Multivariate Analysis HW3

106070020

2021年4月13日

Q3

(a)

```
no2<-c(12, 9, 5, 8, 8, 12, 12, 21, 11, 13, 10, 12, 18, 11, 8, 9, 7, 16, 13,
       9, 14, 7, 13, 5, 10, 7, 11, 7, 9, 7, 10, 12, 8, 10, 6, 9, 6, 13, 9, 8, 11, 6)
o3<-c(8, 5, 6, 15, 10, 12, 15, 14, 11, 9, 3, 7, 10, 7, 10, 10, 7, 4, 2, 5,
      4, 6, 11, 2, 23, 6, 11, 10, 8, 2, 7, 8, 4, 24, 9, 10, 12, 18, 25, 6, 14, 5)

x<-cbind(no2, o3)
# t(x)
#mean
I<-matrix(rep(1, 42), 42, 1)
# I
xbar<-1/42*t(x)%*%I
# xbar
#var-covar matrix
cov<-1/41*t(x)%*%(diag(42)-(1/42)*I%*%t(I))%*%x
# cov
sqdist<-mahalanobis(x, xbar, cov)
sqsortdist<-sort(sqdist)
sqsortdist
```

```
##  [1]  0.1224973  0.1224973  0.1379719  0.1388339  0.1388339  0.1901188
##  [7]  0.3159498  0.4135364  0.4135364  0.4606524  0.4606524  0.4760726
## [13]  0.6228096  0.6370218  0.6592206  0.6592206  0.7032485  0.7874152
## [19]  0.8162468  0.8856041  0.8987982  1.0360061  1.0360061  1.1471939
## [25]  1.1848895  1.3566301  1.4584229  1.6282902  1.8013611  1.8984708
## [31]  2.2488867  2.3770610  2.7741416  2.7782596  3.0089122  3.4437748
## [37]  4.7646873  5.6494392  6.1488606  7.0857237  8.4730649 10.6391792
```

(b)

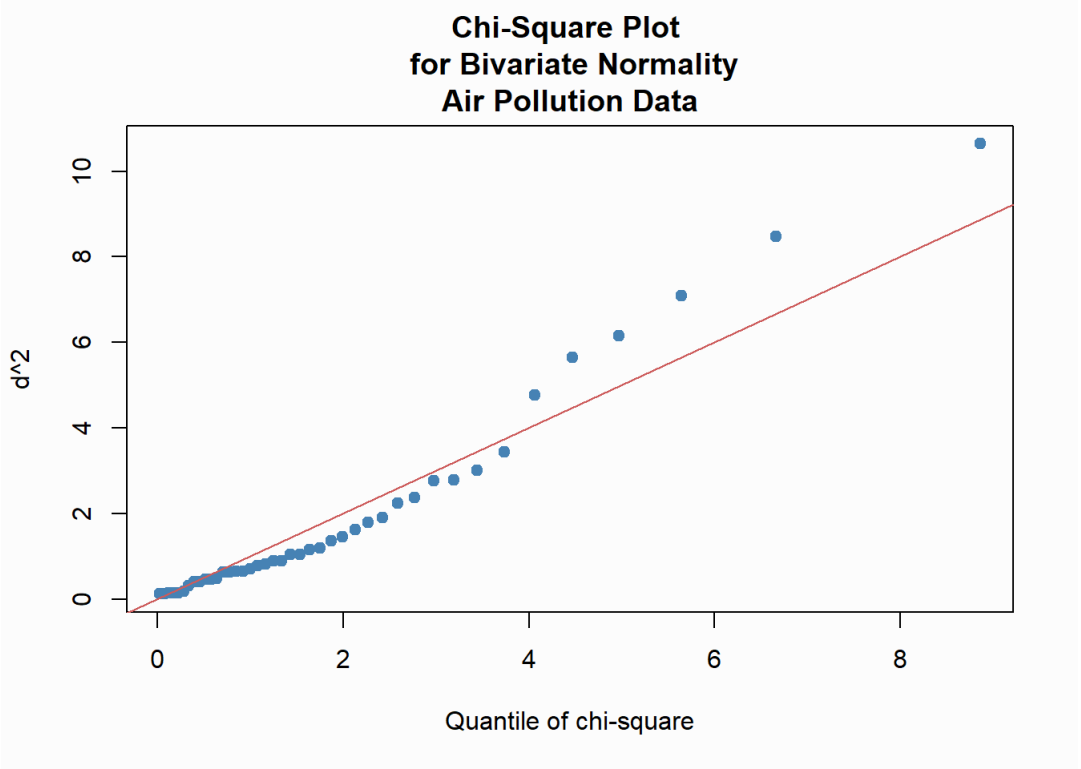
```
#(b)
qcp<-qchisq(.5, 2)
t<-NULL
for(i in 1:length(sqsortdist)){
  if (sqsortdist[i]<=qcp){
    t[i]<-1
  } else {
    t[i]<-0
  }
}
prop.normality.test<-mean(t)
prop.normality.test
```

```
## [1] 0.6190476
```

