HW3-SOLUTION

1. Consider the two 2-dimensional data sets from two populations...

(a) Calculate the linear discriminant function.

$$\bar{y}_1 = (3,6)', \quad \bar{y}_2 = (5,8)', \quad S_{pl} = \frac{1}{n_1+n_2-2}\sum_k^2 \sum_{j=1}^{n_k}(y_{kj}-\bar{y}_k)(y_{kj}-\bar{y}_k)' = \begin{pmatrix} 1 & 1 \\ 1 & 2 \end{pmatrix}.$$

The discriminant function coefficient is

$$a = S_{pl}^{-1}(\bar{y}_1 - \bar{y}_2) = \begin{pmatrix} 2 & -1 \\ -1 & 1 \end{pmatrix}\begin{pmatrix} -2 \\ -2 \end{pmatrix} = \begin{pmatrix} -2 \\ 0 \end{pmatrix}.$$

The discriminant function is $a'y = -2y_1$ where $y = (y_1, y_2)'$.

(b) Classify the observation...

y_0 should be allocated to G_1 since

$$a'y_0 = -4 > -8 = a'\left(\frac{\bar{y}_1 + \bar{y}_2}{2}\right).$$

2. Show the following based on Fisher's LDA.

(a) The maximized statistical distance between the transformed sample mean...

Define $\mathbf{S}_{pl}$ as follows

$$\mathbf{S}_{pl} = \frac{1}{n_1+n_2-2}\left[(n_1-1)\,\mathbf{S}_1 + (n_2-1)\,\mathbf{S}_2\right]$$

To maximize the square distance between $\bar{z}_1$ and $\bar{z}_2$ is

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

By Cauchy-Schwarx inequality, the maximal value is

$$(\bar{\mathbf{y}}_1 - \bar{\mathbf{y}}_2)'\,\mathbf{S}_{pl}^{-1}(\bar{\mathbf{y}}_1 - \bar{\mathbf{y}}_2)$$

Then the maximal statistical distance betweem $\bar{z}_1$ and $\bar{z}_2$ is d_1

$$d_1 = \sqrt{(\bar{\mathbf{y}}_1 - \bar{\mathbf{y}}_2)'\,\mathbf{S}_{pl}^{-1}(\bar{\mathbf{y}}_1 - \bar{\mathbf{y}}_2)}$$

The statistical distance between $\bar{\mathbf{y}}_1$ and $\bar{\mathbf{y}}_2$ is d_2

$$\begin{aligned} d_2 &\propto \sqrt{(\bar{\mathbf{y}}_1 - \bar{\mathbf{y}}_2)'\,\mathbf{S}_{pl}^{-1}(\bar{\mathbf{y}}_1 - \bar{\mathbf{y}}_2)} \\ &= \sqrt{(\bar{\mathbf{y}}_1 - \bar{\mathbf{y}}_2)'\,\mathbf{S}_{pl}^{-1}(\bar{\mathbf{y}}_1 - \bar{\mathbf{y}}_2)} \end{aligned}$$

Thus, d_1 is proportional to d_2

(b) Fisher's allocation rule is indeed to compare the statistical distance between the new observation...

Fisher's allocation rule is allocating the new obersvation to G_1 if

$$(\overline{\mathbf{y}}_1-\overline{\mathbf{y}}_2)'\mathbf{S}_{pl}^{-1}\mathbf{y}_0 \geq \frac{1}{2}(\overline{\mathbf{y}}_1-\overline{\mathbf{y}}_2)'\mathbf{S}_{pl}^{-1}(\overline{\mathbf{y}}_1+\overline{\mathbf{y}}_2)$$

The LHS equals to

$$(\overline{\mathbf{y}}_1-\mathbf{y}_0)'\mathbf{S}_{pl}^{-1}\mathbf{y}_0+(\mathbf{y}_0-\overline{\mathbf{y}}_2)'\mathbf{S}_{pl}^{-1}\mathbf{y}_0$$

The RHS equals to

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

The the inequality becomes

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

The RHS equals to

$$\begin{aligned}
&\frac{1}{2}(\mathrm{y}_0-\overline{\mathrm{y}}_2+\mathrm{y}_0-\overline{\mathrm{y}}_1)'\mathrm{S}_{pl}^{-1}(\overline{\mathrm{y}}_1-\overline{\mathrm{y}}_2)\\
=&\frac{1}{2}\left[(\mathrm{y}_0-\overline{\mathrm{y}}_2)'\mathrm{S}_{pl}^{-1}(\overline{\mathrm{y}}_1-\mathrm{y}_0)+(\mathrm{y}_0-\overline{\mathrm{y}}_2)'\mathrm{S}_{pl}^{-1}(\mathrm{y}_0-\overline{\mathrm{y}}_2)+(\mathrm{y}_0-\overline{\mathrm{y}}_1)'\mathrm{S}_{pl}^{-1}(\overline{\mathrm{y}}_1-\mathrm{y}_0)+(\mathrm{y}_0-\overline{\mathrm{y}}_1)'\mathrm{S}_{pl}^{-1}(\mathrm{y}_0-\overline{\mathrm{y}}_2)\right.\\
=&\frac{1}{2}(\mathrm{y}_0-\overline{\mathrm{y}}_2)'\mathrm{S}_{pl}^{-1}(\mathrm{y}_0-\overline{\mathrm{y}}_2)-\frac{1}{2}(\overline{\mathrm{y}}_1-\mathrm{y}_0)'\mathrm{S}_{pl}^{-1}(\overline{\mathrm{y}}_1-\mathrm{y}_0)
\end{aligned}$$

Then the inequality becomes

$$(\mathbf{y}_0-\overline{\mathbf{y}}_2)'\mathbf{S}_{pl}^{-1}(\mathbf{y}_0-\overline{\mathbf{y}}_2)\geq(\overline{\mathbf{y}}_1-\mathbf{y}_0)'\mathbf{S}_{pl}^{-1}(\overline{\mathbf{y}}_1-\mathbf{y}_0)$$

3. Profile analysis pertains to situations in which a battery of...

(a) Test whether the profiles of two populations are parallel.

