Factor Analysis

Factor analysis aims at reducing the dimensionality of a data set, i.e. the dimensions/variables.

We want to summarize variables into factors (information loss).

- Variables within factors should be highly correlated.
- Factors should be independent from each other, i.e. poorly correlated.

Data requirements

- metric scale level no outliers sufficient correlation between variables (|r|>0.3)
- linear relationship between variables
- at least 5/10 observations per variable
- underlying data should be homogeneous and theoretically consistent, i.e. logical relationship between variables

Steps of factors analysis

1. Calculation and evaluation of correlation matrix
2. Extraction of factors
3. Determination of communalities
4. Number of factors
5. Factor interpretation
6. Determination of factor scores

1. Calculation and evaluation of correlation matrix

Generate the correlation matrix for the variables in the analysis.

- Kaiser-Meyer-Olkin (KMO) criterion: How useful are variables for a factor analysis?

  Shows overall common variability of variables => should be larger than 0.5

- Bartlett-test of sphericity : are correlations between variables significantly different from zero?

  It tests for identity correlation matrix

2.Extraction of factors

Principal component analysis (PCA) and principle axis factoring (PAF)

- PCA most commonly used => best method for information reduction

  Generates the factors as linear combinations of the variables used in the analysis.

  We obtain factor loadings which are correlation coefficients between variables and factors.

- PCA generates factors such that the factor loadings are maximized within an iterative process.

3.Determination of communalities

Factor loadings are correlation coefficients between factors and original variables, i.e. range from -1 to 1. High loading => variable contributes much to the explanation of the factor.

Squared factor loading: What percentage of the variability in a variable is explained by a factor.

Communality: the share of a variable's variability explained by all factors together Calculated by the sum of the squared loadings of the variable across all factors

4.Number of factors

Eigenvalue of a factor: The sum of all squared loadings for one factor

- It gives us the number of average variables that the factor represents

Due to the nature of factor analysis, we lose information contained in the original variables when using factors instead of variables.

- The smaller the number off factors, the larger the information loss.

There is trade-off between the advantages and disadvantages of factor analysis when determining the optimal number of factors.

1. Eigenvalue criterion: Eigenvalues should be above one.

```
Factor explains more variation than an average variable.
```

2. Elbow criterion: Exclude all factors in the flat region and use factors up to the elbow.

5. Factor interpretation

Look at the factor loadings to identify variables that load on the factor.

We can assign names and check for theoretical consistency.

Sometimes factor composition does not make much sense. In that case use factor rotation.

6. Determination of factor scores

We can use factors to

- discover variable structures and create categories of variables.
- for further analysis such as regression models in cases of small samples(=> Determine factor scores).

Application

Let us analyze a data set containing information on characteristics of 50 pastries

```
library(readr)
food <- read.csv("https://userpage.fu-berlin.de/soga/300/30100_data_sets/food-texture.csv",
                 row.names = "X")
head(food)
```

```
##                      Oil    Density    Crispy    Fracture    Hardness
##  B110  16.5         2955          10                23              97
##  B136  17.7         2660          14                 9             139
##  B171  16.2         2870          12                17             143
##  B192  16.7         2920          10                31              95
##  B225  16.3         2975          11                26             143
## B237 19.1    2790     13       16       189
```

Evaluation of the correlation matrix

```
cor(food, use="complete.obs")
```

```
##                    Oil     Density      Crispy    Fracture     Hardness
## Oil         1.00000000 -0.7500240   0.5930863 -0.5337392 -0.09604521
## Density    -0.75002399  1.0000000  -0.6709460  0.5721324  0.10793720
## Crispy      0.59308631 -0.6709460   1.0000000 -0.8439650  0.41109340
## Fracture   -0.53373917  0.5721324  -0.8439650  1.0000000 -0.37335844
## Hardness -0.09604521  0.1079372  0.4110934 -0.3733584  1.00000000
```

