Multivariate Midterm

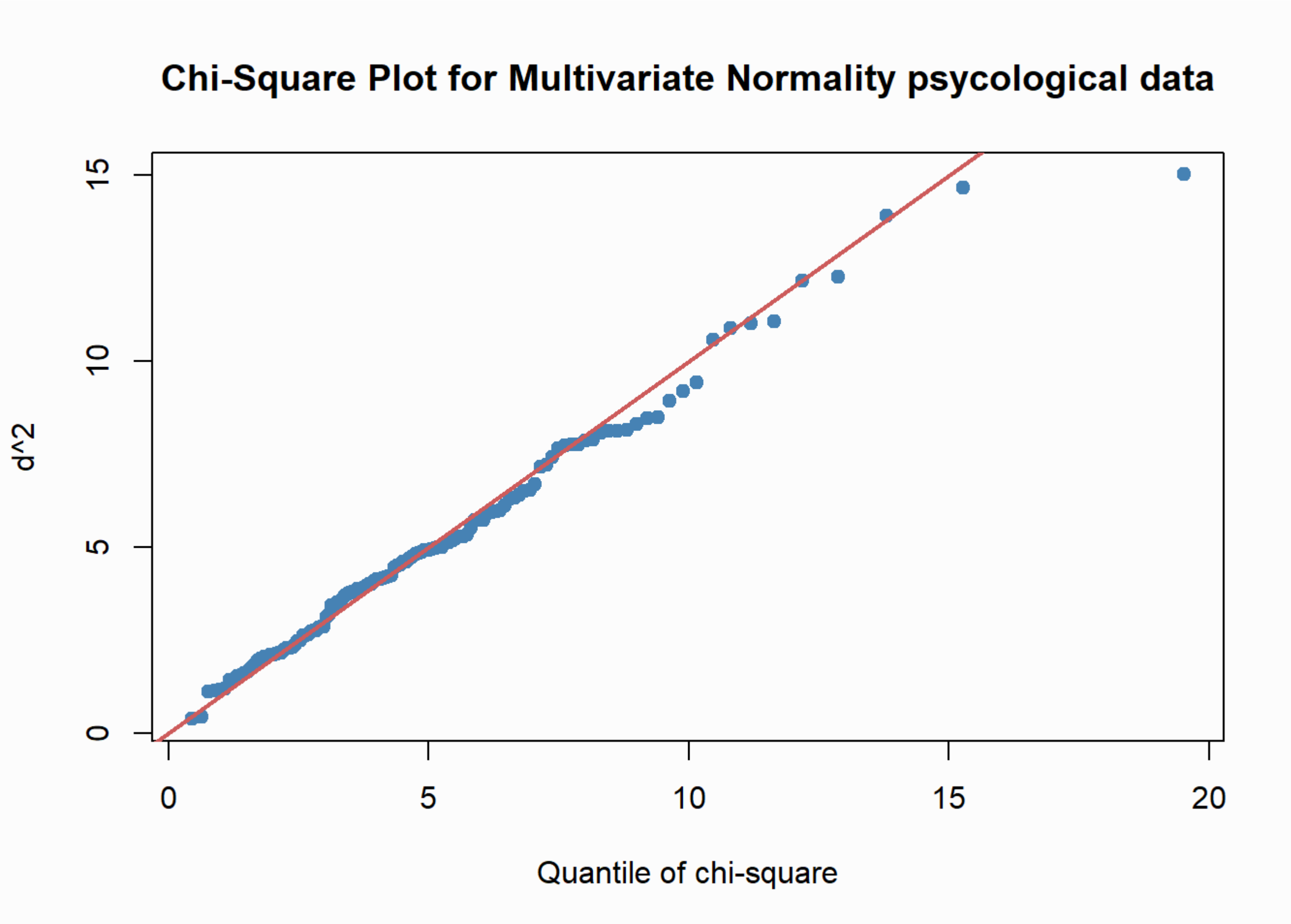
106070020

2021年5月5日

5.

(a)

```
dat <- as.matrix(read.table("C:/Users/eva/Desktop/作業 上課資料(清大)/大四下/多變量/mid/T4-6.DAT"
), header=T)
I<-matrix(rep(1, 130), 130, 1)
xbar<-1/130*t(dat)%*%I
X1<-xbar[1:5,]
S<-var(dat)
S1<-S[1:5,1:5]
dat1_5=dat[,1:5]
sqdist2<-mahalanobis(dat1_5, X1, S1)
sqsortdist2<-sort(sqdist2)
# sqsortdist2
n<-130
prob2<-(c(1:n)-0.2)/n
q2<-qchisq(prob2,5)
par(bg="gray99")
plot(q2,sqsortdist2,xlab='Quantile of chi-square',ylab='d^2',
     main = "Chi-Square Plot for Multivariate Normality psycological data", col = "steelblue", p
ch = 19)
abline(a=0, b=1, col="indianred", lwd=2)
```



> Answer: The QQ-plot and scatter plot are showing above. From the plot, I conclude that the data follows Multivariate normal distribution, as most of the scatter points are along with the line.

