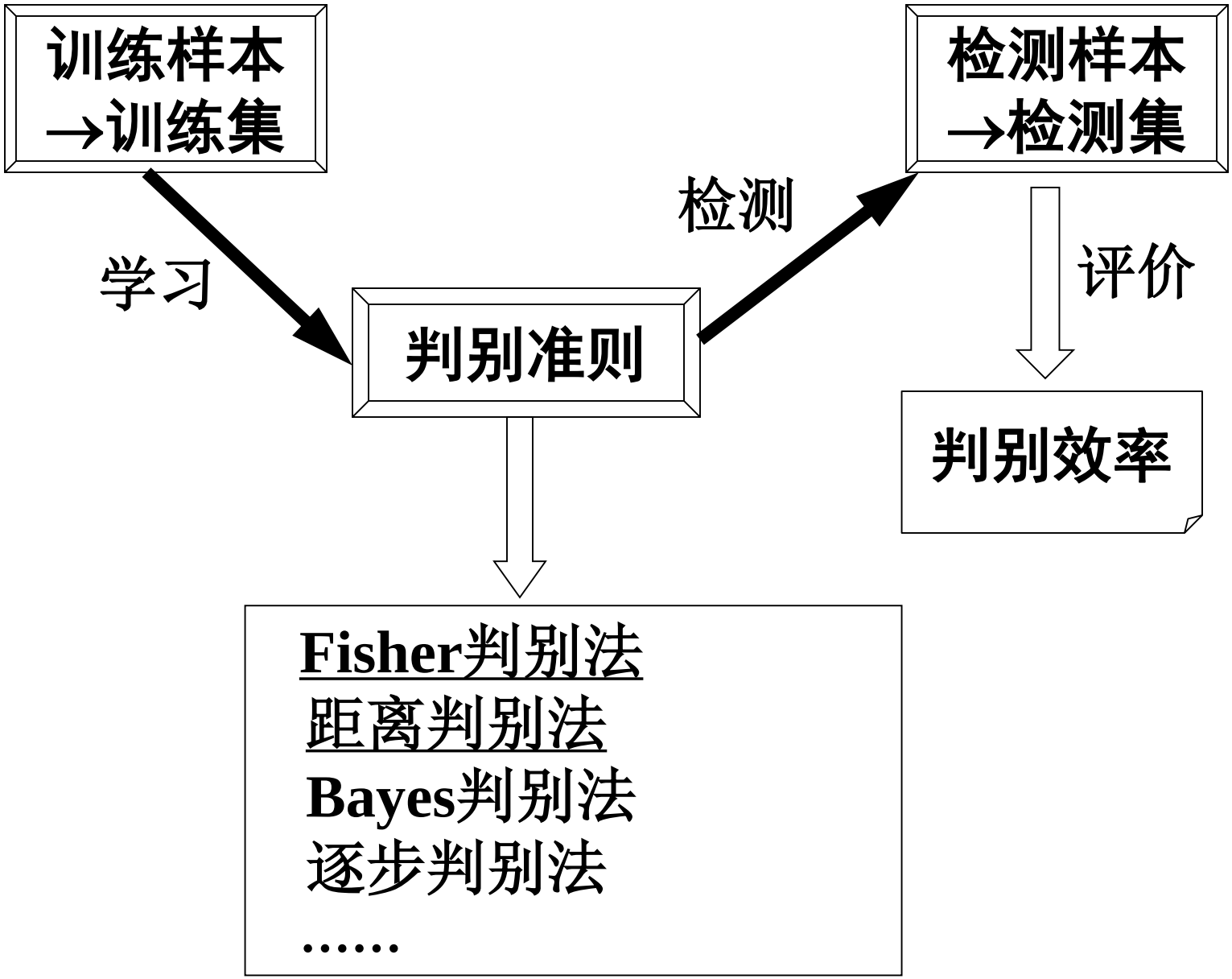
第六章 判别分析及R使用
(Discriminant Analysis)

- §6.1 判别分析的概念
- §6.2 Fisher线性判别
- §6.3 距离判别
- §6.4 Bayes判别
- §6.5 逐步判别
- §6.6 判别分析几点说明

§6.1 判别分析的概念

提出的问题	分组	判别变量
信用评级	风险等级：高、低	年龄、收入、贷款额、工作延续时间
新产品的成功前景	经济收益：盈利、亏损	产品新颖度、市场信息、价格和技术
选民分析	共和党、民主党	纳税、就业、医疗、裁军等观点
挑选销售人员	业绩：好、坏	教育、年龄、性格和身体特征


