Assignment 1: Task 2 - EFA

Group 11

16.10.2018

0. Load Libraries

```
library(psych)
```

1. Problem Statement

The correlation matrix illustrates the information about anxiety symptoms which are regarded as the reflection to estimate the overall anxiety related to screening exercises for several commonly occurring medical conditions. A reasonable and accurate exploratory factor analysis can can be regarded as a method to identify a small number of common but unobservable factors and solve the initial questions. This paper will focus on the relationship between each factor and determine a reasonable number of factors that captures the information available in the data. A reasonable and accurate exploratory factor analysis can help researchers identify a small number of common but unobservable factors and solve the initial questions. The second section includes the description of the data and the visualisation of the correlation matrix. Then, the assumptions which can be applied to this case are demonstrated. Furthermore, exploratory factor analysis is implemented with fundamental algorithms of the method. The last section focuses on the interpretion the solution that explained the optimal factors.

2. Descriptive Statistics

There are some preparations of the data which are done before implementing exploratory factor analysis (EFA).

```
# Import data
corr <-read.delim("data/screening.txt", header = TRUE, sep="",dec = ".", skipNul = FALSE)
```

The first should be eliminated because it only contains X_name values. It cannnot be applied to the EFA as the results will be heavily biased.

```
# Eliminate first column
corr <- subset(corr, select = -c(X_name_))
```

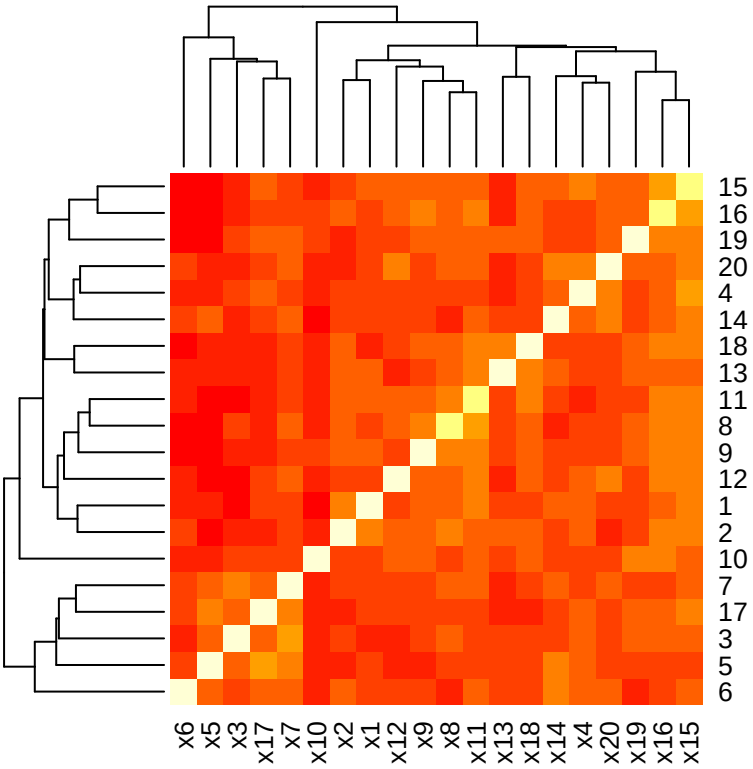
The following function processes the supplied lower correlation matrix into the desired format.

```
m <- matrix(NA,20,20)
m[lower.tri(m,diag=TRUE)] <- 1:10

makeSymm <- function(m) {
   m[upper.tri(m)] <- t(m)[upper.tri(m)]
   return(m)
}
```

```
corr <- makeSymm(corr)

corr <- makeSymm(corr)
corr<- as.matrix(corr)
heatmap(corr)
```



According to the heatmap, lighter colors represent a higher correlation between corresponding two variables. Variable x4 is highly correlated to variable x15, and the correlation between x5 and x10 is relatively weak. Most of the variables are equally high correlated.

