# MVA_HW3

陈子睿 15220212202842

4/22/24

## 1 Discrimination and Classification

(a) The discriminant function coefficient is given by:

$$a = S_{pl}^{-1}(\bar{y}_1 - \bar{y}_2) = (-2, 0)'.$$

Hence the discriminant function is:

$$Z = a'Y = -2Y_1.$$

The calculation is done by R, here's the code:

```
Y1 <- matrix(c(3,7,2,4,4,7), nrow = 3, ncol = 2, byrow = TRUE)
Y2 <- matrix(c(6,9,5,7,4,8), nrow = 3, ncol = 2, byrow = TRUE)
S1 <- cov(Y1)
S2 <- cov(Y2)
Y1_mean <- colMeans(Y1)
Y2_mean <- colMeans(Y2)
Spl <- (2 * S1 + 2 * S2)/4
dis_coefficient <- solve(Spl) %*% (Y1_mean - Y2_mean)
dis_coefficient
     [,1]
[1,]   -2
[2,]    0
```

(b) To classify the new observation into group $G_1$ or $G_2$, by the Fisher's allocation rule for two groups, we need to compare $z_0$ to $\frac{1}{2}(\bar{Z}_1 + \bar{X}_2)$. We can compute by R:

```
> y0 <- c(2,7)
> Z1_mean <- t(dis_coefficient) %*% Y1_mean
> Z2_mean <- t(dis_coefficient) %*% Y2_mean
> z0 <- t(dis_coefficient) %*% y0
> (Z1_mean + Z2_mean) / 2
     [,1]
```

```
[1,]   -8
> z0
     [,1]
[1,]   -4
```

We can see that $z_0 > \frac{1}{2}(\bar{Z}_1 + \bar{X}_2)$, therefore we allocate $y_0$ to $G_1$.

# 2 Some propositions based on Fisher's LDA

(a) The statistical distance between the transformed group is:

$$\frac{(\bar{z}_1 - \bar{z}_2)^2}{S_z^2} = \frac{[a'(\bar{y}_1 - \bar{y}_2)]^2}{a' S_{pl} a}.$$

The maximum occurs when

$$a = S_{pl}^{-1}(\bar{y}_1 - \bar{y}_2).$$

In such case, the maximized statistical distance is

$$d_{max}^2 = (\bar{y}_1 - \bar{y}_2)' S_{pl}^{-1}(\bar{y}_1 - \bar{y}_2)$$

However, the statistical distance between $\bar{y}_1$ and $\bar{y}_2$ is

$$d^2(\bar{y}_1, \bar{y}_2) = (\bar{y}_1 - \bar{y}_2)' S_{mix}^{-1}(\bar{y}_1 - \bar{y}_2),$$

where $S_{mix} = \frac{n_1+n_2-2}{n_1+n_2-1} S_{pl}$. Then

$$\begin{aligned} d^2(\bar{y}_1, \bar{y}_2) &= (\bar{y}_1 - \bar{y}_2)' S_{mix}^{-1}(\bar{y}_1 - \bar{y}_2) \\ &= (\bar{y}_1 - \bar{y}_2)' (\frac{n_1 + n_2 - 2}{n_1 + n_2 - 1} S_{pl})^{-1}(\bar{y}_1 - \bar{y}_2) \\ &= (\bar{y}_1 - \bar{y}_2)' \frac{n_1 + n_2 - 1}{n_1 + n_2 - 2} S_{pl}^{-1}(\bar{y}_1 - \bar{y}_2) \\ &= \frac{n_1 + n_2 - 1}{n_1 + n_2 - 2}(\bar{y}_1 - \bar{y}_2)' S_{pl}^{-1}(\bar{y}_1 - \bar{y}_2). \end{aligned}$$

Therefore, we've shown that

$$d(\bar{y}_1, \bar{y}_2) = \sqrt{\frac{n_1 + n_2 - 1}{n_1 + n_2 - 2}} d_{max},$$

i.e.

$$d(\bar{y}_1, \bar{y}_2) \propto d_{max}.$$

(b) Recall that the Fisher's allocation rule is allocating the new observation $y_0$ to $G_1$ if

$$(\bar{y}_1 - \bar{y}_2)' S_{pl}^{-1} y_0 \geq \frac{1}{2}(\bar{y}_1 - \bar{y}_2)' S_{pl}^{-1}(\bar{y}_1 + \bar{y}_2).$$

The LHS equals to

$$\begin{aligned} LHS &= (\bar{y}_1 - \bar{y}_2)' S_{pl}^{-1} y_0 \\ &= (\bar{y}_1 - y_0)' S_{pl}^{-1} y_0 - (\bar{y}_2 - y_0)' S_{pl}^{-1} y_0. \end{aligned}$$

The RHS equals to

$$\begin{aligned}
RHS &= \frac{1}{2}(\bar{y}_1 - \bar{y}_2)' S_{pl}^{-1}(\bar{y}_1 + \bar{y}_2) \\
&= \frac{1}{2}[(\bar{y}_1 - y_0) - (\bar{y}_2 - y_0)]' S_{pl}^{-1}(\bar{y}_1 + \bar{y}_2) \\
&= \frac{1}{2}(\bar{y}_1 - y_0)' S_{pl}^{-1}(\bar{y}_1 + \bar{y}_2) - \frac{1}{2}(\bar{y}_2 - y_0)' S_{pl}^{-1}(\bar{y}_1 + \bar{y}_2) \\
&= \frac{1}{2}(\bar{y}_1 - y_0)' S_{pl}^{-1}(\bar{y}_1 - \bar{y}_2) + (\bar{y}_1 - y_0)' S_{pl}^{-1}\bar{y}_2 - \frac{1}{2}(\bar{y}_2 - y_0)' S_{pl}^{-1}(\bar{y}_1 - \bar{y}_2) + (\bar{y}_2 - y_0)' S_{pl}^{-1}\bar{y}_2.
\end{aligned}$$