> Almost 62% of squared distances fall within the approximate 50% probability contour of a bivariate normal distribution. We fail to reject the assumption of bivariate normality here.

(c)

```
#(c)
n<-42
prob<-(c(1:n)-0.5)/n
q1<-qchisq(prob,2)
par(bg="gray99")
plot(q1,sqsortdist,xlab='Quantile of chi-square',ylab='d^2',
     main = "Chi-Square Plot \n for Bivariate Normality\nAir Pollution Data", col = "steelblue", pch = 19)
abline(a=0, b=1, col="indianred")
```



Q4

(a)

```
l2<-c(141, 140, 145, 146, 150, 142, 139)
l3<-c(157, 168, 162, 159, 158, 140, 171)
l4<-c(168, 174, 172, 176, 168, 178, 176)
l5<-c(183, 170, 177, 171, 175, 189, 175)
bear.length<-cbind(l2, l3, l4, l5)

simult.ci<-function(x, n, p){
  crit.value<-sqrt(((p*(n-1))/(n-p)*qf(.05, p, n-p, lower.tail = F)))
  paste("(",mean(x)-crit.value*sqrt(var(x)/n),",", mean(x)+crit.value*sqrt(var(x)/n),")")
}
n<-7
p<-4
simult.ci(l2, n, p) #Length 2 T-SQUARE CI
```

```
## [1] "( 130.685102442155 , 155.886326129274 )"
```

```
simult.ci(l3, n, p)
```

```
## [1] "( 127.021626824303 , 191.549801747126 )"
```

```
simult.ci(l4, n, p)
```

```
## [1] "( 160.308158196709 , 185.977556089005 )"
```

```
simult.ci(l5, n, p)
```

```
## [1] "( 155.37486709837 , 198.910847187344 )"
```

(b)

```
#(b)
successive.diff<-function(x, i, j, n, p){
  mean.dif<-mean(x[,j]-x[,i])
  crit.value<-sqrt(((p*(n-1))/(n-p)*qf(.05, p, n-p, lower.tail = F)))
  var<-var(x)
  lb<-mean.dif-crit.value*sqrt((var[i,i]+var[j,j]-2*var[i,j])/n)
  ub<-mean.dif+crit.value*sqrt((var[i,i]+var[j,j]-2*var[i,j])/n)
  paste("(",lb, ",", ub, ")")

}

successive.diff(bear.length, 1, 2, 7, 4) #Length 3- Length 2 T-SQUARE CI
```

```
## [1] "( -21.2264919444937 , 53.2264919444937 )"
```

```
successive.diff(bear.length, 2, 3, 7, 4) #Length 4- Length 3 T-SQUARE CI
```

```
## [1] "( -22.7307683981843 , 50.44505411247 )"
```

```
successive.diff(bear.length, 3, 4, 7, 4) #Length 5- Length 4 T-SQUARE CI
```

```
## [1] "( -20.6538465602516 , 28.6538465602516 )"
```

(c)

```
#(c)
a.mat<-matrix(c(l3-l2,l5-l4), 7, 2)
a.mat
```

```
##      [,1] [,2]
## [1,]   16   15
## [2,]   28   -4
## [3,]   17    5
## [4,]   13   -5
## [5,]    8    7
## [6,]   -2   11
## [7,]   32   -1
```

```
var(a.mat)
```