3. Assumptions

When dealing with exploratory factor analysis, there is one fundamental assumption which should be satisfied, which is that the the factors are not correlated, implying an orthogonal model. Additionally, the data should be standardized, in others words, the variables have to be either mean centered and standardized. Since we are given a correlation matrix as input, this assumption can be regarded as true. Moreover, there should be a linear relationship between each variable used in this case. This means that the variables need to be at least reasonably related to each other, otherwise there will be no difference between the number of factors and the number of initial variables. For example, if there is no linear relationship between 20 variables in this case, exploratory factor analysis will be meaningless. Therefore, we assume that the data is standardized and there exists linear relationship between each variable.

```
# Perform Kaiser's MSA to evaluate appropriateness of data
KMO(corr)
```

```
## Kaiser-Meyer-Olkin factor adequacy
```

```
## Call: KMO(r = corr)
## Overall MSA =  0.95
## MSA for each item =
##   x1   x2   x3   x4   x5   x6   x7   x8   x9  x10  x11  x12  x13  x14  x15
## 0.95 0.94 0.94 0.95 0.89 0.95 0.95 0.96 0.97 0.97 0.96 0.96 0.95 0.96 0.96
##  x16  x17  x18  x19  x20
## 0.96 0.93 0.96 0.94 0.95
```

```
# Kaiser MSA = 0.95 > 0.8 --> appropriate data
```

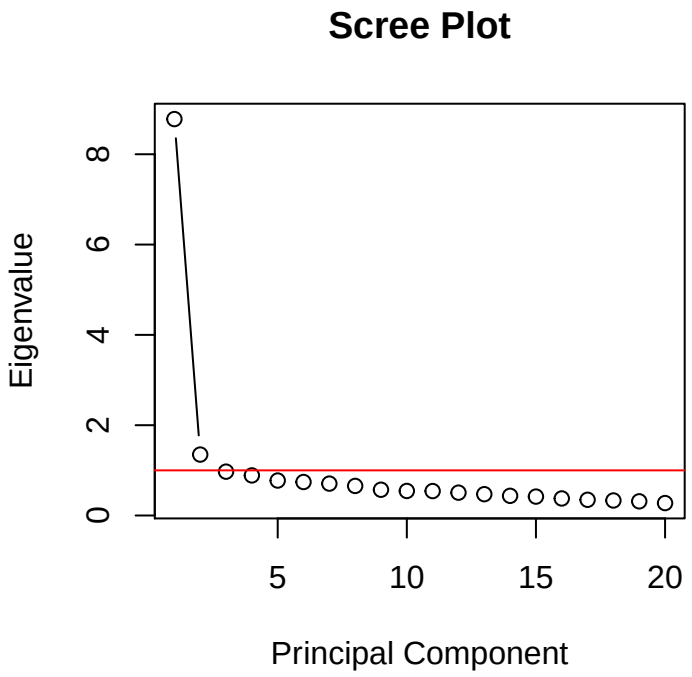
4. Method

There are different methods of obtaining a factor model, such as principal component factoring, iterative principal components factoring or the maximum likelihood method. First, the number of relevant factors to be extracted has to be determined. This can be inferred from the eigenvalues λ_i and eigenvectors ϵ_i of the observed correlation matrix R^{obs}. There are a number of rules of thumb which can be applied to the computed values:

- Retain only those factors with an eigenvalue larger than 1 (Guttman-Kaiser rule)
- Make a scree-plot and extract the amount of factors before the knee point of the slope
- Horn's parallel procedure

```
# Define the amount of factors
eval <- eigen(corr)$values

plot(eval, xlab = "Principal Component", ylab = "Eigenvalue",
     type = "b", main = "Scree Plot")
abline(h=1,col="red")
```



