

□ 应用统计学丛书

Large Dimensional Statistical Analysis

大维统计分析

白志东 郑术蓉 姜丹丹



高等教育出版社
HIGHER EDUCATION PRESS

□ 应用统计学丛书

Large Dimensional Statistical Analysis

大维统计分析

白志东 郑术蓉 姜丹丹

DAWEI TONGJI FENXI



图书在版编目(CIP)数据

大维统计分析/白志东,郑术蓉,姜丹丹编著. --
北京:高等教育出版社,2012.5
ISBN 978-7-04-034830-9

I. ①大… II. ①白… ②郑… ③姜… III. ①统计
分析 IV. ①C812

中国版本图书馆 CIP 数据核字(2012)第 063747 号

策划编辑 王丽萍
责任校对 金 辉

责任编辑 李华英
责任印制 朱学忠

封面设计 王凌波

版式设计 于 婕

出版发行 高等教育出版社
社 址 北京市西城区德外大街 4 号
邮政编码 100120
印 刷 保定市中画美凯印刷有限公司
开 本 787mm × 1092mm 1/16
印 张 34
字 数 700 千字
购书热线 010-58581118

咨询电话 400-810-0598
网 址 <http://www.hep.edu.cn>
<http://www.hep.com.cn>
网上订购 <http://www.landaco.com>
<http://www.landaco.com.cn>
版 次 2012 年 5 月第 1 版
印 次 2012 年 5 月第 1 次印刷
定 价 79.00 元

本书如有缺页、倒页、脱页等质量问题,请到所购图书销售部门联系调换
版权所有 侵权必究
物 料 号 34830-00

前言

在一元统计分析中, 对每个样本个体只观测一个指标值。如果个体包含多重特性时, 一元统计很难做出比较精确的推断。例如研究小孩身体发育, 如果仅观测样本中每个小孩的身高, 则仅根据身高信息, 很难对小孩的身体发育情况得到精确的结论。因此在实际问题中通常要观测每个样本个体的多个指标值, 进行多元统计分析。如研究小孩的身体发育, 我们需要观测样本中每个小孩的身高、体重、腰围、胸围等多个指标值。多个指标蕴含的信息量大, 有助于做出精确的统计推断。因此, 我们需要将一元统计中发展起来的统计推断理论推广到多元场合。

另外, 需要指出的是, 在 20 世纪 40 年代, 即多元分析理论建立和发展的时期, 因为实际问题中收集的样本量很小, 与一元统计相似, 多元分析也是在多元正态分布假定下发展起来的。这是因为在正态分布假定下很多理论都可以得到精确的结果, 如统计量的精确分布等。为保持这一历史事实, 也为了小样本统计的需要, 我们对于多元分析的基本理论, 也在多元正态的假定下来介绍。但是, 随着科学技术的发展, 实际问题中能收集到大量的样本, 例如, 在无线电电子学中, 1 秒钟可以收集 500 个样本; 在股票市场中, 1 秒钟也能收集数百个数值。以前因为样本量小不能解决的问题就可以在大样本下得到渐近的结果。特别是近数十年来, 随着计算机技术的飞速发展和广泛应用, 人们得以搜集、储存和处理大维巨量数据。例如, 金融市场上的资料的维数可以高达数千或数万; 手机通讯资料的维数可达百万乃至千万计; 生物学中关于 DNA 的资料维数更是大到难以想象。高速计算技术的发展使得人们可以对大维巨量数据进行计算和统计推断。它给我们带来了巨大的好处, 同时, 也带来了巨大的挑战: 经典的极限理论是在维数固定, 而样本容量趋于无穷的假定下得到的。首先, 许多经典的多元统计方法应用于大维资料时, 它们或者根本不可以应用, 如点估计的偏差很大, 第一类错误可能会趋于 1 等, 这在点估计和假设检验中是根本不可能被接受的; 或者这些经典方法即使可以应用, 其功效也会很低。其次, 当维数不是很大时, 人们早就创立了各种降维方法, 例如变量选择、主成分分析、因子分析等。

但是这些方法根本不能适用于维数非常大的情况。作为本书的特点,除基本理论部分外,对于各种多元统计方法,我们尽可能引进一些传统统计分析方法的修正,使之更适用于多元统计分析。由于许多修正方法依赖于大维随机矩阵的谱分析理论,所以我们在附录中简要介绍该理论中的一些主要结果。