To test whether the profiles of two populations are parallel is equivalent to test

$$H_0:\exists K_0 \text{ s.t. } \mu_1=\mu_2+K_0 \leftrightarrow H_1:\forall K:\mu_1\neq\mu_2+K,$$

where the null hypothesis can be rewritten as

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

Exploiting the mapping $y\mapsto Cy$, the test statistic is given by

$$\frac{n_1n_2}{n_1+n_2}(\bar{y}_1-\bar{y}_2)'C'(CS_{pl}C')^{-1}C(\bar{y}_1-\bar{y}_2)\sim T^2(p-1,n_1+n_2-2),$$

and the corresponding rejection region is

$$\left\{(y_1,y_2):\frac{n_1n_2}{n_1+n_2}(\bar{y}_1-\bar{y}_2)'C'(CS_{pl}C')^{-1}C(\bar{y}_1-\bar{y}_2)>T_\alpha^2(p-1,n_1+n_2-2)\right\}.$$

(b) Test whether the total measurements are the same between the two population.

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

$$H_0 : 1_p'\mu_1 = 1_p'\mu_2 \leftrightarrow H_1 : 1_p'\mu_1 \neq 1_p'\mu_2.$$

Exploiting the mapping $y \mapsto 1_p'y$, the test statistic is given by

$$\frac{n_1 n_2}{n_1 + n_2}(\bar{y}_1 - \bar{y}_2)'1_p(1_p'S_{pl}1_p)^{-1}1_p'(\bar{y}_1 - \bar{y}_2) \sim T^2(1, n_1 + n_2 - 2),$$

or equivalently,

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

and the corresponding rejection region is

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

(c) Assuming the profiles are parallel, test whether the profiles are linear. (Hint: First show that the linearity test can be written as...

Assuming parallelism, to test whether the profiles are linear are parallel is equivalent to test

$$H_0 : \exists k \forall j : \frac{1}{2}(\mu_{1j} + \mu_{2j}) = (j-1)k + \frac{1}{2}(\mu_{11} + \mu_{21}) \leftrightarrow H_1 : \forall k \exists j : \frac{1}{2}(\mu_{1j} + \mu_{2j}) \neq (j-1)k + \frac{1}{2}(\mu_{11} + \mu_{21}),$$

where the null hypothesis can be rewritten as

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

Exploiting the mapping $y_1 \mapsto Cy_1$ and $y_2 \mapsto -Cy_2$, respectively, the test statistic is given by

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

and the corresponding rejection region is

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

(d) Following (c), Let...

$$C = \begin{pmatrix} 1 & -2 & 1 & \\ & 1 & -2 & 1 \end{pmatrix} \implies \frac{n_1 n_2}{n_1 + n_2}(\bar{y}_1 + \bar{y}_2)'C'(CS_{pl}C')^{-1}C(\bar{y}_1 + \bar{y}_2)$$

$$= 15(-0.1, -1.4)\begin{pmatrix} 3.67 & 2.02 \\ 2.02 & 2.4 \end{pmatrix}\begin{pmatrix} -0.1 \\ -1.4 \end{pmatrix}\begin{vmatrix} 2.4 & -2.02 \\ -2.02 & 3.67 \end{vmatrix}^{-1} = \frac{79.5945}{4.7276}$$

$$= 16.83613 > 6.428522 = \frac{(58)(2)}{57}F_{0.05}(2, 57) = T^2_{0.05}(4-2, 30+30-2),$$

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

4. (R exercise) In the first phase of a study of the cost of transporting milk from farms to dairy plants, a survey was taken of firms engaged in milk transportation. Cost data on . . .

```r
rm(list = ls())
y <- read.table("./cost.dat")
y
```

```
##       V1    V2    V3       V4
## 1  16.44 12.43 11.23 gasoline
## 2   7.19  2.70  3.92 gasoline
## 3   9.92  1.35  9.75 gasoline
## 4   4.24  5.78  7.78 gasoline
## 5  11.20  5.05 10.67 gasoline
## 6  14.25  5.78  9.88 gasoline
## 7  13.50 10.98 10.60 gasoline
## 8  13.32 14.27  9.45 gasoline
## 9  29.11 15.09  3.28 gasoline
## 10 12.68  7.61 10.23 gasoline
## 11  7.51  5.80  8.13 gasoline
## 12  9.90  3.63  9.13 gasoline
## 13 10.25  5.07 10.17 gasoline
## 14 11.11  6.15  7.61 gasoline
## 15 12.17 14.26 14.39 gasoline
## 16 10.24  2.59  6.09 gasoline
## 17 10.18  6.05 12.14 gasoline
## 18  8.88  2.70 12.23 gasoline
## 19 12.34  7.73 11.68 gasoline
## 20  8.51 14.02 12.01 gasoline
## 21 26.16 17.44 16.89 gasoline
## 22 12.95  8.24  7.18 gasoline
## 23 16.93 13.37 17.59 gasoline
## 24 14.70 10.78 14.58 gasoline
## 25 10.32  5.16 17.00 gasoline
## 26  8.98  4.49  4.26 gasoline
## 27  9.70 11.59  6.83 gasoline
## 28 12.72  8.63  5.59 gasoline
## 29  9.49  2.16  6.23 gasoline
## 30  8.22  7.95  6.72 gasoline
## 31 13.70 11.22  4.91 gasoline
## 32  8.21  9.85  8.17 gasoline
## 33 15.86 11.42 13.06 gasoline
## 34  9.18  9.18  9.49 gasoline
## 35 12.49  4.67 11.94 gasoline
## 36 17.32  6.86  4.44 gasoline
## 37  8.50 12.26  9.11   diesel
## 38  7.42  5.13 17.15   diesel
## 39 10.28  3.32 11.23   diesel
## 40 10.16 14.72  5.99   diesel
## 41 12.79  4.17 29.28   diesel
## 42  9.60 12.72 11.00   diesel
## 43  6.47  8.89 19.00   diesel
## 44 11.35  9.95 14.53   diesel
## 45  9.15  2.94 13.68   diesel
## 46  9.70  5.06 20.84   diesel
## 47  9.77 17.86 35.18   diesel
## 48 11.61 11.75 17.00   diesel
## 49  9.09 13.25 20.66   diesel
## 50  8.53 10.14 17.45   diesel
## 51  8.29  6.22 16.38   diesel
## 52 15.90 12.90 19.09   diesel
## 53 11.94  5.69 14.77   diesel
## 54  9.54 16.77 22.66   diesel
## 55 10.43 17.65 10.66   diesel
## 56 10.87 21.52 28.47   diesel
## 57  7.13 13.22 19.44   diesel
## 58 11.88 12.18 21.20   diesel
## 59 12.03  9.22 23.09   diesel
```