Rearranging the terms and the inequality becomes

$$(y_0 - \bar{y}_2)' S_{pl}^{-1}(y_0 - \bar{y}_2) \geq (y_0 - \bar{y}_1)' S_{pl}^{-1}(y_0 - \bar{y}_1).$$

(c) To verify that the solution of two-population LDA is indeed a special case of the several-population case, first we recall the objective function of the several-population case:

$$a = argmax_a \frac{a'Ba}{a'Wa} = \frac{a'\left[\sum_{k=1}^{g}(\bar{y}_k - \bar{y})(\bar{y}_k - \bar{y})'\right]a}{a'\left[\sum_{k=1}^{g}\sum_{i=1}^{n_k}(\bar{y}_{ki} - \bar{y}_k)(\bar{y}_{ki} - \bar{y}_k)'\right]a}.$$

While the denominator is the quadratic form of the "between-group variance", the numerator is the quadratic form of the "within-group variance". In the two-population case, the "between-group variance" reduces to

$$\begin{aligned}
B &= \sum_{k=1}^{g}(\bar{y}_k - \bar{y})(\bar{y}_k - \bar{y})' = [\bar{y}_1 - (\bar{y}_1 + \bar{y}_2)/2][\bar{y}_1 - (\bar{y}_1 + \bar{y}_2)/2]' + [\bar{y}_2 - (\bar{y}_1 + \bar{y}_2)/2][\bar{y}_2 - (\bar{y}_1 + \bar{y}_2)/2]' \\
&= \frac{1}{4}(\bar{y}_1 - \bar{y}_2)(\bar{y}_1 - \bar{y}_2)' + \frac{1}{4}(\bar{y}_2 - \bar{y}_1)(\bar{y}_2 - \bar{y}_1)' \\
&= \frac{1}{2}(\bar{y}_1 - \bar{y}_2)(\bar{y}_1 - \bar{y}_2)'.
\end{aligned}$$

And the "within-group variance" reduces to

$$\begin{aligned}
W &= \left[\sum_{k=1}^{g}\sum_{i=1}^{n_k}(\bar{y}_{ki} - \bar{y}_k)(\bar{y}_{ki} - \bar{y}_k)'\right] \\
&= (n_1 - 1)S_1 + (n_2 - 1)S_2 \\
&= (n_1 + n_2 - 2)S_{pl}.
\end{aligned}$$

By the FOC of the objective function,

$$\frac{\partial}{\partial a}\frac{a'Ba}{a'Wa} = 0 \iff W^{-1}Ba = \lambda a.$$

Which indicates that $a$ is the corresponding eigenvector of a positive eigenvalue of the matrix $W^{-1}B$. We let $\gamma = (\bar{y}_1 - \bar{y}_2)'a$ be a scalar, thus $Ba = \frac{1}{2}(\bar{y}_1 - \bar{y}_2)(\bar{y}_1 - \bar{y}_2)'a = \frac{1}{2}\gamma(\bar{y}_1 - \bar{y}_2)$. Therefore the FOC becomes $\frac{1}{2}W^{-1}\gamma(\bar{y}_1 - \bar{y}_2) = \lambda a$, namely $a = \frac{\gamma}{2\lambda(n_1+n_2-2)}S_{pl}^{-1}(\bar{y}_1 - \bar{y}_2)$. Since the discriminant coefficient possesses the scaling invariance property, $a = S_{pl}^{-1}(\bar{y}_1 - \bar{y}_2)$, indicating that the solution of two-population LDA is indeed a special case of the several-population case.

# 3 Formulate tests for profile analysis

(a) To test whether the profiles of two populations are parallel is equivalent to test that

$$H_0 : \exists c \in \mathbb{R} \text{ such that } \mu_1 = \mu_2 + c\mathbf{1}_p.$$

Or equivalently

$$H_0 : C\mu_1 = C\mu_2, \quad C = (-\mathbf{1}_{p-1}, I_{p-1})_{p-1,p}.$$

Consider transformation $y \to Cy$, then the test statistic is given by

$$T^2 = \frac{n_1 n_2}{n_1 + n_2}[C(\bar{y}_1 - \bar{y}_2)]'(CS_{pl}C')^{-1}[C(\bar{y}_1 - \bar{y}_2)] \sim T^2(p-1, n_1 + n_2 - 2).$$

The corresponding rejection region is given by

$$\left\{(y_1, y_2) : \frac{n_1 n_2}{n_1 + n_2}(\bar{y}_1 - \bar{y}_2)'C'(CS_{pl}C')^{-1}C(\bar{y}_1 - \bar{y}_2) > T^2_\alpha(p-1, n_1 + n_2 - 2)\right\}.$$

(b) To test whether the total measurements are the same between the two population is equivalent to test

$$H_0 : \mathbf{1}_p'\mu_1 = \mathbf{1}_p'\mu_2 \text{ v.s. } H_1 : \mathbf{1}_p'\mu_1 \neq \mathbf{1}_p'\mu_2.$$

Still, we consider transformation $y \to \mathbf{1}_p'y$, then the Hotelling's $T^2$ test reduces to the univariate t-test, the test statistic is

$$T = \frac{\mathbf{1}_p'(\bar{y}_1 - \bar{y}_2)}{\sqrt{\mathbf{1}_p'S_{pl}\mathbf{1}_p(\frac{1}{n_1} + \frac{1}{n_2})}} \sim t(n_1 + n_2 - 2).$$

