Lab 4 732A97

Raymond Sseguya

2019-12-13

Question: Canonical correlation analysis by utilizing suitable software

Data

```
library(dplyr); library(knitr)

S <- as.matrix(read.table("P10-16.dat"))
R <- cov2cor(S); kable(R)
```

V1	V2	V3	V4	V5
1.0000000	0.2444071	0.0705263	0.1870842	0.0938948
0.2444071	1.0000000	0.5067278	-0.0346643	-0.0584436
0.0705263	0.5067278	1.0000000	0.3744425	-0.0536807
0.1870842	-0.0346643	0.3744425	1.0000000	0.2032893
0.0938948	-0.0584436	-0.0536807	0.2032893	1.0000000

10.16. Andrews and Herzberg [1] give data obtained from a study of a comparison of nondiabetic and diabetic patients. Three primary variables,

$$X_1^{(1)} = \text{glucose intolerance}$$
$$X_2^{(1)} = \text{insulin response to oral glucose}$$
$$X_3^{(1)} = \text{insulin resistance}$$

and two secondary variables,

$$X_1^{(2)} = \text{relative weight}.$$
$$X_2^{(2)} = \text{fasting plasma glucose}$$

were measured. The data for $n = 46$ nondiabetic patients yield the covariance matrix

$$\mathbf{S} = \left[\begin{array}{c|c} \mathbf{S}_{11} & \mathbf{S}_{12} \\ \hline \mathbf{S}_{21} & \mathbf{S}_{22} \end{array}\right] = \left[\begin{array}{ccc|cc} 1106.000 & 396.700 & 108.400 & .787 & 26.230 \\ 396.700 & 2382.000 & 1143.000 & -.214 & -23.960 \\ 108.400 & 1143.000 & 2136.000 & 2.189 & -20.840 \\ \hline .787 & -.214 & 2.189 & .016 & .216 \\ 26.230 & -23.960 & -20.840 & .216 & 70.560 \end{array}\right]$$

Determine the sample canonical variates and their correlations. Interpret these quantities. Are the first canonical variates good summary measures of their respective sets of variables? Explain. Test for the significance of the canonical relations with $\alpha = .05$.

Figure 1: Assignment4

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

```
inverse_sqrtm <- function(M){
  stopifnot(is.matrix(M))

  U <- eigen(M)$vectors
  D <- diag(eigen(M)$values)
  # print ( U%*%D%*%solve(U) )  # should be same as M
  M_sqrt_inv <- U%*%solve(sqrt(D))%*%solve(U)
  return(M_sqrt_inv)
}


ccm_Func <- function(M, p, q){

   M11=M[1:p, 1:p]; M12=M[1:p, (p+1):(p+q)]
   M21=t(M[1:p, (p+1):(p+q)]); M22=M[(p+1):(p+q), (p+1):(p+q)]
   CCM1=inverse_sqrtm(M11)%*%M12%*%solve(M22)%*%M21%*%inverse_sqrtm(M11)
   squared_rhos <- eigen(CCM1)$values
   # rhos <- sqrt(squared_rhos)
   e_vectors <- eigen(CCM1)$vectors
   a_s <- inverse_sqrtm(M11)%*%e_vectors
   f_vectors <- inverse_sqrtm(M22)%*%M21%*%inverse_sqrtm(M11)%*%e_vectors
   b_s <- inverse_sqrtm(M22)%*%f_vectors

   return(list(squared_rhos=squared_rhos, a_s=a_s, b_s=b_s))
}

R_res = ccm_Func(R,p=3,q=2)

test_statistic <- function(n,p,q, squared_rhos){
  -(n-1-0.5*(p+q+1))*log(prod(1-squared_rhos))  }
n = 46; p=3; q=2
t <- test_statistic(n=n, p=p, q=q, squared_rhos=R_res$squared_rhos)
c <- qchisq(p=1-0.05, df=p*q)

kable(t(c(t,c, as.character(t>c) )),
      col.names=c("Test Statistic", "Critical Value", "Check t > c"), digits = 3)
```

Test Statistic	Critical Value	Check t > c
13.7494849041102	12.591587243744	TRUE

We REJECT the null hypothesis at 5% significance level. Therefore we can say that that is SIGNIFICANT correlation between the groups of variables. The test statistic 13.7494849 is GREATER than the critical value 12.5915872.

b) How many pairs of canonical variates are significant?

```
options(digits=22); sqrt(R_res$squared_rhos)
```

```
## [1] 0.51734494531675435 0.12550820734296467                 NaN
```

There are two significant canonical correlations 0.517344945316754, 0.125508207342965 because they are not very close to zero.

c) Interpret the "significant" squared canonical correlations.

Tip: Read section "Canonical Correlations as Generalizations of Other Correlation Coefficients".

"Glucose intolerance" and "insulin response to oral glucose"" have a recognizable effect on "relative weight" and "fasting plasma resistance" among patients. Patients who are glucose intolerant will tend to be fatter.

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

```
options(digits=5)
kable(R_res$a_s, col.names=sapply(1:ncol(R_res$a_s),
                    FUN=function(i) paste0("a_",i)) )
```

a_1	a_2	a_3
0.43568	0.82318	0.44769
-0.70467	-0.45475	0.85224
1.08146	-0.40057	-0.14478

```
kable(R_res$b_s, col.names=sapply(1:ncol(R_res$b_s),
                    FUN=function(i) paste0("b_",i)) )
```

b_1	b_2	b_3
0.52781	-0.00596	0
-0.08326	0.12658	0

```
options(digits=3)
```

Because in this case the correlation matrix was used, U and V are the canonical variates and they are expressed as $U = aZ$ and $V = bZ$ where Z is the standardized variates and a and b are the vectors of the coefficients of the linear combinations of the standardized variates that summarise the correlation between the variables.

A 0.436 **increase** in "Glucose intolerance" and a -0.705 **decrease** in "insulin response to oral glucose" are

significantly correlated with a 0.528 **increase** in "relative weight" and a -0.083 **decrease** in "fasting plasma resistance"

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

Tip: Read section "Proportions of Explained Sample Variance".

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