训练样本
→训练集
检测样本
→检测集
学习
检测
评价
判别准则
判别效率
Fisher判别法
距离判别法
Bayes判别法
逐步判别法
……

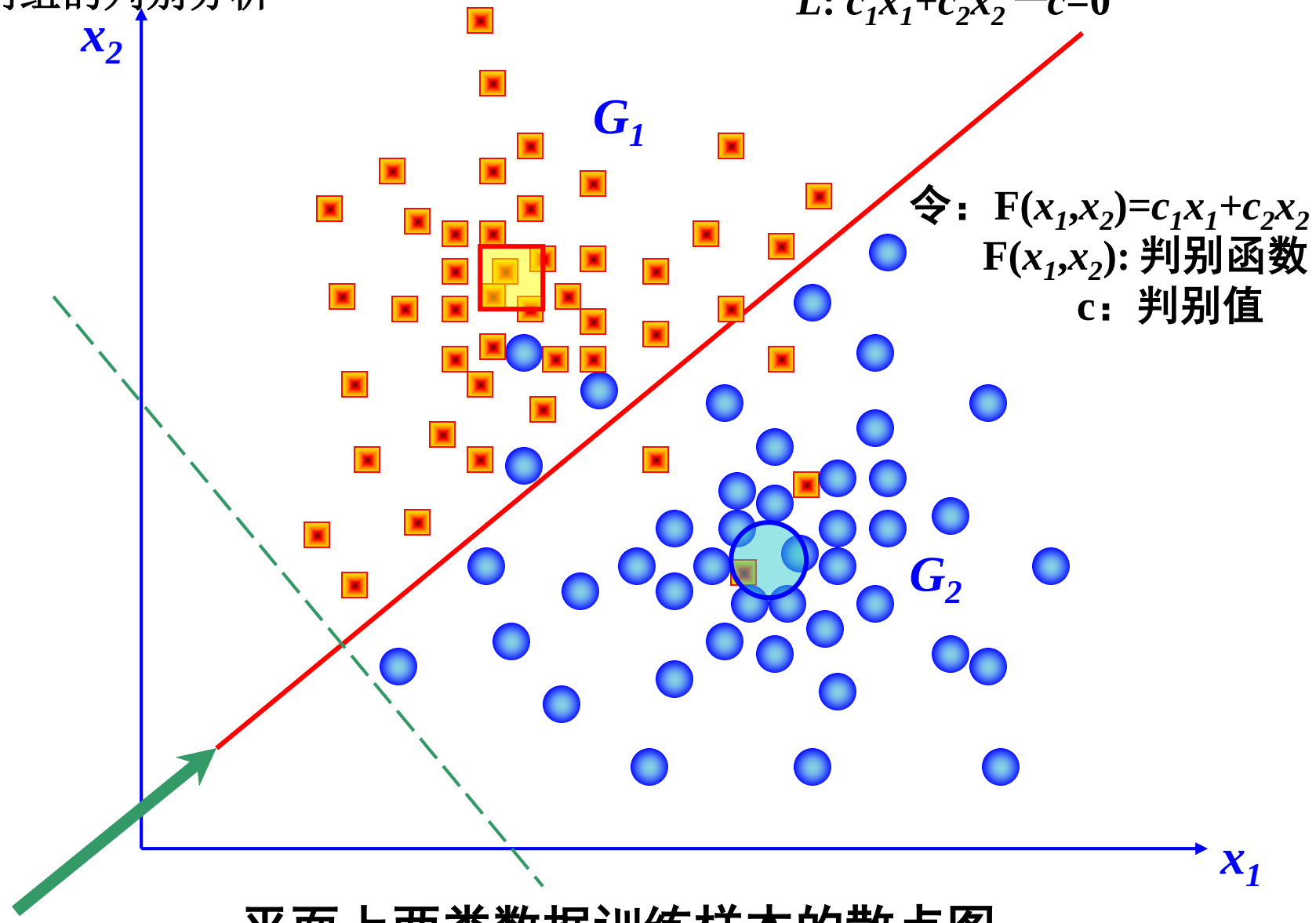
§6.2 Fisher线性判别法

Fisher判别的基本思想

将 k 组p维的数据投影到某一个方向，使得投影后的组与组之间尽可能地分开。

下面以两组为例进行说明：

只有两组的判别分析



平面上两类数据训练样本的散点图

（两组数据样本在平面上存在一个合理的分界线*L*）

已知：数据有**p**个变量，每个数据点为**p**维向量**X**：

$$X(x_1, x_2, ..., x_p)$$

已知总体数据分为两类： G_1和G_2 ， 总体G_1有**m**个样本点，总体G_2有**n**个样本点。

目标：

求解在**p**维空间中总体G_1和总体G_2的最优分界平面。

			变量			
			1	2	…	p
总体G_1 (i=1, …, m)	1	$X_1^{(1)}$	$x_{11}^{(1)}$	$x_{12}^{(1)}$	…	$x_{1p}^{(1)}$
		…	…	…	…	…
	i	$X_i^{(1)}$	$x_{i1}^{(1)}$	$x_{i2}^{(1)}$	…	$x_{ip}^{(1)}$
		…	…	…	…	…
	m	$X_m^{(1)}$	$x_{m1}^{(1)}$	$x_{m2}^{(1)}$	…	$x_{mp}^{(1)}$
总体G_2 (j=1, …, n)	1	$X_1^{(2)}$	$x_{11}^{(2)}$	$x_{12}^{(2)}$	…	$x_{1p}^{(2)}$
		…	…	…	…	…
	j	$X_j^{(2)}$	$x_{j1}^{(2)}$	$x_{j2}^{(2)}$	…	$x_{jp}^{(2)}$
		…	…	…	…	…
	n	$X_n^{(2)}$	$x_{n1}^{(2)}$	$x_{n2}^{(2)}$	…	$x_{np}^{(2)}$

定义线性判别函数为：

$$F(x_1, x_2, ..., x_p) = C_1 x_1 + C_2 x_2 + ... + C_p x_p$$

其中C_i (i = 1, 2, …, p)为常数（待定系数）。

若判别值为 C ， 对于任何未知数据点$X(x_1, x_2, …, x_p)$，代入判别函数，依据 F $(x_1, x_2, …, x_p)$与C值的比较，可以判别点X属于哪一类。

1、确定待定系数C_i (i = 1, 2, …, p)

2、确定判别值C

确定待定系数C_i

将类G_1的m个点、类G_2的n个点分别代入判别函数：

$$y_i^{(1)} = C_1 x_{i1}^{(1)} + C_2 x_{i2}^{(1)} + \ldots + C_p x_{ip}^{(1)} \qquad i = 1, \ldots, m$$

$$y_j^{(2)} = C_1 x_{j1}^{(2)} + C_2 x_{j2}^{(2)} + \ldots + C_p x_{jp}^{(2)} \qquad j = 1, \ldots, n$$

记 $\bar{y}^{(1)} = \frac{1}{m}\sum_{i=1}^{m} y_i^{(1)}$ $\bar{y}^{(2)} = \frac{1}{n}\sum_{j=1}^{n} y_j^{(2)}$

令：
$$\delta_A = (\bar{y}^{(1)} - \bar{y}^{(2)})^2$$

δ_A与G_1和G_2两类点的几何中心的距离相关。显然，判别函数$F(x_1, x_2, \ldots, x_p)$应该使δ_A值越大越好。

令：
$$\delta_B = \sum_{i=1}^{m}\left(y_i^{(1)} - \bar{y}^{(1)}\right)^2 + \sum_{j=1}^{n}\left(y_j^{(2)} - \bar{y}^{(2)}\right)^2$$

δ_B与G_1和G_2两类点的相对于各自几何中心的离差相关。显然，判别函数$F(x_1, x_2, \ldots, x_p)$应该使δ_B值越小越好。

构造函数I：

$$I = I(C_1, C_2, ..., C_p) = \frac{\delta_A}{\delta_B} = \frac{\left(\bar{y}^{(1)} - \bar{y}^{(2)}\right)^2}{\sum_{i=1}^{m}\left(y_i^{(1)} - \bar{y}^{(1)}\right)^2 + \sum_{j=1}^{n}\left(y_j^{(2)} - \bar{y}^{(2)}\right)^2}$$

选择合适的待定系数C_i (i = 1, 2, …, p)，
使得函数I(C_1, C_2, …, C_p)达到极大值。

$$\frac{\partial I}{\partial C_i} = 0 \qquad i = 1, 2, ..., p$$

定理6.1: 线性组合 $Y = C'X = (\overline{X_1} - \overline{X_2})'S_p^{-1}X$ 对所有可能的线性系数向量，使得目标函数 I 达到最大（同书上**P151**）

这里， $S_p = \dfrac{(m-1)S_1 + (n-1)S_2}{m+n-2}$

$$S_1 = \frac{1}{m-1}\sum_{i=1}^{m}(X_i^{(1)} - \bar{X}_1)(X_i^{(1)} - \bar{X}_1)'$$

$$S_2 = \frac{1}{n-1}\sum_{j=1}^{n}(X_j^{(2)} - \bar{X}_2)(X_j^{(2)} - \bar{X}_2)'$$

样本协方差矩阵

确定判别值C

判别函数已知，不妨写成：

$$y = C_1 x_1 + C_2 x_2 + \ldots + C_p x_p$$

把两类均值代入判别函数

$$\bar{y}^{(1)} = \frac{1}{m}\sum_{i=1}^{m} y_i^{(1)} \qquad \bar{y}^{(2)} = \frac{1}{n}\sum_{j=1}^{n} y_j^{(2)}$$

对G_1、G_2的$(m+n)$个点的判别函数值取总体的平均值：

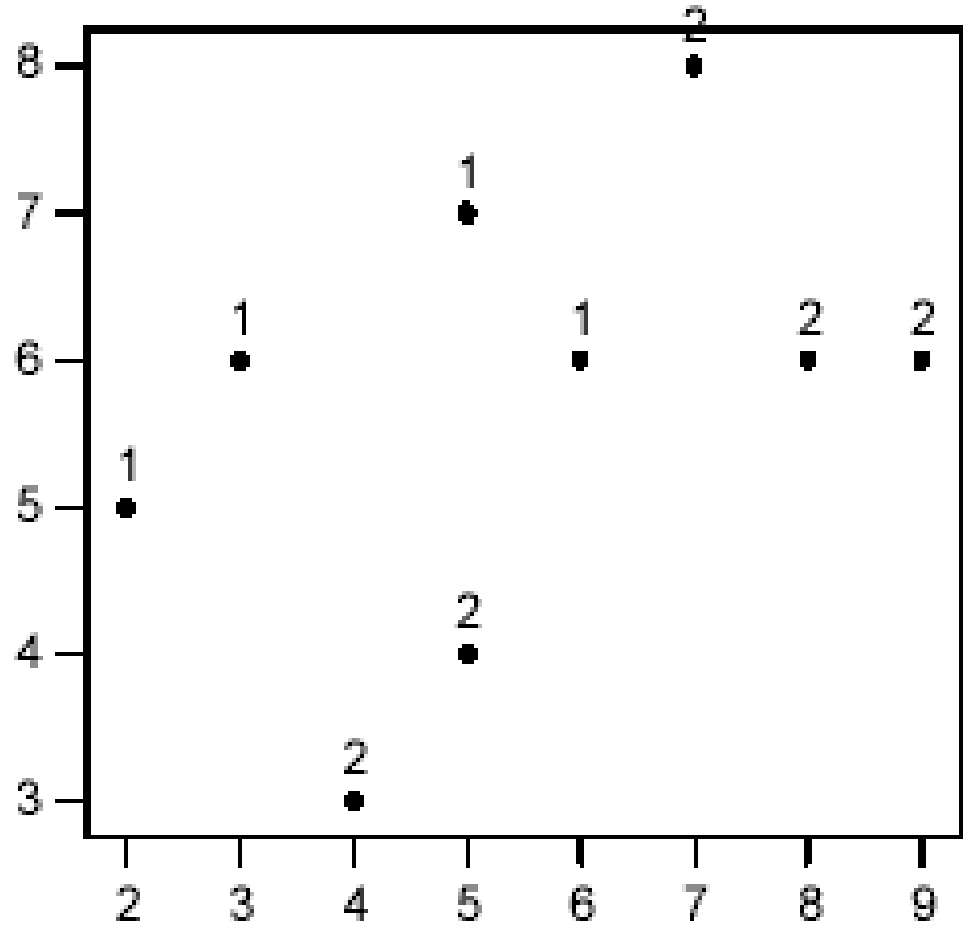
$$\mu = \frac{1}{m+n}\left(\sum_{i=1}^{m} y_i^{(1)} + \sum_{j=1}^{n} y_j^{(2)}\right) = \frac{1}{m+n}\left(m\bar{y}^{(1)} + n\bar{y}^{(2)}\right)$$