(a) Test for differences in the mean cost vectors at the significance level 0.01.

```
y1 <- y[y$V4 == "gasoline",1:3]
y2 <- y[y$V4 == "diesel",1:3]
library(ICSNP)
```

```
## Loading required package: mvtnorm
```

```
## Loading required package: ICS
```

```
HotellingsT2(y1, y2) # Here, the statistic T.2 has an F distribution.
```

```
##
##  Hotelling's two sample T2-test
##
## data:  y1 and y2
## T.2 = 16.375, df1 = 3, df2 = 55, p-value = 1e-07
## alternative hypothesis: true location difference is not equal to c(0,0,0)
```

Thus, the null hypothesis that $\mu_1 = \mu_2$ can be rejected with level 0.01.

(b) If the hypothesis of equal cost vectors is rejected, conduct the univariate tests at the same significance level. What is your conclusion?

```
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 results above, $\mu_{11} = \mu_{21}$ and $\mu_{12} = \mu_{22}$ cannot be rejected with level 0.01 while $\mu_{13} = \mu_{23}$ can be rejected with level 0.01, which implies that the capital might contribute the most to make gasoline trucks differ from diesel.

(c) If the hypothesis of equal cost vectors is rejected, find the linear combination of mean components most responsible for the rejection. Interpret the coefficients in the linear combination.

```
library(MASS)
lda(V4 ~ V1 + V2 + V3, y)
```

```
## 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
```

According to the results above, the desired linear combination is $0.134Y_1 - 0.070Y_2 - 0.167Y_3$, from which we can see that the capital contributed the most to make gasoline trucks differ from diesel and it coincides with results in (b).

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

```r
y11 <- y1[1:23,]
```

(d-a) Test for differences in the mean cost vectors at the significance level 0.01.

```r
HotellingsT2(y11 - y2) # Here, the statistic T.2 has an F distribution.
```

```
##
##  Hotelling's one sample T2-test
##
## data:  y11 - y2
## T.2 = 15.899, df1 = 3, df2 = 20, p-value = 1.603e-05
## alternative hypothesis: true location is not equal to c(0,0,0)
```

Thus, the null hypothesis that $\delta = 0$ can be rejected with level 0.01.

(d-b) If the hypothesis of equal cost vectors is rejected, conduct the univariate tests at the same significance level. What is your conclusion?

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

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

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

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

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

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

From the results above, $\mu_{d1} = 0$ and $\mu_{d2} = 0$ cannot be rejected with level 0.01 while $\mu_{d3} = 0$ can be rejected with level 0.01, which implies that the capital might contribute the most to make gasoline trucks differ from diesel.

(d-c) If the hypothesis of equal cost vectors is rejected, find the linear combination of mean components most responsible for the rejection. Interpret the coefficients in the linear combination.

```
library(MASS)
lda(V4 ~ V1 + V2 + V3, y[c(1:23,37:59),])
```

```
## 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
```

According to the results above, the desired linear combination is $0.140Y_1 - 0.069Y_2 - 0.153Y_3$, from which we can see that the capital contributed the most to make gasoline trucks differ from diesel and it coincides with results in (d-b).

5. (R exercise) The tail lengths in millimeters...

```
rm(list = ls())
y1 <- read.table("./male")
y1
```

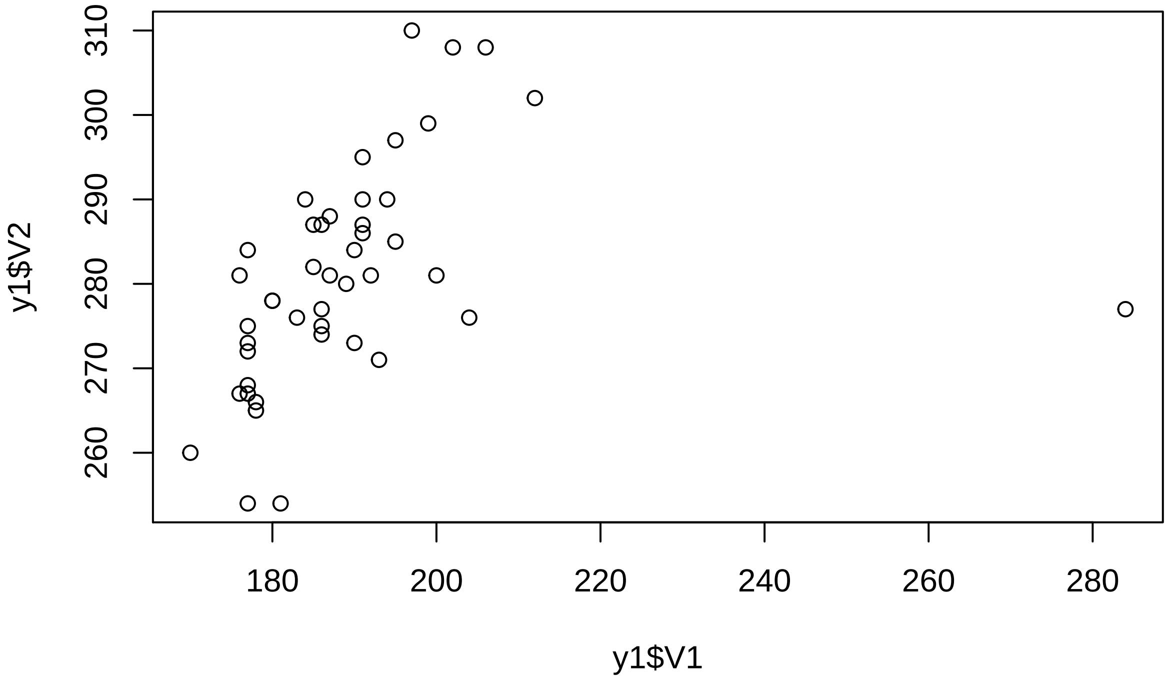
```
##     V1  V2
## 1  180 278
## 2  186 277
## 3  206 308
## 4  184 290
## 5  177 273
## 6  177 284
## 7  176 267
## 8  200 281
## 9  191 287
## 10 193 271
## 11 212 302
## 12 181 254
## 13 195 297
## 14 187 281
## 15 190 284
## 16 185 282
## 17 195 285
## 18 183 276
## 19 202 308
## 20 177 254
## 21 177 268
## 22 170 260
## 23 186 274
## 24 177 272
## 25 178 266
## 26 192 281
## 27 204 276
## 28 191 290
## 29 178 265
## 30 177 275
## 31 284 277
## 32 176 281
## 33 185 287
## 34 191 295
## 35 177 267
## 36 197 310
## 37 199 299
## 38 190 273
## 39 180 278
## 40 189 280
## 41 194 290
## 42 186 287
## 43 191 286
## 44 187 288
## 45 186 275
```

