HomeWork 4-5

Nicholas Anthony Small

```
data("heptathlon",package="HSAUR2")
mydata <- heptathlon[-25,-8] # remove the PNG as an outlier and the last vari
able (the final scores)
```

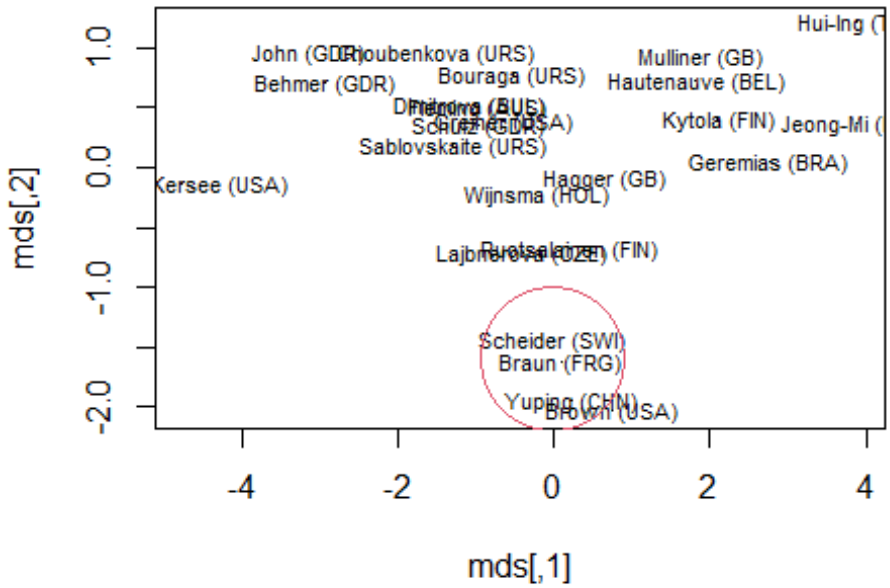
a) Create a scaled distance matrix for observations.

```
d <- dist(scale(mydata))
```

b) Perform a graphical MDS analysis on the resulting "distance" matrix of part a. Label the points using the row names (set an appropriate cex (size) for a better view). Who is the most similar athlete to Scheider(SWI)?

```
# Braun is the most similar athlete to Scheider

mds <- cmdscale(d)
plot(mds, pch = ". ")
text(mds, labels = rownames(mydata), cex = 0.7)
points(0, -1.6, cex = 10, col =2)
```



c) Use the correlation matrix of the data. Convert the correlation matrix to a distance matrix by computing (1-correlation) . Explain why the resulting matrix represents "distances" between variables.

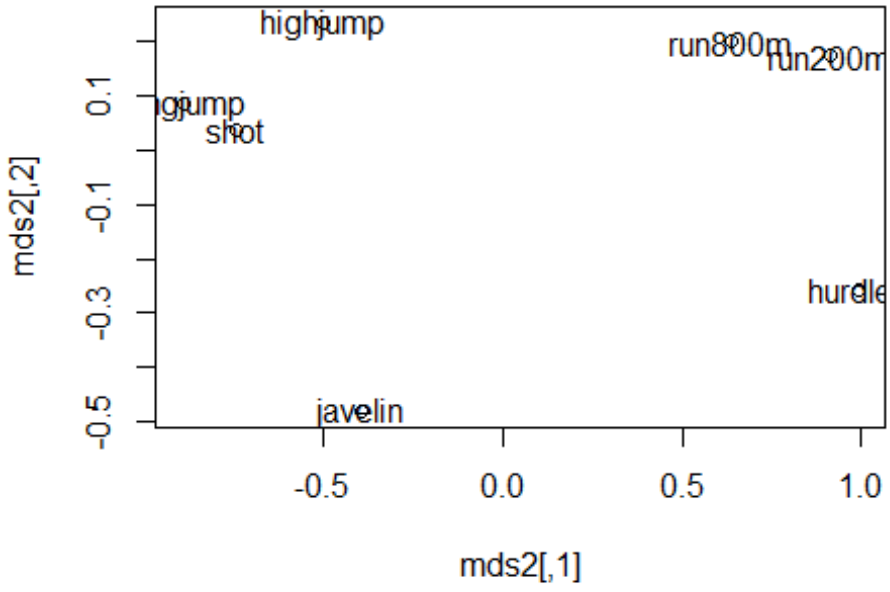
```
# The matrix represents the correlation between each variable based on the di
stances. When the distance is large it relates to a low or negative correlati
on and if it is small there is a high correlation.

v.dist <- 1-cor(mydata)
round(v.dist, 2)
```

```
##          hurdles highjump shot run200m longjump javelin run800m
## hurdles     0.00     1.58 1.77    0.17     1.89    1.33    0.44
## highjump    1.58     0.00 0.54    1.39     0.34    0.65    1.15
## shot        1.77     0.54 0.00    1.67     0.22    0.66    1.41
## run200m     0.17     1.39 1.67    0.00     1.81    1.47    0.43
## longjump    1.89     0.34 0.22    1.81     0.00    0.71    1.52
## javelin     1.33     0.65 0.66    1.47     0.71    0.00    1.26
## run800m     0.44     1.15 1.41    0.43     1.52    1.26    0.00
```

