Multivariate Statistics with R

Exploratory Factor Analysis

Aja Murray, Aja.Murray@ed.ac.uk

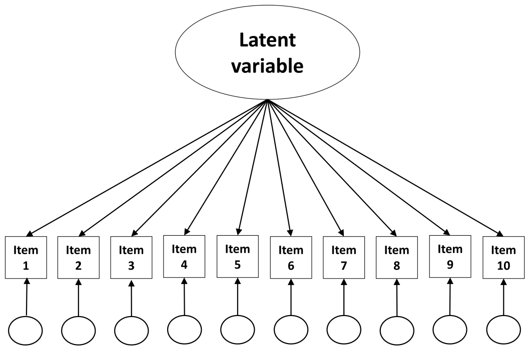
Exploratory factor analysis

- EFA used for identifying the number and nature of dimensions that describe a psychological construct and their inter-relations
- Procedurally similar to PCA but differs in important ways
  - *Uses only the common variance in its calculation*
  - *Can give quite different results to PCA under some circumstances*
  - *The resulting dimensions are called **factors***
  - *EFA based on a **latent variable model***

Latent variable models

- Divides the world into **observed variables** and **latent variables** (factors)
  - *Observed variables can be measured directly*
    - e.g., scores on IQ subtests
  - *Latent variables inferred based on patterns of observed variable associations*
    - e.g., Spearman's *g*
- Latent variables generate the correlations between observed variables
  - *e.g., higher g causes higher subtest scores*
- Observed variables are imperfect **indicators** (measures) of latent variables
  - *Observed variable scores have both a systematic and a random error component*

Latent variable models as an SEM diagram



- Latent variables are ellipses
- Observed variables are rectangles
- Single-headed arrows go from the latent variables to the observed variables
- There are also unique variances for the observed variables

Doing EFA

- Like PCA, there are a number of decisions:
  - *How many factors?*
  - *Which rotation?*
  - ***Which extraction method?***
- In EFA we also have to choose an extraction/estimation method

How many factors?

- As in PCA, we can use the following tools to help us decide how many factors to retain:
  - *Scree test*
  - *Parallel analysis*
  - *MAP test*
- It is also important to examine the factor solutions for varying numbers of factors
  - *Which solutions make more sense based on our background knowledge of the construct?*
  - *Do some solutions have deficiencies such as minor factors?*

Our running example

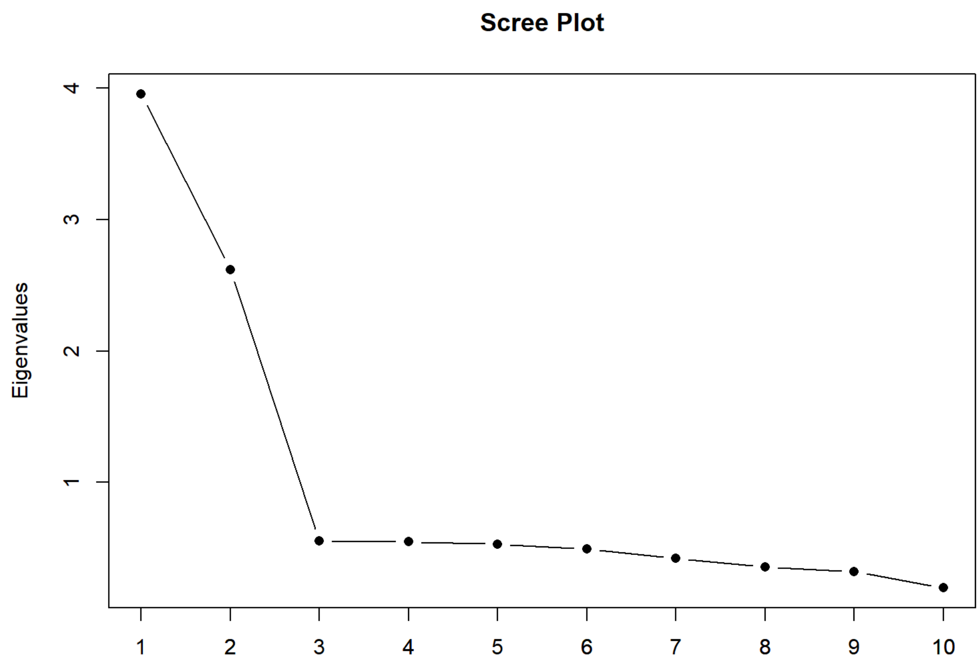
- Let’s return to our aggression example and now run an EFA
- We had n=1000 participants with data on the following 10 items:

