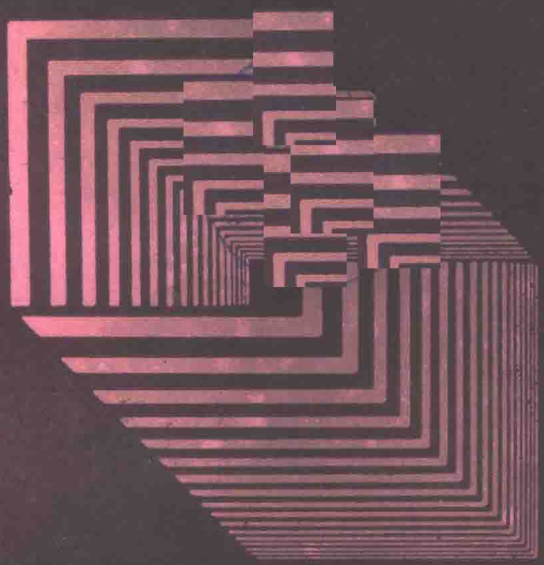
实用判别分析

孙尚拱　潘恩沛 编著

科学出版社



实用判别分析

孙尚拱　潘恩沛　编著

科学出版社

1990

内 容 简 介

本书主要介绍解决不同实际问题的各种判别方法，包括正态变量的两母体判别法、正态变量的多母体判别法、费歇判别法、分布未知时的判别法、离散型变量判别法、逻辑斯谛回归分析及混合型变量判别法．重点放在方法、概念及结果的解释上．书中也反映了近几年判别分析领域中最新研究成果．

本书可供生物学、医学、社会科学工作者及一般从事实际工作的统计工作者参考．

实 用 判 别 分 析

孙尚拱 潘恩沛 编著

责任编辑 毕 颖

科学出版社出版

北京东黄城根北街16号

中国科学院印刷厂印刷

新华书店北京发行所发行 各地新华书店经售

1990年1月第一版 开本：787×1092 1/32

1990年1月第一次印刷 印张：12 5/8

印数：0001-880 字数：285 000

ISBN 7-03-001279-8/O·287

定价：9.50元

绪　言

判别分析是多变量分析方法中较为成熟的一类方法，近年来已在我国社会科学和自然科学各个领域获得了广泛的应用，并取得了丰硕的成果.

判别分析根据两个或多个已知不同母体的抽样结果，按照所确定的准则，建立数学模型或函数，从而用来判别任一新观测到的样品应归属哪一个母体. 比如，天气预报中判别"晴"和"雨"，地质找矿中判别"矿"与"非矿"，医学中判别"肺癌"和"非肺癌"，等等，都是二母体判别问题. 判别问题也可称为分类问题或决策问题.

为了区分不同的母体，就必须获得一系列的可观测的变量(或称"指标"、"特征"、"因素"等). 这样，判别分析的第二项任务就是检验两个或多个母体在所观测到的指标上是否存在显著差异. 比如，肺癌患者、一般肺病患者及无肺病者的肺组织中多种微量元素（如铁、锌、砷等）含量是否有显著的不同？如果有，又是哪些元素的差异最为突出？这在医学中常称之为因素分析问题. 其它领域也有类似的问题. 另外，当测量到的变量很多时，从判别效果和经济方面考虑，都必须加以挑选，力求去掉冗余的变量. 这又是判别分析的一项基本内容，习惯上一般称为"变量的筛选".

应该指出，作为一种统计方法，判别分析所处理的问题一般都是机制不甚清楚或者基本不了解的复杂问题，当然一切判别结果都带有一定的不确定性. 这就有错判率估计的问题.

本书是为实际工作者撰写的．我们力求在不长的篇幅中介绍较为丰富的、适合于不同领域应用的判别法的模型及算法．目前，在多元统计应用的中文文献中，对判别错判率的介绍较为零散，对判别方法的蒙特卡洛模拟结果也极少涉及．而这些内容对于实际工作者能有针对性地选择判别法模型以及正确地解释所获得的计算结果，都是很有用的，因此我们也花了一定的篇幅较为系统地介绍了这些方面的知识．

