28/2/2022

Exploratory Factor Analysis

AM: P3622004

Mihalis Galanakis

Exploratory Factor Analysis

Mihalis Galanakis

28/2/2022

Objective

The data in the folder named data.txt refer to counts from different variables in a population of women, aiming to gain useful insights and explore different ways to conduct exploratory factor analysis.

Read in the data

```
data <- read.table('C:/Users/mihal/OneDrive/data.txt',sep=",")

# Statistical Learning Project
# The following Dataset involves predicting the onset of diabetes withi
n 5 years in a women population given medical details.
# It is a binary (2-class) classification problem. The number of observ
ations for each class is not balanced.
# There are 768 observations with 8 input variables and 1 output variab
le. Missing values are believed to be encoded with zero values.
# The variable names are as follows:

# Number of times pregnant.
# Plasma glucose concentration a 2 hours in an oral glucose tolerance t
est.
# Diastolic blood pressure (mm Hg).
# Triceps skinfold thickness (mm).
# 2-Hour serum insulin (mu U/ml).
# Body mass index (weight in kg/(height in m)^2).
# Diabetes pedigree function.
# Age (years).
# Class variable (0 or 1).
```

Descriptive statistics

```
str(data)
```

```
## 'data.frame':    768 obs. of  9 variables:
##  $ V1: int  6 1 8 1 0 5 3 10 2 8 ...
##  $ V2: int  148 85 183 89 137 116 78 115 197 125 ...
##  $ V3: int  72 66 64 66 40 74 50 0 70 96 ...
##  $ V4: int  35 29 0 23 35 0 32 0 45 0 ...
##  $ V5: int  0 0 0 94 168 0 88 0 543 0 ...
##  $ V6: num  33.6 26.6 23.3 28.1 43.1 25.6 31 35.3 30.5 0 ...
##  $ V7: num  0.627 0.351 0.672 0.167 2.288 ...
##  $ V8: int  50 31 32 21 33 30 26 29 53 54 ...
##  $ V9: int  1 0 1 0 1 0 1 0 1 1 ...
```

```
dim(data)
```

```
## [1] 768   9
```

```
summary(data)
```

```
##        V1               V2              V3               V4        
##  Min.   : 0.000   Min.   :  0.0   Min.   :  0.00   Min.   : 0.00  
##  1st Qu.: 1.000   1st Qu.: 99.0   1st Qu.: 62.00   1st Qu.: 0.00  
##  Median : 3.000   Median :117.0   Median : 72.00   Median :23.00  
##  Mean   : 3.845   Mean   :120.9   Mean   : 69.11   Mean   :20.54  
##  3rd Qu.: 6.000   3rd Qu.:140.2   3rd Qu.: 80.00   3rd Qu.:32.00  
##  Max.   :17.000   Max.   :199.0   Max.   :122.00   Max.   :99.00  
##        V5              V6              V7               V8        
##  Min.   :  0.0   Min.   : 0.00   Min.   :0.0780   Min.   :21.00  
##  1st Qu.:  0.0   1st Qu.:27.30   1st Qu.:0.2437   1st Qu.:24.00  
##  Median : 30.5   Median :32.00   Median :0.3725   Median :29.00  
##  Mean   : 79.8   Mean   :31.99   Mean   :0.4719   Mean   :33.24  
##  3rd Qu.:127.2   3rd Qu.:36.60   3rd Qu.:0.6262   3rd Qu.:41.00  
##  Max.   :846.0   Max.   :67.10   Max.   :2.4200   Max.   :81.00  
##        V9        
##  Min.   :0.000  
##  1st Qu.:0.000  
##  Median :0.000  
##  Mean   :0.349  
##  3rd Qu.:1.000  
##  Max.   :1.000
```

```
colnames(data) <- c("t.pregnant","plasma","bl.press","tr.thick","serum.
ins","bmi","diab","age","class")
head(data,5)
```

```
##   t.pregnant plasma bl.press tr.thick serum.ins  bmi  diab age class
## 1          6    148       72       35         0 33.6 0.627  50     1
## 2          1     85       66       29         0 26.6 0.351  31     0
## 3          8    183       64        0         0 23.3 0.672  32     1
## 4          1     89       66       23        94 28.1 0.167  21     0
## 5          0    137       40       35       168 43.1 2.288  33     1
```

```
tail(data,5)
```

```
##     t.pregnant plasma bl.press tr.thick serum.ins  bmi  diab age cla
ss
## 764         10    101       76       48       180 32.9 0.171  63
0
## 765          2    122       70       27         0 36.8 0.340  27
0
## 766          5    121       72       23       112 26.2 0.245  30
0
## 767          1    126       60        0         0 30.1 0.349  47
1
```

```
## 768          1     93       70       31         0 30.4 0.315  23
0

class <- data$class
t.pregnant <- data$t.pregnant


expl_data <- data[,2:8]
# We assume that the zeros in the variable times.pregnant are not missi
ng, hence we don't replace zeros with "NA"

head(expl_data,10)

##    plasma bl.press tr.thick serum.ins  bmi  diab age
## 1     148       72       35         0 33.6 0.627  50
## 2      85       66       29         0 26.6 0.351  31
## 3     183       64        0         0 23.3 0.672  32
## 4      89       66       23        94 28.1 0.167  21
## 5     137       40       35       168 43.1 2.288  33
## 6     116       74        0         0 25.6 0.201  30
## 7      78       50       32        88 31.0 0.248  26
## 8     115        0        0         0 35.3 0.134  29
## 9     197       70       45       543 30.5 0.158  53
## 10    125       96        0         0  0.0 0.232  54

expl_data[expl_data==0] <- NA
# Replace the zeros with "NA"

df <- data.frame(t.pregnant,expl_data,class)
# Create the transformed dataframe
head(df,5)

##   t.pregnant plasma bl.press tr.thick serum.ins  bmi  diab age class
## 1          6    148       72       35        NA 33.6 0.627  50     1
## 2          1     85       66       29        NA 26.6 0.351  31     0
## 3          8    183       64       NA        NA 23.3 0.672  32     1
## 4          1     89       66       23        94 28.1 0.167  21     0
## 5          0    137       40       35       168 43.1 2.288  33     1

# View the first 5 obs of the df
tail(df,5)

##     t.pregnant plasma bl.press tr.thick serum.ins  bmi  diab age cla
ss
## 764         10    101       76       48       180 32.9 0.171  63
0
## 765          2    122       70       27        NA 36.8 0.340  27
0
## 766          5    121       72       23       112 26.2 0.245  30
0
## 767          1    126       60       NA        NA 30.1 0.349  47
```

```
1
## 768          1     93       70       31        NA 30.4 0.315  23
0

# View the last 5 obs of the df
dim(df)

## [1] 768   9
```

Since we are not interested in using an imputed dataset we omit the missing values

```
newdf <- na.omit(df)
# Create the final dataframe that doesn't include missing values!
```

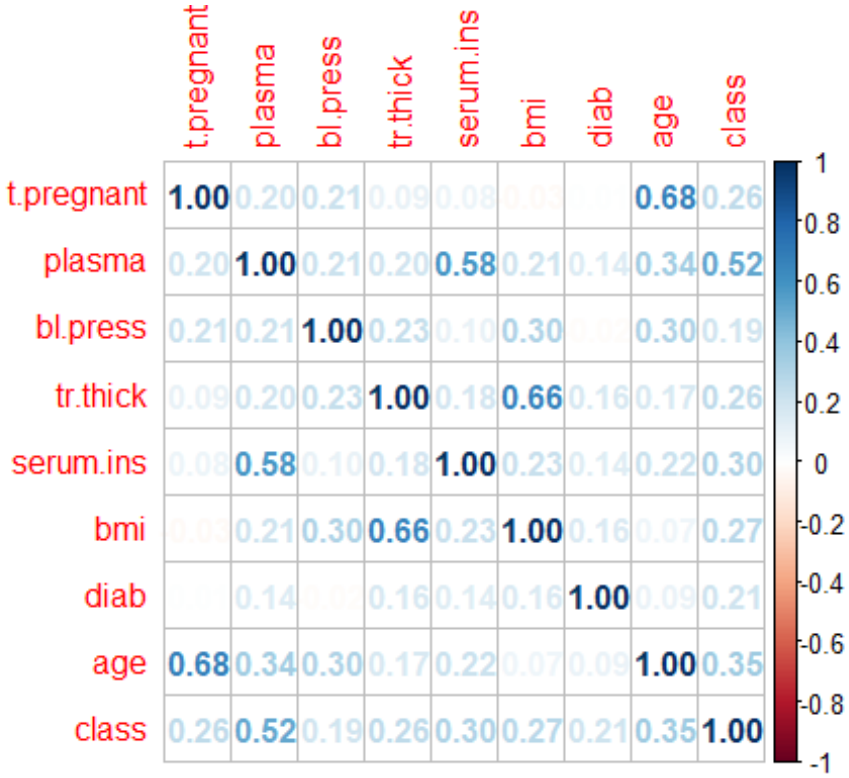
Getting insights about our dataset named newdf

```
head(newdf)

##    t.pregnant plasma bl.press tr.thick serum.ins  bmi  diab age clas
s
## 4           1     89       66       23        94 28.1 0.167  21
0
## 5           0    137       40       35       168 43.1 2.288  33
1
## 7           3     78       50       32        88 31.0 0.248  26
1
## 9           2    197       70       45       543 30.5 0.158  53
1
## 14          1    189       60       23       846 30.1 0.398  59
1
## 15          5    166       72       19       175 25.8 0.587  51
1

str(newdf)

## 'data.frame':    392 obs. of  9 variables:
##  $ t.pregnant: int  1 0 3 2 1 5 0 1 1 3 ...
##  $ plasma    : int  89 137 78 197 189 166 118 103 115 126 ...
##  $ bl.press  : int  66 40 50 70 60 72 84 30 70 88 ...
##  $ tr.thick  : int  23 35 32 45 23 19 47 38 30 41 ...
##  $ serum.ins : int  94 168 88 543 846 175 230 83 96 235 ...
##  $ bmi       : num  28.1 43.1 31 30.5 30.1 25.8 45.8 43.3 34.6 39.3
...
##  $ diab      : num  0.167 2.288 0.248 0.158 0.398 ...
##  $ age       : int  21 33 26 53 59 51 31 33 32 27 ...
##  $ class     : int  0 1 1 1 1 1 1 0 1 0 ...
##  - attr(*, "na.action")= 'omit' Named int [1:376] 1 2 3 6 8 10 11 12
13 16 ...
##   ..- attr(*, "names")= chr [1:376] "1" "2" "3" "6" ...

summary(newdf)
```

```
##    t.pregnant         plasma         bl.press        tr.thick
##  Min.   : 0.000   Min.   : 56.0   Min.   : 24.00   Min.   : 7.00
##  1st Qu.: 1.000   1st Qu.: 99.0   1st Qu.: 62.00   1st Qu.:21.00
##  Median : 2.000   Median :119.0   Median : 70.00   Median :29.00
##  Mean   : 3.301   Mean   :122.6   Mean   : 70.66   Mean   :29.15
##  3rd Qu.: 5.000   3rd Qu.:143.0   3rd Qu.: 78.00   3rd Qu.:37.00
##  Max.   :17.000   Max.   :198.0   Max.   :110.00   Max.   :63.00
##    serum.ins           bmi             diab             age
##  Min.   : 14.00   Min.   :18.20   Min.   :0.0850   Min.   :21.00
##  1st Qu.: 76.75   1st Qu.:28.40   1st Qu.:0.2697   1st Qu.:23.00
##  Median :125.50   Median :33.20   Median :0.4495   Median :27.00
##  Mean   :156.06   Mean   :33.09   Mean   :0.5230   Mean   :30.86
##  3rd Qu.:190.00   3rd Qu.:37.10   3rd Qu.:0.6870   3rd Qu.:36.00
##  Max.   :846.00   Max.   :67.10   Max.   :2.4200   Max.   :81.00
##      class
##  Min.   :0.0000
##  1st Qu.:0.0000
##  Median :0.0000
##  Mean   :0.3316
##  3rd Qu.:1.0000
##  Max.   :1.0000
```