  1. *I hit someone*
  2. *I kicked someone*
  3. *I shoved someone*
  4. *I battered someone*
  5. *I physically hurt someone on purpose*
  6. *I deliberately insulted someone*
  7. *I swore at someone*
  8. *I threatened to hurt someone*
  9. *I called someone a nasty name to their face*
  10. *I shouted mean things at someone*

How many aggression factors? Scree test

- We can plot the eigenvalues and look for a kink in the plot:

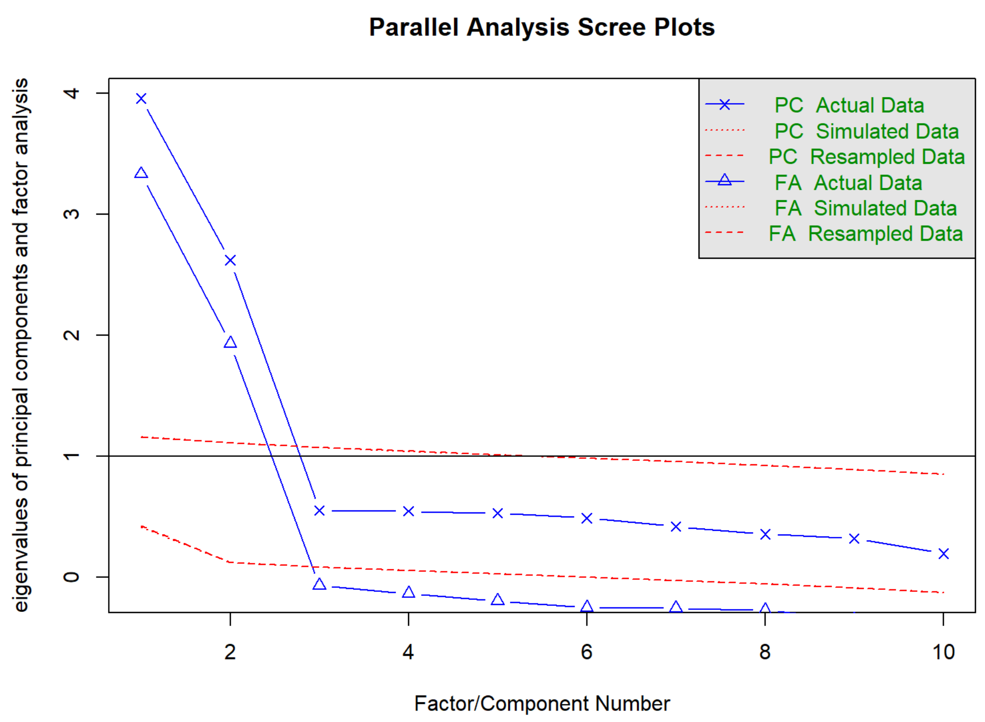
```
eigenvalues<-eigen(cor(agg.items))$values
plot(eigenvalues, type = 'b', pch = 16,
     main = "Scree Plot", xlab="", ylab="Eigenvalues")
axis(1, at = 1:10, labels = 1:10)
```



How many aggresion factors? Parallel analysis

- We can conduct a parallel analysis using fa.parallel( )from the psych package:

```
library(psych)
fa.parallel(agg.items, n.iter=500)
```