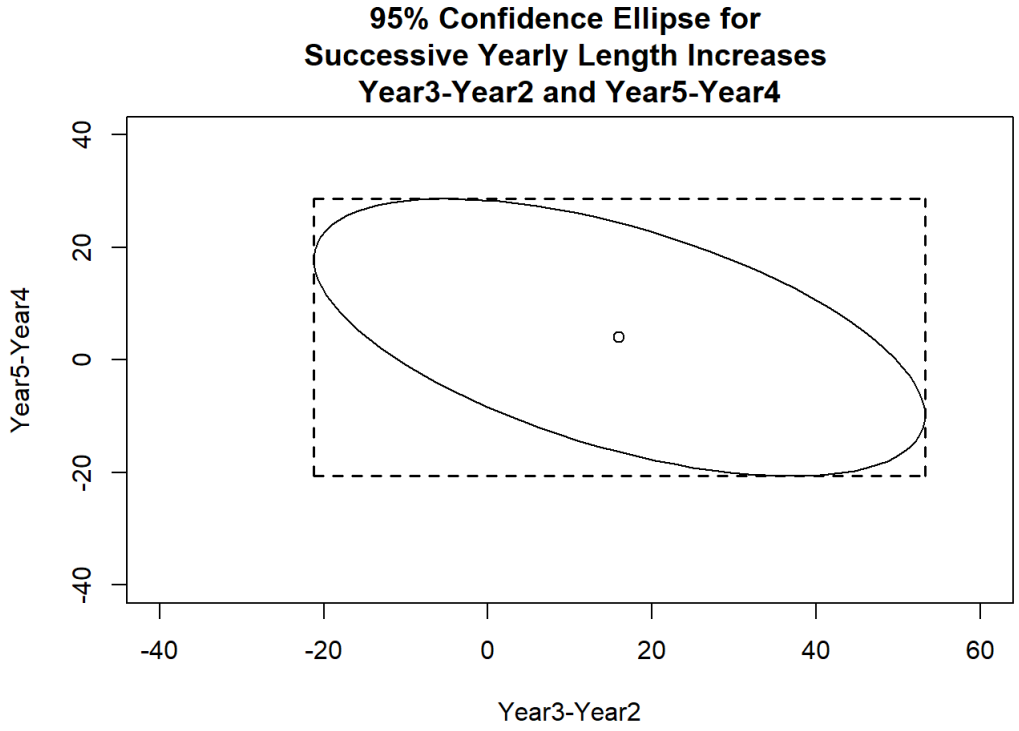
```
##           [,1]      [,2]
## [1,] 133.00000 -49.66667
## [2,] -49.66667  58.33333
```

```
library(ellipse)
```

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

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

```
dbar<-matrix(c(16, 4), 2, 1)
n<-7
p<-4
S<-var(a.mat)
plot(ellipse(S,centre=dbar,t=sqrt(((n-1)*p/(n*(n-p)))*qf(0.95,p,n-p))),type="l",xlim=c(-40,60),ylim=c(-40,40),
     main="95% Confidence Ellipse for \nSuccessive Yearly Length Increases \nYear3-Year2 and Year5-Year4", xlab = "Yea
r3-Year2",
     ylab = "Year5-Year4")
points(dbar[1,],dbar[2,])
lines(x=c(-21.2264919444937 , 53.2264919444937), y=c(-20.6538465602516,-20.6538465602516), lty=2, lwd=1.5)
lines(x=c(-21.2264919444937 , 53.2264919444937), y=c(28.6538465602516, 28.6538465602516), lty=2, lwd=1.5)
lines(x=c(-21.2264919444937, -21.2264919444937), y=c(-20.6538465602516, 28.6538465602516), lty=2, lwd=1.5)
lines(x=c(53.2264919444937, 53.2264919444937), y=c(-20.6538465602516, 28.6538465602516), lty=2, lwd=1.5)
```



