Multivariate Statistics HW3

Namwoo Kwon 20236002

2023-03-20

Quiz 5.4

Imagine that you have been employed as a research assistant by a faculty member doing some empirical analysis.

This professor has a data set with six correlated variables (n=100), and he has asked you to conduct a factor analysis on these data.

You have been provided with the following information :

- The variables are simply labeled X1 to X6, The professor wants you to be informed only by the patterns of association you observe in the data (and not by the names of the variables)
- According to the professor, most of the published empirical research in his area is based on the common factor model
- The professor intends to use the data in a subsequent analysis. Therefore, it is important that he can easily interpret the results you present to him

The raw data are in the file SIX_VARIABLES

```
#getwd()
setwd('/Users/namwoo/Desktop/UNIST/lecture/1-1/Multivaraiate Statistics')
```

```
df <- read.csv("SIX_VARIABLES.csv")
head(df)
```

```
##          X1       X2       X3       X4       X5       X6
## 1   0.49464  0.84748 -0.18870  0.22944  0.31475  0.14086
## 2   1.29702  0.64319  1.00948  0.55360  2.12976  2.82855
## 3 -0.47604  0.63695  0.01731  0.56861  1.02323  0.86964
## 4   2.26378  1.17854  2.66852  2.61012  3.92767  3.30061
## 5   0.97278  1.24157 -1.10611 -1.69492 -2.00753 -1.50181
## 6 -2.62733 -3.17792 -0.99573 -0.48252 -0.11812 -0.75417
```

```
tail(df)
```

```
##           X1       X2       X3       X4       X5       X6
## 95   0.95214  0.31101 -1.66021 -2.22263 -1.67761 -2.20977
## 96  -0.17752 -0.35893  0.41261 -0.31711  0.72249  0.05213
## 97  -0.15654 -0.32270 -2.27084 -3.60791 -2.27287 -1.84273
## 98   0.05233  0.25244 -0.22527 -0.66256 -0.84718  0.60513
## 99   0.90883  0.46759 -0.75479 -1.46050 -0.98723 -0.09253
## 100 -1.11111 -0.69541 -1.36972 -1.56176 -0.28748 -1.92193
```

```
str(df)
```

```
## 'data.frame':    100 obs. of  6 variables:
##  $ X1: num  0.495 1.297 -0.476 2.264 0.973 ...
##  $ X2: num  0.847 0.643 0.637 1.179 1.242 ...
##  $ X3: num  -0.1887 1.0095 0.0173 2.6685 -1.1061 ...
##  $ X4: num  0.229 0.554 0.569 2.61 -1.695 ...
##  $ X5: num  0.315 2.13 1.023 3.928 -2.008 ...
##  $ X6: num  0.141 2.829 0.87 3.301 -1.502 ...
```

```
summary(df)
```

```
##        X1                X2                X3                X4
##  Min.   :-3.4819   Min.   :-3.1779   Min.   :-3.21068   Min.   :-3.71032
##  1st Qu.:-0.5304   1st Qu.:-0.6903   1st Qu.:-1.00846   1st Qu.:-0.83814
##  Median : 0.2769   Median : 0.3932   Median : 0.04046   Median : 0.09073
##  Mean   : 0.1843   Mean   : 0.2024   Mean   : 0.14209   Mean   : 0.05370
##  3rd Qu.: 1.0642   3rd Qu.: 1.0619   3rd Qu.: 1.09874   3rd Qu.: 1.04891
##  Max.   : 4.0083   Max.   : 4.4932   Max.   : 4.00683   Max.   : 3.79377
##        X5                X6
##  Min.   :-3.76336   Min.   :-3.13323
##  1st Qu.:-0.99647   1st Qu.:-0.93801
##  Median :-0.15167   Median :-0.08321
##  Mean   :-0.06415   Mean   : 0.11452
##  3rd Qu.: 1.04373   3rd Qu.: 1.11026
##  Max.   : 3.92767   Max.   : 4.27324
```

The professor has asked you to provide him with the answers to the following questions :

Quiz a :

- How many factors would you extract from these six variables?
- Clearly explain the reasons behind your decision.

