Homework 8

Darian-Florian Voda

2022-12-22

Loading packages

```
library(psych)
library(rela)
```

Exercise 87: Example from sociology

- A survey was designed at the Institute of Sociology at the University of Marburg and conducted at two metalworking companies in Hesse on "attitudes towards foreigners". The respondents were presented with 15 statements. The answers were each given on a seven-point scale, intended to represent complete disagreement (1) to complete agreement (7). The results of 90 respondents were stored in variables a1 to a15 of the PCA-foreigner.txt file.
- Questions of questionnaire
  - a1 The integration of foreigners must be improved.
  - a2 Refugee money must be reduced.
  - a3 German money should be spent on German issues.
  - a4 Germany is not the social welfare office of the world.
  - a5 Good coexistence must be striven for.
  - a6 The right of asylum should be restricted.
  - a7 Germans are becoming a minority.
  - a8 The right of asylum must be protected throughout Europe.
  - a9 Xenophobia harms the German economy.
  - a10 Housing should be created for Germans first.
  - a11 We are foreigners too, almost everywhere.
  - a12 Multicultural means multicriminal.
  - a13 The boat is full.
- Level of Agreement
  - 1 Strongly disagree
  - 2 Disagree
  - 3 Somewhat disagree
  - 4 Neither agree or disagree
  - 5 Somewhat agree
  - 6 Agree
  - 7 Strongly agree
- Aim of the study

  - 1 The aim of this study is to use the data-set PCA-foreigner.txt and perform a PCA to identify the main components/dimensions.
  - 2 Interpretation and naming of the detected underlying components.
  - 3 Then, perform a reliability analyses of the extracted dimensions.
  - 4 As a next step, scores for every of the new dimensions should be calculated and stored as new variables.
  - 5 Finally, find associations between the new variables and the other variables in the data set.

- Research questions

  - 1 How many factors can be extracted?
  - 2 Which items can be assigned to the extracted components?
  - 3 What is the reliability (Crohnbach's Alpha) of each of the new components?
  - 4 Is there an association between the extracted components and
    * the satisfaction with own economic circumstances?
    * age?
    * gender?
    * socio-political commitment?
    * the position in the company?

```
foreignsurvey<-read.delim("C:/Users/daria/OneDrive/Desktop/Master - AppDS/Statistics/Datasets-20221007/
                      stringsAsFactors=F)


head(foreignsurvey)
```

```
##   nr ecosituation commitment position yearbirth    sex a1 a2 a3 a4 a5 a6 a7 a8
## 1  1    uncertain         no Employee 1951-1960   male  6  5  4  2  6  5  2  5
## 2  2          yes  uncertain Employee 1961-1970 female  6  6  4  3  7  7  2  7
## 3  3    uncertain         no Employee 1961-1970 female  5  7  7  6  6  6  6  5
## 4  4          yes  uncertain Employee 1961-1970   male  7  6  3  2  7  7  2  5
## 5  5          yes         no Employee 1961-1970   male  7  5  5  2  7  2  2  6
## 6  6    uncertain        yes Employee 1951-1960   male  7  7  3  1  7  7  3  7
##   a9 a10 a11 a12 a13 a14 a15
## 1  6   2   7   4   3   2   1
## 2  7   4   7   4   1   1   1
## 3  5   7   5   4   7   4   1
## 4  2   1   2   1   1   1   1
## 5  7   1   7   1   1   1   1
## 6  7   4   7   2   5   1   2
```

```
### Find number of principal components

# Remove NA vals
foreignsurvey = na.omit(foreignsurvey)

# Remove nr, ecosituation, commitment, position, yearbirth, sex

foreignsurvey$nr = NULL
foreignsurvey$ecosituation = NULL
foreignsurvey$commitment = NULL
foreignsurvey$position = NULL
foreignsurvey$yearbirth = NULL
```

```r
foreignsurvey$sex = NULL

# KMO + Bartlett statistics

paf.obj <- paf(as.matrix(foreignsurvey))
cat("KMO statistics:", paf.obj$KMO, " Bartlett statistics:",  paf.obj$Bartlett, "\n")
```

```
## KMO statistics: 0.80049  Bartlett statistics: 513.68
```

```r
paste("KMO: 0.800 - very good")
```

```
## [1] "KMO: 0.800 - very good"
```