Kaiser-Meyer-Olkin criterion

```
library(psych)

## Warning: package 'psych' was built under R version 4.3.3

KMO(food)
```

```
##             Kaiser-Meyer-Olkin             factor             adequacy
##              Call:              KMO(r              =              food)
##           Overall           MSA           =                      0.71
##           MSA           for           each           item           =
```

```
##                  Oil      Density         Crispy   Fracture   Hardness
##     0.82     0.71     0.67     0.79     0.43
```

Bartlett-test of sphericity

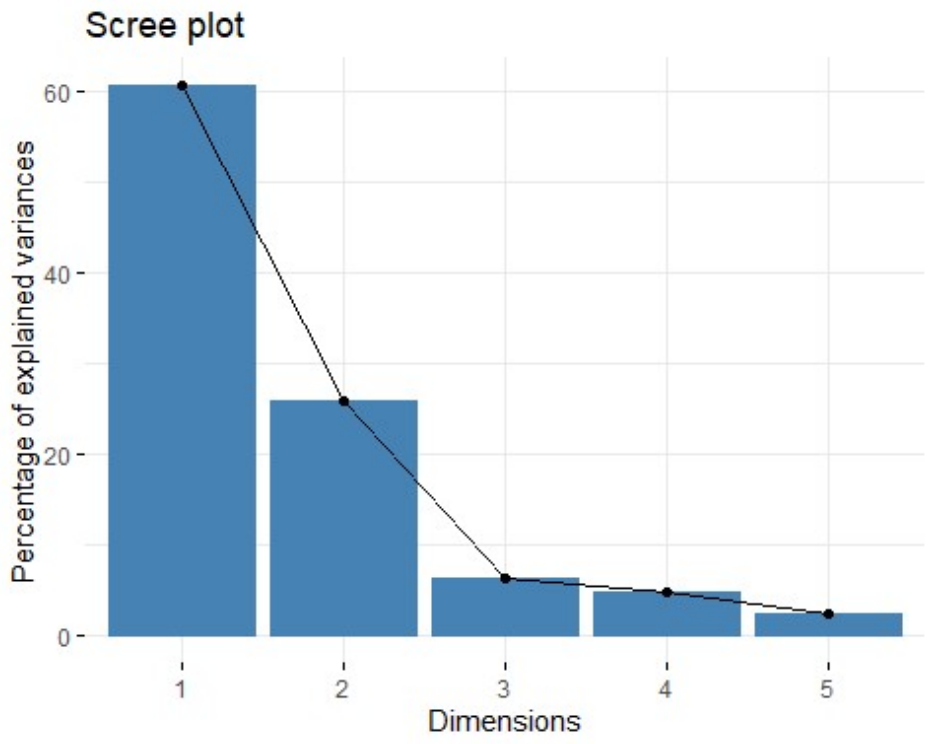
```
library(REdaS)

## Warning: package 'REdaS' was built under R version 4.3.3

## Loading required package: grid

bart_spher(food, use = c("complete.obs"))

##                  Bartlett's         Test         of         Sphericity
##
##    Call:    bart_spher(x    =    food,    use    =    c("complete.obs"))
##
##                                              X2        =        154.994
##                                                  df        =         10
## p-value < 2.22e-16
```

