Final project

STA 135

Instructor: Xiaodong Li

Student Name: Xuecheng Zhang

Student ID: 915942842

1. Introduction

Goal: Assessing whether pulmonary function in non-pathological population varies according to gender.

Methods: Asked subjects to run on a treadmill until exhaustion. Samples of air were collected at definite intervals and the gas contents analyzed. The results on 4 measures of oxygen consumption for 25 males and 25 females.

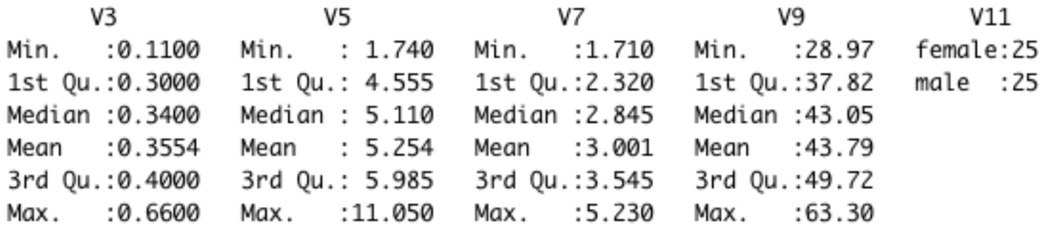
* V3 = resting volume O2(L/min) * V5 = resting volume O2(mL/kg/min)

* V7 = maximum volume O2 (L/min) * V9 = maximum volume O2 (mL/kg/min)

Maximal oxygen consumption reflects cardiorespiratory fitness and endurance capacity in exercise performance.

2. Summary

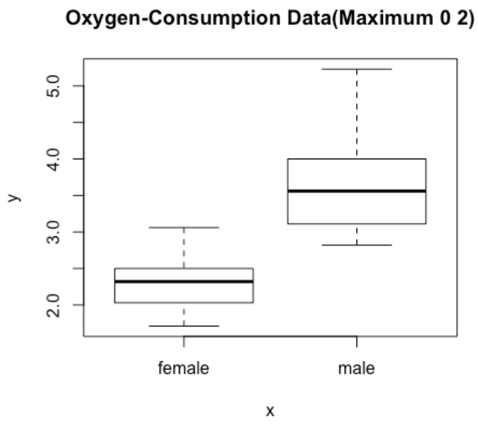
2.1. <u>the summary table of data:</u>

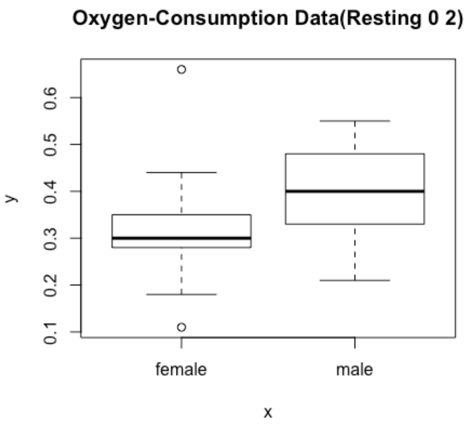


```
       V3               V5               V7              V9             V11
 Min.   :0.1100   Min.   : 1.740   Min.   :1.710   Min.   :28.97   female:25
 1st Qu.:0.3000   1st Qu.: 4.555   1st Qu.:2.320   1st Qu.:37.82   male  :25
 Median :0.3400   Median : 5.110   Median :2.845   Median :43.05
 Mean   :0.3554   Mean   : 5.254   Mean   :3.001   Mean   :43.79
 3rd Qu.:0.4000   3rd Qu.: 5.985   3rd Qu.:3.545   3rd Qu.:49.72
 Max.   :0.6600   Max.   :11.050   Max.   :5.230   Max.   :63.30
```

Form data we got that for the non-pathological population, the mean of Resting O2 and the mean of Maximum O2 in L/min are around 3, but the gap between Resting O2 and the mean of Maximum O2 in ml/kg/min is large. It means that after heavy exercise, the oxygen consumption of non-pathological population increased.

2.2. <u>A scatterplot with the DATA points Labelled by Resting and Maximum Group:</u>





Moreover, the resting volume O2 for Male and Female is kind same. After heavy exercise, male will need more O2 than female.

