# Assignment 4: Canonical correlation analysis

Group 12

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

*14/12/2019*

## Question: Canonical correlation analysis

$(\rho_k^{*2}, e_k)$ and $(\rho_k^{*2}, f_k)$ are the eigenvalue-eigenvectors pairs of the matrices:

$\Sigma_{11}^{-1/2}\Sigma_{12}\Sigma_{22}^{-1/2}\Sigma_{21}\Sigma_{11}^{-1/2}$ and $\Sigma_{22}^{-1/2}\Sigma_{21}\Sigma_{11}^{-1/2}\Sigma_{12}\Sigma_{22}^{-1/2}$ respectively.

The kth pair of canonical variables is given by $U_k = e_k^T\Sigma_{11}^{-1/2}X^{(1)}$ and $V_k = f_k^T\Sigma_{22}^{-1/2}X^{(2)}$.

- The first two canonical correlations are: $\rho_1^* = 0.51734$ and $\rho_2^* = 0.12551$. (There are only two canonical correlations, k=min(p,q)=min(3,2)=2.)

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

Test for the significance of the canonical relations with $\alpha = 0.05$ Observed test statistic (Barlett correction).

$H_0 : \Sigma_{12} = 0$,
$H_1 : \Sigma_{12} \neq 0$

$-(n - 1 - \frac{1}{2}(p + q + 1)) \ln \prod_{i=1}^{k}(1 - \hat{\rho}_i^{*2}) > \chi_{pq}^2(\alpha)$

```
## 13.74948 > 12.59159
```

where k is the number of canonical correlations we take into consideration

We reject the $H_0$ with significance level $\alpha = 0.05$. There is some association between the groups of variables because not all canonical correlations are zero. Therefore the use of canonical correlation analysis is justified.

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

$H_0 : \rho_1^* \neq 0$, $\hat{\rho}_2^* = 0$

$-(n - 1 - \frac{1}{2}(p + q + 1)) \ln \prod_{i=2}^{k}(1 - \hat{\rho}_i^{*2}) > \chi_{(p-1)(q-1)}^2(\alpha)$

```
## 0.6668632 < 5.991465
```

we cannot reject the $H_0$ with significance level $\alpha = 0.05$.

From the above tests we get that only the first pair of canonical variables is significant.

## c) Interpret the "significant" squared canonical correlations.

The canonical correlation $\rho_k^*$ generalizes the correlation between two sets of variables, $X^{(1)}$ and $X^{(2)}$.

The canonical correlations are also the multiple correlation coefficients of $U_k$ with $X^{(2)}$ or the multiple correlation coefficients of $V_k$ with $X^{(1)}$. Because of its multiple correlation coefficient interpretation, the kth squared canonical correlation $\rho_k^{*2}$ is the proportion of the variance of canonical variate $U_k$ 'explained' by the set $X^{(2)}$. It is also the proportion of the variance of canonical variate $V_k$ 'explained' by the set $X^{(1)}$. Therefore, $\rho_k^{*2}$ is often called the shared variance between the two sets $X^{(1)}$ and $X^{(2)}$. The largest value, $\rho_1^{*2}$, is sometimes regarded as a measure of set 'overlap'. (book p548)

The 26.76% is the percentage of the variance of canonical variate $U_1$ and $V_1$ 'explained' by the set $X^{(2)}$ and $X^{(1)}$ respectively.

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

- Raw Canonical Coefficients for $X^{(1)}$

```
## glucose intolerance    insulin response  insulin resistance
##             0.01310            -0.01444             0.02340
```

Using the raw canonical variables it is not clear to interprate the outcome, thus we will use the standardized canonical variables.

- Standardized Canonical Coefficients for $X^{(1)}$

```
## glucose intolerance    insulin response  insulin resistance
##             0.43568            -0.70467             1.08146
```

- Standardized Canonical Coefficients for $X^{(2)}$

```
##        relative weight fasting plasma glucose
##               -1.02022                0.16094
```

- Correlation between $X^{(1)}$ and their Canonical Variable $U_1$

```
## glucose intolerance    insulin response  insulin resistance
##             0.33973            -0.05018             0.75511
```

- Correlation between $X^{(2)}$ and their Canonical Variable $V_1$

```
##        relative weight fasting plasma glucose
##               -0.98751               -0.04646
```

The first pair of canonical variables, or first canonical variate pair, is the pair of linear combinations $U_1, V_1$ which maximize the correlation.

According to the coefficients of $U_1$ the variables that contribute the most are $X_3^{(1)}$ and $X_2^{(1)}$, i.e. $U_1$ is primarily a insuline response to oral glucose and an insuline resistance variable. The indicators of medical problems are the primary variables $X^{(1)}$. While $V_1$ represents the relative weight variable. From the correlation between $X^{(1)}$ and their Canonical Variable $U_1$ we notice that insuline resistance is the most correlated variable with $U_1$. So we can interpret it as an indicator of insuline resistance problems. On the other hand, from the correlation between $X^{(2)}$ and their Canonical Variable $V_1$ we notice that is most correlated with the relative weight variable.

