第五章 判别分析

- §5.1 引言
- §5.2 距离判别
- §5.3 贝叶斯判别
- §5.4 费希尔判别
- §5.5 逐步判别

判别分析的目标

- 目标1（预测方面）：分类（或分配）。

  在已知历史上用某些方法已把研究对象分成若干组（亦称类或总体）的情况下，来判定新的观测样品应归属的组别。

- 目标2（描述方面）：分离。

  就是用图形（通常二维，有时三维或一维，一般通过降维实现）方法或代数方法描述来自各组的样品之间的差异性，最大限度地分离各组。

§5.1 引言

- 要判定一个样品的归属，理想的情况似乎是能够获得完备的用于分类的信息，以作出准确的判断。但这往往是不太现实的，因为
  - 要获得完备的信息可能根本做不到（如《红楼梦》后四十回的作者到底是谁）
  - 要做破坏性的试验（如欲获知某电子仪器的寿命）
  - 成本高昂（如许多疾病只有通过代价高昂的手术才能确诊）。
- 实践中往往是依据不完备信息来进行判别分类的。

判别分类的例子

- 有偿付力与无偿付力的财产责任保险公司。

  测量变量：总资产，股票与债券价值，股票与债券的市值，损失支出，盈余，签定的保费金额。

- 非溃疡胃病组（胃功能紊乱者）与控制组（“正常”者）。

  测量变量：焦虑、依赖性、罪恶感、完美主义的量度。

- 两种野草。

测量变量：萼片与花瓣的长度，花瓣裂缝的深度，苞的长度，花粉直径。

- 新产品的速购者与迟购者。

测量变量：教育，收入，家庭大小，过去更换品牌的次数。

- 良好信用与不良信用风险。

测量变量：收入，年龄，信用卡数目，家庭规模。

本章讨论的判别分析

- 每一组中所有样品的p维指标值 $\boldsymbol{x}=\left(x_1,x_2,\cdots,x_p\right)'$ 构成了该组的一个p元总体分布。
- 我们对新样品$\boldsymbol{x}$进行的判别归类将在很大程度上依赖于各组的总体分布或其分布特征。
- 距离判别和贝叶斯（Bayes）判别只能用于分类。
- 费希尔（Fisher）判别即可用于分类，也可用于分离，且更多地用于后者。
- 这些都是基于判别变量为定量变量的。

§5.2 距离判别

- 一、两组距离判别
- 二、多组距离判别

一、两组距离判别

- 设组π_1和π_2的均值分别为$\boldsymbol{\mu}_1$和$\boldsymbol{\mu}_2$，协差阵分别为$\boldsymbol{\Sigma}_1$和$\boldsymbol{\Sigma}_2(\boldsymbol{\Sigma}_1,\boldsymbol{\Sigma}_2>0)$，$\boldsymbol{x}$是一个新样品（$p$维），现欲判断它来自哪一组。
- （基于马氏距离的）判别规则：

$$\begin{cases} \boldsymbol{x} \in \pi_1, & 若 d^2(\boldsymbol{x},\pi_1) \le d^2(\boldsymbol{x},\pi_2) \\ \boldsymbol{x} \in \pi_2, & 若 d^2(\boldsymbol{x},\pi_1) > d^2(\boldsymbol{x},\pi_2) \end{cases}$$

- 1. $\boldsymbol{\Sigma}_1=\boldsymbol{\Sigma}_2=\boldsymbol{\Sigma}$时的判别
- 2. $\boldsymbol{\Sigma}_1\ne\boldsymbol{\Sigma}_2$时的判别

1. $\boldsymbol{\Sigma}_1=\boldsymbol{\Sigma}_2=\boldsymbol{\Sigma}$时的判别

$$
\begin{aligned}
& d^2(\boldsymbol{x},\pi_1)-d^2(\boldsymbol{x},\pi_2) \\
&=(\boldsymbol{x}-\boldsymbol{\mu}_1)'\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_1)-(\boldsymbol{x}-\boldsymbol{\mu}_2)'\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_2) \\
&=\boldsymbol{x}'\boldsymbol{\Sigma}^{-1}\boldsymbol{x}-2\boldsymbol{x}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_1+\boldsymbol{\mu}_1'\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_1-\left(\boldsymbol{x}'\boldsymbol{\Sigma}^{-1}\boldsymbol{x}-2\boldsymbol{x}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_2+\boldsymbol{\mu}_2'\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_2\right) \\
&=2\boldsymbol{x}'\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_2-\boldsymbol{\mu}_1)+\boldsymbol{\mu}_1'\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_1-\boldsymbol{\mu}_2'\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_2 \\
&=2\boldsymbol{x}'\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_2-\boldsymbol{\mu}_1)+(\boldsymbol{\mu}_1+\boldsymbol{\mu}_2)'\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2) \\
&=-2\left(\boldsymbol{x}-\frac{\boldsymbol{\mu}_1+\boldsymbol{\mu}_2}{2}\right)'\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2) \\
&=-2(\boldsymbol{x}-\bar{\boldsymbol{\mu}})'\boldsymbol{a}=-2\boldsymbol{a}'(\boldsymbol{x}-\bar{\boldsymbol{\mu}})
\end{aligned}
$$

其中$\bar{\boldsymbol{\mu}}=\frac{1}{2}(\boldsymbol{\mu}_1+\boldsymbol{\mu}_2)$，$\boldsymbol{a}=\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)$。

- 令$W(\boldsymbol{x})=\boldsymbol{a}'(\boldsymbol{x}-\bar{\boldsymbol{\mu}})$，则上述判别规则可简化为

$$\begin{cases}\boldsymbol{x}\in\pi_1, & 若W(\boldsymbol{x})\geq 0\\ \boldsymbol{x}\in\pi_2, & 若W(\boldsymbol{x})<0\end{cases}\qquad(5.2.3)$$

称$W(\boldsymbol{x})$为两组距离判别的（线性）判别函数，称$\boldsymbol{a}$为判别系数向量。

误判概率

- 误判概率

$$P(2|1)=P\left(W(\boldsymbol{x})<0\,|\,\boldsymbol{x}\in\pi_1\right)$$

$$P(1|2)=P\left(W(\boldsymbol{x})\geq 0\,|\,\boldsymbol{x}\in\pi_2\right)$$

- 设$\pi_1\sim N_p(\boldsymbol{\mu}_1,\boldsymbol{\Sigma})$, $\pi_2\sim N_p(\boldsymbol{\mu}_2,\boldsymbol{\Sigma})$，则

$$P(2|1)=P(1|2)=\Phi\left(-\frac{\Delta}{2}\right)$$

其中$\Delta=\sqrt{(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)'\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)}$是两组之间的马氏距离。

- 可见，两个正态组越是分开（即Δ越大），两个误判概率就越小，此时的判别效果也就越佳。当两个正态组很接近时，两个误判概率都将很大，这时作判别分析就没有什么实际意义了。

组之间是否已过于接近的界定

- 我们可对假设H_0：$\boldsymbol{\mu}_1=\boldsymbol{\mu}_2$，$H_1$：$\boldsymbol{\mu}_1\neq\boldsymbol{\mu}_2$进行检验，若检验接受原假设$H_0$，则说明两组均值之间无显著差异，此时作判别分析一般会是徒劳的。
- 若检验拒绝 H_0，则两组均值之间虽然存在显著差异，但这种差异对进行有效的判别分析未必足够大，此时还应看误判概率是否超过了一个合理的水平。

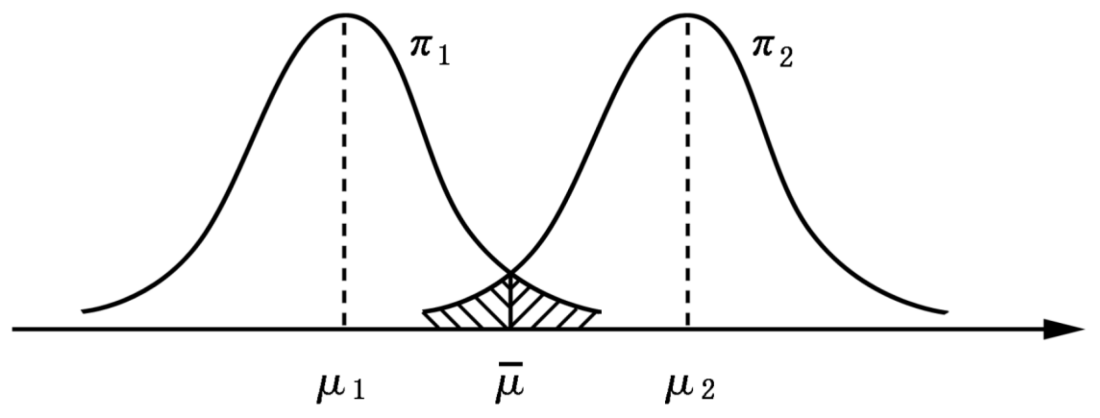
❖ 例5.2.1 设$p=1$，π_1和π_2的分布分别为$N(\mu_1,\sigma^2)$和$N(\mu_2,\sigma^2)$，μ_1,μ_2,σ^2均已知，$\mu_1<\mu_2$，则判别系数$a=(\mu_1-\mu_2)/\sigma^2<0$，

判别函数：
$$W(x)=a(x-\bar{\mu})$$

判别规则：
$$\begin{cases} x\in\pi_1, & 若x\le\bar{\mu} \\ x\in\pi_2, & 若x>\bar{\mu} \end{cases}$$

误判概率：
$$P(2|1)=P(1|2)=\Phi\left(-\frac{\Delta}{2}\right)=\Phi\left(\frac{\mu_1-\mu_2}{2\sigma}\right)$$

误判概率图示：



抽取样本估计有关未知参数

- 设$\boldsymbol{x}_{11},\boldsymbol{x}_{12},\cdots,\boldsymbol{x}_{1n_1}$是来自组$\pi_1$的样本，$\boldsymbol{x}_{21},\boldsymbol{x}_{22},\cdots,\boldsymbol{x}_{2n_2}$是来自组$\pi_2$的样本，$n_1+n_2-2\ge p$，则$\boldsymbol{\mu}_1$和$\boldsymbol{\mu}_2$的一个无偏估计分别为

$$\bar{\boldsymbol{x}}_1=\frac{1}{n_1}\sum_{j=1}^{n_1}\boldsymbol{x}_{1j} \quad 和 \quad \bar{\boldsymbol{x}}_2=\frac{1}{n_2}\sum_{j=1}^{n_2}\boldsymbol{x}_{2j}$$

$\boldsymbol{\Sigma}$的一个联合无偏估计为

$$\boldsymbol{S}_p=\frac{(n_1-1)\boldsymbol{S}_1+(n_2-1)\boldsymbol{S}_2}{n_1+n_2-2}$$

其中

$$\boldsymbol{S}_i=\frac{1}{n_i-1}\sum_{j=1}^{n_i}\left(\boldsymbol{x}_{ij}-\bar{\boldsymbol{x}}_i\right)\left(\boldsymbol{x}_{ij}-\bar{\boldsymbol{x}}_i\right)', \quad i=1,2$$

- 实际使用的判别函数为

$$\hat{W}(\boldsymbol{x})=\hat{\boldsymbol{a}}'(\boldsymbol{x}-\bar{\boldsymbol{x}}) \tag{5.2.5}$$

这里$\bar{\boldsymbol{x}}=\frac{1}{2}(\bar{\boldsymbol{x}}_1+\bar{\boldsymbol{x}}_2),\hat{\boldsymbol{a}}=\boldsymbol{S}_p^{-1}(\bar{\boldsymbol{x}}_1-\bar{\boldsymbol{x}}_2)$。其判别规则为

$$\begin{cases}\boldsymbol{x}\in\pi_1, & 若\hat{W}(\boldsymbol{x})\geq 0\\ \boldsymbol{x}\in\pi_2, & 若\hat{W}(\boldsymbol{x})<0\end{cases}$$

- 若π_1和π_2都为正态组，则两个误判概率$P(2|1)$和$P(1|2)$可估计为

$$\hat{P}(2|1)=\hat{P}(1|2)=\Phi\left(-\frac{\hat{\Delta}}{2}\right)$$

其中 $\hat{\Delta}=\sqrt{(\bar{\boldsymbol{x}}_1-\bar{\boldsymbol{x}}_2)'\boldsymbol{S}_p^{-1}(\bar{\boldsymbol{x}}_1-\bar{\boldsymbol{x}}_2)}$。

