

**Instructions:**

Please prepare your submission in a document (Word or PDF) and clearly **label all answers and output** with their corresponding question number and part.

You should aim to answer questions concisely and focus only on the relevant aspects. Only include a figure if it is relevant to answering the question. A strict **maximum of one figure per question part** may be included. For example, a separate figure may be included for both Q1a and Q1b, but neither part can have more than one figure.

Any software output, such as figure, should be included in the document (either at the relevant question, or in an appendix). All such output **MUST** be referred to in your answer. For example, “The scatterplot in figure 1a shows a strong negative relationship between ...”. **Any software output that is not referred to in your answer should not be included.**

Take care to write any equations clearly and using a suitably sized font. You may consider using the equation editor in Word, but it is not required.

Any relevant R code should be included in an appendix. It **MUST** have documentation to indicate the purpose of the code, and be labelled with its corresponding question number and part.

## Question 1

A naturalist for the Alaska Fish and Game Department studies grizzly bears with the goal of maintaining a healthy population. An analysis of measurements on  $N = 61$  bears produced the following covariance matrix called **bears**:

```

      3266.46 1343.97  731.54 1175.50  162.68  238.37
      1343.97 324.25  179.28 281.17  39.15      117.73
      731.54  80.17  39.15  63.73   9.95      56.80
      1175.50 117.73  56.80  94.85  13.88      21.26
      162.68      94.85  13.88      21.26
      238.37      21.26

```

The variables are X1: weight (kg), X2: body length (cm), X3: neck circumference (cm), X4: girth (cm), X5: head length (cm) and X6: head width (cm).

- Read the data into R and explain why the matrix elements **bears**[2,3] and **bears**[3,2] are equal.
- Verify that **bears** is a covariance matrix.
- Conduct a principal component analysis on the covariance matrix, **bears**, and show your R code. Can the data be effectively summarised in fewer than six dimensions? Justify your answer.
- Based on the information provided in the question, and your results to part c, interpret the meaning of the first principal component.
- Which two variables have, respectively, the most and least influence on the value of first principal component? Justify your answer.

## Question2

The file **wine.csv** contains data on concentrations of 13 different chemicals in wines grown in the same region in Italy. They were derived from three different cultivars: a grape variety which has been selectively cultivated. The data were collected by the Italian Wine Association (IWA) and is available on Canvas.

You are contacted by the IWA and asked to analyse the data. They would like you to study the chemicals related to smell (`chem6`, `chem7`, `chem9`), taste (`chem3`, `chem5`) and alcohol content (`chem8`). The IWA inform you that each of these variables is as important as any other, and ask you to focus on summarising the information in this set of 6 variables by a smaller number of components.

Answer the following questions or complete the task.

- a. Prepare the dataset for the analysis by extracting the relevant variables and check for any missing values.
- b. Examine the mean, median, standard deviation, and range of the data. Which aspect of these descriptive statistics is particularly relevant for the analysis asked for by the IWA? Justify your answer.
- c. Produce a plot to visualise potential univariate outliers in the data that will be used as input to a principal component analysis (PCA) and comment on the key features seen in the results.
- d. Use the relevant Mahalanobis distance measure to identify potential multivariate outliers (do not plot them). How many are there? Based on the number of these potential outliers, should they be investigated further? Justify your answers.
- e. Generate a scatterplot matrix of the relevant data for the PCA such that there are scatterplots in one set of off-diagonal elements, correlation coefficients in the other set of off-diagonal elements, and histograms on the diagonal of the matrix. Briefly comment on what the plot shows for each of these three aspects **and** what, if any, their relevance is for a PCA.
- f. Assuming that the data have been checked and there is no need to remove outliers or make any other changes, conduct a PCA and decide how many PCs to retain in your results. Justify your answer.
- g. State the value of the second largest eigenvalue found in your analysis of part f. Interpret its meaning in a PCA.
- h. Write the retained PCs from part f in equation form as linear combinations of the original variables.
- i. To what extent do the first two PCs measure different aspects in the data? Justify your answer.

### Question 3

The dataset **household2.csv** contains responses to a questionnaire on different aspects of housing. In the nine questions, Q1, ..., Q9, participants were asked to rate their satisfaction with their living conditions on a scale from 1 to 7, with 7 indicating most satisfaction and 1 indicating most dissatisfaction. There was no missing data in the satisfaction data that was collected. The dataset is available on Canvas.

The primary goal of the analysis is to identify any dimensions that underly satisfaction with living conditions. A secondary goal is to use these results to measure satisfaction with living conditions in other similar populations or on other future occasions.

Answer the following questions or complete the task.

