Lab 4 732A97

Raymond Sseguya

2019-12-12

Question: Canonical correlation analysis by utilizing suitable software

Data

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

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

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)
}

# S; inverse_sqrtm(S)
S11=S[1:3,1:3]; S12=S[1:3,4:5]; S21=t(S[1:3,4:5]); S22=S[4:5,4:5]
CCM=inverse_sqrtm(S11)%*%S12%*%solve(S22)%*%S21%*%inverse_sqrtm(S11); CCM
```

```
##             [,1]        [,2]        [,3]
## [1,]  0.04680679 -0.03742592  0.08138543
## [2,] -0.03742592  0.03060065 -0.07167427
## [3,]  0.08138543 -0.07167427  0.20599066
```

```
squared_rhos <- eigen(CCM)$values; 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 = squared_rhos); t
```

```
## [1] 13.74948
```

```
c <- qchisq(p=1-0.05, df=p*q); c
```

```
## [1] 12.59159
```

```
t > c
```

```
## [1] TRUE
```

We REJECT the null hypothesis at 5% significance level. Therefore we can say that that is NO association 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(squared_rhos)
```

```
## [1] 5.1734494531675446e-01 1.2550820734296467e-01 4.3148498136107650e-09
```

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. Patients who are glucose intolerant will tend to be fatter.

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

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.