732A97 - Multivariate statistical methods

Lab 4 - "Canonical correlation analysis"

Alexander Karlsson (aleka769)
Ruben Munoz (rubmu773)
Mariano Mariani (marma330)
Raymond Sseguya (rayss753)

2019-12-15

Contents

Canonical correlation analysis **3**
a) Test association between levels 3
b) Significant canonical variate pairs 5
c) Interpretation of squared canonical correlations 7
d) Interpretation of canonical variates using suitable method 8
e) Does significant canonical variates give good summary of data? 10
f) Comment on the success of the analysis 11

Packages used can be seen below.

```
library(dplyr)
library(ggplot2)
library(CCA)
library(expm)
theme_set(theme_bw())
```

Data used for this lab:

```
fullNames = c("Glucose Intolerance (primary)",
              "Insulin Response to Oral Glucose (primary)",
              "Insulin Resistance (primary)",
              "Relative Weight (secondary)",
              "Fasting Plasma Glucose (secondary)")

patientData = read.table(file = "P10-16.DAT",
                         col.names = c("p_GI",
                                       "p_IRtOG",
                                       "p_IR",
                                       "s_RW",
                                       "s_FPG"))
```

Canonical correlation analysis

The loaded data is summarized in a covariance matrix:

$$S = \begin{bmatrix} S_{11} & S_{12} & S_{13} & S_{14} & S_{15} \\ S_{21} & S_{22} & S_{23} & S_{24} & S_{25} \\ S_{31} & S_{32} & S_{33} & S_{34} & S_{35} \\ S_{41} & S_{42} & S_{43} & S_{44} & S_{45} \\ S_{51} & S_{52} & S_{53} & S_{54} & S_{55} \end{bmatrix} = \begin{bmatrix} \mathbf{S_{11}} & \mathbf{S_{12}} \\ \mathbf{S_{21}} & \mathbf{S_{22}} \end{bmatrix}$$

where the coloring of the matrix entries represent block. Number of points in the original data that was used to create this matrix is $n = 46$ (given in description). **The aim of canonical correlation analysis is to...**

```
S = as.matrix(patientData)
rownames(S) = colnames(S)

n = 46    # Number of data points used to calculate the covariance (block) matrix
p = 3     # Nr of primary variables
q = 2     # Nr of secondary variables
a = 0.05  # Significance level for tests
```

a) Test association between levels

The first thing that should be tested when canonical correlation analysis is considered is the association between levels:

$$H_0 : \Sigma_{12} = \Sigma_{X^{(1)},X^{(2)}} = 0$$
$$H_a : \Sigma_{12} = \Sigma_{X^{(1)},X^{(2)}} \neq 0$$

In case the null hypothesis cannot be rejected, there is no point in performing canonical correlation analysis, because no linear combinations $\mathbf{a}'X^{(1)}$ and $\mathbf{b}'X^{(2)}$ will yield a canonical correlation betwen levels that is separated from 0. The test is conducted with a likelihood ratio, where the determinants from the two levels block-covariances are used. The sampling distribution below follows a $\chi^2_{[p\cdot q]}(\alpha)$-distribution:

$$\begin{aligned} \chi^2_{test} &= [correction]\left[\ln\left(\frac{Generalized\ variance\ under\ H_0}{Unrestricted\ generalized\ variance}\right)\right] \\ &= [n]\left[\ln\left(\frac{|\mathbf{S}_{11}||\mathbf{S}_{22}|}{|\mathbf{S}|}\right)\right] \end{aligned}$$

Generalized variance under H_0 means that the blue entries defined in the matrix (full sample covariance matrix above) is set to 0. If the null hypothesis is rejected ($\chi^2_{test} \geq \chi^2_{crit}$), there is an association between levels.

```
S11 = S[1:3, 1:3]
S22 = S[4:5, 4:5]

chisq_df = p*q
chisq_crit = qchisq(1-a, chisq_df)

Likelihood_ratio = (det(S11) * det(S22)) / det(S)

chisq_test = n * log(Likelihood_ratio)
```

```r
data.frame(chisq_crit,
           chisq_test,
           ifelse(test = chisq_test > chisq_crit,
                  yes  = "Reject Null",
                  no   = "Fail to reject Null")) %>%
  knitr::kable(. ,align = "c",
               col.names = c("Critical value","Test statistic","Conclusion"))
```

Critical value	Test statistic	Conclusion
12.59159	15.05896	Reject Null

On 5 % significance, we reject the null hypothesis. Thus, we conclude that there is at least one significant canonical variate between levels $X^{(1)}$: (`Glucose Intolerance (primary)`, `Insulin Response to Oral Glucose (primary)`, `Insulin Resistance (primary)`) and $X^{(2)}$: ((`Relative Weight (secondary)`, `Fasting Plasma Glucose (secondary)`)).

b) Significant canonical variate pairs

We proceed by testing individual canonical correlations $\hat{\rho}_i^*$ with the following hypotheses:

$$\begin{aligned} H_0: &\quad \hat{\rho}_1^* \neq 0, \quad ..., \quad \hat{\rho}_k^* \neq 0, \quad \hat{\rho}_{k+1}^* = 0, \quad \hat{\rho}_p^* = 0 \\ H_a: &\quad \hat{\rho}_i^* = 0 \quad (\text{for all } i \geq k+1) \end{aligned}$$