- a. Download the dataset and select the variables needed for an exploratory factor analysis (EFA) of satisfaction with living conditions. Explain why the variable `Gender` should not be included in the EFA.
- b. Based only on the problem and data description given above in the question notes (and not on any results), why would an EFA initially be preferred to a PCA to achieve the analysis goals?
- c. Show for all variables to be included in the EFA that their recorded values are valid based on the data description given above in the question notes.
- d. Assuming that there is no need to remove outliers, show that the data are suitable for an EFA, and decide how many factors should be investigated in your EFA. Justify your answers. Conduct an EFA on the correlation matrix with as many factors as you think are justified. Is simple structure achieved in the unrotated solution? Justify your answer.
- e. What is the value of the second largest eigenvalue for the analysis done in part d? Interpret its meaning in an EFA.
- f. Does a varimax rotation of the results found in part d improve the solution? Justify your answer.
- g. Compare the unrotated solution of part d with an oblimin rotation of the results. Which of these two would be the preferred solution to answer the analysis goals?
- h. Based on your chosen solution to part g, what would you recommend as a next step (you don't need to conduct this action) to achieve a solution with simple structure? Justify your answer.
- i. By inspecting the communalities seen with the results for part d, what is there to indicate that an EFA is a preferred analysis strategy to a PCA with these data? Justify your answer.

# Answers

```
# install package
#install.packages(c('psych'))

library(psych)
```

## Question 1

```
# reading covariance matrix into R
bears <- read.table('bears_covariance.txt', sep = ',', header = T)
```

a.

```
bears[2,3]; bears[3,2]
```

```
## [1] 324.25
```

```
## [1] 324.25
```

```
all.equal(bears[2,3], bears[3,2])
```

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

Both are equal because they are mirror images of themselves or symmetric to each other. For example, within the rows, the second variable is body length while the third variable along the columns is neck circumference (bears[2,3]). Also, along the rows, the third variable is neck circumference while the second variable along the columns is body length.

b.

```
# verifying that it is a covariance matrix
isSymmetric.matrix(unname(as.matrix(bears)))
```

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

The covariance above is a covariance matrix

c. Performing PCA

```

# PCA
# Getting the eigen values and vectors for the covariance matrix
pca <- eigen(bears, symmetric = T)

# get eigen values and vectors
eigen_vectors <- pca$vectors
eigen_values <- pca$values

# get the scaled variances
scaled_variance <- eigen_values/ sum(eigen_values)

scaled_variance

## [1] 0.9582856531 0.0326211775 0.0069159101 0.0017366358 0.0003245417
## [6] 0.0001160817

```

The data can be summarized in a fewer number of variables. From the PCA result above, the first principal component has the highest variance of about 95%. Hence, the first principal component (weights) can be used as the only variable since about 96% of the variance comes from its direction.

d.

The first principal component is where the variability in the data is most. From the result in part c, it accounts for about 96% of the variance in the data.

e.

```

eigen_vectors <- as.data.frame(eigen_vectors)
names(eigen_vectors) <- paste0('PC', 1:6)
row.names(eigen_vectors) <- names(bears)

print.data.frame(eigen_vectors)

##              PC1              PC2              PC3              PC4              PC5
## weight          0.84933911  0.47083174  0.22660616 -0.07426037  0.008692249
## body_length      0.36855162 -0.84607780  0.36813225 -0.01275430  0.110783709
## neck_circumference 0.19413158 -0.05812709 -0.30314314  0.92838758  0.012288753
## girth            0.31467769 -0.21674766 -0.84857608 -0.35506048  0.082353215
## head_length       0.04391809 -0.06035442 -0.00181524  0.06016158 -0.440118616
## head_width        0.06445825 -0.09202640 -0.03388032 -0.05226733 -0.887138069
##              PC6
## weight          0.0002016017
## body_length      0.0191050346
## neck_circumference 0.0705967427
## girth            -0.0326656761
## head_length       -0.8928053752
## head_width        0.4432635475

```

From above, the variable with the most influence in the first principal component is weight, while the variable with the least influence is head\_length

## Question 2

a.

```
wine <- read.csv('wine.csv', )

# extract relevant data
data <- wine[c('chem3', 'chem5', 'chem6', 'chem7', 'chem8', 'chem9')]

# checking for missing values
sapply(data, function(x) sum(is.na(x)))

## chem3 chem5 chem6 chem7 chem8 chem9
##      0      0      0      0      0      0
```