3. Analysis:

3.1. Two-sample Hotelling T^2-test:

Two independent p-variate random samples with the same population covariance $\Sigma_1 = \Sigma_2 = \Sigma$

$$X_{11\dots},X_{1n1} \sim N_p(\vec{\mu}_1,\Sigma_1)$$

$$X_{21}, \dots ,X_{2n2} \sim N_p(\vec{\mu}_2,\Sigma_2)$$

Let

$$x_{11}, \dots , x_{1n1} \text{ (X are vector)}$$

$$x_{21}, \dots , x_{2n2}\text{(X are vector)}$$

be two observed samples with the summary statistics $\vec{x}_1$, $\vec{x}_2$, S_1 and S_2. We want to test the hypothesis:

$$H_0: \overrightarrow{\mu_1} = \overrightarrow{\mu_2}$$

Then we have:

$$\overrightarrow{\overline{X_1}}\text{-}\overrightarrow{\overline{X_2}} \sim N_p\left(\overrightarrow{\mu_1} - \overrightarrow{\mu_2}, \left(\frac{1}{n_1} + \frac{1}{n_2}\right)\Sigma\right)$$

and

$$\left(\left(\overrightarrow{\overline{X_1}} - \overrightarrow{\overline{X_2}}\right) - (\overrightarrow{\mu_1} - \overrightarrow{\mu_2})\right)^T \left(\left(\frac{1}{n1} + \frac{1}{n2}\right) S_{pooled}\right)^{-1} \left(\left(\overrightarrow{\overline{X_1}} - \overrightarrow{\overline{X_2}}\right) - (\overrightarrow{\mu_1} - \overrightarrow{\mu_2})\right)$$

= 96.37322

then the test is

$$\frac{(n_1 + n_2 - 2)p}{n_1 + n_2 - 1 - p} F_{p, n_1 + n_2 - 1 - p}$$

= 11.00262

We reject $H_0: \overrightarrow{\mu_1} = \overrightarrow{\mu_2}$ at α = 0.05, so we use the simultaneous confidence intervals to check the significant components.

3.2. <u>Construct simultaneous confidence intervals</u>

3.2.1 simultaneous confidence intervals based on T^2:

Two Independent random samples with the same population variance $\Sigma_1 = \Sigma_2 = \Sigma$

We know:

$$X_{11},, X_{1n1} \sim N_p(\overrightarrow{\mu_1}, \Sigma)$$

$$X_{21} ..., X_{2n2} \sim N_p(\overrightarrow{\mu_2}, \Sigma)$$

Let

$$x_{11}, \dots x_{1n1}$$

$$x_{21}, \dots x_{2n2}$$

be two samples with the summary statistics $\overline{x_1}$, $\overline{x_2}$ and $S_1{}^2$, $S_2{}^2$.

We want to test the hypothesisH_0: $\overrightarrow{\mu_1}$ =$\overrightarrow{\mu_2}$. The random sample means obey the following sampling normal distributions and $\overline{X_1}$, $\overline{X_2}$ are independent

we have:

$$\overrightarrow{\overline{X_1}}\text{-}\overrightarrow{\overline{X_2}} \sim N_p\left(\overrightarrow{\mu_1} - \overrightarrow{\mu_2}, \left(\frac{1}{n_1}+\frac{1}{n_2}\right)\Sigma\right)$$

Then we holds:

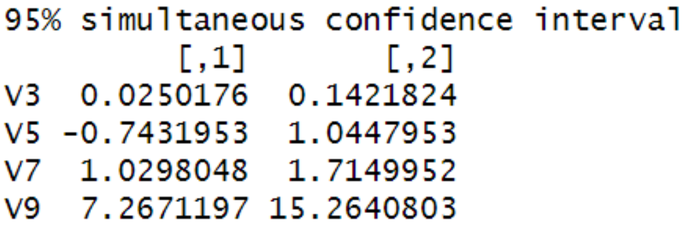
$$\frac{\left(\overrightarrow{\overline{X_1}}-\overrightarrow{\overline{X_2}}\right)-\left(\overrightarrow{\mu_1}-\overrightarrow{\mu_2}\right)}{diag(S_{pooled})\sqrt{\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}} \sim t_{n_1+n_2-2}$$

Based on the two observed samples, the 95% confidence interval for $\overrightarrow{\mu_1} - \overrightarrow{\mu_2}$ is

$$\left[\overline{x_1} - \overline{x_2} - diag(S_{pooled})\sqrt{\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}\; t_{n_1+n_2-2}{}^{(\alpha)}, \overline{x_1} - \overline{x_2} \right.$$

$$\left. + diag(S_{pooled})\sqrt{\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}\; t_{n_1+n_2-2}{}^{(\alpha)}\right]$$

Then we got：



```
95% simultaneous confidence interval
         [,1]       [,2]
V3  0.0250176  0.1421824
V5 -0.7431953  1.0447953
V7  1.0298048  1.7149952
V9  7.2671197 15.2640803
```

According to 95% simultaneous confidence interval based on T^2, we know only V5(the resting O_2 (ml/kg/min)) cover 0. So V3(Resting O_2 (l/min)), V7(max O_2 (l/min)),V9(max O_2 (ml/kg/min)) have significant differences.