d) Perform a graphical MDS analysis on the resulting "distance" matrix of part c. Label the points using the column names (set an appropriate cex (size) for a better view). What variables are more similar (related) to each other?

```
# Shot and Longjump has a high correlation, thus making them similar.
# Hurdles and run200m is also similar
mds2 <- cmdscale(v.dist)
plot(mds2)
text(mds2, labels = colnames(mydata))
```



Problem 2

```
grad <- read.csv("http://westfall.ba.ttu.edu/isqs6348/Rdata/pgs.csv", header
= T)
```

Two variables of interest are FacTeaching, a 1, 2, 3, 4, 5 ratings of teaching at TTU by the student, and COL, the college from which the student graduated. Perform a correspondence analysis of these two variables as follows.

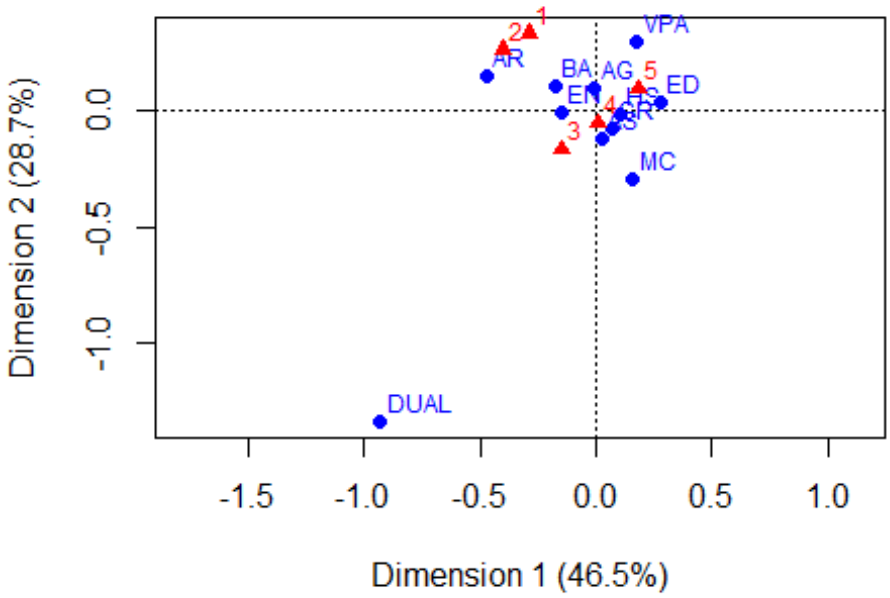
a) Construct the contingency table showing counts of students in all combinations of these two variables.

```
tbl = table(grad$COL, grad$FacTeaching)
tbl
```

