MVS Lab 4

Joshua Hudson, Carles Sans, Karolina Ziomek

18 December 2017

Question 1: Canonical correlation analysis by utilizing suitable software

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 chose the must suitable one for the analysis. Supplement with own code where necessary

```
#setwd("~/LIU/Semester3/P2/MVS/L4")

#data1 <- read.table(paste0(getwd(), "/P10-16.DAT"))
data1<- read.table("C:/Users/Carles/Desktop/MasterStatistics-MachineLearning/Master_subjects/Multivaria
colnames(data1) <- rownames(data1) <- c("gluc intol", "ins resp", "ins resi", "weight", "plasma gluc")

S <- as.matrix(data1)
```

First we built a function for performing CCA using a covariance matrix only.

```
mycca <- function(S, split) {
  #cca using a covariance matrix
  #"split"" defines the number of columns/row for X(1); the rest are for X(2)
  size <- dim(S)[1]
  p <- split
  q <- size - p
  S11 <- S[1:p, 1:p]
  S12 <- S[1:p, (p+1):size]
  S21 <- S[(p+1):size, 1:p]
  S22 <- S[(p+1):size, (p+1):size]
  #check p<=q condition, if not swap X(1) and X(2)
  if (p > q) {
    p <- q
    q <- size-p
    cov11 <- S11
    cov12 <- S12
    S11 <- S22
    S12 <- S21
    S21 <- cov12
    S22 <- cov11
  }
  #function for matrix powers
  "%^%" <- function(x, n)
    with(eigen(x), vectors %*% (values^n * t(vectors)))
  #end function

  mat1 <- (S11%^%(-0.5))%*%S12%*%(S22%^%(-1))%*%S21%*%(S11%^%(-0.5))
  mat2 <- (S22%^%(-0.5))%*%S21%*%(S11%^%(-1))%*%S12%*%(S22%^%(-0.5))
  #canonical correlations
  cancor <- sqrt(sort(eigen(mat1)$value, decreasing=TRUE))
  #coefficients
  A <- matrix(nrow = p, ncol = p)
  B <- matrix(nrow = p, ncol = q)
  for (k in 1:p) {
    A[k ,] <- t(eigen(mat1)$vector[, k])%*%(S11%^%(-0.5))
    B[k, ] <- t(eigen(mat2)$vector[, k])%*%(S22%^%(-0.5))
  }
  colnames(A) <- colnames(S11)
  colnames(B) <- colnames(S22)

  #data/canonicals correlations
  R_Ux1 <- A%*%S11%*%(diag(diag(S11))%^%(-0.5))
  colnames(R_Ux1) <- colnames(S11)
  R_Vx2 <- B%*%S22%*%(diag(diag(S22))%^%(-0.5))
  colnames(R_Vx2) <- colnames(S22)
  R_Ux2 <- A%*%S12%*%(diag(diag(S22))%^%(-0.5))
  colnames(R_Ux2) <- colnames(S22)
  R_Vx1 <- B%*%S21%*%(diag(diag(S11))%^%(-0.5))
  colnames(R_Vx1) <- colnames(S11)

  R <- list(Ux1 = R_Ux1, Vx2 = R_Vx2, Ux2 = R_Ux2, Vx1 = R_Vx1)
  #reassemble S
  newS <- rbind(cbind(S11, S12), cbind(S21, S22))
  return(list(cancor=cancor, coefsX1 = A, coefsX2 = B, R = R, covmat = newS))

}
```

(a) Test at the 5% level if there is any association between the groups of variables

Below are the canonical correlations found using our function, as well as the result of the 5% significance test on association between groups.

```
#run mycca
cca <- mycca(S, 3)
cc <- cca$cancor
S <- cca$covmat
print(cc)
```

```
## [1] 0.5173449 0.1255082
```

```
#a) 5% significance test on association between groups
n <- 46
p <- 2
q <- 3
teststat <- -(n-1-0.5*(p+q+1))*log((1-cc[1]^2)*(1-cc[2]^2))
chistat <- qchisq(0.95, p*q)
#test
if (teststat > chistat) {print("Reject H0, there is association")} else {print("Accept H0, there is no a
```

```
## [1] "Reject H0, there is association"
```

b) How many pairs of canonical variates are significant?

```
teststat2 <- -(n-1-0.5*(p+q+1))*log(1-cc[2]^2)
chistat2 <- qchisq(0.95, (p-1)*(q-1))
if (teststat2 > chistat2) {print("Reject H0, the 2nd canonicial variate pair is significant")} else {pr
```

```
## [1] "Accept H0, the 2nd canonical variate pair is not significant"
```