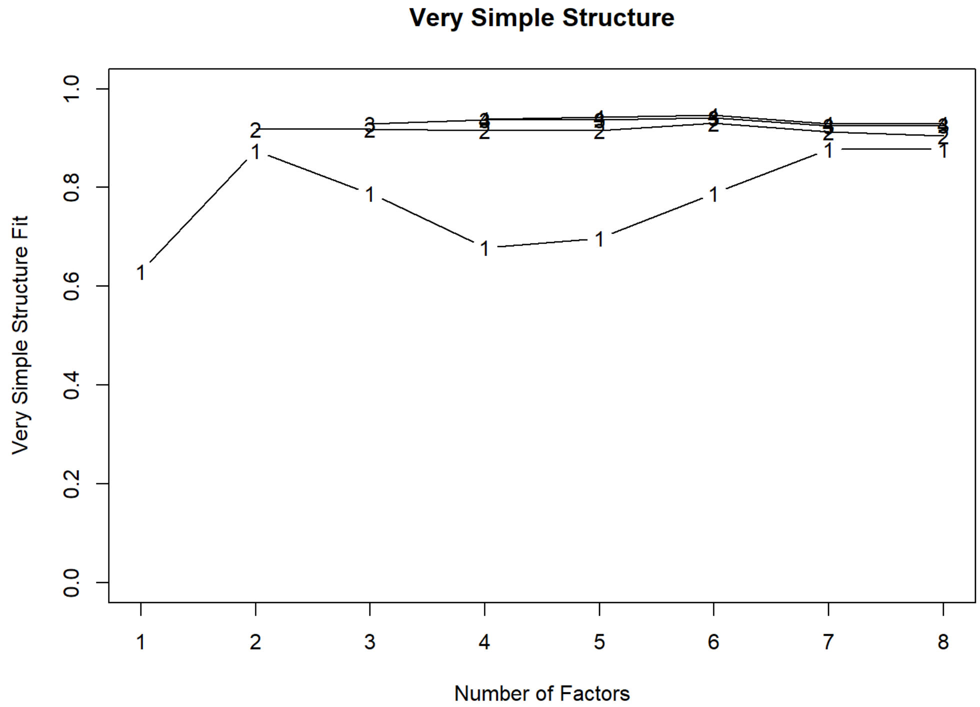



```
## Parallel analysis suggests that the number of factors =  2  and the number of components =  2
```

How many aggression factors? MAP

- We can conduct a MAP test using vss( ):

```
library(psych)
vss(agg.items)
```




```
##
## Very Simple Structure
## Call: vss(x = agg.items)
## Although the VSS complexity 1 shows  8  factors, it is probably more reasonable to think about  2  factors
## VSS complexity 2 achieves a maximimum of 0.93  with  6  factors
##
## The Velicer MAP achieves a minimum of 0.03  with  2  factors
## BIC achieves a minimum of  NA  with  2  factors
## Sample Size adjusted BIC achieves a minimum of  NA  with  2  factors
##
## Statistics by number of factors
##   vss1 vss2   map dof   chisq prob sqresid  fit RMSEA  BIC SABIC complex
## 1 0.63 0.00 0.150  35 2.4e+03 0.00     8.9 0.63  0.26 2142  2253     1.0
## 2 0.88 0.92 0.029  26 1.2e+01 0.99     2.0 0.92  0.00 -168   -85     1.0
## 3 0.79 0.92 0.054  18 6.8e+00 0.99     1.7 0.93  0.00 -118   -60     1.1
## 4 0.68 0.92 0.106  11 3.4e+00 0.98     1.5 0.94  0.00  -73   -38     1.2
## 5 0.70 0.92 0.162   5 2.2e+00 0.83     1.4 0.94  0.00  -32   -16     1.2
## 6 0.79 0.93 0.247   0 4.3e-03   NA     1.3 0.95    NA   NA    NA     1.3
## 7 0.88 0.91 0.328  -4 2.2e-07   NA     1.7 0.93    NA   NA    NA     1.1
## 8 0.88 0.91 0.611  -7 1.7e-06   NA     1.7 0.93    NA   NA    NA     1.1
##    eChisq    SRMR  eCRMS  eBIC
## 1 4.3e+03 2.2e-01 0.2475 4047
## 2 4.2e+00 6.9e-03 0.0090 -175
## 3 2.1e+00 4.8e-03 0.0076 -122
## 4 9.1e-01 3.2e-03 0.0064  -75
## 5 5.5e-01 2.5e-03 0.0074  -34
## 6 8.1e-04 9.5e-05     NA   NA
## 7 4.9e-08 7.4e-07     NA   NA
## 8 5.7e-07 2.5e-06     NA   NA
```