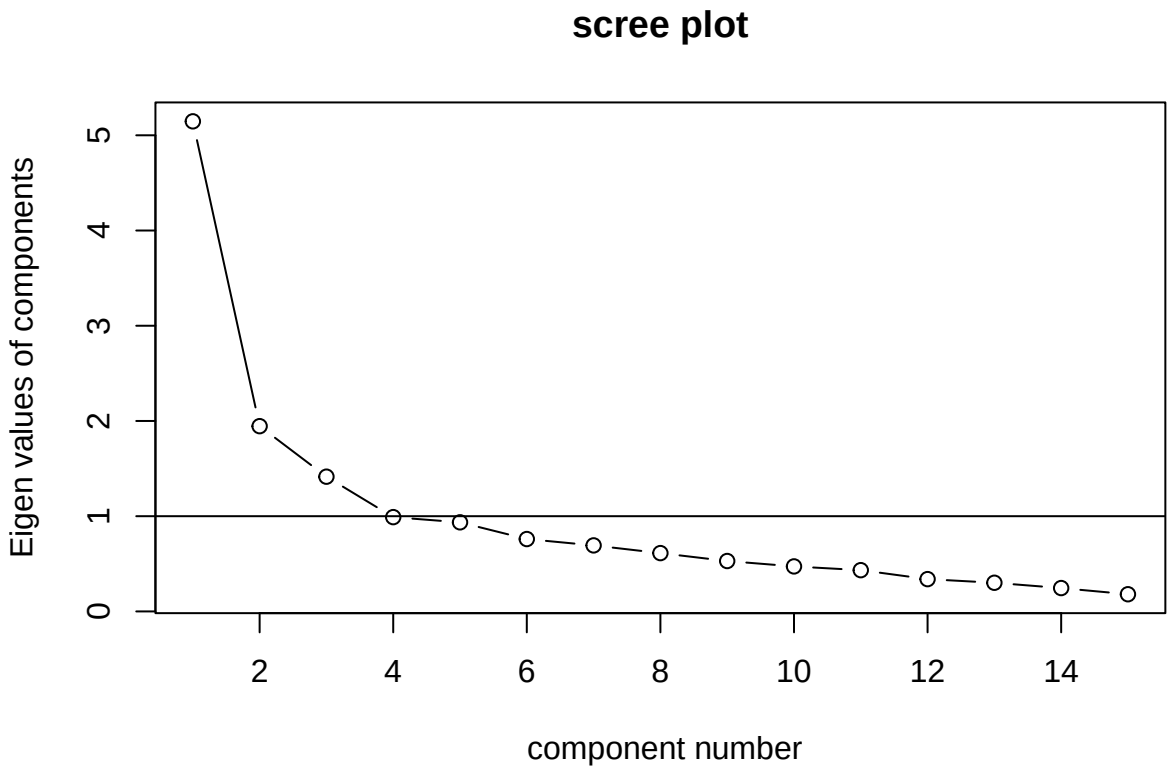
```r
bart <- cortest.bartlett(cor(foreignsurvey), n = nrow(foreignsurvey))
unlist(bart)
```

```
##      chisq    p.value         df
## 5.1368e+02 4.0071e-55 1.0500e+02
```

```r
paste("Bartlett: < 0.05, thus we can use PCA test since not all correlation are zero")
```

```
## [1] "Bartlett: < 0.05, thus we can use PCA test since not all correlation are zero"
```

```r
# VSS Scree on data

VSS.scree(foreignsurvey)
```




```
paste("Our 15 variables measure 3 underlying components")
```

```
## [1] "Our 15 variables measure 3 underlying components"
```

```
# Extract four components and do not perform rotation

pca.foreign <- principal(foreignsurvey, 3, rotate = "none")
pca.foreign$criteria <- NULL
pca.foreign
```

```
## Principal Components Analysis
## Call: principal(r = foreignsurvey, nfactors = 3, rotate = "none")
## Standardized loadings (pattern matrix) based upon correlation matrix
##       PC1   PC2   PC3   h2   u2 com
## a1  -0.72  0.35  0.04 0.65 0.35 1.5
## a2  -0.29  0.64 -0.01 0.50 0.50 1.4
## a3   0.66  0.31 -0.31 0.63 0.37 1.9
## a4   0.64  0.20  0.05 0.45 0.55 1.2
## a5  -0.60  0.62  0.00 0.75 0.25 2.0
## a6   0.39  0.39 -0.53 0.58 0.42 2.7
## a7   0.64  0.40  0.00 0.57 0.43 1.7
## a8  -0.54  0.43  0.35 0.60 0.40 2.7
## a9  -0.27  0.41  0.27 0.31 0.69 2.5
## a10  0.57  0.33 -0.32 0.54 0.46 2.3
## a11 -0.23  0.18  0.41 0.25 0.75 2.0
```