本书在第一章第一节首先简要地介绍各种常用的判别准则．以后的章节基本上按观察变量的形式及其概率分布形式分别加以论述．第一章是正态变量下两母体的判别法，第二章是正态变量下的多母体判别法．前两章重点是讨论观察变量呈多维正态分布且具有等协方差阵时的判别问题，常称之为贝叶斯（Bayes）判别法．第二章还着重指出变量的正态性条件可以极大地放宽．第三章是费歇（Fisher）判别法，这种方法出现时间最早，对变量类型及其概率分布形式未作明确限定．但在正态等协方差阵时，其判别函数与贝叶斯判别函数又完全相同，因此很多文献常简单地把两母体时的判别函数称之为费歇判别函数．第四章是概率分布未知时的判别法，它特别适用于各母体间协方差阵不等的情况，但所处理的主要是连续性变量．第五章讨论离散变量判别法，重点是讨论如何利用离散变量在结构上的特点去构造判别函数．第六章讨论混合型变量的判别法，重点是逻辑斯谛判别法（也称逻辑斯谛回归），此外还介绍位置模型判别法．

我们约定：在本书中将客体上观察到的变量向量称之为样品（或样本），把它所属类别称为母体，并记 π_i 为第 i 个母体．在 π_i 中抽取的全部样品称为样本．我们总是假定样本的抽取是符合统计学上的随机性及独立性条件的．当抽取的样品数较少时称之为小样本，数量大时则称之为大样本．

判别分析种类繁多，本书当然无法全部介绍．我们力求能较全面地综合近年来国内外在判别分析研究方面的新成果．书中许多内容是作者的研究成果。

作　者

目　录

第一章　正态变量的两母体判别法

第一节　构造判别函数的主要准则

在判别分析中，大多数的判别方法都是通过判别函数来完成的，而构造判别函数的出发点或准则又是多种多样的.

本章讨论两个母体 π_1, π_2 的情况. 记 $\boldsymbol{X}=(x_1, x_2, \cdots, x_p)'$ 为观测向量，它相当于一个样本(或称样品). 记 q_1, q_2 分别为母体 π_1, π_2 的先验概率(以后简称"先验率")，即 $P(\pi_i)=q_i$，其中 $i=1, 2$. 它的意义是：如把两个母体 π_1, π_2 全部混合，则随机抽得的一个样品 $\boldsymbol{X}$ 属于 π_1 的概率为 q_1，属于 π_2 的概率为 q_2，显然，这里有 $q_1+q_2=1$.

一、平均损失最小准则

设判别函数为 $u=u(\boldsymbol{X})$，记 $\boldsymbol{X}$ 的定义域为 D，现将空间 D 分成两部分 D_1, D_2，显然 D_1 与 D_2 之和为 D. 规定判别方法为

$$\begin{cases}\text{若 } \boldsymbol{X} \text{ 落入 } D_1, \text{ 则判 } \boldsymbol{X} \text{ 属 } \pi_1\\ \text{若 } \boldsymbol{X} \text{ 落入 } D_2, \text{ 则判 } \boldsymbol{X} \text{ 属 } \pi_2\end{cases} \tag{1.1}$$

记 $C(2/1)$ 为把母体 π_1 中的样品错误地判属 π_2 所造成的损失值，$C(1/2)$ 为把母体 π_2 中的样品错误地判归 π_1 所造成的损失值，又 $P(D_2/\pi_1)$ 表示应属 π_1 的样品点落入 D_2 的概率，$P(D_1/\pi_2)$ 意义类似，则一个待判样品 $\boldsymbol{X}$ 被错误分类而造成的平均损失值 R 为 $C(2/1)$ 与 $C(1/2)$ 的概率平均值，即

$$R = C(2/1) \cdot P(D_2/\pi_1) + C(1/2) \cdot P(D_1/\pi_2) \quad (1.2)$$

一般 $C(2/1)$ 和 $C(1/2)$ 人为地确定.

一个"好"的判别函数 $u(\boldsymbol{X})$ 应能使 R 达到最小. 这一准则称为平均损失最小准则,又称为贝叶斯准则,由此在判别分析中使 (1.2) 中平均损失为最小的判别函数 $u(\boldsymbol{X})$ 就被称为贝叶斯判别函数.

平均错分率最小准则: 作为上述准则的一个重要的特例,如果我们在分类时不考虑"损失",而仅关心任一待判样品 $\boldsymbol{X}$ 被错误分类的概率,我们可以取在平均意义上错分概率为最小作为一种判别准则,这个准则被称为平均错分率最小准则. 此时只要在 (1.2) 式中取 $C(1/2) = C(2/1) = 1$ 即可.

由于一般实际问题中损失值 $C(j/i)$ 很难定量化,因此,人们常用平均错分率作为判别效果好坏的尺度. 这样,平均损失最小准则常被平均错分率最小准则所代替,后者显然是最令人感兴趣的准则.