- 该误判概率的估计是有偏的，但大样本时偏差的影响是可以忽略的。

误判概率的非参数估计

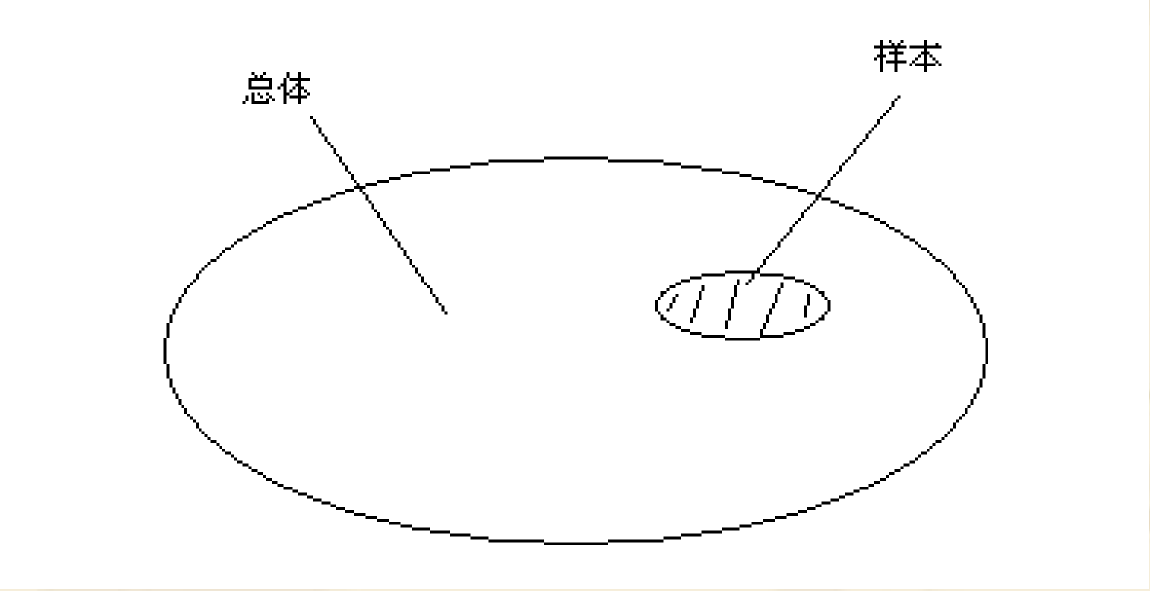
- 若两组不能假定为正态组，则$P(2|1)$ 和 $P(1|2)$ 可以用样本中样品的误判比例来估计，通常有如下三种非参数估计方法：
- (1)回代法
- 令$n(2|1)$为样本中来自π_1而误判为π_2的个数，$n(1|2)$为样本中来自π_2而误判为π_1的个数，则$P(2|1)$ 和$P(1|2)$ 可估计为

$$\hat{P}(2|1)=\frac{n(2|1)}{n_1},\quad \hat{P}(1|2)=\frac{n(1|2)}{n_2}$$

- 该方法简单、直观，且易于计算。但遗憾的是，它给出的估计值通常偏低，除非n_1和n_2都非常大。

出现乐观估计的原因

- 同样的样本信息被重复使用。判别函数自然对构造它的样本数据有更好的适用性，以致出现偏低的误判率。



- (2)划分样本

- 将整个样本一分为二，一部分作为训练样本，用于构造判别函数，另一部分用作验证样本，用于对该判别函数进行评估。误判概率用验证样本的被误判比例来估计，其估计是无偏的。
- 该方法的两个主要缺陷：

(i)需要用大样本；

(ii)该方法构造的判别函数只用了部分样本数据，与使用全部样本数据构造的判别函数相比，损失了过多有价值的信息，其效用自然不如后者，表现为前者的误判概率通常将高于后者的，而后者的误判概率才是我们真正感兴趣的。该缺陷随样本容量的增大而逐渐减弱，甚至可基本忽略。

(3) 交叉验证法（或称刀切法）

- 从组π_1中取出$\boldsymbol{x}_{1j}$，用该组的其余n_1-1个观测值和组π_2的n_2个观测值构造判别函数，然后对$\boldsymbol{x}_{1j}$进行判别，$j=1,2,\cdots,n_1$。同样，从组π_2中取出$\boldsymbol{x}_{2j}$，用这一组的其余n_2-1个观测值和组π_1的n_1个观测值构造判别函数，再对$\boldsymbol{x}_{2j}$作出判别，$j=1,2,\cdots,n_2$。
- 令

$n^*(2|1)$——样本中来自π_1而误判为π_2的个数

$n^*(1|2)$——为样本中来自π_2而误判为π_1的个数

则两个误判概率$P(2|1)$和$P(1|2)$的估计量为

$$\hat{P}(2|1)=\frac{n^*(2|1)}{n_1},\quad \hat{P}(1|2)=\frac{n^*(1|2)}{n_2}$$

它们都是接近无偏的估计量。

- 以上所述误判概率的这三种非参数估计方法同样适用于其它的判别方法或判别情形，并且可类似地推广到多组的情形。

2. $\Sigma_1 \neq \Sigma_2$时的判别

- 判别规则：

$$\begin{cases} \boldsymbol{x} \in \pi_1, & 若 d^2(\boldsymbol{x}, \pi_1) \le d^2(\boldsymbol{x}, \pi_2) \\ \boldsymbol{x} \in \pi_2, & 若 d^2(\boldsymbol{x}, \pi_1) > d^2(\boldsymbol{x}, \pi_2) \end{cases}$$

- 也可采用另一种形式：选择判别函数为

$$W(\boldsymbol{x}) = d^2(\boldsymbol{x}, \pi_1) - d^2(\boldsymbol{x}, \pi_2)$$

$$= (\boldsymbol{x} - \boldsymbol{\mu}_1)' \Sigma_1^{-1} (\boldsymbol{x} - \boldsymbol{\mu}_1) - (\boldsymbol{x} - \boldsymbol{\mu}_2)' \Sigma_2^{-1} (\boldsymbol{x} - \boldsymbol{\mu}_2)$$

它是**x**的二次函数，相应的判别规则为

$$\begin{cases} \boldsymbol{x} \in \pi_1, & 若 W(\boldsymbol{x}) \le 0 \\ \boldsymbol{x} \in \pi_2, & 若 W(\boldsymbol{x}) > 0 \end{cases} \qquad (5.2.10)$$

❖ 例5.2.2 在例5.2.1中，设π_1和π_2这两个组的方差不相同，分别为σ_1^2和σ_2^2，这时

$$d(x,\pi_i)=\frac{|x-\mu_i|}{\sigma_i},\quad i=1,2$$

当$\mu_1<x<\mu_2$时，判别函数可简单地取为

$$W(x)=d(x,\pi_1)-d(x,\pi_2)=\frac{x-\mu_1}{\sigma_1}-\frac{\mu_2-x}{\sigma_2}$$

$$=\frac{\sigma_1+\sigma_2}{\sigma_1\sigma_2}\left(x-\frac{\sigma_2\mu_1+\sigma_1\mu_2}{\sigma_1+\sigma_2}\right)=\frac{\sigma_1+\sigma_2}{\sigma_1\sigma_2}\left(x-\mu^*\right)$$

式中

$$\mu^*=\frac{\sigma_2\mu_1+\sigma_1\mu_2}{\sigma_1+\sigma_2}$$

它是μ_1与μ_2的加权平均，常称为阈值点，如图5.2.2所示。

- 判别规则为

$$\begin{cases} x \in \pi_1, & 若 x \le \mu^* \\ x \in \pi_2, & 若 x > \mu^* \end{cases}$$

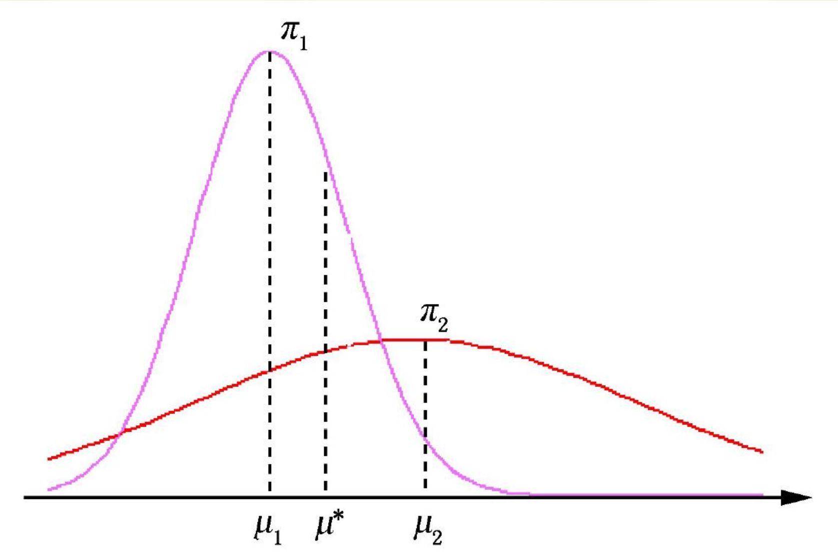


图5. 2. 2　方差不同时两组判别的阈值点

- 实际应用中，$\boldsymbol{\mu}_1$和$\boldsymbol{\mu}_2$，$\boldsymbol{\Sigma}_1$和$\boldsymbol{\Sigma}_2$一般都是未知的，可由相应的样本值代替。

二、多组距离判别

- 设有k个组$\pi_1,\pi_2,\cdots,\pi_k$，它们的均值分别是$\boldsymbol{\mu}_1,\boldsymbol{\mu}_2,\cdots,\boldsymbol{\mu}_k$，协方差矩阵分别是$\boldsymbol{\Sigma}_1(>0),\boldsymbol{\Sigma}_2(>0),\cdots,\boldsymbol{\Sigma}_k(>0)$，$\mathbf{x}$到总体$\pi_i$的平方马氏距离为

$$d^2(\mathbf{x},\pi_i)=(\mathbf{x}-\boldsymbol{\mu}_i)'\boldsymbol{\Sigma}_i^{-1}(\mathbf{x}-\boldsymbol{\mu}_i),\quad i=1,2,\cdots,k$$

判别规则为

$$\mathbf{x}\in\pi_l,\quad 若d^2(\mathbf{x},\pi_l)=\min_{1\le i\le k}d^2(\mathbf{x},\pi_i)$$

- 该判别规则不受变量单位的影响。
- 若$\boldsymbol{\Sigma}_1=\boldsymbol{\Sigma}_2=\cdots=\boldsymbol{\Sigma}_k=\boldsymbol{\Sigma}$，则上述判别规则可简化。

$$d^2(\mathbf{x},\pi_i)=(\mathbf{x}-\boldsymbol{\mu}_i)'\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}_i)=\mathbf{x}'\boldsymbol{\Sigma}^{-1}\mathbf{x}-2\boldsymbol{\mu}_i'\boldsymbol{\Sigma}^{-1}\mathbf{x}+\boldsymbol{\mu}_i'\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_i$$

$$=\mathbf{x}'\boldsymbol{\Sigma}^{-1}\mathbf{x}-2(\boldsymbol{I}_i'\mathbf{x}+c_i)$$

其中$\boldsymbol{I}_i=\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_i,c_i=-\frac{1}{2}\boldsymbol{\mu}_i'\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_i,i=1,2,\cdots,k$，判别规则简化为

$$\boldsymbol{x} \in \pi_l, \quad \text{若} \boldsymbol{I}_l'\boldsymbol{x} + c_l = \max_{1 \le i \le k}\left(\boldsymbol{I}_i'\boldsymbol{x} + c_i\right) \tag{5.2.14}$$

这里$\boldsymbol{I}_i'\boldsymbol{x}+c_i$为线性判别函数。

- 当组数k=2时，可将上式写成

$$\begin{cases} \boldsymbol{x} \in \pi_1, & \text{若} \boldsymbol{I}_1'\boldsymbol{x} + c_1 \ge \boldsymbol{I}_2'\boldsymbol{x} + c_2 \\ \boldsymbol{x} \in \pi_2, & \text{若} \boldsymbol{I}_1'\boldsymbol{x} + c_1 < \boldsymbol{I}_2'\boldsymbol{x} + c_2 \end{cases} \tag{5.2.15}$$

- 它等价于(5.2.3)式的判别规则：

$$\begin{cases} \boldsymbol{x} \in \pi_1, & \text{若} W(\boldsymbol{x}) \ge 0 \\ \boldsymbol{x} \in \pi_2, & \text{若} W(\boldsymbol{x}) < 0 \end{cases}$$

因为

$$W(\boldsymbol{x}) = \left(\boldsymbol{I}_1'\boldsymbol{x} + c_1\right) - \left(\boldsymbol{I}_2'\boldsymbol{x} + c_2\right)$$

- 实践中$\boldsymbol{\mu}_1,\boldsymbol{\mu}_2,\cdots,\boldsymbol{\mu}_k$和$\boldsymbol{\Sigma}_1,\boldsymbol{\Sigma}_2,\cdots,\boldsymbol{\Sigma}_k$一般都是未知的，它们的值可由相应的样本估计值代替。设$\boldsymbol{x}_{i1},\boldsymbol{x}_{i2},\cdots,\boldsymbol{x}_{in_i}$是从组$\pi_i$中抽取的一个样本，则$\boldsymbol{\mu}_i$可估计为

$$\bar{\boldsymbol{x}}_i=\frac{1}{n_i}\sum_{j=1}^{n_i}\boldsymbol{x}_{ij}$$

$(i=1,2,\cdots,k)$。

$\boldsymbol{\Sigma}_1=\boldsymbol{\Sigma}_2=\cdots=\boldsymbol{\Sigma}_k=\boldsymbol{\Sigma}$的情形

- $\boldsymbol{\Sigma}$的联合无偏估计为

$$\boldsymbol{S}_p=\frac{1}{n-k}\sum_{i=1}^{k}\left(n_i-1\right)\boldsymbol{S}_i$$

其中$n=n_1+n_2+\cdots+n_k$，$n-k\geq p$，$\boldsymbol{S}_i=\frac{1}{n_i-1}\sum_{j=1}^{n_i}(\boldsymbol{x}_{ij}-\bar{\boldsymbol{x}}_i)(\boldsymbol{x}_{ij}-\bar{\boldsymbol{x}}_i)'$

为第i组的样本协方差矩阵。

- 实际应用中使用的判别规则是

$$\boldsymbol{x}\in\pi_l,\quad 若\hat{\boldsymbol{I}}_l'\boldsymbol{x}+\hat{c}_l=\max_{1\leq i\leq k}\left(\hat{\boldsymbol{I}}_i'\boldsymbol{x}+\hat{c}_i\right)\qquad(5.2.17)$$

其中$\hat{\boldsymbol{I}}_i=\boldsymbol{S}_p^{-1}\bar{\boldsymbol{x}}_i,\hat{c}_i=-\frac{1}{2}\bar{\boldsymbol{x}}_i'\boldsymbol{S}_p^{-1}\bar{\boldsymbol{x}}_i,i=1,2,\cdots,k$。

$\boldsymbol{\Sigma}_1,\boldsymbol{\Sigma}_2,\cdots,\boldsymbol{\Sigma}_k$不全相等的情形

- $\boldsymbol{\Sigma}_i$可估计为$\boldsymbol{S}_i(i=1,2,\cdots,k)$。
- 实际应用中使用的判别规则是

$$\boldsymbol{x}\in\pi_l,\quad \text{若}\,\hat{d}^2(\boldsymbol{x},\pi_l)=\min_{1\le i\le k}\hat{d}^2(\boldsymbol{x},\pi_i) \qquad (5.2.18)$$