Answer

- **Let's check an analysis of the number of component or factors to retain in an principal component analysis.**
- **According to optimal coordinates oc & the acceleration factor af, Number of components/factors to retain is one.**
- **And according to parallel analysis & the Kaiser rule, Number of components/factors to retain is two.**
- **But, The number of components or factors I recommend is three. The reasons for this are :**
    1. Using as many components or factors as possible, within reason, can reflect a lot of information in the original data.
    2. Therefore, based on the Kaiser rule and parallel analysis, we can use two factors.
    3. However, an exploratory factor analysis with two factors would be rejected by the p-value of the chi square statistic.
    4. And in uniquenesses, we can see that there are variables with a value greater than 0.5
    5. If you check the correlation coefficients for each variable, there are three cases where the correlation coefficients are high between different variables.
    6. Using 3 factors does not cause problems with p-values and uniquenesses.

```
#install.packages("nFactors")
library(nFactors)
```

```
## Loading required package: lattice
```

```
##
## Attaching package: 'nFactors'
```

```
## The following object is masked from 'package:lattice':
##
##     parallel
```

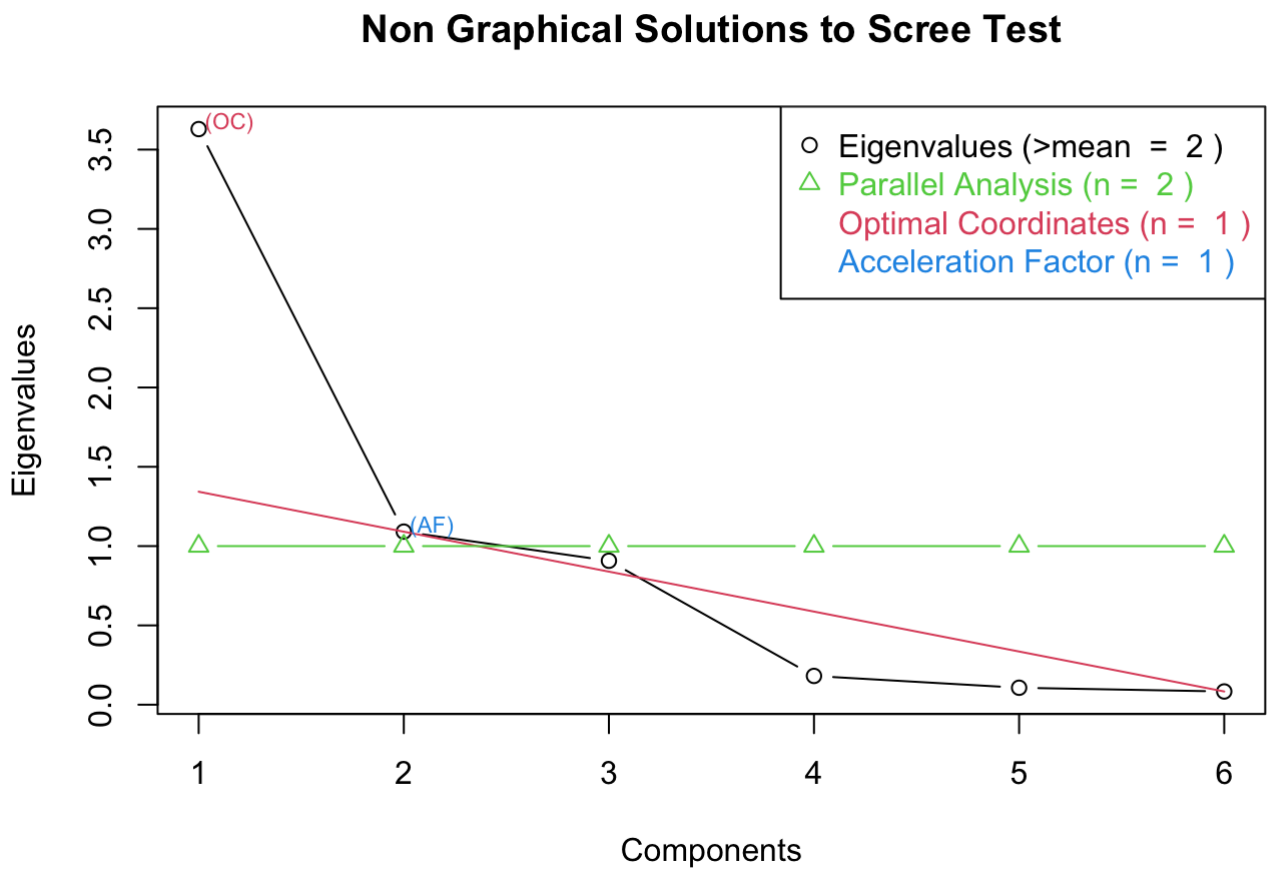
Description of output values

- noc : Determine the appropriate number of components or factors by finding an optimal point in the scree plot that better reflects the eigenvalues after the point where the rate of decline of the eigenvalues slows down dramatically.
- naf : A method for calculating the rate of decline of eigenvalues in a scree plot, and selecting the number of components or factors at the point where this rate of decline changes the most.
- nparallel : Generate random data similar to the real data, and compare the eigenvalues extracted from this random data with the eigenvalues extracted from the real data to determine the number of principal components.
- nkaiser : Components or factor with eigenvalues greater than 1 are assumed to represent significant dimensions in the real data, so choosing a priori the number of principal components is about eliminating unnecessary dimensions and selecting those that best reflect the structure of the real data.

```
nScree(df)
```

```
##   noc naf nparallel nkaiser
## 1   1   1         2       2
```