Examining the factor solutions

- Finally, we draw on information from the factor solutions themselves
- We run a series of factor analysis models with different numbers of factors
- Look at the loadings and factor correlations:
  - *Are important distinctions blurred when the number of factors is smaller?*
  - *Are there minor or ‘methodological’ factors when the number of factors is larger?*
  - *Are the factor correlations very high?*
  - *Do the factor solutions make theoretical sense?*
- In this case, given the MAP, scree and parallel analysis results we would likely want to examine the 1,2 and 3 factor solutions

Conducting EFA in R

- We can run our factor analyses using the fa() function
- The first argument is the dataset with the items we want to factor analyse
- We also need to mention the number of factors we want to extract, e.g., nfactors=1

```
onef<-fa(agg.items, nfactors=1) #EFA with 1 factor
```

The one-factor solution

- To help us choose an optimal number of factors, we can look at the one-factor solution…

```
onef<-fa(agg.items, nfactors=1) #EFA with 1 factor
onef$loadings #inspect the factor loadings
```

```
##
## Loadings:
##        MR1
## item1  0.473
## item2  0.500
## item3  0.434
## item4  0.440
## item5  0.499
## item6  0.553
## item7  0.749
## item8  0.737
## item9  0.658
## item10 0.621
##
##                  MR1
## SS loadings    3.333
## Proportion Var 0.333
```

The two-factor solution

- And compare with the two-factor solution…

```
library(psych)
twof<-fa(agg.items, nfactors=2, rotate='oblimin') #EFA with 2 factors
```

```
## Loading required namespace: GPArotation
```

```
twof$loadings ##inspect the factor loadings
```

```
##
## Loadings:
##        MR1    MR2
## item1         0.698
## item2         0.798
## item3         0.677
## item4         0.656
## item5         0.836
## item6   0.686
## item7   0.879
## item8   0.914
## item9   0.653  0.115
## item10  0.730
##
##                  MR1   MR2
## SS loadings    3.040 2.730
## Proportion Var 0.304 0.273
## Cumulative Var 0.304 0.577
```

```
twof$Phi  ## inspect the factor correlations
```

```
##           MR1       MR2
## MR1 1.0000000 0.2164953
## MR2 0.2164953 1.0000000
```

The three-factor solution

- And the three-factor solution

```
library(psych)
threef<-fa(agg.items, nfactors=3, rotate='oblimin') #EFA with 3 factors
threef$loadings #inspect the factor loadings
```

```
##
## Loadings:
##        MR1    MR2    MR3
## item1                0.996
## item2         0.806
## item3         0.710
## item4         0.657
## item5         0.789
## item6  0.686
## item7  0.879
## item8  0.913
## item9  0.654  0.120
## item10 0.730
##
##                  MR1   MR2   MR3
## SS loadings    3.039 2.224 0.997
## Proportion Var 0.304 0.222 0.100
## Cumulative Var 0.304 0.526 0.626
```

```
threef$Phi # inpsect the factor correlations
```

```
##           MR1       MR2       MR3
## MR1 1.0000000 0.2095680 0.1782584
## MR2 0.2095680 1.0000000 0.7019552
## MR3 0.1782584 0.7019552 1.0000000
```

Factor extraction in EFA

- **Factor extraction** refers to the method of deriving the factors
- PCA is itself an extraction method
- In EFA there are a number of factor extraction options:
  - *principal axis factoring*
  - *ordinary least squares (OLS)*
  - *weighted least squares (WLS)*
  - *minres*
  - *maximum likelihood (ML)*

Principal axis factoring (PAF)

- Traditional method
- An eigendecomposition of a reduced form of correlation matrix
  - *Diagonals are replaced by communalities*
  - *Communalities estimates used as starting point*
    - Based on e.g. multiple squared R
  - *Iteratively updated across successive PAFs*
  - *Process terminates when estimates change little across iterations*
- Focus on common rather than all variance is key EFA vs PCA distinction

Other extraction methods

- **OLS** finds the factor solution that minimises difference between observed and model-implied covariance matrices
  - *specifically, minimises the sum of squared residuals*
- **WLS** up-weights the variables with higher communalities
- **minres** ignores the diagonals
- **ML** finds the factor solution that maximises the likelihood of the observed covariance matrix

Which to use?

