，我们引进向量组线性相关性的概念.

对向量α和向量组 $\alpha_1,\ \alpha_2,\ \cdots,\ \alpha_S$，如果有一组数$k_1,\ k_2,\ \cdots,\ k_S$使得

$$\alpha=k_1\alpha_1+k_2\alpha_2+\cdots+k_s\alpha_s, \qquad (1.1.2)$$

则称α可用$\alpha_1,\ \alpha_2,\ \cdots,\ \alpha_S$**线性表出**.

对于向量组$\alpha_1,\ \alpha_2,\ \cdots,\ \alpha_s$，如果存在一组不全为零的数$k_1,\ k_2,\ \cdots,\ k_S$使得

$$\theta=k_1\alpha_1+k_2\alpha_2+\cdots+k_s\alpha_s, \qquad (1.1.3)$$

则称向量组$\alpha_1,\ \alpha_2,\ \cdots,\ \alpha_S$**线性相关**，否则，称它们**线性无关**. 换句话说，如果只有一组全为零的数$k_1,\ k_2,\ \cdots,\ k_s$使得式(1.1.3)成立，则称$\alpha_1,\ \alpha_2,\ \cdots,\ \alpha_s$线性无关.

例如，向量组$\alpha_1=(1,\ 0,\ 1)$，$\alpha_2=(1,\ -1,\ 0)$，$\alpha_3=(0,\ 1,\ 1)$线性相关，因为有

$$1\cdot\alpha_1+(-1)\cdot\alpha_2+(-1)\cdot\alpha_3=\theta.$$

而向量组$\alpha_1=(1,-1,1)$，$\alpha_2=(0,-1,1)$，$\alpha_3=(2,0,1)$线性无关，因为要使得

$$k_1\alpha_1+k_2\alpha_2+k_3\alpha_3=\theta,$$

等价地，要使得齐次线性方程组

$$\begin{cases} k_1 \quad +2k_3=0, \\ -k_1-k_2 \quad =0, \\ k_1+k_2+k_3=0 \end{cases}$$

成立，只能是$k_1=k_2=k_3=0$。

容易证明：

定理1.1.1 向量组$\alpha_1,\cdots,\alpha_s$线性相关的充要条件是其中至少有一个向量能用其余向量线性表示。

向量组的线性相关可作如下几何解释。设向量α与β线性相关，即能表示成$\alpha=k\beta$（或者$\beta=\mu\alpha$），这说明α与β共线。反之，若向量α与β共线，即可表成$\alpha=k\beta$（或者$\beta=\mu\alpha$），由定理1.1.1，说明α与β线性相关。所以，两向量线性相关与否，就是它们共线与否（图1.1.2(*a*)）。三向量α_1，α_2，α_3线性相关时，即存在不全为零的数k_1，k_2，k_3（不妨设$k_1\neq 0$）使得

$$k_1\alpha_1+k_2\alpha_2+k_3\alpha_3=\theta,$$

则

$$\alpha_1=-\frac{k_2}{k_1}\alpha_2-\frac{k_3}{k_1}\alpha_3.$$

可见向量α_1为向量$-\frac{k_2}{k_1}\alpha_2$与$-\frac{k_3}{k_1}\alpha_3$之和，说明向量α_1在由α_2，α_3所决定的平面上，或者说三向量共面。反之，若三向量共面，则其中一个向量必定可用其余两个向量线性表示，由定理1.1.1，这三向量线性相关。于是，三向量线性相关与这三向量共面等价（图1.1.2(*b*)）。

向量组相关性的这一几何解释，还可以推广到更多个向量的情形。

下面再介绍向量组的一个重要概念，关于向量组的**最大线性无**