

## 4.5. Discriminant analysis

### 1. Separation:

Suppose there are  $k$  populations,

$$\begin{aligned} x_{11}, x_{12}, \dots, x_{1n_1} &: \text{population 1} \\ x_{21}, x_{22}, \dots, x_{2n_2} &: \text{population 2} \\ &\vdots \\ x_{k1}, x_{k2}, \dots, x_{kn_k} &: \text{population } k \end{aligned}$$

where  $n_1 + \dots + n_k = n$ .

Let  $\bar{x}_j$  be the sample mean for the population  $j$ ,  $j = 1, \dots, k$ , and

$$\bar{x} = \frac{\sum_{j=1}^k \sum_{i=1}^{n_j} x_{ji}}{n}.$$

The sample between matrix

$$B = \sum_{j=1}^k n_j (\bar{x}_j - \bar{x})(\bar{x}_j - \bar{x})^t.$$

The sample within group matrix  $W$  is

$$W = \sum_{j=1}^k \sum_{i=1}^{n_j} (x_{ji} - \bar{x}_j)(x_{ji} - \bar{x}_j)^t$$

**Note:**

$$S_{pooled} = \frac{W}{n-k} = \frac{\sum_{j=1}^k \sum_{i=1}^{n_j} (x_{ji} - \bar{x}_j)(x_{ji} - \bar{x}_j)^t}{n-k} \equiv \text{pooled estimate}$$

**Important result:**

Let  $e_1, e_2, \dots, e_s$  be the eigenvectors of  $W^{-1}B$  corresponding to the eigenvalues  $\hat{\lambda}_1 \geq \hat{\lambda}_2 \geq \dots \geq \hat{\lambda}_s > 0$ . Then,  $\hat{a}_j, j = 1, \dots, s$ , are the scaled eigenvectors satisfying  $\hat{a}_j^t S_{pooled} \hat{a}_j = 1$ . That is,

$$\hat{a}_j = \frac{e_j}{\sqrt{\hat{e}_j^t S_{pooled} \hat{e}_j}}$$

**Example (Splus):**

```
spe=c(rep("s",50),rep("c",50),rep("v",50))
irdata=list(spe=spe,ir=ir)
irdata$spe
```

```

irdata$ir
xmean1=apply(irdata$ir[irdata$spe=="s",],2,mean)    #  $\bar{x}_1$ 
xmean2=apply(irdata$ir[irdata$spe=="c",],2,mean)    #  $\bar{x}_2$ 
xmean3=apply(irdata$ir[irdata$spe=="v",],2,mean)    #  $\bar{x}_3$ 
xmean=apply(irdata$ir,2,mean)                         #  $\bar{x}$ 
b1=50*(xmean1-xmean)%*%t(xmean1-xmean)             #  $n_1(\bar{x}_1 - \bar{x})(\bar{x}_1 - \bar{x})^t$ 
b2=50*(xmean2-xmean)%*%t(xmean2-xmean)             #  $n_2(\bar{x}_2 - \bar{x})(\bar{x}_2 - \bar{x})^t$ 
b3=50*(xmean3-xmean)%*%t(xmean3-xmean)             #  $n_3(\bar{x}_3 - \bar{x})(\bar{x}_3 - \bar{x})^t$ 

b=b1+b2+b3                                         #  $B = \sum_{j=1}^3 n_j (\bar{x}_j - \bar{x})(\bar{x}_j - \bar{x})^t$ 

sum1=49*var(irdata$ir[irdata$spe=="s",])           #  $\sum_{i=1}^{50} (x_{1i} - \bar{x}_1)(x_{1i} - \bar{x}_1)^t$ 
sum2=49*var(irdata$ir[irdata$spe=="c",])           #  $\sum_{i=1}^{50} (x_{2i} - \bar{x}_2)(x_{2i} - \bar{x}_2)^t$ 
sum3=49*var(irdata$ir[irdata$spe=="v",])           #  $\sum_{i=1}^{50} (x_{3i} - \bar{x}_3)(x_{3i} - \bar{x}_3)^t$ 
w=sum1+sum2+sum3                                    #  $W$ 
invw=solve(w)                                       #  $W^{-1}$ 
spool=w/(50+50+50-3)                               #  $S_{pooled} = W / (n_1 + n_2 + n_3 - 3)$ 
eectors=eigen(invW%*%b)$vectors                  #  $e_1, e_2, e_3, e_4$ 
## By important result

```

$$\# \hat{a}_1 = \frac{e_1}{\sqrt{\hat{e}_1^t S_{pooled} \hat{e}_1}}, \hat{a}_2 = \frac{e_2}{\sqrt{\hat{e}_2^t S_{pooled} \hat{e}_2}}$$

```

a1hat=eectors[,1]/sqrt(t(eectors[,1])%*%spool%*%eectors[,1])
a2hat=eectors[,2]/sqrt(t(eectors[,2])%*%spool%*%eectors[,2])