No missing values in any of the extracted variables

b.

```
get_summary_stat <- function(x){
  mean <- mean(x)
  median <- median(x)
  sd <- sd(x)
  range <- max(x) - min(x)
  c(mean=mean, median=median, sd=sd, range=range)
}

# get summary data
sapply(data, get_summary_stat)

##          chem3    chem5    chem6    chem7    chem8    chem9
## mean  2.366517 99.74157 2.295112 2.0292697 0.3618539 1.5908989
## median 2.360000 98.00000 2.355000 2.1350000 0.3400000 1.5550000
## sd     0.274344 14.28248 0.625851 0.9988587 0.1244533 0.5723589
## range  1.870000 92.00000 2.900000 4.7400000 0.5300000 3.1700000
```

The mean and standard deviation of the chemicals would be very relevant for the analysis asked by IWA. This is because, the mean will give information about the average concentration of chemicals in the wines grown within the same region in Italy. Similarly, the standard deviation will provide information about how the concentrations of these chemicals deviate from themselves.

c.

```
par(mfrow=c(2,3))
for (col in names(data)) boxplot(wine[col], main=col, ylab='Concentration')
```

The boxplot (figure 1) above shows the concentration distributions for smell, taste and alcohol chemicals. From the plot, chem9, a smell chemical has two values that are outliers. Also, for the taste chemicals, both have outliers in them, while for alcohol chemicals, there are no outliers seen in them.

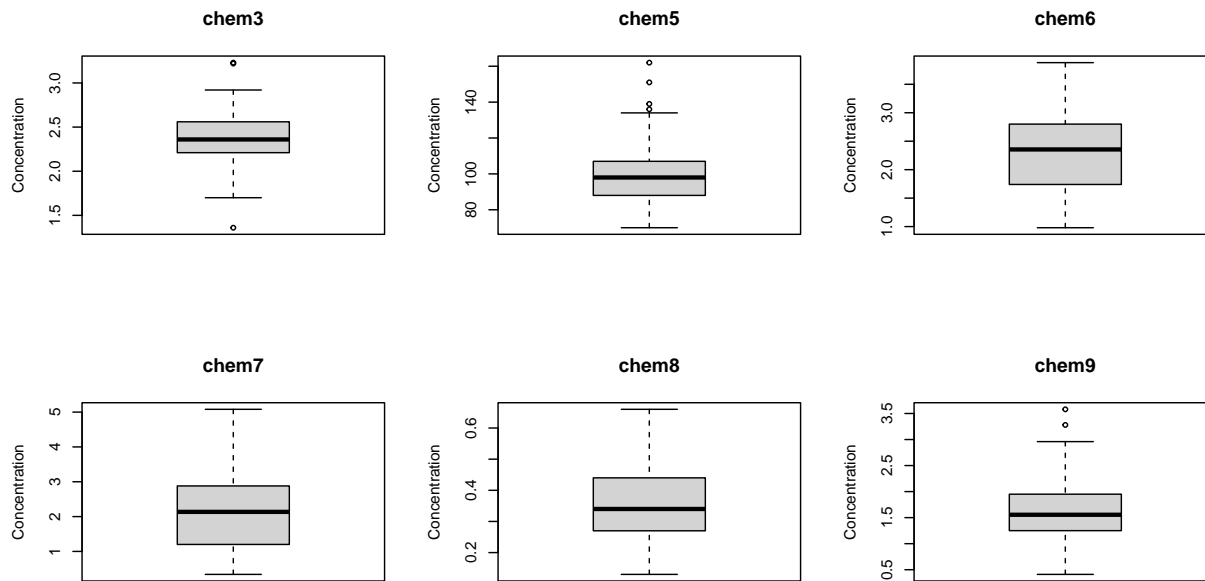


Figure 1: Figure 1: Checking for Outliers

#### d. Mahalanobis distance

```
m_obis <- mahalanobis(data, colMeans(data), cov(data))
pvalues <- pchisq(m_obis, df=ncol(data)-1, lower.tail = F)

# outliers are considered as values < 0.001
cat('Number of observations with outliers is: ', sum(pvalues < 0.001), '\n\n')
```

```
## Number of observations with outliers is: 5
```

```
# print the data points
print.data.frame(data[which(pvalues < 0.001),])
```

```
##      chem3 chem5 chem6 chem7 chem8 chem9
## 60    1.36    88    1.98    0.57    0.28    0.42
## 70    1.75   151    1.85    1.28    0.14    2.50
## 96    2.20   162    2.50    2.27    0.32    3.28
## 111   1.82   107    3.18    2.58    0.24    3.58
## 122   3.23   119    3.18    5.08    0.47    1.87
```

There are 5 observations that are outliers. They should not be investigated further since the number of observations found to contain outliers is only a small number (~2.8%) out of the 178 records.