二、费歇准则

这是与"距离"概念相联系的一个准则. 如果已知判别函数 $u(\boldsymbol{X})$,则理论上在母体 π_1 中 $u(\boldsymbol{X})$ 的均值 $\bar{u}_1$ 和方差 σ_1^2 是一定的,同样的道理,在母体 π_2 中 $u(\boldsymbol{X})$ 的均值 $\bar{u}_2$ 和方差 σ_2^2 也是一定的.

费歇准则的思想是: 要使得母体 π_1 与 π_2 间距离 $(\bar{u}_1 - \bar{u}_2)^2$ 尽可能地大,而同时使各母体内方差 σ_1^2 和 σ_2^2 尽可能地小,即在

$$I = \frac{(\bar{u}_1 - \bar{u}_2)^2}{\sigma_1^2 + \sigma_2^2} \quad (1.3)$$

为极大的条件下,求得判别函数 $u(\boldsymbol{X})$. 这是判别分析历史

上最老的准则.

三、最小平方准则（LMS）

记 $f_i(X)$ 为母体 π_i 下 X 的概率密度函数，$i=1,2$，又我们希望母体 π_i 内判别函数 $u(X)$ 的值都尽可能地集中在某个定值 a_i 附近. 最小平方准则是要求 $u(X)$ 离理想值 a_1 和 a_2 的偏差之平均平方为最小，即要求

$$c^2 = q_1\int [u(X)-a_1]^2 f_1(X)dX + q_2\int [u(X) - a_2]^2 f_2(X)dX \tag{1.4}$$

为最小. 上述积分均为 p 维重积分，p 是指标（或称为变量、因素等）的个数.

四、库尔贝克（Kullback）准则

这是一个与信息论有关的准则. 记 $f_i(u)$ 为母体 π_i 内判别函数 $u=u(X)$ 的概率密度，$i=1,2$，则对 $u(X)$ 在两母体之一的判别中有利于 π_1 的平均信息为

$$I(1:2) = \int f_1(u)\ln\frac{f_1(u)}{f_2(u)}du \tag{1.5}$$

同样，有利于 π_2 的判别之平均信息为

$$I(2:1) = \int f_2(u)\ln\frac{f_2(u)}{f_1(u)}du \tag{1.6}$$

再定义 $u(X)$ 的散度 $J(1:2)$ 如下：

$$J(1:2) \triangleq I(1:2) + I(2:1) = \int (f_1(u)-f_2(u))\ln\frac{f_1(u)}{f_2(u)}du \tag{1.7}$$

在使 $J(1:2)$ 达到极大的条件下求判别函数 $u(\boldsymbol{X})$ 的准则称为库尔贝克准则，由此所得到的 $u(\boldsymbol{X})$ 就称为库尔贝克判别函数。

五、不确定性准则

这个准则是直接来源于信息论中熵的概念。设判别函数为 $u=u(\boldsymbol{X})$，又判别规则要求把样本空间 D 划分成 D_1 和 D_2 两部分，记 $P(\pi_i, D_j)$ 为 π_i 的样本点 X 落在空间 D_j 内的概率，$P(\pi_i|D_j)$ 为样本点已落在空间 D_j 中而被判属于母体 π_i 的条件概率，则

$$H(\pi|D)=-\sum_{i=1}^{2}\sum_{j=1}^{2}P(\pi_i, D_j)\ln P(\pi_i|D_j) \qquad (1.8)$$

表示 D 被划分成 D_1，D_2 后母体 π_1，π_2 仍然有不确定性的程度。这样，使(1.8)式表示的不确定性达到极小的函数 $u(\boldsymbol{X})$，就称为不确定性准则下的最佳判别函数。

六、最大似然准则

如记母体 π_i 中已抽取的样品为 $\boldsymbol{X}_1^{(1)}, \boldsymbol{X}_2^{(1)}, \cdots, \boldsymbol{X}_{n_1}^{(1)}$，母体 π_2 中已抽取的样品为 $\boldsymbol{X}_1^{(2)}, \boldsymbol{X}_2^{(2)}, \cdots, \boldsymbol{X}_{n_2}^{(2)}$，记 $P(\boldsymbol{X}|\pi_i)$ 为已知 $\boldsymbol{X}$ 属于母体 π_i 的情况下 $\boldsymbol{X}$ 出现的概率或概率密度，而称

$$L=\prod_{i=1}^{2}\prod_{j=1}^{n_i}P(\boldsymbol{X}_j^{(i)}|\pi_i) \qquad (1.9)$$

