GR5223 - HW5

Mathieu Sauterey - UNI: mjs2364

April 8, 2018

Problem 1

A sample of n = 140 seventh-grade children received four tests on x_1 = reading speed, x_2 = reading power, y_1 = arithmetic speed, and y_2 = arithmetic power. The correlations for performance are given below.

```
# Creates the sample correlation matrix given by the problem
R <- matrix(c(1,0.6328,0.2412,0.0586, 0.6328, 1, -0.0553,0.0655,
              0.2412, -0.0553, 1, 0.4248, 0.0586, 0.0655, 0.4248, 1),
            nrow = 4, ncol = 4)
```

(a) Find the sample canonical correlation coefficients and sample canonical variables.

```
# We create the necessary sample correlation matrices of R
Rxx <- R[1:2,1:2]
Ryy <- R[3:4,3:4]
Rxy <- R[1:2,3:4]


## Now we need to calculate K_hat = Rxx^(-1/2)*Rxy*Ryy^(-1/2)

# First we calculate Rxx^(-1/2) and Ryy^(-1/2)
Rxx_n05 <- eigen(Rxx)$vectors %*% diag(1/sqrt(eigen(Rxx)$values)) %*% t(eigen(Rxx)$vectors)
Ryy_n05 <- eigen(Ryy)$vectors %*% diag(1/sqrt(eigen(Ryy)$values)) %*% t(eigen(Ryy)$vectors)

# Second, we get K_hat
K_hat <- Rxx_n05 %*% Rxy %*% Ryy_n05

## Then we calculate the sample canonical correlation coefficients

# We complete SVD on K_hat*t(K_hat)

G <- svd(K_hat)$u
L <- svd(K_hat)$d
D <- svd(K_hat)$v

# We calculate the correlation vectors
A_hat <- t(G) %*% Rxx_n05
rownames(A_hat) <- c("a1","a2")
print(A_hat)
```

```
##          [,1]      [,2]
## a1 -1.2568447 1.0253167
## a2  0.2970177 0.7852413
```

```
B_hat <- t(D) %*% Ryy_n05
rownames(B_hat) <- c("b1","b2")
print(B_hat)
```

```
##           [,1]      [,2]
## b1 -1.10447218 0.4527216
## b2 -0.01818009 1.0075875
```

```
# Finally we get the canonical correlation coefficients from L
names(L) <- c("r1","r2")
print(L)
```

```
##         r1         r2
## 0.39450592 0.06884787
```

Based on the results above we find the following sample canonical variables:

η_1=-1.257x_1+1.025x_2 and ϕ_1=-1.104y_1+0.452y_2

η_2=0.297x_1+0.785x_2 and ϕ_2=-0.018y_1+1.007y_2

The sample canonical correlation coefficients given above are r_1=0.3945 r_2=0.06885

(b) Find the p-value for a test of H0 : Rxy = 0.

Testing $H_0 : \sum_{xy}$=0 is equivalent to testing the population canonical correlation coefficients $H_0 : \rho_1 = \rho_2 = 0$.

We have to assume multivariate normality from here on.

```
p <- 2
q <- 2
n <- 140

# Let's calculate the Wilks' likelihood ratio test statistic
Lambda_0 <- prod(1-L^2)

# Calculates the modified test statistics approx ~ chi-square with df = pq
ModLambda_0 <-  -(n-0.5*(p+q+3))*log(Lambda_0)

# Calculates the p-value for the test
1 - pchisq(ModLambda_0, df=p*q)
```

```
## [1] 9.00415e-05
```

The p-value is essentially zero so we reject the null hypothesis $H_0 : \rho_1 = \rho_2 = 0$.

(c) If you reject H_0 : rho_1 = rho_2 = 0 at the alpha = 0.05 level of significance, find the p-value for a test of H0 : rho_2 = 0.