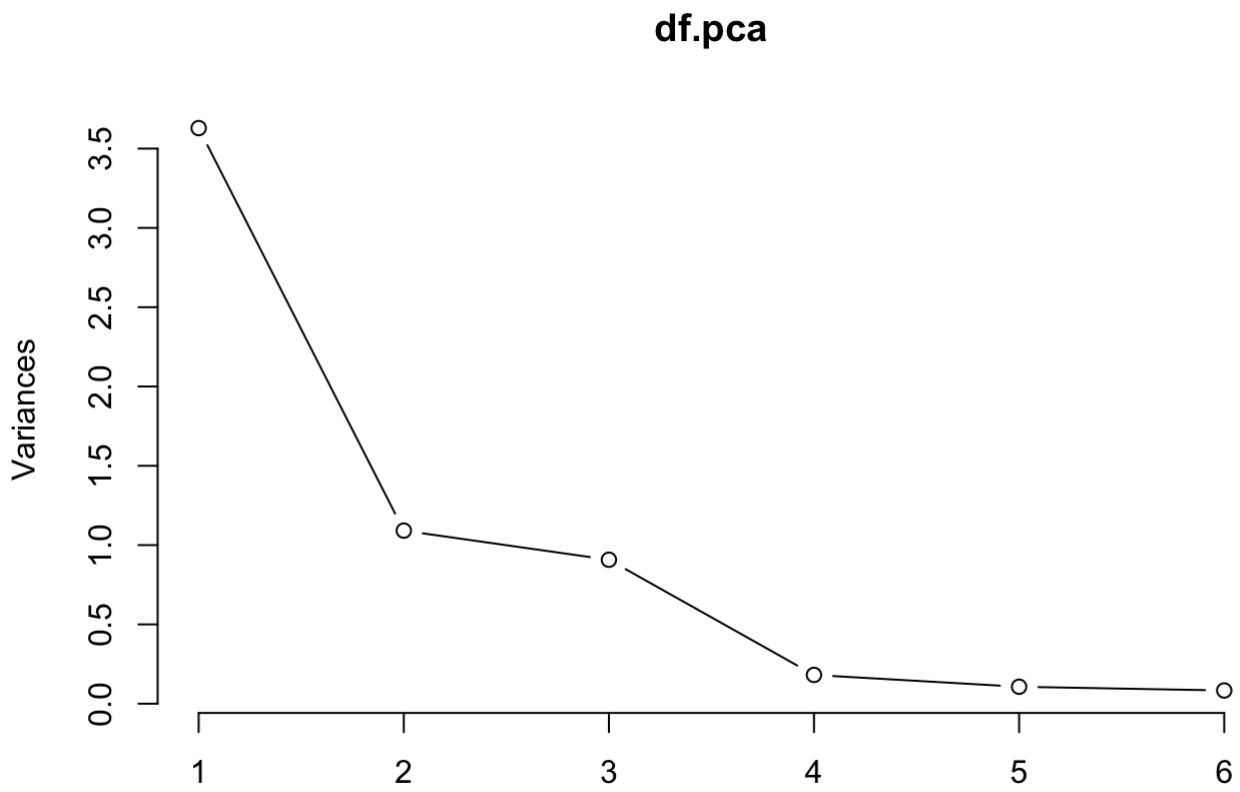
```
plotnScree(nScree(df))
```



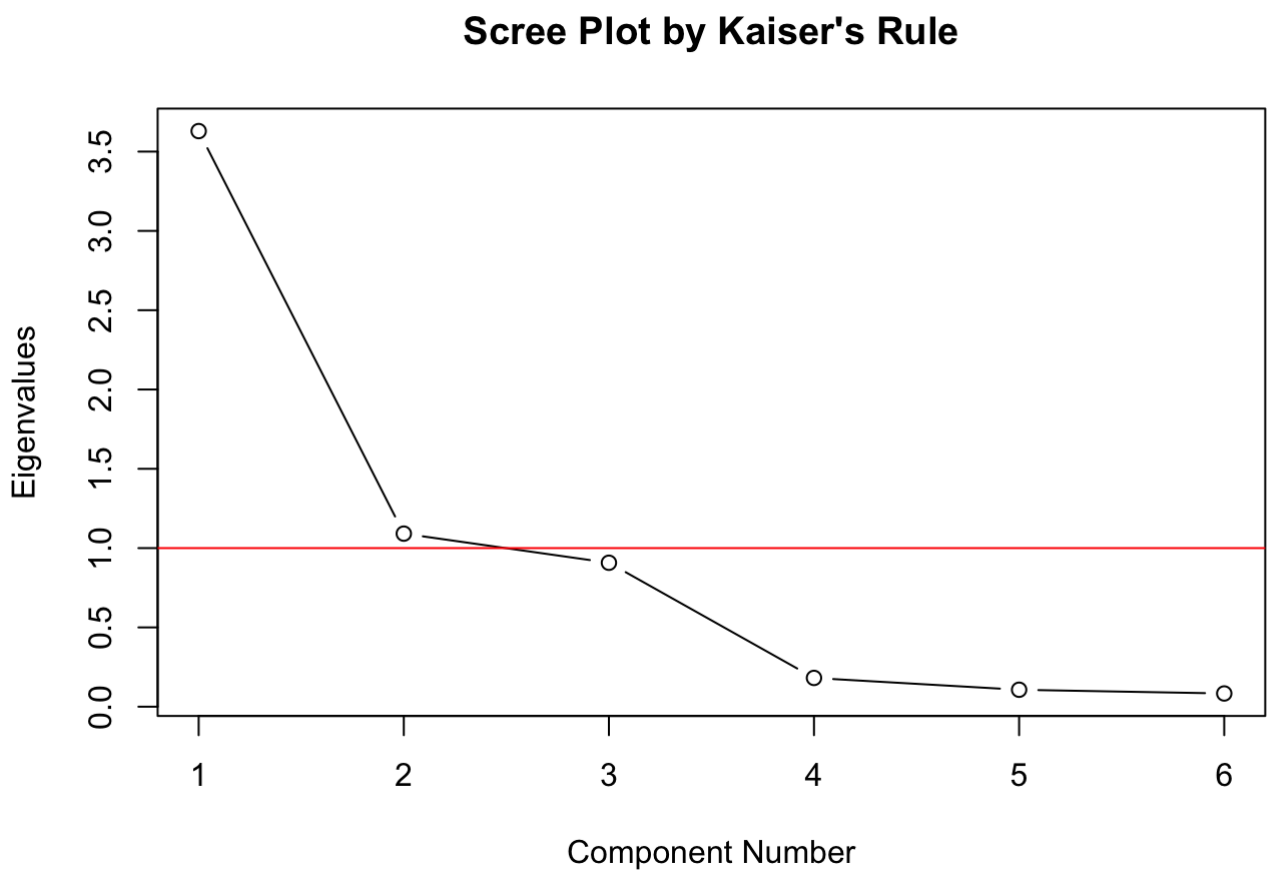

```
eigen(cor(df))
```

```
## eigen() decomposition
## $values
## [1] 3.62933576 1.09128940 0.90760960 0.18141900 0.10693845 0.08340778
##
## $vectors
##            [,1]       [,2]       [,3]       [,4]       [,5]        [,6]
## [1,] -0.3989716  0.5300910 -0.2171500  0.5150365 -0.4165504  0.27183507
## [2,] -0.3796204  0.6007336 -0.1194725 -0.5198417  0.3915189 -0.23918897
## [3,] -0.4315721 -0.1222734  0.5350538  0.2082631 -0.1926092 -0.65729774
## [4,] -0.4241110 -0.1413400  0.5572358 -0.1361441  0.1966615  0.65759416
## [5,] -0.4092421 -0.4109041 -0.3852970 -0.4928257 -0.5211149  0.02805369
## [6,] -0.4038642 -0.3928774 -0.4396863  0.3996450  0.5708449 -0.06030691
```

```
df.pca <- prcomp(df,
                 center=T,
                 scale=T)

plot(df.pca, type="l")
```