```
y2 <- read.table("./female")
y2
```

```
##     V1  V2
## 1  191 284
## 2  197 285
## 3  208 288
## 4  180 273
## 5  180 275
## 6  188 280
## 7  210 283
## 8  196 288
## 9  191 271
## 10 179 257
## 11 208 289
## 12 202 285
## 13 200 272
## 14 192 282
## 15 199 280
## 16 186 266
## 17 197 285
## 18 201 295
## 19 190 282
## 20 209 305
## 21 187 285
## 22 207 297
## 23 178 268
## 24 202 271
## 25 205 285
## 26 190 280
## 27 189 277
## 28 211 310
## 29 216 305
## 30 189 274
## 31 173 271
## 32 194 280
## 33 198 300
## 34 180 272
## 35 190 292
## 36 191 286
## 37 196 285
## 38 207 286
## 39 209 303
## 40 179 261
## 41 186 262
## 42 174 245
## 43 181 250
## 44 189 262
## 45 188 258
```

(a) Plot the male hook-billed kite data as a scatterplot, and visually check for outliers.

```
plot(y1$V1, y1$V2)
```



It can be seen from the scatterplot that one point (which is corresponding to the 31th sample) is apparently an outlier.

(b) Test for equality of mean vectors for the populations of male and female hookbilled kites. Set…

(b-a) You may want to eliminate any outlier found in Part (a) for the male hook-billed kite data before conducting this test.

```
y1a <- y1[c(1:30, 32:45),]
library(ICSNP)
HotellingsT2(y1a, y2) # Here, the statistic T.2 has an F distribution.
```

```
##
##  Hotelling's two sample T2-test
##
## data:  y1a and y2
## T.2 = 12.339, df1 = 2, df2 = 86, p-value = 1.944e-05
## alternative hypothesis: true location difference is not equal to c(0,0)
```

Thus, the null hypothesis that $\mu_1 = \mu_2$ can be rejected with level 0.05.

```
library(MASS)
ya <- data.frame(V1 = c(y1a[,1], y2[,1]), V2 = c(y1a[,2], y2[,2]),
                 group = factor(c(rep("male", nrow(y1a)), rep("female", nrow(y2)))))
lda(group ~ V1 + V2, ya)
```

```
## Call:
## lda(group ~ V1 + V2, data = ya)
##
## 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
```

According to the results above, the desired linear combination is $-0.148X_1 + 0.088X_2$.

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

```r
y1b <- y1
y1b[31,1] <- y1b[31,1] - 100 # Imputation
HotellingsT2(y1b, y2) # Here, the statistic T.2 has an F distribution.
```

```
##
##  Hotelling's two sample T2-test
##
## data:  y1b and y2
## T.2 = 12.685, df1 = 2, df2 = 87, p-value = 1.464e-05
## alternative hypothesis: true location difference is not equal to c(0,0)
```

Thus, the null hypothesis that $\mu_1 = \mu_2$ can be rejected with level 0.05, again.

```r
yb <- data.frame(V1 = c(y1b[,1], y2[,1]), V2 = c(y1b[,2], y2[,2]),
                 group = factor(c(rep("male", nrow(y1b)), rep("female", nrow(y2)))))
lda(group ~ V1 + V2, yb)
```

```
## Call:
## lda(group ~ V1 + V2, data = yb)
##
## Prior probabilities of groups:
## female   male
##    0.5    0.5
##
## Group means:
##              V1       V2
## female 193.6222 279.7778
## male   187.0889 280.8667
##
## Coefficients of linear discriminants:
##            LD1
## V1 -0.14874629
## V2  0.08830973
```

According to the results above, the desired linear combination is $-0.149X_1 + 0.088X_2$.

(b-c) Does it make any difference in this case how outliers for the male hook-billed kite data are treated?

From (b-a) and (b-b), we can see that it made little difference in this case how outliers for the male hook-billed kite data are treated.

(c) Determine and draw the 95% confidence region for. . .

From now on, I will eliminate the outlier found in Part (a). Write the confidence region as

$$\left\{\mu_1-\mu_2 : [(\mu_1-\mu_2)-(\bar{y}_1-\bar{y}_2)]' S_{pl}^{-1} [(\mu_1-\mu_2)-(\bar{y}_1-\bar{y}_2)] \le \left(\frac{1}{n_1}+\frac{1}{n_2}\right) T_\alpha^2(p, n_1+n_2-2)\right\}.$$

```
n1 <- nrow(y1a)
n2 <- nrow(y2)
p <- ncol(y1a)
cm <- colMeans(y1a) - colMeans(y2)
cm
```