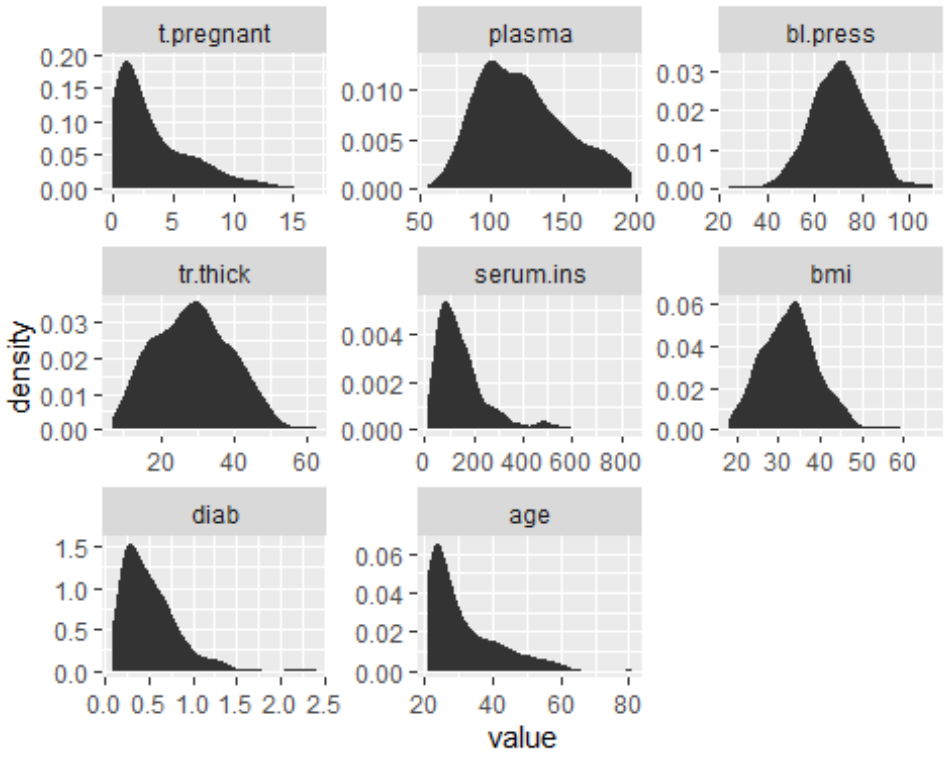
Heatmap of the correlations of our dataset

```
library(corrplot)

## corrplot 0.92 loaded

new.corrmatrix <- cor(newdf)
corrplot(new.corrmatrix, method = 'number')
```



Normality check on dataset newdf

```
library(reshape2)
library(ggplot2)
df1 <- melt(newdf[,-9])

## No id variables; using all as measure variables

ggplot(data = df1, aes(x = value)) +
  stat_density() +
  facet_wrap(~variable, scales = "free")
```




```
# Small multiple chart
# Not a great picture overall as regards normality, probably a transfor
mation would be a more appropriate choice
```

Extracting the correlations of our dataset

```
cor.data <- cor(newdf[,-9])
cor.data
```

```
##              t.pregnant    plasma   bl.press  tr.thick  serum.ins
bmi
## t.pregnant  1.000000000 0.1982910  0.2133548 0.0932094 0.07898363 -0
.02534728
## plasma      0.198291043 1.0000000  0.2100266 0.1988558 0.58122301  0
.20951592
## bl.press    0.213354775 0.2100266  1.0000000 0.2325712 0.09851150  0
.30440337
## tr.thick    0.093209397 0.1988558  0.2325712 1.0000000 0.18219906  0
.66435487
## serum.ins   0.078983625 0.5812230  0.0985115 0.1821991 1.00000000  0
.22639652
## bmi        -0.025347276 0.2095159  0.3044034 0.6643549 0.22639652  1
.00000000
## diab        0.007562116 0.1401802 -0.0159711 0.1604985 0.13590578  0
.15877104
## age         0.679608470 0.3436415  0.3000389 0.1677611 0.21708199  0
.06981380
```

```
##                    diab        age
## t.pregnant  0.007562116 0.67960847
## plasma      0.140180180 0.34364150
## bl.press   -0.015971104 0.30003895
## tr.thick    0.160498526 0.16776114
## serum.ins   0.135905781 0.21708199
## bmi         0.158771043 0.06981380
## diab        1.000000000 0.08502911
## age         0.085029106 1.00000000

# The correlation matrix of dataset called newdf without the 9th variab
le, which is the class (0 refers to
# women that won't develop diabetes, 1 refers to women that will develo
p diabetes)
```

Correlations check (both for partial and simple correlations)

```
library(psych)

##
## Attaching package: 'psych'

## The following objects are masked from 'package:ggplot2':
##
##     %+%, alpha

KMO(cor.data)

## Kaiser-Meyer-Olkin factor adequacy
## Call: KMO(r = cor.data)
## Overall MSA =  0.62
## MSA for each item =
## t.pregnant     plasma   bl.press   tr.thick  serum.ins        bmi
diab
##       0.56       0.66       0.75       0.61       0.62       0.58
0.76
##        age
##       0.61

# Kaiser provided the following values for interpreting results:
# * 0.00 to 0.49 unacceptable
# * 0.50 to 0.59 miserable
# * 0.60 to 0.69 mediocre
# * 0.70 to 0.79 middling
# * 0.80 to 0.89 meritorious
# * 0.90 to 1.00 marvelous

# KMO is a measure of sampling adequacy
# We observe that Overall MSA is 0.62, which is mediocre to satisfactor
y!
# We also notice that MSA for t.pregnant variable is 0.56, which is acc
```

```
eptable but we will remove this variable in order to improve the fit
# The same applies for variable bmi since we observe that MSA for bmi i
s 0.58
new.cor.data <- cor(newdf[,-c(1,6,9)])
new.cor.data

##              plasma   bl.press  tr.thick serum.ins        diab
age
## plasma    1.0000000  0.2100266 0.1988558 0.5812230  0.14018018 0.343
64150
## bl.press  0.2100266  1.0000000 0.2325712 0.0985115 -0.01597110 0.300
03895
## tr.thick  0.1988558  0.2325712 1.0000000 0.1821991  0.16049853 0.167
76114
## serum.ins 0.5812230  0.0985115 0.1821991 1.0000000  0.13590578 0.217
08199
## diab      0.1401802 -0.0159711 0.1604985 0.1359058  1.00000000 0.085
02911
## age       0.3436415  0.3000389 0.1677611 0.2170820  0.08502911 1.000
00000

KMO(new.cor.data)

## Kaiser-Meyer-Olkin factor adequacy
## Call: KMO(r = new.cor.data)
## Overall MSA =  0.64
## MSA for each item =
##    plasma  bl.press  tr.thick serum.ins      diab       age
##      0.61      0.64      0.72      0.60      0.67      0.72

# Slightly improved the Overall MSA, valued at 0.64, which is more dece
nt



cortest.bartlett(new.cor.data, n=392)

## $chisq
## [1] 306.7945
##
## $p.value
## [1] 2.161594e-56
##
## $df
## [1] 15

# Bartlett's test compares the correlation matrix to an identity matrix
(a matrix filled with zeroes).
# We observe that p_value ~= 0 hence we reject the null hypothesis
# Thus, the correlation matrix is not equal to the idd and we can move
on to the factor analysis
```

 # Factor analysis with k=1 (one factor)

```
fit <- factanal(x=newdf[,-c(1,6,9)] ,factors = 1)
fit

##
## Call:
## factanal(x = newdf[, -c(1, 6, 9)], factors = 1)
##
## Uniquenesses:
##    plasma  bl.press  tr.thick serum.ins      diab       age
##     0.273     0.932     0.927     0.557     0.968     0.835
##
## Loadings:
##           Factor1
## plasma    0.853
## bl.press  0.261
## tr.thick  0.270
## serum.ins 0.665
## diab      0.180
## age       0.407
##
##                Factor1
## SS loadings      1.508
## Proportion Var   0.251
##
## Test of the hypothesis that 1 factor is sufficient.
## The chi square statistic is 48.27 on 9 degrees of freedom.
## The p-value is 2.28e-07

# We desire the communalities to be as high as possible (Communalities
refer to the complementary of the Uniqueness)
# thus, We want the uniqueness to be as low as possible
# We observe that the Uniqueness in all variables other than plasma is
extremely high so we gather that the
# percentage of the variability of the correlations (of the 6 variables
) explained by the factor1 is extremely low! Indications of a poor fit!
# The proportion of the total variance explained is 25.1%, which is ext
remely low. Extra indications that we need to consider more factors!
# p_value ~= 0 hence we reject the null hypothesis, therefore the fit i
s definately not good!
# Last but not least the SS loadings refer to the sum of squares of the
loadings of factor 1
```

Factor analysis with k=2 (two factors)

```
fit_2 <- factanal(x=newdf[,-c(1,6,9)] ,factors = 2)
fit_2

##
## Call:
```

```
## factanal(x = newdf[, -c(1, 6, 9)], factors = 2)
##
## Uniquenesses:
##    plasma  bl.press  tr.thick serum.ins      diab       age
##     0.283     0.148     0.897     0.527     0.962     0.793
##
## Loadings:
##           Factor1 Factor2
## plasma     0.797   0.286
## bl.press           0.920
## tr.thick   0.181   0.265
## serum.ins  0.670   0.155
## diab       0.194
## age        0.292   0.349
##
##                Factor1 Factor2
## SS loadings      1.245   1.145
## Proportion Var   0.207   0.191
## Cumulative Var   0.207   0.398
##
## Test of the hypothesis that 2 factors are sufficient.
## The chi square statistic is 10.25 on 4 degrees of freedom.
## The p-value is 0.0365

# We observe that the fit in this case is closer to being acceptable th
an before (p_value 0.0365)
# We also observe some improvements in Uniqueness (in bl.press the grea
test change!)
# In this case the cumulative proportion of variance is equal to 39.8%,
which is not great at all, yet it's better than the 25.1% we had before
!
```