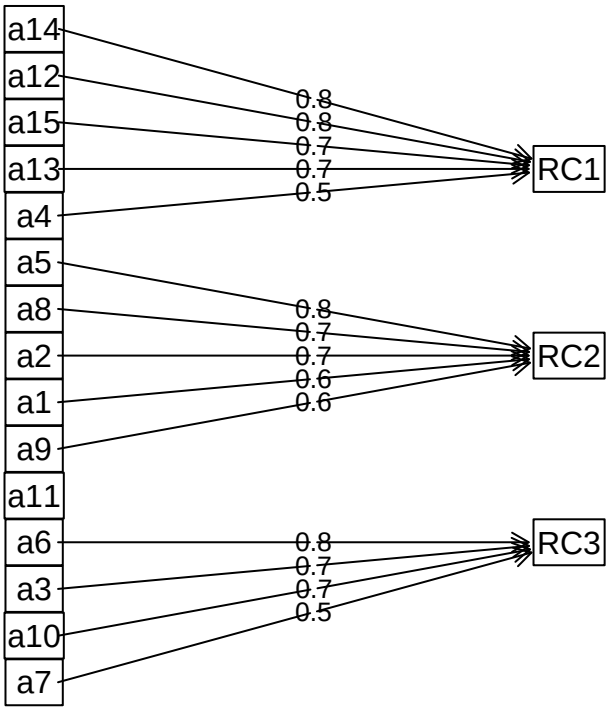
```
## a12  0.74  0.00  0.39 0.69 0.31 1.5
## a13  0.79  0.26  0.10 0.70 0.30 1.2
## a14  0.67 -0.02  0.52 0.72 0.28 1.9
## a15  0.63  0.13  0.37 0.55 0.45 1.7
##
##                        PC1  PC2  PC3
## SS loadings           5.15 1.95 1.41
## Proportion Var        0.34 0.13 0.09
## Cumulative Var        0.34 0.47 0.57
## Proportion Explained  0.60 0.23 0.17
## Cumulative Proportion 0.60 0.83 1.00
##
## Mean item complexity =  1.9
## Fit based upon off diagonal values = 0.95
```

```
# Rotate with varimax pca

pca.foreign.r <- principal(foreignsurvey, 3)
pca.foreign.r$criteria <- NULL
print(pca.foreign.r, cut = 0.5, sort = TRUE, digits = 2)
```

```
## Principal Components Analysis
## Call: principal(r = foreignsurvey, nfactors = 3)
## Standardized loadings (pattern matrix) based upon correlation matrix
##     item   RC1   RC2   RC3   h2   u2 com
## a14   14  0.84             0.72 0.28 1.1
## a12   12  0.80             0.69 0.31 1.2
## a15   15  0.73             0.55 0.45 1.1
## a13   13  0.69             0.70 0.30 1.8
## a4     4  0.53             0.45 0.55 1.9
## a5     5        0.78       0.75 0.25 1.4
## a8     8        0.72       0.60 0.40 1.3
## a2     2        0.66       0.50 0.50 1.3
## a1     1        0.63       0.65 0.35 2.1
## a9     9        0.55       0.31 0.69 1.1
## a11   11                   0.25 0.75 2.1
## a6     6              0.76 0.58 0.42 1.0
## a3     3              0.71 0.63 0.37 1.5
## a10   10              0.69 0.54 0.46 1.3
## a7     7  0.53        0.54 0.57 0.43 2.0
##
##                        RC1  RC2  RC3
## SS loadings           3.47 2.54 2.50
## Proportion Var        0.23 0.17 0.17
## Cumulative Var        0.23 0.40 0.57
## Proportion Explained  0.41 0.30 0.29
## Cumulative Proportion 0.41 0.71 1.00
##
## Mean item complexity =  1.5
## Fit based upon off diagonal values = 0.95
```

```
# fa diagram

fa.diagram(pca.foreign.r, cut = 0.5, cex = 0.8, rsize = 0.5, main = "FA Diagram")
```