显然，μ值是两类点的判别函数值的加权平均，处于两类判别函数平均值之间，也等价于两类点的总体几何中心的判别函数值。因此，将判别值C取为μ值：

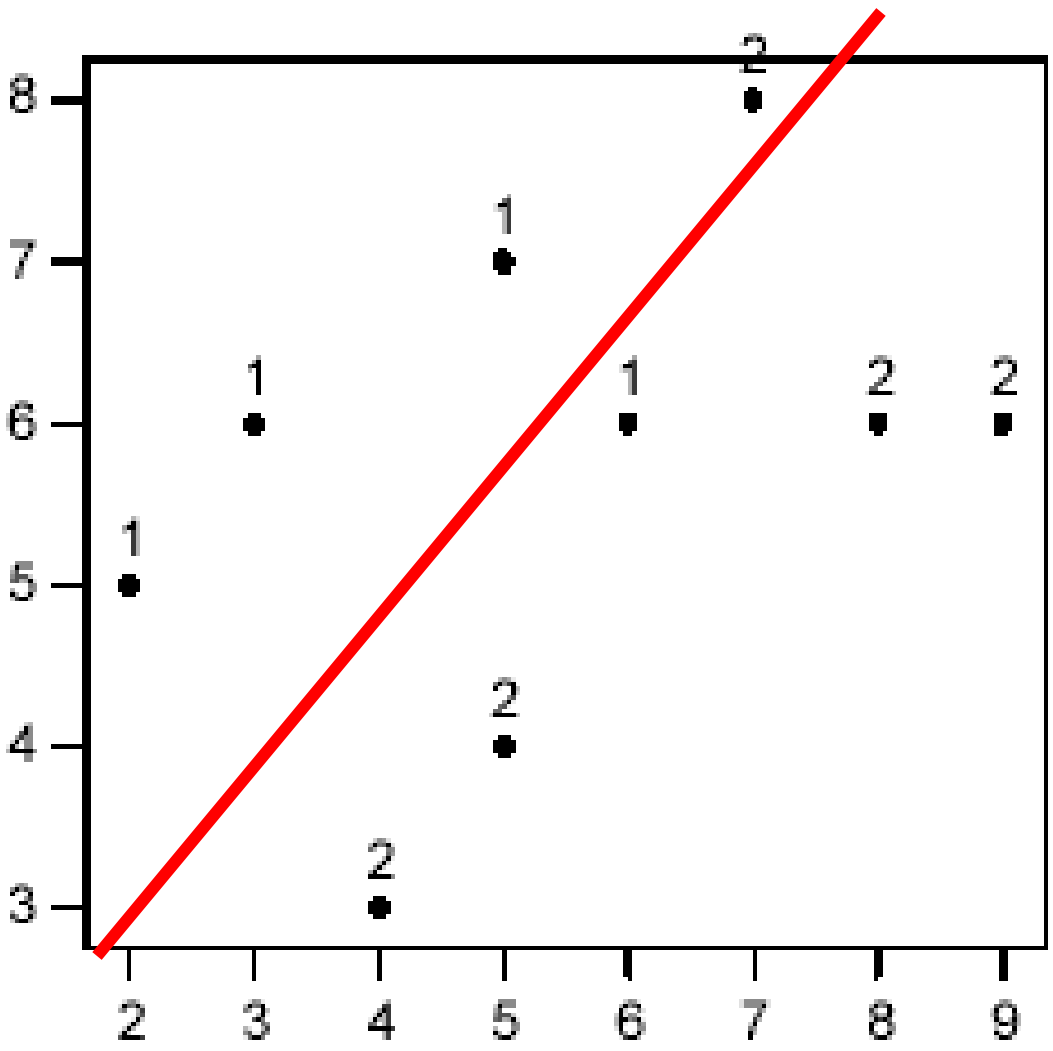
$$C = \frac{m\bar{y}^{(1)} + n\bar{y}^{(2)}}{m+n}$$

Fisher线性判别的应用举例

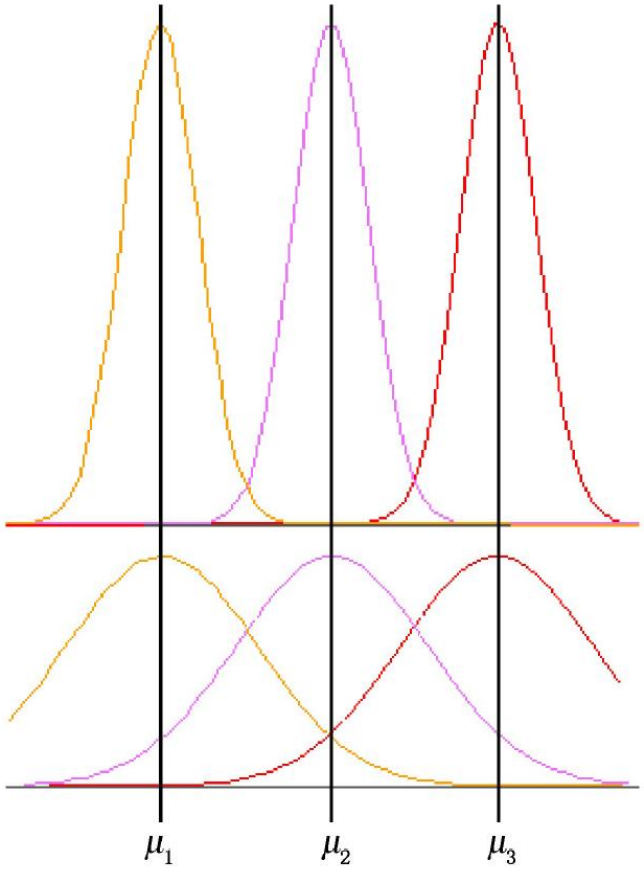
序号	x_1	x_2	类别
1	5	7	1
2	4	3	2
3	7	8	2
4	8	6	2
5	3	6	1
6	2	5	1
7	6	6	1
8	9	6	2
9	5	4	2



$$-1.5080x_1 + 1.5418x_2 = 0.5264$$



三组之间的分离程度



多组别Fisher判别函数（了解）

- 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$\boldsymbol{\Sigma}_1=\boldsymbol{\Sigma}_2=\cdots=\boldsymbol{\Sigma}_k=\boldsymbol{\Sigma}$。

- 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

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

- 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$\boldsymbol{S}_p=\frac{1}{n-k}\boldsymbol{E}$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

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

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

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

- 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

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

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。
- $\Delta(\boldsymbol{t}_i)=\lambda_i$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$r(\leq s)$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。

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，对来自各组样品的分离情况进行观测评估或用目测法对新样品的归属进行辨别。
- 当r=3时，可利用有关统计软件，让样本中来自不同组的样品点呈现不同颜色（或不同形状）以区分各组，然后作（三维）旋转图从多角度来观测评估各组之间的分离效果或辨别新样品的归属，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。
- 如各组的分离情况，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6.1】今天和昨天湿温差 x1及气温差x2是预报明天下雨否的其中两个重要因子， 试建立Fisher线性判别函数.

如测得今天 x1=8.1, x2=2.0试报明天是雨天还是晴天？

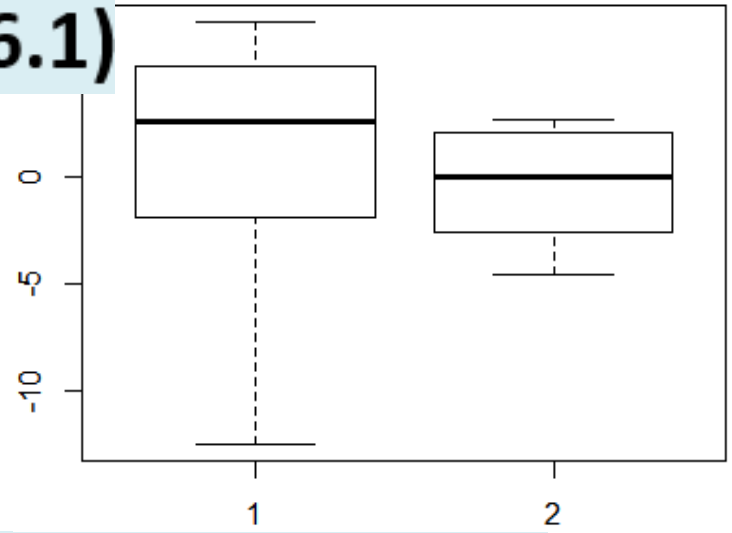
表 6.2 雨天和晴天的湿温差 x1 和气温差 x2

雨天			晴天		
组别	x1	x2	组别	x1	x2
1	−1.9	3.2	2	0.2	6.2
1	−6.9	0.4	2	−0.1	7.5
1	5.2	2.0	2	0.4	14.6
1	5.0	2.5	2	2.7	8.3
1	7.3	0.0	2	2.1	0.8
1	6.8	12.7	2	−4.6	4.3
1	0.9	−5.4	2	−1.7	10.9
1	−12.5	−2.5	2	−2.6	13.1
1	1.5	1.3	2	2.6	12.8
1	3.8	6.8	2	−2.8	10.0

一、基本统计分析

```
d6.1=read.table("clipboard",header=T)
```

```
boxplot(x1~G,d6.1)
```



