判别分析

P·A·拉亨布鲁克 著

李从珠 译

群众出版社

1988年·北京

判别分析

P·A·拉亨布鲁克 著

李从珠 译

内 容 简 介

本书主要介绍判别分析的方法和应用。

主要内容包括：判别分析的基本思想；判别函数的评价；线性判别函数的稳健性；非正态和非参数方法；多重组问题以及其他判别分析方法。书中列举许多判别分析的实例，每章后都有习题。

本书可供从事刑事技术、生物、医学、气象、地质、心理学、考古学、经济学、教育学等方面的应用研究工作者和大专院校统计专业的师生参考。

判 别 分 析

P·A·拉亨布鲁克著　　李从珠译

*

群众出版社出版
新华书店北京发行所发行　各地新华书店经售
北京大郊亭印刷厂印刷

*

开本：787×1092毫米　1/32　印张：6.625 字数：140千字
1988年8月第1版　　1988年8月第1次印刷
印数：0001—6000　　定价：2.55元
书号：ISBN 7—5014—0312—O/D·194

译 者 序

近年来判别分析在我国很多领域中得到了广泛应用，许多实际工作者希望有一本介绍判别分析的好书。译者在运用判别分析方法处理刑事技术问题时发现著名统计学家Pefer A. Lachenbruch的《判别分析》是本比较好的书。该书不仅较详细地介绍了判别分析方法，同时指出了该方法的适用范围和注意事项，并引用了大量范围广泛的实例。故将此书推荐给国内读者。

本书写的深入浅出，仅要求读者具有初等概率的基础知识。某些数学推导需要矩阵代数，但对实际工作者来说这并不是必须的。书末附有大量的参考文献，这对想深入了解判别分析的读者是非常有益的。

在翻译过程中，对原书的错误及不妥之处作了订正。但因水平有限，难免仍有缺点和错误，欢迎读者批评指正。

译者感谢王柱、杨振海二位学友，尤其是北京大学卢崇飞老师在校阅译稿中提出的许多宝贵意见。

译者

一九八五年元月

序　言

这是本应用判别分析及其有关问题的入门书。是写给在实际工作中使用判别分析而又希望有一本简明扼要介绍判别分析的书的读者，也适用于希望得到这方面较深入知识的统计学家的用户。本书不是为想全面了解判别分析理论发展的读者写的；这方面的读者可以参阅 Das Gupta（142a）①的近期论文。

近年来，工程技术人员在特征识别和图象识别领域中做了大量的工作。遗憾的是这方面的许多工作没有引起统计学家的注意，因此导致了某些重复的劳动。有关判别分析方面的许多参考文献列于书后的参考书目中。

本书所需要的数学和统计知识是相当初等的。读者仅需具备如分布和初等概率这样一些基本概念。某些数学推导需要矩阵代数的知识。然而，作者希望没有这方面的知识也能阅读下去。

本书的安排如下：第一章给出了有关判别分析的一般介绍，并进行了基本推证。第二章讨论了判别函数的估计。第三章涉及到线性判别函数的稳健性。第四章包含了某些非参数方法。在第五章中讨论了多重组问题。其他方法包含在第六章中，包括序贯方法、变量的选择和Bayesian方法。书末附有相当广泛的参考书目。（以下略译）

①(142a)表示书末参考书目中的序号，下同。译者注

符　号

符号	定义
X	$K\times 1$维矢量的观察值
Π_i	总体i
μ_i	Π_i中$K\times 1$均值矢量
Σ_i	Π_i中$K\times K$协方差矩阵
$\bar{X}_i$	Π_i中$K\times 1$样本均值矢量
S_i	Π_i中$K\times K$样本协方差矩阵
p_i	来自Π_i的一个观察值的先验概率
K	变量个数
$D_T(X)=(X-\frac{1}{2}(\mu_1+\mu_2))'\Sigma^{-1}(\mu_1-\mu_2)$	总体的判别函数（参数已知）
$D_S(X)=(X-\frac{1}{2}(\bar{X}_1+\bar{X}_2))'S^{-1}(\bar{X}_1-\bar{X}_2)$	样本的判别函数（参数未知）
$f_i(X)$	在Π_i中X的密度函数
P_i	来自Π_i的一个观察值错误分类的概率
Φ	累计正态分布函数
n_i	来自Π_i的样本大小
δ^2	$(\mu_1-\mu_2)'\Sigma^{-1}(\mu_1-\mu_2)=$ Mahalanobis δ^2距离（参数已知）
D^2	$(\bar{X}_1-\bar{X}_2)'S^{-1}(\bar{X}_1-\bar{X}_2)=$ Mahalanobis D^2距离（参数未知）
g	总体的个数