3.2.1. Bonferroni simultaneous confidence intervals:

Two random sample:

$$\mathrm{X}_{11}...,\mathrm{X}_{1n1} \sim \mathrm{N}_{\mathrm{p}}(\overrightarrow{\mu_1}, \Sigma_1)$$

$$\mathrm{X}_{21}...,\mathrm{X}_{2n2} \sim \mathrm{N}_{\mathrm{p}}(\overrightarrow{\mu_2}, \Sigma_2)$$

let

$$x_{11j}, \dots x_{1n1j} \sim \mathrm{N}_{\mathrm{p}}(\overrightarrow{\mu_{1j}}, {\sigma_{1j}}^2)$$

$$x_{21j}, \dots x_{2n2j} \sim \mathrm{N}_{\mathrm{p}}(\overrightarrow{\mu_{2j}}, {\sigma_{2j}}^2)$$

By the assumption of equal variances: , ${\sigma_{1j}}^2$=${\sigma_{2j}}^2$=${\sigma_j}^2$, j=1…p and the independence of the two samples, we have

$$\overline{\overrightarrow{X_{1j}}}\text{-}\overline{\overrightarrow{X_{2j}}} \sim N_p\left(\overrightarrow{\mu_{1j}} - \overrightarrow{\mu_2 j}, \left(\frac{1}{n_1}+\frac{1}{n_2}\right){\sigma_j}^2,\right)$$

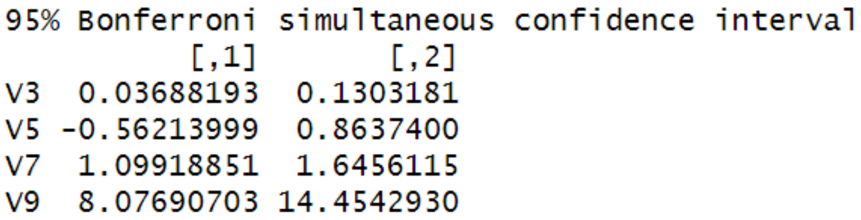
we have the following sampling distribution result

$$\frac{\left(\overline{\overrightarrow{X_{1j}}}-\overline{\overrightarrow{X_{2j}}}\right)-\left(\overrightarrow{\mu_{1j}}-\overrightarrow{\mu_{2j}}\right)}{s_{pooled,j}\sqrt{\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}} \sim t_{n_1+n_2-2}$$

Based on the two observed samples, the 95% confidence interval for $\overrightarrow{\mu_1} - \overrightarrow{\mu_2}$ is

$$\left[\overline{x_{1J}} - \overline{x_{2J}} - S_{pooled,j}\sqrt{\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}\; t_{n_1+n_2-2}{}^{\frac{\alpha}{2p}}, \overline{x_{1J}} - \overline{x_{2J}} + S_{pooled,j}\sqrt{\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}\; t_{n_1+n_2-2}{}^{\frac{\alpha}{2p}}\right]$$

Then we got：



```
95% Bonferroni simultaneous confidence interval
          [,1]       [,2]
V3  0.03688193  0.1303181
V5 -0.56213999  0.8637400
V7  1.09918851  1.6456115
V9  8.07690703 14.4542930
```

According to 95% simultaneous confidence interval based on Bonferroni correction, the answer is same, only V5(the resting O_2 (ml/kg/min)) cover 0. However, the Bonferroni is narrower.

3.3 Principal Component Analysis

3.3.1 We would like to explain the variance-covariance structure of a set of variables by a few linear combinations of these variables.

Let

$$\vec{X} = \begin{bmatrix} X1 \\ \vdots \\ Xp \end{bmatrix}$$

be a random vector and

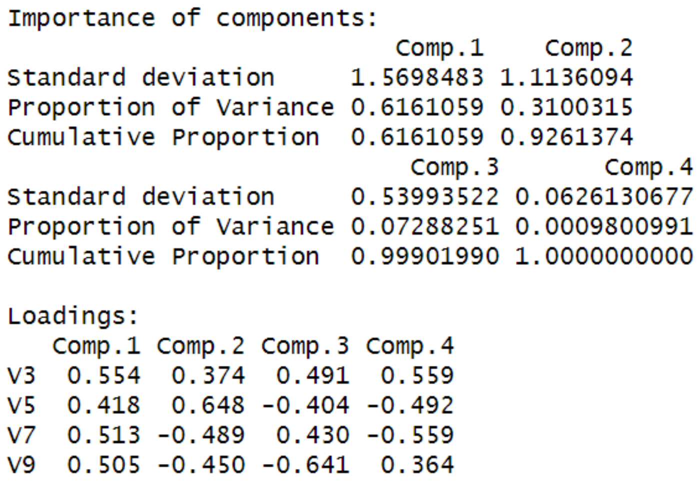
$$\vec{a} = \begin{bmatrix} a1 \\ \vdots \\ a_p \end{bmatrix}$$

be a deterministic vector to be determined. The first principal component for $\vec{X}$ is defined as

$$Y_1 = \vec{a}_1{}^T\vec{X}+\vec{a}_{11}{}^T X_1 + \cdots + \vec{a}_{1p}{}^T X_P$$

$$Y_2 = \vec{a}_2{}^T\vec{X}+\vec{a}_{21}{}^T X_1 + \cdots + \vec{a}_{2p}{}^T X_P$$