```
##        V1        V2
## -6.463131  1.176768
```

```
S <- ((n1 - 1) * var(y1a) + (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
```

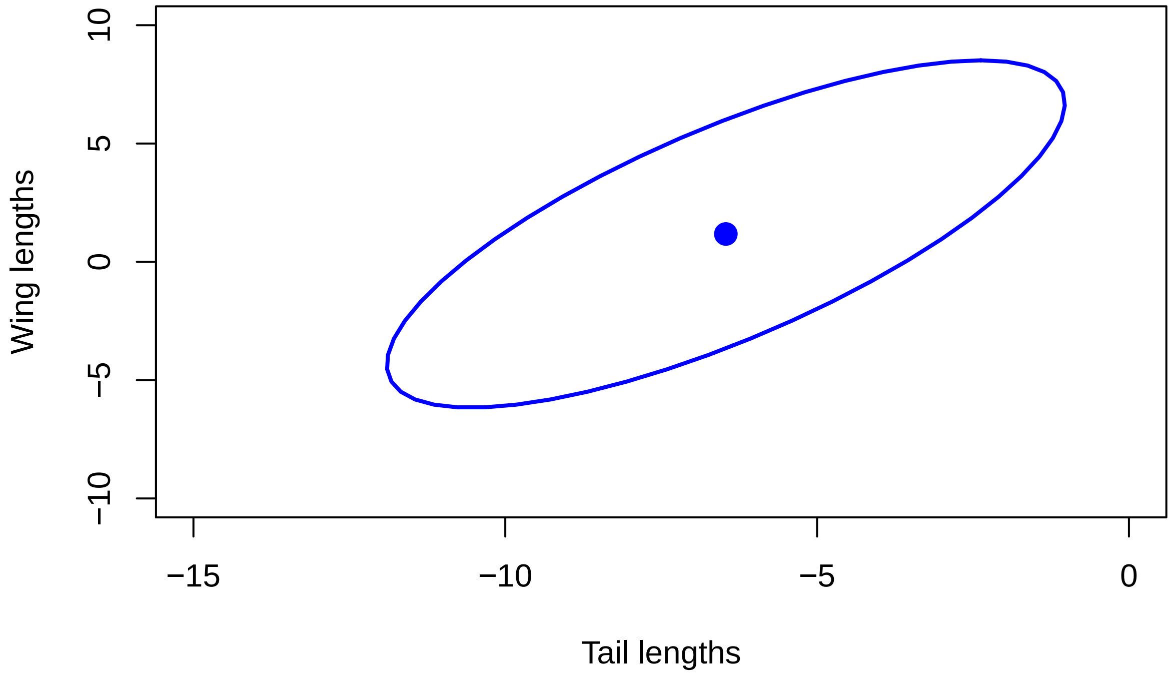
Using the computation results above, an 95% region estimate for $\mu_1 - \mu_2$ is given by

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

```
library(car)
```

```
## Loading required package: carData
```

```
plot(0, 0, type = "n", xlim = c(-15, 0), ylim = c(-10, 10),
     xlab = "Tail lengths", ylab = "Wing lengths")
ellipse(center = cm, shape = S, radius = sqrt(RHS))
```



(d) Are male or female birds generally larger?

From the discriminant functions given in (b) and the confidence region given in (c), we can see that tail length contributed the most to make male birds differ from female, which results in a fact that female birds are generally larger.

6. (R exercise.) The admission officer of a business school has used an index of. . .

```r
rm(list = ls())
Sys.setlocale('LC_ALL','C')
```

```
## [1] "C/C/C/C/C/zh_CN.UTF-8"
```

```r
y <- read.table("./gpa-gmat.DAT")
y
```

```
##      V1  V2 V3
## 1  2.96 596  1
## 2  3.14 473  1
## 3  3.22 482  1
## 4  3.29 527  1
## 5  3.69 505  1
## 6  3.46 693  1
## 7  3.03 626  1
## 8  3.19 663  1
## 9  3.63 447  1
## 10 3.59 588  1
## 11 3.30 563  1
## 12 3.40 553  1
## 13 3.50 572  1
## 14 3.78 591  1
## 15 3.44 692  1
## 16 3.48 528  1
## 17 3.47 552  1
## 18 3.35 520  1
## 19 3.39 543  1
## 20 3.28 523  1
## 21 3.21 530  1
## 22 3.58 564  1
## 23 3.33 565  1
## 24 3.40 431  1
## 25 3.38 605  1
## 26 3.26 664  1
## 27 3.60 609  1
## 28 3.37 559  1
## 29 3.80 521  1
## 30 3.76 646  1
## 31 3.24 467  1
## 32 2.54 446  2
## 33 2.43 425  2
## 34 2.20 474  2
## 35 2.36 531  2
## 36 2.57 542  2
## 37 2.35 406  2
## 38 2.51 412  2
## 39 2.51 458  2
## 40 2.36 399  2
## 41 2.36 482  2
## 42 2.66 420  2
## 43 2.68 414  2
## 44 2.48 533  2
## 45 2.46 509  2
## 46 2.63 504  2
## 47 2.44 336  2
## 48 2.13 408  2
## 49 2.41 469  2
## 50 2.55 538  2
## 51 2.31 505  2
## 52 2.41 489  2
## 53 2.19 411  2
## 54 2.35 321  2
## 55 2.60 394  2
## 56 2.55 528  2
## 57 2.72 399  2
## 58 2.85 381  2
## 59 2.90 384  2
## 60 2.86 494  3
## 61 2.85 496  3
## 62 3.14 419  3
## 63 3.28 371  3
## 64 2.89 447  3
## 65 3.15 313  3
## 66 3.50 402  3
## 67 2.89 485  3
## 68 2.80 444  3
## 69 3.13 416  3
## 70 3.01 471  3
## 71 2.79 490  3
## 72 2.89 431  3
## 73 2.91 446  3
## 74 2.75 546  3
## 75 2.73 467  3
## 76 3.12 463  3
## 77 3.08 440  3
## 78 3.03 419  3
## 79 3.00 509  3
## 80 3.03 438  3
## 81 3.05 399  3
## 82 2.85 483  3
## 83 3.01 453  3
## 84 3.03 414  3
## 85 3.04 446  3
```

(a) Calculate the group means, overall means, and pooled sample covariance matrix.