(d)

```
#(d)
critical_value<-qt(.05/14, 6, lower.tail = F)
critical_value
```

```
## [1] 3.997061
```

```
bonferonni.cis<-function(m, x, n){
  critical_value<-qt(.05/(2*m), n-1, lower.tail = F)
  paste("(",mean(x)-critical_value*sqrt(var(x)/n),",", mean(x)+critical_value*sqrt(var(x)/n),")")
}

bonferonni.cis(7, l2, 7)#CI FOR LENGTH 2
```

```
## [1] "( 137.388361957808 , 149.18306661362 )"
```

```
bonferonni.cis(7, l3, 7)#CI FOR LENGTH 3
```

```
## [1] "( 144.185440294212 , 174.385988277217 )"
```

```
bonferonni.cis(7, l4, 7)#CI FOR LENGTH 4
```

```
## [1] "( 167.135947109617 , 179.149767176097 )"
```

```
bonferonni.cis(7, l5, 7)#CI FOR LENGTH 5
```

```
## [1] "( 166.9549783036 , 187.330735982114 )"
```

```
successive.diff.bon<-function(x, i, j, n, m){
  mean.dif<-mean(x[,j]-x[,i])
  critical_value<-qt(.05/(2*m), n-1, lower.tail = F)
  var<-var(x)
  lb<-mean.dif-critical_value*sqrt((var[i,i]+var[j,j]-2*var[i,j])/n)
  ub<-mean.dif+critical_value*sqrt((var[i,i]+var[j,j]-2*var[i,j])/n)
  paste("(",lb, ",", ub, ")")

}
successive.diff.bon(bear.length, 1, 2, 7, 7) # Year 3 - Year 2
```

```
## [1] "( -1.42278404051086 , 33.4227840405109 )"
```

```
successive.diff.bon(bear.length, 2, 3, 7, 7) # Year 4 - Year 3
```

```
## [1] "( -3.26677194109882 , 30.9810576553845 )"
```

```
successive.diff.bon(bear.length, 3, 4, 7, 7) # Year 5 - Year 4
```

```
## [1] "( -7.53852060590654 , 15.5385206059065 )"
```