Factor analysis with k=3 (three factors)

```
fit_3 <-  factanal(x=newdf[,-c(1,6,9)] ,factors = 3)
fit_3

##
## Call:
## factanal(x = newdf[, -c(1, 6, 9)], factors = 3)
##
## Uniquenesses:
##    plasma  bl.press  tr.thick serum.ins      diab       age
##     0.008     0.005     0.743     0.632     0.866     0.809
##
## Loadings:
##           Factor1 Factor2 Factor3
## plasma     0.980   0.124   0.129
## bl.press           0.993
## tr.thick   0.118   0.208   0.447
## serum.ins  0.557           0.237
```

```
## diab       0.102           0.350
## age        0.293   0.271   0.177
##
##                Factor1 Factor2 Factor3
## SS loadings      1.387   1.122   0.428
## Proportion Var   0.231   0.187   0.071
## Cumulative Var   0.231   0.418   0.489
##
## The degrees of freedom for the model is 0 and the fit was 0

# The last section of the function output shows the results of a hypoth
esis test. The null hypothesis, H0, is that
# the number of factors in the model, in our example 2 factors, is suff
icient to capture the full dimensionality of
# the data set. Conventionally, we reject H0 if the p-value is less tha
n 0.05. Such a result indicates that the
# number of factors is too small. In contrast, we do not reject H0 if t
he p-value exceeds 0.05. Such a result
# indicates that there are likely enough (or more than enough) factors
capture the full dimensionality of the dataset
# (Teetor 2011). The high p-value in our example above leads us to not
reject the H0, and indicates that we fitted
# an appropriate model. This hypothesis test is available thanks to our
method of estimation, maximum likelihood.
# Notice there is no entry for certain variables. That is because R doe
s not print loadings less than 0.1.
# This is meant to help us spot groups of variables.
# Definitely not 3 factors!

apply(fit_3$loadings^2,1,sum)

##    plasma  bl.press  tr.thick serum.ins      diab       age
## 0.9918226 0.9950000 0.2568286 0.3679858 0.1340721 0.1907028

# Another way to calculate the Communalities

1-apply(fit_3$loadings^2,1,sum)

##      plasma    bl.press    tr.thick   serum.ins        diab
age
## 0.008177356 0.004999987 0.743171372 0.632014214 0.865927941 0.809297
231

# Another way to calculate the Uniqueness

scores1 <- factor.scores(newdf[,-c(1,6,9)], fit)$scores
head(scores1)

##       Factor1
## 4  -1.0254369
## 5   0.4334774
```

```
## 7  -1.2322724
## 9   2.6234436
## 14  2.9987084
## 15  1.1000477
```

```
tail(scores1)
```

```
##          Factor1
## 752 -0.171304901
## 754  2.038896609
## 756  0.273237395
## 761 -1.151836968
## 764  0.008059989
## 766 -0.193785008
```

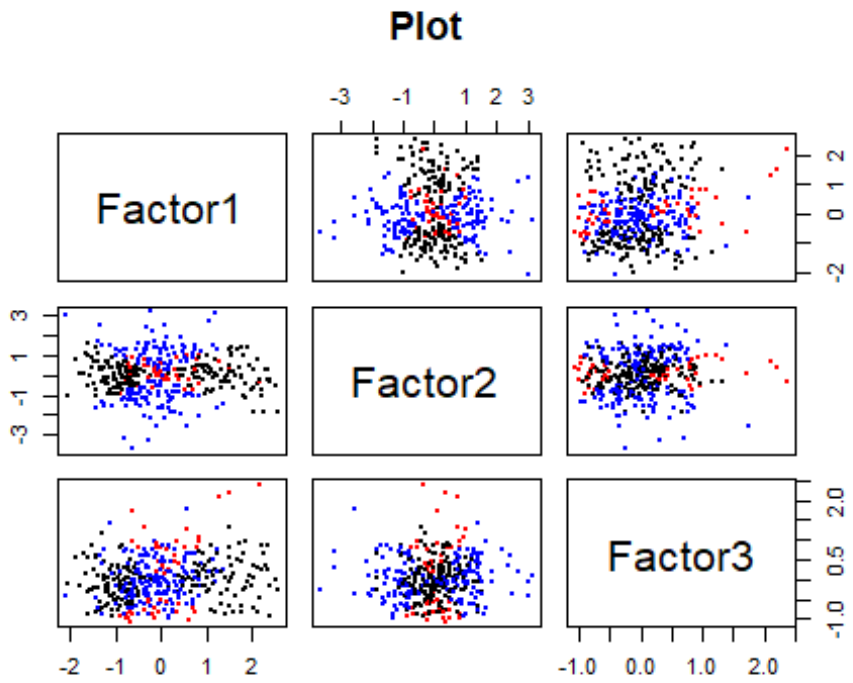
```
length(scores1)
```

```
## [1] 392
```

```
# As expected, 392
```

Plot of the scores using as label the class

```
scores <- factanal(newdf[,-c(1,6,9)], 3,
                   rotation="varimax",
                   scores = "regression")

fa.plot(scores$scores, labels=newdf[,-c(1,6,9)]$Class, pch=18, cex=0.3)
```



Let's try some rotations!

```
# Through factor rotation, we can make the output more understandable a
nd is usually necessary to facilitate the
# interpretation of factors. The aim is to find a simple solution (in o
ther words a solution that has simple structure!) that each factor has
a small number of large loadings and a large number of zero (or small)
loadings
# Note that the different rotations won't have an impact on the model f
it, it is expected to remain the same!
library(GPArotation)
fit_4 <- fa(newdf[,-c(1,6,9)], nfactors=2, n.obs=392,rotate="quartimax"
)
# Implementing the quartimax rotation with 2 factors
fit_4

## Factor Analysis using method =  minres
## Call: fa(r = newdf[, -c(1, 6, 9)], nfactors = 2, n.obs = 392, rotate
= "quartimax")
## Standardized loadings (pattern matrix) based upon correlation matrix
##             MR1   MR2    h2   u2 com
## plasma     0.82  0.13 0.687 0.31   1
## bl.press   0.11  0.89 0.797 0.20   1
## tr.thick   0.25  0.23 0.115 0.88   2
## serum.ins 0.69  0.01 0.480 0.52   1
## diab       0.21 -0.01 0.044 0.96   1
## age        0.35  0.30 0.212 0.79   2
```

```
##
##                        MR1  MR2
## SS loadings           1.39 0.94
## Proportion Var        0.23 0.16
## Cumulative Var        0.23 0.39
## Proportion Explained  0.60 0.40
## Cumulative Proportion 0.60 1.00
##
## Mean item complexity =  1.3
## Test of the hypothesis that 2 factors are sufficient.
##
## The degrees of freedom for the null model are  15  and the objective
function was  0.79 with Chi Square of  306.79
## The degrees of freedom for the model are 4  and the objective functi
on was  0.03
##
## The root mean square of the residuals (RMSR) is  0.03
## The df corrected root mean square of the residuals is  0.07
##
## The harmonic number of observations is  392 with the empirical chi s
quare  13.49  with prob <  0.0091
## The total number of observations was  392  with Likelihood Chi Squar
e =  10.82  with prob <  0.029
##
## Tucker Lewis Index of factoring reliability =  0.912
## RMSEA index =  0.066  and the 90 % confidence intervals are  0.019 0
.115
## BIC =  -13.06
## Fit based upon off diagonal values = 0.98
## Measures of factor score adequacy
##                                                MR1  MR2
## Correlation of (regression) scores with factors   0.87 0.89
## Multiple R square of scores with factors          0.76 0.79
## Minimum correlation of possible factor scores     0.52 0.59

# We observe that the cumulative percentage of the variance explained i
s 39%
# What's MR, ML, PC etc.? These are factors, and the name merely reflec
ts the fitting method, e.g. minimum residual,
# maximum likelihood, principal components. The default is minimum resi
dual, so in this case MR.
# h2: the amount of variance in the item/variable explained by the (ret
ained) factors. It is the sum of the squared
# loadings, a.k.a. communality. u2: 1 - h2. residual variance, a.k.a. u
niqueness
# SS loadings: These are the eigenvalues, the sum of the squared loadin
gs. In this case where we are using a
# correlation matrix, summing across all factors would equal the number
of variables used in the analysis
# The table beneath the loadings shows the proportion of variance expla
```

```
ined by each factor. The row Cumulative Var
# gives the cumulative proportion of variance explained. These numbers
range from 0 to 1. The row Proportion Var
# gives the proportion of variance explained by each factor, and the ro
w SS loadings gives the sum of squared
# loadings. This is sometimes used to determine the value of a particul
ar factor. A factor is worth keeping if the
# SS loading is greater than 1 (Kaiser's rule).
# null model: The degrees of freedom for the null model that assumes no
correlation structure.
# objective function: The value of the function that is minimized by a
specific procedure.
# model: The one you're actually interested in. Where p = Number of ite
ms, nf = number of factors then: degrees of
# freedom
fit_4$PVAL

## [1] 0.02861607

# p_value equal to 0.028 indicates that the fit is not good in this cas
e either!



fit_5 <- fa(newdf[,-c(1,6,9)], nfactors=2, n.obs=392,rotate="equamax")
fit_5

## Factor Analysis using method =  minres
## Call: fa(r = newdf[, -c(1, 6, 9)], nfactors = 2, n.obs = 392, rotate
= "equamax")
## Standardized loadings (pattern matrix) based upon correlation matrix
##            MR1   MR2    h2   u2 com
## plasma    0.82  0.15 0.687 0.31 1.1
## bl.press  0.09  0.89 0.797 0.20 1.0
## tr.thick  0.25  0.23 0.115 0.88 2.0
## serum.ins 0.69  0.03 0.480 0.52 1.0
## diab      0.21 -0.01 0.044 0.96 1.0
## age       0.34  0.31 0.212 0.79 2.0
##
##                        MR1  MR2
## SS loadings           1.37 0.96
## Proportion Var        0.23 0.16
## Cumulative Var        0.23 0.39
## Proportion Explained  0.59 0.41
## Cumulative Proportion 0.59 1.00
##
## Mean item complexity =  1.3
## Test of the hypothesis that 2 factors are sufficient.
##
## The degrees of freedom for the null model are  15  and the objective
function was  0.79 with Chi Square of  306.79
```

```
## The degrees of freedom for the model are 4  and the objective functi
on was  0.03
##
## The root mean square of the residuals (RMSR) is  0.03
## The df corrected root mean square of the residuals is  0.07
##
## The harmonic number of observations is  392 with the empirical chi s
quare  13.49  with prob <  0.0091
## The total number of observations was  392  with Likelihood Chi Squar
e =  10.82  with prob <  0.029
##
## Tucker Lewis Index of factoring reliability =  0.912
## RMSEA index =  0.066  and the 90 % confidence intervals are  0.019 0
.115
## BIC =  -13.06
## Fit based upon off diagonal values = 0.98
## Measures of factor score adequacy
##                                                  MR1  MR2
## Correlation of (regression) scores with factors   0.87 0.89
## Multiple R square of scores with factors          0.76 0.80
## Minimum correlation of possible factor scores     0.52 0.59
```

```
fit_5$PVAL
```

```
## [1] 0.02861607
```

```
# Same p_value as expected
```

```
fit_6 <- fa(newdf[,-c(1,6,9)] , nfactors=2, n.obs=392,rotate="promax")
fit_6
```

```
## Factor Analysis using method =  minres
## Call: fa(r = newdf[, -c(1, 6, 9)], nfactors = 2, n.obs = 392, rotate
= "promax")
## Standardized loadings (pattern matrix) based upon correlation matrix
##             MR1   MR2    h2   u2 com
## plasma     0.80  0.05 0.687 0.31 1.0
## bl.press  -0.31  1.01 0.797 0.20 1.2
## tr.thick   0.16  0.23 0.115 0.88 1.8
## serum.ins  0.73 -0.07 0.480 0.52 1.0
## diab       0.23 -0.04 0.044 0.96 1.1
## age        0.23  0.30 0.212 0.79 1.8
##
##                        MR1  MR2
## SS loadings           1.28 1.06
## Proportion Var        0.21 0.18
## Cumulative Var        0.21 0.39
## Proportion Explained  0.55 0.45
## Cumulative Proportion 0.55 1.00
##
```

```
##  With factor correlations of
##      MR1  MR2
## MR1 1.00 0.51
## MR2 0.51 1.00
##
## Mean item complexity =  1.3
## Test of the hypothesis that 2 factors are sufficient.
##
## The degrees of freedom for the null model are  15  and the objective
function was  0.79 with Chi Square of  306.79
## The degrees of freedom for the model are 4  and the objective functi
on was  0.03
##
## The root mean square of the residuals (RMSR) is  0.03
## The df corrected root mean square of the residuals is  0.07
##
## The harmonic number of observations is  392 with the empirical chi s
quare  13.49  with prob <  0.0091
## The total number of observations was  392  with Likelihood Chi Squar
e =  10.82  with prob <  0.029
##
## Tucker Lewis Index of factoring reliability =  0.912
## RMSEA index =  0.066  and the 90 % confidence intervals are  0.019 0
.115
## BIC =  -13.06
## Fit based upon off diagonal values = 0.98
## Measures of factor score adequacy
##                                                  MR1  MR2
## Correlation of (regression) scores with factors   0.88 0.91
## Multiple R square of scores with factors          0.77 0.82
## Minimum correlation of possible factor scores     0.54 0.65

fit_6$PVAL

## [1] 0.02861607

# Same p_value as expected


fit_7 <- fa(newdf[,-c(1,6,9)] , nfactors=3, n.obs=392,rotate="quartimax
")
fit_7

## Factor Analysis using method =  minres
## Call: fa(r = newdf[, -c(1, 6, 9)], nfactors = 3, n.obs = 392, rotate
= "quartimax")
## Standardized loadings (pattern matrix) based upon correlation matrix
##            MR1   MR2   MR3   h2     u2 com
## plasma    0.99  0.10 -0.02 0.99 0.0052 1.0
## bl.press  0.11  0.99  0.01 1.00 0.0047 1.0
## tr.thick  0.19  0.21  0.42 0.25 0.7463 1.9
```

```
## serum.ins 0.59  0.03  0.15 0.37 0.6329 1.1
## diab      0.15 -0.04  0.33 0.14 0.8645 1.4
## age       0.32  0.26  0.13 0.19 0.8094 2.3
##
##                        MR1  MR2  MR3
## SS loadings           1.50 1.11 0.33
## Proportion Var        0.25 0.18 0.05
## Cumulative Var        0.25 0.43 0.49
## Proportion Explained  0.51 0.38 0.11
## Cumulative Proportion 0.51 0.89 1.00
##
## Mean item complexity =  1.5
## Test of the hypothesis that 3 factors are sufficient.
##
## The degrees of freedom for the null model are  15  and the objective
function was  0.79 with Chi Square of  306.79
## The degrees of freedom for the model are 0  and the objective functi
on 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  392 with the empirical chi s
quare  0.02  with prob <  NA
## The total number of observations was  392  with Likelihood Chi Squar
e =  0.02  with prob <  NA
##
## Tucker Lewis Index of factoring reliability =  -Inf
## Fit based upon off diagonal values = 1
## Measures of factor score adequacy
##                                                MR1  MR2   MR3
## Correlation of (regression) scores with factors   1.00 1.00  0.55
## Multiple R square of scores with factors          0.99 0.99  0.30
## Minimum correlation of possible factor scores     0.99 0.99 -0.40
```