目　录

第一章　判别分析的基本思想

判别分析的基本问题是根据不知所属的观察值X，去判别该观察值属于两个（或多个）不同组中的哪一组。对某些问题来说，关于X在两组中的分布有相当完全的信息可以利用。此时，我们可利用这些信息，并且处理问题时把分布看作已知的。然而，在多数情况下，X分布的信息来源于这些组群的相对来说是较小的样本，因此所采用的方法略有不同。在判别分析方法的实际应用中出现的其他问题是：

1．怎样合理地制定判别原则？

2．违背关于判别原则所作假定的稳健性如何？

3．在判别原则中应选择什么样的变量？

本书将讨论这些问题和其他一些问题。

某些作者的观点是把判别分析作为描述和检验组间差异的方法。所包括的检验同多元方差分析是一样的。这种观点能够在Cooley和Lohnes（117）中看到。我的观点是，判别分析讨论以低误差率来判别一个未知观察值属于哪一组。用于判别的函数或函数组可能与多元方差分析方法中所使用的函数或函数组相一致。

在判别分析背景中，有这样的假定：我们能够以某种方式对原始数据正确地进行分类。即，存在某个变量或某些变量使我们能够按已确定的组分组。例如，研究肺癌时，一个医生习惯于将其分组为“有癌”和“无癌”。在预言一个心

脏病人继续生存或不幸死亡的研究中，人们在观察病人一段时间后，区分他们为继续生存和不幸死亡。使用这些变量预测一个病人属于哪一组不是容易的。在肺癌的例子中，由于实验费用和实验规模的限制，人们更倾向于研究那些有很大可能患这种病的人。在心脏病的例子中，由于时间滞后，在病人死亡之前不可能告知该病人属于哪一组。于是人们将使用另外的变量，这些变量是低代价且发现病人时就可使用。幸运的是这些变量对于作出准确的判别是足够灵敏的。另外若能早期诊断出某种疾病，给予适当治疗，可以重新编组。（例如，在心脏病的例子中，对死亡的早期预测，可使某些病人得到及时治疗。）

现在给出判别分析问题某些类型的例子。

1．将一个心电图图迹划分成5毫秒的区间，并在这些点上读值。另外，测量QRS综合波的长度。根据这些信息，判别一个病人是“正常”还是“非正常”（即，一个人的心脏是有病还是没病）。因为不同年龄和性别的人心电图是不同的，故必须对病人划分年龄——性别组。能够较容易地以某种方式综合这些数据为一个单一数，而并不看个别的测量值，利用这个单一数，判断病人是属于正常组还是非正常组（503）。

2．一个住院诊断为心肌梗塞的人，测得了收缩压、舒张压、心率、猝发次数和平均动脉压之值，有可能预测该病人“继续生存”吗？能使用这些测量数据算出该病人继续生存的概率吗？

3．在一个区域的五个气象站我们有许多预报变量。这些包括：能见度、云层高度、东西风分量、南北风分量、云层

复盖总量和最近三小时的气压变化。根据这些观测，我们希望预报两小时内机场上空云层状况。我们必须由此确定机场是否关闭、低仪表飞行、高仪表飞行、低开放或高开放。这是一个多重组判别问题的例子（373）。

4．一个地质学家获得了一个海滩沉积颗粒大小的均值、方差、倚斜度和峭度。如何使用这些统计量确定该海滩原来是波形成的还是风形成的？粒度大小的分布有差异吗(216)？

5．在一项城郊发展规划中，出现这样的问题：这个区域发展的最佳方案是什么？这个问题也可以考虑为确定该区域应是何种类型。变量应当包括：到最近的 25 万人以上城市的距离、湖泊的面积、森林占陆地的百分比、到最近主要飞机场的距离等等。根据这些变量，这个区域是规划为游览区还是工业区（70）？

6．在病人是死亡还是生存的问题中，什么是重要因素？能对生存作出精确的预测吗？对提供的大量数据对该问题已进行了研究（1a）。

7．考古学家①得到一个头盖骨，想知道它是属于 2 万年前居住在该地区的一个种族，还是居住在附近的后代。根据对这个头盖骨的测量，与对来自于两个总体的每一头盖骨集合所作的测量，可作出判断。