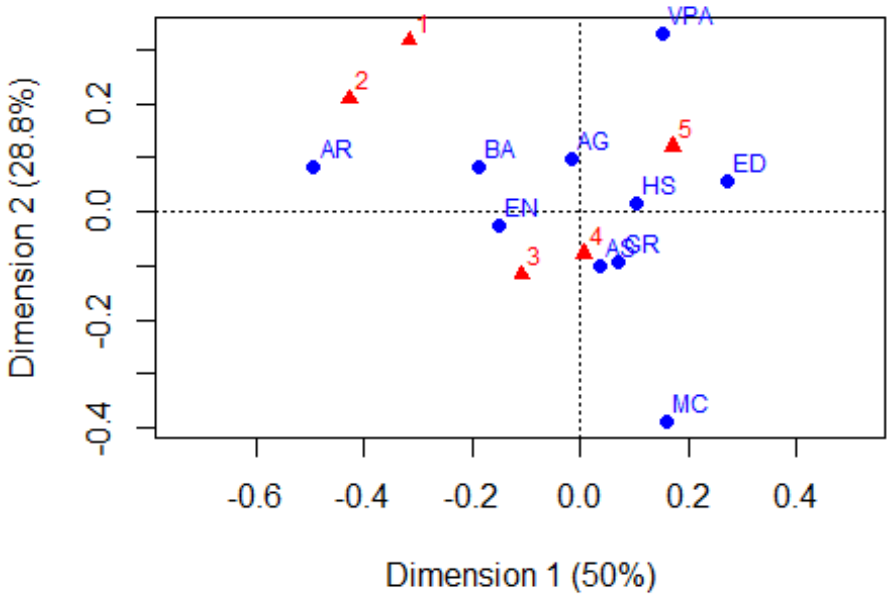
```
##
##          1   2   3   4   5
##   AG     4  15  26  78  56
##   AR     3   4   6  16   4
##   AS    12  24 124 290 171
##   BA     9  28  44 116  66
##   DUAL   0   0   2   0   0
##   ED     3   6  26 113  93
##   EN     5  36  65 168  86
##   GR     0   3   8  27  15
##   HS     1   5  17  41  33
##   MC     0   0   3  25   6
##   VPA    4   7  10  37  44
```

b) Construct the correspondence analysis (CA) plot and comment on the outlier, in light of your table in A. Then remove the outlier data you discovered and re-construct the CA plot.

```
#This includes an outlier of DUAL which has only 2 students. This will skew t
he data and should be removed.
library(ca)
grad.ca <- ca(tbl)
plot(grad.ca)
```




```
#Removing the outlier of Dual and reconstructing the plot
tbl2 <- tbl[-5,]
grad.ca2 <- ca(tbl2)
plot(grad.ca2)
```



c) Pick three colleges in your graph, two of which are close to each other, and the third of which is far from your first two. Find the three conditional distributions of rating for your three colleges, and interpret the distance between the graph points in terms of "distances" between those three conditional distributions

```
# We can conclude that
#P(5 | ED ) \approv P(5 | HS)
#P(5 | AR ) < P(5 | ED)
#P(1 | AR ) > P(1 | ED)


#Proof
cond.tbl <- prop.table(tbl2, margin = 1) ;
round(cond.tbl,2)

##
##          1    2    3    4    5
##   AG  0.02 0.08 0.15 0.44 0.31
##   AR  0.09 0.12 0.18 0.48 0.12
##   AS  0.02 0.04 0.20 0.47 0.28
##   BA  0.03 0.11 0.17 0.44 0.25
##   ED  0.01 0.02 0.11 0.47 0.39
##   EN  0.01 0.10 0.18 0.47 0.24
##   GR  0.00 0.06 0.15 0.51 0.28
##   HS  0.01 0.05 0.18 0.42 0.34
##   MC  0.00 0.00 0.09 0.74 0.18
##   VPA 0.04 0.07 0.10 0.36 0.43
```

```r
cond.tbl[c(2, 5, 8)]
```

```
## [1] 0.09090909 0.01244813 0.01030928
```

```r
# We can conclude that The probability of ED = 5 is close to the probability
of HS = 5
# The probability of AR = 5 is less than the probability of ED = 5
# The probability of AR = 2 is greater than the probability of ED = 1

# P(5 | ED) = 0.3859
# P(5 | HS) = 0.3402
# P(1 | AR) = 0.0909 > P(1 | ED) = 0.0124
```