The corresponding rejection region is given by

$$\left\{(y_1, y_2) : \left|\frac{\mathbf{1}_p'(\bar{y}_1 - \bar{y}_2)}{\sqrt{\mathbf{1}_p'S_{pl}\mathbf{1}_p(\frac{1}{n_1} + \frac{1}{n_2})}}\right| > t_{\alpha/2}(n_1 + n_2 - 2)\right\}.$$

(c) Assume the mean vector are parallel, to test whether the profiles are linear is equivalent to test

$$H_0 : C\mu_1 = -C\mu_2, C = \begin{pmatrix} 1 & -2 & 1 & & \\ & 1 & -2 & 1 & \\ & & \ddots & \ddots & \ddots \\ & & & 1 & -2 & 1 \end{pmatrix}$$

Consider transformation $y \to Cy$, then the test statistic is given by

$$T^2 = \frac{n_1 n_2}{n_1 + n_2}[C(\bar{y}_1 - \bar{y}_2)]'(CS_{pl}C')^{-1}[C(\bar{y}_1 - \bar{y}_2)] \sim T^2(p-1, n_1 + n_2 - 2).$$

The corresponding rejection region is given by

$$\left\{(y_1, y_2) : \frac{n_1 n_2}{n_1 + n_2}(\bar{y}_1 - \bar{y}_2)'C'(CS_{pl}C')^{-1}C(\bar{y}_1 - \bar{y}_2) > T^2_\alpha(p-1, n_1 + n_2 - 2)\right\}.$$

(d) Following (c), the test statistic is computed by

$$16.83613 > 6.428522,$$

which indicates that the linearity could be rejected with level 0.05.

# 4 Milk tranportation question

(a) To test for differences in the mean cost vectors at the significance level 0.01:

```
> library(Hotelling)
> y1 <- y[y$V4 == "gasoline",1:3]
> y2 <- y[y$V4 == "diesel",1:3]
> result <- hotelling.test(y1,y2)
> print(result)
Test stat:  50.913
Numerator df:  3
Denominator df:  55
P-value:  1e-07
```

Thus we reject $H_0$ at the significance level 0.01.

(b) The univariate tests can be conducted by

```
> t.test(y1$V1,y2$V1, var.equal = TRUE)

        Two Sample t-test

data:  y1$V1 and y2$V1
t = 1.9904, df = 57, p-value = 0.05135
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.01276331  4.23868118
sample estimates:
mean of x mean of y
 12.21861  10.10565
```

```
    > t.test(y1$V2,y2$V2, var.equal = TRUE)

        Two Sample t-test

data:  y1$V2 and y2$V2
t = -2.1791, df = 57, p-value = 0.03348
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -5.084617 -0.214731
sample estimates:
mean of x mean of y
  8.11250  10.76217
```

```
> t.test(y1$V3,y2$V3, var.equal = TRUE)

        Two Sample t-test

data:  y1$V3 and y2$V3
t = -6.2326, df = 57, p-value = 5.966e-08
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -11.333433  -5.821664
sample estimates:
mean of x mean of y
 9.590278 18.167826
```

From the R result above, we can't reject the $H_0$ that $\mu_{11} = \mu_{21}$ and $\mu_{12} = \mu_{22}$ at the significance level 0.01.

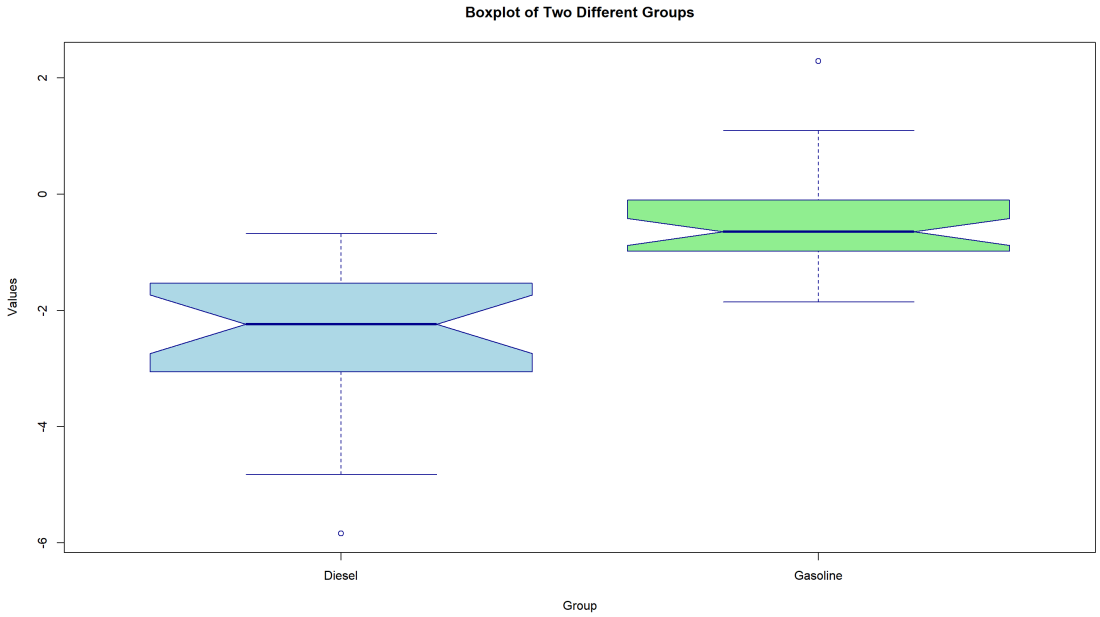
(c) To obtain the linear discriminant coefficient:

```
> library(MASS)
> fit <- lda(V4 ~ V1 + V2 + V3, data = y)
> print(fit)
Call:
lda(V4 ~ V1 + V2 + V3, data = y)
Prior probabilities of groups:
 diesel  gasoline
0.3898305 0.6101695
Group means:
                V1       V2        V3
diesel   10.10565 10.76217 18.167826
gasoline 12.21861  8.11250  9.590278
Coefficients of linear discriminants:
         LD1
V1  0.13374629
V2 -0.07030203
V3 -0.16739189
```

