

Assignment 1: EFA

Group 11

The Date

1. Problem Statement Task 2

exploratory factor analysis: • explain the correlation structure among observed variables • try to find underlying dimensions that can explain the observed correlations • example: the correlation between scores on mathematics, statistics and physics exams can be explained because they all measure somehow quantitative intelligence

1. State the problem
2. Descriptive Statistics (to check data, to find outliers)
3. Test (or at least state) the assumptions of the method, if any
4. Conduct the method (describe in more detail the “best” approach you have found)
5. Interpret the solution
6. Compare the results briefly with alternative solutions, if any
7. Conclusion

2. Descriptive Statistics

```
library(psych)

corr <- read.delim("data/screening.txt", header = TRUE, sep=" ", dec = ".", skipNul = FALSE)
corr <- subset(corr, select = -c(X_name_))

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)
```

3. Assumptions of Methods

assuming standardized data and factors + uncorrelated factors

```
# 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

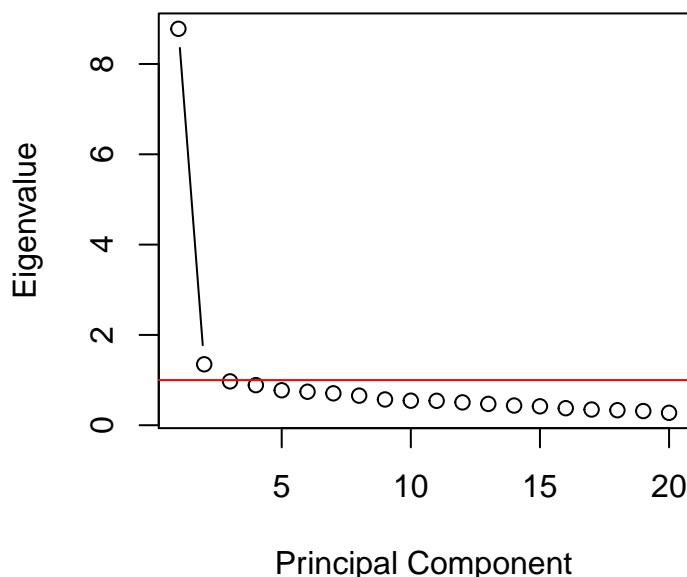
There are different methods of obtaining a factor model, such as principal component factoring, iterative principal components factoring or the maximum likelihood method. First, the number of relevant factors to be extracted has to be determined. This can be inferred from the eigenvalues λ_i and eigenvectors 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)
- Horn's parallel procedure
- Make a scree-plot and extract the amount of factors before the knee point of the slope

```
# 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")
```

Scree Plot



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 initial 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 identifying 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
##
## Correlation of (regression) scores with factors      PA1  PA2
## Multiple R square of scores with factors            0.91 0.86
## Minimum correlation of possible factor scores        0.83 0.73
##                                                       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 ambiguous. 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.

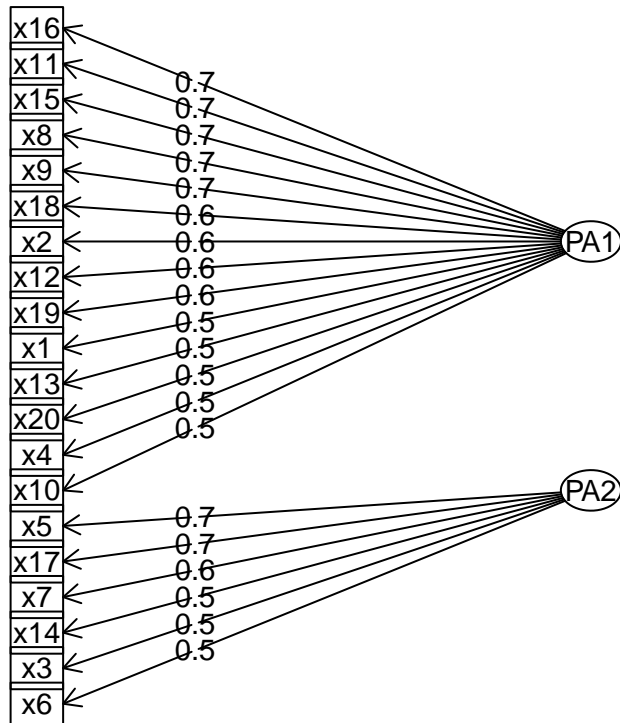
In regard of rotation approaches, varimax provided the better possible factor explanation as 14 variables were loading higher on the first factor and the rest on the second. Quartimax resulted in all but one variable loading on the first factor, which we did not consider a useful split.

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.

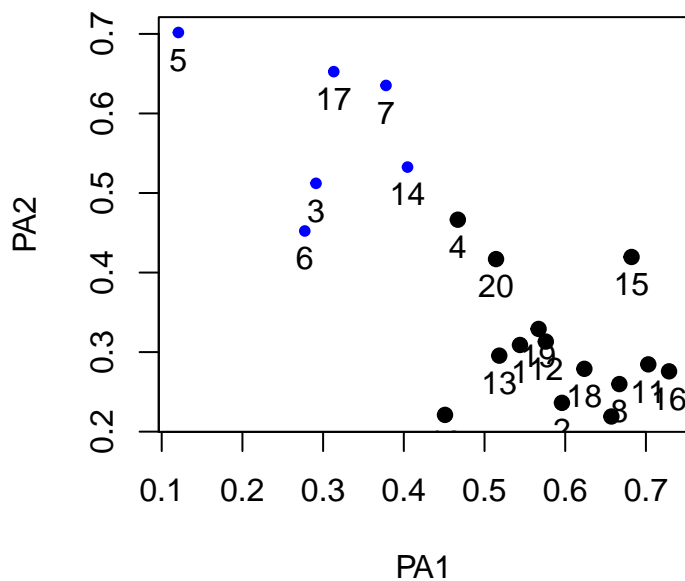
```
# Structural diagram
fa.diagram(fa.out.pa, main = "Structural diagram")
```

Structural diagram



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

Factor loadings



From the factor loadings plot we can graphically verify that the variables load higher on one of the factors and very few (x4 and x20) share similarly high loadings 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”.