Then we got :

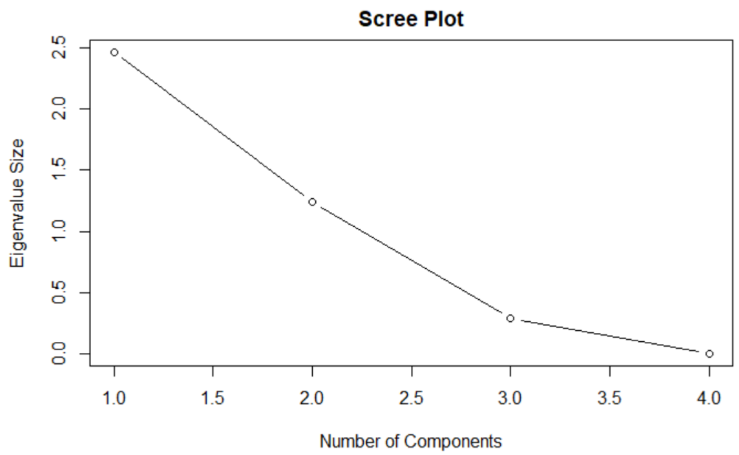


```
Importance of components:
                          Comp.1    Comp.2
Standard deviation     1.5698483 1.1136094
Proportion of Variance 0.6161059 0.3100315
Cumulative Proportion  0.6161059 0.9261374
                           Comp.3       Comp.4
Standard deviation     0.53993522 0.0626130677
Proportion of Variance 0.07288251 0.0009800991
Cumulative Proportion  0.99901990 1.0000000000

Loadings:
   Comp.1 Comp.2 Comp.3 Comp.4
V3  0.554  0.374  0.491  0.559
V5  0.418  0.648 -0.404 -0.492
V7  0.513 -0.489  0.430 -0.559
V9  0.505 -0.450 -0.641  0.364
```

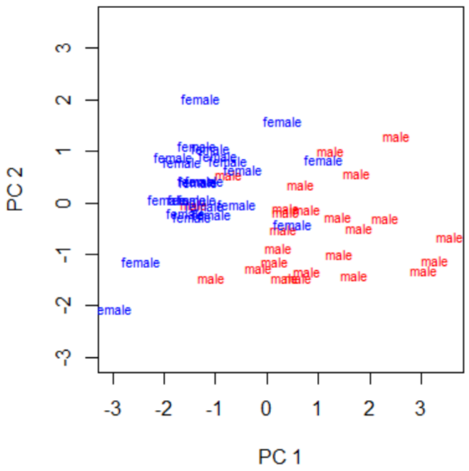
From the Cumulative Proportion we got Comp.1 = 0.6161, and Comp.2 = 0.9261, 0.9261 is > 0.9, we choose first and second Principal Component.

3.3.2 <u>A Scree Plot:</u>



the "elbow" occurs at point 3, but from the summery we got PC1 and PC2 is good enough for us.

3.3.3 <u>Plotting the PC scores for the sample data in the space of the first two principal components:</u>



Then from this plot, for PC1 male is on the positive side and female is on the negative side; on the contrary for PC2 female is on the positive side.

3.4 Linear discriminant analysis

3.4.1 compute pooled estimate for the covariance matrix and plot decision boundary:

There are two classes π_1 and π_2 corresponding to two distributions D1 and D2. Suppose the density of $D_i$is $f_i(\vec{x})$For simplicity, in this class we assume the prior of π_1 and π_2 are 1/2 and 1/2.

The distribution of $\overrightarrow{X}$ is denoted as the mixture

$$\frac{1}{2}D_1+\frac{1}{2}D_2.$$

Linear Discriminant Analysis:

Let π_1: $N(\vec{\mu}_1,\Sigma)$, and π_2: $N_p(\vec{\mu}_2,\Sigma)$, and we assume $\Sigma_1 = \Sigma_2 = \Sigma$ then

$$f_1(\vec{x})=\frac{1}{2\pi^{\frac{p}{2}}|\Sigma|^{\frac{1}{2}}}\exp\left(-\frac{1}{2}(\vec{x}-\overrightarrow{\mu_1})^T\Sigma^{-1}(\vec{x}-\overrightarrow{\mu_1})\right)$$

$$f_2(\vec{x})=\frac{1}{2\pi^{\frac{p}{2}}|\Sigma|^{\frac{1}{2}}}\exp\left(-\frac{1}{2}(\vec{x}-\overrightarrow{\mu_2})^T\Sigma^{-1}(\vec{x}-\overrightarrow{\mu_2})\right)$$

