Assignment 4: Canonical correlation analysis

Group 12

Dávid Hrabovszki (davhr856), Laura Julia Melis (lauju103), Spyridon Dimitriadis (spydi472), Vasileia Kampouraki (vaska979)

15/12/2019

Question: Canonical correlation analysis by utilizing suit- able 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.

A.

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

To analyze if there is any association between the groups of variables we will test the following hypothesis:

$$H_0 : \Sigma_{12} = 0 \quad (\rho_1^* = \rho_2^* = 0)$$
$$H_1 : \Sigma_{12} \neq 0$$

We will reject H_0 at significance level $\alpha = 0.05$ if:

$$c = -\big(n - \frac{1}{2}(p+q+3)\big) \ln \big[\prod_{i=1}^{p} (1 - \hat{\rho_i^*}^2) \big] > \chi^2_{pq}(\alpha)$$

Given that we have $p = 3$ primary variables and $q = 2$ secondary variables, there are $k = min(p, q) = min(3, 2) = 2$ canonical correlations ρ_k^*.

After manually computing the necessary operations, we obtain that the values of the canonical correlations are: $\rho_1^* = 0.5173449$ and $\rho_2^* = 0.1255082$

Now, we can calculate the values of the teststatistic and the critical value:

The results are that c= 12.88 > 12.59 thus we reject the null hypothesis. We can conclude that not all canonical correlations are zero: there are associations between the groups of variables.

B.

b) How many pairs of canonical variates are significant?

Given that $\rho_k^* = \mathrm{Cor}(U_k, V_k)$, to analyze if the canonical variates are significant we will test if ρ_1^* is non-zero.

We will reject H_0 at significance level $\alpha = 0.05$ if:

$$c = -\big(n - \frac{1}{2}(p+q+1)\big) \ln \big[\prod_{i=k+1}^{p} (1 - \hat{\rho_i^*}^2) \big] \; \chi^2_{(p-k)(q-k)}(\alpha)$$

The results obtained are c= 13.39 > 5.99 so we reject the null hypothesis. This suggests that only the first pair of canonical variables is significant.

C.

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

The kth squared canonical correlation r…k2 is the proportion of the variance of canonical variate Uk "explained" by the set X122. It is also the proportion of the variance of canonical variate Vk "explained" by the set X112. Therefore, r…k2 is often called the shared variance between the two sets X112 and X122.

D.

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

E.

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

F.

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

Appendix

```r
knitr::opts_chunk$set(echo = TRUE, message = F, warning = F, error = F)
# NEEDED LIBRARIES
library(ggplot2)
library(tidyr)
library(gridExtra)
library(car)
library(heplots)
library(MASS)
library(fmsb)

RNGversion('3.5.1')
##### Question a #####
# Importing the data
S = read.table("P10-16.dat") # covariance matrix
S = as.matrix(S)
p <- 3
q <- 2

# S11: covariances in Set1 of p variables
S11 <- as.matrix(S[1:3,1:3])

# S22: covariances in Set2 of q variables
S22 <- as.matrix(S[4:5,4:5])

# S12 and S21: pxq covariances between the sets.
S12 <- as.matrix(S[1:3,4:5])
S21 <-  as.matrix(S[4:5,1:3])

# With: p=3,q=2
S11eig <- eigen(S11, symmetric=TRUE)
```

```r
S11sqrt <- S11eig$vectors %*% diag(1/sqrt(S11eig$values)) %*% t(S11eig$vectors)

S22eig <- eigen(S22, symmetric=TRUE)
S22sqrt <- S22eig$vectors %*% diag(1/sqrt(S22eig$values)) %*% t(S22eig$vectors)

Xmat <- S11sqrt %*% S12 %*% solve(S22) %*% S21 %*% S11sqrt
Ymat <- S22sqrt %*% S21 %*% solve(S11) %*% S12 %*% S22sqrt
Xeig <- eigen(Xmat, symmetric=TRUE)
Yeig <- eigen(Ymat, symmetric=TRUE)

#The two canonical correlations
r <- sqrt(Yeig$values)

# Computing the statistic and the critical value:
c <- -42*log(1-r[1]^2)*(1-r[2]^2)
chisq <- qchisq(0.95,df=6)
##### Question b #####
newc <- -43*log(1-r[1]^2)
newchisq <- qchisq(0.95,df=2)
```