



武汉大学
百年名典

多元统计分析引论

■ 张尧庭 方开泰 著



根燦葉茂 實亦聲光
山高水長 流風長美



WUHAN UNIVERSITY PRESS

武汉大学出版社

014031664

0212.4
28

武汉大学
百年名典

多元统计 分析引论

张尧庭 方开泰 著

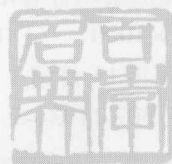


武汉大学出版社
WUHAN UNIVERSITY PRESS



北航

C1720228



0212-4

28

图书在版编目(CIP)数据

多元统计分析引论/张尧庭,方开泰著. —武汉:武汉大学出版社,
2013. 11

武汉大学百年名典

ISBN 978-7-307-11934-5

I. 多… II. ①张… ②方… III. 多元分析—统计分析—研究
IV. O212.4

中国版本图书馆 CIP 数据核字(2013)第 242535 号

责任编辑:顾素萍

责任校对:鄢春梅

版式设计:马 佳

出版发行: **武汉大学出版社** (430072 武昌 珞珈山)

(电子邮件: cbs22@whu.edu.cn 网址: www.wdp.com.cn)

印刷: 湖北恒泰印务有限公司

开本: 720 × 1000 1/16 印张: 32.25 字数: 462 千字 插页: 4

版次: 2013 年 11 月第 1 版 2013 年 11 月第 1 次印刷

ISBN 978-7-307-11934-5 定价: 86.00 元

版权所有, 不得翻印; 凡购我社的图书, 如有质量问题, 请与当地图书销售部门联系调换。

张尧庭（1933—2007年），1933年出生于上海，1951年9月进入清华大学数学系学习，1952年高校院系调整后进入北京大学数学力学系学习。1956年9月获学士学位，留校任北大数学力学系助教，1962年升任讲师。1978年4月至1994年3月先后在武汉大学数学系、统计系和管理学院任教，1980年被破格提升为正教授。曾任武汉大学统计系主任、管理学院院长、概率统计博士生导师，兼任中国统计学会理事、湖北统计学会副理事长、武汉市科协副主席。1994年3月调入上海财经大学，任教授、数量经济学博士生导师，同时兼任中国人民大学、浙江大学等高校的兼职教授。

张尧庭教授从事数理统计的研究和应用，在多元分析理论、部分平衡不完全区组设计、广义相关系数及应用方面，取得研究成果。累计出版学术专著27本，发表学术论文75篇。在著书立说的同时，还十分注重统计理论的推广和应用。

方开泰

1940年生于江苏泰州，1957—1963年就读于北京大学，随后在中国科学院数学所攻读研究生，1967年毕业留所工作。1980年作为访问学者在美国耶鲁大学、斯坦福大学两年。1985—1986年被邀请为瑞士联邦理工大学客座教授，1988年为美国北卡罗尼亚州大学的访问教授。1985年批准为概率统计博士生导师。1984—1992年，任中国科学院应用数学所副所长。1993年至2006年1月，是香港浸会大学数学系的讲座教授、统计研究与咨询中心主任，其间2002—2005年担任数学系系主任。2006年至今任北京师范大学—香港浸会大学联合国际学院(UIC)教授、统计与计算智能研究所所长。

研究领域主要涉及试验设计、多元分析、数据挖掘在统计中的应用，已出版专著22本，发表论文260多篇，是均匀设计创始人之一。曾经担任许多国际和国内学术期刊的副主编，自2010年以来，担任高等教育出版社《高等教育现代统计学系列教材》的主编。获得许多奖励，与王元院士合作的项目“均匀设计理论、方法及其应用”项目获2008 国家自然科学二等奖。1992年和2001年方开泰教授分别获美国数理统计学院和美国统计学会选为院士（Elected Fellow）。

《武汉大学百年名典》出版前言

百年武汉大学,走过的是学术传承、学术发展和学术创新的辉煌路程;世纪珞珈山水,承沐的是学者大师们学术风范、学术精神和学术风格的润泽。在武汉大学发展的不同年代,一批批著名学者和学术大师在这里辛勤耕耘,教书育人,著书立说。他们在学术上精品、上品纷呈,有的在继承传统中开创新论,有的集众家之说而独成一派,也有的学贯中西而独领风骚,还有的因顺应时代发展潮流而开学术学科先河。所有这些,构成了武汉大学百年学府最深厚、最深刻的学术底蕴。