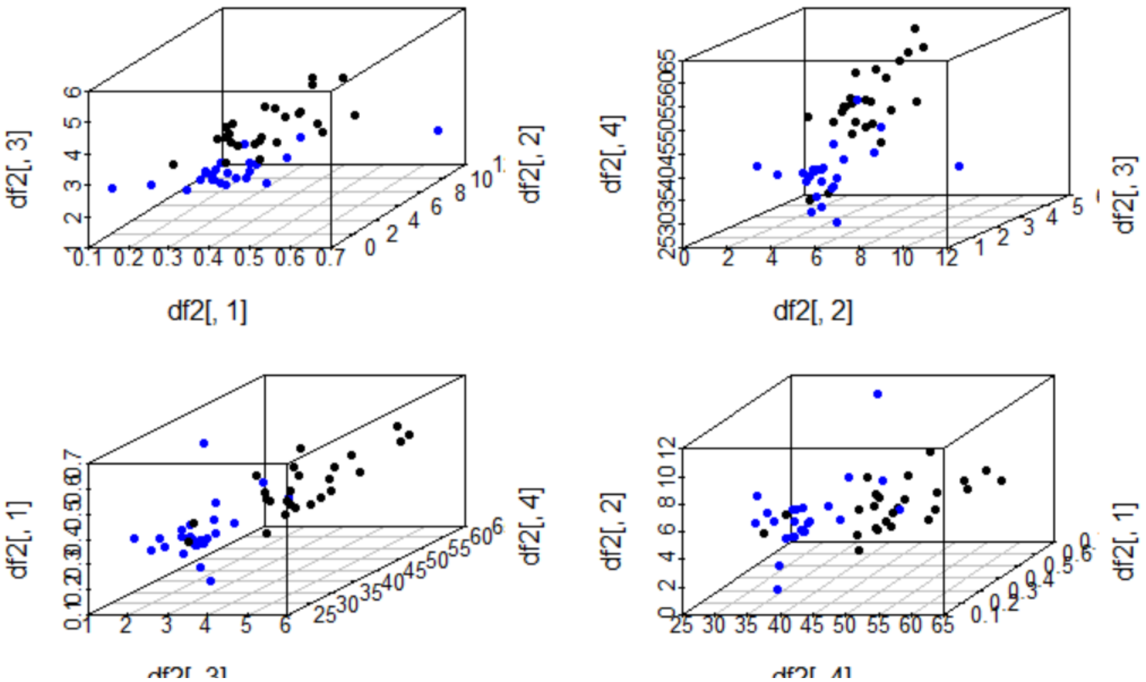
This gives us:

$$\log\frac{f_1(\vec{x})}{f_2(\vec{x})}=(\overrightarrow{\mu_1}-\overrightarrow{\mu_2})^T\Sigma^{-1}\left(\vec{x}-\frac{1}{2}(\overrightarrow{\mu_1}+\overrightarrow{\mu_2})\right)$$

The conditional probability criterion gives the following classification rule $\vec{x}$ is assigned to Class 1 if

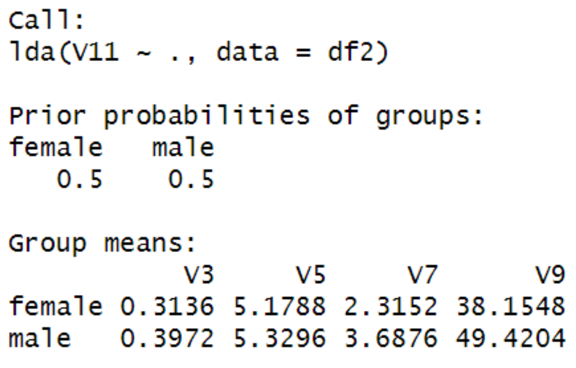
$$(\overrightarrow{\mu_1}-\overrightarrow{\mu_2})^T\Sigma^{-1}\left(\vec{x}-\frac{1}{2}(\overrightarrow{\mu_1}+\overrightarrow{\mu_2})\right)\geq 0$$

and Class 2 otherwise.



From the 3d scatterplot, we compare three at one time then we got four plots, and from the plot we can see clearly male and female are separated by a bound line.