5.(b)

```
library(dplyr)
dat<-as.data.frame(dat)
dat1_5=as.data.frame(dat1_5)
low<-filter(.data=dat1_5, dat$V7 == "1")
medium<-filter(.data=dat1_5, dat$V7 == "2")
s1<-data.frame(type=rep(1,nrow(low)),score1=low$V1,score2=low$V2,score3=low$V3,score4=low$V4,sco
re5=low$V5)
s2<-data.frame(type=rep(2,nrow(medium)),score1=medium$V1,score2=medium$V2,score3=medium$V3,score
4=medium$V4,score5=medium$V5)
pmd<-rbind(s1,s2)
tvart<-t.test(cbind(pmd$score1,pmd$score2,pmd$score3,pmd$score4,pmd$score5)~pmd$type , var.equal
= T)
tvart
```

```
##
##  Two Sample t-test
##
## data:  cbind(pmd$score1, pmd$score2, pmd$score3, pmd$score4, pmd$score5) by pmd$type
## t = 1.1308, df = 648, p-value = 0.2586
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.381364  1.416996
## sample estimates:
## mean in group 1 mean in group 2
##        15.98333        15.46552
```

```
tvarf<-t.test(cbind(pmd$score1,pmd$score2,pmd$score3,pmd$score4,pmd$score5)~pmd$type , var.equal
= F)
tvarf
```

```
##
##  Welch Two Sample t-test
##
## data:  cbind(pmd$score1, pmd$score2, pmd$score3, pmd$score4, pmd$score5) by pmd$type
## t = 1.1275, df = 611.53, p-value = 0.26
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.3841306  1.4197628
## sample estimates:
## mean in group 1 mean in group 2
##        15.98333        15.46552
```

> Answer:
>
> - H0 : True difference in means is equal to 0
> - H1 : True difference in means is not equal to 0

- Testing statistics:
  - With equal variance: t = 1.1308, df = 648, p-value = 0.2586
  - Without equal variance: t = 1.1275, df = 611.53, p-value = 0.26
- We cannot reject H0 with or without the equal variance.

5.(c)

```
#(c)
male<-filter(.data=dat, dat$V6 == "1")
female<-filter(.data=dat, dat$V6 == "2")
m_male<-c(mean(male$V4),mean(male$V5))
males11<-var(male$V4)
males12<-var(male$V4,male$V5)
males22<-var(male$V5)
malestr2<-rbind(c(males11,males12),c(males12,males22))

m_fem<-c(mean(female$V4),mean(female$V5))
fems311<-var(female$V4)
fems312<-var(female$V4,female$V5)
fems322<-var(female$V5)
femstr3<-rbind(c(fems311,fems312),c(fems312,fems322))
library(ellipse)
```

```
##
## Attaching package: 'ellipse'
```

```
## The following object is masked from 'package:graphics':
##
##     pairs
```