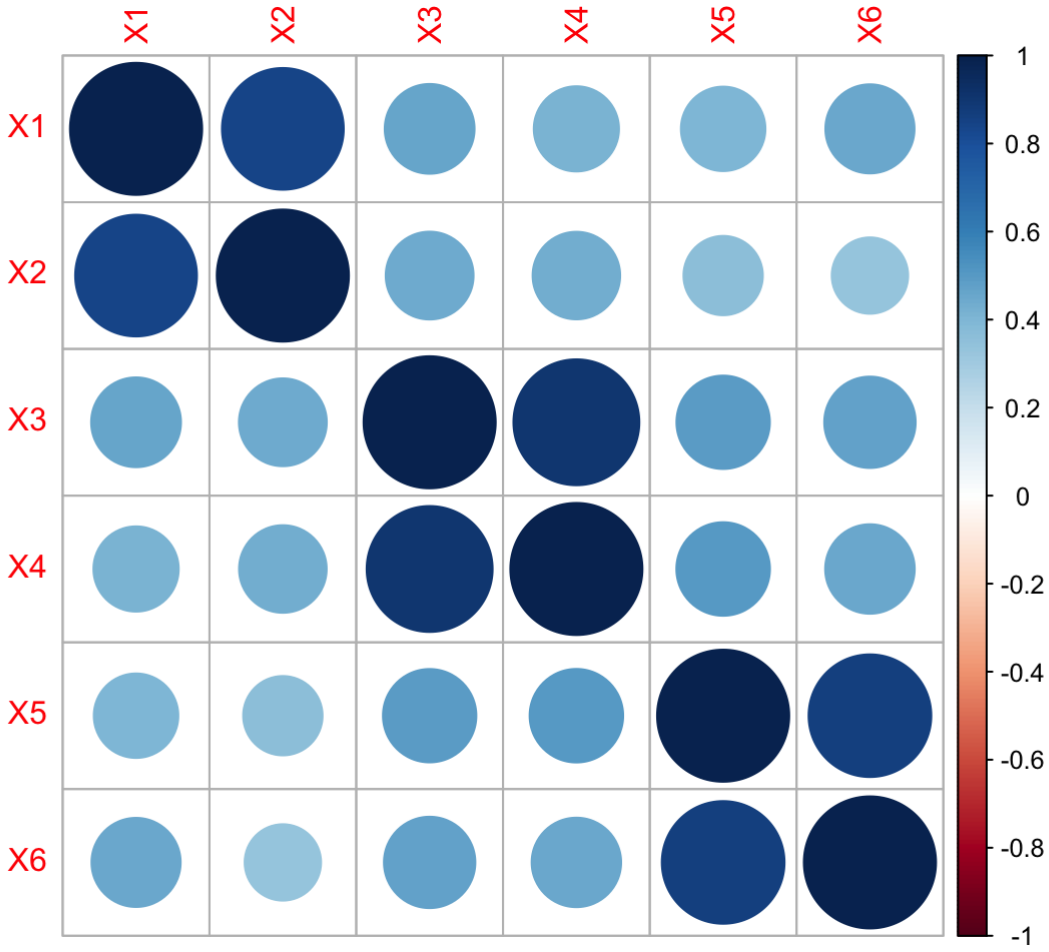

```
plot(eigen(cor(df))$values,
     main="Scree Plot by Kaiser's Rule",
     xlab="Component Number",
     ylab="Eigenvalues",
     type="b")

abline(h=1,
       col='red')
```




```
#install.packages('corrplot')
library(corrplot)
```

```
## corrplot 0.92 loaded
```

```
corrplot(cor(df), order='hclust')
```




```
factanal(df,
         factors=2,
         rotation='none')
```

```
##
## Call:
## factanal(x = df, factors = 2, rotation = "none")
##
## Uniquenesses:
##    X1    X2    X3    X4    X5    X6
## 0.005 0.269 0.106 0.076 0.676 0.681
##
## Loadings:
##    Factor1 Factor2
## X1  0.996
## X2  0.854
## X3  0.501   0.801
## X4  0.458   0.845
## X5  0.429   0.375
## X6  0.471   0.312
##
##                Factor1 Factor2
## SS loadings      2.589   1.598
## Proportion Var   0.431   0.266
## Cumulative Var   0.431   0.698
##
## Test of the hypothesis that 2 factors are sufficient.
## The chi square statistic is 106.84 on 4 degrees of freedom.
## The p-value is 3.44e-22
```

```
factanal(df,
         factors=3,
         rotation='none')
```

```
##
## Call:
## factanal(x = df, factors = 3, rotation = "none")
##
## Uniquenesses:
##    X1    X2    X3    X4    X5    X6
## 0.005 0.257 0.161 0.005 0.239 0.005
##
## Loadings:
##    Factor1 Factor2 Factor3
## X1  0.781  -0.556   0.275
## X2  0.679  -0.402   0.346
## X3  0.779   0.398   0.271
## X4  0.789   0.519   0.322
## X5  0.749          -0.437
## X6  0.806          -0.587
##
##                Factor1 Factor2 Factor3
## SS loadings      3.510   0.909   0.908
## Proportion Var   0.585   0.152   0.151
## Cumulative Var   0.585   0.737   0.888
##
## The degrees of freedom for the model is 0 and the fit was 0.0425
```