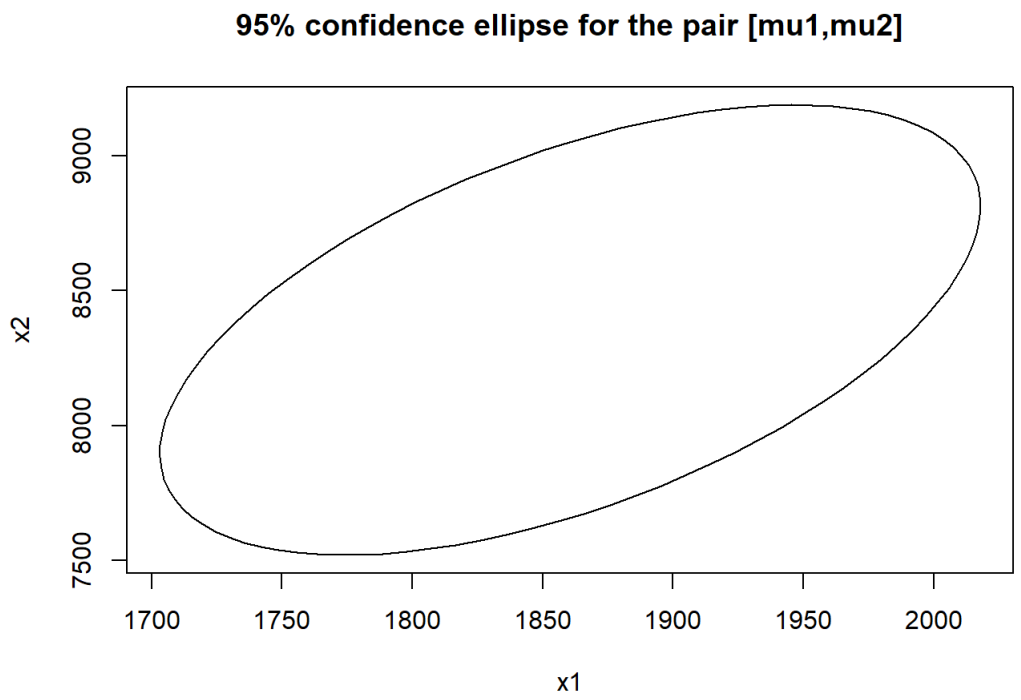
Q5

(a)

```
#Q5(a)
x1<-c(1232,1115,2205,1897,1932,1612,1598,1804,1752,2067,2365,1646,1579,1880,1773,1712,1932,1820,1900,2426,1558,1470,18
58,1587,2208,1487,2206,2332,2540,2322)
x2<-c(4175,6652,7612,10914,10850,7627,6954,8365,9469,6410,10327,7320,8196,9709,10370,7749,6818,9307,6457,10102,7414,75
56,7833,8309,9559,6255,10723,5430,12090,10072)
lumber<-cbind(x1,x2)
# t(lumber)
I<-matrix(rep(1, 30), 30, 1)
# I
xbar<-1/30*t(lumber)%*%I
xbar
```

```
##        [,1]
## x1 1860.500
## x2 8354.133
```

```
S<-var(lumber)
library(ellipse)
conf.ellipse<-ellipse(S/nrow(lumber), centre=xbar, level=0.95)
plot(conf.ellipse, type="l", main="95% confidence ellipse for the pair [mu1,mu2]")
```



(b)

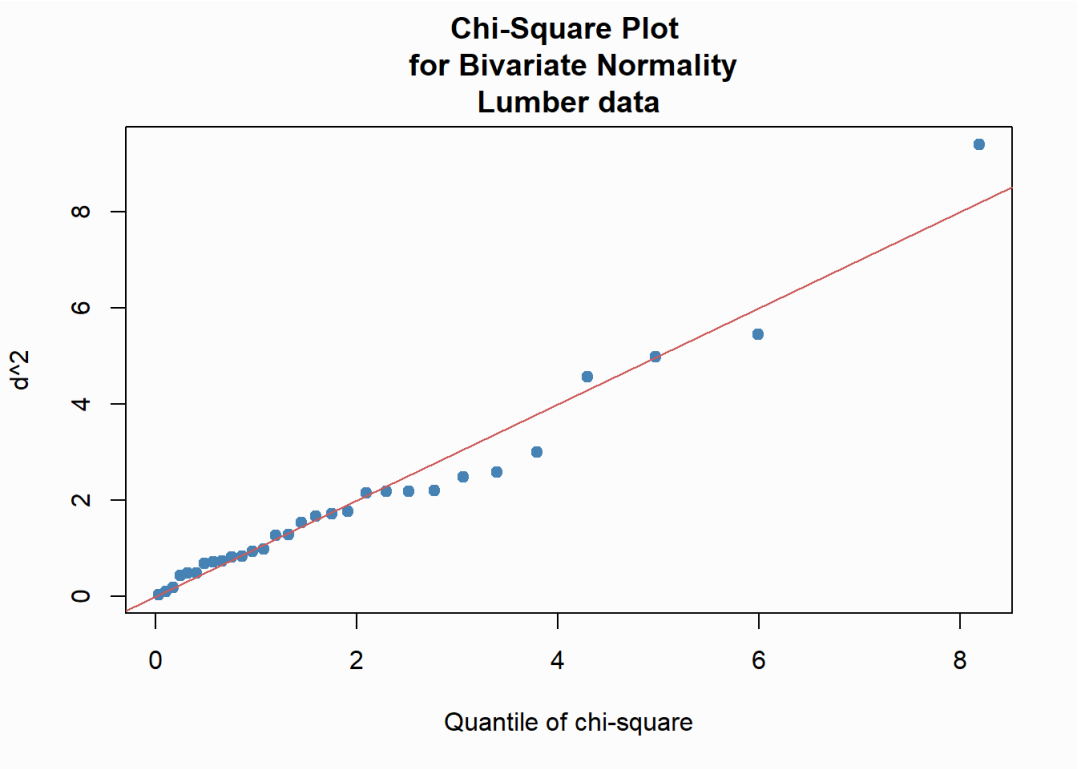
```
#(b)
p = ncol(S)
n = nrow(lumber)
nullmean = c(2000, 10000)
d = xbar-nullmean
t2 <- n*t(d)%*%solve(S)%*%d;
t2mod <- (n-p)*t2/(p*(n-1))
pval <- 1- pf(t2mod,p,n-p)
ifelse(pval<0.05, "Reject", "Fail to Reject")
```

```
##      [,1]
## [1,] "Reject"
```

As the result shown above, based on α=0.05, we can reject the hypothesis that the true mean values of tail length and wing length are 2000 and 10000, respectively. In indeed, the confidence ellipse analysis in part (a) also give the same conclusion, as we can see the “point” corresponds to tail length = 2000 and wing length = 10000 fail ro fall inside the ellipse.