e.



```
psych::pairs.panels(data, lm=FALSE, smooth=F, density=F, pch=20, hist.col='steelblue',
jiggle=F, ellipses=F, breaks=25, cex.cor=0.7, stars=T)
```

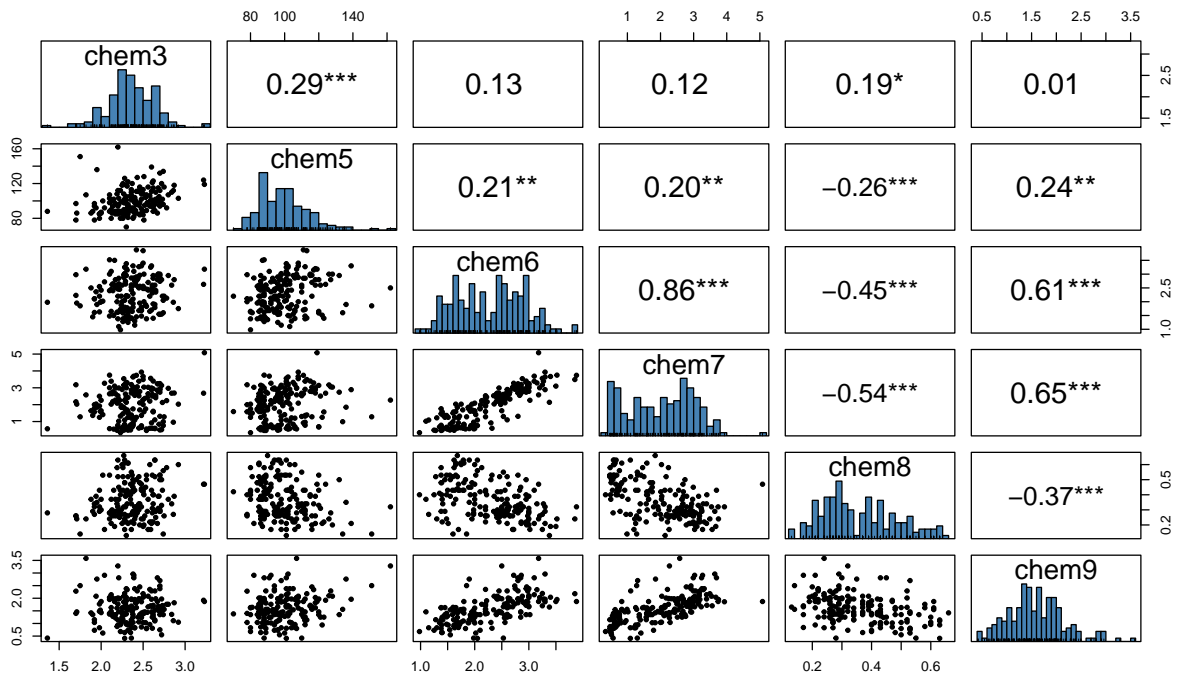


Figure 2: Figure 2: ScatterPlot Matrix

Figure 2, shows the scatterplot matrix of all the variables in the data. The lower matrix shows a scatterplot showing the relationship between one variable with the other. On the other side of the matrix is the correlation coefficient scores and their statistical significance (pvalues). The correlation coefficient is a numerical value that shows the linear relationship between one variable and another. Some of the values show that they are statistically significant at 5% threshold, while in some others, a relationship exists but it shows that these relationships are not statistically significant at 5% level. At the diagonal is the histogram chart showing the distribution of each of the variables (chemicals). From the above plot, there is a weak to strong relationship between one variable and another, except for chem 3 and chem 9.

f.

```
# performing PCA
pca <- prcomp(data, scale. = T)
summary(pca)
```

```
## Importance of components:
##              PC1      PC2      PC3      PC4      PC5      PC6
## Standard deviation  1.6994  1.1156  0.9260  0.73794  0.58473  0.35135
## Proportion of Variance 0.4813  0.2074  0.1429  0.09076  0.05699  0.02057
## Cumulative Proportion 0.4813  0.6887  0.8317  0.92244  0.97943  1.00000
```

```
pca$rotation
```

		PC1	PC2	PC3	PC4	PC5	PC6
##	chem3	-0.0683442	-0.799074970	0.3042731	-0.3363680	0.3840889	0.05965232
##	chem5	-0.2376977	-0.518833910	-0.7046603	0.2829717	-0.3051573	-0.06760970
##	chem6	-0.5241291	0.008255488	0.2788713	-0.1184179	-0.4972287	0.62144359
##	chem7	-0.5415729	0.052491399	0.2382900	-0.1673048	-0.2044400	-0.75985178
##	chem8	0.3945918	-0.292062832	0.5034399	0.5367731	-0.4440547	-0.14224943
##	chem9	-0.4638035	0.064547416	0.1512238	0.6903944	0.5226261	0.08982632

From the cumulative variance proportion, 5 principal components is sufficient to be retained since they can account for a combined 98% variance in the data.

g.