We reject $H_0 : \rho_1 = \rho_2 = 0$ at the $\alpha = 0.05$ level of significance, so we test $H_0 : \rho_2 = 0$.

```
k <- min(p,q)
s <- 1

# Let's calculate the Wilks' likelihood ratio test statistic
Lambda_s <- prod(1-L[(s+1):k]^2)
```

```
# Calculates the modified test statistics approx ~ chi-square with df = (p-s)(q-s)
ModLambda_1 <-  -(n-0.5*(p+q+3))*log(Lambda_s)

# Calculates the p-value for the test
1 - pchisq(ModLambda_1, df=(p-s)*(q-s))
```

```
## [1] 0.4206308
```

The p-value = 0.42 so we fail to reject $H_0 : \rho_2 = 0$ at the $\alpha = 0.05$ level of significance.

(d) Does reading ability (as measured by the two tests) correlate with arithmetic ability (as measured by the two tests)? Discuss.

Since we proved above that ρ_2 is not statistically significant, then only ρ_1 shows a significant correlation between linear combinations of the reading ability and arithmetic ability. Indeed, the first (and unique) canonical variate pair has a correlation of 0.3945 and roughly highlights that the difference between reading speed and reading power is correlated with the difference between arithmetic speed and arithmetic power. Thus, better reading ability does not imply better arithmetic ability, but larger difference between reading speed and power implies a larger difference between arithemtic speed and power.

Problem 2

A random sample of n = 70 families will be surveyed to determine the association between certain "demographic" variables and certain "consumption" variables.

Define the Criterion set of variables by:

y_1 = annual frequency of dining at a restaurant y_2 = annual frequency of attending movies

and the Predictor set by

x_1 = age of head of household x_2 = annual family income x_3 = educational level of head of household

Suppose 70 observations on the preceding variables give the sample correlation matrix below:

```
# Creates the sample correlation matrix given by the problem
R2 <- matrix(c(1,0.37,0.21,0.26, 0.33, 0.37, 1,0.35,
               0.67, 0.59, 0.21, 0.35, 1, 0.34, 0.34, 0.26, 0.67,0.34,1, 0.8, 0.33,
               0.59, 0.34, 0.80, 1),
             nrow = 5, ncol = 5)
```

(a) Determine the sample canonical correlation coefficients, and find the p-value for a test of the null hypothesis H0 : Sigma_XY = 0

```
p2 <- 2
q2 <- 3
n2 <- 70

# We create the necessary sample correlation matrices of R
R2xx <- R2[1:q2,1:q2]
R2yy <- R2[(nrow(R2)-p2+1):nrow(R2),(nrow(R2)-p2+1):nrow(R2)]
R2xy <- R2[(1:q2),(nrow(R2)-p2+1):nrow(R2)]
```

```
## Now we need to calculate K2_hat = R2xx^(-1/2)*R2xy*R2yy^(-1/2)

# First we calculate R2xx^(-1/2) and R2yy^(-1/2) using spectral decomposition
R2xx_n05 <- eigen(R2xx)$vectors %*% diag(1/sqrt(eigen(R2xx)$values)) %*% t(eigen(R2xx)$vectors)
R2yy_n05 <- eigen(R2yy)$vectors %*% diag(1/sqrt(eigen(R2yy)$values)) %*% t(eigen(R2yy)$vectors)

# Second, we get K2_hat
K2_hat <- R2xx_n05 %*% R2xy %*% R2yy_n05

## Then we calculate the sample canonical correlation coefficients

# We complete SVD on K2_hat*t(K2_hat)

G2 <- svd(K2_hat)$u
L2 <- svd(K2_hat)$d
D2 <- svd(K2_hat)$v

# Finally we get the canonical correlation coefficients from L
names(L2) <- c("r1","r2")
print(L2)
```

```
##        r1        r2
## 0.6879479 0.1868654
```

The sample canonical correlation coefficients given above are r_1=0.6879 r_2=0.1869

Testing H_0 : $\sum_{xy}$=0 is equivalent to testing the population canonical correlation coefficients H_0 : $\rho_1 = \rho_2$ =0.