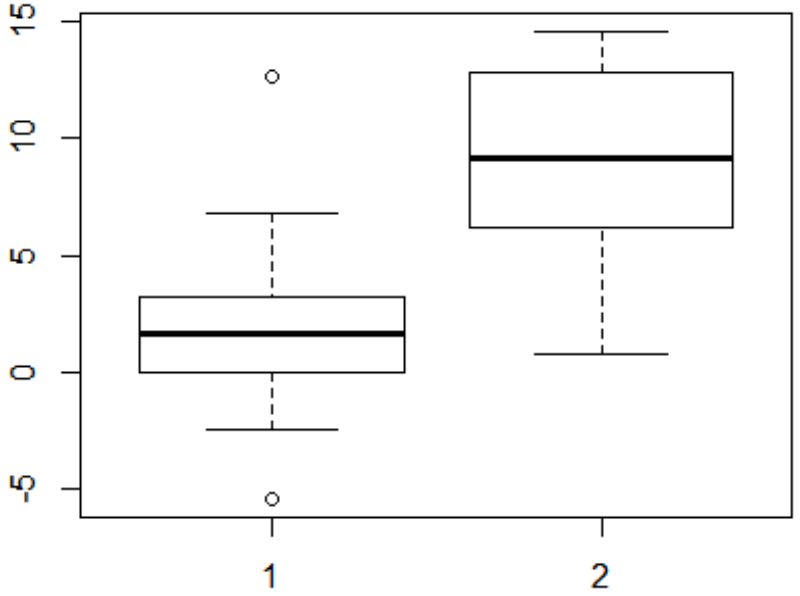

```
t.test(x1~G,d6.1)
```

```
Welch Two Sample t-test

data:  x1 by G
t = 0.6, df = 12, p-value = 0.6
```

	G	x1	x2
1	1	−1.9	3.2
2	1	−6.9	0.4
		5.2	2.0
		5.0	2.5
5	1	7.3	0.0
6	1	6.8	12.7
7	1	0.9	−5.4
8	1	−12.5	−2.5
9	1	1.5	1.3
10	1	3.8	6.8
11	2	0.2	6.2
12	2	−0.1	7.5
13	2	0.4	14.6
14	2	2.7	8.3
15	2	2.1	0.8
16	2	−4.6	4.3
17	2	−1.7	10.9
18	2	−2.6	13.1
19	2	2.6	12.8
20	2	−2.8	10.0

```
boxplot(x2~G,d6.1)
```




```
t.test(x2~G,d6.1)
```

```
Welch Two Sample t-test

data:  x2 by G
t = -3.3, df = 18, p-value = 0.005
```

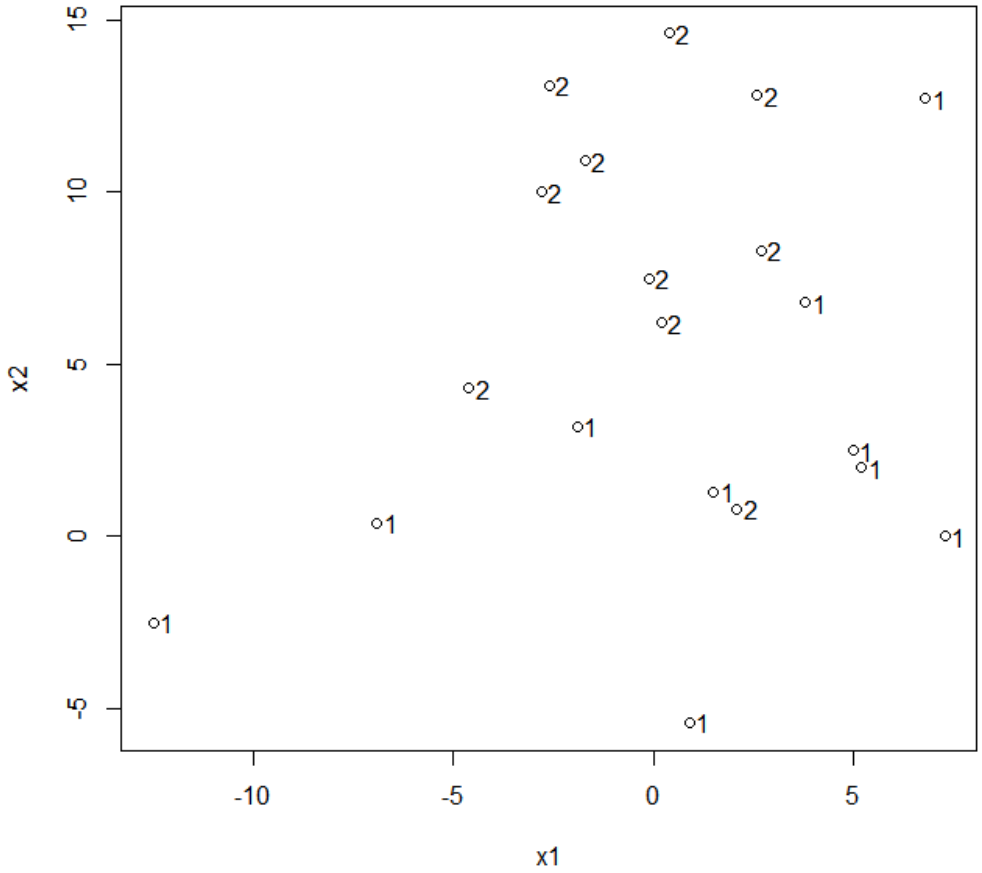
二、Logistic模型分析

```
summary(glm(G-1~x1+x2,family=binomial,d6.1))
```

```
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   -2.076      1.108   -1.87    0.061 .
x1            -0.196      0.146   -1.34    0.179
x2             0.381      0.168    2.27    0.023 *
---
Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

三、Fisher判别分析

```
attach(d6.1)    #绑定数据
plot(x1,x2); text(x1,x2,G,adj=-0.5)    #标识点所属类别 G
```



线性判别分析函数 lda 的用法

lda(formula, data, ...)

formula 形如 **y ~ x1 + x2 + ...** 的公式框架，**data** 数据框

```
library(MASS)
ld=lda(G~x1+x2);ld
```

```
Coefficients of linear discriminants:
     LD1
x1 -0.1035
x2  0.2248
```

```
lp=predict(ld)
G1=lp$class
data.frame(G,G1)
```

	G	G1
1	1	1
2	1	1
3	1	1
4	1	1
5	1	1
6	1	2
7	1	1
8	1	1
9	1	1
10	1	1
11	2	2
12	2	2
13	2	2
14	2	2
15	2	1
16	2	2
17	2	2
18	2	2
19	2	2
20	2	2

```
tab1=table(G,G1);tab1
```

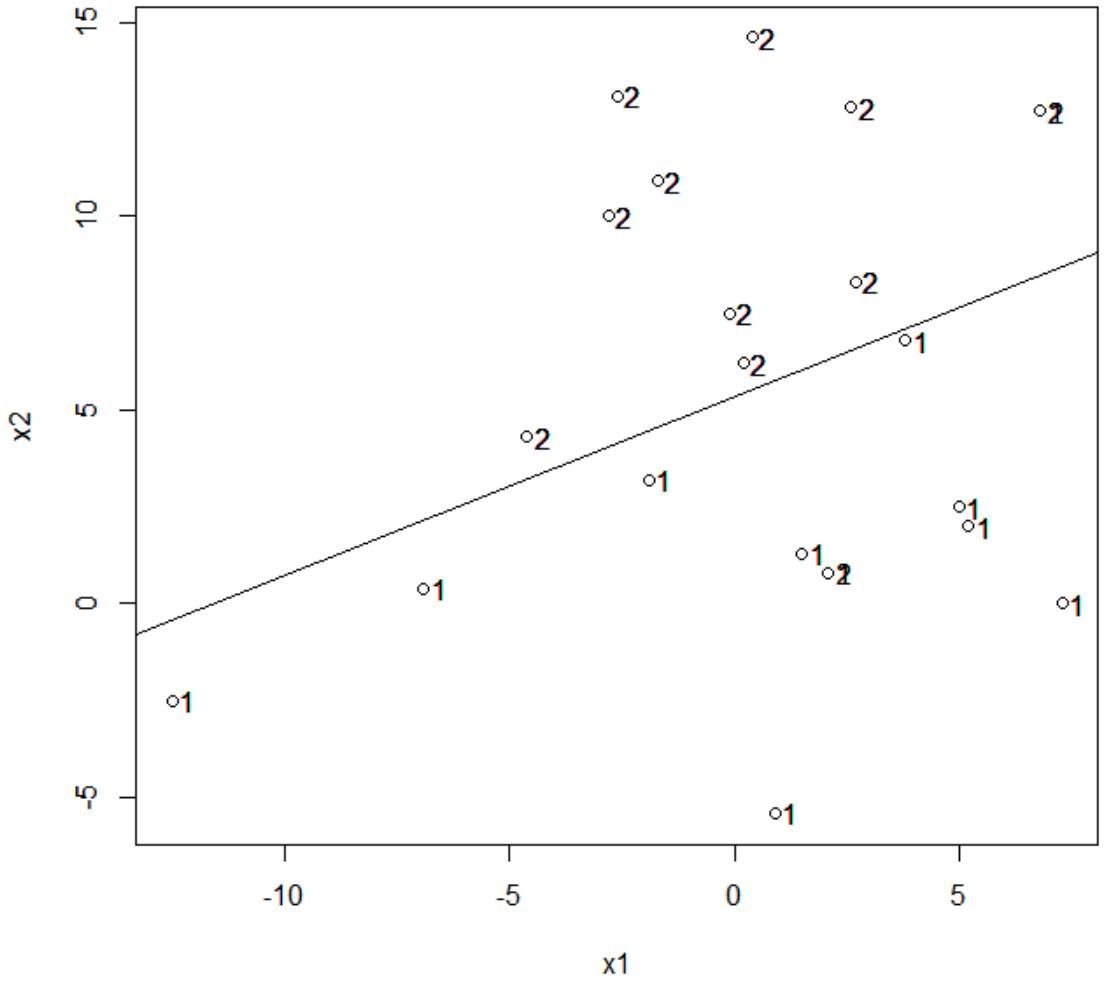
```
   G1
G    1 2
  1 9 1
  2 1 9
```

表 6.3　线性判别的判别效果

原始分类	判别分类		合计
	1	2	
1	9	1	10
2	1	9	10
合计	10	10	20