Quiz b :

- **How much of the information from the original set of six variables is accounted for by these factors?**

Answer

- **For each of the different rotation techniques, let's run an exploratory factor analysis to generate two factors.**
- **Let's look at desirable factor load patterns**
  - Most of the loadings on any specific factor (column) should be small (as close to zero as possible), and only a few loadings should be large in absolute value.
  - A specific row of the loadings matrix, containing the loadings of a given variable with each factor, should display nonzero loadings on only one or no more than a few factors.
  - Any pair of factors (columns) should exhibit different patterns of loadings. Otherwise, one could not distinguish the two factors represented by these columns.
- **It seems preferable to rotate them via the promax technique.**
  - With Promax, you can better account for differences between the factors of an X variable compared to other rotation techniques.
  - And for each variable, it's clear to see what factors are best annualized
- **In conclusion, based on promax, each factor has a variance of 0.314, 0.289 and 0.287 of the original data. And the two factors explain a total of 0.889 of the variance in the original data.**

```
factanal(df, factors=3, rotation="varimax")
```

```
##
## Call:
## factanal(x = df, factors = 3, rotation = "varimax")
##
## Uniquenesses:
##    X1    X2    X3    X4    X5    X6
## 0.005 0.257 0.161 0.005 0.239 0.005
##
## Loadings:
##    Factor1 Factor2 Factor3
## X1 0.172   0.241   0.952
## X2 0.255   0.128   0.813
## X3 0.834   0.271   0.265
## X4 0.946   0.240   0.205
## X5 0.288   0.805   0.174
## X6 0.199   0.957   0.199
##
##                Factor1 Factor2 Factor3
## SS loadings      1.808   1.769   1.751
## Proportion Var   0.301   0.295   0.292
## Cumulative Var   0.301   0.596   0.888
##
## The degrees of freedom for the model is 0 and the fit was 0.0425
```

```
#install.packages("GPArotation")
library(GPArotation)
```

```
factanal(df, factors=3, rotation="oblimin")
```

```
##
## Call:
## factanal(x = df, factors = 3, rotation = "oblimin")
##
## Uniquenesses:
##    X1    X2    X3    X4    X5    X6
## 0.005 0.257 0.161 0.005 0.239 0.005
##
## Loadings:
##    Factor1 Factor2 Factor3
## X1          0.998
## X2  0.103   0.841
## X3  0.862
## X4  1.014
## X5  0.107           0.819
## X6                  1.012
##
##                Factor1 Factor2 Factor3
## SS loadings      1.797   1.708   1.705
## Proportion Var   0.300   0.285   0.284
## Cumulative Var   0.300   0.584   0.868
##
## Factor Correlations:
##         Factor1 Factor2 Factor3
## Factor1   1.000  -0.467  -0.501
## Factor2  -0.467   1.000   0.430
## Factor3  -0.501   0.430   1.000
##
## The degrees of freedom for the model is 0 and the fit was 0.0425
```

```
factanal(df, factors=3, rotation="promax")
```

```
##
## Call:
## factanal(x = df, factors = 3, rotation = "promax")
##
## Uniquenesses:
##    X1    X2    X3    X4    X5    X6
## 0.005 0.257 0.161 0.005 0.239 0.005
##
## Loadings:
##    Factor1 Factor2 Factor3
## X1           1.008
## X2           0.844
## X3  0.881
## X4  1.040
## X5                   0.820
## X6                   1.018
##
##                Factor1 Factor2 Factor3
## SS loadings      1.884   1.732   1.719
## Proportion Var   0.314   0.289   0.287
## Cumulative Var   0.314   0.603   0.889
##
## Factor Correlations:
##         Factor1 Factor2 Factor3
## Factor1   1.000   0.506   0.423
## Factor2   0.506   1.000   0.533
## Factor3   0.423   0.533   1.000
##
## The degrees of freedom for the model is 0 and the fit was 0.0425
```