The second largest eigenvalue is the second principal component. It is the direction where the variance is the most after the first component. It is usually the direction which is not accounted by the first principal component and is orthogonal to it. From the value, the second eigenvalue accounts for about 21% of the variance in the data.

h.

$$y = PC_1 + PC_2 + PC_3 + PC_4 + PC_5$$

where,

$$PC_1 = -0.07 * X_{chem3} - 0.24 * X_{chem5} - 0.52 * X_{chem6} - 0.54 * X_{chem7} + 0.39 * X_{chem8} - 0.46 * X_{chem9}$$

$$PC_2 = -0.80 * X_{chem3} - 0.51 * X_{chem5} + 0.01 * X_{chem6} + 0.05 * X_{chem7} - 0.29 * X_{chem8} + 0.06 * X_{chem9}$$

$$PC_3 = 0.30 * X_{chem3} - 0.70 * X_{chem5} + 0.28 * X_{chem6} + 0.24 * X_{chem7} + 0.50 * X_{chem8} + 0.15 * X_{chem9}$$

$$PC_4 = -0.34 * X_{chem3} + 0.28 * X_{chem5} - 0.12 * X_{chem6} - 0.17 * X_{chem7} + 0.54 * X_{chem8} + 0.69 * X_{chem9}$$

$$PC_5 = 0.38 * X_{chem3} - 0.31 * X_{chem5} - 0.50 * X_{chem6} - 0.20 * X_{chem7} - 0.44 * X_{chem8} + 0.52 * X_{chem9}$$

i.

From the cumulative variance proportion, the first two PCs measure about 69% different aspects in the data.

### Question 3

a.

```
households <- read.csv('household2.csv')

# selecting relevant data
households <- households[paste0('Q', 1:9)]
```

Gender should not be included because it is nominal data and does not fit conceptually since the aim is to identify the different aspects of housing that appeal much to individuals, irrespective of gender.

**b.**

An EFA will be preferred over PCA because the aim is to find out if there are any dimensions that underlie satisfaction with living conditions.

**c.**

```
# Bartlett test for sphericity

cortest.bartlett(cor(households), n=nrow(households))

## $chisq
## [1] 11961.44
##
## $p.value
## [1] 0
##
## $df
## [1] 36

bartlett.test(households)

##
## Bartlett test of homogeneity of variances
##
## data: households
## Bartlett's K-squared = 22.343, df = 8, p-value = 0.004318
```

From the result of the Bartlett test for sphericity, the p-value is less than 0.05, hence we will reject the null hypothesis that there's no correlation amongst variables and conclude that at 5% significance level there is evidence to show that there's correlation between variables.

**d.**

```
# Kaiser-Meyer-Olkin factor adequacy
KMO(households)
```

```
## Kaiser-Meyer-Olkin factor adequacy
## Call: KMO(r = households)
## Overall MSA = 0.96
## MSA for each item =
##   Q1   Q2   Q3   Q4   Q5   Q6   Q7   Q8   Q9
## 0.96 0.96 0.95 0.96 0.96 0.96 0.96 0.97 0.96
```

From the Kaiser-Meyer-Olkin factor adequacy result, the overall MSA score is 0.96 showing very excellent suitability for EFA. Also from their individual MSA scores, they all have scores above 0.9. Therefore, all factors should be investigated for EFA.