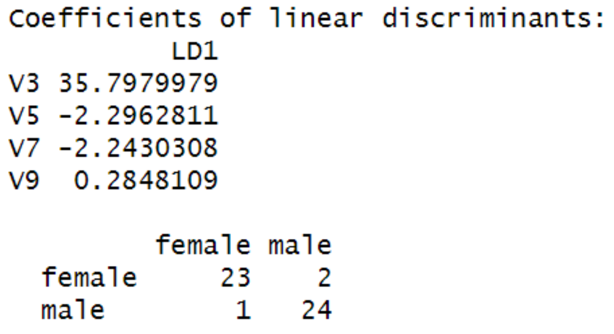
3.4.2 Determine how well the model fits:



```
Call:
lda(V11 ~ ., data = df2)

Prior probabilities of groups:
female   male
   0.5    0.5

Group means:
           V3     V5     V7      V9
female 0.3136 5.1788 2.3152 38.1548
male   0.3972 5.3296 3.6876 49.4204
```



```
Coefficients of linear discriminants:
           LD1
V3  35.7979979
V5  -2.2962811
V7  -2.2430308
V9   0.2848109

          female male
 female       23    2
 male          1   24
```

For Female group all the sample are positive but except two sample on the positive.

for Male group all sample are positive expect one sample.

4 Interpretation:

4.1 Two-sample Hotelling T^2-test:

Since we only have two sample, we start at Hotelling's T^2 test. The result for O_2: H_0: $\overrightarrow{\mu_1} = \overrightarrow{\mu_2}$ is reject. It's means the overall mean of female is not equal to overall mean of male. Since we reject the null, we use the simultaneous confidence intervals to check significant components.

4.2 T^2-test: we got

V5(the resting O_2 (ml/kg/min)) ∈ [-0.7431953, 1.0447953], it cover 0 shows us that the mean of male's resting $O_2$is equal to mean of female's resting O_2.

V3 ∈ [0.0250176, 0.1421824] it not cover 0 shows us that the mean of male's resting O_2 (l/min)is not equal to mean of female's resting O_2.

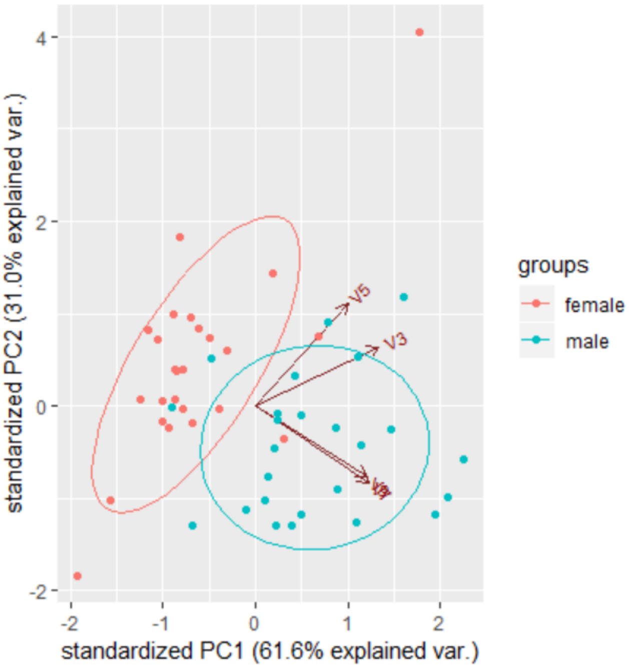
V7 ∈ [1.0298048, 1.7149952] it not cover 0 shows us that the mean of male's max O_2 (l/min)is not equal to mean of female's max O_2.

V9 ∈ [7.2671197, 15.2640803] it not cover 0 shows us that the mean of male's resting O_2 (ml/kg/min)is not equal to mean of female's resting O_2.

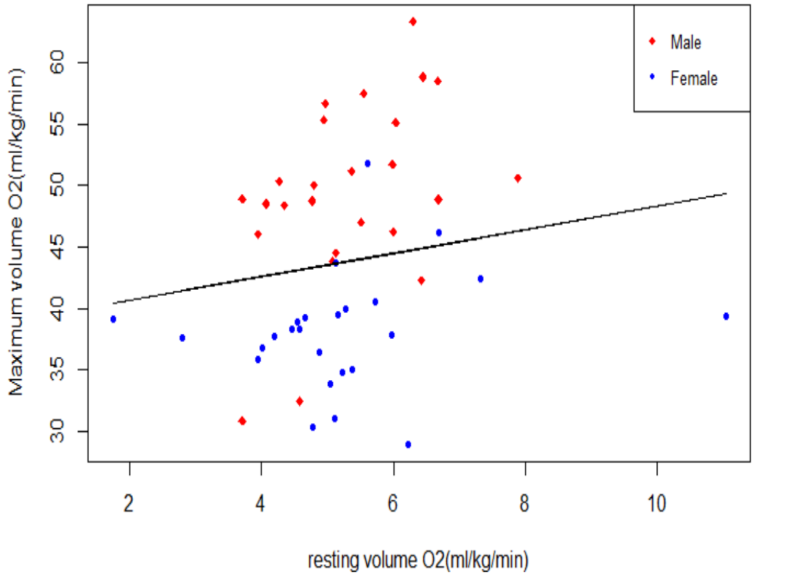
In Bonferroni correction : we got the same conclusion but the for test μ_1, μ_2 μ_3 μ_4 Bonferroni is narrow than T^2-test.