Quiz c :

- **Clearly (but succinctly) explain the relationship between the chosen factors and the original variables.**

Answer

- **X3 and X4 have a strong positive relationship with Factor 1.**
- **X1 and X2 have a strong positive relationship with Factor 2.**
- **X5 and X6 have a strong positive relationship with Factor 3.**

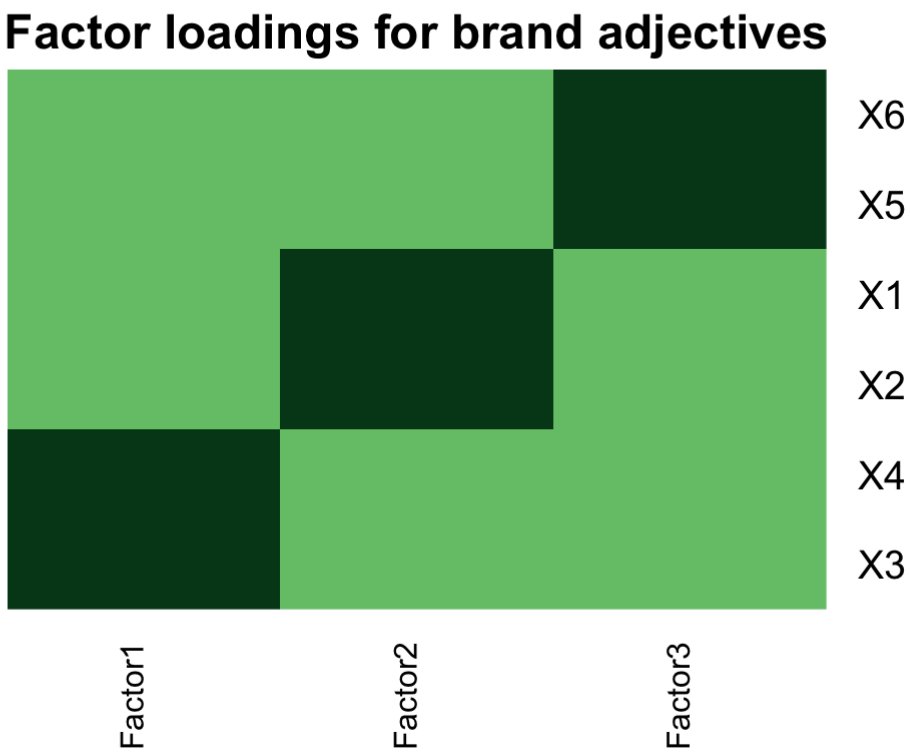
```
library(gplots)
```

```
##
## Attaching package: 'gplots'
```