FA Diagram




```
# a14 Foreigners out.
# a12 Multicultural means multicriminal.
# a15 Foreigner integration is genocide.
# a13 The boat is full.
# a4 Germany is not the social welfare office of the world.
#
# C1: Discriminatory behavior
#
#
# a5 Good coexistence must be striven for.
# a8 The right of asylum must be protected throughout Europe.
# a2 Refugee money must be reduced.
# a1 The integration of foreigners must be improved.
# a9 Xenophobia harms the German economy.
#
# C2: Economical and Integrity based behaviour
#
# a6 The right of asylum should be restricted.
# a3 German money should be spent on German issues.
# a10 Housing should be created for Germans first.
# a7 Germans are becoming a minority.
#
# C3: Patriotic behavior

# C1: Discriminatory behavior
alpha(subset(foreignsurvey, select = c(a14, a12, a15, a13, a4)), check.keys =TRUE)
```

```
##
## Reliability analysis
## Call: alpha(x = subset(foreignsurvey, select = c(a14, a12, a15, a13,
##     a4)), check.keys = TRUE)
##
##   raw_alpha std.alpha G6(smc) average_r S/N  ase mean  sd median_r
##       0.82      0.83    0.82       0.5 4.9 0.03  3.1 1.4     0.49
##
##     95% confidence boundaries
##          lower alpha upper
## Feldt     0.75  0.82  0.87
## Duhachek  0.76  0.82  0.88
##
##  Reliability if an item is dropped:
##     raw_alpha std.alpha G6(smc) average_r S/N alpha se  var.r med.r
## a14      0.78      0.78    0.75      0.48 3.6    0.037 0.0123  0.48
## a12      0.76      0.78    0.74      0.47 3.5    0.039 0.0118  0.48
## a15      0.79      0.80    0.77      0.51 4.1    0.036 0.0086  0.48
## a13      0.76      0.78    0.74      0.47 3.5    0.042 0.0151  0.46
## a4       0.82      0.83    0.80      0.56 5.1    0.029 0.0043  0.55
##
##  Item statistics
##      n raw.r std.r r.cor r.drop mean  sd
## a14 90  0.76  0.80  0.75   0.67  1.7 1.3
## a12 90  0.80  0.82  0.76   0.68  2.7 1.8
## a15 90  0.76  0.76  0.68   0.60  2.6 1.9
## a13 90  0.84  0.81  0.76   0.70  4.1 2.1
## a4  90  0.69  0.67  0.54   0.49  4.2 2.0
##
## Non missing response frequency for each item
##        1    2    3    4    5    6    7 miss
## a14 0.66 0.17 0.09 0.02 0.03 0.02 0.01    0
## a12 0.34 0.22 0.10 0.18 0.06 0.06 0.04    0
## a15 0.44 0.17 0.07 0.12 0.08 0.07 0.06    0
## a13 0.19 0.10 0.09 0.17 0.13 0.13 0.19    0
## a4  0.13 0.10 0.14 0.18 0.12 0.12 0.20    0
```

```
paste("Alpha: 0.82, which is good")
```

```
## [1] "Alpha: 0.82, which is good"
```

```
# C2: Economical and Integrity based behaviour
alpha(subset(foreignsurvey, select = c(a5, a8, a2, a1, a9)), check.keys =TRUE)
```

```
##
## Reliability analysis
## Call: alpha(x = subset(foreignsurvey, select = c(a5, a8, a2, a1, a9)),
##     check.keys = TRUE)
##
##   raw_alpha std.alpha G6(smc) average_r S/N   ase mean  sd median_r
##       0.74      0.76    0.75      0.38 3.1 0.045  5.7 1.1     0.38
##
##     95% confidence boundaries
```

```
##          lower alpha upper
## Feldt     0.64  0.74  0.81
## Duhachek  0.65  0.74  0.82
##
##  Reliability if an item is dropped:
##    raw_alpha std.alpha G6(smc) average_r S/N alpha se var.r med.r
## a5      0.62      0.64    0.59      0.31 1.8    0.066 0.015  0.30
## a8      0.67      0.70    0.68      0.37 2.3    0.059 0.031  0.36
## a2      0.72      0.74    0.72      0.42 2.9    0.051 0.025  0.44
## a1      0.67      0.69    0.65      0.36 2.2    0.058 0.015  0.38
## a9      0.77      0.78    0.74      0.47 3.5    0.040 0.012  0.47
##
##  Item statistics
##     n raw.r std.r r.cor r.drop mean  sd
## a5 90  0.82  0.84  0.83   0.71  6.2 1.3
## a8 90  0.73  0.74  0.64   0.55  5.4 1.5
## a2 90  0.66  0.65  0.51   0.43  5.7 1.6
## a1 90  0.73  0.76  0.71   0.57  5.5 1.3
## a9 90  0.62  0.57  0.39   0.33  5.5 1.8
##
## Non missing response frequency for each item
##       1    2    3    4    5    6    7 miss
## a5 0.02 0.01 0.01 0.03 0.13 0.24 0.54    0
## a8 0.03 0.01 0.04 0.14 0.23 0.22 0.31    0
## a2 0.06 0.01 0.00 0.09 0.19 0.26 0.40    0
## a1 0.02 0.01 0.03 0.12 0.28 0.29 0.24    0
## a9 0.04 0.08 0.03 0.09 0.11 0.19 0.46    0
```