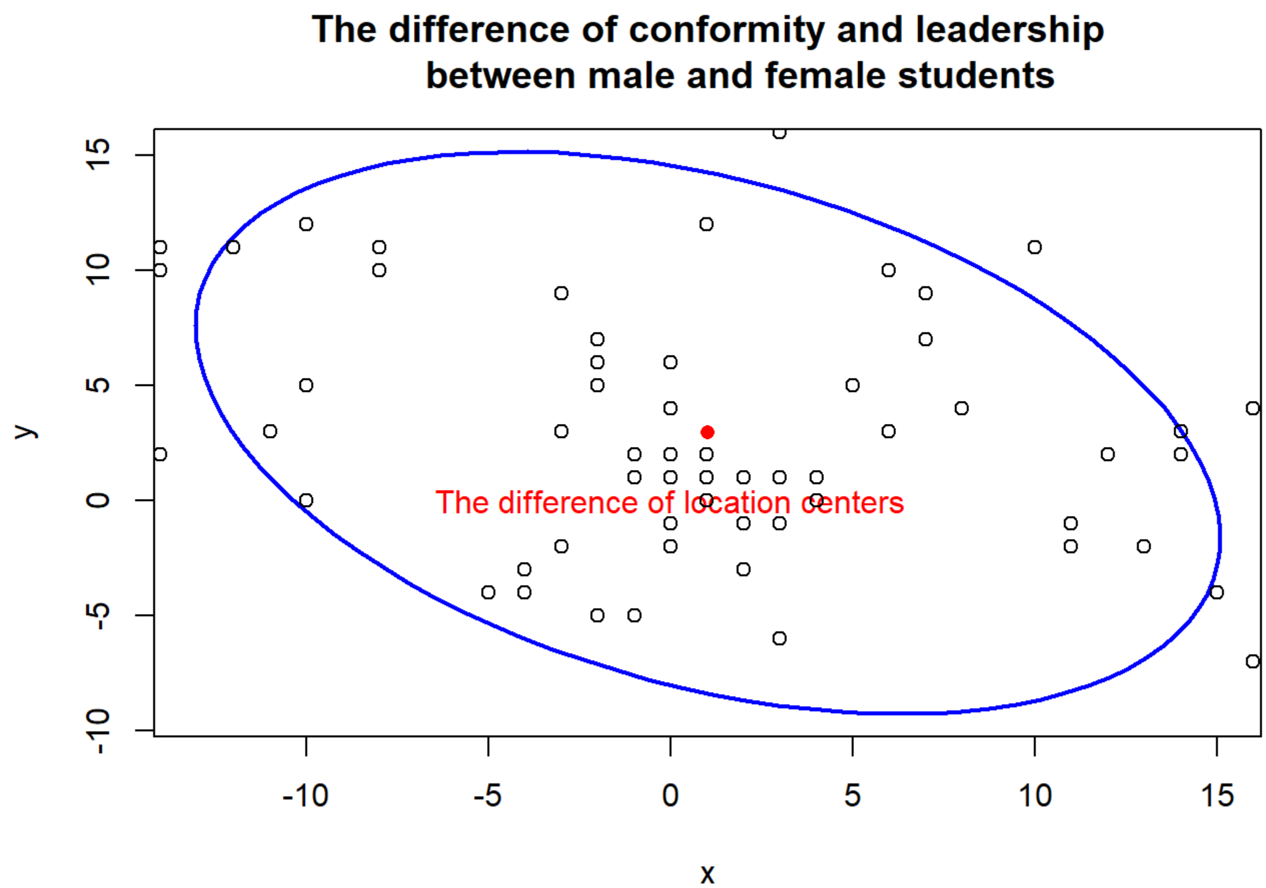
```
nrow(male)
```

```
## [1] 62
```

```
nrow(female)
```

```
## [1] 68
```

```
pps_pool<-(62-1)/(62+68-2)*malestr2+(68-1)/(62+68-2)*femstr3
d.bar<-m_male-m_fem
conf.ellipse<-ellipse(pps_pool,centre=d.bar, level=0.95)
{plot(conf.ellipse, type="l", main='The difference of conformity and leadership
      between male and female students', lwd=2, col='blue')
  points(x=d.bar[1],y=d.bar[2],pch=16, col="red")
  text(x=0, y=0, 'The difference of location centers', col='red')
  points(x=(male$V4-female$V4), y=(male$V5-female$V5), col='black')}
```




```
direction<-eigen(pps_pool)
direction$vectors
```

```
##            [,1]       [,2]
## [1,] -0.8218166 -0.5697521
## [2,]  0.5697521 -0.8218166
```

```
direction$values
```

```
## [1] 40.51221 17.48889
```

```
major_axis<-sqrt(direction$values[1]*qf(.95, 2,128)*2*129/(130*128))
minor_axis<-sqrt(direction$values[2]*qf(.95, 2,128)*2*129/(130*128))
```

Answers

- Direction: the two eigenvectors of the covariance matrix: pps_pool
- The points which outside the ellipse: 7
- calculate the length of major and minor length of the ellipse and them calculate it.

5.(d)

```
#PCA
data<-dat[1:4]
pc <- prcomp(data)
variance_of_4_pc<-(pc$sdev)^2
variance_of_4_pc
```

```
## [1] 62.065754 28.350993 16.489521  8.244715
```

```
pc$rotation
```

```
##            PC1        PC2        PC3       PC4
## V1 -0.65822747  0.2984504 -0.4114669 0.5553008
## V2 -0.02701086 -0.6077049  0.4677757 0.6412106
## V3  0.52514006 -0.3313295 -0.7429085 0.2500714
## V4  0.53873456  0.6571475  0.2448835 0.4668558
```

```
cov_mat<-pc$rotation
e<-eigen(var(data))
e$values
```

```
## [1] 62.065754 28.350993 16.489521  8.244715
```