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

The proportions of total (standardized) sample variances 'explained' by the first canonical variates $U_1$ and $V_1$.

$R^2_{z^{(1)}|\hat{U}_1} = \frac{1}{3}\Sigma_{k=1}^{3} r^2_{\hat{U}_1, z_k^{(1)}}$

```
## [1] 0.2293766
```

$R^2_{z^{(2)}|\hat{V}_1} = \frac{1}{2}\Sigma_{k=1}^{2} r^2_{\hat{V}_1, z_k^{(1)}}$

```
## [1] 0.4886645
```

The first sample canonical variate of the $X^{(1)}$ set accounts for 22.93% of the set's total sample variance. The first sample canonical variate of the $X^{(2)}$ set explains 48.86% of the set's total sample variance. We might thus infer that $V_1$ is a 'better' representative of its set than $U_1$ is of its set.

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

```
## [1] 0.26765
```

The proportion of the variance of canonical variate $U_1$, $V_1$ "explained" by the set $X^{(2)}$ and $X^{(1)}$ respectively is 26.76%.

This canonical correlation analysis was not particularly successful.

# Appendix

```
knitr::opts_chunk$set(echo = TRUE, message = F, warning = F, error = F)
S = as.matrix(read.table("P10-16.dat"))
n = 46
p = 2
q =3
S11 = as.matrix(S[1:3, 1:3])
S12 = as.matrix(S[1:3, 4:5])
S21 = t(S12)
S22 = as.matrix(S[4:5, 4:5])
# sample canonical variets
eig11 <- eigen(S11)
S11inv <- solve(S11)
S11invsqtr <- eig11$vectors %*% diag(sqrt(eig11$values)^(-1)) %*% t(eig11$vectors)

eig22 <- eigen(S22)
S22inv <- solve(S22)
S22invsqtr <- eig22$vectors %*% diag(sqrt(eig22$values)^(-1)) %*% t(eig22$vectors)
# NOTE: invsqtr=inverse square root

# the canonical correlations
rho <- sqrt(eigen(S11invsqtr %*% S12 %*% S22inv %*% S21 %*% S11invsqtr)$values)


# rho
# columns contain the eigenvectors
ei <- eigen(S11invsqtr %*% S12 %*% S22inv %*% S21 %*% S11invsqtr)$vectors
fi <- eigen(S22invsqtr %*% S21 %*% S11inv %*% S12 %*% S22invsqtr)$vectors

sdS11 = sqrt(diag(S11))
sdS22 = sqrt(diag(S22))

# sample canonical varietes
rawU = t(ei) %*% S11invsqtr
rawV = t(fi) %*% S22invsqtr
U = t(ei) %*% S11invsqtr %*% diag(sdS11)  # *sd to standardized
V = t(fi) %*% S22invsqtr %*% diag(sdS22)  # p542

# the first pair of canonical variables
# U[1,]
# V[1,]
# p 566
# Test between the first and second
cat(-(n-1-(1/2)*(p+q+1))*log((1-rho[1]^2)*(1-rho[2]^2)),">",
    qchisq(0.05,df=p*q,lower.tail = F))
p1=p-1
q1=q-1
# for the secondary
cat(-(n-1-(1/2)*(p+q+1))*log((1-rho[2]^2)),"<",
qchisq(0.05,df=p1*q1,lower.tail = F))
# the proportion of the variance of canonical variates
# round(rho[1]^2,5)
```

```r
# the first raw canonical variable
rawwU = round(t(rawU)[,1:2], 5)
rownames(rawwU) = c("glucose intolerance", "insulin response", "insulin resistance")
colnames(rawwU) = c("rawU1","U2")
rawwU[,1]


# the first pair of canonical variables
stdU = round(t(U[1:2,]), 5)
rownames(stdU) = c("glucose intolerance", "insulin response", "insulin resistance")
colnames(stdU) = c("U1","U2")
stdU[,1]


stdV = round(t(V),5)  # p544,546,547
rownames(stdV) = c("relative weight", "fasting plasma glucose")
colnames(stdV) = c("V1", "V2")
stdV[,1]
corU = t(rawU %*% S11 %*% solve(diag(sdS11)))[,1:2]
rownames(corU) = c("glucose intolerance", "insulin response", "insulin resistance")
colnames(corU) = c("U1", "U2")

round(corU[,1],5)
corV = t(rawV %*% S22 %*% solve(diag(sdS22)))
rownames(corV) = c("relative weight", "fasting plasma glucose")
colnames(corV) = c("V1", "V2")

round(corV[,1],5)
mean(corU[,1]^2)
mean(corV[,1]^2)
# the proportion of the variance of canonical variates
round(rho[1]^2,5)
```