Problem 3

Use the Daily stock returns data set. The columns are companies; Man1, Man2, Man3 are manufacturing companies; Serv1, Serv2, Serv3, Serv4 are service companies.

```r
stock <- read.csv("https://bit.ly/3egKiMU")
stock = stock*100
```

a) Perform an exploratory factor analysis (EFA) using two factors.

```r
efa <- factanal(stock, factor=2, scores = "regression")
efa
```

```
##
## Call:
## factanal(x = stock, factors = 2, scores = "regression")
##
## Uniquenesses:
##  Man1  Man2  Man3 Serv1 Serv2 Serv3 Serv4
## 0.764 0.513 0.520 0.900 0.714 0.788 0.574
##
## Loadings:
##       Factor1 Factor2
## Man1  0.462   0.148
## Man2  0.674   0.180
## Man3  0.676   0.150
## Serv1 0.131   0.288
## Serv2 0.128   0.519
## Serv3 0.131   0.441
## Serv4 0.136   0.639
##
##                Factor1 Factor2
## SS loadings      1.195   1.032
## Proportion Var   0.171   0.147
## Cumulative Var   0.171   0.318
##
## Test of the hypothesis that 2 factors are sufficient.
## The chi square statistic is 5.1 on 8 degrees of freedom.
## The p-value is 0.747
```

```
# The P-Value is 0.747
efa$PVAL

## objective
## 0.7467672
```

b) Interpret the p-value reported in your EFA.

P is > than 0.05. The really high p-value means that we fail to reject the null hypothesis (In an EFA, the null hypothesis # is that the model described by the factor we have found predicts the data well).

c) What are the factors (latent variables) in this model? Name them.

The latent variables are combinations of Man1, Man2, Man3, Serv1, Serv2, Serv3 and Serv4

d) Write the EFA regression model for variable Man1. For example

$$Man1 = 0.462f1 + 0.148f2 + e$$

Based on the EFA outputs, what are a and b?

e) In the model of part d, determine the variance of the error term e.

```
# The Variance of error is 0.764 which is very high compared to the others.
efa$uniquenesses[1]

##      Man1
## 0.7642471
```

f) What is the correlation between f2 and Serve2?

```
# The correlation of f2 and Swerve2 is 0.689.
cor(efa$scores[,2], stock$Serv2)

## [1] 0.6887696
```

g) Compare the EFA approximated correlation matrix versus the actual correlation matrix. Report RMSE. What do you conclude?

```
# Comparison of EFA approximated correlation matrix vs the actual correlation
matrix.
# The actual correlation has negative correlation which the approximate corre
lation are all positive.

f.loading <- efa$loadings[,1:2]
corHat <- f.loading %*% t(f.loading) + diag(efa$uniquenesses)
round(corHat,2)
```

```
##       Man1 Man2 Man3 Serv1 Serv2 Serv3 Serv4
## Man1  1.00 0.34 0.34  0.10  0.14  0.13  0.16
## Man2  0.34 1.00 0.48  0.14  0.18  0.17  0.21
## Man3  0.34 0.48 1.00  0.13  0.16  0.15  0.19
## Serv1 0.10 0.14 0.13  1.00  0.17  0.14  0.20
## Serv2 0.14 0.18 0.16  0.17  1.00  0.25  0.35
## Serv3 0.13 0.17 0.15  0.14  0.25  1.00  0.30
## Serv4 0.16 0.21 0.19  0.20  0.35  0.30  1.00
```

```
corr <- cor(mydata)
round(corr,2)
```

```
##          hurdles highjump  shot run200m longjump javelin run800m
## hurdles     1.00    -0.58 -0.77    0.83    -0.89   -0.33    0.56
## highjump   -0.58     1.00  0.46   -0.39     0.66    0.35   -0.15
## shot       -0.77     0.46  1.00   -0.67     0.78    0.34   -0.41
## run200m     0.83    -0.39 -0.67    1.00    -0.81   -0.47    0.57
## longjump   -0.89     0.66  0.78   -0.81     1.00    0.29   -0.52
## javelin    -0.33     0.35  0.34   -0.47     0.29    1.00   -0.26
## run800m     0.56    -0.15 -0.41    0.57    -0.52   -0.26    1.00
```

```
# The calculated RMSE is 0.594 chance of error in the correlation.
rmse =sqrt(mean((corHat-corr)^2)); rmse
```

```
## [1] 0.5940684
```