其中

$$\hat{d}^2(\boldsymbol{x},\pi_i)=(\boldsymbol{x}-\bar{\boldsymbol{x}}_i)'\boldsymbol{S}_i^{-1}(\boldsymbol{x}-\bar{\boldsymbol{x}}_i),\quad i=1,2,\cdots,k$$

判别分类是否有效

- 除非各组均值向量之间有明显的差异，否则就不适合作判别分类。
- 在各组数据满足一定的条件下，可先进行多元方差分析。
- 如果检验没有发现均值间有显著差异，则此时再作判别分类将是白费精力。
- 如果检验结果有显著差异，则可考虑再进行判别分类，但并不意味着所作的判别一定有效，最终还得看一下误判概率。

采用线性还是二次判别函数的策略

- (1)一般而言，如果各组的样本容量普遍较小，则选择线性判别函数应是一个较好的策略。相反地，如果各组的样本容量都非常大，则更倾向于采用二次判别函数。
- (2)对$\boldsymbol{\Sigma}_1,\boldsymbol{\Sigma}_2,\cdots,\boldsymbol{\Sigma}_k$作齐次性检验，即检验假设

$$H_0:\boldsymbol{\Sigma}_1=\boldsymbol{\Sigma}_2=\cdots=\boldsymbol{\Sigma}_k,\quad H_1:\boldsymbol{\Sigma}_1,\boldsymbol{\Sigma}_2,\cdots,\boldsymbol{\Sigma}_k\text{不全相等}$$

- 即使检验所需的正态性假定能够满足，检验的结果也只能作为重要的参考依据，而不宜作为决定性的依据，最终还是应视具体的情况而定。

- (3)我们有时也凭直觉判断一下计算出的$\boldsymbol{S}_1,\boldsymbol{S}_2,\cdots,\boldsymbol{S}_k$是否比较接近，以决定是否应假定各组的协方差矩阵相等。
- (4)如果对使用线性还是二次判别函数拿不准，则可以同时采用这两种方法分别进行判别，然后用交叉验证法来比较其误判概率的大小，以判断到底采用哪种方法更为合适。但小样本情形下得到的误判概率估计不够可靠。

例5.2.3

- 对破产的企业收集它们在破产前两年的年度财务数据，同时对财务良好的企业也收集同一时期的数据。数据涉及四个变量：x_1=现金流量/总债务，x_2=净收入/总资产，x_3=流动资产/流动债务，以及x_4=流动资产/净销售额。数据列于表5.2.1，I组为破产企业，II组为非破产企业。

表5. 2. 1　　破产状况数据

编号	组别	x_1	x_2	x_3	x_4	编号	组别	x_1	x_2	x_3	x_4
1	I	-0.45	-0.41	1.09	0.45	24	II	0.38	0.11	3.27	0.35
2	I	-0.56	-0.31	1.51	0.16	25	II	0.19	0.05	2.25	0.33
3	I	0.06	0.02	1.01	0.4	26	II	0.32	0.07	4.24	0.63
4	I	-0.07	-0.09	1.45	0.26	27	II	0.31	0.05	4.45	0.69
5	I	-0.1	-0.09	1.56	0.67	28	II	0.12	0.05	2.52	0.69
6	I	-0.14	-0.07	0.71	0.28	29	II	-0.02	0.02	2.05	0.35
7	I	0.04	0.01	1.5	0.71	30	II	0.22	0.08	2.35	0.4
8	I	-0.07	-0.06	1.37	0.4	31	II	0.17	0.07	1.8	0.52
9	I	0.07	-0.01	1.37	0.34	32	II	0.15	0.05	2.17	0.55
10	I	-0.14	-0.14	1.42	0.43	33	II	-0.1	-0.01	2.5	0.58
11	I	-0.23	-0.3	0.33	0.18	34	II	0.14	-0.03	0.46	0.26
12	I	0.07	0.02	1.31	0.25	35	II	0.14	0.07	2.61	0.52
13	I	0.01	0	2.15	0.7	36	II	0.15	0.06	2.23	0.56
14	I	-0.28	-0.23	1.19	0.66	37	II	0.16	0.05	2.31	0.2
15	I	0.15	0.05	1.88	0.27	38	II	0.29	0.06	1.84	0.38
16	I	0.37	0.11	1.99	0.38	39	II	0.54	0.11	2.33	0.48
17	I	-0.08	-0.08	1.51	0.42	40	II	-0.33	-0.09	3.01	0.47
18	I	0.05	0.03	1.68	0.95	41	II	0.48	0.09	1.24	0.18
19	I	0.01	0	1.26	0.6	42	II	0.56	0.11	4.29	0.44
20	I	0.12	0.11	1.14	0.17	43	II	0.2	0.08	1.99	0.3
21	I	-0.28	-0.27	1.27	0.51	44	II	0.47	0.14	2.92	0.45
22	II	0.51	0.1	2.49	0.54	45	II	0.17	0.04	2.45	0.14
23	II	0.08	0.02	2.01	0.53	46	II	0.58	0.04	5.06	0.13

- 使用线性判别函数进行判别

$$\bar{\boldsymbol{x}}_1=\begin{pmatrix}-0.0690\\-0.0814\\1.3667\\0.4376\end{pmatrix},\quad \bar{\boldsymbol{x}}_2=\begin{pmatrix}0.2352\\0.0556\\2.5936\\0.4268\end{pmatrix}$$

$$20\boldsymbol{S}_1=\begin{pmatrix}0.8826&0.5695&0.6899&0.0829\\0.5695&0.4201&0.5204&0.0688\\0.6899&0.5204&3.2861&0.6556\\0.0829&0.0688&0.6556&0.8916\end{pmatrix}$$

$$24\boldsymbol{S}_2=\begin{pmatrix}1.1292&0.2042&1.7983&-0.1609\\0.2042&0.0570&0.2060&0.0044\\1.7983&0.2060&25.1226&0.7832\\-0.1609&0.0044&0.7832&0.6331\end{pmatrix}$$

$\boldsymbol{\Sigma}$的联合估计为

$$\boldsymbol{S}_p=\frac{1}{44}(20\boldsymbol{S}_1+24\boldsymbol{S}_2)=\begin{pmatrix} 0.0457 & 0.0176 & 0.0566 & -0.0018 \\ 0.0176 & 0.0108 & 0.0165 & 0.0017 \\ 0.0566 & 0.0165 & 0.6457 & 0.0327 \\ -0.0018 & 0.0017 & 0.0327 & 0.0347 \end{pmatrix}$$

$$\boldsymbol{S}_p^{-1}=\begin{pmatrix} 67.9692 & -106.2364 & -3.8556 & 12.2182 \\ -106.2364 & 262.2058 & 3.6899 & -21.5137 \\ -3.8556 & 3.6899 & 1.9020 & -2.1693 \\ 12.2182 & -21.5137 & -2.1693 & 32.5632 \end{pmatrix}$$

$$\hat{\boldsymbol{I}}_1=\boldsymbol{S}_p^{-1}\overline{\boldsymbol{x}}_1=\begin{pmatrix} 4.035 \\ -18.387 \\ 1.616 \\ 12.194 \end{pmatrix},\quad \hat{\boldsymbol{I}}_2=\boldsymbol{S}_p^{-1}\overline{\boldsymbol{x}}_2=\begin{pmatrix} 5.295 \\ -10.020 \\ 3.306 \\ 9.949 \end{pmatrix}$$

$$\hat{c}_1=-\frac{1}{2}\bar{\boldsymbol{x}}_1'\boldsymbol{S}_p^{-1}\bar{\boldsymbol{x}}_1=-4.382,\quad \hat{c}_2=-\frac{1}{2}\bar{\boldsymbol{x}}_2'\boldsymbol{S}_p^{-1}\bar{\boldsymbol{x}}_2=-6.754$$

于是

$$\hat{\boldsymbol{I}}_1'\boldsymbol{x}+\hat{c}_1=4.035x_1-18.387x_2+1.616x_3+12.194x_4-4.382$$

$$\hat{\boldsymbol{I}}_2'\boldsymbol{x}+\hat{c}_2=5.295x_1-10.020x_2+3.306x_3+9.949x_4-6.754$$

对某个未判企业$\boldsymbol{x}=(-0.16, -0.10, 1.45, 0.51)'$，计算得

$$\hat{\boldsymbol{I}}_1'\boldsymbol{x}+\hat{c}_1=5.373,\quad \hat{\boldsymbol{I}}_2'\boldsymbol{x}+\hat{c}_2=3.268$$

按线性判别函数规则，该企业被判为破产企业。

表5. 2. 3 判别情况

真实组 \ 判别为	I	II
I	18	3
II	1	24

- 在表5.2.3中，估计的误判概率为

$$\hat{P}(2|1)=\frac{n(2|1)}{n_1}=\frac{3}{21}=0.143,\quad \hat{P}(1|2)=\frac{n(1|2)}{n_2}=\frac{1}{25}=0.04$$

- 使用交叉验证法，判别情况列于表5.2.4。

表5. 2. 4 判别情况

真实组 \ 判别为	I	II
I	18	3
II	2	23

- 在表5.2.4中，估计的误判概率为

$$\hat{P}(2|1)=\frac{3}{21}=0.143,\quad \hat{P}(1|2)=\frac{2}{25}=0.08$$

- 如果使用二次判别函数进行判别，则由回代法算出的误判率为

$$\hat{P}(2|1)=\frac{2}{21}=0.095,\quad \hat{P}(1|2)=\frac{1}{25}=0.04$$

由交叉验证法估算出的误判概率为

$$\hat{P}(2|1)=\frac{4}{21}=0.190,\quad \hat{P}(1|2)=\frac{1}{25}=0.04$$

§5.3 贝叶斯判别

- 一、最大后验概率法
- 二、最小期望误判代价法

距离判别不合适的一个例子

- 研究的指标是英语六级考试成绩（满分为710分）
- π_1（校研究生组）：N_1=2000, μ_1=500
  π_2（校本科生组）：N_2=8000, μ_2=400
- 研究生组中$x \geq 500$的有1000人，
  本科生组中$x \geq 500$的有2000人。
- 某该校学生的x=500，试判别该生归属哪一组。
- 距离判别显然不妥，应考虑利用先验概率：

$$p_1 = \frac{2000}{10000} = 0.2, \quad p_2 = \frac{8000}{10000} = 0.8$$

一、最大后验概率法

- 设有k个组$\pi_1, \pi_2,\cdots, \pi_k$，且组$\pi_i$的概率密度为$f_i(\boldsymbol{x})$，样品$\boldsymbol{x}$来自组$\pi_i$的先验概率为$p_i$,i=1,2,⋯,k，满足$p_1+p_2+\cdots+p_k=1$。则$\boldsymbol{x}$属于π_i的后验概率为

$$P(\pi_i \mid \boldsymbol{x}) = \frac{p_i f_i(\boldsymbol{x})}{\sum_{j=1}^{k} p_j f_j(\boldsymbol{x})}, \quad i = 1, 2, \cdots, k$$

- 最大后验概率法是采用如下的判别规则：

$$\boldsymbol{x} \in \pi_l, \quad 若 P(\pi_l \mid \boldsymbol{x}) = \max_{1 \le i \le k} P(\pi_i \mid \boldsymbol{x}) \qquad (5.3.2)$$

❖ 例5.3.1　设有π_1,π_2和π_3三个组，欲判别某样品$\boldsymbol{x}_0$属于何组，已知p_1=0.05，p_2=0.65，p_3=0.30，$f_1(\boldsymbol{x}_0)$=0.10, $f_2(\boldsymbol{x}_0)$=0.63，$f_3(\boldsymbol{x}_0)$=2.4。现计算$\boldsymbol{x}_0$属于各组的后验概率如下：

$$P(\pi_1|\boldsymbol{x}_0)=\frac{p_1f_1(\boldsymbol{x}_0)}{\sum_{i=1}^{3}p_if_i(\boldsymbol{x}_0)}=\frac{0.05\times0.10}{0.05\times0.10+0.65\times0.63+0.30\times2.4}$$

$$=\frac{0.005}{1.1345}=0.004$$

$$P(\pi_2|\boldsymbol{x}_0)=\frac{p_2f_2(\boldsymbol{x}_0)}{\sum_{i=1}^{3}p_if_i(\boldsymbol{x}_0)}=\frac{0.65\times0.63}{1.1345}=0.361$$

$$P(\pi_3|\boldsymbol{x}_0)=\frac{p_3f_3(\boldsymbol{x}_0)}{\sum_{i=1}^{3}p_if_i(\boldsymbol{x}_0)}=\frac{0.30\times2.4}{1.1345}=0.635$$

所以应将$\boldsymbol{x}_0$判为组π_3。

皆为正态组的情形

- 设$\pi_i \sim N_p(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$，$\boldsymbol{\Sigma}_i>0$, $i=1,2,\cdots,k$。这时，组π_i的概率密度为

$$f_i(\boldsymbol{x})=(2\pi)^{-p/2}|\boldsymbol{\Sigma}_i|^{-1/2}\exp[-0.5d^2(\boldsymbol{x},\pi_i)]$$

其中

$$d^2(\boldsymbol{x},\pi_i)=(\boldsymbol{x}-\boldsymbol{\mu}_i)'\boldsymbol{\Sigma}_i^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_i)$$

是$\boldsymbol{x}$到π_i的平方马氏距离。

- 以下各情形下后验概率的具体计算公式。

- 当$p_1=p_2=\cdots=p_k=1/k$，$\boldsymbol{\Sigma}_1=\boldsymbol{\Sigma}_2=\cdots=\boldsymbol{\Sigma}_k=\boldsymbol{\Sigma}$时，

$$P(\pi_i \mid \boldsymbol{x})=\frac{\exp\left[-\frac{1}{2}d^2(\boldsymbol{x},\pi_i)\right]}{\sum_{j=1}^{k}\exp\left[-\frac{1}{2}d^2(\boldsymbol{x},\pi_j)\right]}$$