```
sum(diag(prop.table(tab1)))
```

```
[1] 0.9
```


15
10
5
0
-5
x2
-10
-5
0
5
x1

```
predict(ld,data.frame(x1=8.1,x2=2))

$class
[1] 1
Levels: 1 2
```

误判概率的非参数估计

可以用样本中样品的误判比例来估计误判概率，通常有如下三种非参数估计方法：

- **(1)回代法**

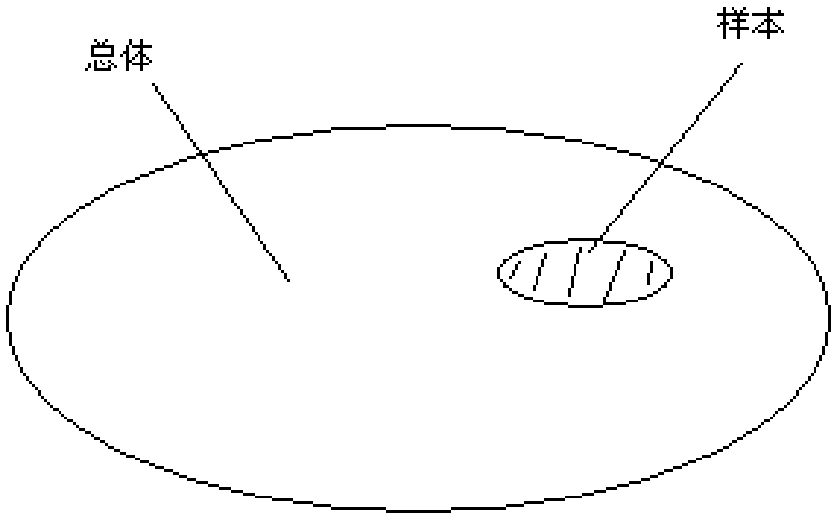
➢ 令$n(2|1)$为样本中来自π_1而误判为π_2的个数，$n(1|2)$为样本中来自π_2而误判为π_1的个数，则$P(2|1)$ 和$P(1|2)$ 可估计为

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

➢ 该方法简单、直观，且易于计算。但遗憾的是，它给出的误判概率的估计值通常偏低，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。

§6.3 距离判别法

距离判别准则：根据已知分类数据，分别计算各类重心，对任给一次观测，若它与第i类重心距离最近，就认为它来自第i类

设有两个总体时：

$$\begin{cases} X \in G_1 & \text{若} d(X,G_1) < d(X,G_2) \\ X \in G_2 & \text{若} d(X,G_1) > d(X,G_2) \\ \text{待判} & \text{若} d(X,G_1) = d(X,G_2) \end{cases}$$

(1) $\Sigma_1 = \Sigma_2 = \Sigma$: $W(X) = D(X,G_2) - D(X,G_1)$ 线性判别

(2) $\Sigma_1 \neq \Sigma_2$: $W(X) = D(X,G_2) - D(X,G_1)$ 二次判别

6.3.1 两总体距离判别

设$\mu_1, \mu_2, \Sigma_1, \Sigma_2$分别为两个类$G_1, G_2$的均值向量和协方差阵

马氏距离：

$$D(X, G_i) = (X - \mu_i)'(\Sigma_i)^{-1}(X - \mu_i),\ i = 1, 2$$

判别准则：

$$\begin{cases} \text{当} D(X, G_1) < D(X, G_2), & \text{则} X \in G_1, \\ \text{当} D(X, G_1) > D(X, G_2), & \text{则} X \in G_2, \\ \text{当} D(X, G_1) = D(X, G_2), & \text{待判。} \end{cases}$$

6.3.1 两总体距离判别

一、等方差阵： 直线判别 当$\Sigma_1 = \Sigma_2 = \Sigma$时

$$
\begin{aligned}
W(X) &= D(X, G_2) - D(X, G_1) \\
&= (X-\mu_2)'\Sigma^{-1}(X-\mu_2) - (X-\mu_1)'\Sigma^{-1}(X-\mu_1) \\
&= 2X'\Sigma^{-1}(\mu_1-\mu_2) - (\mu_1+\mu_2)\Sigma^{-1}(\mu_1-\mu_2) \\
&= 2[X - 1/2(\mu_1+\mu_2)]'\Sigma^{-1}(\mu_1-\mu_2)
\end{aligned}
$$

$$W(X) = b_0 + b_1 X$$

$$b_0 = -1/2(\mu_1+\mu_2)'\Sigma^{-1}(\mu_1-\mu_2)$$