Then the discriminant function is given by $0.134Y_1 - 0.070Y_2 - 0.167Y_3$. Projecting the data by the discriminant function, and we draw a boxplot to see its variation:

```
> coefficients <- fit$scaling
> z1 <- as.matrix(y1) %*% coefficients
> z2 <- as.matrix(y2) %*% coefficients
> data <- c(z1,z2)
```

```
> group <- factor(c(rep("Gasoline", length(z1)), rep("Diesel", length(z2))))
> boxplot(data ~ group,
+         main = "Boxplot of Two Different Groups",
+         xlab = "Group",
+         ylab = "Values",
+         col = c("lightblue", "lightgreen"),
+         border = "darkblue",
+         notch = TRUE)
```




(d) Now we only consider the first 23 gasoline trucks and the 23 diesel trucks.

(d-a)

```
> result1 <- hotelling.test(y11,y2)
> result1
Test stat:  35.484
Numerator df:  3
Denominator df:  42
P-value:  1.461e-05
```

Still, the null hypothesis can be rejected at the significance level 0.01.

(d-b)

```
> t.test(y11$V1,y2$V1,paired = TRUE)
```

```
^^IPaired t-test
```

```
data:  y11$V1 and y2$V1
t = 1.8363, df = 22, p-value = 0.07986
alternative hypothesis: true mean difference is not equal to 0
95 percent confidence interval:
 -0.318098  5.235489
sample estimates:
mean difference
       2.458696
```

```
    > t.test(y11$V2,y2$V2,paired = TRUE)
```

```
^^IPaired t-test

data:  y11$V2 and y2$V2
t = -1.8108, df = 22, p-value = 0.08385
alternative hypothesis: true mean difference is not equal to 0
95 percent confidence interval:
 -5.5441801  0.3754845
sample estimates:
mean difference
      -2.584348
```

```
    > t.test(y11$V3,y2$V3,paired = TRUE)
```

```
^^IPaired t-test

data:  y11$V3 and y2$V3
t = -5.3766, df = 22, p-value = 2.127e-05
alternative hypothesis: true mean difference is not equal to 0
95 percent confidence interval:
 -11.196016  -4.963114
sample estimates:
mean difference
      -8.079565
```

From the results above,$_{d1} = 0$ and $_{d2} = 0$ cannot be rejected with level 0.01 while $_{d3} = 0$ can be rejected with level 0.01.

(d-c)

```
    > fit1 <- lda(V4 ~ V1 + V2 + V3, data = y[c(1:23,37:59),])
```

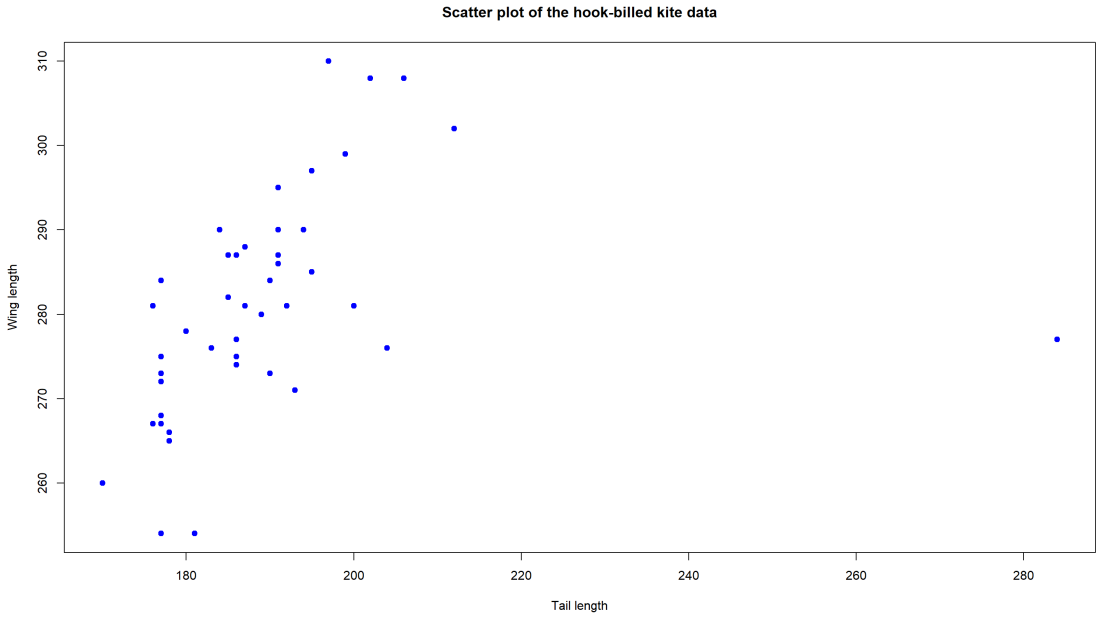
```
    > print(fit1)
Call:
lda(V4 ~ V1 + V2 + V3, data = y[c(1:23, 37:59), ])

Prior probabilities of groups:
  diesel gasoline
     0.5      0.5

Group means:
               V1        V2       V3
diesel   10.10565 10.762174 18.16783
gasoline 12.56435  8.177826 10.08826

Coefficients of linear discriminants:
           LD1
V1  0.13960617
V2 -0.06896152
V3 -0.15286688
```

Then the discriminant function is given by $0.140Y_1 - 0.069Y_2 - 0.153Y_3$.