```
# group means
y1 <- y[y$V3 == 1, 1:2]
y1bar <- colMeans(y1)
y1bar
```

```
##         V1         V2
##   3.403871 561.225806
```

```
y2 <- y[y$V3 == 2, 1:2]
y2bar <- colMeans(y2)
y2bar
```

```
##       V1       V2
##   2.4825 447.0714
```

```
y3 <- y[y$V3 == 3, 1:2]
y3bar <- colMeans(y3)
y3bar
```

```
##         V1         V2
##   2.992692 446.230769
```

```
# overall mean
ybar <- (y1bar + y2bar + y3bar) / 3
ybar
```

```
##         V1         V2
##   2.959688 484.842668
```

```
# group sample covariance matrices
s1 <- var(y1)
s1
```

```
##            V1           V2
## V1 0.04355785 5.809677e-02
## V2 0.05809677 4.618247e+03
```

```
s2 <- var(y2)
s2
```

```
##              V1          V2
## V1  0.03364907   -1.192037
## V2 -1.19203704 3891.253968
```

```
s3 <- var(y3)
s3
```

```
##              V1          V2
## V1  0.02969246   -5.403846
## V2 -5.40384615 2246.904615
```

```
# pooled sample covariance matrix
n1 <- nrow(y1)
n2 <- nrow(y2)
n3 <- nrow(y3)
W <- (n1 - 1) * s1 + (n2 - 1) * s2 + (n3 - 1) * s3
Spl <- W / (nrow(y) - 3)
Spl
```

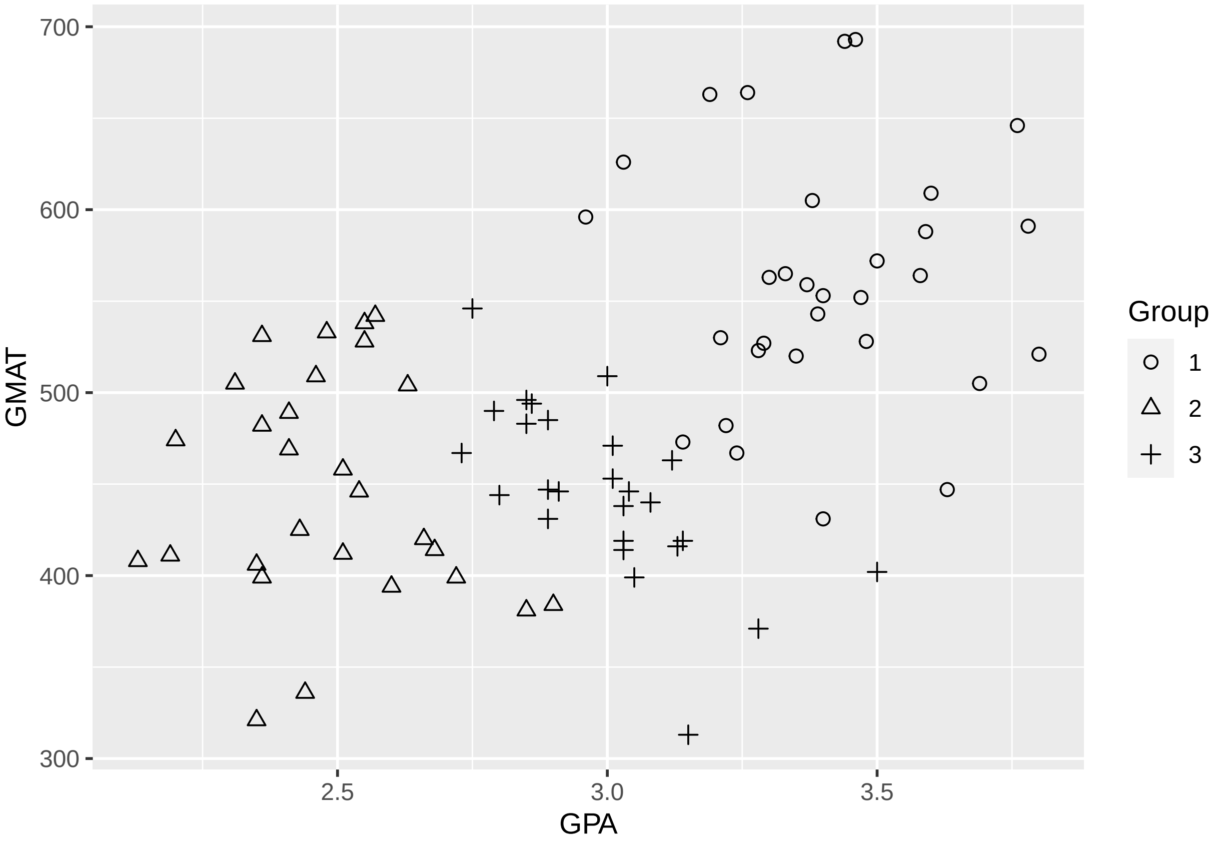
```
##              V1          V2
## V1  0.03606795   -2.018759
## V2 -2.01875915 3655.901121
```

(b) Obtain the scatterplot between GPA and GMAT, and label the three groups. Comment.