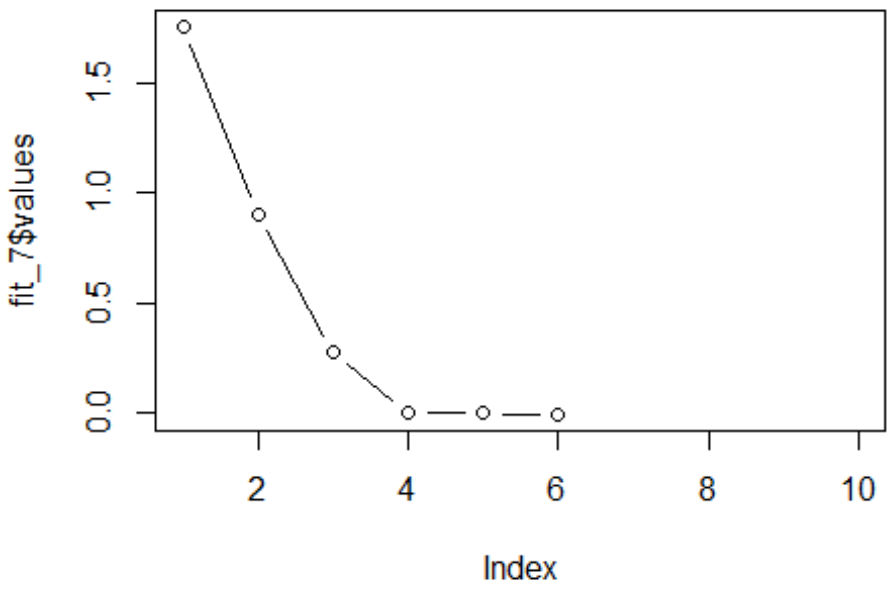
```
fit_7$PVAL
```

```
## [1] NA
```

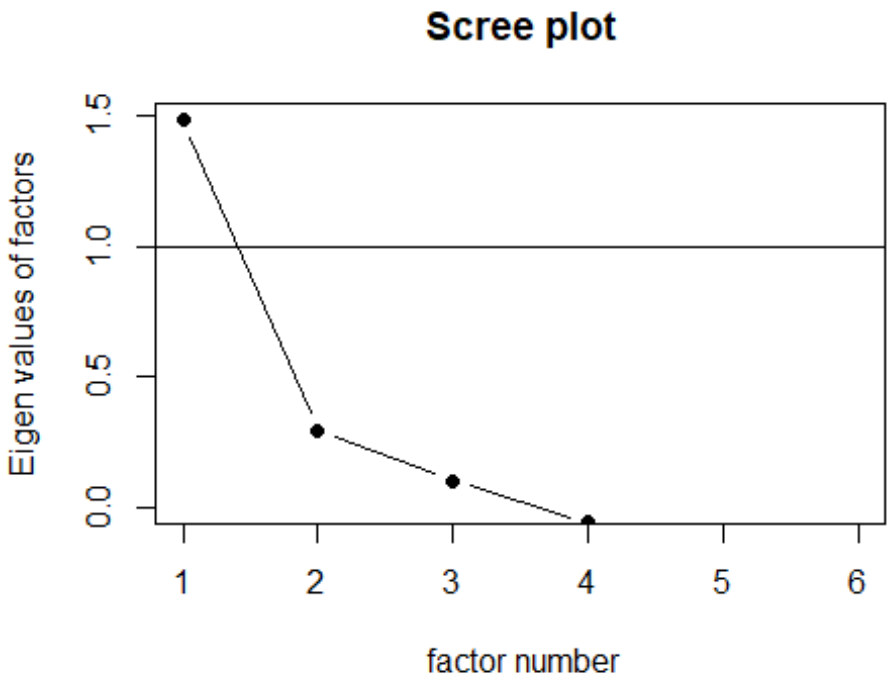
```
# Definitely not 3 factors!
```

Screeplots

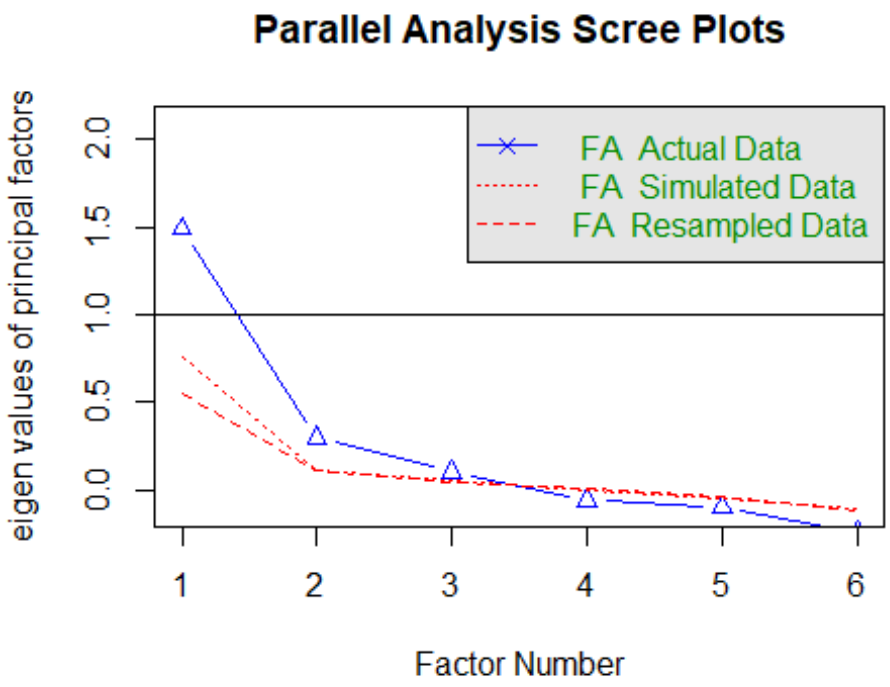
```
plot(fit_7$values, type = "b", xlim = c(1, 10))
```




```
# We construct a scree plot to aid with the selection of the number of
factors. From this plot, we see that the
# eigenvalues drop precipitously after factor 1 (maybe even 2)


scree(newdf[,-c(1,6,9)], pc=FALSE)
```




```
# Second way to provide a scree plot
# Use pc=FALSE for factor analysis


fa.parallel(newdf[,-c(1,6,9)], fa="fa")
```

