Multivariate Statistical Methods - Lab 04

Maximilian Pfundstein (maxpf364), Hector Plata (hecpl268), Aashana Nijhawan(aasni448), Lakshidaa Saigiridharan (laksa656)

2019-12-15

Contents

1 Canonical correlation analysis by utilizing suit- able software

Look at the data described in Exercise 10.16 of *Johnson, Wichern*. You may find it in the file `P10-16.DAT`. The data for 46 patients are summarized in a covariance matrix, which will be analyzed in R. Read through the description of the different R packages and functions so you may chose the most suitable one for the analysis. Supplement with own code where necessary.

```
data = read.table("P10-16.DAT")
head(data)
```

```
##         V1       V2       V3     V4      V5
## 1 1106.000  396.700  108.400  0.787  26.230
## 2  396.700 2382.000 1143.000 -0.214 -23.960
## 3  108.400 1143.000 2136.000  2.189 -20.840
## 4    0.787   -0.214    2.189  0.016   0.216
## 5   26.230  -23.960  -20.840  0.216  70.560
```

1.1 Association Between Groups

Task: Test at the 5 percent level if there is any association between the groups of variables.

```
myFun = function(Sigma, p=3, q=2, n=46, alpha=0.05) {
  S11 = as.matrix(Sigma[1:p,1:p])
  S22 = as.matrix(Sigma[(p+1):(p+q),(p+1):(p+q)])
  S = as.matrix(Sigma)

  test_statistics = n * log((det(S11) * det(S22) / det(S)))
  critical_value = qchisq(1 - alpha, df=p*q)

  return(test_statistics > critical_value)
}
```

```
myFun(data)
```

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

So we reject H_0 which means that we reject: $H_0 : \Sigma_{12} = 0$.

1.2 Number of Cononical Significant Variables

Task: How many pairs of canonical variates are significant?

```
give_me_Rho_sq_Plx = function(M, p=2, q=2) {
  R11 = as.matrix(M[1:p,1:p])
  R12 = as.matrix(M[1:p,(p+1):(p+q)])
  R21 = as.matrix(M[(p+1):(p+q), 1:p])
  R22 = as.matrix(M[(p+1):(p+q),(p+1):(p+q)])
  res = eigen(solve(sqrtm(R11)) %*% R12 %*% solve(R22) %*% R21 %*% solve(sqrtm(R11)))
  return(res$values)
}

significant_k = function(Sigma, alpha=0.05, n=46, p=3, q=2) {

  k_max = min(p, q)
  Rho_sq = give_me_Rho_sq_Plx(Sigma, p=p, q=q)

  for (k in 1:k_max) {

    test_statistics = - (n - 1 - 0.5 * (p + q + 1)) * log(prod(1 - Rho_sq[(k+1):p]))
    critical_value = qchisq(1 - alpha, df=(p-k)*(q-k))

    if (test_statistics >= critical_value)
      return(k)
  }
  return(k_max)
}
```

The amount of significant canonical variates is:

```
significant_k(data)
```

```
## [1] 2
```

1.3 Interpretation of the Significant Squared Canonical Correlations

Task: Interpret the "significant" squared canonical correlations.

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

```
rhos = give_me_Rho_sq_Plx(data, 3, 2)
print(rhos[1:2])
```

```
## [1] 0.26764579 0.01575231
```

Answer: Because of its multiple correlation coefficient interpretation, the kth *squared* canonical correlation ρ_k^{*2} is the proportion of the variance of canonical variate U_k "explained" by the set $\mathbf{X}^{(2)}$. It is also the proportion of the variance of canonical variate V_k "explained" by the set $\mathbf{X}^{(1)}$. Therefore, ρ_k^{*2} is often called the *shared variance* between the two sets $\mathbf{X}^{(1)}$ and $\mathbf{X}^{(2)}$ The largest value, ρ_1^{*2}, is sometimes regarded as a measure of set "overlap".

This means that 26.7% of the variance of the first canonical variate U_1 is explained by the set $X^{(2)}$. The same interpretation goes to the second squared canonical correlation, 1.5% of the variance of the first canonical variate U_1 is explaine by the set $X^{(2)}$.

1.4 Interpretation of Canonial Variates

Task: Interpret the canonical variates by using the coefficients and suitable correlations.

Answer: From the results below we see that the correlation between the first two canonical variables is about 0.5. This suggest, given a1 and b1 that being glucose intolerant and having a high insuline resistance with low insuline response to oral glucose is interconected with the weight.