这些例子的共同点是根据与组有关的数据，判别个体属于哪一组。在某些情况下，希望能够找到这样一个给定变量的子集，据此能够作出最佳判别。

研究这样一个问题：某学生在高中最后一年进行三门考

①原文将archaeologist误为archeologist　译者注

试。根据算术考试成绩x_1，英语成绩x_2和有关课程成绩x_3，对学生将来学习的课程提出建议〔Porebski(418)〕。学生有四种可能的选择：工程、建筑、艺术和商业。从“大伦敦区的初级技术学院新生”中取了一个大样本并计算了样本均值和联合协方差矩阵。这些给在表1—1与表1—2中。

表 1—1 按行业的考试均值

行业	样本大小	$\bar{x}_1$: 算术	$\bar{x}_2$: 英语	$\bar{x}_3$: 有关课程
工程	404	27.88	98.36	33.60
建筑	400	20.65	85.43	31.51
艺术	258	15.01	80.31	32.01
商业	286	24.38	94.94	26.69

来源：Porebski (418)

表 1—2 协方差矩阵〔$S=(S_{ij})$〕

考试	x_1	x_2	x_3
x_1	55.58	33.77	11.66
x_2	33.77	360.04	14.53
x_3	11.66	14.53	69.21

来源：Porebski (418)

假定该生希望在学工程和建筑之间进行选择。帮助这个学生的一种方式是看他与已经选择了行业的学生进行比较。假定该生成绩为（x_1，x_2，x_3）且行业均值成绩为（$\bar{x}_{1E}$，$\bar{x}_{2E}$，$\bar{x}_{3E}$）和（$\bar{x}_{1B}$，$\bar{x}_{2B}$，$\bar{x}_{3B}$）。在学生成绩和行业均值成绩之间的Euclidian距离是平方和：

$$D_E^2 = \sum_{i=1}^{3} (x_i - \bar{x}_{iE})^2$$

$$D_B^2 = \sum_{i=1}^{3} (x_i - \bar{x}_{iB})^2 \qquad (1—1)$$

该生的成绩较接近哪一个行业的均值，人们应当建议他选择那个行业。但这并不很正确，因为相关的存在对不同的考试能产生不同的权。

将使用的另一种距离测度是通过协方差函数来权衡观察值。这等价于Fisher (172) 提出的方法。试求观察值的线性组合，它给出两组间差的平方相对于两组内方差的最大值（即，它极大化$\frac{d^2}{v}$）。假定该线性组合是

$$Z = a_1x_1 + a_2x_2 + a_3x_3 \qquad (1—2)$$

则组间差被估计为

$$\begin{aligned} d = \bar{Z}_E - \bar{Z}_B &= a_1(\bar{x}_{1E} - \bar{x}_{1B}) + a_2(\bar{x}_{2E} - \bar{x}_{2B}) \\ &\quad + a_3(\bar{x}_{3E} - \bar{x}_{3B}) \\ &= a_1(7.23) + a_2(12.93) + a_3(2.09) \end{aligned} \qquad (1—3)$$

Z的方差是

$$\begin{aligned} v &= \sum_i \sum_j a_i a_j S_{ij} \\ &= a_1^2(55.58) + 2a_1a_2(33.77) + 2a_1a_3(11.66) \\ &\quad + a_2^2(360.04) + 2a_2a_3(14.53) + a_3^2(69.21) \end{aligned} \qquad (1—4)$$

系数a_1，a_2，a_3可通过极大化d^2/v而求得。即通过d^2/v依次对每一个系数求导，并令其等于0。例如，

$$\frac{\partial d^2/v}{\partial a_1} = \left(2vd \frac{\partial d}{\partial a_1} - d^2 \frac{\partial v}{\partial a_1} \right) / v^2 = 0$$