- PAF is a good option
- minres can provide EFA solutions when other methods fail
  - *minres is the default for the fa( ) function*
- choice of extraction method usually makes little difference if:
  - *communalities are similar*
  - *sample size is large*
  - *the number of variables is large*

PAF

- We can do a factor analysis with PAF by setting fm='pa' in the fa() function:

```
library(psych)
twof<-fa(agg.items, nfactors=2, rotate='oblimin', fm='pa') #EFA with 2 factors
twof$loadings ##inspect the factor loadings
```

```
##
## Loadings:
##         PA1    PA2
## item1          0.698
## item2          0.798
## item3          0.677
## item4          0.656
## item5          0.836
## item6   0.687
## item7   0.879
## item8   0.913
## item9   0.653  0.115
## item10  0.730
##
##                  PA1   PA2
## SS loadings    3.040 2.730
## Proportion Var 0.304 0.273
## Cumulative Var 0.304 0.577
```

```
twof$Phi  ## inspect the factor correlations
```

```
##           PA1       PA2
## PA1 1.0000000 0.2165792
## PA2 0.2165792 1.0000000
```

minres

- minres is the default method but we can also explicitly set fm='minres':

```
library(psych)
twof<-fa(agg.items, nfactors=2, rotate='oblimin', fm='minres') #EFA with 2 factors
twof$loadings ##inspect the factor loadings
```

```
##
## Loadings:
##        MR1    MR2
## item1         0.698
## item2         0.798
## item3         0.677
## item4         0.656
## item5         0.836
## item6   0.686
## item7   0.879
## item8   0.914
## item9   0.653  0.115
## item10  0.730
##
##                  MR1   MR2
## SS loadings    3.040 2.730
## Proportion Var 0.304 0.273
## Cumulative Var 0.304 0.577
```

```
twof$Phi  ## inspect the factor correlations
```

```
##           MR1       MR2
## MR1 1.0000000 0.2164953
## MR2 0.2164953 1.0000000
```

Factor rotation

- Like in PCA:
  - *Rotation needed to make solution interpretable*
  - *Main choice is between oblique vs orthogonal*
  - *Oblique often preferable as allows correlated or uncorrelated*
  - *Orthogonal rotation yields one loading matrix*
  - *Oblique yields both pattern and structure loading matrices*
    - Pattern matrix is usually used as basis for interpretation

Interpreting the factor solution

- Label factors on basis of high loading items

```
library(psych)
twof<-fa(agg.items, nfactors=2, rotate='oblimin', fm='minres') #EFA with 2 factors
twof$loadings ##inspect the factor loadings
```

```
##
## Loadings:
##        MR1    MR2
## item1         0.698
## item2         0.798
## item3         0.677
## item4         0.656
## item5         0.836
## item6   0.686
## item7   0.879
## item8   0.914
## item9   0.653  0.115
## item10  0.730
##
##                  MR1   MR2
## SS loadings    3.040 2.730
## Proportion Var 0.304 0.273
## Cumulative Var 0.304 0.577
```

Interpreting the factor solution

- Factor 1 could be labelled *verbal aggression* and factor 2 could be labelled *physical aggression*

1. *I hit someone*
2. *I kicked someone*
3. *I shoved someone*
4. *I battered someone*
5. *I physically hurt someone on purpose*
6. *I deliberately insulted someone*
7. *I swore at someone*
8. *I threatened to hurt someone*
9. *I called someone a nasty name to their face*
10. *I shouted mean things at someone*

The magnitude of factor loadings

- How large are the loadings?
- Larger loadings suggest that the variables are ‘better’ markers of the underlying factors
- Comfrey & Lee (1992) offered the following rules of thumb:
  - *.71 (50% overlapping variance) are considered excellent*
  - *.63 (40% overlapping variance) is very good*
  - *.55 (30% overlapping variance) is good*
  - *.45 (20% overlapping variance) is fair*
  - *.32 (10% overlapping variance) is poor*

The magnitude of factor correlations

- How distinct are the factors?

```
library(psych)
twof<-fa(agg.items, nfactors=2, rotate='oblimin', fm='minres') #EFA with 2 factors
twof$Phi  ## inspect the factor correlations
```

```
##           MR1       MR2
## MR1 1.0000000 0.2164953
## MR2 0.2164953 1.0000000
```

How much variance is accounted for by the factors?