Here, we are testing if there are more than one significant canonical correlation. We do so by setting $k = 1$ initially, and check the result. For these tests, Barlett's correction is used to improve the χ^2-approximation of the sampling distribution:

$$\textit{Barlett's correction:} \quad \big[\textit{correction}\big] = \big[-(n - 1 - \frac{1}{2}(p + q + 1))\big]$$

And we compute the test statistic according to:

$$\begin{aligned} \chi^2_{test} &= \big[\textit{correction}\big]\Big[ln\Big(\textit{Likelihood ratio}\Big)\Big] \\ &= \big[-(n - 1 - \frac{1}{2}(p + q + 1))\big]\Big[ln\Big(\prod_{i=1}^{p}\Big(1 - (\hat{\rho}^*)^2\Big)\Big)\Big] \end{aligned}$$

This statistic is compared to $\chi^2_{(p-k)\cdot(q-k)}(\alpha)$, where we reject H_0 if $\chi^2_{test} \geq \chi^2_{crit}$. The individual squared canonical correlations $(\hat{\rho}_i^*)^2$ are the eigenvalues of the eigendecomposition of the matrix multiplication below:

$$\begin{aligned} &\mathbf{C} = \mathbf{S}_{11}^{-1}\mathbf{S}_{12}\mathbf{S}_{22}^{-1}\mathbf{S}_{21} \\ &Eigen(\mathbf{C}) \longrightarrow (\vec{e}_i, \lambda_i) \\ &(\hat{\rho}_i^*)^2 = \lambda_i \end{aligned}$$

Now, we just compute everything that's mentioned above and check the test result:

```
# Define k:
k = 1

# Define block matrices:
S12 = S[1:3, 4:5]
S21 = t(S12)

# Compute eigen decomposition to get canonical correlations:
C   = solve(S11) %*% S12 %*% solve(S22) %*% S21
E   = eigen(C)
rho = E$values  # These are the squared canonical correlations

# Compute critical value:
chisq_df   = (p - k) * (q - k)
chisq_crit = qchisq(1 - a, chisq_df)

# Compute correction:
Barlett_corr = -(n - 1 - (p + q + 1)/2)

# Compute (for k = 1) test statistic:
Likelihood_ratio = log(prod(1 - rho[(k+1):p]))
chisq_test       = Barlett_corr * Likelihood_ratio
```

```r
# Test result to output:
data.frame(chisq_crit,
           chisq_test,
           ifelse(test = chisq_test > chisq_crit,
                  yes  = "Reject Null",
                  no   = "Fail to reject Null")) %>%
  knitr::kable(., align = "c",
               col.names = c("Critical value",
                             "Test statistic",
                             "Conclusion"))
```

Critical value	Test statistic	Conclusion
5.991465	0.6668632	Fail to reject Null

As can be seen above, we fail to reject the Null hypothesis on 5 % significance level. In other words, we have no reason to believe we have any more significant canonical variates.

c) Interpretation of squared canonical correlations

With only one canonical variate significant, only the first one will be interpreted. From the course book, we have the following result:

$$\rho_k^* = corr(U_k, V_k) = \max_a \ \text{corr}(\mathbf{a}'\mathbf{X}^{(1)}, V_k) = \max_b \ \text{corr}(U_k, \mathbf{b}'\mathbf{X}^{(2)})$$

for $k = 1, ..., p$. The squared canonical correlations $\hat{\rho}_k^* = \max_a \ \text{corr}(\mathbf{a}'\mathbf{X}^{(1)}, V_k)$ is the proportion of variance the canonical variate U_k is explained by the set (secondary variables) $\mathbf{X}^{(2)}$. In the formula above, we can interpret "explanation" in two "directions"; how canonical variate U_k is explained by $\mathbf{b}'\mathbf{X}^{(2)}$ and how V_k is explained by $\mathbf{a}'\mathbf{X}^{(1)}$. The proportion will be the same in both directions.

```
(rho %>% round(5)) %>% matrix(., 1) %>%
  knitr::kable(., col.names = c("Rho 1","Rho 2","Rho 3"), align = 'c')
```

Rho 1	Rho 2	Rho 3
0.26765	0.01575	0

We see that $Corr(U_1, V_1) \approx 26.8$. 26.8 % of variance in the canonical variate U_1 is explained by the other canonical variate V_1 and vice versa.

d) Interpretation of canonical variates using suitable method

The canonical variate can also be written in form of standardized variables:

$$U_1 = \vec{a}_1 Z^{(1)}$$
$$V_1 = \vec{b}_1 Z^{(2)}$$

We don't have access to the data $Z^{(1)}$ or $Z^{(2)}$ directly, so the expression written as above will do. However, we can see for which variables the canonical correlation vectors have high scalars, and thus see which variables are important.