$$b_1 = \Sigma^{-1}(\mu_1-\mu_2)$$

$$a' = (\bar{X}_1 - \bar{X}_2)'S_p^{-1}$$

$$
\begin{cases}
\text{当} W(X) > 0\text{，则} X \in G_1 \\
\text{当} W(X) < 0\text{，则} X \in G_2 \\
\text{当} W(X) = 0\text{，待判}
\end{cases}
$$

6.3.1 两总体距离判别

二、异方差阵： 曲线判别 当$\Sigma_1 \neq \Sigma_2$时

$$W(X) = D(X, G_2) - D(X, G_1)$$

$$= (X-\mu_2)'(\Sigma_2)^{-1}(X-\mu_2) - (X-\mu_1)'(\Sigma_1)^{-1}(X-\mu_1)$$

二次判别函数 **qda** 的用法

qda(formula, data, ...)

formula 一个形如 **groups ~ x1 + x2 + ...** 的公式框架，**data** 数据框

(qda: quadratic discriminant analysis)

6.3.2 多总体距离判别

一、协方差矩阵相同： 线性判别

$$
\begin{aligned}
D(X,G_i) &= (X-\mu_i)'\Sigma^{-1}(X-\mu_i) \\
&= X'\Sigma^{-1}X - 2\mu_i'\Sigma^{-1}X + \mu_i'\Sigma^{-1}\mu_i \\
&= X'\Sigma^{-1}X - 2(b_iX + b_0) \\
&= X'\Sigma^{-1}X - 2Z_i
\end{aligned}
$$

$Z_i = b_0 + b_iX$ 当$Z_i = \max\limits_{1\le j\le k}(Z_j)$，则$X \in G_i$

6.3.2 多总体距离判别

二、协方差矩阵不同： 非线性判别

$$D(X,G_i)=(X-\mu_i)'\Sigma_i^{-1}(X-\mu_i)$$

$$\text{当}D(X,G_i)=\min_{1\le j\le k}D(X,G_j)\text{，则}X\in G_i$$

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

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

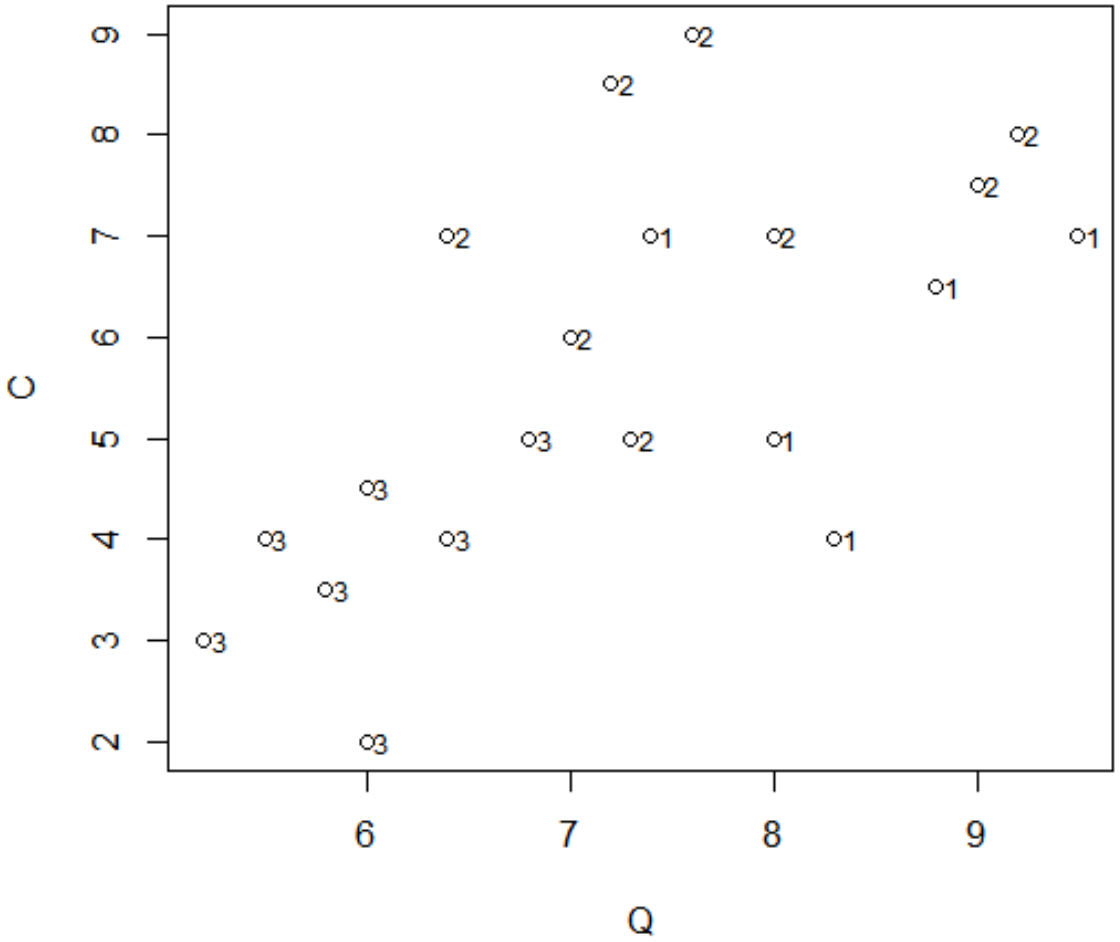
➢即使检验所需的正态性假定能够满足，检验的结果也只能作为重要的参考依据，而不宜作为决定性的依据，最终还是应视具体的情况而定。

- (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6.3】在某市场抽取20种牌子的电视机中，5种畅销，8种平销，另外7种滞销。按电视质量评分Q、功能评分C和销售价格P三项指标衡量，销售状态：1为畅销，2为平销，3为滞销。据此建立判别函数，并根据判别准则进行回判。

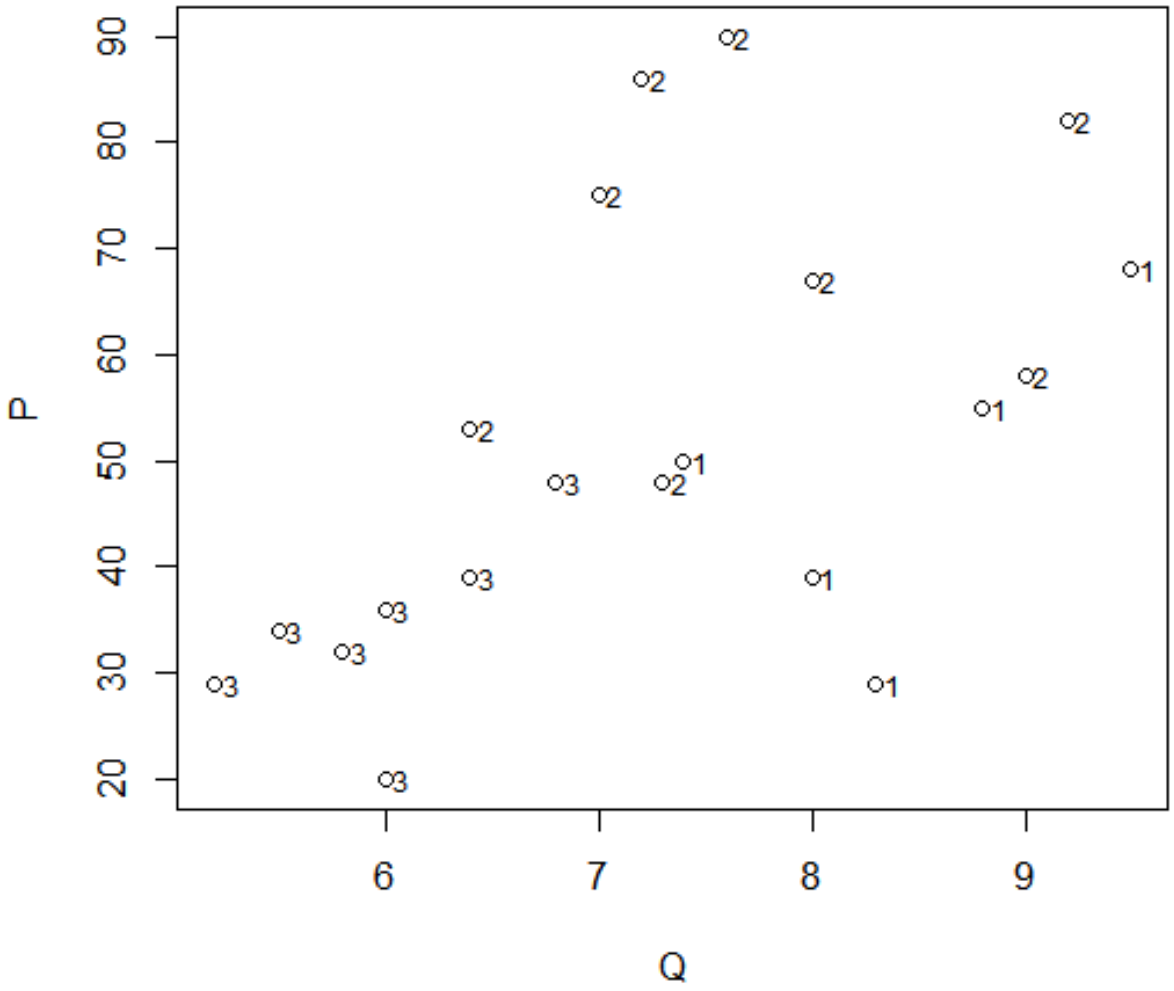
G	Q	C	P
1	8.3	4	29
1	9.5	7	68
1	8	5	39
1	7.4	7	50
1	8.8	6.5	55
2	9	7.5	58
2	7	6	75
2	9.2	8	82
2	8	7	67
2	7.6	9	90
2	7.2	8.5	86
2	6.4	7	53
2	7.3	5	48
3	6	2	20
3	6.4	4	39
3	6.8	5	48
3	5.2	3	29
3	5.8	3.5	32
3	5.5	4	34
3	6	4.5	36