- 当$p_1=p_2=\cdots=p_k=1/k$，而$\boldsymbol{\Sigma}_1,\boldsymbol{\Sigma}_2,\cdots,\boldsymbol{\Sigma}_k$不全相等时，

$$P(\pi_i \mid \boldsymbol{x}) = \frac{\exp\left\{-\frac{1}{2}\left[d^2(\boldsymbol{x},\pi_i)+\ln|\boldsymbol{\Sigma}_i|\right]\right\}}{\sum_{j=1}^{k}\exp\left\{-\frac{1}{2}\left[d^2(\boldsymbol{x},\pi_j)+\ln|\boldsymbol{\Sigma}_j|\right]\right\}}$$

- 当$\boldsymbol{\Sigma}_1=\boldsymbol{\Sigma}_2=\cdots=\boldsymbol{\Sigma}_k=\boldsymbol{\Sigma}$，而$p_1,p_2,\cdots,p_k$不全相等时，

$$P(\pi_i \mid \boldsymbol{x}) = \frac{\exp\left\{-\frac{1}{2}\left[d^2(\boldsymbol{x},\pi_i)-2\ln p_i\right]\right\}}{\sum_{j=1}^{k}\exp\left\{-\frac{1}{2}\left[d^2(\boldsymbol{x},\pi_j)-2\ln p_j\right]\right\}}$$

- 当$p_1,p_2,\cdots,p_k$不全相等，$\boldsymbol{\Sigma}_1,\boldsymbol{\Sigma}_2,\cdots,\boldsymbol{\Sigma}_k$也不全相等时，

$$P(\pi_i \mid \boldsymbol{x}) = \frac{\exp\left\{-\frac{1}{2}\left[d^2(\boldsymbol{x},\pi_i)+\ln|\boldsymbol{\Sigma}_i|-2\ln p_i\right]\right\}}{\sum_{j=1}^{k}\exp\left\{-\frac{1}{2}\left[d^2(\boldsymbol{x},\pi_j)+\ln|\boldsymbol{\Sigma}_j|-2\ln p_j\right]\right\}}$$

- 上述各情形的后验概率可统一表达为

$$P(\pi_i \mid \boldsymbol{x}) = \frac{\exp\left[-\frac{1}{2}D^2(\boldsymbol{x},\pi_i)\right]}{\sum_{j=1}^{k}\exp\left[-\frac{1}{2}D^2(\boldsymbol{x},\pi_j)\right]}, \quad i=1,2,\cdots,k$$

其中

$$D^2(\boldsymbol{x},\pi_i)=d^2(\boldsymbol{x},\pi_i)+g_i+h_i$$

$$g_i = \begin{cases} \ln\left|\boldsymbol{\Sigma}_i\right|, & \text{若}\boldsymbol{\Sigma}_1,\boldsymbol{\Sigma}_2,\cdots,\boldsymbol{\Sigma}_k\text{不全相等} \\ 0, & \text{若}\boldsymbol{\Sigma}_1=\boldsymbol{\Sigma}_2=\cdots=\boldsymbol{\Sigma}_k=\boldsymbol{\Sigma} \end{cases}$$

$$h_i = \begin{cases} -2\ln p_i, & \text{若}p_1,p_2,\cdots,p_k\text{不全相等} \\ 0, & \text{若}p_1=p_2=\cdots=p_k=\frac{1}{k} \end{cases}$$

$$i=1,2,\cdots,k$$

- 称$D^2(\boldsymbol{x}, \pi_i)$为$\boldsymbol{x}$到$\pi_i$的广义平方距离。在正态性假定下，上述判别规则也可等价地表达为

$$\boldsymbol{x} \in \pi_l, \quad 若D^2(\boldsymbol{x}, \pi_l) = \min_{1\le i\le k} D^2(\boldsymbol{x}, \pi_i)$$

- 当$\boldsymbol{\Sigma}_1=\boldsymbol{\Sigma}_2=\cdots=\boldsymbol{\Sigma}_k=\boldsymbol{\Sigma}$时，上述后验概率公式可简化为

$$P(\pi_i \mid \boldsymbol{x}) = \frac{\exp(\boldsymbol{I}_i'\boldsymbol{x} + c_i + \ln p_i)}{\sum_{j=1}^{k} \exp(\boldsymbol{I}_j'\boldsymbol{x} + c_j + \ln p_j)}, \quad i = 1, 2, \cdots, k$$

其中$\boldsymbol{I}_i=\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_i$，$c_i=-0.5\boldsymbol{\mu}_i'\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_i$, $i=1,2,\cdots,k$。此时，判别规则等价于

$$\boldsymbol{x} \in \pi_l, \quad 若\boldsymbol{I}_l'\boldsymbol{x} + c_l + \ln p_l = \max_{1\le i\le k}(\boldsymbol{I}_i'\boldsymbol{x} + c_i + \ln p_i)$$

- 如果我们对$\boldsymbol{x}$来自哪一组的先验信息一无所知或难以确定，则一般可取$p_1=p_2=\cdots=p_k=1/k$。这时，判别规则简化为

$$\boldsymbol{x} \in \pi_l, \quad 若\boldsymbol{I}_l'\boldsymbol{x} + c_l = \max_{1\le i\le k}(\boldsymbol{I}_i'\boldsymbol{x} + c_i)$$

- 实际应用中，以上各式中的$\boldsymbol{\mu}_i$和$\boldsymbol{\Sigma}_i(i=1,2,\cdots,k)$一般都是未知的，需用相应的样本估计值代替。
- 例5.3.2 在例5.2.3中，已知破产企业所占的比例约为10%，即可取$p_1=0.1$，$p_2=0.9$，假定两组均为正态，且$\boldsymbol{\Sigma}_1=\boldsymbol{\Sigma}_2=\boldsymbol{\Sigma}$，则未判企业$\boldsymbol{x}=(-0.16, -0.10, 1.45, 0.51)'$的后验概率为

$$P(\pi_1|\boldsymbol{x})=\frac{\exp\left(\hat{\boldsymbol{I}}_1'\boldsymbol{x}+\hat{c}_1+\ln p_1\right)}{\exp\left(\hat{\boldsymbol{I}}_1'\boldsymbol{x}+\hat{c}_1+\ln p_1\right)+\exp\left(\hat{\boldsymbol{I}}_2'\boldsymbol{x}+\hat{c}_2+\ln p_2\right)}$$

$$=\frac{\exp(5.373+\ln 0.1)}{\exp(5.373+\ln 0.1)+\exp(3.268+\ln 0.9)}=\frac{e^{3.07}}{e^{3.07}+e^{3.163}}=\frac{21.542}{45.183}=0.477$$

$$P(\pi_2|\boldsymbol{x})=\frac{\exp\left(\hat{\boldsymbol{I}}_2'\boldsymbol{x}+\hat{c}_2+\ln p_2\right)}{\exp\left(\hat{\boldsymbol{I}}_1'\boldsymbol{x}+\hat{c}_1+\ln p_1\right)+\exp\left(\hat{\boldsymbol{I}}_2'\boldsymbol{x}+\hat{c}_2+\ln p_2\right)}=\frac{23.641}{45.183}=0.523$$

由于$P(\pi_1|\boldsymbol{x})<P(\pi_2|\boldsymbol{x})$，所以该企业被判为非破产企业，这与例5.2.3的结果正好相反，这正是先验概率的作用结果。

二、最小期望误判代价法

- 例子：

π_1：合格的药，π_2：不合格的药

对于新样品$\boldsymbol{x}$

$$P(\pi_1 \mid \boldsymbol{x}) = 0.6, \quad P(\pi_2 \mid \boldsymbol{x}) = 0.4$$

该问题中，两种误判造成的损失一般是明显不同的，只是根据后验概率的大小进行判别是不太合适的。

- 1.两组的一般情形
- 2.两个正态组的情形
- 3.多组的情形

1.两组的一般情形

- 设组π_1和π_2的概率密度函数分别为$f_1(\boldsymbol{x})$和$f_2(\boldsymbol{x})$，组π_1和π_2的先验概率分别为p_1和p_2，$p_1+p_2=1$。又设将来自π_i的$\boldsymbol{x}$判为π_l的代价为$c(l|i)$, $l,i=1,2$，可用代价矩阵表示为

		判别为	
		π_1	π_2
真实组	π_1	0	$c(2\|1)$
	π_2	$c(1\|2)$	0

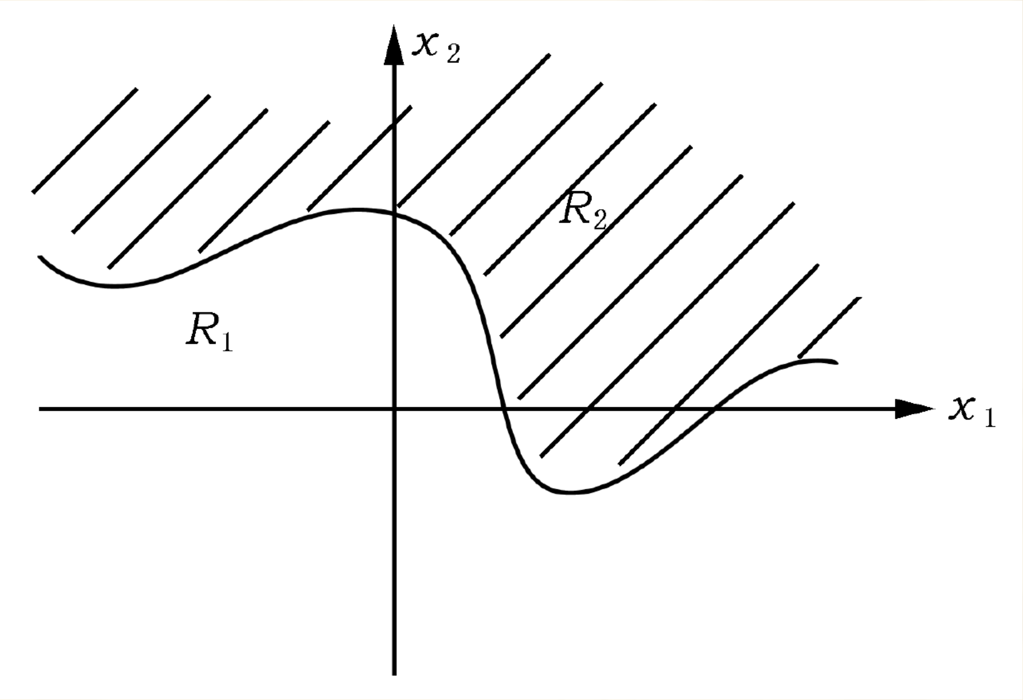
- 对于给定的判别规则，令

$$R_1=\{\boldsymbol{x}:\text{判别归属}\pi_1\},\quad R_2=\{\boldsymbol{x}:\text{判别归属}\pi_2\}$$

显然

$$R_1\cup R_2=\Omega,\quad R_1\cap R_2=\varnothing$$

$$\boldsymbol{x}\in R_1\Leftrightarrow\text{判}\boldsymbol{x}\in\pi_1\quad,\quad\boldsymbol{x}\in R_2\Leftrightarrow\text{判}\boldsymbol{x}\in\pi_2$$



- 将π_1中的样品$\boldsymbol{x}$误判到π_2的条件概率为

$$P(2|1)=P(\boldsymbol{x}\in R_2|\ \boldsymbol{x}\in\pi_1)=\int_{R_2} f_1(\boldsymbol{x})\mathrm{d}\boldsymbol{x}$$

将π_2中的样品$\boldsymbol{x}$误判到π_1的条件概率为

$$P(1|2)=P(\boldsymbol{x}\in R_1|\ \boldsymbol{x}\in\pi_2)=\int_{R_1} f_2(\boldsymbol{x})\mathrm{d}\boldsymbol{x}$$

- 期望误判代价

$$
\begin{aligned}
ECM &= E\left[c(l\,|\,i)\right] \\
&= c(1|1)P(i=1,l=1)+c(2|1)P(i=1,l=2) \\
&\quad +c(1|2)P(i=2,l=1)+c(2|2)P(i=2,l=2) \\
&= c(2|1)P(\boldsymbol{x}\in\pi_1,\boldsymbol{x}\in R_2)+c(1|2)P(\boldsymbol{x}\in\pi_2,\boldsymbol{x}\in R_1) \\
&= c(2|1)P(\boldsymbol{x}\in R_2\,|\,\boldsymbol{x}\in\pi_1)P(\boldsymbol{x}\in\pi_1) \\
&\quad +c(1|2)P(\boldsymbol{x}\in R_1\,|\,\boldsymbol{x}\in\pi_2)P(\boldsymbol{x}\in\pi_2) \\
&= c(2|1)P(2|1)p_1+c(1|2)P(1|2)p_2
\end{aligned}
$$

- 最小期望误判代价法采用的是使*ECM*达到最小的判别规则：

$$
\begin{cases}
\boldsymbol{x}\in\pi_1, & 若\ \dfrac{f_1(\boldsymbol{x})}{f_2(\boldsymbol{x})}\geq\dfrac{c(1|2)p_2}{c(2|1)p_1} \\[2ex]
\boldsymbol{x}\in\pi_2, & 若\ \dfrac{f_1(\boldsymbol{x})}{f_2(\boldsymbol{x})}<\dfrac{c(1|2)p_2}{c(2|1)p_1}
\end{cases}
\qquad (5.3.13)
$$

误判代价之比

- 最小*ECM*规则需要三个比值：
  密度函数比、误判代价比和先验概率比。
- 在这些比值中，误判代价比最富有实际意义。
- 例1

  π_1：应该做手术 ，π_2：不应该做手术
- 例2

  π_1：硕士毕业后应继续攻读博士

  π_2：硕士毕业后应直接找工作

❖ 例5.3.3　设组π_1和π_2的概率密度函数分别为$f_1(\boldsymbol{x})$和$f_2(\boldsymbol{x})$，又知$c(1|2)$=12个单位，$c(2|1)$=4个单位，根据以往经验给出p_1=0.6，p_2=0.4，则最小*ECM*判别规则为

$$\begin{cases} \boldsymbol{x} \in \pi_1, & 若 \dfrac{f_1(\boldsymbol{x})}{f_2(\boldsymbol{x})} \geq \dfrac{12 \times 0.4}{4 \times 0.6} = 2 \\ \boldsymbol{x} \in \pi_2, & 若 \dfrac{f_1(\boldsymbol{x})}{f_2(\boldsymbol{x})} < \dfrac{12 \times 0.4}{4 \times 0.6} = 2 \end{cases}$$

