Lab 4 732A97

Raymond Sseguya

2019-12-14

Question: Canonical correlation analysis by utilizing suitable software

Data

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

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

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

```r
R <- cov2cor(S)
## ![Assignment4](10p16.PNG)
```

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

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

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 $\alpha = .05$.

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, z=FALSE, tested=FALSE, sigf=2){

   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


   if(tested==FALSE){
     e_vectors <- eigen(CCM1)$vectors
     return(list(squared_rhos=squared_rhos))

   }else{
   e_vectors <- eigen(CCM1)$vectors[,1:sigf]

   if(z==FALSE){
      a_s <- t(e_vectors)%*%inverse_sqrtm(M11)
      ff <- (1/(sqrt(squared_rhos)))[1:sigf]
      f_vectors <- ff*inverse_sqrtm(M22)%*%M21%*%inverse_sqrtm(M11)%*%e_vectors
      b_s <- t(f_vectors)%*%inverse_sqrtm(M22)
   }else{
     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)) }
}

Res = ccm_Func(S,p=3,q=2); Res$squared_rhos
```

```
## [1] 2.676458e-01 1.575231e-02 1.861793e-17
```

```
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=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=7); sqrt(Res$squared_rhos)
```

```
## [1] 5.173449e-01 1.255082e-01 4.314850e-09
```

There are two significant canonical correlations 0.5173449, 0.1255082 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.

```
U_V = ccm_Func(S,p=3,q=2, tested=TRUE, sigf = 2)

options(digits=10)
kable(U_V$a_s, col.names=sapply(1:ncol(U_V$a_s),
                    FUN=function(i) paste0("a_",i)) )
```

a_1	a_2	a_3
0.0131006541	-0.0144382541	0.0233997230
0.0247524811	-0.0093175253	-0.0086672162

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

b_1	b_2
8.0622708375	-0.0029998004
-0.1094472452	0.1192659025

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

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

A 0.013 **change** in "Glucose intolerance" and a -0.014 **change** in "insulin response to oral glucose" are significantly correlated with a 8.062 **change** in "relative weight" and a -0.003 **change** 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”.

Basing on formula 10-37 on page 562 of the 6th edition of the text book, we can make the evaluation.

```
U_1 <- (1/p)*sum((U_V$a_s[1,])^2); V_1 <- (1/q)*sum((U_V$b_s[1,])^2)
U_2 <- (1/p)*sum((U_V$a_s[2,])^2); V_2 <- (1/q)*sum((U_V$b_s[2,])^2)
kable( t(c(U_1, V_1, U_2, V_2)), col.names = c("U_1","V_1", "U_2", "V_2"))
```

U_1	V_1	U_2	V_2
0	32.5	0	0.013

Something is wrong with my calculations but I understood the idea.

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

I think the canonical correlation analysis has failed because the results are very wierd for me.