本书面向统计学专业及与统计学相关专业的本科生、硕士生、博士生和科研人员。对于本科生来说,我们建议讲授如下章节: 2.1 ~ 2.4, 3.1 ~ 3.2, 5.1 ~ 5.3, 6.1 ~ 6.5, 11.1 ~ 11.2, 12.1 ~ 12.3, 14.1 ~ 14.7。对于硕士生建议增加章节: 2.5, 3.3 ~ 3.5, 4.1 ~ 4.4, 5.4 ~ 5.6, 6.6 ~ 6.9, 7.1 ~ 7.9, 8.1 ~ 8.10, 9.1 ~ 9.6, 10.1 ~ 10.8, 11.3 ~ 11.4, 12.4 ~ 12.8, 13 章和 14 章。博士生除讲授上述内容外,建议自学所有章节,特别注意各章中的大维统计分析部分。

在此真诚地感谢东北师范大学张宝学教授、胡果荣博士、宋海燕博士、胡江博士在本书完成过程中所做的工作。

作 者

2011 年 9 月 7 日

目 录

记号与约定	i
第一章 引论	1
1.1 多元统计分析	1
1.2 多元正态分布	2
1.3 大维统计分析	2
1.4 大维随机矩阵的谱分析	3
第二章 多元正态分布	5
2.1 引论	5
2.2 多元正态分布的定义	6
2.2.1 标准 p 元正态分布	6
2.2.2 一般 p 元正态分布的定义	6
2.2.3 多元正态分布的特征函数、矩母函数和密度	7
2.2.4 二元正态分布的密度公式	7
2.2.5 多元正态分布的相关系数和相关系数矩阵	8
2.3 多元正态分布的性质	8
2.3.1 多元正态分布族在线性变换下的性质	8
2.3.2 多元正态分布密度的等高线	9
2.3.3 正态随机变量线性组合的分布、独立性及边缘分布	10
2.4 条件分布和多重相关系数	12
2.4.1 条件分布	12
2.4.2 多重相关系数	14
2.4.3 偏相关的一些公式	16

2.5	多元正态分布的二次型及其独立性	17
2.5.1	二次型的矩和矩母函数	18
2.5.2	线性型、二次型相互独立的充要条件	19
2.6	复多元正态分布的定义及基本性质	22
2.6.1	复数运算的补充代数知识	22
2.6.2	复多元正态分布的定义和性质	23
2.6.3	极大似然估计	24
2.7	练习题	25
第三章	均值向量与协方差矩阵的估计	31
3.1	引论	31
3.2	均值向量和协方差矩阵的极大似然估计	32
3.3	协方差矩阵已知时, 样本均值向量的分布及统计推断	36
3.3.1	分布理论	36
3.3.2	协方差矩阵已知时, 关于均值向量的检验和置信域	37
3.3.3	非中心化 χ^2 分布与功效函数	40
3.4	均值向量估计的性质	41
3.4.1	极大似然估计的性质	41
3.4.2	Bayes 与 minmax 估计	43
3.5	均值向量的改进估计	45
3.5.1	引论	45
3.5.2	James-Stein 估计量	45
3.5.3	协方差矩阵已知时任意二次损失函数下的估计	48
3.6	练习题	51
第四章	样本相关系数的分布与应用	57
4.1	引论	57
4.2	二元样本相关系数	58
4.2.1	总体相关系数为零时样本相关系数的分布及不相关的假设检验	58
4.2.2	总体相关系数非零时样本相关系数的分布, 假设检验和置信区间	61
4.2.3	样本相关系数与 Fisher z 的渐近分布	66
4.3	偏相关系数, 条件分布	69
4.3.1	偏相关系数的估计	69
4.3.2	样本偏相关系数的分布	73

4.3.3	偏相关系数的假设检验和置信区间	75
4.4	多重相关系数	75
4.4.1	多重相关系数的估计	75
4.4.2	总体多重相关系数为零时样本多重相关系数的分布	79
4.4.3	总体多重相关系数非零时样本多重相关系数的分布	82
4.4.4	多重相关检验的某些最优性	85
4.5	多重相关系数的大维表现	85
4.5.1	大维情形下样本多重相关系数的极限	86
4.5.2	大维情形下多重相关系数的中心极限定理	87
4.5.3	大维情形下关于多重相关系数的假设检验与置信区间	89
4.6	练习题	89
第五章	T^2 统计量	96
5.1	引论	96
5.2	T^2 统计量的推导及其分布	97
5.2.1	T^2 统计量作为似然比准则函数的推导	97
5.2.2	T^2 统计量的分布	98
5.3	T^2 统计量的应用	101
5.3.1	检验单总体均值向量等于某个给定的向量	101
5.3.2	均值向量的置信域	101
5.3.3	均值向量的所有线性组合的一致置信区间	102
5.3.4	两样本问题	103
5.3.5	多个样本的问题	103
5.3.6	关于对称性的一个问题	104
5.3.7	改进的均值向量估计	105
5.4	T^2 在备择假设下的分布及势函数	106
5.5	协方差矩阵不等时的两样本问题	108
5.6	T^2 检验的一些最优性质	111
5.6.1	最优不变检验	111
5.6.2	可容许检验	113
5.7	大维情形下的均值检验问题	117
5.7.1	Dempster 的非精确检验	118
5.7.2	白 - Saranadasa 的渐近正态检验	121
5.7.3	陈 - 秦改进的检验	124
5.7.4	模拟结果和评论	129
5.8	练习题	131