武汉大学历年累积的学术精品、上品,不仅凸现了武汉大学“自强、弘毅、求是、拓新”的学术风格和学术风范,而且也丰富了武汉大学“自强、弘毅、求是、拓新”的学术气派和学术精神;不仅深刻反映了武汉大学有过的人文社会科学和自然科学的辉煌的学术成就,而且也从多方面映现了20世纪中国人文社会科学和自然科学发展的最具代表性的学术成就。高等学府,自当以学者为敬,以学术为尊,以学风为重;自当在尊重不同学术成就中增进学术繁荣,在包容不同学术观点中提升学术品质。为此,我们纵览武汉大学百年学术源流,取其上品,掬其精华,结集出版,是为《武汉大学百年名典》。

“根深叶茂,实大声洪。山高水长,流风甚美。”这是董必武同志1963年11月为武汉大学校庆题写的诗句,长期以来为武汉大学师生传颂。我们以此诗句为《武汉大学百年名典》的封面题词,实是希望武汉大学留存的那些泽被当时、惠及后人的学术精品、上品,能在现时代得到更为广泛的发扬和传承;实是希望《武汉大学百年名典》这一恢宏的出版工程,能为中华优秀传统文化的积累和当代中国学术的繁荣有所建树。

《武汉大学百年名典》编审委员会

出版说明

本书系我社《武汉大学百年名典》之一。本书展现了作者在多元统计分析方面所做的重要工作。为方便读者阅读本书，成书过程中在版式上作了一些修改，如对向量和矩阵用黑体表示，实数域上的 n 维向量空间表示为 $\mathbf{R}^n$ ，除此之外，整体内容基本保持原貌。

武汉大学出版社

2013 年 11 月

序 言

多元统计分析是数理统计学中近二十多年来迅速发展的一个分支。由于电子计算机使用日益广泛，多元分析的方法也很快地应用到各个领域。在国外，从自然科学到社会科学的许多方面，都已证实了多元分析方法是一种很有用的数据处理方法；在我国，多元分析对于地质、气象、水文、国家标准和误差分析等许多方面的研究工作都取得了很大的成绩，引起了广泛的注意。

但是，目前非常缺少系统介绍多元分析的书，现有的书或者过于偏重理论，或者过于偏重单纯地介绍方法。而迫切需要的则是这样一本书，它使得搞数学的人可以从中看到多元分析方法的实际应用，使得搞实际工作的人可以从中看到相应的一些理论。我们正是朝着这个目标来努力的，并希望本书能作为高等学校高年级学生和研究生入门书，也可以作为实际工作者的参考书。我们假定读者已具备一元统计分析的知识。

全书共分九章，第一章系统介绍多元分析中常用的矩阵知识。本章内容大多只阐述结论，而不给出证明。第二章到第五章，介绍多元正态分布以及常用的方差分析、回归分析和判别分析等方法。第六章、第七章采用比较一般的形式来介绍因子分析和线性模型的内容，读者在熟悉第二章到第五章内容的基础上能更好地理解第六、第七两章比较概括抽象的结果。第八章介绍聚类分析的各种典型的方法。第九章专门讨论统计量的分布。

本书收集了我国数学工作者的成果，特别是许宝騄先生在多元分析方面的奠基性的成果。

最后，我们感谢参加多元分析讨论班的同志们，他们在讨论中给

了我们许多帮助，特别要感谢陈希孺同志，他对本书初稿提出了许多宝贵的意见。

由于水平有限，书中肯定有很多缺点和错误，请读者批评指正。

武汉大学 张尧庭
应用数学研究所 方开泰

1979年3月

目 录

第一章 矩阵	1
1.1 线性空间	1
1.2 内积和投影	3
1.3 矩阵的基本性质	7
1.4 分块矩阵的代数运算	15
1.5 特征根及特征向量	20
1.6 对称阵	26
1.7 非负定阵	31
1.8 广义逆	36
1.9 计算方法	46
1.10 矩阵微商	56
1.11 矩阵的标准型	59
1.12 矩阵内积空间	62
第二章 多元正态分布	67
2.1 定义	67
2.2 正态分布的矩	71
2.3 条件分布和独立性	74
2.4 多元正态分布的参数估计	80
2.5 μ 和 V 的极大似然估计的性质	87
2.6 多维正态分布的特征	100