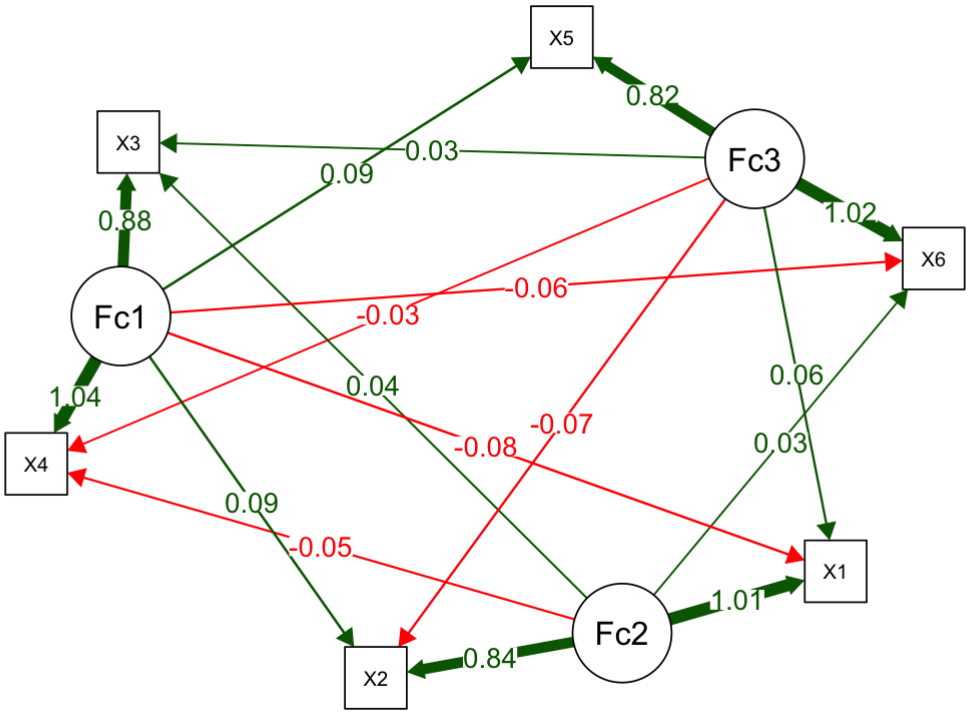
```
## The following object is masked from 'package:stats':
##
##     lowess
```

```
library(RColorBrewer)
```

```
df.fa.ob = factanal(df, factors=3, rotation="promax")

heatmap.2(df.fa.ob$loadings,
          col=brewer.pal(9, "Greens") ,
          trace="none" ,
          key=FALSE ,
          dend="none",
          Colv=FALSE ,
          cexCol = 1.2, main="\n\n\n\n\nFactor loadings for brand adjectives")
```




```
library(semPlot)
semPaths(df.fa.ob,
         what="est" ,
         layout='spring',
         residuals=FALSE,
         cut =0.01,
         posCol=c("white" , "darkgreen") ,
         negCol=c("white" , "red"),
         edge.label.cex =1)
```



Quiz d :

- Using your proposed factor solution, how would you describe the differences between the first two observations in the sample?

Answer

- Both samples are positively associated with three factors.
- However, the first sample seems to be the most correlated with factor 2, but the numbers are not as hig
- The second sample shows a strong association with factors 2 and 3, and a very strong association with factor 3.

```
head(df, 2)
```

```
##        X1      X2       X3      X4      X5      X6
## 1 0.49464 0.84748 -0.18870 0.22944 0.31475 0.14086
## 2 1.29702 0.64319  1.00948 0.55360 2.12976 2.82855
```

```
df.fa.ob <- factanal(df,
                     factors=3,
                     rotation="promax",
                     scores="Bartlett")

head(df.fa.ob$score, 2)
```

```
##        Factor1   Factor2    Factor3
## [1,] 0.1154070 0.2329578 0.02075837
## [2,] 0.4064354 0.7102254 1.79202902
```

```
library(scatterplot3d)
df.score.2 = head(df.fa.ob$score, 2)
scatterplot3d(df.score.2,
              xlim=c(-2, 2),
              ylim=c(-2, 2),
              zlim=c(-2, 2),
              angle=40,
              pch=19,
              color='red',
              )
```