(c)

```
#(c)
sqdist2<-mahalanobis(lumber, xbar, S)
sqsortdist2<-sort(sqdist2)
# sqsortdist2
n<-30
prob2<-(c(1:n)-0.5)/n
q2<-qchisq(prob2,2)
par(bg="gray99")
plot(q2,sqsortdist2,xlab='Quantile of chi-square',ylab='d^2',
     main = "Chi-Square Plot \n for Bivariate Normality\nLumber data", col = "steelblue", pch = 19)
abline(a=0, b=1, col="indianred")
```



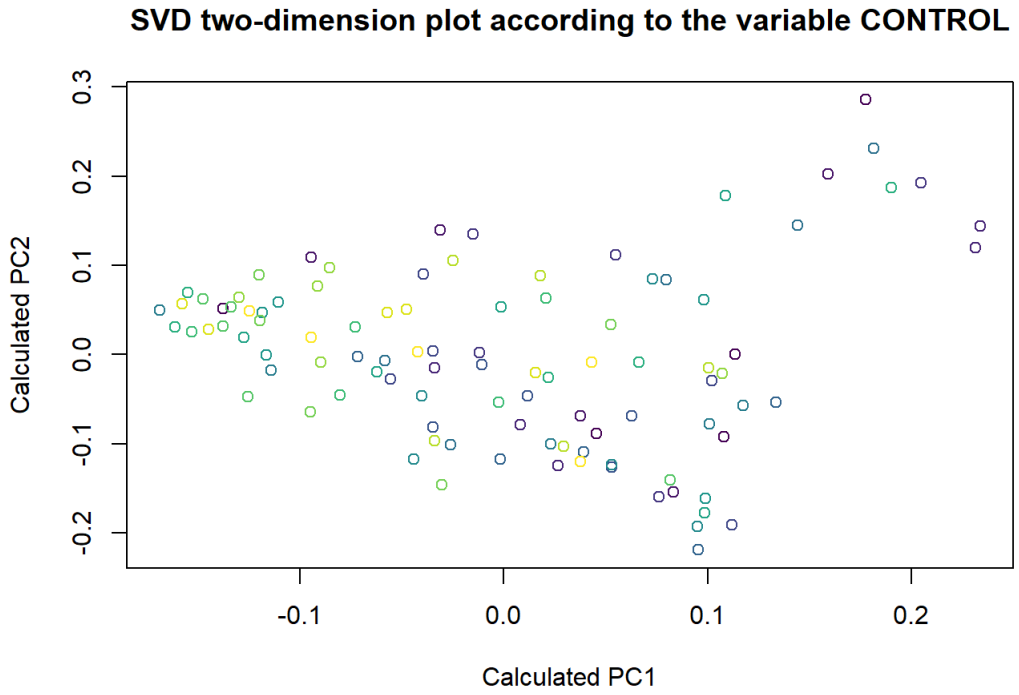
The QQ-plot and scatter plot are showing above. From the plot, I conclude that the data follows bivariate normal distribution, as most of the scatter points are along with the line.

Q6

```
data = read.csv("C:/Users/eva/Desktop/作業 上課資料(清大)/大四下/多變量/hw3/collegeData_hw3_prob5.csv")
library(tidyverse)
```

```
## Warning: package 'dplyr' was built under R version 4.0.5
```

```
data1 = cbind(data["UGDS_WHITE"],data["UGDS_BLACK"],data["UGDS_HISP"]
              ,data["UGDS_ASIAN"],data["UGDS_AIAN"],data["UGDS_NHPI"],data["DEP_STAT_PCT_IND"],data["IND_INC_PCT_LO"]
              ,data["DEP_INC_PCT_LO"],data["INC_PCT_M1"],data["INC_PCT_M2"],data["INC_PCT_H1"],data["INC_PCT_H2"]
              ,data["PAR_ED_PCT_MS"],data["PAR_ED_PCT_HS"],data["PAR_ED_PCT_PS"],data["FEMALE"],data["MARRIED"]
              ,data["DEPENDENT"],data["VETERAN"],data["UGDS_MEN"],data["UGDS_WOMEN"]) %>% na.omit()

# data1
library("viridis")
#SVD
cx <- sweep(data1, 2, colMeans(data1), "-")
# cx
sv <- svd(cx)
plot(sv$u[, 1], sv$u[, 2], col = viridis(20), xlab='Calculated PC1', ylab='Calculated PC2',
     main="SVD two-dimension plot according to the variable CONTROL")
```