```
library(ggplot2)
```

```
## Warning: package 'ggplot2' was built under R version 3.6.2
```

```
ggplot(y, aes(x = V1, y = V2, shape = factor(V3))) + geom_point(size = 2) +
  labs(x = "GPA", y = "GMAT", shape = "Group") +
  scale_shape_manual(values = c(1, 2, 3))
```



As illustrated in the graph above, group 1 performed best in both GPA and GMAT, group 3 second, and then group 2. These three groups seem to be laid on an oblique line in order.

(c) Assuming equal covariance matrices, conduct Fisher's LDA. DO NOT use R package. Start with...

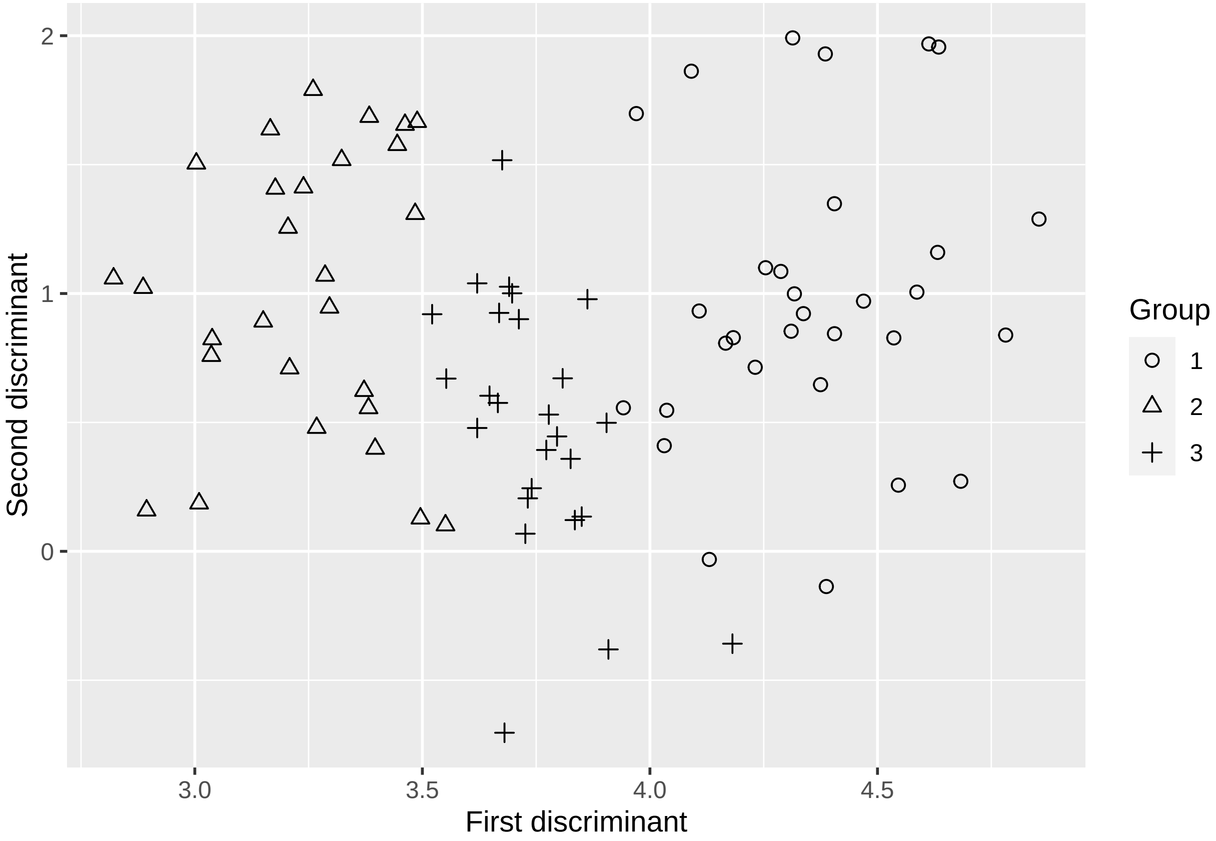
```
B <- (y1bar - ybar) %*% t(y1bar - ybar) + (y2bar - ybar) %*% t(y2bar - ybar) +
  (y3bar - ybar) %*% t(y3bar - ybar)
WinvB <- solve(W) %*% B
WinvB.eig <- eigen(WinvB)
WinvB.eig
```

```
## eigen() decomposition
## $values
## [1] 0.191251866 0.007064677
##
## $vectors
##             [,1]         [,2]
## [1,] 0.999998564 -0.999969462
## [2,] 0.001694796  0.007815072
```

Based on the computation results above, the Fisher's linear discriminants are

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

```
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, shape = factor(V3))) + geom_point(size = 2) +
  labs(x = "First discriminant", y = "Second discriminant", shape = "Group") +
  scale_shape_manual(values = c(1, 2, 3))
```



Apparently, the scatterplot above is somewhat like (not necessarily) a rotation of that in (b). Now the groups seem to be laid on a straight line parallel to x axis in order.

(d) Further assuming bivariate normality, and assume the Admit, DO not admit, and Borderline groups are predetermined to be proportional to 3:6:1. Classify the new observation (3.21,497)0 into one of the three groups. DO NOT use R package.

Assuming equal covariance matrices and sample information, the best choice of allocation is given by

$$\arg\max_k \quad f_k(y_0)p_k = \arg\max_k \quad \log(p_k) - \frac{1}{2}\log|\Sigma| - \frac{1}{2}(y_0 - \mu_k)'\Sigma^{-1}(y_0 - \mu_k)$$

$$= \arg\max_k \quad \log(p_k) - \frac{1}{2}(y_0 - \mu_k)'\Sigma^{-1}(y_0 - \mu_k) \doteq \arg\max_k \quad \log(p_k) - \frac{1}{2}(y_0 - \bar{y}_k)'S_{pl}^{-1}(y_0 - \bar{y}_k).$$

```
y0 <- c(3.21, 497)
# group 1
lpf1 <- log(.3) - c(t(y0-y1bar) %*% solve(Spl) %*% (y0-y1bar)) / 2
lpf1
```

```
## [1] -2.520486
```

```r
# group 2
lpf2 <- log(.6) - c(t(y0-y2bar) %*% solve(Spl) %*% (y0-y2bar)) / 2
lpf2
```

```
## [1] -9.007393
```

```r
# group 3
lpf3 <- log(.1) - c(t(y0-y3bar) %*% solve(Spl) %*% (y0-y3bar)) / 2
lpf3
```

```
## [1] -3.516147
```

Based on the computation results above, $y_0 = (3.21, 497)'$ should be classified to group 1.

(e) Conduct (d) using lda function in R. Compare the results.

```r
library(MASS)
predict(lda(V3 ~ V1 + V2, y , prior=c(.3, .6, .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
```

Apparently, $y_0 = (3.21, 497)'$ should be classified to group 1, which coincides with the result in (d). Indeed, using the computation results in (d), the same posterior probabilities can be given by the following codes.