```
paste("Alpha: 0.74, which is acceptable, 0.77 without a9, so we can optionally drop a9")
```

```
## [1] "Alpha: 0.74, which is acceptable, 0.77 without a9, so we can optionally drop a9"
```

```
alpha(subset(foreignsurvey, select = c(a5, a8, a2, a1)), check.keys =TRUE)
```

```
##
## Reliability analysis
## Call: alpha(x = subset(foreignsurvey, select = c(a5, a8, a2, a1)),
##     check.keys = TRUE)
##
##   raw_alpha std.alpha G6(smc) average_r S/N  ase mean  sd median_r
##       0.77      0.78    0.74      0.47 3.5 0.04  5.7 1.1     0.47
##
##     95% confidence boundaries
##          lower alpha upper
## Feldt     0.68  0.77  0.84
## Duhachek  0.69  0.77  0.85
##
##  Reliability if an item is dropped:
##    raw_alpha std.alpha G6(smc) average_r S/N alpha se  var.r med.r
## a5      0.66      0.67    0.58      0.40 2.0    0.062 0.0080  0.38
## a8      0.72      0.74    0.68      0.48 2.8    0.052 0.0257  0.47
## a2      0.78      0.78    0.72      0.54 3.6    0.041 0.0081  0.50
## a1      0.70      0.70    0.62      0.44 2.4    0.056 0.0034  0.47
```

```
##
##  Item statistics
##     n raw.r std.r r.cor r.drop mean  sd
## a5 90  0.82  0.84  0.79   0.68  6.2 1.3
## a8 90  0.77  0.76  0.63   0.56  5.4 1.5
## a2 90  0.72  0.70  0.53   0.47  5.7 1.6
## a1 90  0.78  0.80  0.73   0.61  5.5 1.3
##
## Non missing response frequency for each item
##       1    2    3    4    5    6    7 miss
## a5 0.02 0.01 0.01 0.03 0.13 0.24 0.54    0
## a8 0.03 0.01 0.04 0.14 0.23 0.22 0.31    0
## a2 0.06 0.01 0.00 0.09 0.19 0.26 0.40    0
## a1 0.02 0.01 0.03 0.12 0.28 0.29 0.24    0
```

```
# C3: Patriotic behavior
alpha(subset(foreignsurvey, select = c(a6, a3, a10, a7)), check.keys =TRUE)
```

```
##
## Reliability analysis
## Call: alpha(x = subset(foreignsurvey, select = c(a6, a3, a10, a7)),
##     check.keys = TRUE)
##
##   raw_alpha std.alpha G6(smc) average_r S/N   ase mean  sd median_r
##       0.74      0.75    0.71      0.42   3 0.044  4.7 1.4     0.38
##
##     95% confidence boundaries
##          lower alpha upper
## Feldt     0.64  0.74  0.82
## Duhachek  0.66  0.74  0.83
##
##  Reliability if an item is dropped:
##     raw_alpha std.alpha G6(smc) average_r S/N alpha se   var.r med.r
## a6       0.73      0.74    0.68      0.48 2.8    0.050 1.9e-02  0.46
## a3       0.62      0.62    0.52      0.35 1.6    0.069 5.9e-05  0.35
## a10      0.66      0.67    0.58      0.40 2.0    0.061 3.1e-03  0.40
## a7       0.72      0.72    0.66      0.46 2.6    0.052 2.3e-02  0.40
##
##  Item statistics
##      n raw.r std.r r.cor r.drop mean  sd
## a6  90  0.69  0.70  0.51   0.45  5.4 1.7
## a3  90  0.82  0.83  0.78   0.66  4.9 1.7
## a10 90  0.78  0.78  0.69   0.57  4.8 1.8
## a7  90  0.73  0.72  0.55   0.48  3.6 1.9
##
## Non missing response frequency for each item
##        1    2    3    4    5    6    7 miss
## a6  0.06 0.04 0.03 0.12 0.19 0.21 0.34    0
## a3  0.04 0.03 0.10 0.22 0.24 0.09 0.27    0
## a10 0.08 0.08 0.03 0.20 0.19 0.19 0.23    0
## a7  0.18 0.16 0.22 0.11 0.14 0.08 0.11    0
```

```
paste("Alpha: 0.74, which is acceptable")
```

```
## [1] "Alpha: 0.74, which is acceptable"
```

```
foreignsurvey$discriminatory <- (foreignsurvey$a14 + foreignsurvey$a12 + foreignsurvey$a15 +
                                   foreignsurvey$a13 + foreignsurvey$a4)/5
foreignsurvey$ecoint <- (foreignsurvey$a5 + foreignsurvey$a8 + foreignsurvey$a2 +
                           foreignsurvey$a1 + foreignsurvey$a9)/5
foreignsurvey$patriotic <- (foreignsurvey$a6 + foreignsurvey$a3 + foreignsurvey$a10 +
                              foreignsurvey$a7)/4

write.table(foreignsurvey, "PCA-foreigner-survey-extended.txt", sep="\t", row.names=FALSE, na="")

summary(foreignsurvey$discriminatory)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1.00    2.05    2.80    3.08    3.80    7.00
```

```
summary(foreignsurvey$ecoint)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1.80    5.20    5.80    5.66    6.40    7.00
```

```
summary(foreignsurvey$patriotic)
```

