Course: Multivariate statistics (AUT23)

Chapter 9: Exploratory factor analysis

9.14.1 Exercise 1: Big-5

To illustrate EFA, let us use the International Personality Item Pool data available in the psych package. It includes 25 personality self-report items following the big 5 personality structure.

The first step is to test if the dataset is suitable for conducting factor analysis. To do so, run the Bartlett's Test of Sphericity and the Kaiser Meyer Olkin (KMO) measure.

Reminder: Bartlett's Test of Sphericity tests whether a matrix (of correlations) is significantly different from an identity matrix (probability that the correlation matrix has significant correlations among at least some of the variables in a dataset). KMO measure indicates the degree to which each variable in a set is predicted without error by the other variables (a KMO value close to 1 indicates that the sum of partial correlations is not large relative to the sum of correlations and so factor analysis should yield distinct and reliable factors).

- Show the code

Once you are confident that the dataset is appropriate for factor analysis, you can explore a factor structure made of 5 (theoretically motivated) latent variables. Start by defining the model.

- Show the code

What do you see? How can you interpret the output?

- Solution: Interpretation

It is possible to visualize the results to ease the interpretation:

- Show the code

Based on this model, you could predict back the scores for each individual for these new variables. This could be useful for further analysis (e.g. regression analysis).

Tips: use the function predict() and give labels to the latent factors. Based on this model, you could predict back the scores for each individual for these new variables. This could be useful for further analysis (e.g. regression analysis).

Tips: use the function predict() and give labels to the latent factors.

- Show the code

9.14.2 Exercise 2: Environmental concerns

We will use survey data from the World Values Survey (WVS) website to investigate human belief and values, especially about environmental (EC) We will analyse Swiss data derived from the much larger WVS cross-country database (2007). The data can be downloaded here.

EC has been measured by a set of 8 items on a four-step Likert Scale. These items are thought to load on three different facets of EC: concerns about one's own community (water quality, air quality and sanitation), concerns about the world at large (fears about global warming, loss of biodiversity and ocean pollution), willingness to pay/do more. We want to answer the following question: Is the three-factor model proposed by the structure of items (or can we assume a one-factor structure)?

Let's first prepare the data and get the table of correlations:

- Show the code

The first step is to test if the dataset is suitable for conducting factor analysis. To do so, run the Bartlett's Test of Sphericity and the Kaiser Meyer Olkin (KMO) measure.

- Show the code

Now, you can explore a factor structure made of 3 (theoretically motivated) latent variables. Start by defining the model.

- Show the code

What do you see? How can you interpret the output?

- Solution: Interpretation

It is possible to visualize the results to ease the interpretation:

- Show the code

Chapter 9: Exploratory factor analysis (answers)

9.14.1 Exercise 1: Big-5

To illustrate EFA, let us use the International Personality Item Pool data available in the psych package. It includes 25 personality self-report items following the big 5 personality structure.

The first step is to test if the dataset is suitable for conducting factor analysis. To do so, run the Bartlett's Test of Sphericity and the Kaiser Meyer Olkin (KMO) measure.

Reminder: Bartlett's Test of Sphericity tests whether a matrix (of correlations) is significantly different from an identity matrix (probability that the correlation matrix has significant correlations among at least some of the variables in a dataset). KMO measure indicates the degree to which each variable in a set is predicted without error by the other variables (a KMO value close to 1 indicates that the sum of partial correlations is not large relative to the sum of correlations and so factor analysis should yield distinct and reliable factors).

➢ Show the code

```
# load the data
data <- psych::bfi[, 1:25] # 25 first columns corresponding to the items
data <- na.omit(data)
# option 1: check suitability
psych::KMO(data)
## Kaiser-Meyer-Olkin factor adequacy
## Call: psych::KMO(r = data)
## Overall MSA =  0.85
## MSA for each item =
##   A1   A2   A3   A4   A5   C1   C2   C3   C4   C5   E1   E2   E3   E4   E5   N1   N2   N3
## 0.75 0.84 0.87 0.88 0.90 0.84 0.80 0.85 0.83 0.86 0.84 0.88 0.90 0.88 0.89 0.78 0.78 0.86
##   N4   N5   O1   O2   O3   O4   O5
## 0.89 0.86 0.86 0.78 0.84 0.77 0.76
bartlett = psych::cortest.bartlett(data)
## R was not square, finding R from data
print(paste0("Chi-2: ", round(bartlett[["chisq"]],2),
             "; p-value: ", bartlett[["p.value"]]))
## [1] "Chi-2: 18146.07; p-value: 0"
```