2.7 多维正态分布函数的计算	104
2.8 例	115
第三章 样本分布的性质和均值与协差阵的检验	120
3.1 二次型分布	120
3.2 维希特(Wishart)分布	136
3.3 与样本协差阵有关的统计量, T^2 和 Λ 统计量	141
3.4 均值的检验	150
3.5 T^2 统计量的优良性	158
3.6 多母体均值的检验	164
3.7 协方差不等时均值的检验	173
3.8 协差阵的检验	175
3.9 独立性检验	186
第四章 判别分析	193
4.1 距离判别	193
4.2 贝叶斯(Bayes)判别	207
4.3 费歇(Fisher)的判别准则	214
4.4 误判概率	222
4.5 附加信息检验	229
4.6 逐步判别	235
4.7 序贯判别	247
第五章 回归分析	252
5.1 问题及模型	252
5.2 最小二乘估计	257
5.3 假设检验	265
5.4 逐步回归	272

5.5 双重筛选逐步回归	285
5.6 回归分析与判别分析的关系	293
第六章 相关	297
6.1 投影	297
6.2 典型相关变量	301
6.3 广义相关系数	310
6.4 主成分分析及主分量分析	316
6.5 因子分析	322
第七章 线性模型	333
7.1 模型	333
7.2 估值	335
7.3 广义线性模型	341
7.4 递推公式	351
7.5 正态线性模型的假设检验	359
7.6 试验设计	370
第八章 聚类分析	384
8.1 相似系数和距离	384
8.2 系统聚类法	392
8.3 系统聚类法的性质	405
8.4 动态聚类法	416
8.5 分解法	431
8.6 有序样品的聚类与预报	436
第九章 统计量的分布	449
9.1 预备知识	449

9.2	$I_m(f r_1, \dots, r_m)$	453
9.3	一元非中心分布	458
9.4	Wishart 分布	461
9.5	广义方差的分布	469
9.6	非中心 T^2 分布	473
9.7	样本相关系数的分布	477
9.8	$S_1 S_2^{-1}$ 特征根的联合分布	483
9.9	结束语	497

参考文献	502
------------	-----

第一章 矩 阵

在这章中我们把矩阵和向量空间中有关多元分析的一些结果介绍给读者,凡是一般教材中有的材料,我们只是罗列一下,大多不给出证明.熟悉这些内容的人可以从第二章看起,想看第一章的读者最好把正文中未给证明的结果试着证明一下以作为练习,这将为以后各章的阅读带来很多方便.

1.1 线 性 空 间

今后我们限定在实数域 $\mathbf{R}$ 上讨论. 用 $\mathbf{R}^n$ 表示实数域 $\mathbf{R}$ 上的全部 n 维向量, $\mathbf{R}^n$ 中的一个向量(有时称为元素) a 有 n 个坐标, 写成

$$a = \begin{pmatrix} a_1 \\ \vdots \\ a_n \end{pmatrix} \text{ 或 } a = (a_1, \dots, a_n)'.$$

1.1.1 向量的运算

对 $\mathbf{R}^n$ 中的向量定义了两种运算:

(i) 加法 设 $a = (a_1, \dots, a_n)'$, $b = (b_1, \dots, b_n)'$, 则

$$a + b \triangleq \textcircled{1} (a_1 + b_1, \dots, a_n + b_n)'.$$

(ii) 数乘 设 c 是一个数, $a = (a_1, \dots, a_n)'$, 则

$$ca \triangleq (ca_1, \dots, ca_n)'.$$

为了方便,我们用 $\mathbf{0}$ 表示坐标全为 0 的向量. 容易证明,对上述定

① 表示该符号左边的内容由右边的符号表示,也可以理解为定义.