Problem 4

Perform factor analysis on questions 22-35 of TTU web survey data (The text description of variables and constructs are available at this link).

```
ttuweb <- read.csv("https://bit.ly/3oNr5qX")
mydata2 <- ttuweb[,22:35]
```

a) There are some missing values in this data. Find the correlation matrix based on pair-wise deletion. You suppose to use this correlation matrix as an input for EFA.

```
Mpwd=cor(mydata2,use="pairwise.complete.obs")
```

b) Perform EFA suggesting two common factors. How would you name those factors?

```
# Factor 1 would be named after a persons "attitude toward TTU"
# Factor 2 would be named after a person "attitude toward web site"

efa2 = factanal(covmat = Mpwd, factors = 2, n.obs = nrow(mydata2))
efa2
```

```
##
## Call:
## factanal(factors = 2, covmat = Mpwd, n.obs = nrow(mydata2))
##
## Uniquenesses:
##   Q22   Q23   Q24   Q25   Q26   Q27   Q28   Q29   Q30   Q31   Q32   Q33
```

```
Q34
## 0.530 0.512 0.582 0.614 0.289 0.355 0.206 0.289 0.625 0.470 0.483 0.506 0.
636
##   Q35
## 0.653
##
## Loadings:
##     Factor1 Factor2
## Q22  0.617  -0.301
## Q23  0.665  -0.214
## Q24  0.615  -0.199
## Q25  0.567  -0.255
## Q26  0.827  -0.165
## Q27  0.788  -0.155
## Q28  0.888
## Q29  0.838
## Q30 -0.153   0.593
## Q31 -0.264   0.679
## Q32 -0.197   0.691
## Q33 -0.126   0.692
## Q34          0.603
## Q35 -0.216   0.548
##
##                Factor1 Factor2
## SS loadings      4.512   2.740
## Proportion Var   0.322   0.196
## Cumulative Var   0.322   0.518
##
## Test of the hypothesis that 2 factors are sufficient.
## The chi square statistic is 329.83 on 64 degrees of freedom.
## The p-value is 1.94e-37

# Demonstrates all values greater than 0.5
print(efa2$loadings, cut = 0.5)

##
## Loadings:
##     Factor1 Factor2
## Q22  0.617
## Q23  0.665
## Q24  0.615
## Q25  0.567
## Q26  0.827
## Q27  0.788
## Q28  0.888
## Q29  0.838
## Q30          0.593
## Q31          0.679
## Q32          0.691
## Q33          0.692
```

```
## Q34          0.603
## Q35          0.548
##
##                Factor1 Factor2
## SS loadings      4.512   2.740
## Proportion Var   0.322   0.196
## Cumulative Var   0.322   0.518
```

