Assignment 4 - Canonical Correlation Analysis

GROUP 03
*Agustin Valencia - **aguva779***
*Bayu Brahmantio - **baybr878***
*Joris van Doorn - **jorva845***
*Marcos F Mourao - **marfr825***

13 December 2019

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.

The given matrix is the following:

V1	V2	V3	V4	V5
1106.000	396.700	108.400	0.787	26.230
396.700	2382.000	1143.000	-0.214	-23.960
108.400	1143.000	2136.000	2.189	-20.840
0.787	-0.214	2.189	0.016	0.216
26.230	-23.960	-20.840	0.216	70.560

Thus, separating the variance-covariance matrix it is obtained that

$$\Sigma_{11} =$$

V1	V2	V3
1106.0	396.7	108.4
396.7	2382.0	1143.0
108.4	1143.0	2136.0

$$\Sigma_{22} =$$

	V4	V5
4	0.016	0.216
5	0.216	70.560

$$\Sigma_{21} =$$

	V1	V2	V3
4	0.787	-0.214	2.189
5	26.230	-23.960	-20.840

$$\Sigma_{12} =$$

V4	V5
0.787	26.23
-0.214	-23.96
2.189	-20.84

It can be computed that

$$M = S_{11}^{-1/2} S_{12} S_{22}^{-1} S_{21} S_{11}^{-1/2} =$$

0.0468068	-0.0374259	0.0813854
-0.0374259	0.0306007	-0.0716743
0.0813854	-0.0716743	0.2059907

Its eigen-decomposition is given by

The first two eigenvalues $\overrightarrow{\alpha}$:

x
0.2676458
0.0157523

The first two eigenvectors $\overrightarrow{e}$:

0.3749438	0.7634104
-0.3220478	-0.4247428
0.8693114	-0.4866191

Also,

$$D = S_{22}^{-1/2} S_{21} S_{11}^{-1} S_{12} S_{22}^{-1/2} =$$

0.2671702	0.0109346
0.0109346	0.0162279

Its eigen-decomposition is given by

Eigenvalues $\overrightarrow{\beta}$:

x
0.2676458
0.0157523

Eigenvectors $\overrightarrow{f}$:

-0.9990556	0.0434508
-0.0434508	-0.9990556

It is defined that

$$\hat{\mathbf{a}}_k' = \hat{\mathbf{e}}_k' S_{11}^{-1/2}$$
$$\hat{\mathbf{b}}_k' = \hat{\mathbf{f}}_k' S_{22}^{-1/2}$$

Thus, based in results above:

$$\hat{\mathbf{a}} =$$

0.0131007	0.0247525
-0.0144383	-0.0093175
0.0233997	-0.0086672

$$\hat{\mathbf{b}} =$$

-8.0655751	0.3751678
0.0191591	-0.1200675

M and D have the same eigenvalues, so the canonical correlations can be calculated from one of

$$[\hat{\rho_1^*}, \hat{\rho_2^*}] =$$

0.5173449	0.1255082

For Hypothesis testing, using $\alpha = 0.05, p = 3, q = 2$ we have that, the obtained critical value is given by

$$\chi_{(1-\alpha,pq)} = 12.59159$$

$$H_0 : \Sigma_{12} = 0$$

The test statistic:

$$-\left(n - 1 - \frac{1}{2}(p+q+1)\right) \ln \prod_{i=1}^{2} (1 - \hat{\rho_i^{*2}})$$

For the obtained canonical correlations it is obtained that

$$test_1 = 13.74948$$
$$test_2 = 0.6668632$$

It is seen that $test_i > 12.59159$. Thus H_0 is rejected.

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

$$H_0 : \Sigma_{12} = 0$$
$$H1 : \Sigma 12 \neq 0$$

We have $p \cdot q = 3 * 2 = 6$ degrees of freedom. For $n = 46$, we use Bartletts approximation and our test statistic is: (fancy formula latechhh here).

b) **How many pairs of canonical variates are significant?**

$$H_0 : \rho_2 \neq 0$$

and

$$H_1 : \rho_2 = 0$$

At $\alpha = 0.05$ significance, we get a test statistic of 0.66, smaller than the critical value of 12.59. We then reject H_0, meaning only the first canonical correlation (ρ_1) is significantly different from 0.