假定在一个新样品$\boldsymbol{x}_0$处算得$f_1(\boldsymbol{x}_0)$=0.36，$f_2(\boldsymbol{x}_0)$=0.24，于是

$$\frac{f_1(\boldsymbol{x}_0)}{f_2(\boldsymbol{x}_0)} = \frac{0.36}{0.24} = 1.5 < 2$$

因此，判$\boldsymbol{x}_0$来自组π_2。

(5.3.13)式的一些特殊情形

- (1)当p_1=p_2=0.5时，(5.3.13)式简化为

$$\begin{cases} \boldsymbol{x} \in \pi_1, & 若\dfrac{f_1(\boldsymbol{x})}{f_2(\boldsymbol{x})} \ge \dfrac{c(1|2)}{c(2|1)} \\ \boldsymbol{x} \in \pi_2, & 若\dfrac{f_1(\boldsymbol{x})}{f_2(\boldsymbol{x})} < \dfrac{c(1|2)}{c(2|1)} \end{cases}$$

- 实际应用中，如果先验概率难以给出，则它们通常被取成相等。

- (2) 当$c(1|2)= c(2|1)$时，(5.3.13)式简化为

$$\begin{cases} \boldsymbol{x} \in \pi_1, & 若 p_1 f_1(\boldsymbol{x}) \geq p_2 f_2(\boldsymbol{x}) \\ \boldsymbol{x} \in \pi_2, & 若 p_1 f_1(\boldsymbol{x}) < p_2 f_2(\boldsymbol{x}) \end{cases} \qquad (5.3.15)$$

- 该式等价于组数k=2时的 (5.3.2)式，即

$$\begin{cases} \boldsymbol{x} \in \pi_1, & 若 P(\pi_1 | \boldsymbol{x}) \geq P(\pi_2 | \boldsymbol{x}) \\ \boldsymbol{x} \in \pi_2, & 若 P(\pi_1 | \boldsymbol{x}) < P(\pi_2 | \boldsymbol{x}) \end{cases}$$

- 实践中，若误判代价比无法确定，则通常取比值为1。
- 记$c(1|2)=c(2|1)=c$，有

$$ECM=c[p_1P(2|1)+p_2P(1|2)]$$

- 此时的最小期望误判代价判别规则即为最小总误判概率判别规则。
- 总的误判概率$=p_1P(2|1)+p_2P(1|2)$

❖ (3)当$\dfrac{p_1}{p_2}=\dfrac{c(1|2)}{c(2|1)}$ [通常的情况是，$p_1=p_2=0.5$且$c(1|2)=c(2|1)$]

时，(5.3.13)式可进一步简化为

$$\begin{cases} \boldsymbol{x}\in\pi_1, & 若f_1(\boldsymbol{x})\ge f_2(\boldsymbol{x}) \\ \boldsymbol{x}\in\pi_2, & 若f_1(\boldsymbol{x})< f_2(\boldsymbol{x}) \end{cases} \qquad (5.3.17)$$

❖ 判别规则(5.3.17)可看成是$c(2|1)p_1=c(1|2)p_2$时的判别规则(5.3.13)，从而它可使

$$ECM=b[P(2|1)+P(1|2)]$$

达到最小，其中$b=c(2|1)p_1=c(1|2)p_2$是一个不依赖于判别规则的常数，故判别规则(5.3.17)可使两个误判概率之和$P(2|1)+P(1|2)$达到最小，或者说可使平均误判概率$0.5P(2|1)+0.5P(1|2)$达到最小，这个平均误判概率也是当$p_1=p_2=0.5$时的总误判概率。

- 例5.3.4 在例5.2.2中，按（5.3.17）式，判别规则可写为

$$\begin{cases} x \in \pi_1, & 若 x \in R_1 \\ x \in \pi_2, & 若 x \in R_2 \end{cases}$$

- 这是一个在两个误判概率之和达到最小的意义上的最优判别规则。

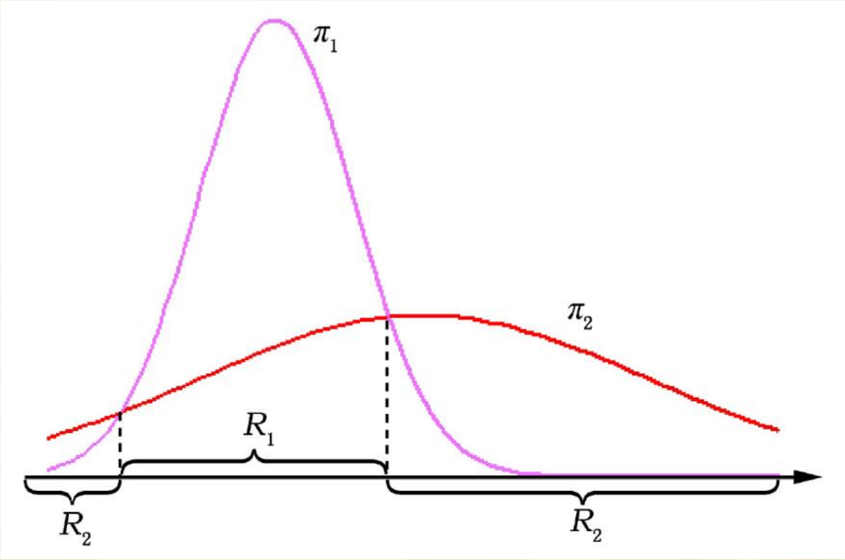


图5. 3. 2 按(5. 3. 17)式判别的组归属区域

2.两个正态组的情形

- 假定$\pi_i \sim N_p(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$, $\boldsymbol{\Sigma}_i>0$, $i=1,2$。
- 当$\boldsymbol{\Sigma}_1=\boldsymbol{\Sigma}_2=\boldsymbol{\Sigma}$时，(5.3.13)式可具体写成

$$\begin{cases} \boldsymbol{x} \in \pi_1, & 若\boldsymbol{a}'(\boldsymbol{x}-\bar{\boldsymbol{\mu}}) \ge \ln\left[\dfrac{c(1|2)p_2}{c(2|1)p_1}\right] \\ \boldsymbol{x} \in \pi_2, & 若\boldsymbol{a}'(\boldsymbol{x}-\bar{\boldsymbol{\mu}}) < \ln\left[\dfrac{c(1|2)p_2}{c(2|1)p_1}\right] \end{cases}$$

其中$\boldsymbol{a}=\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)$，$\bar{\boldsymbol{\mu}}=\dfrac{1}{2}(\boldsymbol{\mu}_1+\boldsymbol{\mu}_2)$。

- 在$p_1=p_2$，$c(1|2)=c(2|1)$的条件下上式将退化为(5.2.3)式。

- 在两组皆为正态组且协差阵相等的情形下，距离判别(5.2.3)等价于不考虑先验概率和误判代价[相当于$p_1=p_2$，$c(1|2)=c(2|1)$]时的贝叶斯判别。
- 重要结论：在上述情形下，判别规则(5.2.3)在使两个误判概率之和（或平均误判概率）达到最小的意义上是最优的。
- 实践中，因未知参数需用样本值替代，故实际所使用的判别规则(5.2.5)只是渐近最优的。
- 当$\boldsymbol{\Sigma}_1\neq\boldsymbol{\Sigma}_2$时，(5.3.13)式可写为

$$\begin{cases} \boldsymbol{x}\in\pi_1, & \text{若} d^2(\boldsymbol{x},\pi_1)-d^2(\boldsymbol{x},\pi_2)\le 2\ln\left[\dfrac{c(2|1)p_1|\boldsymbol{\Sigma}_2|^{1/2}}{c(1|2)p_2|\boldsymbol{\Sigma}_1|^{1/2}}\right] \\ \boldsymbol{x}\in\pi_2, & \text{若} d^2(\boldsymbol{x},\pi_1)-d^2(\boldsymbol{x},\pi_2)> 2\ln\left[\dfrac{c(2|1)p_1|\boldsymbol{\Sigma}_2|^{1/2}}{c(1|2)p_2|\boldsymbol{\Sigma}_1|^{1/2}}\right] \end{cases}$$

其中$d^2(\boldsymbol{x},\pi_i)=(\boldsymbol{x}-\boldsymbol{\mu}_i)'\boldsymbol{\Sigma}_i^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_i)$, $i=1,2$。

- 在$p_1=p_2$，$c(1|2)=c(2|1)$的条件下上式可简化为

$$\begin{cases} \boldsymbol{x} \in \pi_1, & 若 d^2(\boldsymbol{x}, \pi_1) - d^2(\boldsymbol{x}, \pi_2) \le 2\ln\left(\dfrac{|\boldsymbol{\Sigma}_2|^{1/2}}{|\boldsymbol{\Sigma}_1|^{1/2}}\right) \\ \boldsymbol{x} \in \pi_2, & 若 d^2(\boldsymbol{x}, \pi_1) - d^2(\boldsymbol{x}, \pi_2) > 2\ln\left(\dfrac{|\boldsymbol{\Sigma}_2|^{1/2}}{|\boldsymbol{\Sigma}_1|^{1/2}}\right) \end{cases} \qquad (5.3.20)$$

- 在两组皆为正态组的情形下，判别规则(5.3.20)在使两个误判概率之和（或平均误判概率）达到最小的意义上是最优的。此时，它当然也就优于(5.2.10)式的距离判别。
- 基于二次函数的判别规则相比线性判别规则，其判别效果更依赖于多元正态性的假定。
- 实践中，为了达到较理想的判别效果，需要时可以考虑先将各组的非正态性数据变换成接近正态性的数据，然后再作判别分析。

3.多组的情形

- 设 $f_i(\boldsymbol{x})$为组π_i的概率密度函数，$i=1,2,\cdots,k$。令

  p_i——组π_i的先验概率，$i=1,2,\cdots,k$。

  $c(l|i)$——将来自π_i的$\boldsymbol{x}$判为π_l的代价, $l,i=1,2,\cdots,k$,
对$l=i,c(i|i)=0,i=1,2,\cdots,k$。

  R_l——所有判为π_l的$\boldsymbol{x}$的集合，$l=1,2,\cdots,k$。

  于是

$$P(l\,|\,i)=P\left(\boldsymbol{x}\in R_l\,|\,\boldsymbol{x}\in\pi_i\right)=\int_{R_l} f_i\left(\boldsymbol{x}\right)\mathrm{d}\boldsymbol{x}$$

- 期望误判代价：

$$
\begin{aligned}
ECM &= E\left[c(l|i)\right] \\
&= \sum_{i=1}^{k}\sum_{l=1}^{k} c(l|i) P\left(\mathbf{x}\in\pi_i, \mathbf{x}\in R_l\right) \\
&= \sum_{i=1}^{k}\sum_{l=1}^{k} c(l|i) P\left(\mathbf{x}\in R_l \mid \mathbf{x}\in\pi_i\right) P\left(\mathbf{x}\in\pi_i\right) \\
&= \sum_{i=1}^{k}\sum_{l=1}^{k} c(l|i) P(l|i) p_i = \sum_{i=1}^{k} p_i \sum_{\substack{l=1 \\ l\neq i}}^{k} c(l|i) P(l|i)
\end{aligned}
$$

- 使ECM达到最小的判别规则是

$$
\mathbf{x}\in\pi_l, \quad 若\sum_{\substack{j=1\\ j\neq l}}^{k} p_j c(l|j) f_j(\mathbf{x}) = \min_{1\le i\le k} \sum_{\substack{j=1\\ j\neq i}}^{k} p_j c(i|j) f_j(\mathbf{x})
$$

❖ 例5.3.5 在例5.3.1中，假定误判代价矩阵为

		判别为		
		π_1	π_2	π_3
真实组	π_1	$c(1\|1)=0$	$c(2\|1)=10$	$c(3\|1)=200$
	π_2	$c(1\|2)=20$	$c(2\|2)=0$	$c(3\|2)=100$
	π_3	$c(1\|3)=60$	$c(2\|3)=50$	$c(3\|3)=0$

➢ 现采用最小*ECM*规则进行判别。

l=1： $p_2c(1|2)\,f_2(\boldsymbol{x}_0)+p_3c(1|3)\,f_3(\boldsymbol{x}_0)$

$=0.65\times20\times0.63+0.30\times60\times2.4=51.39$

l=2： $p_1c(2|1)\,f_1(\boldsymbol{x}_0)+p_3c(2|3)\,f_3(\boldsymbol{x}_0)$

$=0.05\times10\times0.10+0.30\times50\times2.4=36.05$

l=3： $p_1c(3|1)\,f_1(\boldsymbol{x}_0)+p_2c(3|2)\,f_2(\boldsymbol{x}_0)$

$=0.05\times200\times0.10+0.65\times100\times0.63=41.95$

由于l=2时为最小值，故将$\boldsymbol{x}_0$判为π_2。

假定所有的误判代价都是相同的，不失一般性，可令$c(l|i)=1$, $l\neq i$, $l,i=1,2,\cdots,k$，则此时

$$ECM=\sum_{i=1}^{k}p_i\sum_{\substack{l=1\\l\neq i}}^{k}P(l|i)=1-\sum_{i=1}^{k}p_iP(i|i)$$

称为总的误判概率。故此时的最小期望误判代价法也可称为最小总误判概率法，并且上式可简化为

$$\boldsymbol{x}\in\pi_l,\quad 若\sum_{\substack{j=1\\j\neq l}}^{k}p_jf_j(\boldsymbol{x})=\min_{1\le i\le k}\sum_{\substack{j=1\\j\neq i}}^{k}p_jf_j(\boldsymbol{x})$$