c) Perform EFA suggesting three common factors. How would you name those factors?

```
# Factor 1 would be named after a persons "attitude toward TTU"
# Factor 2 would be named after a person "attitude toward web site"
# Factor 3 would be named after a person "attitude toward TTU"

efa3 = factanal(covmat = Mpwd, factors = 3, n.obs = nrow(mydata2))
efa3

##
## Call:
## factanal(factors = 3, covmat = Mpwd, n.obs = nrow(mydata2))
##
## Uniquenesses:
##   Q22   Q23   Q24   Q25   Q26   Q27   Q28   Q29   Q30   Q31   Q32   Q33
Q34
## 0.444 0.304 0.380 0.476 0.305 0.376 0.162 0.215 0.603 0.459 0.480 0.512 0.
627
##   Q35
## 0.653
##
## Loadings:
##     Factor1 Factor2 Factor3
## Q22  0.369  -0.256   0.595
## Q23  0.351  -0.142   0.743
## Q24  0.306  -0.126   0.715
## Q25  0.291  -0.196   0.633
## Q26  0.722  -0.189   0.371
## Q27  0.671  -0.173   0.379
## Q28  0.839  -0.106   0.349
## Q29  0.837  -0.132   0.260
## Q30 -0.131   0.614
## Q31 -0.201   0.687  -0.167
## Q32 -0.140   0.693  -0.141
## Q33          0.679  -0.156
## Q34          0.610
## Q35 -0.110   0.537  -0.216
##
##                Factor1 Factor2 Factor3
## SS loadings      2.906   2.687   2.410
## Proportion Var   0.208   0.192   0.172
## Cumulative Var   0.208   0.399   0.572
##
```

```
## Test of the hypothesis that 3 factors are sufficient.
## The chi square statistic is 151.4 on 52 degrees of freedom.
## The p-value is 1.2e-11

print(efa3$loadings, cut = 0.5)

##
## Loadings:
##     Factor1 Factor2 Factor3
## Q22                  0.595
## Q23                  0.743
## Q24                  0.715
## Q25                  0.633
## Q26  0.722
## Q27  0.671
## Q28  0.839
## Q29  0.837
## Q30          0.614
## Q31          0.687
## Q32          0.693
## Q33          0.679
## Q34          0.610
## Q35          0.537
##
##                Factor1 Factor2 Factor3
## SS loadings      2.906   2.687   2.410
## Proportion Var   0.208   0.192   0.172
## Cumulative Var   0.208   0.399   0.572
```

d) What rotation method is used in factanal as a default method? Explain what that rotation does?

Varimax Rotation is the default rotation method of factor analysis. Varimax rotation is used to simplify # the expression of a particular sub-space Where the actual coordinate system is unchanged, but it is the orthogonal basis that is being rotated to align with those coordinates.

e) Repeat part b (EFA with two factors) without rotation (inside factanal put rotation = "none"). Will you end up with the same names for your factors?

```
# Yes you will end up with the same names for the factors.

efa4 = factanal(covmat = Mpwd, factors = 2, n.obs = nrow(mydata2), rotation =
"none")
efa4

##
## Call:
```

```
## factanal(factors = 2, covmat = Mpwd, n.obs = nrow(mydata2), rotation = "no
ne")
##
## Uniquenesses:
##   Q22   Q23   Q24   Q25   Q26   Q27   Q28   Q29   Q30   Q31   Q32   Q33
Q34
## 0.530 0.512 0.582 0.614 0.289 0.355 0.206 0.289 0.625 0.470 0.483 0.506 0.
636
##   Q35
## 0.653
##
## Loadings:
##     Factor1 Factor2
## Q22  0.683
## Q23  0.698
## Q24  0.646
## Q25  0.620
## Q26  0.832   0.139
## Q27  0.792   0.134
## Q28  0.859   0.238
## Q29  0.817   0.209
## Q30 -0.353   0.500
## Q31 -0.487   0.542
## Q32 -0.429   0.577
## Q33 -0.363   0.602
## Q34 -0.222   0.561
## Q35 -0.395   0.436
##
##                Factor1 Factor2
## SS loadings      5.362   1.889
## Proportion Var   0.383   0.135
## Cumulative Var   0.383   0.518
##
## Test of the hypothesis that 2 factors are sufficient.
## The chi square statistic is 329.83 on 64 degrees of freedom.
## The p-value is 1.94e-37

print(efa4$loadings, cut = 0.5)

##
## Loadings:
##     Factor1 Factor2
## Q22  0.683
## Q23  0.698
## Q24  0.646
## Q25  0.620
## Q26  0.832
## Q27  0.792
## Q28  0.859
## Q29  0.817
```

```
## Q30           0.500
## Q31           0.542
## Q32           0.577
## Q33           0.602
## Q34           0.561
## Q35
##
##                  Factor1 Factor2
## SS loadings        5.362   1.889
## Proportion Var     0.383   0.135
## Cumulative Var     0.383   0.518
```