```
d6.3=read.table("clipboard",header=T)
attach(d6.3)    #绑定数据
plot(Q,C);text(Q,C,G3,adj=-0.8,cex=0.75)
```

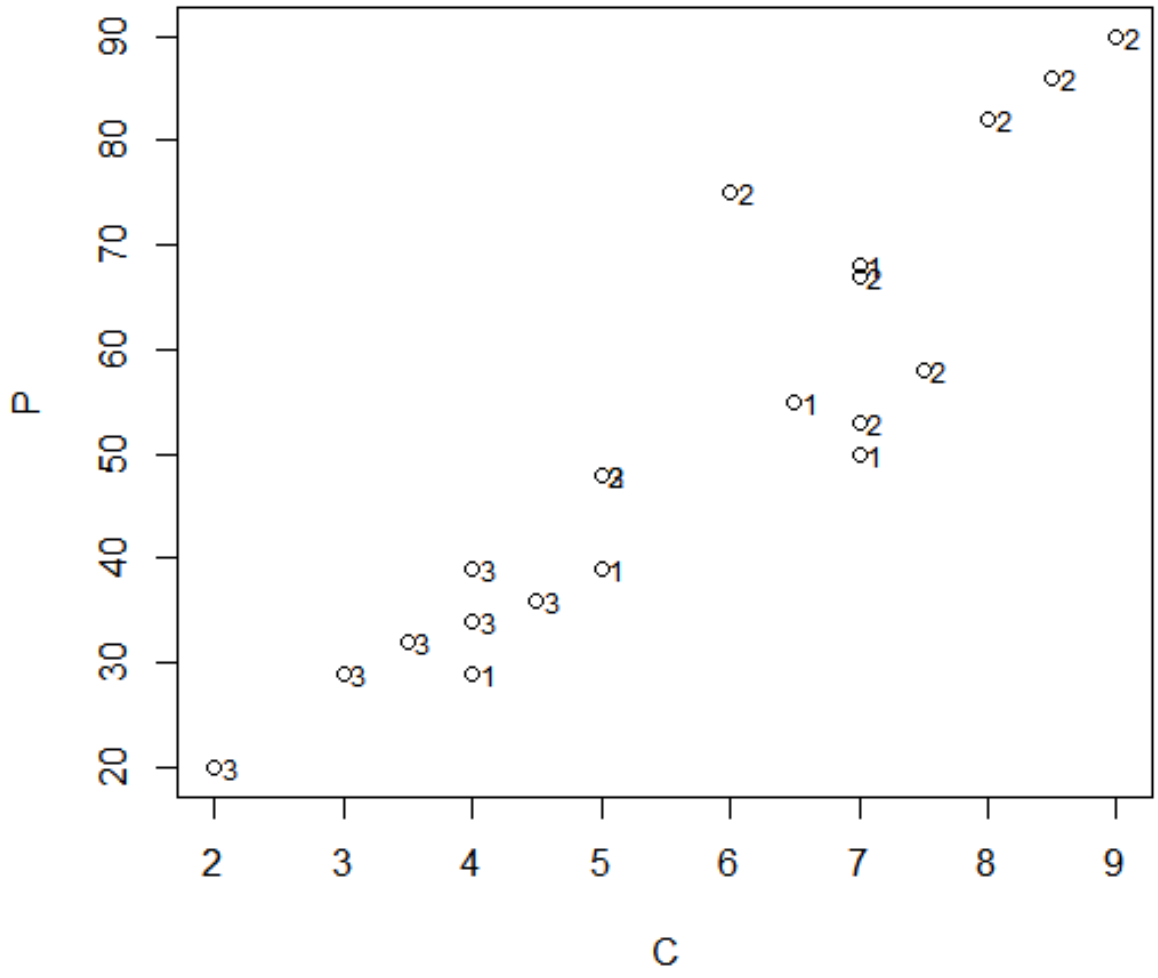


	Q	C	P	G3
1	8.3	4.0	29	1
2	9.5	7.0	68	1
3	8.0	5.0	39	1
4	7.4	7.0	50	1
5	8.8	6.5	55	1
6	9.0	7.5	58	2
7	7.0	6.0	75	2
8	9.2	8.0	82	2
9	8.0	7.0	67	2
10	7.6	9.0	90	2
11	7.2	8.5	86	2
12	6.4	7.0	53	2
13	7.3	5.0	48	2
14	6.0	2.0	20	3
15	6.4	4.0	39	3
16	6.8	5.0	48	3
17	5.2	3.0	29	3
18	5.8	3.5	32	3
19	5.5	4.0	34	3
20	6.0	4.5	36	3

plot(Q,P);text(Q,P,G3,adj=-0.8,cex=0.75)



```
plot(C,P);text(C,P,G3,adj=-0.8,cex=0.75)
```



1. 线性判别（等方差）

```
ld3=lda(G3~Q+C+P); ld3
```

Coefficients of linear disc

	LD1	LD2
Q	-0.8117	0.8841
C	-0.6309	0.2013
P	0.0158	-0.0878

```
lp3=predict(ld3); lG3=lp3$class
data.frame(G3,lG3)
```

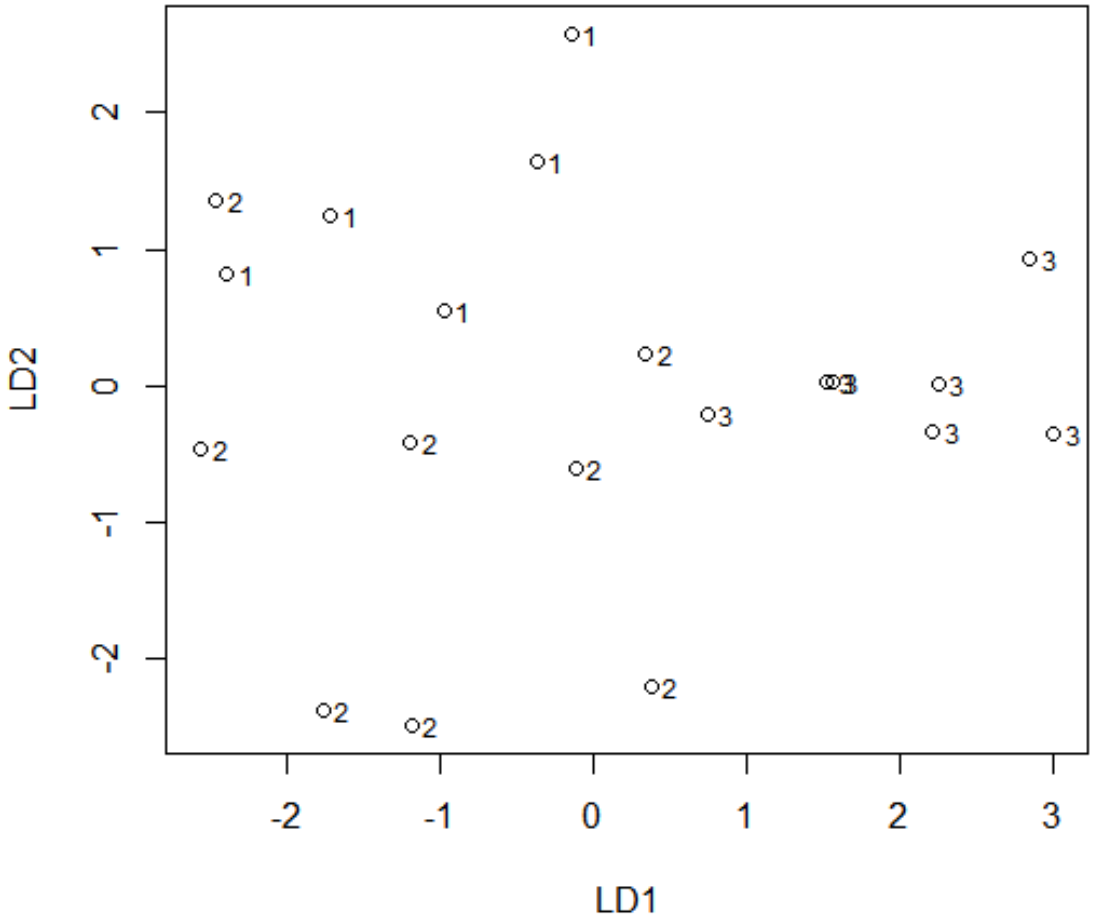
	G3	lG3
1	1	1
2	1	1
3	1	1
4	1	1
5	1	1
6	2	1
7	2	2
8	2	2
9	2	2
10	2	2
11	2	2
12	2	2
13	2	3
14	3	3
15	3	3
16	3	3
17	3	3
18	3	3
19	3	3
20	3	3

```
ltab3=table(G3,lG3)
ltab3
```

```
    lG3
G3   1 2 3
1    5 0 0
2    1 6 1
3    0 0 7
```

[1] 0.9

```
plot(lp3$x); text(lp3$x[,1],lp3$x[,2],lG3,adj=-0.8,cex=0.75)
```



```
predict(ld3,data.frame(Q=8,C=7.5,P=65))
```

```
$class
[1] 2
Levels: 1 2 3
```

2. 二次判别（异方差）

```
qd3=qda(G3~Q+C+P); qd3
qp3=predict(qd3)
qG3=qp3$class
data.frame(G3,lG3,qG3)
qtab3=table(G3,lG3))
qtab3
```