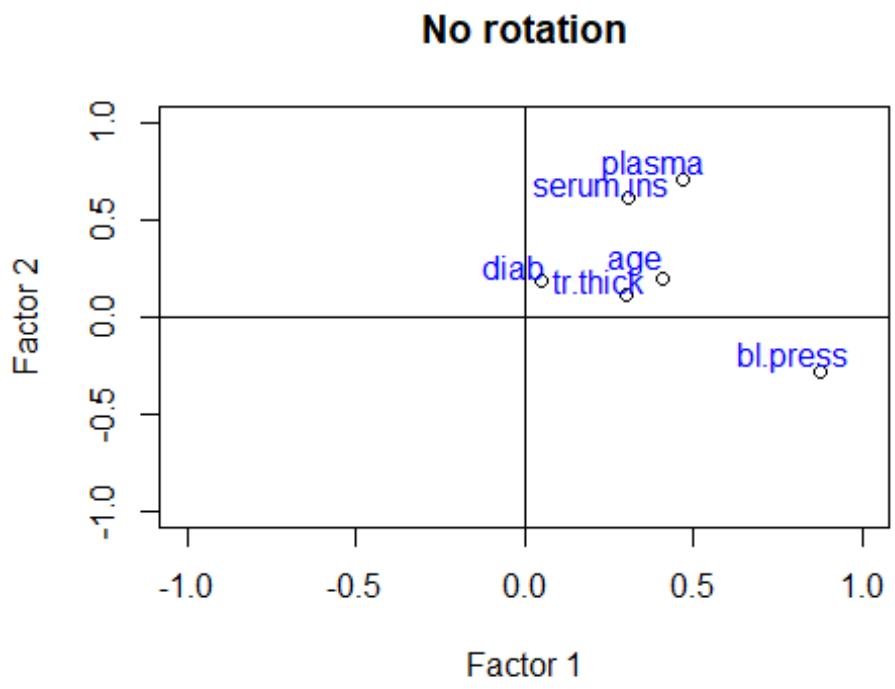



```
## Parallel analysis suggests that the number of factors =  3  and the
number of components =  NA

# The eigenvalue method ("Kaiser's rule") is telling us that 3 factors
may be best. Parallel analysis is revealing
# only two factors
```

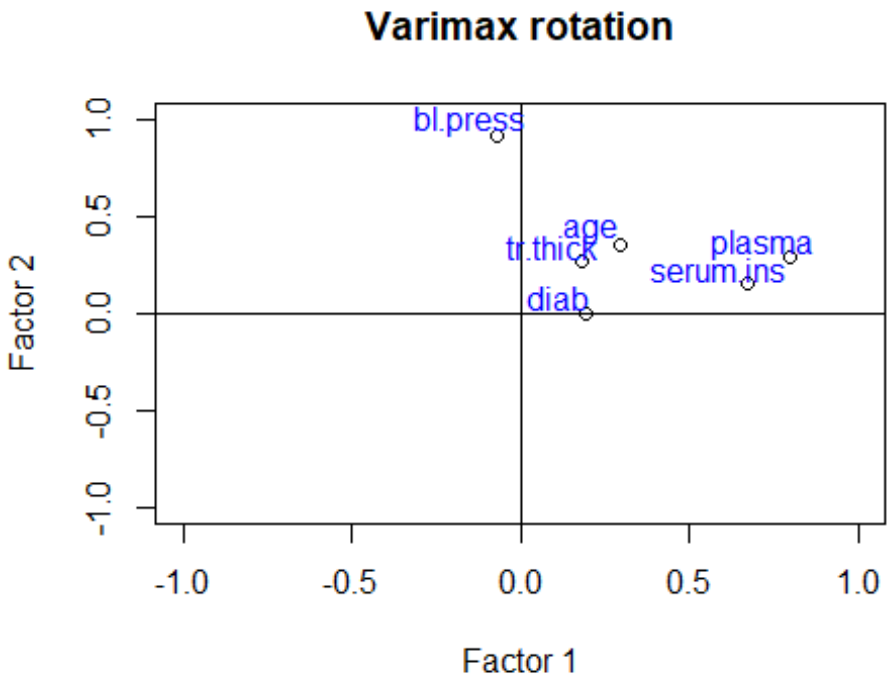
Interpretation of the factors

```
med.data.none <- factanal(newdf[,-c(1,6,9)] , factors = 2, rotation = "
none")
med.data.varimax <- factanal(newdf[,-c(1,6,9)] , factors = 2, rotation
= "varimax")
med.data.promax <- factanal(newdf[,-c(1,6,9)] , factors = 2, rotation =
"promax")
med.data.equamax <- factanal(newdf[,-c(1,6,9)] , factors = 2, rotation
= "equamax")
# Let's get a better picture of the factors along with the 4 rotations

plot(med.data.none$loadings[,1],
     med.data.none$loadings[,2],
     xlab = "Factor 1",
     ylab = "Factor 2",
     ylim = c(-1,1),
     xlim = c(-1,1),
     main = "No rotation")
abline(h = 0, v = 0)
```

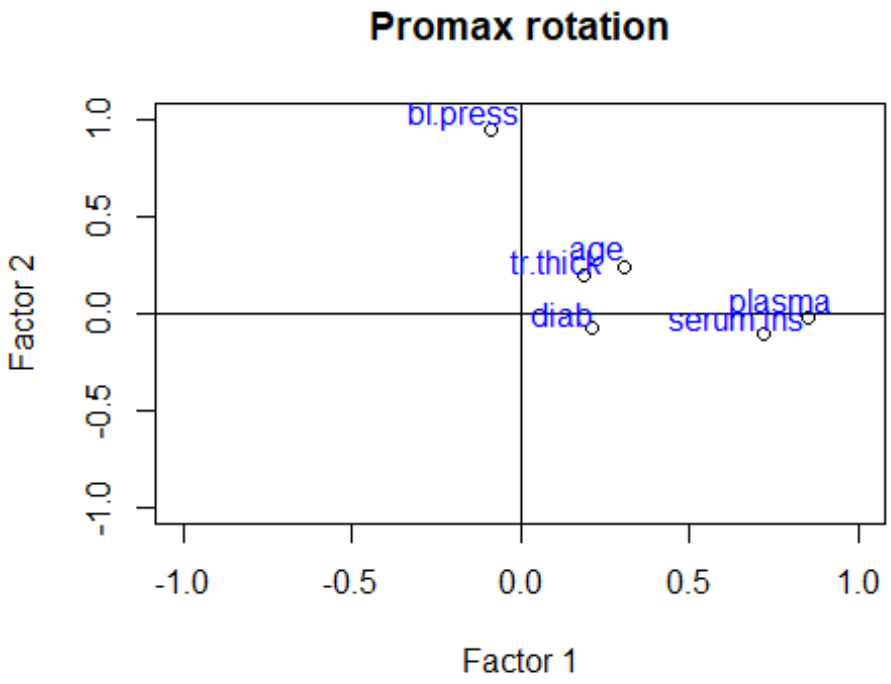
```
text(med.data.none$loadings[,1]-0.08,
     med.data.none$loadings[,2]+0.08,
     colnames(newdf[,-c(1,6,9)]),
     col="blue")
```



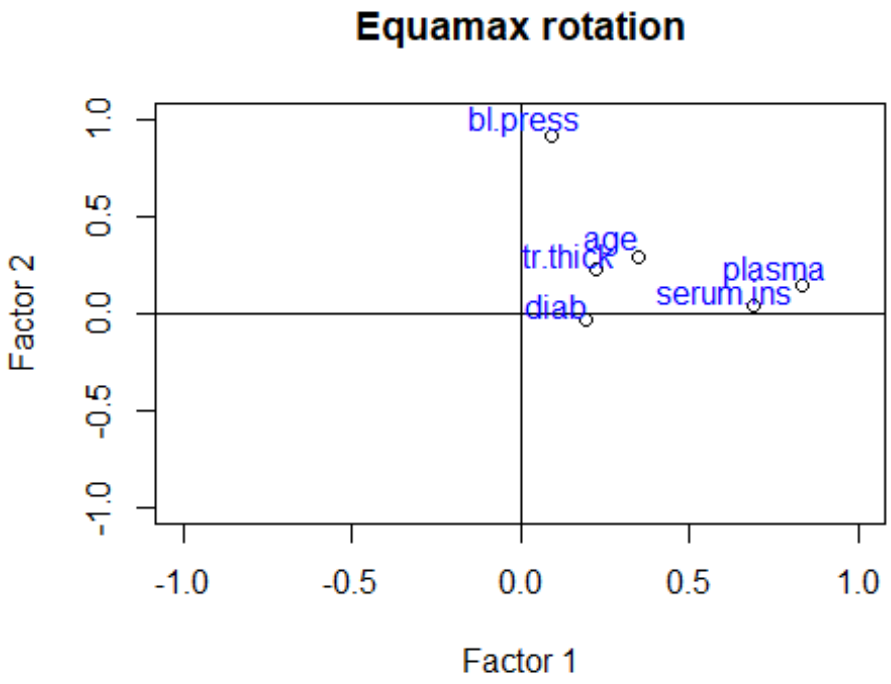

```
plot(med.data.varimax$loadings[,1],
     med.data.varimax$loadings[,2],
     xlab = "Factor 1",
     ylab = "Factor 2",
     ylim = c(-1,1),
     xlim = c(-1,1),
     main = "Varimax rotation")
abline(h = 0, v = 0)

text(med.data.varimax$loadings[,1]-0.08,
     med.data.varimax$loadings[,2]+0.08,
     colnames(newdf[,-c(1,6,9)]),
     col="blue")
```




```
plot(med.data.promax$loadings[,1],
     med.data.promax$loadings[,2],
     xlab = "Factor 1",
     ylab = "Factor 2",
     ylim = c(-1,1),
     xlim = c(-1,1),
     main = "Promax rotation")
abline(h = 0, v = 0)

text(med.data.promax$loadings[,1]-0.08,
     med.data.promax$loadings[,2]+0.08,
     colnames(newdf[,-c(1,6,9)]),
     col="blue")
```




```
plot(med.data.equamax$loadings[,1],
     med.data.equamax$loadings[,2],
     xlab = "Factor 1",
     ylab = "Factor 2",
     ylim = c(-1,1),
     xlim = c(-1,1),
     main = "Equamax rotation")
abline(h = 0, v = 0)

text(med.data.equamax$loadings[,1]-0.08,
     med.data.equamax$loadings[,2]+0.08,
     colnames(newdf[,-c(1,6,9)]),
     col="blue")
```




```
# Now comes the tricky aspect in factor analysis: Interpreting the fact
ors themselves. If two variables both have large loadings for the same
factor, then we know they have something in common. As a researcher, we
have to understand the data and its meaning in order to give a name to
that common ground.
```