e.

```
# fitting an EFA model
efa_model <- fa(households, nfactors = ncol(households), fm='minres', rotate = 'none')

efa_model
```

```
## Factor Analysis using method = minres
## Call: fa(r = households, nfactors = ncol(households), rotate = "none",
##       fm = "minres")
## Standardized loadings (pattern matrix) based upon correlation matrix
##   MR1   MR2   MR3   MR4   MR5   MR6   MR7   MR8   MR9   h2   u2 com
## Q1 0.93 -0.09 -0.12 0.08 0.00 -0.07 0.04 -0.01 0 0.90 0.095 1.1
## Q2 0.89 -0.16 0.18 0.09 0.02 -0.02 -0.01 0.03 0 0.87 0.133 1.2
## Q3 0.87 0.29 -0.03 -0.01 0.04 0.08 0.02 0.02 0 0.86 0.144 1.2
## Q4 0.86 0.25 0.15 0.07 -0.05 -0.02 0.00 -0.02 0 0.83 0.166 1.3
## Q5 0.92 -0.08 -0.15 0.01 -0.10 0.03 0.00 0.02 0 0.89 0.112 1.1
## Q6 0.94 0.08 0.00 -0.13 -0.05 -0.07 -0.02 0.00 0 0.92 0.082 1.1
## Q7 0.90 -0.15 0.12 -0.13 0.03 0.01 0.03 0.00 0 0.87 0.133 1.1
## Q8 0.90 0.06 -0.12 0.01 0.11 -0.04 -0.03 0.00 0 0.84 0.158 1.1
## Q9 0.91 -0.17 -0.01 0.01 0.00 0.09 -0.03 -0.03 0 0.87 0.131 1.1
##
##
## SS loadings      MR1 MR2 MR3 MR4 MR5 MR6 MR7 MR8 MR9
## Proportion Var   0.82 0.03 0.01 0.01 0.00 0.00 0.00 0.00 0.00
## Cumulative Var   0.82 0.85 0.86 0.86 0.87 0.87 0.87 0.87 0.87
## Proportion Explained 0.94 0.03 0.02 0.01 0.00 0.00 0.00 0.00 0.00
## Cumulative Proportion 0.94 0.97 0.98 0.99 1.00 1.00 1.00 1.00 1.00
##
## Mean item complexity = 1.1
## Test of the hypothesis that 9 factors are sufficient.
##
## The degrees of freedom for the null model are 36 and the objective function was 12.01 with Chi Sq
## The degrees of freedom for the model are -9 and the objective function was 0
##
## The root mean square of the residuals (RMSR) is 0
## The df corrected root mean square of the residuals is NA
##
## The harmonic number of observations is 1001 with the empirical chi square 0 with prob < NA
## The total number of observations was 1001 with Likelihood Chi Square = 0 with prob < NA
##
```

```
## Tucker Lewis Index of factoring reliability = 1.003
## Fit based upon off diagonal values = 1
## Measures of factor score adequacy
##
## Correlation of (regression) scores with factors    MR1  MR2  MR3  MR4  MR5
## Multiple R square of scores with factors          0.99 0.80 0.69 0.58 0.42
## Minimum correlation of possible factor scores      0.98 0.64 0.48 0.34 0.18
##                                                    0.97 0.28 -0.05 -0.32 -0.64
##
## Correlation of (regression) scores with factors    MR6  MR7  MR8 MR9
## Multiple R square of scores with factors          0.44 0.21 0.15 0
## Minimum correlation of possible factor scores      0.19 0.05 0.02 0
##                                                    -0.61 -0.91 -0.96 -1
```

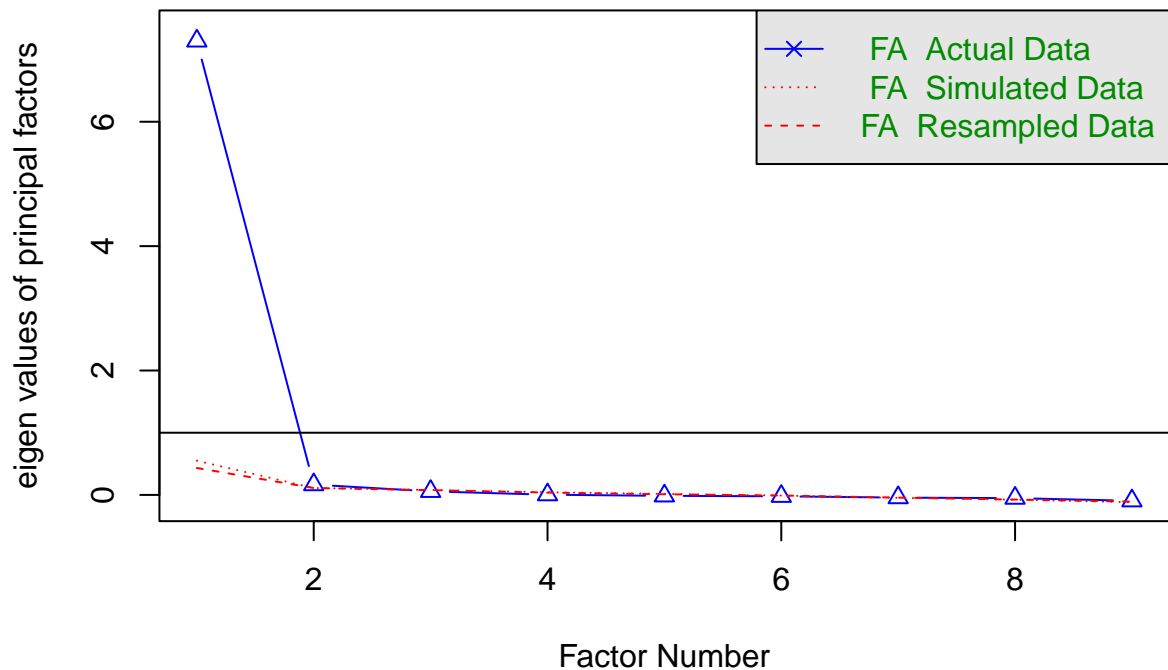
```
efa_model$loadings
```

```
##
## Loadings:
##   MR1  MR2  MR3  MR4  MR5  MR6  MR7  MR8  MR9
## Q1  0.932      -0.117
## Q2  0.895 -0.159 0.176
## Q3  0.873 0.291
## Q4  0.862 0.246 0.148
## Q5  0.921      -0.153
## Q6  0.942      -0.133
## Q7  0.901 -0.153 0.115 -0.132
## Q8  0.900      -0.117      0.105
## Q9  0.910 -0.173
##
##
##   MR1  MR2  MR3  MR4  MR5  MR6  MR7  MR8  MR9
## SS loadings  7.358 0.249 0.118 0.056 0.029 0.028 0.006 0.003 0.000
## Proportion Var 0.818 0.028 0.013 0.006 0.003 0.003 0.001 0.000 0.000
## Cumulative Var 0.818 0.845 0.858 0.865 0.868 0.871 0.872 0.872 0.872
```