The 2nd pair is not significant so we will not consider it going forward.

c) Interpret the "significant" squared canonical correlations. Tip: Read section "Canonical Correlations as Generalizations of Other Correlation Coefficients".

The 1st squared canonical correlation 0.27 is both: the proportion of the canonical variate U_1's variance explained by the 3 primary variables (glucose intolerance, insulin response to oral glucose and insulin resistance) and the proportion of the canonical variate V_1's variance explained by the 2 secondary variables (relative weight and fasting plasma glucose). It is seen as a measure of overlap between the 2 sets of variables mentioned.

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

Below is the coefficient vector for the U_1 scores correlations with the original variable.

```
#score weights
print("Score weights")
```

```
## [1] "Score weights"
```

```
cca$coefsX1[1, ]
```

```
##      weight plasma gluc
## -8.06557508  0.01915905
```

```
#correlations
print("Correlation U1, X1")
```

```
## [1] "Correlation U1, X1"
```

```
cca$R$Ux1[1, ]
```

```
##      weight plasma gluc
## -0.98750694 -0.04646446
```

```
print("Correlation V1, X1")
```

```
## [1] "Correlation V1, X1"
```

```
cca$R$Vx1[1, ]
```

```
##      weight plasma gluc
##  0.51088173  0.02403815
```

We can see that the U_1 score is strongly weighted by relative weight and very lightly weighted by fasting plasma glucose. The correlation between relative weight and the V_1 is 0.51, meaning that relative weight is fairly correlated to the variables in set 2, the primary variables.

Below is the coefficient vector for the V_k scores and correlations with the original variables.

```
#score weights
print("Score weights")
```

```
## [1] "Score weights"
```

```
cca$coefsX2[1, ]
```

```
##  gluc intol    ins resp    ins resi 
##  0.01310065 -0.01443825  0.02339972
```

```
#correlations
print("Correlation U1, X2")
```

```
## [1] "Correlation U1, X2"
```

```
cca$R$Ux2[1, ]
```

```
## gluc intol   ins resp   ins resi 
## -0.1757567  0.0259597 -0.3906542
```

```
print("Correlation V1, X2")
```

```
## [1] "Correlation V1, X2"
```

```
cca$R$Vx2[1, ]
```

```
## gluc intol   ins resp   ins resi 
##  0.3397282 -0.0501787  0.7551136
```

We can see that the resulting scores will be very low when the data is plugged in. All 3 coefficients are roughly the same in size, insulin resistance slightly more influential than the 2 others. The correlation between U_1 and X2 the primary variables shows that Insulin resistance has a relatively high correlation with the secondary variables (relative weight and fasting plasma glucose), Glucose intolerance also to some degree but Insulin Resistance hardly any (correlation close to 0).

e) Are the "significant" canonical variates good summary measures of the respective data sets? Tip: Read section "Proportions of Explained Sample Variance".

```
propvar_X1 <- (sum(cca$R$Ux1[1, ]^2))/p
propvar_X2 <- (sum(cca$R$Vx2[1, ]^2))/q
```

U_1, the first canonical variate for the 2ndary variables (glucose intolerance, etc), explains 49% of this set's sample variance. V_1, the first canonical variate for the primary variables (relative weight, etc), explains 23% of this set's sample variance. Neither provides a very good summary of their respective sets, but U_1 does it better.

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

The CCA found that the “relative weight” variable from the secondary group was correlated with 0.51 to the primary variables. Also the “glucose intolerance” and “insulin resistance” variables from the primary set are correlated with -0.18 and -0.39 to the secondary set. However the proportion of variance showed that the representation of the sets by the CCA score is rather poor, so the results should be taken with a pinch of salt.