为母体 π_1，π_2 上样品的似然函数。由于概率 $P(\boldsymbol{X}|\pi_i)$ 与判别函数 $u(\boldsymbol{X})$ 有关，在使(1.9)式中 L 为最大的条件下求得的 $u(\boldsymbol{X})$，就称之为最大似然判别函数。

除了上面的与建立函数 $u = u(\boldsymbol{X})$ 有关的判别准则外，还有一些判别方法并不直接从建立判别函数出发，例如下面的两个判别准则就是如此.

七、最大概率准则

如果我们用任何一种方法已经估计出每个母体中 $\boldsymbol{X}$ 的概率 $P(\boldsymbol{X}|\pi_i)$，$i = 1, 2$，则利用条件概率公式得

$$P(\pi_i|\boldsymbol{X}) = \frac{P(\boldsymbol{X}|\pi_i) \cdot P(\pi_i)}{P(\boldsymbol{X}|\pi_1) \cdot P(\pi_1) + P(\boldsymbol{X}|\pi_2) \cdot P(\pi_2)} \quad i = 1, 2 \tag{1.10}$$

若先验率 $P(\pi_i)$ 为 q_i 且事先已知，可直接代入上式，若不知道，实际计算时可取 $q_1 = q_2 = 0.5$.

对每个新样本点 $\boldsymbol{X}$，只要计算 $P(\pi_1|\boldsymbol{X})$ 和 $P(\pi_2|\boldsymbol{X})$，比较这二者大小，即判 $\boldsymbol{X}$ 属于与较大概率相应的母体.

由于上式中对 $i = 1, 2$ 的两式分母是公共的，所以比较 $P(\pi_i|\boldsymbol{X})$ 时实际上只要比较分子即可. 由此得最大概率判别法为

$$\left.\begin{array}{l}\text{若 } P(\boldsymbol{X}|\pi_1) \cdot q_1 \geqslant P(\boldsymbol{X}|\pi_2) \cdot q_2, \text{ 则判 } \boldsymbol{X} \text{ 属 } \pi_1 \\ \text{否则，判 } \boldsymbol{X} \text{ 属 } \pi_2\end{array}\right\} \tag{1.11}$$

实际上将 (1.10) 式中概率 $P(\boldsymbol{X}|\pi_i)$ 换成相应的密度函数仍可使用.

八、马氏距离最小准则

马哈拉诺比斯 (Mahalanobis) 距离是一种广义距离，本书中一律简称为马氏距离.

记母体 π_i 的均值向量为 $\boldsymbol{\mu}^{(i)}$，协方差阵为 $\sum_i$，则一个新

样本点 X 到母体 π_i 的马氏距离定义为

$$d^2(X, \pi_i) \triangleq (X-\mu^{(i)})'\Sigma_i^{-1}(X-\mu^{(i)})$$

式中 $(X-\mu^{(i)})'$ 为 $(X-\mu^{(i)})$ 的转置，Σ_i^{-1} 为矩阵 Σ_i 的逆，往后类似符号将不再说明.

$$\left.\begin{array}{r}\text{若 } d^2(X, \pi_1) \leqslant d^2(X, \pi_2), \text{则判 } X \text{属 } \pi_1\\ \text{否则，判 } X \text{属 } \pi_2\end{array}\right\} \quad (1.12)$$

在判别分析中，判别准则和判别方法很多，上述是其中最有名的一些. 对于同一批数据采用不同的准则或方法来处理，其结果一般不会相同. 可是在统计理论研究中早已证明：在两母体时，如 X 在 π_1 及 π_2 中的分布服从多维正态分布且有相同的协方差阵（在本书中凡不致引起误会时这个假定简称为正态等协方差阵假定），则上述八个判别准则之效果是相同的.

第二节　二次型判别函数与刀切法

一、关于正态性和等协方差阵性

这里我们约定：观测向量 X 在各母体中的概率分布服从多维正态分布，而且各协方差阵都相同.