```

```

a1x=ir%*%a1hat                                     #  $\hat{a}_1^t x_{ji}$ 
a2x=ir%*%a2hat                                     #  $\hat{a}_2^t x_{ji}$ 
plot(a1x,a2x)                                     # separation based on the first two sample discriminant

```

## 2. Classification:

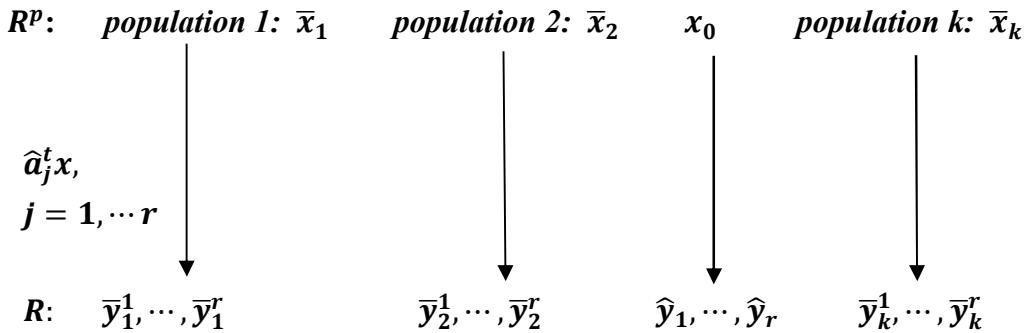
Fisher's classification method for several populations is as follows:

For an observation  $x_0$ , Fisher's classification procedure based on the first  $r \leq s$  sample discriminants is to allocate  $x_0$  to the population  $l$  if

$$\sum_{j=1}^r (\hat{y}_j - \bar{y}_l^j)^2 = \sum_{j=1}^r [\hat{a}_j^t (x_0 - \bar{x}_l)]^2 \leq \sum_{j=1}^r [\hat{a}_j^t (x_0 - \bar{x}_i)]^2 = \sum_{j=1}^r (\hat{y}_j - \bar{y}_i^j)^2, i \neq l,$$

where  $\hat{y}_j = \hat{a}_j^t x_0$  and  $\bar{y}_i^j = \hat{a}_j^t \bar{x}_i, j = 1, \dots, r; i = 1, \dots, k$ .

**Intuition of Fisher's method:**



**Example (Splus):**

```

x0=c(5,3,1,1)
yhat1=a1hat%*%x0                      # \hat{y}_1
yhat2=a2hat%*%x0                      # \hat{y}_2
y1bar1=a1hat%*%xmean1                  # \bar{y}_1^1
y1bar2=a2hat%*%xmean1                  # \bar{y}_1^2
y2bar1=a1hat%*%xmean2                  # \bar{y}_2^1
y2bar2=a2hat%*%xmean2                  # \bar{y}_2^2
y3bar1=a1hat%*%xmean3                  # \bar{y}_3^1
y3bar2=a2hat%*%xmean3                  # \bar{y}_3^2
dis1=(yhat1-y1bar1)^2+(yhat2-y1bar2)^2   # \sum_{j=1}^2 (\hat{y}_j - \bar{y}_1^j)^2
dis2=(yhat1-y2bar1)^2+(yhat2-y2bar2)^2   # \sum_{j=1}^2 (\hat{y}_j - \bar{y}_2^j)^2
dis3=(yhat1-y3bar1)^2+(yhat2-y3bar2)^2   # \sum_{j=1}^2 (\hat{y}_j - \bar{y}_3^j)^2
c(dis1,dis2,dis3)

```

**Note:**

To calculate the error rate as using the linear discriminant method for the iris data, the code is given below:

**Example (Splus):**

```

### Discriminant Analysis
y1bar1=a1hat%*%xmean1
y1bar2=a2hat%*%xmean1
y2bar1=a1hat%*%xmean2
y2bar2=a2hat%*%xmean2
y3bar1=a1hat%*%xmean3

```

```

y3bar2=a2hat%*%xmean3
cls=rep(0,150)
for(i in 1:nrow(ir))
{
  x0=ir[i,]
  yhat1=a1hat%*%x0
  yhat2=a2hat%*%x0
  dis1=(yhat1-y1bar1)^2+(yhat2-y1bar2)^2
  dis2=(yhat1-y2bar1)^2+(yhat2-y2bar2)^2
  dis3=(yhat1-y3bar1)^2+(yhat2-y3bar2)^2
  cls[i]=order(c(dis1,dis2,dis3))[1]
}
correct1=c(rep(1,50),rep(2,50),rep(3,50))
errorrate1=sum(cls!=correct1)/length(cls)
errorrate1

### Cluster Method
correct2=c(rep(2,50),rep(3,50),rep(1,50))
errorrate2=sum(kmeans(ir,3)$cluster!=correct2)/length(correct2)
errorrate2

```