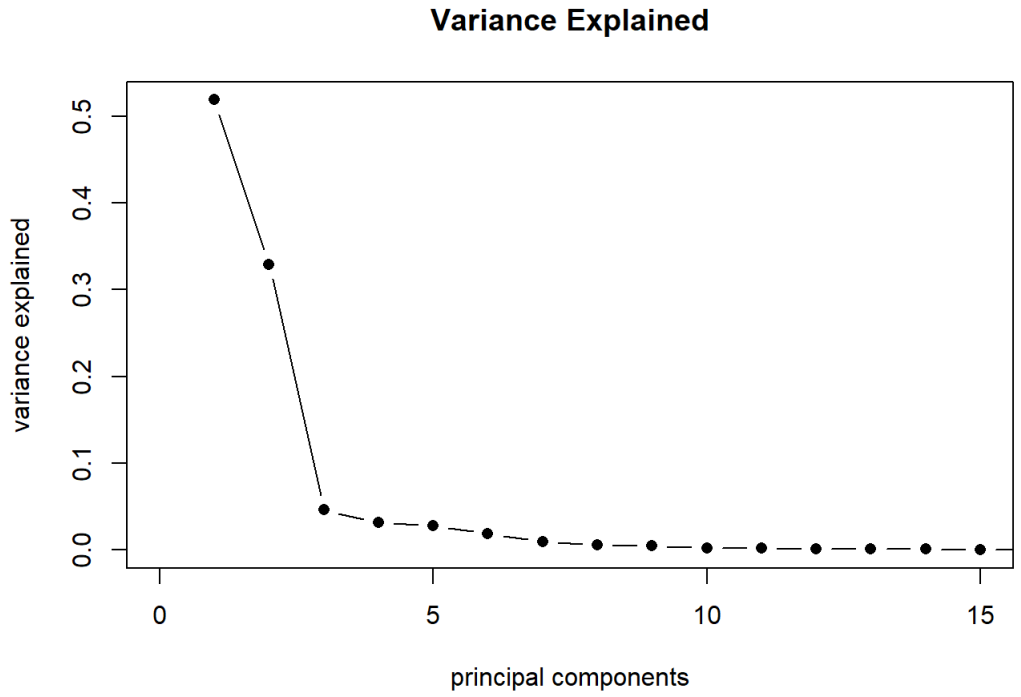

```
#PCA
# pc <- prcomp(data1)
# # head(pc$x[, 1:2])
# plot(pc$x[, 1], pc$x[, 2], xlab='PC1', ylab='PC2',col = viridis(20), main="PCA outcome")

#Var explain
var<-sv$d^2/sum(sv$d^2)
var
```

```
##  [1] 5.188720e-01 3.292979e-01 4.590491e-02 3.150160e-02 2.824223e-02
##  [6] 1.832626e-02 9.672657e-03 5.835047e-03 4.912880e-03 2.205250e-03
## [11] 1.793036e-03 1.340342e-03 8.724556e-04 6.488673e-04 4.086704e-04
## [16] 1.155381e-04 2.264914e-05 1.633681e-05 1.134046e-05 2.376569e-21
## [21] 2.442891e-32 2.779788e-33
```

```
var[1]+var[2]
```

```
## [1] 0.8481699
```

```
plot(sv$d^2/sum(sv$d^2), xlim = c(0, 15), type = "b", pch = 16, xlab = "principal components",
     ylab = "variance explained", main="Variance Explained")
```



According the variance analysis above, we got the sum of variance of the PCA1 and PCA2 is 0.8481699, which means that the two variables controled can explain about 85% of the total variance in the data, so having two variables controled is an ideal number of PCA. Furthermore, as PCA1 can explain more variance than PCA2, the scale that data distribute along the x axis is larger than that of PCA2 does.