第六章 判别分析	136
6.1 判别问题	136
6.2 判别的准则	137
6.2.1 初步考虑	137
6.2.2 两个总体的情形	137
6.3 概率分布已知的两个总体的判别方法	139
6.3.1 先验分布已知的情形	139
6.3.2 先验概率未知的情形	140
6.4 两个已知多元正态分布的判别	141
6.5 参数未知时两个正态总体的判别	144
6.5.1 判别准则	144
6.5.2 判别准则的分布	145
6.5.3 判别准则的渐近分布	146
6.5.4 判别准则的另外一种推导	147
6.5.5 似然比准则	147
6.5.6 不变性	149
6.6 错判概率	150
6.6.1 准则 W 的错判概率的渐近展开	150
6.6.2 准则 Z 的错判概率的渐近展开	153
6.7 多个总体的判别	154
6.8 多个多元正态分布的判别	157
6.8.1 基本理论	157
6.8.2 一个例子	159
6.9 两个已知的具有不同协方差矩阵的多元正态总体的判别	161
6.9.1 似然方法	161
6.9.2 线性方法	162
6.10 大维判别分析	166
6.10.1 A- 准则与 D- 准则	166
6.10.2 两个正态总体时 D- 准则的错判概率	167
6.10.3 两个正态总体时 A- 准则的错判概率	169
6.10.4 A- 准则与 D- 准则的比较与评论	170
6.11 练习题	171
第七章 样本协方差矩阵的分布与广义方差	174
7.1 引论	174
7.2 Wishart 分布	175

7.3	Wishart 分布的性质	178
7.3.1	Wishart 分布的特征函数	178
7.3.2	Wishart 矩阵的和	178
7.3.3	Wishart 矩阵的线性变换	179
7.3.4	Wishart 分布的边缘分布	179
7.3.5	条件分布	180
7.4	Cochran 定理	180
7.5	广义方差	182
7.5.1	广义方差的定义	182
7.5.2	样本广义方差的分布	185
7.5.3	样本广义方差的渐近分布	187
7.6	当总体协方差矩阵是对角矩阵时全体相关系数的分布	187
7.7	逆 Wishart 分布和协方差矩阵的 Bayes 估计	189
7.7.1	逆 Wishart 分布	189
7.7.2	协方差矩阵的 Bayes 估计	189
7.8	协方差矩阵的改良估计	192
7.9	非中心化 Wishart 分布	196
7.10	大维架构下样本广义方差的性质与统计推断	200
7.10.1	大维架构下样本广义方差的极限与中心极限定理	200
7.10.2	大维架构下关于广义方差的假设检验和区间估计	202
7.11	练习题	204
第八章	一般线性假设的检验及方差分析	209
8.1	引论	209
8.2	多元线性回归的参数估计	210
8.2.1	极大似然估计, 最小二乘估计	210
8.2.2	β 和 $\hat{\Sigma}$ 的分布	212
8.3	关于回归系数线性假设的似然比检验准则	215
8.3.1	似然比准则	215
8.3.2	几何解释	216
8.3.3	标准型	218
8.4	原假设下似然比准则的分布	219
8.4.1	分布的刻画	219
8.4.2	准则的矩	222
8.4.3	一些特殊情形下的分布	224
8.4.4	似然比方法	226