Considering the Scree Plot we should only retain the first factor. However, we also took the Kaiser-Guttman rule into account and decided to keep two factors. Factor analysis can be executed with different factoring methods to extract the latent variables. We used principal factoring and maximum likelihood. They both employ iterative approaches of estimating the correlation matrix from the observed Matrix. Principal factoring assumes that the intitial communalities are 1, meaning that there is no error at the starting point. In each iteration, these values then replace the diagonal in the correlation matrix which is used to recompute the set of factors. Maximum likelihood assumes a normal distribution of the dataset and iteratively adjusts distribution parameters to better fit the model to the observed data.
As there is an infinite number of different factoring solutions, rotations are applied to find the best possible interpretation of the model. For orthogonal models, which is one of our assumptions for this task, the most common procedures are varimax and quartimax. The latter focuses on identifing factor structure such that all variables have fairly high loadings on a few factors and have near zero loadings on the other factors. Varimax on the other hand tries to maximize the variance of loadings for each factor, such that every factor has high loadings on a few variables and low loadings for the other variables.

```
# Perform Factor Analysis
fa.out.ml <- fa(r = corr, nfactors = 2, fm="ml", rotate = "varimax", residuals = TRUE, SMC=FALSE,
                max.iter = 10)

fa.out.pa <- fa(r = corr, nfactors = 2, fm="pa", rotate = "varimax", residuals = TRUE, SMC=FALSE)
fa.out.pa
```

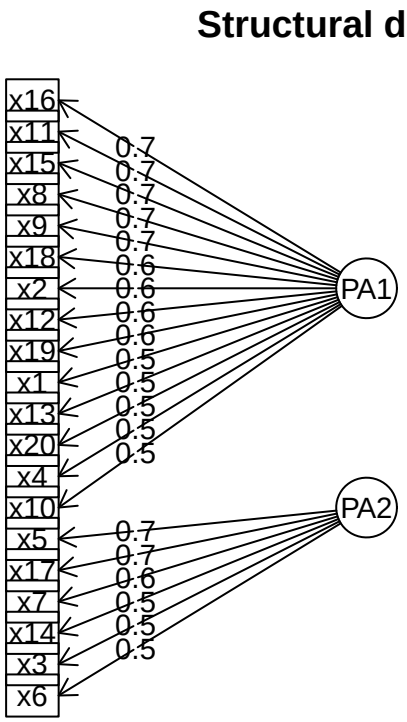
```
## Factor Analysis using method =  pa
## Call: fa(r = corr, nfactors = 2, rotate = "varimax", residuals = TRUE,
##     SMC = FALSE, fm = "pa")
## Standardized loadings (pattern matrix) based upon correlation matrix
##      PA1  PA2   h2   u2 com
## x1  0.54 0.31 0.39 0.61 1.6
## x2  0.60 0.24 0.41 0.59 1.3
## x3  0.29 0.51 0.35 0.65 1.6
## x4  0.47 0.47 0.44 0.56 2.0
## x5  0.12 0.70 0.51 0.49 1.1
## x6  0.28 0.45 0.28 0.72 1.7
## x7  0.38 0.64 0.55 0.45 1.6
## x8  0.67 0.26 0.51 0.49 1.3
## x9  0.66 0.22 0.48 0.52 1.2
## x10 0.45 0.22 0.25 0.75 1.5
## x11 0.70 0.28 0.58 0.42 1.3
## x12 0.58 0.31 0.43 0.57 1.5
## x13 0.52 0.30 0.36 0.64 1.6
## x14 0.40 0.53 0.45 0.55 1.9
## x15 0.68 0.42 0.64 0.36 1.7
## x16 0.73 0.28 0.61 0.39 1.3
## x17 0.31 0.65 0.52 0.48 1.4
## x18 0.62 0.28 0.47 0.53 1.4
## x19 0.57 0.33 0.43 0.57 1.6
## x20 0.51 0.42 0.44 0.56 1.9
##
##                        PA1  PA2
## SS loadings           5.60 3.48
## Proportion Var        0.28 0.17
## Cumulative Var        0.28 0.45
## Proportion Explained  0.62 0.38
## Cumulative Proportion 0.62 1.00
```

```
##
## Mean item complexity =  1.5
## Test of the hypothesis that 2 factors are sufficient.
##
## The degrees of freedom for the null model are  190  and the objective function was  9.5
## The degrees of freedom for the model are 151  and the objective function was  0.89
##
## The root mean square of the residuals (RMSR) is  0.04
## The df corrected root mean square of the residuals is  0.04
##
## Fit based upon off diagonal values = 0.99
## Measures of factor score adequacy
##                                                  PA1  PA2
## Correlation of (regression) scores with factors   0.91 0.86
## Multiple R square of scores with factors          0.83 0.73
## Minimum correlation of possible factor scores     0.66 0.47
```