From the proportion of variance, the optimal number of factor is 1, since it contributes about 82% of the variance. We will use the `fa.parallel` function to determine the optimal number of factors

```
# selecting optimal number of factor
set.seed(4)
fa.parallel(households, fa='fa')
```

## Parallel Analysis Scree Plots



```
## Parallel analysis suggests that the number of factors = 2 and the number of components = NA
```

Parallel analysis suggests that the number of factors is 2.

```
# refitting with 2 factors
# fitting an EFA model
efa_model <- fa(households, nfactors = 2, fm='minres', rotate = 'none')

efa_model$loadings
```

```
##
## Loadings:
##      MR1    MR2
## Q1  0.930
## Q2  0.890 -0.140
## Q3  0.876  0.317
## Q4  0.856  0.191
## Q5  0.918
## Q6  0.940
## Q7  0.897 -0.136
## Q8  0.899
## Q9  0.911 -0.178
##
##              MR1    MR2
## SS loadings  7.326 0.228
```

```
## Proportion Var 0.814 0.025
## Cumulative Var 0.814 0.839
```

From the loadings, a simple structure seems to have been achieved in the unrotated form, since all items load mostly on only one factor. For the second factor, the loadings of the items are typically less than 0.3, except in one item with less value.

f.

Based on the eigen values, the eigen values with the second largest value is 0.228. It gives the amount of variance contributed by each item to the second factor. From the loadings in e, we see that the variables contributing to the second factor include Q2, Q3, Q4, Q7 and Q9.

g.

```
# FA with varimax rotation
```

```
efa_model.varimax <- fa(cor(households), nfactors = 2, fm='minres', rotate = 'varimax')
print(efa_model.varimax)
```

```
## Factor Analysis using method = minres
## Call: fa(r = cor(households), nfactors = 2, rotate = "varimax", fm = "minres")
## Standardized loadings (pattern matrix) based upon correlation matrix
##      MR1  MR2  h2  u2 com
## Q1 0.76 0.54 0.87 0.13 1.8
## Q2 0.76 0.48 0.81 0.19 1.7
## Q3 0.45 0.81 0.87 0.13 1.6
## Q4 0.52 0.70 0.77 0.23 1.8
## Q5 0.74 0.55 0.85 0.15 1.8
## Q6 0.67 0.67 0.89 0.11 2.0
## Q7 0.77 0.48 0.82 0.18 1.7
## Q8 0.64 0.63 0.81 0.19 2.0
## Q9 0.81 0.46 0.86 0.14 1.6
##
##
##      MR1  MR2
## SS loadings      4.29 3.27
## Proportion Var    0.48 0.36
## Cumulative Var    0.48 0.84
## Proportion Explained 0.57 0.43
## Cumulative Proportion 0.57 1.00
##
## Mean item complexity = 1.8
## Test of the hypothesis that 2 factors are sufficient.
##
## The degrees of freedom for the null model are 36 and the objective function was 12.01
## The degrees of freedom for the model are 19 and the objective function was 0.16
##
## The root mean square of the residuals (RMSR) is 0.01
## The df corrected root mean square of the residuals is 0.02
##
```

```
## Fit based upon off diagonal values = 1
## Measures of factor score adequacy
##
## Correlation of (regression) scores with factors    MR1  MR2
## Multiple R square of scores with factors          0.90 0.87
## Minimum correlation of possible factor scores      0.82 0.77
##                                                    0.64 0.53
```

```
efa_model.varimax$loadings
```

```
##
## Loadings:
##      MR1  MR2
## Q1 0.762 0.542
## Q2 0.765 0.477
## Q3 0.455 0.813
## Q4 0.522 0.705
## Q5 0.740 0.547
## Q6 0.666 0.667
## Q7 0.768 0.484
## Q8 0.644 0.630
## Q9 0.805 0.462
##
##
##              MR1  MR2
## SS loadings    4.288 3.266
## Proportion Var 0.476 0.363
## Cumulative Var 0.476 0.839
```