4.3 Principal Component Analysis

Since we have 4 variables, we would like to explain the variance-covariance structure of a set of variables by a few linear combinations of these variables. We got Comp.1 = 0.6161, and Comp.2 = 0.9261, and 0.9261 is > 0.9, we choose first and second Principal Component. So PC1 is the O_2process from rest to end of the exercise, it seems that 61.6% of the variation in the data are related to differences in this process. From this plot, we know the male form a distinct cluster to the right and female form a distinct cluster to the left. V3(resting volume O_2 (L/min)), V5(resting volume O_2 (L/min)) are slight toward male population, V7(maximum volume O_2 (L/min)),V9(maximum volume O_2 (mL/kg/min)) total toward female population. V7,V9 can describe more details and separate two population.

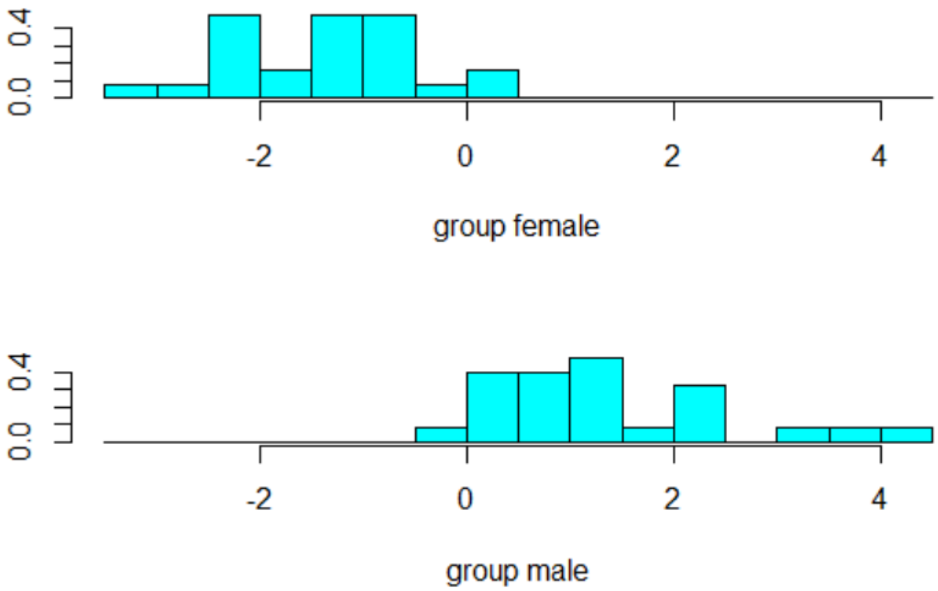


4.4 Linear discriminant analysis



There are two classes Male and Female corresponding to two distributions D1 and D2. Simplicity, in this class we assume the prior of Male and Female are 1/2 and 1/2. We need to find a function that can separate Male and Female. For example, we only consider

V5(resting volume O_2 (L/min)) and V9(maximum volume O_2 (mL/kg/min)), we got this plot on the left. As we can see this black line is divied Male and Female tow population, but not prefect: there are two females in male group, and three males in female group, and still some points on the black line, those points we can not separate clearly.



For the 3D plots we got sepatation for 4 varibles, in this plot, the 0 is the line to separate famle and male. In Female Group we have 2 sample are divided to wrong group, and in Male Group there is only 1 sample is divided to wrong group. In general, this modle have 94% accuracy to saparation.

5 Conclusion

When male and female breathe normally, there is little difference in the oxygen content they inhale, which may be equal at some time. After vigorous exercise, the oxygen content of male and female increased significantly, and the oxygen volume of men inhaled was significantly higher than that of women. So men in non-pathological groups need more oxygen during exercise. Among the four variables, V7(maximum volume O_2 (L/min)),V9(maximum volume O_2 (mL/kg/min))more clearly describe men's oxygen demand during exercise.Thus, the important criteria of male and female oxygen demand for maximum volume O_2 (L/min)), and maximum volume O_2 (mL/kg/min) variables are explained.