让$\sum_{j=1}^{k}p_jf_j(\boldsymbol{x})$减去上面等式的两边，即有更简洁的形式：

$$\boldsymbol{x}\in\pi_l,\quad 若p_lf_l(\boldsymbol{x})=\max_{1\le i\le k}p_if_i(\boldsymbol{x})$$

- 它与(5.3.2)式是等价的。因此，此时的最小总误判概率法等同于最大后验概率法，或者说，最大后验概率法可看成是所有误判代价均相同时的最小期望误判代价法。
- 当$p_1=p_2=\cdots=p_k=1/k$时，上式又进一步简化为

$$\boldsymbol{x} \in \pi_l, \quad 若 f_l(\boldsymbol{x}) = \max_{1 \le i \le k} f_i(\boldsymbol{x})$$

该判别规则实际上也是一种极大似然法。

❖ 注　令

B={误判}，A_i={样品来自π_i}，i=1,2,⋯,k

则总的误判概率为

$$P(B)=\sum_{i=1}^{k}P(A_i)P(B|A_i)=\sum_{i=1}^{k}p_i\sum_{\substack{l=1\\l\neq i}}^{k}P(l|i)$$

总的正确判别概率为

$$P(\bar{B})=1-P(B)=1-\sum_{i=1}^{k}p_i\sum_{\substack{l=1\\l\neq i}}^{k}P(l|i)$$

$$=1-\sum_{i=1}^{k}p_i\left[1-P(i|i)\right]=\sum_{i=1}^{k}p_iP(i|i)$$

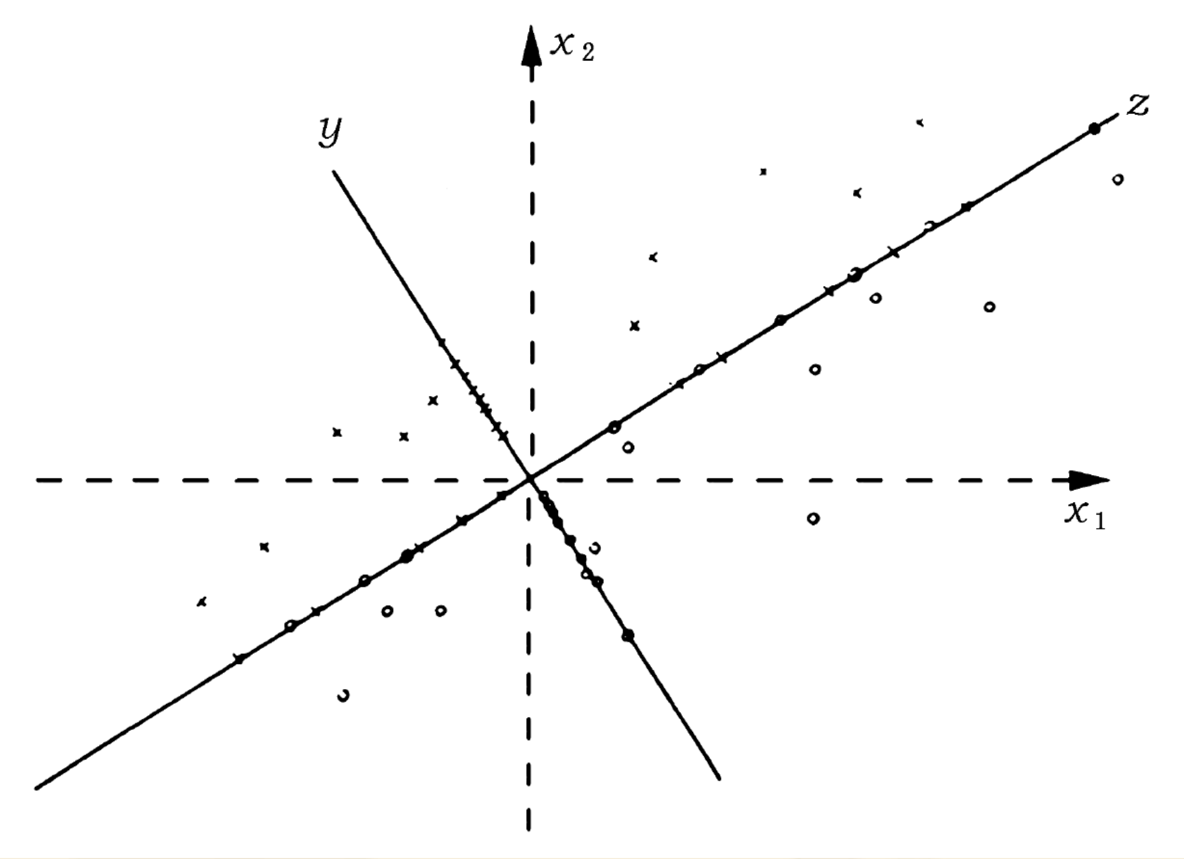
§5.4 费希尔判别

- 一、费希尔判别的基本思想
- 二、费希尔判别函数
- 三、判别函数得分图
- 四、判别规则

一、费希尔判别的基本思想

- 费希尔判别（或称典型判别）的基本思想是投影（或降维）：用p维向量$\boldsymbol{x}=(x_1,x_2,\cdots,x_p)'$的少数几个线性组合（称为费希尔判别函数或典型变量）
$$y_1=\boldsymbol{a}_1'\boldsymbol{x},y_2=\boldsymbol{a}_2'\boldsymbol{x},\cdots,y_r=\boldsymbol{a}_r'\boldsymbol{x}$$
（一般r明显小于p）来代替原始的p个变量$x_1,x_2,\cdots,x_p$，以达到降维的目的，并根据这r个判别函数$y_1,y_2,\cdots,y_r$对样品的归属作出判别或将各组分离。
- 成功的降维将使样品的归类或组的分离更为方便和有效，并且可以对前两个或前三个判别函数作图，从直观的几何图形上区别各组。

一个说明性的二维例子



二、费希尔判别函数

- 设来自组π_i的p维观测值为$\boldsymbol{x}_{ij}$，$j=1,2,\cdots,n_i$，$i=1,2,\cdots,k$，将它们共同投影到某一p维常数向量$\boldsymbol{a}$上，得到的投影点可分别对应线性组合$y_{ij}=\boldsymbol{a}'\boldsymbol{x}_{ij}$，$j=1,2,\cdots,n_i$，$i=1,2,\cdots,k$。

$$\bar{y}_i=\frac{1}{n_i}\sum_{j=1}^{n_i}y_{ij}=\boldsymbol{a}'\bar{\boldsymbol{x}}_i$$

$$\bar{y}=\frac{1}{n}\sum_{i=1}^{k}\sum_{j=1}^{n_i}y_{ij}=\frac{1}{n}\sum_{i=1}^{k}n_i\bar{y}_i=\boldsymbol{a}'\bar{\boldsymbol{x}}$$

式中$n=\sum_{i=1}^{k}n_i,\ \bar{\boldsymbol{x}}_i=\frac{1}{n_i}\sum_{j=1}^{n_i}\boldsymbol{x}_{ij},\ \bar{\boldsymbol{x}}=\frac{1}{n}\sum_{i=1}^{k}n_i\bar{\boldsymbol{x}}_i$。

- 费希尔判别需假定$\boldsymbol{\Sigma}_1=\boldsymbol{\Sigma}_2=\cdots=\boldsymbol{\Sigma}_k=\boldsymbol{\Sigma}$。

三组之间的分离程度

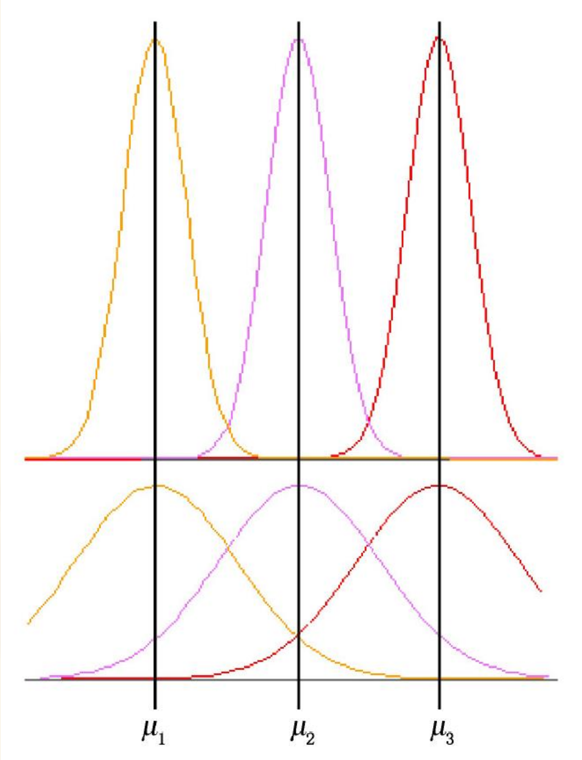


图5.4.2 三组之间的分离程度

- y_{ij}的组间平方和及组内平方和分别为

$$SSTR=\sum_{i=1}^{k}n_i\left(\bar{y}_i-\bar{y}\right)^2=\sum_{i=1}^{k}n_i\left(\boldsymbol{a}'\bar{\boldsymbol{x}}_i-\boldsymbol{a}'\bar{\boldsymbol{x}}\right)^2=\boldsymbol{a}'\boldsymbol{H}\boldsymbol{a}$$

$$SSE=\sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(y_{ij}-\bar{y}_i\right)^2=\sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(\boldsymbol{a}'\boldsymbol{x}_{ij}-\boldsymbol{a}'\bar{\boldsymbol{x}}_i\right)^2=\boldsymbol{a}'\boldsymbol{E}\boldsymbol{a}$$

- 式中

$$\boldsymbol{H}=\sum_{i=1}^{k}n_i\left(\bar{\boldsymbol{x}}_i-\bar{\boldsymbol{x}}\right)\left(\bar{\boldsymbol{x}}_i-\bar{\boldsymbol{x}}\right)'$$

$$\boldsymbol{E}=\sum_{i=1}^{k}\left(n_i-1\right)\boldsymbol{S}_i=\sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(\boldsymbol{x}_{ij}-\bar{\boldsymbol{x}}_i\right)\left(\boldsymbol{x}_{ij}-\bar{\boldsymbol{x}}_i\right)'$$

- 可用来反映y_{ij}的组之间分离程度的一个量是

$$\Delta(\boldsymbol{a})=\frac{SSTR}{SSE}=\frac{\boldsymbol{a}'\boldsymbol{H}\boldsymbol{a}}{\boldsymbol{a}'\boldsymbol{E}\boldsymbol{a}}$$

- 在约束条件$\boldsymbol{a}'\boldsymbol{S}_p\boldsymbol{a}=1$下，寻找$\boldsymbol{a}$，使得$\Delta(\boldsymbol{a})$达到最大，其中 $\boldsymbol{S}_p=\dfrac{1}{n-k}\boldsymbol{E}$ 是$\boldsymbol{\Sigma}$的联合无偏估计。
- 设$\boldsymbol{E}^{-1}\boldsymbol{H}$的全部非零特征值依次为$\lambda_1\geq\lambda_2\geq\cdots\geq\lambda_s>0$，这里$s=\text{rank}(\boldsymbol{H})$，且有

$$s\leq\min(k-1,p)$$

相应的特征向量依次记为$\boldsymbol{t}_1,\boldsymbol{t}_2,\cdots,\boldsymbol{t}_s$（标准化为$\boldsymbol{t}_i'\boldsymbol{S}_p\boldsymbol{t}_i=1$，$i=1,2,\cdots,s$）。

- 当$\boldsymbol{a}_1=\boldsymbol{t}_1$时$\Delta(\boldsymbol{a}_1)$达到最大值$\lambda_1$。所以，选择投影到$\boldsymbol{t}_1$上能使各组的投影点最大限度地分离，称$y_1=\boldsymbol{t}_1'\boldsymbol{x}$为费希尔第一线性判别函数，简称第一判别函数。
- 在许多情况下（如k或p是大的），仅仅使用第一判别函数也许不够，应考虑建立$y_2=\boldsymbol{a}_2'\boldsymbol{x}$，且满足

$$\text{Cov}(y_1,y_2)=\text{Cov}(\boldsymbol{t}_1'\boldsymbol{x},\boldsymbol{a}_2'\boldsymbol{x})=\boldsymbol{t}_1'\boldsymbol{\Sigma}\boldsymbol{a}_2=0$$

- 用$\boldsymbol{S}_p$代替未知的$\boldsymbol{\Sigma}$，于是在约束条件

$$\boldsymbol{t}_1'\boldsymbol{S}_p\boldsymbol{a}_2=0\text{（或}\boldsymbol{t}_1'\boldsymbol{E}\boldsymbol{a}_2=0\text{）}$$

下寻找$\boldsymbol{a}_2$，使得$\Delta(\boldsymbol{a}_2)$达到最大。当$\boldsymbol{a}_2=\boldsymbol{t}_2$时$\Delta(\boldsymbol{a}_2)$达到最大值$\lambda_2$，称$y_2=\boldsymbol{t}_2'\boldsymbol{x}$为第二判别函数。一般地，我们要求第$i$个线性组合$y_i=\boldsymbol{a}_i'\boldsymbol{x}$不重复前$i-1$个判别函数中的信息，即

$$\operatorname{Cov}\left(y_j,y_i\right)=\operatorname{Cov}\left(\boldsymbol{t}_j'\boldsymbol{x},\boldsymbol{a}_i'\boldsymbol{x}\right)=\boldsymbol{t}_j'\boldsymbol{\Sigma}\boldsymbol{a}_i=0,\quad j=1,2,\cdots,i-1$$

- 用$\boldsymbol{S}_p$替代$\boldsymbol{\Sigma}$，上式变为

$$\boldsymbol{t}_j'\boldsymbol{S}_p\boldsymbol{a}_i=0\text{（或}\boldsymbol{t}_j'\boldsymbol{E}\boldsymbol{a}_i=0\text{）},\quad j=1,2,\cdots,i-1$$