- We can also check how much variance overall is accounted for by the factors

```
twof
```

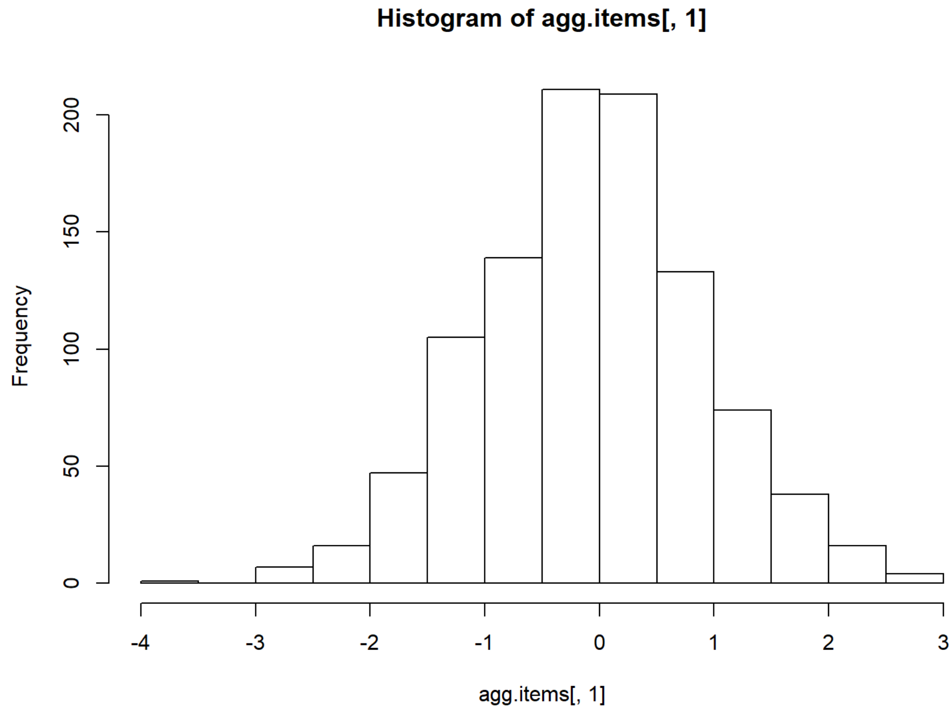
```
## Factor Analysis using method =  minres
## Call: fa(r = agg.items, nfactors = 2, rotate = "oblimin", fm = "minres")
## Standardized loadings (pattern matrix) based upon correlation matrix
##          MR1   MR2   h2   u2 com
## item1   0.03  0.70 0.50 0.50 1.0
## item2   0.00  0.80 0.64 0.36 1.0
## item3   0.00  0.68 0.46 0.54 1.0
## item4   0.02  0.66 0.44 0.56 1.0
## item5  -0.02  0.84 0.69 0.31 1.0
## item6   0.69 -0.05 0.46 0.54 1.0
## item7   0.88  0.02 0.78 0.22 1.0
## item8   0.91 -0.03 0.83 0.17 1.0
## item9   0.65  0.12 0.47 0.53 1.1
## item10  0.73 -0.01 0.53 0.47 1.0
##
##                        MR1  MR2
## SS loadings           3.05 2.74
## Proportion Var        0.30 0.27
## Cumulative Var        0.30 0.58
## Proportion Explained  0.53 0.47
## Cumulative Proportion 0.53 1.00
##
##  With factor correlations of
##      MR1  MR2
## MR1 1.00 0.22
## MR2 0.22 1.00
##
## Mean item complexity =  1
## Test of the hypothesis that 2 factors are sufficient.
##
## The degrees of freedom for the null model are  45  and the objective function was  4.85 with Chi Square of  4820.
## The degrees of freedom for the model are 26  and the objective function was  0.01
##
## The root mean square of the residuals (RMSR) is  0.01
## The df corrected root mean square of the residuals is  0.01
##
## The harmonic number of observations is  1000 with the empirical chi square  4.25  with prob <  1
## The total number of observations was  1000  with Likelihood Chi Square =  12.07  with prob <  0.99
##
## Tucker Lewis Index of factoring reliability =  1.005
## RMSEA index =  0  and the 90 % confidence intervals are  0 0
## BIC =  -167.53
## Fit based upon off diagonal values = 1
## Measures of factor score adequacy
##                                                   MR1  MR2
## Correlation of (regression) scores with factors   0.96 0.93
## Multiple R square of scores with factors          0.92 0.87
## Minimum correlation of possible factor scores     0.84 0.74
```