Let's try exploring different methods

```
fit_8 <- fa(newdf[,-c(1,6,9)], nfactors=2, n.obs=392,rotate="quartimax"
,fm="ols")
# Implementing the quartimax rotation using ols method
fit_8

## Factor Analysis using method =  ols
## Call: fa(r = newdf[, -c(1, 6, 9)], nfactors = 2, n.obs = 392, rotate
= "quartimax",
##     fm = "ols")
## Standardized loadings (pattern matrix) based upon correlation matrix
##              1     2    h2   u2 com
## plasma    0.82  0.13 0.687 0.31   1
## bl.press  0.11  0.89 0.797 0.20   1
## tr.thick  0.25  0.23 0.115 0.88   2
## serum.ins 0.69  0.01 0.480 0.52   1
## diab      0.21 -0.01 0.044 0.96   1
## age       0.35  0.30 0.212 0.79   2
##
##                        [,1] [,2]
```

```
## SS loadings           1.39 0.94
## Proportion Var        0.23 0.16
## Cumulative Var        0.23 0.39
## Proportion Explained  0.60 0.40
## Cumulative Proportion 0.60 1.00
##
## Mean item complexity =  1.3
## Test of the hypothesis that 2 factors are sufficient.
##
## The degrees of freedom for the null model are  15  and the objective
function was  0.79 with Chi Square of  306.79
## The degrees of freedom for the model are 4  and the objective functi
on was  0.03
##
## The root mean square of the residuals (RMSR) is  0.03
## The df corrected root mean square of the residuals is  0.07
##
## The harmonic number of observations is  392 with the empirical chi s
quare  13.49  with prob <  0.0091
## The total number of observations was  392  with Likelihood Chi Squar
e =  10.82  with prob <  0.029
##
## Tucker Lewis Index of factoring reliability =  0.912
## RMSEA index =  0.066  and the 90 % confidence intervals are  0.019 0
.115
## BIC =  -13.06
## Fit based upon off diagonal values = 0.98
## Measures of factor score adequacy
##                                                 [,1] [,2]
## Correlation of (regression) scores with factors   0.87 0.89
## Multiple R square of scores with factors          0.76 0.79
## Minimum correlation of possible factor scores     0.52 0.59

# We observe that the results do not differ
fit_8$PVAL

## [1] 0.0286169

# p_value equal to 0.028 indicates that the fit is not good in this cas
e either!


fit_9 <- fa(newdf[,-c(1,6,9)], nfactors=2, n.obs=392,rotate="equamax",f
m="ols")
fit_9

## Factor Analysis using method =  ols
## Call: fa(r = newdf[, -c(1, 6, 9)], nfactors = 2, n.obs = 392, rotate
= "equamax",
##     fm = "ols")
## Standardized loadings (pattern matrix) based upon correlation matrix
```

```
##              1     2    h2   u2 com
## plasma    0.82  0.15 0.687 0.31 1.1
## bl.press  0.09  0.89 0.797 0.20 1.0
## tr.thick  0.25  0.23 0.115 0.88 2.0
## serum.ins 0.69  0.03 0.480 0.52 1.0
## diab      0.21 -0.01 0.044 0.96 1.0
## age       0.34  0.31 0.212 0.79 2.0
##
##                        [,1] [,2]
## SS loadings           1.37 0.96
## Proportion Var        0.23 0.16
## Cumulative Var        0.23 0.39
## Proportion Explained  0.59 0.41
## Cumulative Proportion 0.59 1.00
##
## Mean item complexity =  1.3
## Test of the hypothesis that 2 factors are sufficient.
##
## The degrees of freedom for the null model are  15  and the objective
function was  0.79 with Chi Square of  306.79
## The degrees of freedom for the model are 4  and the objective functi
on was  0.03
##
## The root mean square of the residuals (RMSR) is  0.03
## The df corrected root mean square of the residuals is  0.07
##
## The harmonic number of observations is  392 with the empirical chi s
quare  13.49  with prob <  0.0091
## The total number of observations was  392  with Likelihood Chi Squar
e =  10.82  with prob <  0.029
##
## Tucker Lewis Index of factoring reliability =  0.912
## RMSEA index =  0.066  and the 90 % confidence intervals are  0.019 0
.115
## BIC =  -13.06
## Fit based upon off diagonal values = 0.98
## Measures of factor score adequacy
##                                                    [,1] [,2]
## Correlation of (regression) scores with factors   0.87 0.89
## Multiple R square of scores with factors          0.76 0.80
## Minimum correlation of possible factor scores     0.52 0.59

fit_9$PVAL

## [1] 0.0286169

# Equally poor fit as the previous one, the p_value as expected is the
same
```

```r
fit_10 <- fa(newdf[,-c(1,6,9)] , nfactors=2, n.obs=392,rotate="promax",
fm="ols")
fit_10
```

```
## Factor Analysis using method =  ols
## Call: fa(r = newdf[, -c(1, 6, 9)], nfactors = 2, n.obs = 392, rotate
= "promax",
##     fm = "ols")
## Standardized loadings (pattern matrix) based upon correlation matrix
##               1     2    h2   u2 com
## plasma     0.80  0.05 0.687 0.31 1.0
## bl.press  -0.31  1.01 0.797 0.20 1.2
## tr.thick   0.16  0.23 0.115 0.88 1.8
## serum.ins  0.73 -0.07 0.480 0.52 1.0
## diab       0.23 -0.04 0.044 0.96 1.1
## age        0.23  0.30 0.212 0.79 1.8
##
##                        [,1] [,2]
## SS loadings            1.28 1.06
## Proportion Var         0.21 0.18
## Cumulative Var         0.21 0.39
## Proportion Explained   0.55 0.45
## Cumulative Proportion 0.55 1.00
##
##  With factor correlations of
##      [,1] [,2]
## [1,] 1.00 0.51
## [2,] 0.51 1.00
##
## Mean item complexity =  1.3
## Test of the hypothesis that 2 factors are sufficient.
##
## The degrees of freedom for the null model are  15  and the objective
function was  0.79 with Chi Square of  306.79
## The degrees of freedom for the model are 4  and the objective functi
on was  0.03
##
## The root mean square of the residuals (RMSR) is  0.03
## The df corrected root mean square of the residuals is  0.07
##
## The harmonic number of observations is  392 with the empirical chi s
quare  13.49  with prob <  0.0091
## The total number of observations was  392  with Likelihood Chi Squar
e =  10.82  with prob <  0.029
##
## Tucker Lewis Index of factoring reliability =  0.912
## RMSEA index =  0.066  and the 90 % confidence intervals are  0.019 0
.115
## BIC =  -13.06
## Fit based upon off diagonal values = 0.98
```

```
## Measures of factor score adequacy
##                                                 [,1] [,2]
## Correlation of (regression) scores with factors   0.88 0.91
## Multiple R square of scores with factors          0.77 0.82
## Minimum correlation of possible factor scores     0.54 0.65

fit_10$PVAL

## [1] 0.0286169

# Still not a desirable fit, same p_value observed just as expected


fit_11 <- fa(newdf[,-c(1,6,9)] , nfactors=2, n.obs=392,rotate="quartima
x",fm="ml")
fit_11

## Factor Analysis using method =  ml
## Call: fa(r = newdf[, -c(1, 6, 9)], nfactors = 2, n.obs = 392, rotate
= "quartimax",
##     fm = "ml")
## Standardized loadings (pattern matrix) based upon correlation matrix
##            ML2   ML1    h2   u2 com
## plasma    0.84  0.13 0.717 0.28 1.0
## bl.press  0.11  0.92 0.852 0.15 1.0
## tr.thick  0.23  0.23 0.103 0.90 2.0
## serum.ins 0.69  0.02 0.473 0.53 1.0
## diab      0.19 -0.03 0.038 0.96 1.1
## age       0.35  0.29 0.207 0.79 1.9
##
##                        ML2  ML1
## SS loadings           1.40 0.99
## Proportion Var        0.23 0.16
## Cumulative Var        0.23 0.40
## Proportion Explained  0.59 0.41
## Cumulative Proportion 0.59 1.00
##
## Mean item complexity =  1.3
## Test of the hypothesis that 2 factors are sufficient.
##
## The degrees of freedom for the null model are  15  and the objective
function was  0.79 with Chi Square of  306.79
## The degrees of freedom for the model are 4  and the objective functi
on was  0.03
##
## The root mean square of the residuals (RMSR) is  0.04
## The df corrected root mean square of the residuals is  0.07
##
## The harmonic number of observations is  392 with the empirical chi s
quare  15  with prob <  0.0047
## The total number of observations was  392  with Likelihood Chi Squar
```

```
e =  10.25  with prob <  0.036
##
## Tucker Lewis Index of factoring reliability =  0.919
## RMSEA index =  0.063  and the 90 % confidence intervals are  0.014 0
.112
## BIC =  -13.64
## Fit based upon off diagonal values = 0.98
## Measures of factor score adequacy
##                                                  ML2  ML1
## Correlation of (regression) scores with factors   0.88 0.92
## Multiple R square of scores with factors          0.78 0.85
## Minimum correlation of possible factor scores     0.55 0.69

fit_11$PVAL

## [1] 0.03646977

# We observe a p_value closer to 0.05 when using as a method ml!
# Let us note that this methodology would have produced better results
had the data been close to normality!
# We recall that our data are nowhere near normality hence we have some
indications about the poor fit!

fit_12 <- fa(newdf[,-c(1,6,9)] , nfactors=2, n.obs=392,rotate="quartima
x",fm="wls")
fit_12

## Factor Analysis using method =  wls
## Call: fa(r = newdf[, -c(1, 6, 9)], nfactors = 2, n.obs = 392, rotate
= "quartimax",
##     fm = "wls")
## Standardized loadings (pattern matrix) based upon correlation matrix
##           WLS1  WLS2   h2    u2 com
## plasma    0.79  0.08 0.63 0.369 1.0
## bl.press  0.33 -0.03 0.11 0.888 1.0
## tr.thick  0.31  0.13 0.11 0.889 1.4
## serum.ins 0.62  0.10 0.40 0.602 1.0
## diab      0.08  0.96 0.93 0.071 1.0
## age       0.45  0.04 0.21 0.793 1.0
##
##                        WLS1 WLS2
## SS loadings            1.43 0.96
## Proportion Var         0.24 0.16
## Cumulative Var         0.24 0.40
## Proportion Explained   0.60 0.40
## Cumulative Proportion  0.60 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  15  and the objective
```

```
function was  0.79 with Chi Square of  306.79
## The degrees of freedom for the model are 4  and the objective functi
on was  0.12
##
## The root mean square of the residuals (RMSR) is  0.07
## The df corrected root mean square of the residuals is  0.13
##
## The harmonic number of observations is  392 with the empirical chi s
quare  55.28  with prob <  2.8e-11
## The total number of observations was  392  with Likelihood Chi Squar
e =  46.28  with prob <  2.2e-09
##
## Tucker Lewis Index of factoring reliability =  0.455
## RMSEA index =  0.164  and the 90 % confidence intervals are  0.124 0
.209
## BIC =  22.39
## Fit based upon off diagonal values = 0.92
## Measures of factor score adequacy
##                                                 WLS1 WLS2
## Correlation of (regression) scores with factors   0.85 0.96
## Multiple R square of scores with factors          0.73 0.93
## Minimum correlation of possible factor scores     0.46 0.85

fit_12$PVAL

## [1] 2.157892e-09

# Different method now, but poor  model fit

fit_13 <- fa(newdf[,-c(1,6,9)] , nfactors=2, n.obs=392,rotate="quartima
x",fm="gls")
fit_13

## Factor Analysis using method =  gls
## Call: fa(r = newdf[, -c(1, 6, 9)], nfactors = 2, n.obs = 392, rotate
= "quartimax",
##     fm = "gls")
## Standardized loadings (pattern matrix) based upon correlation matrix
##           GLS1  GLS2   h2    u2 com
## plasma    0.83  0.07 0.70 0.301 1.0
## bl.press  0.32 -0.03 0.10 0.899 1.0
## tr.thick  0.29  0.13 0.10 0.896 1.4
## serum.ins 0.63  0.09 0.41 0.593 1.0
## diab      0.08  0.96 0.94 0.065 1.0
## age       0.43  0.04 0.19 0.810 1.0
##
##                        GLS1 GLS2
## SS loadings            1.47 0.96
## Proportion Var         0.25 0.16
## Cumulative Var         0.25 0.41
## Proportion Explained   0.60 0.40
```

```
## Cumulative Proportion 0.60 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  15  and the objective
function was  0.79 with Chi Square of  306.79
## The degrees of freedom for the model are 4  and the objective functi
on was  0.11
##
## The root mean square of the residuals (RMSR) is  0.07
## The df corrected root mean square of the residuals is  0.13
##
## The harmonic number of observations is  392 with the empirical chi s
quare  55.83  with prob <  2.2e-11
## The total number of observations was  392  with Likelihood Chi Squar
e =  43  with prob <  1e-08
##
## Tucker Lewis Index of factoring reliability =  0.497
## RMSEA index =  0.158  and the 90 % confidence intervals are  0.117 0
.202
## BIC =  19.12
## Fit based upon off diagonal values = 0.92
## Measures of factor score adequacy
##                                                  GLS1 GLS2
## Correlation of (regression) scores with factors   0.88 0.97
## Multiple R square of scores with factors          0.77 0.93
## Minimum correlation of possible factor scores     0.55 0.87

fit_13$PVAL

## [1] 1.032324e-08

# Another method, still a poor model fit

fit_14 <- fa(newdf[,-c(1,6,9)] , nfactors=2, n.obs=392,rotate="quartima
x",fm="pa")

## maximum iteration exceeded

fit_14

## Factor Analysis using method =  pa
## Call: fa(r = newdf[, -c(1, 6, 9)], nfactors = 2, n.obs = 392, rotate
= "quartimax",
##     fm = "pa")
## Standardized loadings (pattern matrix) based upon correlation matrix
##            PA1   PA2    h2   u2 com
## plasma    0.81  0.12 0.669 0.33 1.0
## bl.press  0.14  0.75 0.590 0.41 1.1
## tr.thick  0.25  0.25 0.125 0.87 2.0
```

```
## serum.ins 0.71 -0.01 0.502 0.50 1.0
## diab      0.21 -0.01 0.042 0.96 1.0
## age       0.35  0.33 0.234 0.77 2.0
##
##                        PA1  PA2
## SS loadings           1.40 0.76
## Proportion Var        0.23 0.13
## Cumulative Var        0.23 0.36
## Proportion Explained  0.65 0.35
## Cumulative Proportion 0.65 1.00
##
## Mean item complexity =  1.4
## Test of the hypothesis that 2 factors are sufficient.
##
## The degrees of freedom for the null model are  15  and the objective
function was  0.79 with Chi Square of  306.79
## The degrees of freedom for the model are 4  and the objective functi
on was  0.03
##
## The root mean square of the residuals (RMSR) is  0.03
## The df corrected root mean square of the residuals is  0.07
##
## The harmonic number of observations is  392 with the empirical chi s
quare  13.65  with prob <  0.0085
## The total number of observations was  392  with Likelihood Chi Squar
e =  10.83  with prob <  0.029
##
## Tucker Lewis Index of factoring reliability =  0.912
## RMSEA index =  0.066  and the 90 % confidence intervals are  0.019 0
.115
## BIC =  -13.06
## Fit based upon off diagonal values = 0.98
## Measures of factor score adequacy
##                                                 PA1  PA2
## Correlation of (regression) scores with factors   0.87 0.78
## Multiple R square of scores with factors          0.75 0.60
## Minimum correlation of possible factor scores     0.51 0.20

fit_14$PVAL

## [1] 0.02855784

# p_value equal to 0.02855784, not a great model fit either

fit_15 <- fa(newdf[,-c(1,6,9)] , nfactors=2, n.obs=392,rotate="quartima
x",fm="uls")
fit_15

## Factor Analysis using method =  uls
## Call: fa(r = newdf[, -c(1, 6, 9)], nfactors = 2, n.obs = 392, rotate
= "quartimax",
```

```
##     fm = "uls")
## Standardized loadings (pattern matrix) based upon correlation matrix
##           ULS1  ULS2    h2   u2 com
## plasma    0.82  0.13 0.687 0.31   1
## bl.press  0.11  0.89 0.797 0.20   1
## tr.thick  0.25  0.23 0.115 0.88   2
## serum.ins 0.69  0.01 0.480 0.52   1
## diab      0.21 -0.01 0.044 0.96   1
## age       0.35  0.30 0.212 0.79   2
##
##                        ULS1 ULS2
## SS loadings            1.39 0.94
## Proportion Var         0.23 0.16
## Cumulative Var         0.23 0.39
## Proportion Explained   0.60 0.40
## Cumulative Proportion  0.60 1.00
##
## Mean item complexity =  1.3
## Test of the hypothesis that 2 factors are sufficient.
##
## The degrees of freedom for the null model are  15  and the objective
function was  0.79 with Chi Square of  306.79
## The degrees of freedom for the model are 4  and the objective functi
on was  0.03
##
## The root mean square of the residuals (RMSR) is  0.03
## The df corrected root mean square of the residuals is  0.07
##
## The harmonic number of observations is  392 with the empirical chi s
quare  13.49  with prob <  0.0091
## The total number of observations was  392  with Likelihood Chi Squar
e =  10.82  with prob <  0.029
##
## Tucker Lewis Index of factoring reliability =  0.912
## RMSEA index =  0.066  and the 90 % confidence intervals are  0.019 0
.115
## BIC =  -13.06
## Fit based upon off diagonal values = 0.98
## Measures of factor score adequacy
##                                                 ULS1 ULS2
## Correlation of (regression) scores with factors   0.87 0.89
## Multiple R square of scores with factors          0.76 0.79
## Minimum correlation of possible factor scores     0.52 0.59

fit_15$PVAL

## [1] 0.02861625

# p_value equal to 0.02861625, not a great model fit either
```