```r
# group 1
exp(lpf1) / (exp(lpf1) + exp(lpf2) + exp(lpf3))
```

```
## [1] 0.7293933
```

```r
# group 2
exp(lpf2) / (exp(lpf1) + exp(lpf2) + exp(lpf3))
```

```
## [1] 0.00111105
```

```r
# group 3
exp(lpf3) / (exp(lpf1) + exp(lpf2) + exp(lpf3))
```

```
## [1] 0.2694956
```

7. (R exercise.) Use the beetle data in the following (data attached as T5_5_FBEETLES.DAT):

```r
rm(list = ls())
y <- read.table(paste('./T5_5_FBEETLES.DAT'))
y
```

```
##    V1 V2  V3  V4  V5  V6
## 1   1  1 189 245 137 163
## 2   2  1 192 260 132 217
## 3   3  1 217 276 141 192
## 4   4  1 221 299 142 213
## 5   5  1 171 239 128 158
## 6   6  1 192 262 147 173
## 7   7  1 213 278 136 201
## 8   8  1 192 255 128 185
## 9   9  1 170 244 128 192
## 10 10  1 201 276 146 186
## 11 11  1 195 242 128 192
## 12 12  1 205 263 147 192
## 13 13  1 180 252 121 167
## 14 14  1 192 283 138 183
## 15 15  1 200 294 138 188
## 16 16  1 192 277 150 177
## 17 17  1 200 287 136 173
## 18 18  1 181 255 146 183
## 19 19  1 192 287 141 198
## 20  1  2 181 305 184 209
## 21  2  2 158 237 133 188
## 22  3  2 184 300 166 231
## 23  4  2 171 273 162 213
## 24  5  2 181 297 163 224
## 25  6  2 181 308 160 223
## 26  7  2 177 301 166 221
## 27  8  2 198 308 141 197
## 28  9  2 180 286 146 214
## 29 10  2 177 299 171 192
## 30 11  2 176 317 166 213
## 31 12  2 192 312 166 209
## 32 13  2 176 285 141 200
## 33 14  2 169 287 162 214
## 34 15  2 164 265 147 192
## 35 16  2 181 308 157 204
## 36 17  2 192 276 154 209
## 37 18  2 181 278 149 235
## 38 19  2 175 271 140 192
## 39 20  2 197 303 170 205
```

(a) Find the discriminant function coefficient vector. Obtain the transformed univariate observations.

```
# the discriminant function coefficient vector
library(MASS)
ld <- lda(V2 ~ V3 + V4 + V5 + V6, y)
ld$scaling
```

```
##             LD1
## V3 -0.09327642
## V4  0.03522706
## V5  0.02875538
## V6  0.03872998
```

```
# the transformed univariate observations
c(as.matrix(y[,3:6]) %*% ld$scaling)
```

```
##  [1] 1.253859 3.450078 0.972349 2.251551 2.269024 2.247743 1.620702
##  [8] 1.919562 3.855255 2.376169 1.452890 1.806247 2.034770 3.116013
## [15] 2.950949 3.017335 2.065899 3.385739 3.924137 7.246776 4.716840
## [22] 7.125274 6.574576 6.942046 7.204548 7.426136 4.065517 5.771684
## [29] 6.376290 7.773206 5.949728 5.423566 7.293037 5.701034 6.382412
## [36] 4.336489 6.296186 4.685068 5.126404
```

(b) Find the discriminant coefficient vector based on the individually standardized observations. Obtain the transformed univariate 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 <- var(y1)
s2 <- var(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))))
ld2 <- lda(G ~ F1 + F2 + F3 + F4, ys)
ld2$scaling
```

```
##           LD1
## F1 -1.1176022
## F2  0.6755773
## F3  0.3127788
## F4  0.5586695
```

```
# the transformed univariate observations
c(as.matrix(ys[,1:4]) %*% ld2$scaling)
```

```
##  [1] 1.253859 3.450078 0.972349 2.251551 2.269024 2.247743 1.620702
##  [8] 1.919562 3.855255 2.376169 1.452890 1.806247 2.034770 3.116013
## [15] 2.950949 3.017335 2.065899 3.385739 3.924137 7.246776 4.716840
## [22] 7.125274 6.574576 6.942046 7.204548 7.426136 4.065517 5.771684
## [29] 6.376290 7.773206 5.949728 5.423566 7.293037 5.701034 6.382412
## [36] 4.336489 6.296186 4.685068 5.126404
```

(c) Compare the results from (a) and (b). Comment.

The discriminant function coefficient vector in (b) is obtained by redistributing the relative importance of that in (a) according to the individual variance. The following codes show the procedure of redistributing. Note that the transformed univariate observations in (a) are the same as those in (b). Therefore, individual standardization does change the interpretation of discriminant function coefficient vector, but does not affect the discrimination result.

```
ld$scaling
```

```
##             LD1
## V3 -0.09327642
## V4  0.03522706
## V5  0.02875538
## V6  0.03872998
```

```
sqrt(diag(Spl))
```

```
##       V3       V4       V5       V6
## 11.98162 19.17779 10.87723 14.42473
```

```
ld$scaling * sqrt(diag(Spl))
```

```
##            LD1
## V3 -1.1176022
## V4  0.6755773
## V5  0.3127788
## V6  0.5586695
```

```
ld2$scaling
```

```
##            LD1
## F1 -1.1176022
## F2  0.6755773
## F3  0.3127788
## F4  0.5586695
```

(d) Calculate…

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

```
##
##  Two 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
```

```
# variable y2
t.test(y1$V4, y2$V4, var.equal = TRUE)
```

```
##
##  Two 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
```

```r
# variable y3
t.test(y1$V5, y2$V5, var.equal = TRUE)
```

```
##
##  Two 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
```

```r
# variable y4
t.test(y1$V6, y2$V6, var.equal = TRUE)
```

```
##
##  Two 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) Compare the results of (a), (b) and (d) as to the contribution of each variable to separation of the groups.

Based on the results in (a), the contribution of each variable is ranked by $Y_1 > Y_4 > Y_2 > Y_3$. Based on the results in (b), the contribution of each variable is ranked by $Y_1 > Y_2 > Y_4 > Y_3$. Based on the results in (d), the contribution of each variable is ranked by $Y_3 > Y_4 > Y_1 > Y_2$.