义的加法和数乘, 下列关系成立:

(A) 当 a, b, c 均 $\in \mathbf{R}^n$ 时, 有

$$a + b = b + a, \quad (a + b) + c = a + (b + c),$$

$$a + 0 = a, \quad a + (-a) = 0.$$

(B) 当 a, b 均 $\in \mathbf{R}^n$, c, c_1, c_2 均为实数时, 有

$$c(a + b) = ca + cb, \quad (c_1 + c_2)a = c_1a + c_2a,$$

$$c_1(c_2a) = c_1c_2a, \quad 1 \cdot a = a.$$

我们称 $\mathbf{R}^n$ 是实数域 $\mathbf{R}$ 上的线性空间或向量空间. 如果 $\mathbf{R}^n$ 中的子集 L , 对加法和数乘这两种运算是封闭的 (即运算的结果仍在 L 中), 且 L 不是空集, 则称 L 是 $\mathbf{R}^n$ 中的一个子空间. 例如

$$L = \{(a, \underbrace{0, \dots, 0}_{n-1 \text{ 个}})' : a \in \mathbf{R}\}$$

就是 $\mathbf{R}^n$ 的一个子空间.

1.1.2 线性相关

线性空间最重要的概念就是线性相关与线性无关.

如有不全为 0 的一组数 $c_1, \dots, c_k$ 使 $c_1a_1 + \dots + c_ka_k = 0$, 则称向量 $a_1, \dots, a_k$ 是线性相关的; 否则, 就称 $a_1, \dots, a_k$ 是线性无关的. 从这个定义可以看出:

(i) 任何含有 0 向量的向量集总是线性相关的;

(ii) $a_1, \dots, a_k$ 线性无关的充要条件是: 如果一组数 $c_1, \dots, c_k$ 使 $c_1a_1 + \dots + c_ka_k = 0$, 则 $c_1 = \dots = c_k = 0$;

(iii) 设 $a_1, \dots, a_k$ 是非零向量, 它们线性相关的充要条件是: 存在 i 使 $a_i = b_1a_1 + b_2a_2 + \dots + b_{i-1}a_{i-1} + b_{i+1}a_{i+1} + \dots + b_ka_k$, 其中 $b_1, \dots, b_k$ 是一组实数. 也即存在 i 使 a_i 是其他向量 $a_1, \dots, a_{i-1}, a_{i+1}, \dots, a_k$ 的线性组合.

给定了 $\mathbf{R}^n$ 中某些向量 $a_1, \dots, a_k$, 考虑由这些向量所有可能的线性组合 $\sum_{i=1}^k c_i a_i$ 组成的集合:

$$\mathcal{L}(a_1, \dots, a_k) = \left\{ \sum_{i=1}^k c_i a_i : c_1, \dots, c_k \text{ 均为实数} \right\},$$

它显然对加法和数乘这两种运算是封闭的, 它是一个子空间, 称它是由向量 $a_1, \dots, a_k$ 生成的子空间.

1.1.3 基

设 $\mathcal{L}$ 是 $\mathbf{R}^n$ 中的一个子空间, 如果存在 $a_1, \dots, a_k$ 使 $\mathcal{L} = \mathcal{L}(a_1, \dots, a_k)$, 且 $a_1, \dots, a_k$ 线性无关, 则称 $a_1, \dots, a_k$ 是 $\mathcal{L}$ 的一组基.

可以证明子空间 $\mathcal{L}$ 中如有两组基, 那么这两组基中向量的个数一定相同, 因此我们把子空间 $\mathcal{L}$ 中一组基所含的向量的个数称为 $\mathcal{L}$ 的维数. 例如

$$e_i = (\underbrace{0, \dots, 0}_{i-1 \text{ 个}}, 1, \underbrace{0, \dots, 0}_{n-i \text{ 个}})', \quad i = 1, 2, \dots, n$$

显然是线性无关的, 并且 $\mathbf{R}^n = \mathcal{L}(e_1, \dots, e_n)$, 因此线性空间 $\mathbf{R}^n$ 的维数是 n . 有时就称 $\mathbf{R}^n$ 为 n 维线性空间. 很显然, $\mathcal{L}(e_1), \dots, \mathcal{L}(e_n)$ 都是一维的子空间; $\mathcal{L}(e_1, e_2), \dots, \mathcal{L}(e_{n-1}, e_n)$ 都是二维的子空间……

从基的定义即可可知, 如果 $a_1, \dots, a_k$ 是子空间 $\mathcal{L}$ 的一组基, 那么 $\mathcal{L}$ 中任一向量 a 都可被 $a_1, \dots, a_k$ 的线性组合来表示, 而且这种表示法是唯一的.

