Group09_report

Mohsin, Ali, Ahmed, Mubarak, Asad

```
covariance_matrix <- read.delim("p10-16.dat" ,sep = "" , header = FALSE)
covariance_matrix<- as.matrix(covariance_matrix)
n = 46
p = 3
q = 2

covariance_matrix
```

```
##            V1       V2       V3     V4      V5
## [1,] 1106.000  396.700  108.400  0.787  26.230
## [2,]  396.700 2382.000 1143.000 -0.214 -23.960
## [3,]  108.400 1143.000 2136.000  2.189 -20.840
## [4,]    0.787   -0.214    2.189  0.016   0.216
## [5,]   26.230  -23.960  -20.840  0.216  70.560
```

a)

```
rho_star1
```

```
## [1] 0.5173449
```

```
rho_star2
```

```
## [1] 0.1255082
```

```
test <- -(n - 1 - 0.5 * (p + q + 1)) * log((1 - rho_star1^2) * (1 - rho_star2
^2))
chi_sq <- qchisq(1 - alpha, df = 6)
test
```

```
## [1] 13.74948
```

```
chi_sq
```

```
## [1] 12.59159
```

Analysis

As 13.74 > 12.59, we can safely reject the null hypothesis.

b)

```
test2 <- -(n - 1 - 0.5 * (p + q + 1)) * log((1 - rho_star2^2))
chi_sq2 <- qchisq((1-alpha), df = 2)
test2
```

```
## [1] 0.6668632
```

```
chi_sq2
```

```
## [1] 5.991465
```

Analysis

Since the value retuned by t-test is less than the value returned by chi-squrae test i-e 0.667<5.991, we accept the hypothesis ρ_2^* equals 0.

This means that the first pair of canonical variables is significant & therefore all values of ρ^* after ρ_2^* have little significance`

c)

```
#rho1 value
rho_star1
```

```
## [1] 0.5173449
```

```
#square of rho1 value
rho_star1^2
```

```
## [1] 0.2676458
```

Analysis

$(\rho_1^*)^2 = 0.2676458$,

This explains the proportion of variance of canonical variate v1 that is explianed by the primary set of variables.

d)

```
paste("U1= ",round(u1[1],3)," z1 ",round(u1[2],3)," z2 +",round(u1[3],3)," z3
",sep="")
```

```
## [1] "U1= 0.436 z1 -0.705 z2 +1.081 z3 "
```

Analysis

On analyzing the canonical variate U1, we observe that z2(insulin response to oral glucose) and z3(insulin resistance)`have the highest values, indicating that z2 and z3 are the significant factors separating diabetic from non diabetic patients.

```
paste("V1= ",round(v1[1],3)," z1 +",round(v1[2],3)," z2",sep="")
```

```
## [1] "V1= -1.02 z1 +0.161 z2"
```

Analysis

Upon analyzing the second canonical variable, we see the term z1 dominates which is basically the relative weight of the person. This means that there is a clear difference in weights between people who have diabetes and those who don't.

Canonical variables correlation

```
correlation_u1

## [1]  0.3397282 -0.0501787  0.7551136

correlation_v1

## [1] -0.98750694 -0.04646446

correlation_u2

## [1] -0.6837882  0.4565378  0.5729495

correlation_v2

## [1] -0.1575755 -0.9989199
```

e)

```
prop_u1

##           [,1]
## [1,] 0.6186465

prop_v1

##           [,1]
## [1,] 0.5333782
```

Analysis

62% of the variance in U1 is illustrated by the first set of variables and 53% of the variance in V1 is illustrated by the secoend set of variables.

f)

Analysis

The analysis is only able to capture 50-60% of the variance which is good to understand the data but we would like the percentage to be a lot higher to fully grasp the key insights about the variance.