Checking the suitability of data for EFA

- The first step in an EFA is actually to check the appropriateness of the data:
  - *Does the data look multivariate normal?*
  - *Do the relations look linear?*
  - *Does the correlation matrix have good factorability?*

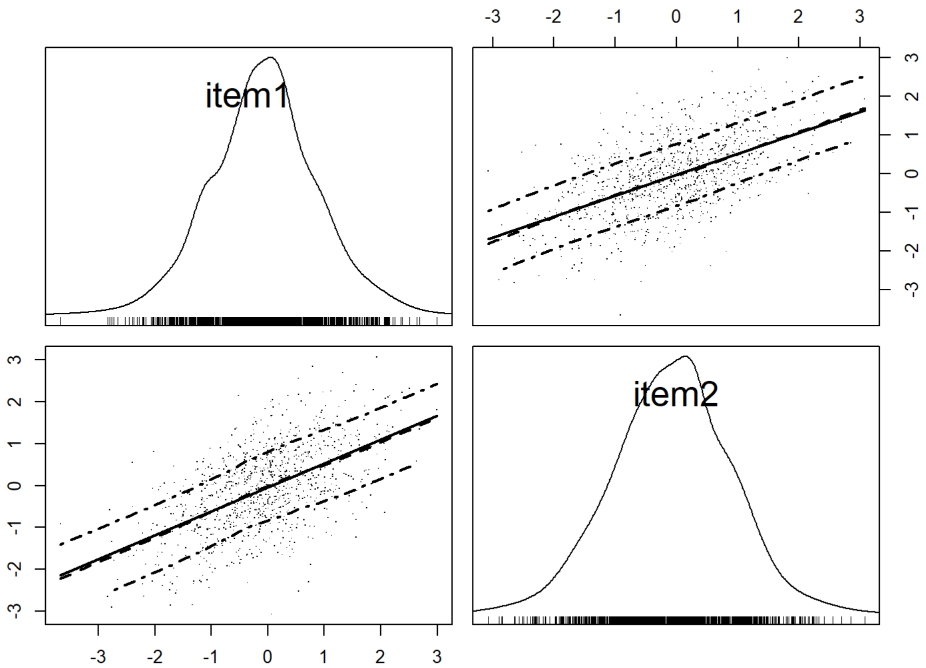
Multivariate normality

- Do the variables have (approximately) continuous measurement scales?
  - *5 or more response options*
- Examining univariate distributions using histograms



Linearity

- Plot linear and lowess lines for pairwise relations and compare



Factorability

- EFA focuses on variance **common** to items
  - *Not much point in an EFA if little variance in common*
- Use Kaiser-Meyer-Olkin (KMO) test
  - *Provides measure of proportion of variance shared between variables*
  - *Can be computed for individual variables or whole correlation matrix*
  - *Overall values >.60 and no variable <.50 is ideal*

KMO in R

```r
KMO(agg.items)
```

```
## Kaiser-Meyer-Olkin factor adequacy
## Call: KMO(r = agg.items)
## Overall MSA =  0.87
## MSA for each item = 
##  item1  item2  item3  item4  item5  item6  item7  item8  item9 item10 
##   0.89   0.86   0.90   0.91   0.84   0.92   0.84   0.82   0.94   0.92 
```

Summary

- Steps in EFA are similar to PCA but…
  - *The underlying theory and interpretation is quite different*
  - *Their results can differ if there is not a lot of common variance*
- EFA involves:
  - *Checking data suitability*
  - *Choosing number of factors*
  - *Factor extraction*
  - *Rotation*
  - *Interpretation of factors*