As for the second pair of canonical variables, it's correlation is about 0.12. Given this and a2 and b2, we can say that theres some relationship between high glucose intolerance with low insule response to oral glucose and insule resistance with low weight and high fsating plasma glucose.

```
S11 = as.matrix(data[1:3,1:3])
S12 = as.matrix(data[1:3,(4):(5)])
S21 = as.matrix(data[(4):(5), 1:3])
S22 = as.matrix(data[(4):(5),(4):(5)])
res = eigen(solve(sqrtm(S11)) %*% S12 %*% solve(S22) %*% S21 %*% solve(sqrtm(S11)))

a1 =  solve(sqrtm(S11)) %*% res$vectors[, 1]
a2 =  solve(sqrtm(S11)) %*% res$vectors[, 2]

b1_prop = solve(S22) %*% S21 %*% a1
b2_prop = solve(S22) %*% S21 %*% a2

b1 = b1_prop / sqrt(t(b1_prop) %*% S22 %*% b1_prop)[1, 1]
b2 = b2_prop / sqrt(t(b2_prop) %*% S22 %*% b2_prop)[1, 1]

print("a1")
```

```
## [1] "a1"
```

```
print(a1)
```

```
##              [,1]
## [1,]  0.01310065
## [2,] -0.01443825
## [3,]  0.02339972
```

```
print("a2")
```

```
## [1] "a2"
```

```
print(a2)
```

```
##               [,1]
## [1,]  0.024752481
## [2,] -0.009317525
## [3,] -0.008667216
```

```
print("b1")
```

```
## [1] "b1"
```

```
print(b1)
```

```
##           [,1]
## V4  8.06557508
## V5 -0.01915905
```

```
print("b2")
```

```
## [1] "b2"
```

```r
print(b2)
```

```
##          [,1]
## V4 -0.3751678
## V5  0.1200675
```

```r
print("Correlation between U1 and V1")
```

```
## [1] "Correlation between U1 and V1"
```

```r
print(sqrt(res$values[1]))
```

```
## [1] 0.5173449
```

```r
print("Correlation between U2 and V2")
```

```
## [1] "Correlation between U2 and V2"
```

```r
print(sqrt(res$values[2]))
```

```
## [1] 0.1255082
```

1.5 Suitability of the Canonical Variates as a Summary Measure

Task: Are the "significant" canonical variates good summary measures of the respective data sets?

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

Answer: The second set of canonical variates are good summary measures of the second standardized dataset since all of its variance is explained by it, since the number of variable is the same as the number of significant canonical variates. As for the first set (U), the total variance of the original standardized variables Z(1) explained by it is almost 40% so it can be said that to some degree is a decent summary.

```r
# Calculating the correlation matrix.
S = as.matrix(data)
D = sqrt(diag(diag(S)))
R = solve(D) %*% S %*% solve(D)

# Getting the canonical coefficient vectors
# of the standardized observations.
S11 = as.matrix(R[1:3,1:3])
S12 = as.matrix(R[1:3,(4):(5)])
S21 = as.matrix(R[(4):(5), 1:3])
S22 = as.matrix(R[(4):(5),(4):(5)])
res = eigen(solve(sqrtm(S11)) %*% S12 %*% solve(S22) %*% S21 %*% solve(sqrtm(S11)))

a1 =  solve(sqrtm(S11)) %*% res$vectors[, 1]
a2 =  solve(sqrtm(S11)) %*% res$vectors[, 2]
a3 =  solve(sqrtm(S11)) %*% res$vectors[, 3]

b1_prop = solve(S22) %*% S21 %*% a1
b2_prop = solve(S22) %*% S21 %*% a2

b1 = b1_prop / sqrt(t(b1_prop) %*% S22 %*% b1_prop)[1, 1]
b2 = b2_prop / sqrt(t(b2_prop) %*% S22 %*% b2_prop)[1, 1]

A_z = as.matrix(cbind(a1, a2, a3), 3, 3)
B_z = as.matrix(cbind(b1, b2), 2, 2)
```

```r
A_z_inv = solve(A_z)
A_z_inv = A_z_inv[1:2, 1:2]
B_z_inv = solve(B_z)

prop_U_set = sum(diag(A_z_inv[, 1] %*% t(A_z_inv[, 1]) +
                      A_z_inv[, 2] %*% t(A_z_inv[, 2]))) / 2

prop_B_set = sum(diag(B_z_inv[, 1] %*% t(B_z_inv[, 1]) +
                      B_z_inv[, 2] %*% t(B_z_inv[, 2]))) / 2

print("Proportion of total standardized sample variance in the first set explained by U1 and U2")
```

```
## [1] "Proportion of total standardized sample variance in the first set explained by U1 and U2"
```

```r
print(prop_U_set)
```

```
## [1] 0.3969631
```

```r
print("Proportion of total standardized sample variance in the first set explained by U1 and U2")
```

```
## [1] "Proportion of total standardized sample variance in the first set explained by U1 and U2"
```

```r
print(prop_B_set)
```

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

1.6 Opinion on the Success of the canonical analysis.

Task: Give your opinion on the success of this canonical correlation analysis.

We think that the canonical correlation analysis was successful since we found two canonical variables that are significant and explain up to some degree, the variability of the original data. It also helped us determine the joint relationship between the variables on the two datasets as seen in task c.