8.4.5 逐步检验方法	227
8.5 似然比准则分布的渐近展开	228
8.5.1 渐近展开的一般理论	228
8.5.2 似然比准则的渐近分布	232
8.5.3 正态逼近	234
8.5.4 F 逼近	235
8.6 线性假设检验的其他准则	236
8.6.1 相对特征根的函数	236
8.6.2 Lawley-Hotelling 的迹准则	237
8.6.3 Bartlett-Nanda-Pillai 的迹准则	239
8.6.4 Roy 的最大根准则	240
8.6.5 功效的比较	241
8.7 置信区间与回归系数矩阵的假设检验	243
8.7.1 假设检验	243
8.7.2 基于 U 的置信区间	244
8.7.3 基于 Lawley-Hotelling 迹的联立置信区间	244
8.7.4 基于 Roy 最大根准则的联立置信区间	245
8.8 对具有相同协方差矩阵的多个正态分布的均值相等的检验	246
8.9 多元方差分析	249
8.10 一些检验的最优性质	254
8.10.1 不变检验的可容许性	254
8.10.2 无偏检验和功效函数的单调性	263
8.11 大维回归分析	267
8.11.1 似然比准则的渐近分布	267
8.11.2 拟似然比准则的渐近分布的稳健性	269
8.11.3 基于最小二乘法的非精确检验	270
8.11.4 模拟比较	271
8.12 练习题	273
第九章 分组变量的独立性检验	280
9.1 引言	280
9.2 分组变量独立性检验的似然比准则	280
9.3 原假设为真时似然比准则的分布	284
9.3.1 分布的刻画	284
9.3.2 准则函数的矩	285
9.3.3 某些特例下的分布	286

9.3.4	似然比准则的分布的渐近展开	286
9.4	其他检验方法	288
9.4.1	其他准则	288
9.4.2	按分块逐步下降检验	289
9.4.3	按照分量的逐步下降检验方法	289
9.4.4	一个例子	291
9.5	变量分为两个集合的情况	292
9.6	似然比检验的基本性质	295
9.6.1	容许性	295
9.6.2	势函数的单调性	296
9.7	大变量组独立性的检验	298
9.7.1	两组变量独立性检验在大维架构下的近似分布	298
9.7.2	多组变量独立性检验在大维架构下的近似分布	300
9.7.3	当变量个数接近于自由度时两大组变量的独立性检验	301
9.7.4	模拟检验	302
9.8	练习题	306
第十章	均值和方差齐性的检验	308
10.1	引论	308
10.2	检验多个协方差矩阵相等的准则	308
10.3	检验多个正态总体同分布	311
10.3.1	检验的准则函数	311
10.3.2	准则函数的分布	312
10.3.3	准则函数的矩	315
10.3.4	逐步检验	317
10.3.5	准则分布的渐近展开	318
10.4	两个总体的情况	321
10.4.1	不变检验	321
10.4.2	方差分量模型	323
10.5	检验协方差矩阵与某个给定矩阵成比例及球形检验	325
10.5.1	假设	325
10.5.2	准则函数	325
10.5.3	准则函数的分布和矩	327
10.5.4	分布的渐近展开	328
10.5.5	不变检验	328
10.5.6	置信域	329

10.6	检验协方差矩阵等于某个给定矩阵	330
10.6.1	准则函数	330
10.6.2	修正似然比准则函数的分布和矩	331
10.6.3	不变检验	333
10.6.4	二次型的置信界	333
10.7	检验均值向量和协方差矩阵分别等于给定的向量和矩阵	334
10.8	可容许检验	336
10.9	均值向量协方差矩阵齐性的大维分析	339
10.9.1	检验一个总体协方差矩阵等于一个给定的矩阵的 似然比检验的修正	339
10.9.2	检验两个总体协方差矩阵相等的似然比检验的修正	340
10.9.3	检验多个总体协方差矩阵相等的似然比检验的修正	341
10.9.4	检验多个正态总体分布相等的似然比检验的修正	343
10.9.5	当维容比靠近 1 时检验多个正态总体分布相等的 修正似然比检验	345
10.10	练习题	347
第十一章	主成分分析	352
11.1	引论	352
11.2	总体主成分的定义及性质	353
11.2.1	总体主成分的定义	353
11.2.2	主成分及其方差的极大似然估计	355
11.2.3	主成分的极大似然估计的计算	356
11.2.4	一个例子	358
11.3	统计推断	360
11.3.1	渐近分布	360
11.3.2	特征向量的置信域	361
11.3.3	特征根的精确置信界	362
11.4	关于协方差矩阵的特征根的假设检验	363
11.4.1	关于若干个最小特征根和的假设检验	363
11.4.2	关于最小特征根的和相对于所有特征根的和的假设检验	364
11.4.3	关于最小特征根相等的假设检验	364
11.5	大维主成分分析	366
11.5.1	离群特征根的极限	367
11.5.2	离群特征向量的极限	369
11.5.3	离群特征根的中心极限定理	371