```
    qG3
G3    1 2 3
   1  5 0 0
   2  0 7 1
   3  0 0 7
```

[1] 0.95

```
predict(qd3,data.frame(Q=8,C=7.5,P=65))
```

```
$class
[1] 2
Levels: 1 2 3
```

	G3	lG3	qG3
1	1	1	1
2	1	1	1
3	1	1	1
4	1	1	1
5	1	1	1
6	2	1	2
7	2	2	2
8	2	2	2
9	2	2	2
10	2	2	2
11	2	2	2
12	2	2	2
13	2	3	3
14	3	3	3
15	3	3	3
16	3	3	3
17	3	3	3
18	3	3	3
19	3	3	3
20	3	3	3

§6.4 贝叶斯判别法

一 、标准的Bayes判别

办公室新来了一个雇员小王，小王是好人还是坏人大家都在猜测。按人们主观意识，一个人是好人或坏人的概率均为0.5。坏人总是要做坏事，好人总是做好事，偶尔也会做一件坏事，一般好人做好事的概率为0.9，坏人做好事的概率为0.2，一天，小王做了一件好事，小王是好人的概率有多大，你现在把小王判为何种人。

$$P(\text{好人}/\text{做好事})$$

$$= \frac{P(\text{好人})P(\text{做好事}/\text{好人})}{P(\text{好人})P(\text{做好事}/\text{好人}) + P(\text{坏人})P(\text{做好事}/\text{坏人})}$$

$$= \frac{0.5\times 0.9}{0.5\times 0.9 + 0.5\times 0.2} = 0.82$$

$$P(\text{坏人}/\text{做好事})$$

$$= \frac{P(\text{坏人})P(\text{做好事}/\text{坏人})}{P(\text{好人})P(\text{做好事}/\text{好人}) + P(\text{坏人})P(\text{做好事}/\text{坏人})}$$

$$= \frac{0.5 \times 0.2}{0.5 \times 0.9 + 0.5 \times 0.2} = 0.18$$

若有k个总体，样本来自第k个总体的后验概率：

$$p(j/x) = \frac{q_j p_j(x)}{\sum_{i=1}^{k} q_i p_i(x)} \qquad j = 1, 2, \cdots, k$$

当第j个后验概率最大时，就判定*x*来自第j个总体

二、 考虑错判损失的Bayes判别分析

$$E(g/x)=\sum_{j\neq i}\frac{q_j p_j(x)}{\sum_{i=1}^{k} q_i p_i(x)}L(g/j) \qquad L(g/j)=\begin{cases}0 & g=j\\ 1 & g\neq j\end{cases}$$

如$E(g/x)=\min\limits_{1\le j\le k} E(j/x)$时，判$x$来自第$g$总体

实际当中计算损失函数不容易，则通常假定损失相同，于是就等价于寻找最大后验概率：

$$P(g/x)\xrightarrow{g}\max \iff E(g/x))\xrightarrow{g}\min$$

实践当中并不是直接计算后验概率，而是计算一个简单判别函数，即哪个判别函数值大，就归哪一类。

最小期望误判代价法

三、 正态总体的Bayes判别

$$p(j/x)=\frac{q_j p_j(x)}{\sum_{i=1}^{k} q_i p_i(x)}$$

1、Bayes判别函数求解

k个总体的先验概率 $q_1, q_2, \cdots, q_k$

密度函数分别为

$$p_j(x)=(2\pi)^{-p/2}|\Sigma_j|^{-1/2}exp[-\frac{1}{2}(x-\mu_j)'\Sigma_j^{-1}(x-\mu_j)]$$

$$ln[q_j p_j(x)]=lnq_j-\tfrac{1}{2}ln(2\pi)^p-\tfrac{1}{2}ln|\Sigma_j|-\tfrac{1}{2}x'\Sigma_j^{-1}x-\tfrac{1}{2}\mu_j'\Sigma_j^{-1}\mu_j+x'\Sigma_j^{-1}\mu_j$$

$$Z(j/x)=lnq_j-\tfrac{1}{2}ln|\Sigma_j|-\tfrac{1}{2}x'\Sigma_j^{-1}x-\tfrac{1}{2}\mu_j'\Sigma_j^{-1}\mu_j+x'\Sigma_j^{-1}\mu_j$$

当$Z(j/x)=\max\limits_{1\le j\le k} Z(j/x)$时，判$x$来自第$j$总体

2、协方差阵相等情形

$$Z(j/x) = lnq_j - \frac{1}{2}ln|\Sigma_j| - \frac{1}{2}x'\Sigma_j^{-1}x - \frac{1}{2}\mu_j'\Sigma_j^{-1}\mu_j + x'\Sigma_j^{-1}\mu_j$$

$$Y(j/x) = lnq_j - \frac{1}{2}\mu_j'\Sigma^{-1}\mu_j + x'\Sigma^{-1}\mu_j$$

当$Y(j/x) = \max\limits_{1\le j\le k} Y(j/x)$时，判$x$来自第$j$总体

3、后验概率的计算

$$p(j/x) = \frac{q_j p_j(x)}{\sum_{i=1}^{k} q_i p_i(x)} = \frac{exp[y(j/x)]}{\sum_{i=1}^{k} exp[y(i/x)]}$$

[例6.3] 在某市场抽取20种牌子的电视机中，5种畅销，8种平销，另外7种滞销。按电视质量评分Q、功能评分C和销售价格P三项指标衡量，销售状态：1为畅销，2为平销，3为滞销。据此建立判别函数，并根据判别准则进行回判。

G	Q	C	P
1	8.3	4	29
1	9.5	7	68
1	8	5	39
1	7.4	7	50
1	8.8	6.5	55
2	9	7.5	58
2	7	6	75
2	9.2	8	82
2	8	7	67
2	7.6	9	90
2	7.2	8.5	86
2	6.4	7	53
2	7.3	5	48
3	6	2	20
3	6.4	4	39
3	6.8	5	48
3	5.2	3	29
3	5.8	3.5	32
3	5.5	4	34
3	6	4.5	36

【例6.4】对例6.3数据应用Bayes判别法进行判别

（1）先验概率相等： $q_1 = q_2 = q_3 = 1/3$

```
ld41=lda(G3~Q+C+P,prior=c(1,1,1)/3)
```

```
Call:
lda(G3~Q+C+P, prior=c(1,1,1)/3)

Prior probabilities of groups:
    1     2     3
0.333 0.333 0.333

Coefficients of linear discrim
      LD1     LD2
Q -0.9231  0.7671
C -0.6522  0.1148
P  0.0274 -0.0848
```

（2）先验概率不等　　$q_1 = 5/20, q_2 = 8/20, q_3 = 7/20$

```
ld42=lda(G3~Q+C+P,prior=c(5,8,7)/20); ld42
```

```
Call:
lda(G3~Q+C+P, prior=c(5,8,7)/20)

Prior probabilities of groups:
   1    2    3
0.25 0.40 0.35

Coefficients of linear discrim
      LD1     LD2
Q -0.8117  0.8841
C -0.6309  0.2013
P  0.0158 -0.0878
```

两种结果比较：

```
Z1=predict(ld41); Z2=predict(ld42)
data.frame(G3,ld41G=Z1$class,ld42G=Z2$class)
```

```
T1=table(G3,Z1$class);T1
```

```
G   1 2 3
  1 5 0 0
  2 1 6 1
  3 0 0 7
```

```
sum(diag(T1))/sum(T1)
```

```
[1] 0.9
```

```
T2=table(G3,Z2$class);T2
```

```
G   1 2 3
  1 5 0 0
  2 1 6 1
  3 0 0 7
```

```
sum(diag(T2))/sum(T2)
```

```
[1] 0.9
```

	G3	ld41G	ld42G				
1	1	1	1	11	2	2	2
2	1	1	1	12	2	2	2
3	1	1	1	13	2	3	3
4	1	1	1	14	3	3	3
5	1	1	1	15	3	3	3
6	2	1	1	16	3	3	3
7	2	2	2	17	3	3	3
8	2	2	2	18	3	3	3
9	2	2	2	19	3	3	3
10	2	2	2	20	3	3	3

两种结果比较

round(Z1$post*100,2)

	1	2	3
1	98.26	0.56	1.19
2	79.42	20.57	0.01
3	93.72	4.31	1.97
4	65.37	33.71	0.91
5	90.52	9.44	0.05
6	92.78	7.21	0.00
7	0.33	86.32	13.34
8	17.75	82.25	0.01
9	18.47	81.05	0.48
10	0.28	99.70	0.02
11	0.22	99.69	0.09
12	11.12	77.98	10.90
13	29.18	32.50	38.32

round(Z2$post*100,2)

	1	2	3
1	97.47	0.88	1.65
2	70.70	29.29	0.01
3	90.66	6.67	2.67
4	54.21	44.73	1.06
5	85.65	14.29	0.06
6	88.93	11.06	0.01
7	0.21	87.90	11.89
8	11.88	88.11	0.01
9	12.41	87.14	0.45
10	0.18	99.81	0.02
11	0.14	99.78	0.08
12	7.36	82.55	10.09
13	21.64	38.57	39.79

```
predict(ld41,data.frame(Q=8,C=7.5,P=65))
$class
[1] 2
Levels: 1 2 3

$posterior
    1     2      3
1 0.3 0.698 0.0018
```

```
predict(ld42,data.frame(Q=8,C=7.5,P=65))
$class
[1] 2
Levels: 1 2 3

$posterior
      1     2       3
1 0.211 0.787 0.00178
```

两种结果比较

小结

1. 判别分析方法是按已知所属组的样本确定判别函数，制定判别规则，然后再判断每一个新样品应属于哪一类。

2. 常用的判别方法有Fisher判别、距离判别、贝叶斯判别等，每个方法根据其出发点不同各有其特点。

3. Fisher类判别对判别变量的分布类型并无要求，而Bayes类判别要变量的分布类型。因此，Fisher类判别较Bayes类判别简单一些。

4. 当两个总体时，若它们的协方差矩阵相同，则距离判别和Fisher判别等价。当变量服从正态分布时，它们还和Bayes判别等价。