1.1.4 直接和

设 $\mathcal{L}$ 是一个子空间, 如果有 k 个子空间 $\mathcal{L}_1, \dots, \mathcal{L}_k$ 使得对每一 $a \in \mathcal{L}$, 能唯一地表示为 $a_1 + \dots + a_k$, 其中 $a_i \in \mathcal{L}_i, i = 1, 2, \dots, k$, 则称 $\mathcal{L}$ 是 $\mathcal{L}_1, \dots, \mathcal{L}_k$ 的直接和, 记为 $\mathcal{L} = \mathcal{L}_1 + \dots + \mathcal{L}_k$. 沿用 1.1.3 中 e_i 的定义, 可以看出 $\mathbf{R}^n = \mathcal{L}(e_1) + \dots + \mathcal{L}(e_n)$.

1.2 内积和投影

$\mathbf{R}^n$ 中任给两个向量 $a = (a_1, \dots, a_n)'$, $b = (b_1, \dots, b_n)'$, 定义 a, b 的内积 $(a, b) \triangleq \sum_{i=1}^n a_i b_i$. 易见内积 (a, b) 满足下列性质:

$$(i) \quad (a, b) = (b, a);$$

(ii) $(a, a) \geq 0; (a, a) = 0 \Leftrightarrow a = 0;$

(iii) $(ca, b) = (a, cb) = c(a, b)$ 对一切 $c \in \mathbf{R}$ 成立;

(iv) $(a, h+g) = (a, h) + (a, g), (h+g, b) = (h, b) + (g, b).$

我们把 (a, a) 的算术平方根称为 a 的长度, 记作 $\|a\|$. 容易验证

$$|(a, b)|^2 \leq \|a\|^2 \|b\|^2 \quad (\text{Schwarz 不等式}), \quad (2.1)$$

$$\|a+b\| \leq \|a\| + \|b\| \quad (\text{三角不等式}). \quad (2.2)$$

1.2.1 标准正交基

如果 $\mathbf{R}^n$ 中的子空间 $\mathcal{L}$ 的基 $a_1, \dots, a_k$ 具有性质:

$$(a_i, a_i) = 1, \quad i = 1, 2, \dots, k,$$

$$(a_i, a_j) = 0, \quad i \neq j, \quad i, j = 1, 2, \dots, k,$$

则称 $a_1, \dots, a_k$ 是 $\mathcal{L}$ 的一组标准正交基, 因为我们把 $\|a\| = 1$ 的向量称为标准化的向量, 有时也称为单位向量(指它的长度是 1, 以它作单位). 当 $(a, b) = 0$ 时, 我们就称 a 与 b 正交, $(a_i, a_j) = 0$ 表示 a_i 与 a_j 正交. 引入克劳涅克尔的 δ 符号:

$$\delta_{ij} = \begin{cases} 1, & \text{当 } i = j, \\ 0, & \text{当 } i \neq j, \end{cases}$$

则 $a_1, \dots, a_k$ 是 $\mathcal{L}$ 的一组标准正交基 $\Leftrightarrow a_1, \dots, a_k$ 均属于 $\mathcal{L}$, 且

$$(a_i, a_j) = \delta_{ij}, \quad i, j = 1, 2, \dots, k.$$

1.2.2 投影

在 $\mathbf{R}^n$ 中, 给定一个向量 a 及子空间 $\mathcal{L}$, 就可考虑 a 在 $\mathcal{L}$ 中的投影. 如果在 $\mathcal{L}$ 中存在 b 使 $\|a-b\| = \inf_{x \in \mathcal{L}} \|a-x\|$, 则称 b 是 a 在 $\mathcal{L}$ 中的投影. 可以证明投影是存在而且唯一的. 下面先证明一条关于投影的重要性质:

b 是 a 在 $\mathcal{L}$ 中的投影 $\Leftrightarrow (a-b, x) = 0$ 对一切 $x \in \mathcal{L}$ 成立.

证 “ $\Rightarrow$ ”. 若有 $x \in \mathcal{L}$ 使 $(a-b, x) \neq 0$, 由于对一切 λ 有

$$\begin{aligned} (a-b, a-b) &\leq (a-b+\lambda x, a-b+\lambda x) \\ &= (a-b, a-b) + 2\lambda(a-b, x) + \lambda^2(x, x), \end{aligned}$$