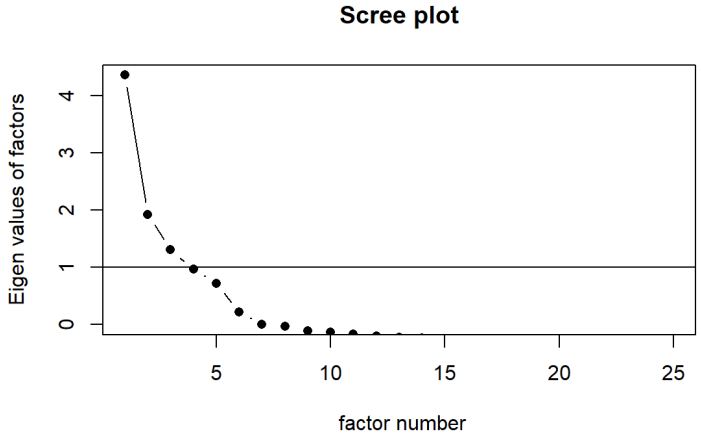
```
# option 2: check suitability

# performance::check_factorstructure(data)
```

Once you are confident that the dataset is appropriate for factor analysis, you can explore a factor structure made of 5 (theoretically motivated) latent variables. Start by defining the model.

➢ Show the code

```
# optional: eigenvalues

ev <- eigen(cor(data))

psych::scree(data, pc=FALSE)

# fit an EFA

efa <- psych::fa(data, nfactors = 5, rotate="oblimin")

## Le chargement a nécessité le package : GPArotation

efa_para <- psych::fa(data, nfactors = 5, rotate="oblimin") |>

  parameters::model_parameters(sort = TRUE, threshold = "max")

efa_para

## # Rotated loadings from Factor Analysis (oblimin-rotation)

##

## Variable | MR2  |  MR1  |  MR3  |  MR5  |  MR4  | Complexity | Uniqueness

## -----------------------------------------------------------------------

## N1       | 0.83 |       |       |       |       |    1.07    |    0.32

## N2       | 0.78 |       |       |       |       |    1.03    |    0.39

## N3       | 0.70 |       |       |       |       |    1.08    |    0.46

## N5       | 0.48 |       |       |       |       |    2.00    |    0.65

## N4       | 0.47 |       |       |       |       |    2.33    |    0.49

## E2       |      | 0.67  |       |       |       |    1.08    |    0.45

## E4       |      | -0.59 |       |       |       |    1.52    |    0.46

## E1       |      | 0.55  |       |       |       |    1.22    |    0.65

## E5       |      | -0.42 |       |       |       |    2.68    |    0.59

## E3       |      | -0.41 |       |       |       |    2.65    |    0.56

## C2       |      |       | 0.67  |       |       |    1.18    |    0.55
```

```
## C4  |  |  | -0.64 |  |  |  1.13  |  0.52
## C3  |  |  | 0.57  |  |  |  1.10  |  0.68
## C5  |  |  | -0.56 |  |  |  1.41  |  0.56
## C1  |  |  | 0.55  |  |  |  1.20  |  0.65
## A3  |  |  |  | 0.68  |  |  1.06  |  0.46
## A2  |  |  |  | 0.66  |  |  1.03  |  0.54
## A5  |  |  |  | 0.54  |  |  1.48  |  0.53
## A4  |  |  |  | 0.45  |  |  1.74  |  0.70
## A1  |  |  |  | -0.44 |  |  1.88  |  0.80
## O3  |  |  |  |  | 0.62  |  1.16  |  0.53
## O5  |  |  |  |  | -0.54 |  1.21  |  0.70
## O1  |  |  |  |  | 0.52  |  1.10  |  0.68
## O2  |  |  |  |  | -0.47 |  1.68  |  0.73
## O4  |  |  |  |  | 0.36  |  2.65  |  0.75
##
## The 5 latent factors (oblimin rotation) accounted for 42.36% of the total variance of the original data (MR2 = 10.31%, MR1 = 8.83%, MR3 = 8.39%, MR5 = 8.29%, MR4 = 6.55%).
```



What do you see? How can you interpret the output?