# 5 Bird data question

(a) To plot the male hook-billed kite data as a scatterplot,

```
rm(list = ls())
y1 <- read.table(paste('C:/Users/Ray Chen/Desktop/MVA/male.DAT'), header =
↪  FALSE) #male data
y2 <- read.table(paste('C:/Users/Ray Chen/Desktop/MVA/female.DAT'), header =
↪  FALSE) #female data
plot(y1$V1, y1$V2, main="Scatter plot of the hook-billed kite data", xlab="Tail
↪  length", ylab="Wing length", pch=19, col="blue")
```




It's obvious that the 31th sample is an outlier.

(b-a) Before testing the equality of mean vectors of the male and female data, we need to eliminate the outlier,

```
> y11 <- y1[c(1:30,32:45),]
> library(Hotelling)
> result <- hotelling.test(y11,y2)
> print(result)
Test stat:  24.965
Numerator df:  2
Denominator df:  86
P-value:  1.944e-05
```

Thus the null hypothesis can be rejected at level 0.05. To do the LDA,

```
> library(MASS)
> y_mix <- data.frame(V1 = c(y11[,1], y2[,1]), V2 = c(y11[,2], y2[,2]), V3 =
↪  factor(c(rep("male", nrow(y11)), rep("female", nrow(y2)))))
> fit <- lda(V3~ V1 + V2 , data = y_mix)
> print(fit)
Call:
lda(V3 ~ V1 + V2, data = y_mix)

Prior probabilities of groups:
  female     male
0.505618 0.494382
```

```
Group means:
             V1       V2
female 193.6222 279.7778
male   187.1591 280.9545

Coefficients of linear discriminants:
           LD1
V1 -0.14784391
V2  0.08819563
```

Then the discriminant function is $-0.148Y_1 + 0.088Y_2$.

(b-b) Alternatively, we may try to interpret the outlier as a misprint and conduct the test with a more reasonable imputation/substitute.

```
> y12 <- y1
> y12[31,1] <- y12[31,1] - 80
> library(Hotelling)
> result1 <- hotelling.test(y12,y2)
> print(result1)
Test stat:  20.712
Numerator df:  2
Denominator df:  87
P-value:  0.0001016
```

Still, the null hypothesis can be rejected at level 0.05, again

```
> library(MASS)
> y_mix <- data.frame(V1 = c(y12[,1], y2[,1]), V2 = c(y12[,2], y2[,2]), V3 =
↪  factor(c(rep("male", nrow(y12)), rep("female", nrow(y2)))))
> fit <- lda(V3~ V1 + V2 , data = y_mix)
> print(fit)
Call:
lda(V3 ~ V1 + V2, data = y_mix)

Prior probabilities of groups:
female   male
   0.5    0.5

Group means:
             V1       V2
female 193.6222 279.7778
```

```
male   187.5333 280.8667

Coefficients of linear discriminants:
           LD1
V1 -0.14242150
V2  0.08473215
```

Then the discriminant function is $-0.142Y_1 + 0.085Y_2$.

To summarize, comparing scenarios (b-a) and (b-b), it appears that the treatment of outliers in the male hook-billed kite data has minimal impact in this instance.

(c) We still eliminate the outlier discussed in (a), the confidence region is

$$\left\{\mu_1 - \mu_2 : [(\mu_1 - \mu_2) - (\bar{y}_1 - \bar{y}_2)]' S_{pl}^{-1} [(\mu_1 - \mu_2) - (\bar{y}_1 - \bar{y}_2)] \leq \left(\frac{1}{n_1} + \frac{1}{n_2}\right) T^2_{.05}(p, n_1 + n_2 - 2)\right\}.$$

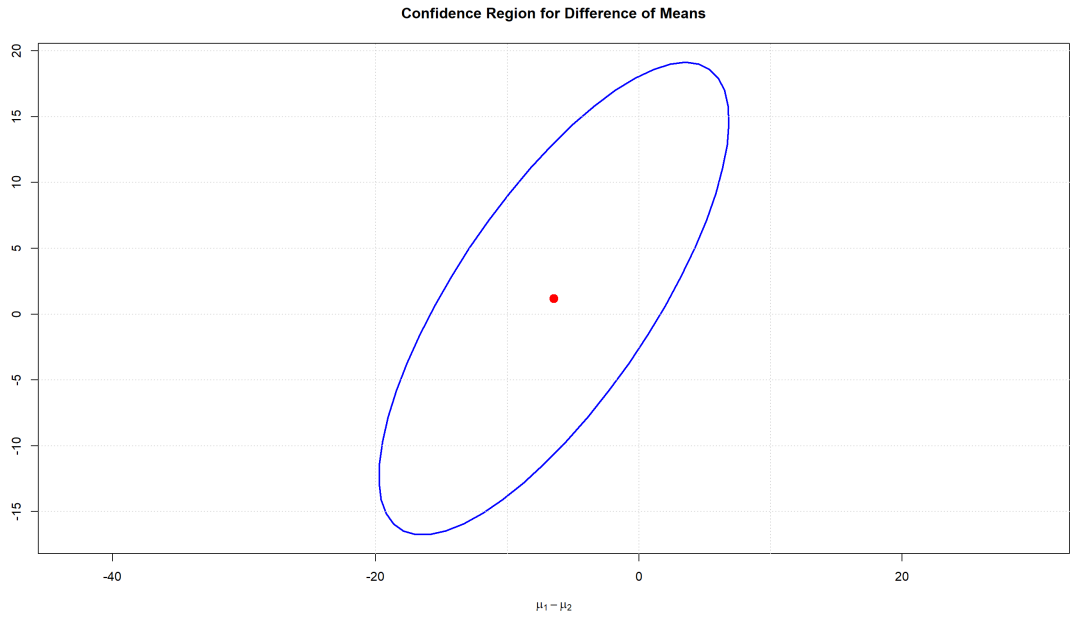
```
> n1 <- nrow(y11)
> n2 <- nrow(y2)
> p <- ncol(y11)
> cM <- colMeans(y11) - colMeans(y2)
> cM
       V1        V2
-6.463131  1.176768
>
> S <- ((n1 - 1) * var(y11) + (n2 - 1) * var(y2)) / (n1 + n2 - 2)
> S.inv <- solve(S)
> S.inv
           V1          V2
V1  0.02200004 -0.01225808
V2 -0.01225808  0.01206854
>
> RHS <- (1 / n1 + 1 / n2) * p * (n1 + n2 - 2) / (n1 + n2 - 1 - p) * qf(.95, p,
↪  n1+n2-1-p)
> RHS
[1] 0.2821595
```