Using the varimax rotation seems not to improve the solution. A simple structure seems not to be achieved since all the variables load on more than one factor.

h.

```
# comparison
# fitting an EFA model with Oblimin rotation
efa_model.oblimin <- fa(households, nfactors = 2, fm='minres', rotate = 'oblimin',
  n.obs = nrow(households))
```

```
## Loading required namespace: GPArotation
```

```
efa_model.oblimin
```

```
## Factor Analysis using method = minres
## Call: fa(r = households, nfactors = 2, n.obs = nrow(households), rotate = "oblimin",
##      fm = "minres")
## Standardized loadings (pattern matrix) based upon correlation matrix
##      MR1  MR2  h2  u2 com
## Q1 0.95 -0.06 0.87 0.13 1.0
## Q2 0.92 -0.11 0.81 0.19 1.0
## Q3 0.78 0.36 0.87 0.13 1.4
## Q4 0.79 0.23 0.77 0.23 1.2
```



```

## Q5 0.93 -0.04 0.85 0.15 1.0
## Q6 0.91 0.11 0.89 0.11 1.0
## Q7 0.93 -0.11 0.82 0.18 1.0
## Q8 0.87 0.09 0.81 0.19 1.0
## Q9 0.95 -0.15 0.86 0.14 1.0
##
##
##          MR1  MR2
## SS loadings      7.25 0.30
## Proportion Var    0.81 0.03
## Cumulative Var    0.81 0.84
## Proportion Explained 0.96 0.04
## Cumulative Proportion 0.96 1.00
##
## With factor correlations of
##          MR1  MR2
## MR1 1.00 0.24
## MR2 0.24 1.00
##
## Mean item complexity = 1.1
## Test of the hypothesis that 2 factors are sufficient.
##
## The degrees of freedom for the null model are 36 and the objective function was 12.01 with Chi Square = 12.01
## The degrees of freedom for the model are 19 and the objective function was 0.16
##
## The root mean square of the residuals (RMSR) is 0.01
## The df corrected root mean square of the residuals is 0.02
##
## The harmonic number of observations is 1001 with the empirical chi square 9.04 with prob < 0.97
## The total number of observations was 1001 with Likelihood Chi Square = 155.61 with prob < 1.8e-3
##
## Tucker Lewis Index of factoring reliability = 0.978
## RMSEA index = 0.085 and the 90 % confidence intervals are 0.073 0.097
## BIC = 24.34
## Fit based upon off diagonal values = 1
## Measures of factor score adequacy
##
##          MR1  MR2
## Correlation of (regression) scores with factors 0.99 0.80
## Multiple R square of scores with factors        0.98 0.63
## Minimum correlation of possible factor scores    0.96 0.27

```

```
efa_model.oblimin$loadings
```

```

##
## Loadings:
##          MR1  MR2
## Q1 0.947
## Q2 0.921 -0.111
## Q3 0.777 0.359
## Q4 0.793 0.229
## Q5 0.929
## Q6 0.911 0.106
## Q7 0.927 -0.107
## Q8 0.874
## Q9 0.953 -0.150

```

```
##
##              MR1   MR2
## SS loadings   7.205 0.252
## Proportion Var 0.801 0.028
## Cumulative Var 0.801 0.828
```

Based on the loadings, the oblimin rotation and the unrotated form seem to be the same. Most of the items load on the first factor, with items having less than 3 loading value loading on the second factor. The oblimin rotation technique and the unrotated technique seem to have simple structures.

i.

To achieve a simple structure, a one factor should be chosen and any of the oblimin rotation or unrotated techniques may be chosen.

j.

By investigating the communality results, EFA seems to be the best technique to use over PCA because, from the communality scores, we are able to understand the variance of each item that can be explained by the factors. For example, about 87% of the variance in Q1 can be explained by all the factors, 81% in Q2, 87% in Q3 and so on.