Extraction of factors, factor loadings and eigenvalues

```
pca=principal(na.omit(food),          nfactors=5,          rotate="none")
pca

##              Principal              Components              Analysis
## Call: principal(r = na.omit(food), nfactors = 5, rotate = "none")
## Standardized loadings (pattern matrix) based upon correlation matrix
##                  PC1     PC2     PC3     PC4     PC5  h2          u2  com
##  Oil             0.80  -0.42    0.37    0.23    0.00    1  5.6e-16  2.2
##  Density    -0.83    0.41    0.01    0.35    0.12    1  3.3e-16  1.9
##  Crispy        0.93    0.22  -0.10  -0.07    0.28    1  1.1e-15  1.3
##  Fracture  -0.88  -0.25    0.30  -0.22    0.15    1  1.1e-15  1.6
##  Hardness    0.27    0.92    0.27  -0.10  -0.08    1  2.2e-16  1.4
##
##                                        PC1    PC2    PC3    PC4    PC5
##  SS  loadings                         3.03   1.30   0.31   0.24   0.12
##  Proportion  Var                     0.61   0.26   0.06   0.05   0.02
##  Cumulative  Var                     0.61   0.87   0.93   0.98   1.00
##   Proportion    Explained        0.61    0.26    0.06    0.05    0.02
##    Cumulative    Proportion    0.61    0.87    0.93    0.98    1.00
##
##       Mean        item        complexity        =                1.7
##  Test  of  the  hypothesis  that  5  components  are  sufficient.
##
##  The  root  mean  square  of  the  residuals  (RMSR)  is     0
##    with  the  empirical  chi  square    0    with  prob  <    NA
##
## Fit based upon off diagonal values = 1
```

Elbow criterion

```
pcae=princomp(na.omit(food),          cor=TRUE,          scores=TRUE)
library(factoextra)

## Warning: package 'factoextra' was built under R version 4.3.3

## Loading required package: ggplot2

##
## Attaching package: 'ggplot2'

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

## Welcome! Want to learn more? See two factoextra-related books at
https://goo.gl/ve3WBa

fviz_eig(pcae)
```



Determination of communalities

```
pcar=principal(na.omit(food),          nfactors=2,          rotate="none")
pcar

##              Principal              Components              Analysis
## Call: principal(r = na.omit(food), nfactors = 2, rotate = "none")
## Standardized loadings (pattern matrix) based upon correlation matrix
##                                PC1        PC2        h2        u2   com
##    Oil                            0.80    -0.42    0.81    0.19    1.5
```

```
##     Density          -0.83          0.41     0.86     0.14      1.4
##    Crispy                0.93         0.22    0.91     0.09     1.1
##      Fracture       -0.88       -0.25       0.83       0.17       1.2
##     Hardness          0.27          0.92     0.91     0.09      1.2
##
##                                                        PC1     PC2
##   SS    loadings                                       3.03    1.30
##    Proportion    Var                                   0.61    0.26
##    Cumulative    Var                                   0.61    0.87
##        Proportion        Explained                0.70         0.30
##          Cumulative          Proportion          0.70          1.00
##
##       Mean        item        complexity        =               1.3
##  Test  of  the  hypothesis  that  2  components  are  sufficient.
##
##  The  root  mean  square  of  the  residuals  (RMSR)  is     0.06
##   with  the  empirical  chi  square    3.55    with  prob  <    0.06
##
## Fit based upon off diagonal values = 0.99

#communalities
pcar$communality

##              Oil      Density        Crispy    Fracture    Hardness
## 0.8123477 0.8596509 0.9097799 0.8348555 0.9102849
```

Let us now perform a factor rotation. The base loadings are

```
pcar

##              Principal              Components              Analysis
## Call: principal(r = na.omit(food), nfactors = 2, rotate = "none")
## Standardized loadings (pattern matrix) based upon correlation matrix
##                                PC1        PC2        h2        u2   com
##   Oil                             0.80    -0.42    0.81    0.19    1.5
##     Density          -0.83          0.41     0.86     0.14      1.4
##    Crispy                0.93         0.22    0.91     0.09     1.1
##      Fracture       -0.88       -0.25       0.83       0.17       1.2
##     Hardness          0.27          0.92     0.91     0.09      1.2
##
##                                                        PC1     PC2
##   SS    loadings                                       3.03    1.30
##    Proportion    Var                                   0.61    0.26
##    Cumulative    Var                                   0.61    0.87
##        Proportion        Explained                0.70         0.30
##          Cumulative          Proportion          0.70          1.00
##
##       Mean        item        complexity        =               1.3
##  Test  of  the  hypothesis  that  2  components  are  sufficient.
##
##  The  root  mean  square  of  the  residuals  (RMSR)  is     0.06
```

```
##    with  the  empirical  chi  square    3.55    with  prob  <    0.06
##
## Fit based upon off diagonal values = 0.99
```