All the code and function are from TA’s notes and lecture nots.

```r
setwd("~/Desktop/")
df2 <- read.csv("T6-12.dat",header = F, sep = ' ')
df2 <- df2[,c(3,5,7,9,11)]
summary(df2)
plot(df2$V11, df2$V3, main="Oxygen-Consumption Data(Resting O 2)")
plot(df2$V11, df2$V7, main="Oxygen-Consumption Data(Maximum O 2)")
library(ICSNP)
male <- df2[df2$V11 == "male",-5]
female <- df2[df2$V11 == "female",-5]
HotellingsT2(male, female)
n<-c(25,25)
p<-4
xmean1<-colMeans(male)
xmean2<-colMeans(female)
d<-xmean1-xmean2
S1<-var(male)
S2<-var(female)
Sp<-((n[1]-1)*S1+(n[2]-1)*S2)/(sum(n)-2)
t2 <- t(d)%*%solve(sum(1/n)*Sp)%*%d
t2
alpha<-0.05
cval <- (sum(n)-2)*p/(sum(n)-p-1)*qf(1-alpha,p,sum(n)-p-1)
cval
alpha<-0.05
male <- df2[df2$V11 == "male",-5]
female <- df2[df2$V11 == "female",-5]
n<-c(50,50)
p<-4
p<-4
xmean1<-colMeans(male)
xmean2<-colMeans(female)
d<-xmean1-xmean2
S1<-var(male)
S2<-var(female)
```

```
Sp<-((n[1]-1)*S1+(n[2]-1)*S2)/(sum(n)-2)
wd<-sqrt(((n[1]+n[2]-2)*p/(n[1]+n[2]-p-1))*qf(1-alpha,p,n[1]+n[2]-p-1))*sqrt
(diag(Sp)*sum(1/n))
Cis<-cbind(d-wd,d+wd)
cat("95% simultaneous confidence interval","\n")
Cis
wd.b<- qt(1-alpha/(2*p),n[1]+n[2]-2) *sqrt(diag(Sp)*sum(1/n))
Cis.b<-cbind(d-wd.b,d+wd.b)
cat("95% Bonferroni simultaneous confidence interval","\n")
Cis.b
attach(df2)
df2.pc <- princomp(df2[,1:4], cor=T)
summary(df2.pc,loadings=T)
plot(1:(length(df2.pc$sdev)),  (df2.pc$sdev)^2, type='b',
     main="Scree Plot", xlab="Number of Components", ylab="Eigenvalue Size")
par(pty="s")
plot(df2.pc$scores[,1], df2.pc$scores[,2], ylim=range(df2.pc$scores[,1]),
     xlab="PC 1", ylab="PC 2", type ='n', lwd=2)
# labeling points with IDs for df2s:
text(df2.pc$scores[,1], df2.pc$scores[,2], labels=V11, cex=0.7, lwd=2,
     col=c(rep("red", times = 25), rep("blue", times=25)) )
library(devtools)
install_github("vqv/ggbiplot")
library(ggbiplot)
X<-df2[,1:4]
groupid<-df2[,5]
ggbiplot(df2.pc,ellipse=TRUE, groups=groupid)
library(rrcov)
par(mar=c(4,4,2,1))
plot(df2$V5,df2$V9,xlab="resting volume O2(ml/kg/min)",ylab="Maximum volume O
2(ml/kg/min)",
     pch=rep(c(18,20),each=25),col=rep(c(2,4),each=25),main="")
legend("topright",legend=c("Male","Female"),pch=c(18,20),col=c(2,4),cex=0.8)

x1<-df2[1:25,c("V5","V9")]
```

```r
x2<-df2[26:50,c("V5","V9")]

# compute sample mean vectors:

x1.mean<-colMeans(x1)

x2.mean<-colMeans(x2)

# compute pooled estimate for the covariance matrix:

S.u<-24*(var(x1)+var(x2))/48

w<-solve(S.u)%*%(x1.mean-x2.mean)

w0<--(x1.mean+x2.mean)%*%w/2

lines(df2[,2],-(w[1]*df2[,2]+w0)/w[2])

library(MASS)

df2.lda <- lda(V11~.,data=df2)

df2.lda# this is very important

df2.pred <- predict(df2.lda)

# Confusion matrix

table(df2$V11,df2.pred$class)

ldahist(data = df2.pred$x[,1], g=df2$V11)

#One can display the 3-dimensional scatterplots.

library(scatterplot3d)

#install.packages("scatterplot3d")

par(mfrow = c(2, 2))

mar0 = c(2, 3, 2, 3)

scatterplot3d(df2[, 1], df2[, 2], df2[, 3], mar = mar0, color = c("blue",
                                                                     "black",
 "red")[df2$V11], pch = 19)

scatterplot3d(df2[, 2], df2[, 3], df2[, 4], mar = mar0, color = c("blue",
                                                                     "black",
 "red")[df2$V11], pch = 19)

scatterplot3d(df2[, 3], df2[, 4], df2[, 1], mar = mar0, color = c("blue",
                                                                     "black",
 "red")[df2$V11], pch = 19)

scatterplot3d(df2[, 4], df2[, 1], df2[, 2], mar = mar0, color = c("blue",
                                                                    "black",
"red")[df2$V11], pch = 19)

detach(package:scatterplot3d)
```