11.5.4	本质离群特征根的估计和统计推断	377
11.5.5	待解决的问题	379
11.6	练习题	379
第十二章	典则相关系数与典则变量	382
12.1	引论	382
12.2	总体典则相关系数与典则变量	383
12.3	典则相关系数与典则变量的估计	388
12.3.1	极大似然估计	388
12.3.2	计算方法	390
12.4	统计推断	392
12.4.1	独立性检验和秩检验	392
12.4.2	典则相关系数的渐近分布	393
12.5	例	394
12.6	与共线性相关的特征数值	395
12.6.1	回归矩阵的典则分析	395
12.6.2	估计	398
12.6.3	典则变量间的关系	399
12.6.4	回归系数矩阵秩的检验	399
12.6.5	线性泛函关系	399
12.7	降秩回归	400
12.8	联立方程模型	403
12.8.1	模型	403
12.8.2	特定零点的确认	404
12.8.3	简化形式的估计	405
12.8.4	方程系数的估计	405
12.8.5	与线性泛函的联系	407
12.8.6	$T \rightarrow \infty$ 时的渐近理论	408
12.9	大维架构下的典则相关分析	410
12.9.1	关于特征根与特征向量的渐近分布的一般理论	410
12.9.2	变量组一大一小时的典则相关分析	415
12.9.3	未解决的问题	420
12.10	练习题	421
第十三章	特征根与特征向量的分布	423
13.1	引论	423

13.2 两个 Wishart 矩阵的情形	423
13.2.1 矩阵分解	423
13.2.2 Jacobi 行列式	426
13.2.3 矩阵 $\mathbf{E}$ 和特征根 $\mathbf{F}$ 的联合分布	428
13.2.4 $\mathbf{A}$ 奇异时的分布	429
13.3 非奇异的 Wishart 矩阵	430
13.4 典则相关系数	435
13.5 Wishart 矩阵与 F 矩阵的特征根与特征向量的渐近分布	436
13.5.1 总体特征根全不相同的情形	436
13.5.2 总体协方差矩阵有复重特征根的情形	438
13.6 两个 Wishart 矩阵情形下的渐近分布	438
13.6.1 非随机情形下相对特征根与特征向量的极限	438
13.6.2 一般场合下两个 Wishart 矩阵相对特征根与特征向量的 极限分布	440
13.7 典则相关系数的渐近分布	443
13.7.1 两个变量集都是随机的情形	443
13.7.2 一个变量集随机而另一个变量集非随机的情形	445
13.7.3 降秩回归估计	447
13.8 练习题	448
第十四章 因子分析	450
14.1 引论	450
14.2 因子分析模型	451
14.2.1 因子分析模型的定义	451
14.2.2 可识别性	453
14.2.3 测量的单位	455
14.3 随机正交因子的极大似然估计	456
14.3.1 极大似然估计	456
14.3.2 模型拟合假设检验	459
14.3.3 估计的渐近分布	461
14.3.4 最小距离方法	461
14.3.5 与主成分分析的关系	462
14.3.6 重心方法	463
14.4 固定因子的估计	464
14.5 因子的解释与变换	464
14.5.1 解释	464

14.5.2	变换	464
14.5.3	正交因子和斜交因子	466
14.6	以限定零元素为可识别性条件的参数估计	467
14.7	因子得分估计	467
14.8	练习题	468
附录 A	矩阵知识	470
A.1	方阵的行列式及其性质	470
A.1.1	方阵的行列式的定义	470
A.1.2	方阵的行列式的性质	471
A.2	矩阵的特征根与特征向量	471
A.2.1	特征根与特征向量的定义	471
A.2.2	实对称矩阵	472
A.2.3	谱分解定理	473
A.2.4	奇异值分解定理	474
A.2.5	Chosky 分解定理	475
A.2.6	相对特征根与同时对角化问题	476
A.3	分块矩阵与向量	477
A.3.1	可逆矩阵的分块求逆	477
A.3.2	分块向量的二次型	477
A.3.3	对称矩阵相合下的标准型	478
A.4	矩阵拉长向量和矩阵的 Kronecker 乘积	479
A.5	关于向量或矩阵的导数	480
A.6	广义逆与投影矩阵	481
A.6.1	广义逆	481
A.6.2	投影矩阵	482
A.7	对称矩阵在相合变换下的 Jacobi 行列式	483
A.8	一般矩阵在仿射变换下的 Jacobi 行列式	485
附录 B	大维随机矩阵谱分析知识	486
B.1	经验谱分布与极限谱分布	486
B.2	随机矩阵极端特征根的极限	492
B.2.1	样本协方差矩阵最大最小特征根的极限	492
B.2.2	谱分离定理	493

B.3 线性谱统计量的中心极限定理	495
B.3.1 样本协方差矩阵的线性谱统计量的中心极限定理	495
B.3.2 F 矩阵的线性谱统计量的中心极限定理	496
B.3.3 定理 B.3.2 的应用	497
参考文献	500
术语索引	517
人名索引	521