Multivariate Lab 4

Ramon Laborda, Ugurcan Lacin, Milda Poceviciute and Henrik Karlsson

12/15/2017

Look at the data described in Exercise 10.16 of Johnson, Wichern. You may find it in the file P10-16.DAT. The data for 46 patients are summarized in a covariance matrix, which will be analyzed in R. Read through the description of the different R packages and functions so you may choose the must suitable one for the analysis. Supplement with own code where necessary.

```
n <- 46 #number of observations
#Covariance matrix
Mcov <- matrix(c(1106, 396.7, 108.4, 0.787, 26.23, 396.7, 2382, 1143, -0.214, -23.96,
                 108.4, 1143, 2136, 2.189, -20.84, 0.787, -0.214, 2.189, 0.016, 0.216,
                 26.23, -23.96, -20.84, 0.216, 70.56), nrow=5, ncol=5)
Mcov
```

```
##          [,1]     [,2]     [,3]   [,4]    [,5]
## [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
```

```
#Transform a covariance matrix into correlation matrix
Mcor <- cov2cor(Mcov)
Mcor
```

```
##            [,1]        [,2]        [,3]        [,4]        [,5]
## [1,] 1.00000000  0.24440707  0.07052630  0.18708423  0.09389478
## [2,] 0.24440707  1.00000000  0.50672780 -0.03466434 -0.05844356
## [3,] 0.07052630  0.50672780  1.00000000  0.37444251 -0.05368065
## [4,] 0.18708423 -0.03466434  0.37444251  1.00000000  0.20328928
## [5,] 0.09389478 -0.05844356 -0.05368065  0.20328928  1.00000000
```

```
##########################
s11 <- Mcov[1:3, 1:3]
s12 <- Mcov[1:3, 4:5]
s21 <- Mcov[4:5, 1:3]
s22 <- Mcov[4:5, 4:5]
r11 <- Mcor[1:3, 1:3]
r12 <- Mcor[1:3, 4:5]
r21 <- Mcor[4:5, 1:3]
r22 <- Mcor[4:5, 4:5]

p <- nrow(s12)
q <- ncol(s12)

##########################
expM <- function(X,e) {
  v <- La.svd(X); #decomposition of a matrix (d,u,vt)
  v$u %*% diag(v$d^e) %*% v$vt
  }
```

Part A

Test at the 5% level if there is any association between the groups of variables.

```
FirstEigVal <- eigen(expM(s11, -0.5) %*% s12 %*% solve(s22) %*% s21 %*% expM(s11, -0.5))[1]
Pstar1 <- sqrt(FirstEigVal[[1]])[1]
```

```
## Warning in sqrt(FirstEigVal[[1]]): NaNs produced
```

```
SecEigVal <- eigen(expM(s22, -0.5) %*% s21 %*% solve(s11)%*% s12 %*% expM(s22, -0.5))
Pstar2 <- sqrt(SecEigVal[[1]])[2]
tVal <- -(46-1-0.5*(3+2+1)) * log((1-Pstar1^2) * (1-Pstar2^2))
chi_sq_val <- qchisq(1-0.05, df=6)
#Avoid NaNs (there is no 3 EigenValues)
#Needs for the critic region
n
```

```
## [1] 46
```

```
Pstar1
```

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

```
Pstar2
```

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

```
tVal
```

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

```
chi_sq_val
```

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

We will test if there is any type of association between both group of variables:

- $X_1^{(1)}$: Glucose intolerance
- $X_2^{(1)}$: Insulin response to oral glucose
- $X_3^{(1)}$: Insulin resistance
- $X_1^{(2)}$: Relative weight
- $X_2^{(2)}$: Fasting plasma glucose

The null hypothesis is that all canonical correlation is zero with a significant association of 5%. The hypothesis will be reject if:

$$(n-1-0.5(p+q+1))ln\prod(1-(\hat{p_i^*}) > \chi^2_{pq}(\alpha)$$

$$(46-1-0.5(3+2+1))ln[(1-0.5173449^2)(1-0.1255082^2)] = 13.74948$$

$$\chi^2_6(0.05) = 12.59159$$

So $13.74948 > 12.59159$ we will reject the null hypothesis. It means that there is association between the first group of variables and the second.