Therefore, 95 confidence region for $\mu_1 - \mu_2$ is given by

$$\left\{\mu_1 - \mu_2 : \left[(\mu_1 - \mu_2) - \begin{pmatrix} -6.46 \\ 1.18 \end{pmatrix}\right]' \begin{pmatrix} 0.022 & -0.012 \\ -0.012 & 0.012 \end{pmatrix} \left[(\mu_1 - \mu_2) - \begin{pmatrix} -6.46 \\ 1.18 \end{pmatrix}\right] \leq 0.282\right\}.$$

```
> library(car)
> ellipse_data <- ellipse(center = cM, shape = S, radius = sqrt(qchisq(0.95, df
↪  = 2) * RHS), draw = FALSE)
```

```
> plot(ellipse_data, type = 'l', asp = 1, lwd = 2, col = 'blue', xlab =
↪  expression(mu[1] - mu[2]), ylab = '', xlim = range(ellipse_data[,1]), ylim =
↪  range(ellipse_data[,2]), main = 'Confidence Region for Difference of Means')
> grid()
> points(cM[1], cM[2], pch = 19, col = 'red', cex = 1.5)
```




(d) From the confidence region described in section (c), it is evident that tail length plays a significant role in distinguishing male birds from female birds. This leads to the observation that female birds are generally larger.

# 6 The admission of a business school

(a) To calculate the group means,

```
> rm(list = ls())
> y <- read.table(paste('C:/Users/Ray Chen/Desktop/MVA/gpa-gmat.DAT'), header =
↪  FALSE)
> # Group means
> y1 <- y[y$V3 == 1, 1:2]
> y1_mean <- colMeans(y1)
> y1_mean
      V1          V2
3.403871 561.225806
> y2 <- y[y$V3 ==2, 1:2]
```

```
> y2_mean <- colMeans(y2)
> y2_mean
      V1       V2
2.4825 447.0714
> y3 <- y[y$V3 == 3, 1:2]
> y3_mean <- colMeans(y3)
> y3_mean
      V1         V2
2.992692 446.230769
```

To calculate the overall meam,

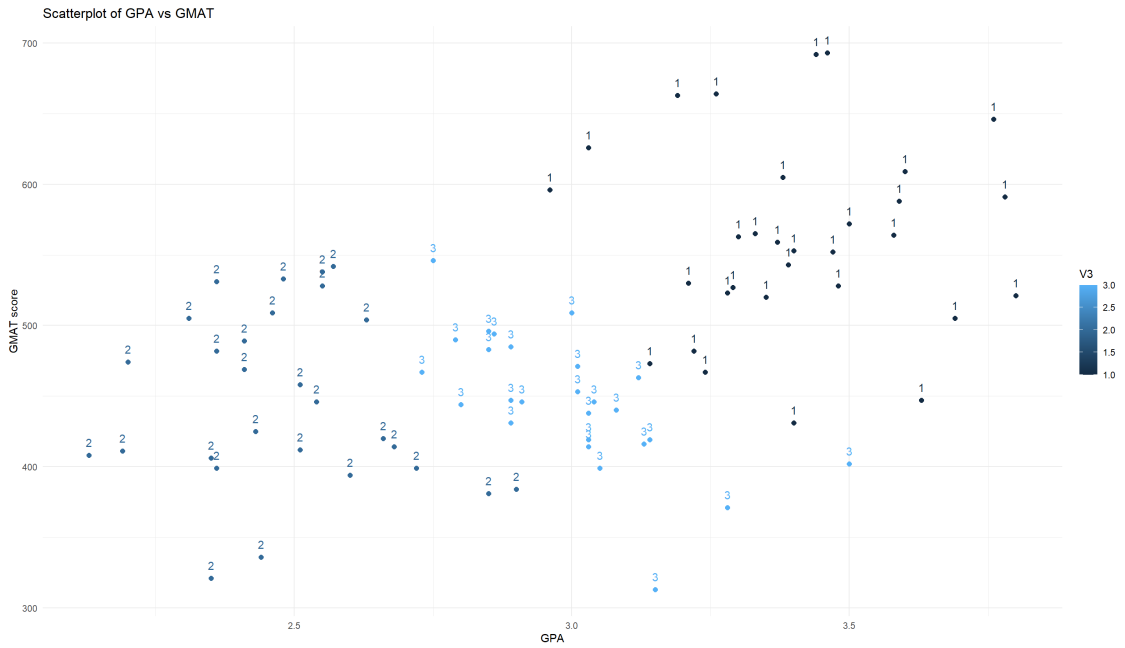
```
> # Overall mean
> y_mean <- (y1_mean + y2_mean + y3_mean) /3
> y_mean
      V1         V2
2.959688 484.842668
```

To calculate the sample pooled covariance matrix,

```
> # pooled sample covariance matrix
> n1 <- nrow(y1)
> n2 <- nrow(y2)
> n3 <- nrow(y3)
> Spl <- ((n1 - 1) * cov(y1) + (n2 - 1) * cov(y2) + (n3 - 1) * cov(y3)) /
↪  (nrow(y) - 3)
> Spl
           V1          V2
V1  0.03606795   -2.018759
V2 -2.01875915 3655.901121
```

(b)

```
library(ggplot2)
ggplot(y, aes(x = V1, y = V2, color = V3)) +
 geom_point(size = 2) +
  geom_text(aes(label = V3), vjust = -1) +
labs(title = "Scatterplot of GPA vs GMAT", x = "GPA", y = "GMAT score") +
theme_minimal()
```




The scatterplot above shows that group 1 had the highest performance in both GPA and GMAT scores, followed by group 3, and then group 2. The arrangement of these groups appears to form a slanted line in sequential order.