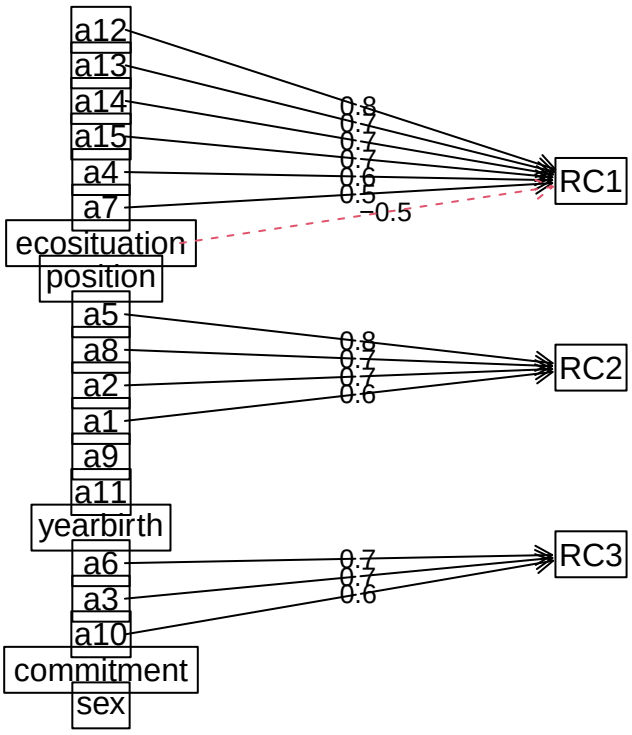
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1.00    3.81    4.75    4.67    5.75    7.00
```