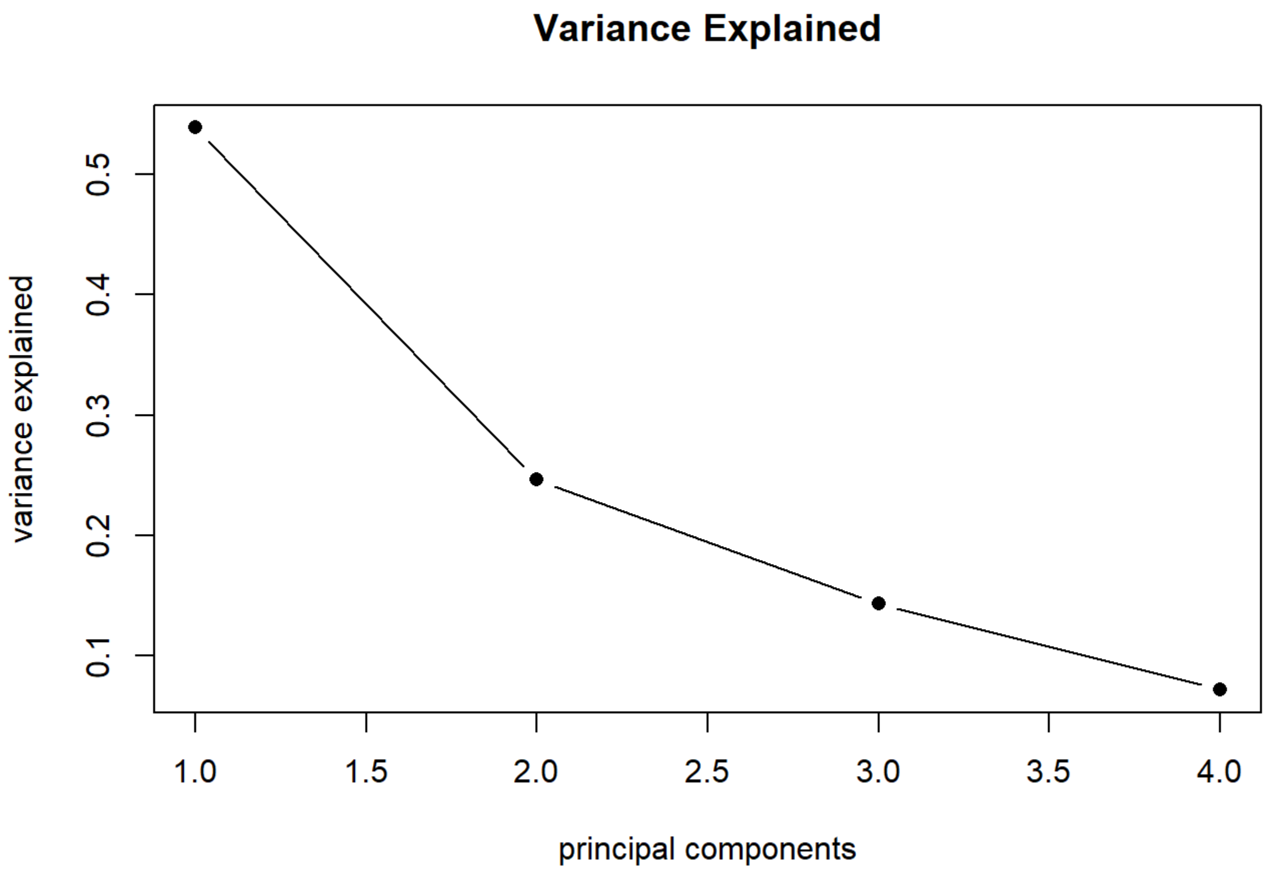
```
e$vectors
```

```
##             [,1]       [,2]       [,3]       [,4]
## [1,]  0.65822747 -0.2984504 -0.4114669 -0.5553008
## [2,]  0.02701086  0.6077049  0.4677757 -0.6412106
## [3,] -0.52514006  0.3313295 -0.7429085 -0.2500714
## [4,] -0.53873456 -0.6571475  0.2448835 -0.4668558
```

```
variance_of_4_pc[1]/sum(variance_of_4_pc)
```

```
## [1] 0.5389946
```

```
plot(pc$sdev^2/sum(pc$sdev^2),type = "b", pch = 16, xlab = "principal components",
     ylab = "variance explained", main="Variance Explained")
```



Answer:

- The variances of four principle components are: 62.065754 28.350993 16.489521 8.244715.
- The largest eigenvalue is 62.065754, so we obtain the first principle component
- Linear model: y1 = e1'v = 0.65822747v1 + 0.02701086v2 - 0.52514006v3 - 0.53873456v4.
- Coefficients are 0.6582275 0.02701086 -0.5251401 -0.5387346.
- Leading relationship: The coefficient of the linear model is the eigenvector of the largest eigenvalue.
- Interpretation of the principle component: The first principle component can explain 54% of the total variance.
- The second principle component cab only explain 25% of the total variance, which is about the half of the first principle component.