(c) To conduct the Fisher's LDA without using R package,

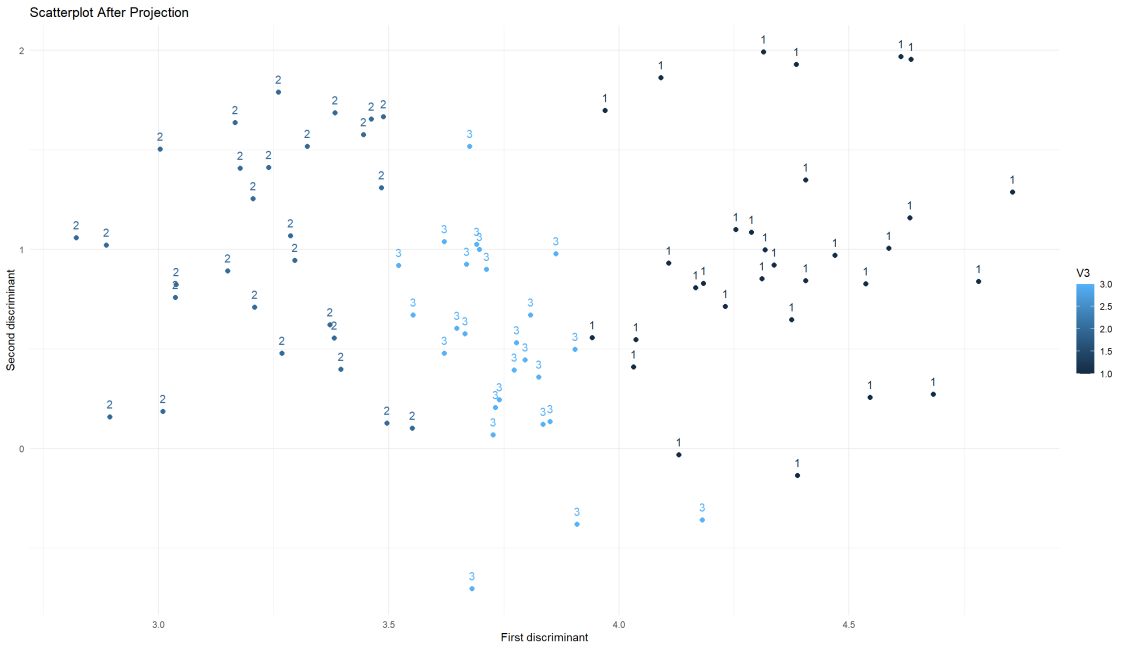
```
> B <- (y1_mean - y_mean) %*% t(y1_mean - y_mean) + (y2_mean - y_mean) %*%
↪  t(y2_mean - y_mean) + (y3_mean - y_mean) %*% t(y3_mean - y_mean)
> W <- ((n1 - 1) * cov(y1) + (n2 - 1) * cov(y2) + (n3 - 1) * cov(y3))
> WinvB <- solve(W) %*% B
> WinvB.eig <- eigen(WinvB)
> WinvB.eig
eigen() decomposition
$values
[1] 0.191251866 0.007064677

$vectors
             [,1]         [,2]
[1,] 0.999998564 -0.999969462
[2,] 0.001694796  0.007815072
```

Based on the above result, the Fisher's linear discriminant functions are

$$\begin{cases} Z_1 = 0.999998564Y_1 - 0.999969462Y_2 \\ Z_2 = 0.001694796Y_1 + 0.007815072Y_2 \end{cases}$$

```
z <- as.matrix(y[,1:2]) %*% WinvB.eig$vectors
zz <- data.frame(V1 = z[,1], V2 = z[,2], V3 = y$V3)
```

```
ggplot(zz, aes(x = V1, y = V2, color = V3)) +
 geom_point(size = 2) +
 geom_text(aes(label = V3), vjust = -1) +
 labs(title = "Scatterplot After Projection", x = "First discriminant", y =
 ↪  "Second discriminant") +
 theme_minimal()
```




(d) From the meaning of the question, it can be inferred that the prior probabilities of the three groups are given by $p_1 = 0.3$ , $p_2 = 0.6$ and $p_3 = 0.1$. Since we specified the prior probabilities, we need to use the Bayes rule for several populations. The best choice of allocation is given by

$$\arg\max_k f_k(y_0)p_k = \arg\max_k \left[\log p_k - \frac{1}{2}(y_0 - \bar{y}_k)' S_{pl}^{-1}(y_0 - \bar{y}_k)\right].$$

```
> y0 <- c(3.21, 497)
> ob1 <- log(0.3) - c(t(y0-y1_mean) %*% solve(Spl) %*% (y0-y1_mean)) / 2
> ob1
[1] -2.520486
> ob2 <- log(0.6) - c(t(y0-y2_mean) %*% solve(Spl) %*% (y0-y2_mean)) / 2
> ob2
[1] -9.007393
> ob3 <- log(0.1) - c(t(y0-y3_mean) %*% solve(Spl) %*% (y0-y3_mean)) / 2
> ob3
[1] -3.516147
```

By the computation of the observations above, we shall classify $y_0$ to group 1.