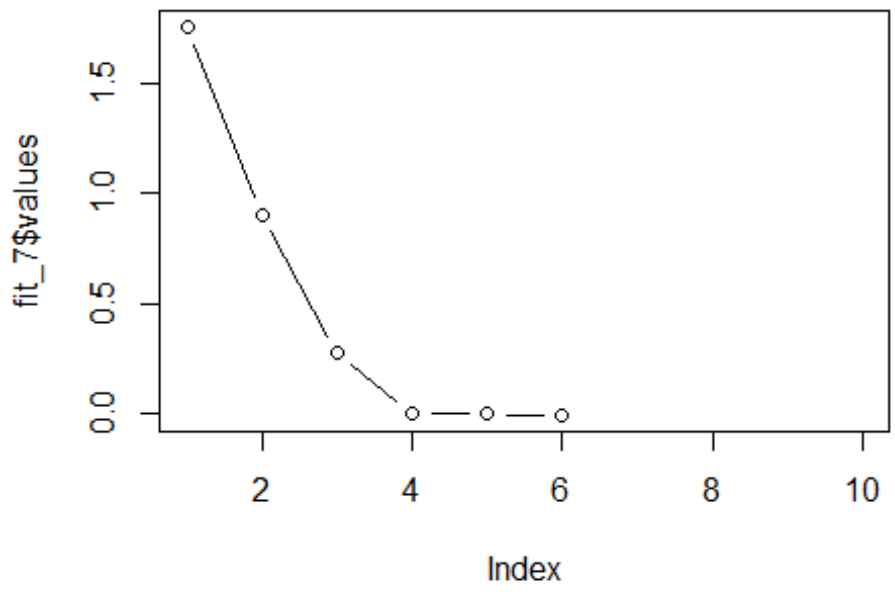
```
fit_16 <- fa(newdf[,-c(1,6,9)] , nfactors=2, n.obs=392,rotate="quartima
x",fm="minchi")
fit_16

## Factor Analysis using method =  minchi
## Call: fa(r = newdf[, -c(1, 6, 9)], nfactors = 2, n.obs = 392, rotate
= "quartimax",
##     fm = "minchi")
## Standardized loadings (pattern matrix) based upon correlation matrix
##            MC2   MC1    h2    u2 com
## plasma    0.85  0.07 0.736 0.264 1.0
## bl.press  0.27 -0.04 0.074 0.926 1.0
## tr.thick  0.25  0.15 0.085 0.915 1.6
## serum.ins 0.65  0.09 0.435 0.565 1.0
## diab      0.08  0.95 0.915 0.085 1.0
## age       0.40  0.05 0.165 0.835 1.0
##
##                        MC2  MC1
## SS loadings           1.46 0.95
## Proportion Var        0.24 0.16
## Cumulative Var        0.24 0.40
## Proportion Explained  0.61 0.39
## Cumulative Proportion 0.61 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  15  and the objective
function was  0.79 with Chi Square of  306.79
## The degrees of freedom for the model are 4  and the objective functi
on was  0.1
##
## The root mean square of the residuals (RMSR) is  0.07
## The df corrected root mean square of the residuals is  0.14
##
## The harmonic number of observations is  392 with the empirical chi s
quare  62.25  with prob <  9.8e-13
## The total number of observations was  392  with Likelihood Chi Squar
e =  40.53  with prob <  3.4e-08
##
## Tucker Lewis Index of factoring reliability =  0.529
## RMSEA index =  0.153  and the 90 % confidence intervals are  0.112 0
.197
## BIC =  16.64
## Fit based upon off diagonal values = 0.91
## Measures of factor score adequacy
##                                                MC2  MC1
## Correlation of (regression) scores with factors   0.89 0.96
## Multiple R square of scores with factors          0.79 0.91
## Minimum correlation of possible factor scores     0.58 0.83
```

```
fit_16$PVAL

## [1] 3.366101e-08

# Poor model fit

library(Rcsdp)
fit_17 <- fa(newdf[,-c(1,6,9)] , nfactors=2, n.obs=392,rotate="quartima
x",fm="minrank")
fit_17

## Factor Analysis using method =  minrank
## Call: fa(r = newdf[, -c(1, 6, 9)], nfactors = 2, n.obs = 392, rotate
= "quartimax",
##     fm = "minrank")
## Standardized loadings (pattern matrix) based upon correlation matrix
##           MRFA1 MRFA2    h2   u2 com
## plasma     0.81  0.11 0.677 0.32 1.0
## bl.press   0.15  0.67 0.466 0.53 1.1
## tr.thick   0.26  0.33 0.176 0.82 1.9
## serum.ins  0.72 -0.03 0.515 0.49 1.0
## diab       0.23 -0.01 0.053 0.95 1.0
## age        0.35  0.38 0.268 0.73 2.0
##
##                        MRFA1 MRFA2
## SS loadings             1.44  0.71
## Proportion Var          0.24  0.12
## Cumulative Var          0.24  0.36
## Proportion Explained    0.67  0.33
## Cumulative Proportion   0.67  1.00
##
## Mean item complexity =  1.3
## Test of the hypothesis that 2 factors are sufficient.
##
## The degrees of freedom for the null model are  15  and the objective
function was  0.79 with Chi Square of  306.79
## The degrees of freedom for the model are 4  and the objective functi
on was  0.03
##
## The root mean square of the residuals (RMSR) is  0.04
## The df corrected root mean square of the residuals is  0.07
##
## The harmonic number of observations is  392 with the empirical chi s
quare  17.52  with prob <  0.0015
## The total number of observations was  392  with Likelihood Chi Squar
e =  12.75  with prob <  0.013
##
## Tucker Lewis Index of factoring reliability =  0.887
## RMSEA index =  0.075  and the 90 % confidence intervals are  0.031 0
.123
```