- 在上述约束条件下寻找$\boldsymbol{a}_i$，使得$\Delta(\boldsymbol{a}_i)$达到最大。当$\boldsymbol{a}_i=\boldsymbol{t}_i$时$\Delta(\boldsymbol{a}_i)$达到最大值$\lambda_i$，称$y_i=\boldsymbol{t}_i'\boldsymbol{x}$为第$i$判别函数，$i=2,3,\cdots,s$。
- 有时我们也使用中心化的费希尔判别函数，即

$$y_i=\boldsymbol{t}_i'\left(\boldsymbol{x}-\bar{\boldsymbol{x}}\right),\quad i=1,2,\cdots,s$$

式中 $\bar{\boldsymbol{x}}=\dfrac{1}{n}\sum_{i=1}^{k}\sum_{j=1}^{n_i}\boldsymbol{x}_{ij}$ 为k个组的总均值。

费希尔判别函数的特点

- (1)各判别函数都具有单位（联合样本）方差;
- (2)各判别函数彼此之间不相关（确切地说，是彼此之间的联合样本协方差为零）；
- (3)判别函数方向$\boldsymbol{t}_1,\boldsymbol{t}_2,\cdots,\boldsymbol{t}_s$并不正交，但作图时仍将它们画成直角坐标系，虽有些变形，但通常并不严重。
- (4)判别函数不受变量度量单位的影响。

- 组数k=2时只有一个判别函数，k=3时最多只有两个判别函数。
- $\Delta(\mathbf{t}_i)=\lambda_i$表明了$y_i$对分离各组的贡献大小，$y_i$在所有$s$个判别函数中的贡献率为

$$\lambda_i \Big/ \sum_{j=1}^{s} \lambda_j$$

- 而前$r(\le s)$个判别函数$y_1,y_2,\cdots,y_r$的累计贡献率为

$$\sum_{i=1}^{r} \lambda_i \Big/ \sum_{i=1}^{s} \lambda_i$$

它表明了$y_1,y_2,\cdots,y_r$的判别能力。
- 在实际应用中，如果前r个判别函数的累计贡献率已达到了一个较高的比例（如75%~95%），则就采用这r个判别函数进行判别。

三、判别函数得分图

- 为作图的目的，一般取r=2，偶尔取r=3，
- 当取r=2时，可将各样品的两个判别函数得分画成平面直角坐标系上的散点图，用目测法对新样品的归属进行辨别或对来自各组样品的分离情况及结构进行观测评估。
- 当r=3时，可作（三维）旋转图从多角度来辨别新样品的归属或观测评估各组之间的分离效果，但其目测效果一般明显不如r=2时清楚。
- 能够利用降维后生成的图形进行直观判别是费希尔判别的最重要应用，图中常常能清晰地展示出丰富的信息，如发现构成各组的结构、离群样品点或数据中的其他异常情况等。

❖ 例5.4.1　费希尔于1936年发表的鸢尾花（Iris）数据被广泛地作为判别分析的例子。数据是对3种鸢尾花：刚毛鸢尾花（第I组）、变色鸢尾花（第II组）和弗吉尼亚鸢尾花（第III组）各抽取一个容量为50的样本，测量其花萼长(x_1)、花萼宽(x_2)、花瓣长(x_3)、花瓣宽(x_4)，单位为mm，数据列于表5.4.1。

表5. 4. 1　　鸢尾花数据

编号	组别	x_1	x_2	x_3	x_4	编号	组别	x_1	x_2	x_3	x_4
1	I	50	33	14	2	⋮	⋮	⋮	⋮	⋮	⋮
2	III	64	28	56	22	141	II	55	23	40	13
3	II	65	28	46	15	142	II	66	30	44	14
4	III	67	31	56	24	143	II	68	28	48	14
5	III	63	28	51	15	144	I	54	34	17	2
6	I	46	34	14	3	145	I	51	37	15	4
7	III	69	31	51	23	146	I	52	35	15	2
8	II	62	22	45	15	147	III	58	28	51	24
9	II	59	32	48	18	148	II	67	30	50	17
10	I	46	36	10	2	149	III	63	33	60	25
⋮	⋮	⋮	⋮	⋮	⋮	150	I	53	37	15	2

- 本题中，$n_1=n_2=n_3=50$，$n=n_1+n_2+n_3=150$。经计算

$$\bar{\boldsymbol{x}}_1=\begin{pmatrix}50.06\\34.28\\14.62\\2.46\end{pmatrix},\quad \bar{\boldsymbol{x}}_2=\begin{pmatrix}59.36\\27.70\\42.60\\13.26\end{pmatrix},\quad \bar{\boldsymbol{x}}_3=\begin{pmatrix}65.88\\29.74\\55.52\\20.26\end{pmatrix}$$

$$\bar{\boldsymbol{x}}=\frac{1}{n}\sum_{i=1}^{3}n_i\bar{\boldsymbol{x}}_i=\begin{pmatrix}58.433\\30.573\\37.580\\11.993\end{pmatrix}$$

$$\boldsymbol{H}=\sum_{i=1}^{3} n_i\left(\overline{\boldsymbol{x}}_i-\overline{\boldsymbol{x}}\right)\left(\overline{\boldsymbol{x}}_i-\overline{\boldsymbol{x}}\right)'$$

$$=\begin{pmatrix} 6321.213 & -1995.267 & 16524.840 & 7127.933 \\ -1995.267 & 1134.493 & -5723.960 & -2293.267 \\ 16524.840 & -5723.960 & 43710.280 & 18677.400 \\ 7127.933 & -2293.267 & 18677.400 & 8041.333 \end{pmatrix}$$

$$\boldsymbol{E}=\sum_{i=1}^{3}\sum_{j=1}^{n_i}\left(\overline{\boldsymbol{x}}_{ij}-\overline{\boldsymbol{x}}_i\right)\left(\overline{\boldsymbol{x}}_{ij}-\overline{\boldsymbol{x}}_i\right)'$$

$$=\begin{pmatrix} 3895.620 & 1363.000 & 2462.460 & 564.500 \\ 1363.000 & 1696.200 & 812.080 & 480.840 \\ 2462.460 & 812.080 & 2722.260 & 627.180 \\ 564.500 & 480.840 & 627.180 & 615.660 \end{pmatrix}$$

$$\boldsymbol{E}^{-1}\boldsymbol{H} = \begin{pmatrix} -3.058 & 1.081 & -8.112 & -3.459 \\ -5.562 & 2.178 & -14.965 & -6.308 \\ 8.077 & -2.943 & 21.512 & 9.142 \\ 10.497 & -3.420 & 27.549 & 11.846 \end{pmatrix}$$

$\boldsymbol{E}^{-1}\boldsymbol{H}$的正特征值个数$s=\min(k-1,p)=\min(2,4)=2$，可求得两个正特征值

$$\lambda_1=32.192,\quad \lambda_2=0.285$$

相应的标准化特征向量（其长度满足$\boldsymbol{t}_i'\boldsymbol{S}_p\boldsymbol{t}_i=1, i=1,2$）

$$\boldsymbol{t}_1 = \begin{pmatrix} -0.083 \\ -0.153 \\ 0.220 \\ 0.281 \end{pmatrix},\quad \boldsymbol{t}_2 = \begin{pmatrix} 0.002 \\ 0.216 \\ -0.093 \\ 0.284 \end{pmatrix}$$

所以，中心化的费希尔判别函数为

$$
\begin{aligned}
y_1 &= \boldsymbol{t}_1'(\boldsymbol{x}-\bar{\boldsymbol{x}}) \\
&= -0.083\times(x_1-58.433)-0.153\times(x_2-30.573) \\
&\quad +0.220\times(x_3-37.580)+0.281\times(x_4-11.993)
\end{aligned}
$$

$$
\begin{aligned}
y_2 &= \boldsymbol{t}_2'(\boldsymbol{x}-\bar{\boldsymbol{x}}) \\
&= 0.002\times(x_1-58.433)+0.216\times(x_2-30.573) \\
&\quad -0.093\times(x_3-37.580)+0.284\times(x_4-11.993)
\end{aligned}
$$

判别函数的组均值为

$$
\bar{y}_{11}=-7.608,\quad \bar{y}_{21}=1.825,\quad \bar{y}_{31}=5.783
$$

$$
\bar{y}_{12}=0.215,\quad \bar{y}_{22}=-0.728,\quad \bar{y}_{32}=0.513
$$

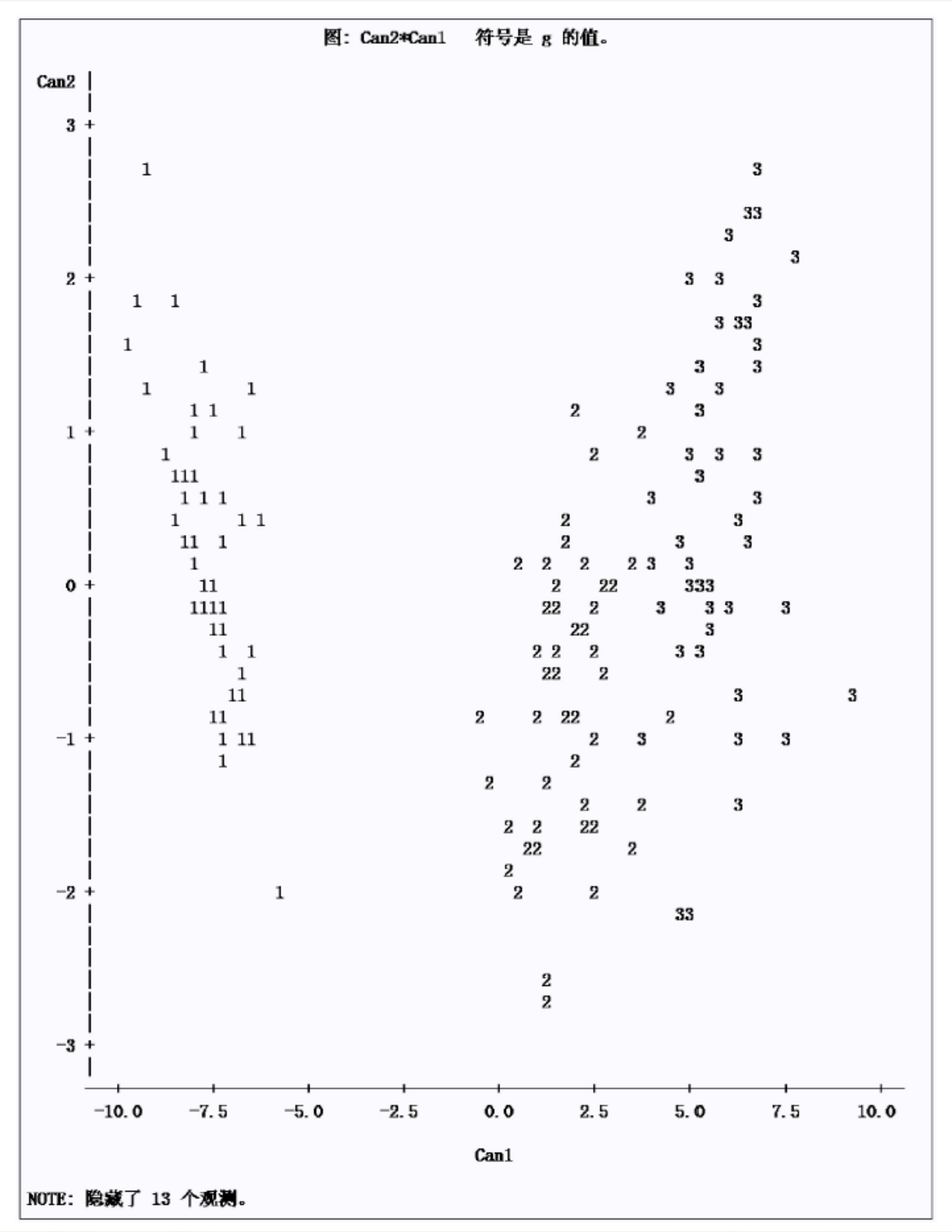


图5. 4. 2　鸢尾花数据两个判别式得分的散点图

- 各组如能在前几个判别函数构成的低维空间中分离得较好，则在原始变量的更高维空间中一般也会分离得好；反之未必。
- 费希尔判别虽是一种很好的降维投影方法，但该方法也有其不适用的场合。

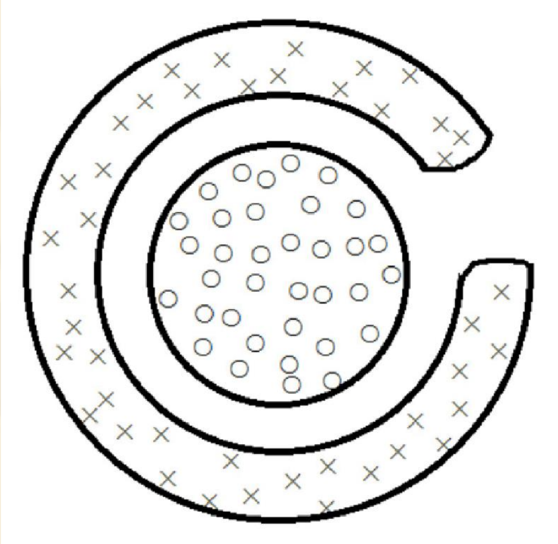


图5.4.4　不适合使用费希尔判别的一个例子

四、判别规则

- *1.一般情形
- 2.两组情形

*1.一般情形

- 由于各判别函数都具有单位方差且彼此不相关，故此时的马氏距离等同于欧氏距离。我们采用距离判别法，依据$(y_1,y_2,\cdots,y_r)$值，判别新样品归属离它最近的那一组。
- 判别规则为

$$\boldsymbol{x} \in \pi_l, \quad 若 \sum_{j=1}^{r}\left(y_j - \bar{y}_{lj}\right)^2 = \min_{1\le i\le k} \sum_{j=1}^{r}\left(y_j - \bar{y}_{ij}\right)^2 \qquad (5.4.7)$$