(e) To conduct (d) using "lda" function in R,

```
> library(MASS)
> predict(lda(V3 ~ V1 + V2, y , prior=c(0.3, 0.6, 0.1)), data.frame(V1 = 3.21,
↪  V2 = 497))
$class
[1] 1
Levels: 1 2 3

$posterior
          1          2         3
1 0.7293933 0.00111105 0.2694956

$x
        LD1       LD2
1 -2.150703 0.4723063
```

Clearly, $y0 = (3.21, 497)'$ is appropriately classified into group 1, aligning with the findings presented in section (d).

# 7 Beetle data problem

(a) To find the discriminant function,

```
> rm(list = ls())
> y <- read.table(paste('C:/Users/Ray Chen/Desktop/MVA/T5_5_FBEETLES.DAT'),
↪  header = FALSE)
> View(y)
> library(MASS)
> result <- lda(V2 ~ V3 + V4 + V5 + V6, y)
> result
Call:
lda(V2 ~ V3 + V4 + V5 + V6, data = y)

Prior probabilities of groups:
        1         2
0.4871795 0.5128205

Group means:
        V3       V4       V5       V6
1 194.4737 267.0526 137.3684 185.9474
```

```
2 179.5500 290.8000 157.2000 209.2500

Coefficients of linear discriminants:
           LD1
V3 -0.09327642
V4  0.03522706
V5  0.02875538
V6  0.03872998
    > c(as.matrix(y[,3:6]) %*% result$scaling)
 [1] 1.253859 3.450078 0.972349 2.251551 2.269024 2.247743 1.620702 1.919562
 ↪  3.855255 2.376169 1.452890 1.806247
[13] 2.034770 3.116013 2.950949 3.017335 2.065899 3.385739 3.924137 7.246776
↪  4.716840 7.125274 6.574576 6.942046
[25] 7.204548 7.426136 4.065517 5.771684 6.376290 7.773206 5.949728 5.423566
↪  7.293037 5.701034 6.382412 4.336489
[37] 6.296186 4.685068 5.126404
```

(b) To find the discriminant coefficient vector based on the individually standardized observations.

```
    > # the discriminant coefficient vector based on the individually standardized
    ↪  observations
    > y1 <- y[y$V2 == 1, 3:6]
    > y2 <- y[y$V2 == 2, 3:6]
    > s1 <- cov(y1)
    > s2 <- cov(y2)
    > n1 <- nrow(y1)
    > n2 <- nrow(y2)
    > W <- (n1 - 1) * s1 + (n2 - 1) * s2
    > Spl <- W / (nrow(y) - 2)
    > y1s <- t(apply(y1, 1, function(t){t / sqrt(diag(Spl))}))
    > y2s <- t(apply(y2, 1, function(t){t / sqrt(diag(Spl))}))
    > ys <- data.frame(F1 = c(y1s[,1], y2s[,1]), F2 = c(y1s[,2], y2s[,2]), F3 =
    ↪  c(y1s[,3], y2s[,3]), F4 = c(y1s[,4], y2s[,4]), G = factor(c(rep(1, n1),
    ↪  rep(2, n2))))
    > result2 <- lda(G ~ F1 + F2 + F3 + F4, ys)
    > result2$scaling
          LD1
F1 -1.1176022
F2  0.6755773
F3  0.3127788
F4  0.5586695
```

```
    > c(as.matrix(ys[,1:4]) %*% result2$scaling)
 [1] 1.253859 3.450078 0.972349 2.251551 2.269024 2.247743 1.620702 1.919562
 ↪  3.855255 2.376169 1.452890 1.806247
[13] 2.034770 3.116013 2.950949 3.017335 2.065899 3.385739 3.924137 7.246776
↪  4.716840 7.125274 6.574576 6.942046
[25] 7.204548 7.426136 4.065517 5.771684 6.376290 7.773206 5.949728 5.423566
↪  7.293037 5.701034 6.382412 4.336489
[37] 6.296186 4.685068 5.126404
```

(c) The coefficient vector for the discriminant function outlined in (b) results from adjusting the relative importance found in (a) according to each variable's individual variance. It's important to note that the transformed univariate observations in (a) remain consistent with those in (b). Hence, while individual standardization alters the interpretation of the discriminant function coefficient vector, it does not impact the overall discrimination outcome.

(d)

```
    > t.test(y1$V3, y2$V3, var.equal = TRUE)

^^ITwo Sample t-test

data:  y1$V3 and y2$V3
t = 3.8879, df = 37, p-value = 0.0004049
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
  7.146246 22.701122
sample estimates:
mean of x mean of y
 194.4737  179.5500

    > t.test(y1$V4, y2$V4, var.equal = TRUE)

^^ITwo Sample t-test

data:  y1$V4 and y2$V4
t = -3.8652, df = 37, p-value = 0.0004326
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -36.19595 -11.29879
sample estimates:
mean of x mean of y
 267.0526  290.8000
```

```
    > t.test(y1$V5, y2$V5, var.equal = TRUE)

^^ITwo Sample t-test

data:  y1$V5 and y2$V5
t = -5.6911, df = 37, p-value = 1.645e-06
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -26.89214 -12.77101
sample estimates:
mean of x mean of y
 137.3684  157.2000

    > t.test(y1$V6, y2$V6, var.equal = TRUE)

^^ITwo Sample t-test

data:  y1$V6 and y2$V6
t = -5.0426, df = 37, p-value = 1.236e-05
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -32.66593 -13.93933
sample estimates:
mean of x mean of y
 185.9474  209.2500
```

(e) According to the findings in section (a), the variables are ranked in order of their contribution as $Y1 > Y4 > Y2 > Y3$. The rankings based on section (b) are $Y1 > Y2 > Y4 > Y3$. From the results provided in section (d), the variables are ranked as $Y3 > Y4 > Y1 > Y2$ in terms of their contribution.