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

The significant squared canonical correlation (ρ_1^2) measures the overlap between $X^{(1)}$ and $X^{(2)}$, the different group of variables.

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

The canonical correlations variables are:

$$U_1 = 0.0131007X_1^{(1)} - 0.0144383X_2^{(1)} + 0.0233997X_3^{(1)}$$

$$V_1 = -8.0655751X_1^{(2)} + 0.0191591X_2^{(2)}$$

We can interpret U_1 as a measure of how well a patient can process sugar (glucose) in their bloodstream. V_1, on the other hand, can be interpreted as how prone the patient is to diabetes.

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

```r
num <- t(a[,1]) %*% sigma12 %*% b[1,]
den <- sqrt(t(a[,1]) %*% sigma11 %*% a[,1]) * sqrt(t(b[,1]) %*% sigma22 %*% b[,1])
r <- num/den
```

The correlation variate explains 44% of the total sample variance. We conclude therefore that the significant canonical variates are not a good summary because there might be nonlinear relations.

The significant canonical variate is *not* a good summary of the dataset, because it only captures approximately 10% of the total sample variance.

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

While the canonical correlation variables have insightful interpretration, the analysis is not good enough given the small summary capability of its variates.

Appendix A - Code

```r
RNGversion('3.5.1')
knitr::opts_chunk$set(echo = TRUE)
library(expm)
library(knitr)
library(CCP)
data <- read.table("./Data/P10-16.DAT")
#number of observations (patients)
n <- 46
#number of primary variables
p <- 3
#number of secondary variables
q <- 2
kable(data)
#separating the variance-covariance matrix
sigma11 <- as.matrix(data[1:3,1:3])
sigma22 <- as.matrix(data[4:5, 4:5])
sigma21 <- as.matrix(data[4:5, 1:3])
sigma12 <- as.matrix(data[1:3, 4:5])

M <- sqrtm(solve(sigma11)) %*% sigma12 %*% solve(sigma22) %*% sigma21 %*% sqrtm(solve(sigma11))
D <- sqrtm(solve(sigma22)) %*% sigma21 %*% solve(sigma11) %*% sigma12 %*% sqrtm(solve(sigma22))

kable(sigma11)
kable(sigma22)
kable(sigma21)
kable(sigma12)
kable(M)
mEigenDecom <- eigen(M)
mEigVect <- mEigenDecom$vectors[,1:2]
mEigVals <- mEigenDecom$values[1:2]
kable(mEigVals)
kable(mEigVect)
kable(D)
dEigenDecom <- eigen(D)
dEigVect <- dEigenDecom$vectors
dEigVals <- dEigenDecom$values
kable(dEigVals)
kable(dEigVect)
a <- t(mEigVect) %*% sqrtm(solve(sigma11))
a <- t(a)
b <- t(dEigVect) %*% sqrtm(solve(sigma22))
b <- t(b)
kable(a)
kable(b)
rho1 <- sqrt(mEigVals)
rho2 <- sqrt(dEigVals)
kable(t(rho1))
test1 <- -(n-1-(0.5*(p+q+1))) * log(prod(1-rho1^2))
#Hypothesis testing
alpha <- 0.05
#critical value
```

```
crit <- qchisq(p = (1-alpha), df = p*q)
#test statistic
test1 <- -(n-1-(0.5*(p+q+1))) * log(prod(1-rho1^2))  #13.74948
test2 <- -(n-1-(0.5*(p+q+1))) * log(1-rho1[2]^2)  #13.74948 they are the same
num <- t(a[,1]) %*% sigma12 %*% b[1,]
den <- sqrt(t(a[,1]) %*% sigma11 %*% a[,1]) * sqrt(t(b[,1]) %*% sigma22 %*% b[,1])
r <- num/den
```