We have to assume multivariate normality from here on.

```
# Let's calculate the Wilks' likelihood ratio test statistic
Lambda2_0 <- prod(1-L2^2)

# Calculates the modified test statistics approx ~ chi-square with df = pq
ModLambda2_0 <-  -(n2-0.5*(p2+q2+3))*log(Lambda2_0)

# Calculates the p-value for the test
1 - pchisq(ModLambda2_0, df=p2*q2)
```

```
## [1] 5.476244e-08
```

The p-value is essentially zero so we reject the null hypothesis H_0 : $\rho_1 = \rho_2 = 0$.

(b) If H0 : Sigma_XY = 0 is rejected at the alpha = 0.05 level, test the null hypothesis that the second and higher canonical correlations are all zero.

We reject $\sum_{XY} = 0$ at the $\alpha = 0.05$ level of significance, so we test H_0 : $\rho_2 = 0$.

```
k2 <- min(p2,q2)
s2 <- 1

# Let's calculate the Wilks' likelihood ratio test statistic
Lambda2_s <- prod(1-L2[(s2+1):k2]^2)
```

```r
# Calculates the modified test statistics approx ~ chi-square with df = (p2-s2)(q2-s2)
ModLambda2_1 <-  -(n2-0.5*(p2+q2+3))*log(Lambda2_s)

# Calculates the p-value for the test
1 - pchisq(ModLambda2_1, df=(p2-s2)*(q2-s2))
```

```
## [1] 0.309463
```

The p-value = 0.301 so we fail to reject $H_0 : \rho_2 = 0$ at the $\alpha = 0.05$ level of significance. Therefore, the first canonical correlation is the only significant one.

(c) Using standardized variables, construct the canonical variables corresponding to the significant (alpha = 0.05) canonical correlation(s).

```r
# We calculate the correlation vectors
A2_hat <- t(-G2) %*% R2xx_n05
rownames(A2_hat) <- c("a1","a2")
print(A2_hat)
```

```
##           [,1]      [,2]       [,3]
## a1  0.04912716 0.8975114  0.1900411
## a2 -1.00029422 0.5836950 -0.2955723
```

```r
B2_hat <- t(-D2) %*% R2yy_n05
rownames(B2_hat) <- c("b1","b2")
print(B2_hat)
```

```
##         [,1]       [,2]
## b1 0.7689274  0.2720729
## b2 1.4786914 -1.6443096
```

Since only ρ_1 is significant at α=0.05, we find the below sample canonical variable pair from the first canonical correlation vectors $\hat{a}$ and $\hat{b}$ computed above :

η_1= $0.0491x_1$+$0.8975x_2$+$0.1900x_3$ and ϕ_1= $0.7689y_1$+$0.2721y_2$

(d) Interpret the canonical variates corresponding to significant (alpha = 0.05) canonical correlation(s).

η_1 is roughly a weighted average of the demographics predictors, which the largest weight assigned to x_2 the annual family income, and the second largest to x_3 the educational level of head of household. This means that these two demographic predictors are most correlated with the consumption criteria.

ϕ_1 is roughly a weighted average of the consumption criteria, which the largest weight assigned to y_1 the annual frequency of dining at a restaurant, and the second largest to y_2 the annual frequency of attending movies. This means that these two consumption predictors are most correlated with the demographic predictors.

We can interpret the above result as follows: η_1 is primarily an indicator of annual family income and it is most correlated with ϕ_1 which primarily represents the annual frequency of dining at a restaurant for a family.

(e) Do the demographic variables have something to say about the consumption variables? Do the consumption variables provide information about the demographic variables?

The first canonical variable pair has a statiscally significant correlation of 0.6879 therefore consumption criteria provide information about demographic predictors, and vice-versa.