```
# association between the extracted components and other variables

foreignsurvey2 = read.delim("C:/Users/daria/OneDrive/Desktop/Master - AppDS/Statistics/Datasets-2022100
                                          stringsAsFactors=F)

# hotfix-encoding for categorical data

foreignsurvey2$ecosituation = as.numeric(factor(foreignsurvey2$ecosituation))
foreignsurvey2$commitment = as.numeric(factor(foreignsurvey2$commitment))
foreignsurvey2$position = as.numeric(factor(foreignsurvey2$position))
foreignsurvey2$yearbirth = as.numeric(factor(foreignsurvey2$yearbirth))
foreignsurvey2$sex = as.numeric(factor(foreignsurvey2$sex))
foreignsurvey2$nr = NULL


# Rotate with varimax pca
pca.foreign.r <- principal(foreignsurvey2, 3)
pca.foreign.r$criteria <- NULL
print(pca.foreign.r, cut = 0.5, sort = TRUE, digits = 2)
```

```
## Principal Components Analysis
## Call: principal(r = foreignsurvey2, nfactors = 3)
```

```
## Standardized loadings (pattern matrix) based upon correlation matrix
##              item   RC1   RC2   RC3    h2   u2 com
## a12            17  0.79             0.682 0.32 1.2
## a13            18  0.75             0.707 0.29 1.5
## a14            19  0.75             0.611 0.39 1.2
## a15            20  0.71             0.525 0.47 1.1
## a4              9  0.59             0.465 0.53 1.6
## a7             12  0.55        0.52 0.571 0.43 2.0
## ecosituation    1 -0.52             0.276 0.72 1.0
## position        3                   0.214 0.79 1.7
## a5             10        0.79       0.724 0.28 1.3
## a8             13        0.71       0.620 0.38 1.5
## a2              7        0.68       0.506 0.49 1.2
## a1              6        0.63       0.635 0.36 2.0
## a9             14                   0.253 0.75 1.1
## a11            16                   0.272 0.73 2.2
## yearbirth       4                   0.100 0.90 2.1
## a6             11              0.70 0.507 0.49 1.1
## a3              8              0.66 0.633 0.37 1.8
## a10            15              0.62 0.514 0.49 1.6
## commitment      2                   0.172 0.83 1.7
## sex             5                   0.079 0.92 1.0
##
##                        RC1  RC2  RC3
## SS loadings           4.03 2.68 2.35
## Proportion Var        0.20 0.13 0.12
## Cumulative Var        0.20 0.34 0.45
## Proportion Explained  0.44 0.30 0.26
## Cumulative Proportion 0.44 0.74 1.00
##
## Mean item complexity =  1.5
## Fit based upon off diagonal values = 0.91
```

```r
# fa diagram
fa.diagram(pca.foreign.r, cut = 0.5, cex = 0.8, rsize = 0.5, main = "FA Diagram")
```



Thus, we can conclude that:

1. There are 15 factors where we can see that only 14 are good for extraction of component analysis (a11 has a low loading based on correlation matrix)
2. There are 3 components which are having the following items:

- PC1 [Discriminatory behavior]: a14, a12, a15, a13, a4
- PC2 [Economical and Integrity based behavior]: a5, a8, a2, a1, a9
- PC3 [Patriotic behavior]: a6, a3, a10, a7

3. The reliability of each new components is the following:

- PC1: 0.82, good internal consistency
- PC2: 0.74/0.77, acceptable internal consistency
- PC3: 0.74, acceptable internal consistency

4. There is no association between the extracted components and categorical variables, but:

- I created a hotfix encoding in order to show the low loading of categorical variables in correlation with other discrete variables
- The results showed that there is no association, but these results are not the most reliable one
- A reliable result would be to calculate an MCA and/or Factorial Analysis of Mixed Data (FAMD) in order to get a better overview regarding the association of the new variables in comparison with the others
- FAMD/MCA is not a subject of this homework [it was not taught at class], thus, I considered that the following analysis can be stopped here, with further research options