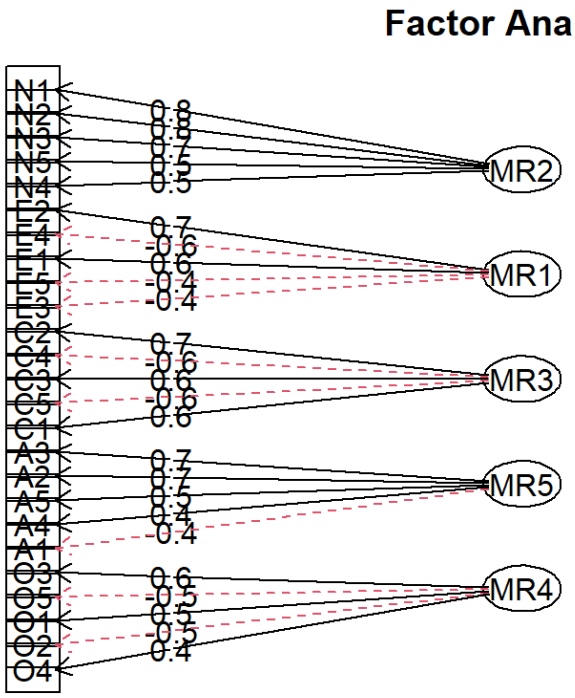
- Solution: Interpretation

The 25 items spread on the 5 latent factors nicely - the famous big 5.

It is possible to visualize the results to ease the interpretation:

➢ Show the code

```
loads <- efa$loadings

psych::fa.diagram(loads)
```



Based on this model, you could predict back the scores for each individual for these new variables. This could be useful for further analysis (e.g. regression analysis).

Tips: use the function predict() and give labels to the latent factors. Based on this model, you could predict back the scores for each individual for these new variables. This could be useful for further analysis (e.g. regression analysis).

Tips: use the function predict() and give labels to the latent factors.

➢ Show the code

```
# predictions <- predict(

#   efa_para,

#   names = c("Neuroticism", "Conscientiousness", "Extraversion", "Agreeableness", "Opennness"),

#   verbose = FALSE

# )

# head(predictions)
```

9.14.2 Exercise 2: Environmental concerns

We will use survey data from the World Values Survey (WVS) website to investigate human belief and values, especially about environmental (EC) We will analyse Swiss data derived from the much larger WVS cross-country database (2007). The data can be downloaded here.

EC has been measured by a set of 8 items on a four-step Likert Scale. These items are thought to load on three different facets of EC: concerns about one's own community (water quality, air quality and sanitation), concerns about the world at large (fears about global warming, loss of biodiversity and ocean pollution), willingness to pay/do more. We want to answer the following question: Is the three-factor model proposed by the structure of items (or can we assume a one-factor structure)?

Let's first prepare the data and get the table of correlations:

➢ Show the code

```
db <- openxlsx::read.xlsx(paste0(getwd(),
                "/data/WV5_Data_Switzerland_Excel_v20201117.xlsx"))
colnames(db) <- gsub(":.*","",colnames(db))
sel <- db |>
  dplyr::select(V108,V109,V110,
                V111,V112,V113,
                V106,V107
                ) |>
  dplyr::rename("water"="V108",
                "air"="V109",
                "sanitation"="V110",
                "warming"="V111",
                "biodiv"="V112",
                "pollution"="V113",
                "taxes"="V106",
                "gov"="V107") |>
  stats::na.omit()
sel = replace(sel, sel==-1, NA)
sel = sel[complete.cases(sel),]
# reverse scale
```

```
for(i in 1:ncol(sel)){sel[,i] <- (sel[,i]-5)*(-1)}
# correlation
round(cor(sel),2)
##            water  air sanitation warming biodiv pollution taxes   gov
## water       1.00 0.73       0.84    0.04   0.06      0.10  0.00  0.09
## air         0.73 1.00       0.74    0.14   0.14      0.15  0.07  0.03
## sanitation  0.84 0.74       1.00    0.06   0.08      0.12  0.00  0.11
## warming     0.04 0.14       0.06    1.00   0.49      0.39  0.20  0.01
## biodiv      0.06 0.14       0.08    0.49   1.00      0.45  0.13  0.09
## pollution   0.10 0.15       0.12    0.39   0.45      1.00  0.14  0.00
## taxes       0.00 0.07       0.00    0.20   0.13      0.14  1.00 -0.23
## gov         0.09 0.03       0.11    0.01   0.09      0.00 -0.23  1.00
```

The first step is to test if the dataset is suitable for conducting factor analysis. To do so, run the Bartlett's Test of Sphericity and the Kaiser Meyer Olkin (KMO) measure.