APPENDIX

```
####Assignment 4######

#read data
covariance_matrix <- read.delim("p10-16.dat" ,sep = "" , header = FALSE)
covariance_matrix<- as.matrix(covariance_matrix)

# part a
#Test at the 5% level if there is any association between the groups of varia
bles.
#Test for correlation between primary and secondary variable with alpha = 0.0
5
#covariance needs scaled and standardized
n = 46
p = 3
q = 2
alpha = 0.05

s11 <- covariance_matrix[1:3, 1:3]
s12 <- covariance_matrix[1:3, 4:5]
s21 <- covariance_matrix[4:5, 1:3]
s22 <- covariance_matrix[4:5, 4:5]


#compute matrix ^(-1/2)

inv_mat <- function(A){
  e <- eigen(A)
  eigen_vecs <- e$vectors
  eigen_vals <- e$values
  res <- matrix(0,nrow=ncol(eigen_vecs),ncol=ncol(eigen_vecs))
  for(i in 1:length(eigen_vals)){
    res <- res + 1/sqrt(eigen_vals[i]) * crossprod(t(eigen_vecs[,i]),t(eigen_
vecs[,i]))
  }
  return(res)
}

s11_sqrt_inv <- inv_mat(s11)
s22_sqrt_inv <- inv_mat(s22)

eigen_val1 <- eigen(s11_sqrt_inv %*% s12 %*% solve(s22) %*% s21 %*% s11_sqrt_
inv)[1]
eigen_val2 <- eigen(s22_sqrt_inv %*% s21 %*% solve(s11) %*% s12 %*% s22_sqrt_
inv)[1]

eigen_vec1 <- eigen(s11_sqrt_inv %*% s12 %*% solve(s22) %*% s21 %*% s11_sqrt_
inv)[2]
```

```r
eigen_vec2 <- eigen(s22_sqrt_inv %*% s21 %*% solve(s11) %*% s12 %*% s22_sqrt_
inv)[2]


rho_star1 <- sqrt(eigen_val1[[1]])[1]
rho_star2 <- sqrt(eigen_val2[[1]])[2]

#compute sample test

test <- -(n - 1 - 0.5 * (p + q + 1)) * log((1 - rho_star1^2) * (1 - rho_star2
^2))


#compute Chi square
chi_sq <- qchisq(1 - alpha, df = 6)

#Since 13.74 > 12.59 so the null hypothesis is rejected ()

#part b
#How many pairs of canonical variates are signifcant?
#test for second canonical correlation since there are only 2 variables
#df= (p-1)(q-1)

test2 <- -(n - 1 - 0.5 * (p + q + 1)) * log((1 - rho_star2^2))
chi_sq2 <- qchisq((1-alpha), df = 2)

#since 0.667 < 5.991, we accept hypothesis that p_star2 is 0, so the
#second pairs is not significant


#part c
#Interpret the "signifcant" squared canonical correlations.
#Tip: Read section "Canonical Correlations as Generalizations of Other Correl
ation Coeficients".

#The canonical correlation that was significant in the test was
#p_star1. p*1 = 0.517345, p*1^2 = 0.2676
#This value could be interpreted as the proportion
#of variance of canonical variate v1 that is explianed by the primary set of
variables.


#part d
#Interpret the canonical variates by using the coefficients and suitable corr
elations.
#find u and v (textbook pg 578)

#raw canonical coefficients for the glucose and insulin
```

```r
x11 <- t(as.matrix(eigen_vec1[[1]][,1])) %*% s11_sqrt_inv
x21 <- t(as.matrix(eigen_vec2[[1]][,1])) %*% s22_sqrt_inv
x21 <- t(as.matrix(eigen_vec1[[1]][,2])) %*% s11_sqrt_inv
x22 <- t(as.matrix(eigen_vec2[[1]][,2])) %*% s22_sqrt_inv


#standardized canonical coefficients for the glucose and insulin
corr_mat <- cov2cor(covariance_matrix)  #scaled correlation

r11 <- corr_mat[1:3, 1:3]
r12 <- corr_mat[1:3, 4:5]
r21 <- corr_mat[4:5, 1:3]
r22 <- corr_mat[4:5, 4:5]

r11_sqrt_inv <- inv_mat(r11)
r22_sqrt_inv <- inv_mat(r22)



evec_r <- eigen(r11_sqrt_inv %*% r12 %*% solve(r22) %*% r21 %*% r11_sqrt_inv)
[2]
evec2_r <- eigen(r22_sqrt_inv %*% r21 %*% solve(r11) %*% r12 %*% r22_sqrt_inv
)[2]

u1 <- t(as.matrix(evec_r[[1]][,1])) %*% r11_sqrt_inv
v1 <- t(as.matrix(evec2_r[[1]][,1])) %*% r22_sqrt_inv
u2 <- t(as.matrix(evec_r[[1]][,2])) %*% r11_sqrt_inv
v2 <- t(as.matrix(evec2_r[[1]][,2])) %*% r22_sqrt_inv


#correlation between the glucose and insulin and their canonical variables

correlation_u1 <- as.vector(u1 %*% r11)
correlation_v1 <- as.vector(v1 %*% r22)

correlation_u2 <- as.vector(u2 %*% r11)
correlation_v2 <- as.vector(v2 %*% r22)

#u1 equation
paste("U1= ",round(u1[1],3)," z1 ",round(u1[2],3)," z2 +",round(u1[3],3)," z3
",sep="")
#v1
paste("V1= ",round(v1[1],3)," z1 +",round(v1[2],3)," z2",sep="")


#part e
#total proportion explained
prop_u1 <- (u1) %*% t(u1)/p # U1
prop_v1 <- (v1) %*% t(v1)/q # U1
```

```
#part f

#conclusion above
```