Part B

How many pairs of canonical variates are significant?

```
tVal2 <- -(46-1-0.5*(3+2+1)) * log(1-Pstar2^2)
chi_sq_val2 <- qchisq(1-0.05, df=2)
tVal2
```

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

```
chi_sq_val2
```

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

We will use a similar test used in part a to test if the second canonical correlation is significantly separated from zero.

$$H_0 : P_2^* = 0$$
$$H_1 : P_2^* \neq 0$$

$$-(46 - 1 - 0.5(3 + 2 + 1))lna(1 - 0.1255082^2) = 0.6668632$$

And

$$\chi_2^2(0.05) = 5.991465$$

So the critical value is lower than the observed test $0.6668632 < 5.991465$, there is no statistical evidence to reject the null hypothesis, so only the first pair of canonical variates is significant.

Part C

Interpret the "significant" squared canonical correlations.

```
Pstar1
```

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

```
Pstar1^2
```

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

The significant squared canonical correlation using the test p_1^{*2} that is equal to $0.5173449^2 = 0.2676457$. The interpretation of this value is that around 26.8% of the variance of canonical variate U_1 is explained by the secondary set of variables. Reversely this value also could be interpreted as the proportion of variance of canonical variate V_1 that is explained by the primary set of variables.

Part D

Interpret the canonical variates by using the coefficients and suitable correlations.

```
FirstEigVec <- eigen(expM(r11, -0.5) %*% r12 %*% solve(r22) %*% r21 %*% expM(r11, -0.5))[2]
SecEigVec <- eigen(expM(r22, -0.5) %*% r21 %*% solve(r11) %*% r12 %*% expM(r22, -0.5))[2]
U1 <- t(as.matrix(FirstEigVec[[1]][,1])) %*% expM(r11, -0.5)
V1 <- t(as.matrix(SecEigVec[[1]][,1])) %*% expM(r22, -0.5)
corrU1 <- as.vector(U1 %*% r11)
```

```r
names(corrU1) <- c("Glucose", "Insulin", "Insulres")
CU1 <- t(data.frame(corrU1))
corrV1 <- as.vector(V1 %*% r22)
names(corrV1) <-c("Weight", "Fasting")
#Needs for the canonical variates
U1
```

```
##           [,1]       [,2]     [,3]
## [1,] 0.4356829 -0.7046696 1.081462
```

```r
V1
```

```
##           [,1]     [,2]
## [1,] -1.020224 0.160936
```

```r
corrU1
```

```
##    Glucose    Insulin   Insulres
##  0.3397282 -0.0501787  0.7551136
```

```r
corrV1
```

```
##      Weight     Fasting
## -0.98750694 -0.04646446
```

$$\hat{U_1} = 0.4356829_{z1}^{(1)} - 0.7046696_{z2}^{(1)} + 1.081462_{z3}^{(1)}$$

$$V_1 = -1.020224_{z1}^{(2)} + 0.160936_{z2}^{(2)}$$

Correlation between U_1 and the primary set: - Glucose intolerance: 0.3397282 - Insulin response: -0.0501787 - Insulin resistance: 0.7551136

Correlation between V_1 and the secondary set: - Relative weight: -0.98750694 - Fasting plasma glucose: -0.04646446

To see how influent are the respective variables we use the canonical variates. We can see that the two insulin variables dominates $\hat{U_1}$ and the variable weight is the one consisting $\hat{V_1}$. If a value is high for the correlation between the variables and their canonical variates that indicate that the variable is closely associated with the canonical variate.

For $\hat{U_1}$ insulin resistance has the strongest correlation followed by glucose intolerance (moderate strong correlation). The last variable, insulin response, is an influent variable for the coefficients in $\hat{U_1}$ but is not closely associated to the variate.

For $\hat{V_1}$ relative weight, is very closely associated to the canonical variate and the second variable fasting plasma glucose has a very weak correlation with the variate.

Part E

Are the "significant" canonical variates good summary measures of the respective data sets?

```
(U1)%*%t(U1)/p # U1
```

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

```
(V1)%*%t(V1)/q # V1
```

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

According to result we can say that, 62% of the variance in U_1 is explained by the first set of variables. For V_1, we can also say that 53% of the variance is explained by the second set of variables.

Part F

Give your opinion on the success of this canonical correlation analysis.

We can comment that analysis is slightly successful. The canonical variables make sense when interpreted although it only seize above 50% - 60%, which can be acceptable as good sign.