众所周知，要判断已有的一批数据是否服从多维正态分布，并不是一件简单的事，至今还没有简易而又实用的方法. 但是，反过来，要肯定数据不服从多维正态分布倒是有一些简易方法，其依据的原理是：如果 $X=(x_1, x_2, \cdots, x_p)'$ 服从 p 维正态分布，则它的每个分量 $x_i(i=1, 2, \cdots, p)$ 必服从一元正态分布. 因此把某个分量（比如 x_1）的 n 个样品值作成直方图，或用文献 [1] 中 §5 介绍的一些简易方法处置，如果断定 x_1 不是正态分布，则我们也就可以断定 X 也不可能是

多维正态分布的了.

应该指出，即使是 X 不服从多维正态分布，形式上贝叶斯判别函数也还是成立的，不过此时的判别函数已失去了贝叶斯意义下(即损失最小意义下)的最佳性. 此外，一些统计检验公式的合理性就必然会受影响. 看来受影响最严重的大概要算理论错判率的公式估计法. 因此，在可能情况下，我们应当尽量把非正态数据通过各种变换变成正态分布. 例如，自然界微量元素常呈对数正态分布，这时我们可以将原始数据取对数，就是一种最简单的办法.

关于多母体等协方差阵的检验，对于大样本情况有一个不太复杂的方法，即巴特利特 (Bartlett) χ^2 检验，现介绍如下.

设有 k 个母体，每个母体中 X 都呈多维正态分布. 记 π_r 中 X 的母体均值向量为 $\mu^{(r)}$，协方差阵为 Σ_r，一般可简记为

$$X|_{\pi_r} \sim N_p(\mu^{(r)}, \Sigma_r) \quad r = 1, 2, \cdots, k \tag{1.13}$$

我们的检验是在 X 为多维正态分布条件下，检验下面的零假设:

$$H_0: \Sigma_1 = \Sigma_2 = \cdots = \Sigma_k = \Sigma \tag{1.14}$$

现以已有的一批样本来进行检验.

设在母体 π_r 中抽取的 n_r 个样品记为 $X_1^{(r)}, X_2^{(r)}, \cdots, X_{n_r}^{(r)}$，令

$$\bar{X} = \frac{1}{n_r} \sum_{i=1}^{n_r} X_i^{(r)} \tag{1.15}$$

$$W_r = \sum_{i=1}^{n_r} (X_i^{(r)} - \bar{X}^{(r)})(X_i^{(r)} - \bar{X}^{(r)})' \tag{1.16}$$

$$S_r = W_r/(n_r - 1) \tag{1.17}$$

$$W = \sum_{r=1}^{k} W_r \tag{1.18}$$

$$S = W/(N-k) \tag{1.19}$$

其中 $n = n_1 + n_2 + \cdots + n_k$，$\bar{X}^{(r)}$ 为 π_r 的样品均值向量，W，及 S_r 分别为 π_r 的样品离差阵及协方差阵. 当(1.14)式零假设 H_0 成立时，则 S 就是公共的母体协方差阵之无偏估计，而且 $\bar{X}^{(r)}$ 同 S_r（因而同 S）是独立的，这都是早有证明的.

记 $|S_r|$ 及 $|S|$ 分别为 S_r 及 S 的行列式，则检验(1.14)式零假设的卡方公式为[1]

$$\chi^2 \approx (N-k)\ln|S| - \sum_{r=1}^{k}(n_r-1)\ln|S_r| - \sum_{r=1}^{k}(n_r-1)\ln\frac{|S|}{|S_r|} \tag{1.20}$$

自由度为 $(k-1)\cdot p\cdot(p+1)/2$. 如果计算出的 χ^2 值大于规定的显著性水平 α 相应的临界值 χ^2_α，则拒绝(1.14)式零假设 H_0，否则接受 H_0.

公式(1.20)并不复杂，主要工作量在于计算行列式，但实际工作者很少使用它. 我们以为，在可能而且必要时还是应该进行检验. 也可能有这种情况：在大多数实际（而不是人造）数据的例子里，常常是不等协方差阵的，而等协方差阵的例子倒是不多见的. 以医学问题为例，如果母体 π_1 表示“病人”，π_2 表示“正常人”，等协方差阵的要求是：“病人”任两项指标间的相关性与对应的“正常人”同样两项指标间的相关性要相同，而且“病人”与“正常人”在任何指标上的波动情形也要相同. 而从常识判断，并且实际数据也表明，这是不大可能的. “病人”的各项指标波动性总是比较大的，而且两项指标间的相关性也不会同于常人. 另外，从(1.20)式可见，只要 $|S_r|$ 与 $|S|$ 略有不同，则 n_r 足够大时总可使(1.20)式的

1) 在某些统计书中有比(1.20)式更精细的公式，但我们认为不必要.