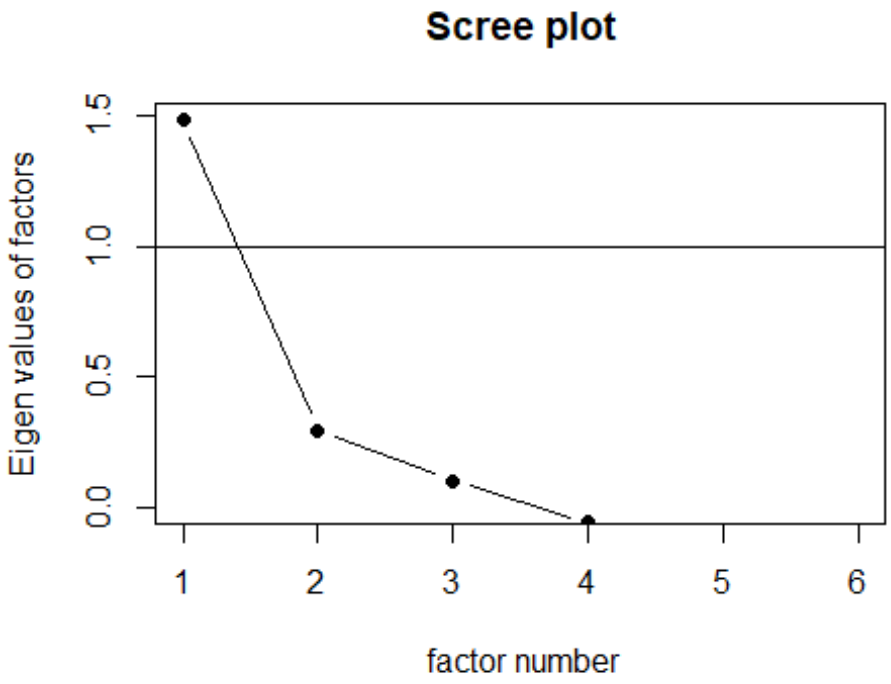
```
## BIC =  -11.14
## Fit based upon off diagonal values = 0.97
## Measures of factor score adequacy
##                                                  MRFA1 MRFA2
## Correlation of (regression) scores with factors   0.88  0.73
## Multiple R square of scores with factors          0.77  0.53
## Minimum correlation of possible factor scores     0.54  0.06

fit_17$PVAL

## [1] 0.01258349

# p_value equal to 0.01258349, indicating not a great fit either
```

Screeplots

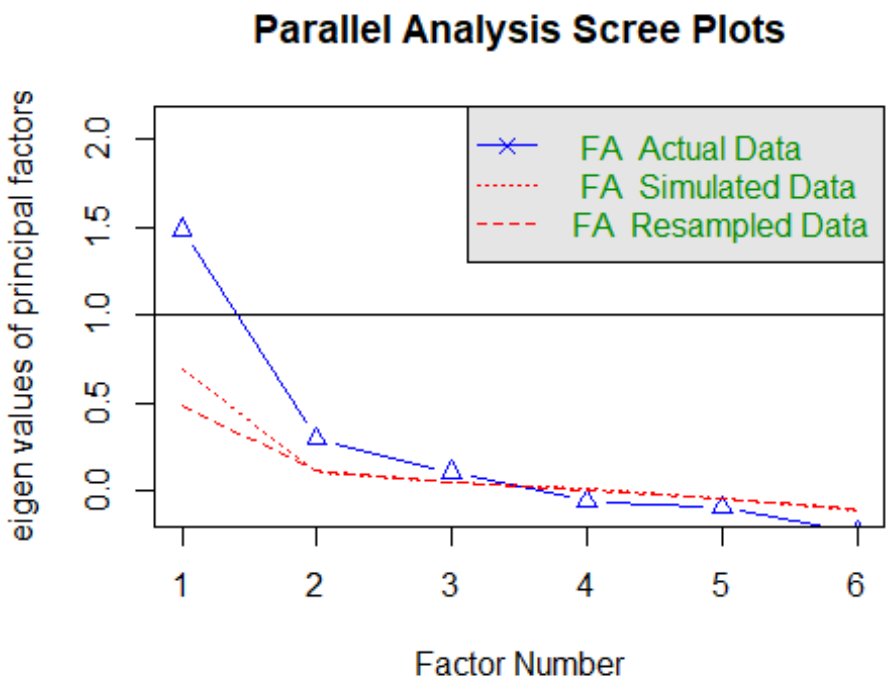
```
plot(fit_7$values, type = "b", xlim = c(1, 10))
```




```
# We construct a scree plot to aid with the selection of the number of
factors. From this plot, we see that the
# eigenvalues drop precipitously after factor 1 (maybe even 2)


scree(newdf[,-c(1,6,9)], pc=FALSE)
```



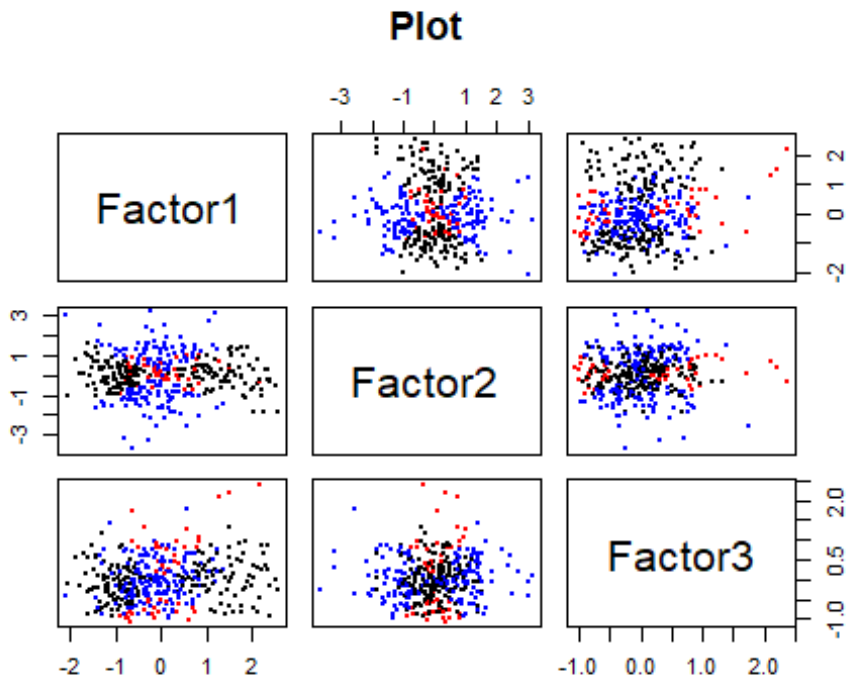

```
# Second way to provide a scree plot
# Use pc=FALSE for factor analysis


fa.parallel(newdf[,-c(1,6,9)], fa="fa")
```




```
## Parallel analysis suggests that the number of factors =  3  and the
number of components =  NA

# The eigenvalue method ("Kaiser's rule") is telling us that 3 factors
may be best. Parallel analysis is revealing
# only two factors
```

Plot of the (new) scores using as label the class (via the ml method)

```
scores_2 <- factanal(newdf[,-c(1,6,9)], 3,
                     rotation="varimax",
                     scores = "regression", method="ml")

fa.plot(scores_2$scores, labels=newdf[,-c(1,6,9)]$Class, pch=18, cex=0.
3)
```



Final suggestions

Our final suggestion will be based after taking into account:

- Model fit (as good as possible even though that's rarely the case!)
- The model that has a simple structure (after having tried many rotations in order to obtain better interpretation!)
- The model that has the 'best' valued Communalities (in other words the max values or the min values of Uniqueness!)

Our choice after conducting the above exploratory factor analysis is going to be fit_11, which produced the following output:

```
## Factor Analysis using method =  ml
## Call: fa(r = newdf[, -c(1, 6, 9)], nfactors = 2, n.obs = 392, rotate
= "quartimax",
##     fm = "ml")
## Standardized loadings (pattern matrix) based upon correlation matrix
##             ML2   ML1    h2   u2 com
## plasma     0.84  0.13 0.717 0.28 1.0
## bl.press   0.11  0.92 0.852 0.15 1.0
## tr.thick   0.23  0.23 0.103 0.90 2.0
## serum.ins 0.69  0.02 0.473 0.53 1.0
## diab       0.19 -0.03 0.038 0.96 1.1
## age        0.35  0.29 0.207 0.79 1.9
```

```
##                         ML2  ML1
## SS loadings           1.40 0.99
## Proportion Var        0.23 0.16
## Cumulative Var        0.23 0.40
## Proportion Explained  0.59 0.41
## Cumulative Proportion 0.59 1.00
##
## Mean item complexity =  1.3
## Test of the hypothesis that 2 factors are sufficient.
## The degrees of freedom for the null model are  15  and the objective
function was  0.79 with Chi Square of  306.79
## The degrees of freedom for the model are 4  and the objective functi
on was  0.03
## The root mean square of the residuals (RMSR) is  0.04
## The df corrected root mean square of the residuals is  0.07
## The harmonic number of observations is  392 with the empirical chi s
quare  15  with prob <  0.0047
## The total number of observations was  392  with Likelihood Chi Squar
e =  10.25  with prob <  0.036
## Tucker Lewis Index of factoring reliability =  0.919
## RMSEA index =  0.063  and the 90 % confidence intervals are  0.014 0
.112
## BIC =  -13.64
## Fit based upon off diagonal values = 0.98
## Measures of factor score adequacy
##                                                    ML2  ML1
## Correlation of (regression) scores with factors   0.88 0.92
## Multiple R square of scores with factors          0.78 0.85
## Minimum correlation of possible factor scores     0.55 0.69

fit_11$PVAL

## [1] 0.03646977
```

We notice that the 1st factor explains the 23% of the total variability while having 2 factors the percentage of the total variability explained increases to 40%!

Note that we wanted to have as few factors as possible! (Should we have more than 2 we wouldn't have made a big difference in comparison to the initial variables!)

We also observe that this model has loadings that are as closer to simple structure (in comparison with the previous models explored)! Moreover, the fit is close to being acceptable (should we decrease the significance level the confidence interval would get wider , thus we would have an acceptable fit!). As regards the loadings now, for the first factor: $\lambda 11=0.84$, $\lambda 21=0.11$, $\lambda 31=0.23$, $\lambda 41=0.69$, $\lambda 51= 0.19$ and $\lambda 61=0.35$. As regards the second factor: $\lambda 12=0.13$, $\lambda 22=0.92$, $\lambda 32=0.23$, $\lambda 42=0.02$, $\lambda 52=-0.03$, $\lambda 62=0.29$.

Last but not least, maybe we could try some transformation (get closer to normality if possible!), or an imputed dataset so as to obtain better results!