➢ Show the code

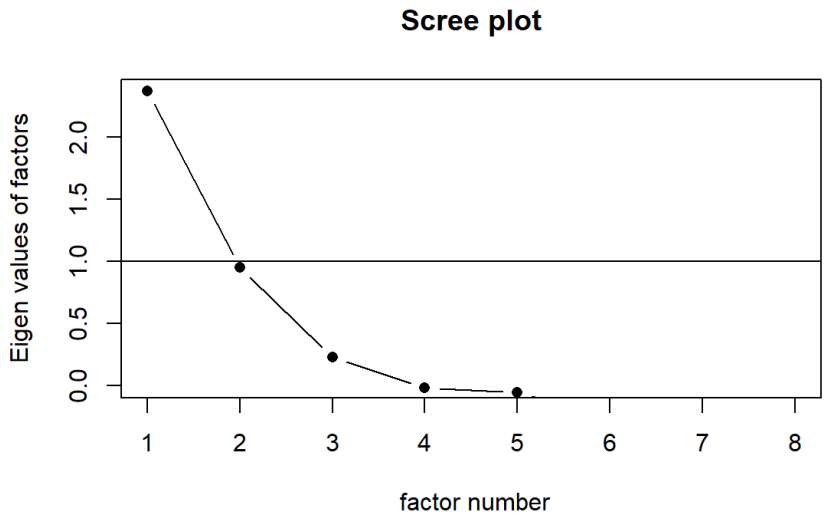
```
# option 1: check suitability
psych::KMO(sel)
## Kaiser-Meyer-Olkin factor adequacy
## Call: psych::KMO(r = sel)
## Overall MSA =  0.72
## MSA for each item =
##     water       air sanitation    warming     biodiv  pollution      taxes       gov
##      0.71      0.84      0.70      0.69      0.66      0.74      0.62      0.50
bartlett = psych::cortest.bartlett(sel)
## R was not square, finding R from data
print(paste0("Chi-2: ", round(bartlett[["chisq"]],2),
             "; p-value: ", bartlett[["p.value"]]))
## [1] "Chi-2: 3331.64; p-value: 0"
# option 2: check suitability
```

```
# performance::check_factorstructure(sel)
```

Now, you can explore a factor structure made of 3 (theoretically motivated) latent variables. Start by defining the model.

➢ Show the code

```
# optional: eigenvalues

ev <- eigen(cor(sel))

psych::scree(sel, pc=FALSE)

# fit an EFA

efa <- psych::fa(sel, nfactors = 3, rotate="oblimin")

## Le chargement a nécessité le package : GPArotation

efa_para <- psych::fa(sel, nfactors = 3, rotate="oblimin") |>

  parameters::model_parameters(sort = TRUE, threshold = "max")

efa_para

## # Rotated loadings from Factor Analysis (oblimin-rotation)

##

## Variable   | MR1  | MR2  |  MR3  | Complexity | Uniqueness

## -----------------------------------------------------------

## sanitation | 0.93 |      |       |    1.01    |    0.14

## water      | 0.92 |      |       |    1.00    |    0.17

## air        | 0.79 |      |       |    1.04    |    0.34

## biodiv     |      | 0.79 |       |    1.02    |    0.40

## warming    |      | 0.64 |       |    1.05    |    0.56

## pollution  |      | 0.57 |       |    1.04    |    0.65

## gov        |      |      | -0.53 |    1.14    |    0.72

## taxes      |      |      | 0.48  |    1.20    |    0.73

##

## The 3 latent factors (oblimin rotation) accounted for 53.70% of the total variance of the original data (MR1 = 29.23%, MR2 = 17.66%, MR3 = 6.81%).
```



What do you see? How can you interpret the output?

- Solution: Interpretation

The 8 items spread on the 3 latent factors nicely. Note that the 'gov' item has a negative loading.

It is possible to visualize the results to ease the interpretation:

- Show the code

```
loads <- efa$loadings

psych::fa.diagram(loads)
```