给出

$$2vd(7.23) - 2d^2(a_1(55.58) + a_2(33.77) + a_3(11.66) = 0 \qquad (1—5)$$

由于系数被确定仅差常数倍（即，若人们由极大化d^2/v来确定系数，那么它的任意倍数也能极大化d^2/v），我们写（1—5）式为

$$55.58a_1 + 33.77a_2 + 11.66a_3 = 7.23 \qquad (1—6)$$

分别对a_2和a_3求导，我们得到

$$33.77a_1 + 360.04a_2 + 14.53a_3 = 12.93$$

$$11.66a_1 + 14.53a_2 + 69.21a_3 = 2.09 \qquad (1—7)$$

这些方程的解是

$$a_1 = 0.1136 \qquad a_2 = 0.0250 \qquad a_3 = 0.0058$$

这里的解不同于Porebski(418)给出的解，这是因为我们利用协方差矩阵，而Porebski利用的是组内平方和与矩阵积。为了分类使用的函数是

$$D(\mathbf{X}) = 0.1136x_1 + 0.0250x_2 + 0.0058x_3 \qquad (1—8)$$

如果最适合于工程的几率与最适合于建筑的几率是一样的，那么最好的分点在两组均值中间。对于工程，D（X）的均值是

$$(0.1136)(27.88) + (0.0250)(98.36) + (0.0058)(33.60) = 5.8210$$

对于建筑，它是

$$(0.1136)(20.65) + (0.0250)(85.43) +$$

$$+(0.0058)(31.51)=4.6643$$

中点是

$$(0.5)(5.8210+4.6643)\approx 5.24$$

于是分类原则是：若$D(\mathbf{X})>5.24$，该生更适合工程；若$D(\mathbf{X})<5.24$，则更适合于建筑。对于所有的组对，能够进行类似的分析。对于某个体最适合哪一组通过顺序地消除一些组能够得到判别（三步）。

判别函数组均值之间的差是$5.8210-4.6643\approx 1.16$。这个量是Mahalanobis（马哈拉诺比斯）距离（以下简称马氏距离——译者注），它能校正相关的影响。表1—3给出了全部马氏距离。可以看出最大间距是工程和艺术（“最容易被分离”），最类似的对是建筑和艺术。再看均值表，我们看到建筑和艺术的算术和英语成绩较低，而与二者有关的课程，成绩是中等的。

表 1—3　成对的马氏距离

	E	B	A
B	1.16		
A	3.31	0.63	
C	0.70	0.90	2.66

来源：Porebski（418）

利用判别函数可以告诉学生他最类似哪一组。因为某些组显著地重叠，在分配中可能存在有误差。例如一个工程的成绩离建筑的均值比离工程的均值更近。当观察值近似正态分布时，可以利用马氏距离得到误差率的一个近似。若将第i组和第j组之间的马氏距离记为d_{ij}，则误差率近似地为1-

$\Phi(\sqrt{d_{ij}}/2) = \Phi(-\sqrt{d_{ij}}/2)$。表1—4给出了这些数据的近似误差率。可以看出，靠近的组比间隔大的那些组有较高的误差率。但使用这个误差率的估计要特别谨慎。如果数据不是正态的，或样本容量较小时，这个方法不好使用。此时需要非常小心。在第二章我们将讨论这一点。

第二个例子是建议高中学生选择学院预科课程或选非学院预科的课程。

表 1—4 Porebski提供数据的近似误差率

	E	B	A
B	0.29		
A	0.18	0.35	
C	0.33	0.32	0.21

在许多高级中学，9—10年级的学生在选择专业时询问指导顾问。一般说来，他们有学生过去学习的记录和一套标准的考试。根据这些数据指导顾问建议学生选择有学院预科的课程或没有学院预科的课程。Lohnes 和 McIntire(339a) 根据教育考试服务机构给出的六门考试报告了令人满意的预测能力。表1—5给出了观察值的均值和标准差。有学院预科的组为404，没有预科组为424。

计算判别函数以后，将观察值代入函数并分类。这就产生了像给在表1—6中那样的明显误差率。这个方法有一个偏差，因为同样的观察值既被用来计算函数又被用来估计它。另一方法，尽管它不总是可能的，这个方法有一个数据分离集合并且能被分类。表1—7给出了这一研究的结果。于是

表 1—5 Lohnes—McIntire所提供数据的均值和标准差

考　试	学院预科 $\bar{x}$	学院预科 S	非学院预科 $\bar{x}$	非学院预科 S
SCAT口语	289.2	11.2	274.6	12.1
SCAT量	302.3	20.0	288.3	19.7
词汇量	156.4	7.2	147.6	8.2
阅读水平	154.7	7.5	146.0	9.0
阅读速度	155.5	8.8	146.0	8.3
阅读理解力	155.1	8.8	145.5	8.2

来源：Lohnes和McIntire(335a)。

这个偏差是0.280－0.256＝0.024①。此时，这个偏差小，但样本容量较小时，这个偏差可能相当大。

表 1—6 错似误差率

判归的组	实际的组		
	学院预科	非学院预科	总　计
学院预科	304	112	416
非学院预科	100	312	412
总　计	404	424	828

$P_1 = 100/404 = 0.248$

$P_2 = 112/424 = 0.264$

$\bar{P} = 212/828 = 0.256$

来源：Lohnes和McIntire (335a)

①原文误为0.280－0.296＝0.024，已改正　译者注