Multivariate Data Analysis
Prof. Dr. Christina Andersson

R Solution Exercise Sheet 5: Canonical Correlation Analysis

Computer Problems:

1. (a)
```
> dim(mmreg)
[1] 600   9
```

   (b)
```
psych <- mmreg[, 2:4]
acad <- mmreg[, 5:8]
```

   (c)
```
require(ggplot2)
require(GGally)
require(CCA)
 colnames(mmreg) <- c("No", "Control", "Concept", "Motivation", "Read", "Write", "Math",
+     "Science", "Sex")
> summary(mmreg)
       No           Control            Concept            Motivation          Read
 Min.   :  1.0   Min.   :-2.23000   Min.   :-2.620000   Min.   :0.0000   Min.   :28.3
 1st Qu.:150.8   1st Qu.:-0.37250   1st Qu.:-0.300000   1st Qu.:0.3300   1st Qu.:44.2
 Median :300.5   Median : 0.21000   Median : 0.030000   Median :0.6700   Median :52.1
 Mean   :300.5   Mean   : 0.09653   Mean   : 0.004917   Mean   :0.6608   Mean   :51.9
 3rd Qu.:450.2   3rd Qu.: 0.51000   3rd Qu.: 0.440000   3rd Qu.:1.0000   3rd Qu.:60.1
 Max.   :600.0   Max.   : 1.36000   Max.   : 1.190000   Max.   :1.0000   Max.   :76.0
     Write           Math           Science           Sex
 Min.   :25.50   Min.   :31.80   Min.   :26.00   Min.   :0.000
 1st Qu.:44.30   1st Qu.:44.50   1st Qu.:44.40   1st Qu.:0.000
 Median :54.10   Median :51.30   Median :52.60   Median :1.000
 Mean   :52.38   Mean   :51.85   Mean   :51.76   Mean   :0.545
 3rd Qu.:59.90   3rd Qu.:58.38   3rd Qu.:58.65   3rd Qu.:1.000
 Max.   :67.10   Max.   :75.50   Max.   :74.20   Max.   :1.000
```

   (d)
```
> cc1 <- cc(psych, acad)
```

   (e)
```
> # display the canonical correlations
> cc1$cor
[1] 0.4640861 0.1675091 0.1039911
```

   (f)
```
> # raw canonical coefficients
> cc1[3:4]
$xcoef
                 [,1]       [,2]       [,3]
Control    -1.2538339 -0.6214775 -0.6616896
Concept     0.3513499 -1.1876867  0.8267209
Motivation -1.2624203  2.0272641  2.0002284

$ycoef
                [,1]         [,2]        [,3]
Read    -0.044620596 -0.004910018  0.02138056
Write   -0.035877112  0.042071471  0.09130733
Math    -0.023417185  0.004229472  0.00939821
Science -0.005025157 -0.085162175 -0.10983502
```

```
Sex     -0.632119239  1.084642482 -1.79464692
```

(g) The raw canonical coefficients are interpreted in a manner analogous to interpreting regression coefficients i.e., for the variable read, a one unit increase in reading leads to a .0446 decrease in the first canonical variate of set 2 when all of the other variables are held constant. Here is another example: being female leads to a .6321 decrease in the dimension 1 for the academic set with the other predictors held constant.

(h)
```
> # compute canonical loadings
> cc2 <- comput(psych, acad, cc1)
>
> # display canonical loadings
> cc2[3:6]
$corr.X.xscores
                  [,1]       [,2]       [,3]
Control    -0.90404632 -0.3896883 -0.1756227
Concept    -0.02084327 -0.7087386  0.7051632
Motivation -0.56715105  0.3508882  0.7451290

$corr.Y.xscores
              [,1]        [,2]        [,3]
Read    -0.3900402 -0.06010654  0.01407660
Write   -0.4067914  0.01086074  0.02647208
Math    -0.3545378 -0.04990916  0.01536586
Science -0.3055607 -0.11336979 -0.02395489
Sex     -0.1689796  0.12645737 -0.05650916

$corr.X.yscores
                   [,1]        [,2]        [,3]
Control    -0.419555308 -0.06527635 -0.01826320
Concept    -0.009673071 -0.11872021  0.07333073
Motivation -0.263206905  0.05877698  0.07748682

$corr.Y.yscores
              [,1]        [,2]       [,3]
Read    -0.8404480 -0.35882539  0.1353635
Write   -0.8765429  0.06483672  0.2545609
Math    -0.7639483 -0.29794886  0.1477612
Science -0.6584139 -0.67679758 -0.2303551
Sex     -0.3641127  0.75492816 -0.5434035
```

These loadings are correlations between variables and the canonical variates.

The above correlations are between observed variables and canonical variables which are known as the canonical loadings. These canonical variates are actually a type of latent variable.

(i)
```
> # tests of canonical dimensions
> ev <- (1 - cc1$cor^2)
>
> n <- dim(psych)[1]
> p <- length(psych)
> q <- length(acad)
> k <- min(p, q)
> m <- n - 3/2 - (p + q)/2
>
> w <- rev(cumprod(rev(ev)))
>
> # initialize
> d1 <- d2 <- f <- vector("numeric", k)
>
> for (i in 1:k) {
+     s <- sqrt((p^2 * q^2 - 4)/(p^2 + q^2 - 5))
+     si <- 1/s
+     d1[i] <- p * q
+     d2[i] <- m * s - p * q/2 + 1
+     r <- (1 - w[i]^si)/w[i]^si
+     f[i] <- r * d2[i]/d1[i]
+     p <- p - 1
+     q <- q - 1
+ }
>
> pv <- pf(f, d1, d2, lower.tail = FALSE)
> (dmat <- cbind(WilksL = w, F = f, df1 = d1, df2 = d2, p = pv))
        WilksL         F df1      df2            p
[1,] 0.7543611 11.715733  15 1634.653 7.497602e-28
[2,] 0.9614300  2.944459   8 1186.000 2.905057e-03
[3,] 0.9891858  2.164612   3  594.000 9.109217e-02
```

As shown in the table above, the first test of the canonical dimensions tests whether all three dimensions are significant (they are, F = 11.72), the next test tests whether dimensions 2 and 3 combined are significant (they are, F = 2.94). Finally, the last test tests whether dimension 3, by itself, is significant (it is not). Therefore dimensions 1 and 2 must each be significant while dimension three is not.