Now we need to compute the eigendecomposition of $\mathbf{R}_{11}^{-1}\mathbf{R}_{12}\mathbf{R}_{22}^{-1}\mathbf{R}_{21}$ and $\mathbf{R}_{22}^{-1}\mathbf{R}_{21}\mathbf{R}_{11}^{-1}\mathbf{R}_{12}$ to get $\vec{a}_1$ and $\vec{b}_1$ respectively. These calculations are done in accordance with p.544 in the course book. Note that we use the function `sqrtm()` from the `expm` package to compute the matrix $\mathbf{R}_{11}^{-1/2}$.

```
R <- cov2cor(S)
R11<-R[1:3,1:3]
R12<-R[1:3,4:5]
R21<-R[4:5,1:3]
R22<-R[4:5,4:5]

C1 <- solve(R11) %*% R12 %*% solve(R22) %*% R21
C2 <- solve(R22) %*% R21 %*% solve(R11) %*% R12

E1 <- eigen(C1)
E2 <- eigen(C2)

e1 = E1$vectors[,1]
e2 = E2$vectors[,1]

# Following calculations from p.544 in course book:
a1 = sqrtm(solve(R11)) %*% e1
b1 = solve(R22) %*% R21 %*% a1
b1_corr = sqrt(t(b1) %*% R22 %*% b1)
b1 = (1 / b1_corr[1,1]) * b1

# print for equations below
as.vector(t(a1)); as.vector(t(b1))
```

```
## [1]  0.4106277 -0.8821876  1.0544531
```

```
## [1]  1.0190941 -0.1410883
```

$$U_1 = 0.4106277 \cdot Z_1^{(1)} - 0.8821876 \cdot Z_2^{(1)} + 1.0544531 \cdot Z_3^{(1)}$$
$$V_1 = 1.0190941 \cdot Z_1^{(2)} - 0.1410883 \cdot Z_2^{(2)}$$

We observe that the variable $Z_3^{(1)}$ (`Insulin Resistance (primary)`) is most important for the canonical variate U_1 and that the variable $Z_1^{(2)}$ (`Relative Weight (secondary)`) is most important for the canonical variate V_1. Below, we see the correlations between canonical covariates:

```
data.frame(t(a1) %*% R11 %>% as.vector(),
           t(b1) %*% R21 %>% as.vector(),
           fullNames[1:3]) %>%
  knitr::kable(., col.names = c("Cor(U1, Primary)", "Cor(V1, Primary)", "Variable"))
```

Cor(U1, Primary)	Cor(V1, Primary)	Variable
0.2693815	0.1774090	Glucose Intolerance (primary)
-0.2475066	-0.0270805	Insulin Response to Oral Glucose (primary)
0.6363841	0.3891659	Insulin Resistance (primary)

```
data.frame(t(a1) %*% R12 %>% as.vector(),
           t(b1) %*% R22 %>% as.vector(),
           fullNames[4:5]) %>%
  knitr::kable(., col.names = c("Cor(U1, Secondary)", "Cor(V1, Secondary)", "Variable"))
```

Cor(U1, Secondary)	Cor(V1, Secondary)	Variable
0.5022345	0.9904124	Relative Weight (secondary)
0.0335103	0.0660826	Fasting Plasma Glucose (secondary)

The canonical covariates U_1 and V_1 seem to have correlations that follow each other for all variables, both primary and secondary. But U_1 have higher correlations for the primary variables (which is logical), and V_2 have higher correlations for secondary variables (also logical).

Furthermore, we see that the variables with high coefficients also have high correlations with the corresponding variates. For example, $Z_3^{(1)}$ have the highest coefficient (1.0544531) for the canonical covariate U_1 and also the highest correlation (0.6363841).

e) Does significant canonical variates give good summary of data?

To see if the canonical variates give a good summary of data, we compute the proportion of explained (standardized) variance for our significant canonical variates U_1 and V_1.

```
U1 = sum(diag(R11)) # Should be 3 (or p...)!
V1 = sum(diag(R22)) # Should be 2 (or q...)!

U1_sum = (a1^2) %>% sum()
V1_sum = (b1^2) %>% sum()

c("% of var in X(1) explained by U1" = U1_sum / U1,
  "% of var in X(2) explained by V1" = V1_sum / V1)
```

```
## % of var in X(1) explained by U1 % of var in X(2) explained by V1
##                        0.6862471                        0.5292294
```

We observe that U_1 does a better job of explaining it's corresponding data set ($Z^{(1)}$) than V_1 does at explaining it's corresponding data set ($Z^{(2)}$). We think that $\approx 68\%$ and $\approx 53\%$ are fairly high, and conclude that the canonical variates does indeed give a good summary of data.

f) Comment on the success of the analysis

For the first question of this lab, we tested if there is at least one significant canonical correlation and in the second question we tested if there are any more significant canonical correlations. These formulas are based on a large sample size. The test conclusion from *1a)* is that we have at least one significant, although the test statistic was fairly close to the critical value. Therefore, with $n = 46$ and a test statistic that is fairly close to the critical value, we can question the validity of the whole analysis.

However, if we deem the results from 1a and 1b as adequate, then the analysis might be considered successful. Especially considering that we have deemed the proportions of the standardized sample variance to be high for each set.