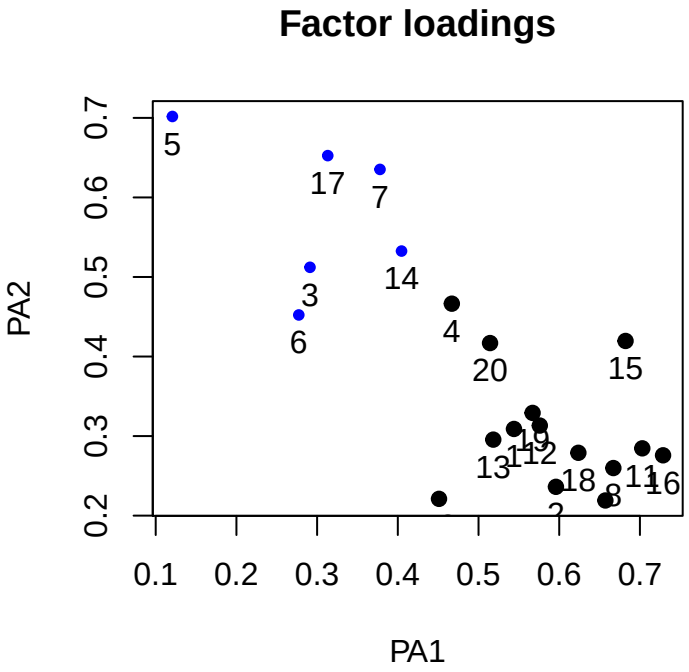
5. Interpretation of Solution

As it is the objective of EFA to explore the variable structure, the labeling of meaningful latent factors is ambigious. Maximum likelihood and principal factoring achieved very similar results, both generating the same root mean square of the residuals (RMSR) of 0.04 and the same variable groupings implied by the factor loadings. The RMSR measure indicates the sum of remaining error components, which is desired to be small. The factor model explained a cumulative variance of 0.45. However, the RMSR is the most important measure for the quality of the model.
In regard of rotation approaches, varimax provided the better possible factor explanation as 14 variables were loading higher on the first factor and 6 on the second. Quartimax resulted in all but one variable loading on the first factor, which we did not consider a useful split.
We thought of two possible interpretations of the structure. Variables loading strong on the first factor PA1 included statements such as "Lack of confidence during tests" (x1), "Heart beating fast during tests" (x18), "Screening bothers me" (x12) which we labeled as "self-confidence during test situations" or "high intense anxiety". The latent factor of the other group of variables (PA2) consisted of statements such as "Thinking about test results" (x3), "The harder I try to contain myself, the less assured I get" (x6), "Defeat myself during tests" (x14) and can be called "self-manipulative thoughts during test situations" or "low intense anxiety".
A full structural diagram displaying the variable groupings can be seen below.

```
# Structural diagram
fa.diagram(fa.out.pa, main = "Structural diagram")
```




```
# Plot of factor loadings
plot(fa.out.pa, title = "Factor loadings")
```



From the factor loadings plot we can graphically verify that the varibles load higher on one of the factors and very few (x4 and x20) share similarly high laodings with both. We can conclude that this factor model captures useful information about the structure of the measured variables. Specifically one group of variables shares the component PA1 of "self-confidence during test situations" or "high intense anxiety" and the other one can be regarded as "self-manipulative thoughts during test situations" or "low intense anxiety".