其中$\bar{y}_{ij} = \boldsymbol{t}_j'\bar{\boldsymbol{x}}_i$, $\bar{\boldsymbol{x}}_i = \dfrac{1}{n_i}\sum_{j=1}^{n_i}\boldsymbol{x}_{ij}$ ，$i=1,2,\cdots,k$ 。该判别规则也可表达为

$$\boldsymbol{x} \in \pi_l, \quad 若 \sum_{j=1}^{r}\left[\boldsymbol{t}_j'\left(\boldsymbol{x} - \bar{\boldsymbol{x}}_l\right)\right]^2 = \min_{1\le i\le k} \sum_{j=1}^{r}\left[\boldsymbol{t}_j'\left(\boldsymbol{x} - \bar{\boldsymbol{x}}_i\right)\right]^2$$

- 如果只使用一个判别函数进行判别（即r=1），则以上判别规则可简化为

$$\boldsymbol{x} \in \pi_l, \quad 若\left|y-\bar{y}_l\right| = \min_{1\le i\le k}\left|y-\bar{y}_i\right|$$

式中y和$\bar{y}_i(i=1,2,\cdots,k)$分别是前面判别规则中的$y_1$和$\bar{y}_{i1}$ $(i=1,2,\cdots,k)$。

- 如果使用所有s个判别函数作判别（即$r=s$），则费希尔判别(5.4.7)等价于距离判别(5.2.17)，自然对各组皆为正态也等价于协方差矩阵相等且先验概率和误判代价也均相同的贝叶斯判别。

❖ 例5.4.2　在例5.4.1中，k=3，取r=s(=2)，使用(5.4.7)式[或(5.2.17)式]进行判别分类。回代法的判别情况列于表5.4.2。

表5. 4. 2　　判别情况

真实组 \ 判别为	I	II	III
I	50	0	0
II	0	48	2
III	0	1	49

所以

$$\hat{P}(2|1)=0,\quad \hat{P}(3|1)=0$$

$$\hat{P}(1|2)=0,\quad \hat{P}(3|2)=\frac{2}{50}=0.04$$

$$\hat{P}(1|3)=0,\quad \hat{P}(2|3)=\frac{1}{50}=0.02$$

这些误判概率是比较低的。

2.两组情形

- 对于两组的判别，费希尔判别函数只有一个，有$r=s=1$，从而(5.4.7)式等价于(5.2.17)式，而后者又退化为(5.2.6)式。
- 因此，两组的费希尔判别等价于协方差矩阵相等的距离判别，对两个正态组也等价于协方差矩阵相等且先验概率和误判代价也均相同的贝叶斯判别。

§5.5 逐步判别

- 逐步判别法是判别分析中一种自动搜索变量子集的方法，它未必最优，但往往却是有效的，是一种应用最广泛的判别变量选择方法。
- 逐步判别法的基本思想及基本步骤类似于回归分析中的逐步回归法。

- 一、附加信息检验
- 二、变量选择的方法

一、附加信息检验

- 设$\mathbf{x}=(\mathbf{x}_1',\mathbf{x}_2')$，其中$\mathbf{x}_1=(x_1,x_2,\cdots,x_r)'$是原先用作判别的变量，而$\mathbf{x}_2=(x_{r+1},x_{r+2},\cdots,x_p)'$是新引入的变量。
- 我们希望知道，在已有$\mathbf{x}_1$用作判别的条件下，$\mathbf{x}_2$所提供的（超越$\mathbf{x}_1$所含信息的）附加信息能否使区分各组的能力有显著的提高。如果没有显著提高，则就认为$\mathbf{x}_2$的引入是不值得的。
- 设有k个组$\pi_1,\pi_2,\cdots,\pi_k$，其$\mathbf{x}$的分布皆为p元正态分布，且具有相同的协方差矩阵。从这k个组中各自独立地抽取一个样本，n为k个组的总样本容量。欲检验

  H_0：各组的$E(\mathbf{x}_2|\mathbf{x}_1)$相等，$H_1$：各组的$E(\mathbf{x}_2|\mathbf{x}_1)$不全相等

- 将组内平方和及叉积和矩阵$\boldsymbol{E}$，组间平方和及叉积和矩阵$\boldsymbol{H}$分块为：

$$\boldsymbol{E}=\begin{array}{cc} \begin{pmatrix} \boldsymbol{E}_{11} & \boldsymbol{E}_{12} \\ \boldsymbol{E}_{21} & \boldsymbol{E}_{22} \end{pmatrix} & \begin{array}{l} r \\ p-r \end{array} \\ \begin{array}{cc} r & p-r \end{array} & \end{array},\quad \boldsymbol{H}=\begin{array}{cc} \begin{pmatrix} \boldsymbol{H}_{11} & \boldsymbol{H}_{12} \\ \boldsymbol{H}_{21} & \boldsymbol{H}_{22} \end{pmatrix} & \begin{array}{l} r \\ p-r \end{array} \\ \begin{array}{cc} r & p-r \end{array} & \end{array}$$

则检验统计量为

$$\Lambda\left(\boldsymbol{x}_2 \mid \boldsymbol{x}_1\right)=\frac{\Lambda\left(\boldsymbol{x}_1, \boldsymbol{x}_2\right)}{\Lambda\left(\boldsymbol{x}_1\right)}$$

其中

$$\Lambda\left(\boldsymbol{x}_1, \boldsymbol{x}_2\right)=\frac{|\boldsymbol{E}|}{|\boldsymbol{E}+\boldsymbol{H}|},\quad \Lambda\left(\boldsymbol{x}_1\right)=\frac{\left|\boldsymbol{E}_{11}\right|}{\left|\boldsymbol{E}_{11}+\boldsymbol{H}_{11}\right|}$$

- 当H_0为真时，$\Lambda(\mathbf{x}_2|\mathbf{x}_1)$服从$\Lambda(p-r,k-1,n-k-r)$。我们特别感兴趣的是$p-r=1$（即$r=p-1$）时的情形，此时

偏Λ统计量

$$\Lambda=\Lambda\left(x_p\mid x_1,x_2,\cdots,x_{p-1}\right)=\frac{\Lambda\left(x_1,x_2,\cdots,x_p\right)}{\Lambda\left(x_1,x_2,\cdots,x_{p-1}\right)}$$

$$\sim\Lambda\left(1,k-1,n-k-p+1\right)$$

偏F统计量

$$F=F\left(x_p\mid x_1,x_2,\cdots,x_{p-1}\right)=\frac{1-\Lambda}{\Lambda}\frac{n-k-p+1}{k-1}$$

$$\sim F\left(k-1,n-k-p+1\right)$$

- 对给定的α，拒绝规则为：

若$F\geq F_\alpha(k-1,n-k-p+1)$，则拒绝$H_0$

二、变量选择的方法

- 判别分析的变量选择方法：前进法、后退法和逐步判别法。
- 前进法开始时没有用作判别的变量，每次选入一个对判别能力的提高有最显著作用的变量，过程只进不出，当不再有未被选入的变量达到临界值时，前进选入的过程停止。
- 后退法的过程与前进法相反，开始时引入所有变量，每次剔除一个对判别能力的提高最不显著的变量，过程只出不进，当余下的变量都达到用作判别的标准时，后退剔除的过程停止。
- 逐步判别法是前进法和后退法的结合，在变量的选择过程中有进有出。实践中，逐步判别法通常最受欢迎。

逐步判别法的基本步骤

- (1)对每个x_i，计算其一元方差分析的F统计量$F(x_i)$，不妨设 $F(x_1)=\max\limits_{i} F(x_i)$，即$x_1$有最大的判别能力。
  若$F(x_1)<F_\alpha(k-1,n-k)$，则表明没有一个变量可以选入；
  若$F(x_1)\geq F_\alpha(k-1,n-k)$，则$x_1$选入，并进入下一步。
- (2)对(1)中每一未选入的变量，计算偏F统计量$F(x_i|x_1)$，不妨设$F(x_2|x_1)=\max\limits_{2\leq i\leq p} F(x_i|x_1)$，即$x_2$对判别能力的提升有最大贡献。若$F(x_2|x_1)<F_\alpha(k-1,n-k-1)$，则选变量过程结束；若$F(x_2|x_1)\geq F_\alpha(k-1,n-k-1)$，则$x_2$选入，并进入下一步。

- 一般地，如已选入了r个变量，不妨设是$x_1,x_2,\cdots,x_r$，并设

$$F\left(x_{r+1} \mid x_1,x_2,\cdots,x_r\right)=\max_{r+1\le i\le p} F\left(x_i \mid x_1,x_2,\cdots,x_r\right)$$

若$F\left(x_{r+1} \mid x_1,x_2,\cdots,x_r\right)<F_\alpha\left(k-1,n-k-r\right)$，则选变量过程结束；若$F\left(x_{r+1} \mid x_1,x_2,\cdots,x_r\right)\ge F_\alpha\left(k-1,n-k-r\right)$，则$x_{r+1}$选入，并进入下一步。

- (3)在第r+1个变量选入后，要重新核实较早选入的r个变量，应将对判别效果不再显著的变量剔除出去。不妨设

$$F\left(x_l \mid x_1,\cdots,x_{l-1},x_{l+1},\cdots,x_{r+1}\right)=\min_{1\le i\le r} F\left(x_i \mid x_1,\cdots,x_{i-1},x_{i+1},\cdots,x_{r+1}\right)$$

若 $F\left(x_l \mid x_1,\cdots,x_{l-1},x_{l+1},\cdots,x_{r+1}\right)\ge F_\alpha\left(k-1,n-k-r\right)$，则没有变量需剔除，回到(2)；若$F\left(x_l \mid x_1,\cdots,x_{l-1},x_{l+1},\cdots,x_{r+1}\right)<$ $F_\alpha(k-1,n-k-r)$，则剔除变量x_l，再对其余$r-1$个变量继续进行核实，直至无变量可剔除为止，然后再回到(2)。

- (4)经过(2)和(3)的不断选入和剔除的过程，最后既不能选进新变量，也不能剔除已选入的变量，变量选择过程到此结束。
- 如果选入变量的临界值$F_{进}$和剔除变量的临界值$F_{出}$相同，则有很小的可能性会使得变量的选入和剔除过程无休止、连续不断地循环进行下去。但只要在确定临界值时让$F_{出}$比$F_{进}$略微小一点，这种可能性就可以被排除。
- 进行逐步判别实际上是在做逐步多元方差分析，在变量的筛选过程中没有任何判别函数被计算。在变量筛选完成后，我们方可以对选择的变量计算判别函数和建立判别规则。

- 例5.5.1　对例5.4.2中的数据作逐步判别，具体步骤如下：
- (1)对每一变量分别计算一元方差分析的F统计量和p值，并列于表5.5.1。x_3第一个选入。

表5.5.1　　F统计量和p值

变量	x_1	x_2	x_3	x_4
F	119.26	49.16	1180.16	960.01
p值	<0.0001	<0.0001	<0.0001	<0.0001

- (2)

$$\Lambda\left(x_i \mid x_3\right)=\frac{\Lambda\left(x_3, x_i\right)}{\Lambda\left(x_3\right)}$$

$$F\left(x_i \mid x_3\right)=\frac{1-\Lambda\left(x_i \mid x_3\right)}{\Lambda\left(x_i \mid x_3\right)} \frac{n-k-1}{k-1}$$

❖ i=1,2,4，计算结果列于表5.5.2。x_2选入。

表5.5.2　　x_3已选入时的偏F统计量和p值

变量	x_1	x_2	x_4
偏F	34.32	43.04	24.77
p值	<0.0001	<0.0001	<0.0001

❖ (3)核实x_3是否因x_2的选入仍保持显著。经计算，$F(x_3|x_2)$=1112.95，p<0.0001，从而保留x_3。

$$\Lambda\left(x_i \mid x_2, x_3\right)=\frac{\Lambda\left(x_2, x_3, x_i\right)}{\Lambda\left(x_2, x_3\right)}$$

$$F\left(x_i \mid x_2, x_3\right)=\frac{1-\Lambda\left(x_i \mid x_2, x_3\right)}{\Lambda\left(x_i \mid x_2, x_3\right)} \frac{n-k-2}{k-1}$$

- i=1,4，结果见表5.5.3。可见，x_4选入。

表5.5.3 x_2, x_3已选入时的偏F统计量和p值

变量	x_1	x_4
偏F	12.27	34.57
p值	<0.0001	<0.0001

- (4)核实x_4选入后早先已选入的x_2和x_3是否还显著，计算偏F统计量$F(x_2|x_3,x_4)$和$F(x_3|x_2,x_4)$，结果列于表5.5.4。可见，x_2和x_3皆保留。继续计算

$$\Lambda\left(x_1 \mid x_2, x_3, x_4\right)=\frac{\Lambda\left(x_1, x_2, x_3, x_4\right)}{\Lambda\left(x_2, x_3, x_4\right)}$$

$$F\left(x_1 \mid x_2, x_3, x_4\right)=\frac{1-\Lambda\left(x_1 \mid x_2, x_3, x_4\right)}{\Lambda\left(x_1 \mid x_2, x_3, x_4\right)} \frac{n-k-3}{k-1}$$

- 可得$F(x_1|x_2,x_3,x_4)=4.72$，$p=0.0103$，故x_1也选入。

表5.5.4　选入x_4后核实x_2和x_3是否还显著的偏F统计量和p值

变量	x_2	x_3
偏F	54.58	38.72
p值	<0.0001	<0.0001

- (5)核实x_1选入后原已选入的x_2,x_3,x_4是否还是显著的，计算偏F统计量$F(x_2|x_1,x_3,x_4)$, $F(x_3|x_1,x_2,x_4)$和$F(x_4|x_1,x_2,x_3)$，结果列于表5.5.5。计算结果表明，已选入的变量无一剔除。

表5.5.5　选入x_1后核实x_2,x_3和x_4是否还显著的偏F统计量和p值

变量	x_2	x_3	x_4
偏F	21.94	35.59	24.90
p值	<0.0001	<0.0001	<0.0001

- (6)

表5.5.6 变量选择过程汇总

步骤	1	2	3	4
变量	x_3	x_2	x_4	x_1
F	1180.16	43.04	34.57	4.72
p值	<0.0001	<0.0001	<0.0001	0.0103