Performing a rotation with *varimax*

```
pcarot=principal(na.omit(food),      nfactors=2,      rotate="varimax")
pcarot
```

```
##              Principal              Components              Analysis
## Call: principal(r = na.omit(food), nfactors = 2, rotate = "varimax")
## Standardized loadings (pattern matrix) based upon correlation matrix
##                                RC1        RC2        h2        u2   com
##    Oil                          -0.90    -0.08    0.81     0.19     1.0
##     Density              0.93          0.05     0.86     0.14     1.0
##     Crispy              -0.77          0.57     0.91     0.09     1.8
##      Fracture            0.71      -0.57      0.83      0.17      1.9
##     Hardness           0.11           0.95     0.91     0.09     1.0
##
##                                                          RC1     RC2
##    SS    loadings                                       2.77    1.56
##     Proportion     Var                                  0.55    0.31
##     Cumulative     Var                                  0.55    0.87
##         Proportion        Explained                0.64         0.36
##           Cumulative           Proportion          0.64          1.00
##
##        Mean        item        complexity        =                1.4
##  Test  of  the  hypothesis  that  2  components  are  sufficient.
##
##  The  root  mean  square  of  the  residuals  (RMSR)  is     0.06
##    with  the  empirical  chi  square    3.55    with  prob  <    0.06
##
## Fit based upon off diagonal values = 0.99
```

We perform a rotation with *oblimin*

```
library(GPArotation)
```

```
## Warning: package 'GPArotation' was built under R version 4.3.3
```

```
##
## Attaching package: 'GPArotation'
```

```
##   The   following   objects   are   masked   from   'package:psych':
##
##     equamax, varimin
```

```
pcarotobl=principal(na.omit(food),     nfactors=2,     rotate="oblimin")
pcarotobl
```

```
##              Principal              Components              Analysis
## Call: principal(r = na.omit(food), nfactors = 2, rotate = "oblimin")
```

```
## Standardized loadings (pattern matrix) based upon correlation matrix
##                              TC1        TC2        h2        u2   com
##    Oil                        -0.90    -0.24    0.81    0.19    1.1
##     Density              0.93          0.22     0.86     0.14     1.1
##     Crispy               -0.80          0.42     0.91     0.09     1.5
##      Fracture            0.74      -0.44      0.83      0.17      1.6
##     Hardness           0.07           0.96     0.91     0.09     1.0
##
##                                                        TC1     TC2
##    SS    loadings                                      2.90    1.43
##     Proportion     Var                                 0.58    0.29
##     Cumulative     Var                                 0.58    0.87
##         Proportion        Explained                 0.67         0.33
##           Cumulative           Proportion          0.67          1.00
##
##                 With         component         correlations         of
##                                                TC1                   TC2
##              TC1                                 1.00               -0.13
##              TC2              -0.13                                  1.00
##
##        Mean        item        complexity        =                 1.3
##  Test   of   the   hypothesis   that   2   components   are   sufficient.
##
##  The   root   mean   square   of   the   residuals   (RMSR)   is     0.06
##    with  the  empirical  chi  square    3.55    with  prob  <    0.06
##
## Fit based upon off diagonal values = 0.99
```

Finally we get the Factor scoores

```
pcarotobl$scores[1:10,]
```

```
##                                      TC1                       TC2
##           B110                      0.65887766           -0.78470238
##           B136           -1.50069373                       0.84507630
##         B171                  0.02948447                   0.75744047
##           B192                      0.87215597           -1.32556823
##         B225                  0.83656328                   0.29603917
##           B237           -0.87516851                       1.46623078
##              B261              -0.95428857                -0.22701476
##           B264                      0.01485081           -2.12889574
##         B353                  0.73094479                   0.05408441
## B360  0.62354087  0.09257588
```