Aarhus University Hospital

AARHUS UNIVERSITY

Factor analysis

Eva Ørnbøl
Statistician

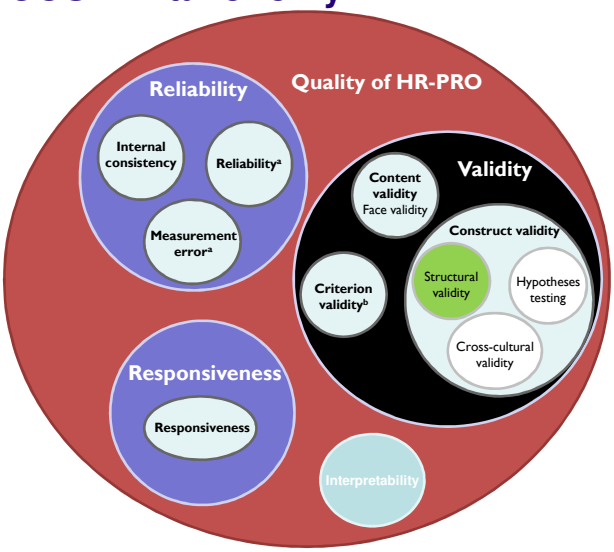
Aarhus University Hospital

AARHUS UNIVERSITY

Outline

- The setting
- EFA versus CFA
- The FA model
- The idea behind EFA
- Key concepts and steps in EFA
  - estimation methods
  - selection of number of factors
  - rotation
- The idea of CFA
- Note concerning assumptions

The COSMIN taxonomy



Introduction

- Structural validity

  "*The degree to which the scores of a measurement instrument are an adequate reflection of the dimensionality of the construct to be measured*"

Factor analysis

- Exploratory Factor Analysis (EFA)
  - Used when:
    - You have no idea of the number of factors involved
    - You have no idea of which items belong to which factor
- Confirmatory Factor Analysis (CFA)
  - Used when:
    - You know the number factors, based on theory or prior research
    - You know which items belong to which factor, again based on theory or prior research

Factor Analysis model

$$x_i = \mu_i + \lambda_{i1}F_1 + \lambda_{i2}F_2 + \cdots \lambda_{iM}F_M + e_i$$

$x_i = \text{item } i, i = 1, \cdots, p$

$\mu_i = \text{mean item } i, i = 1, \cdots, p$

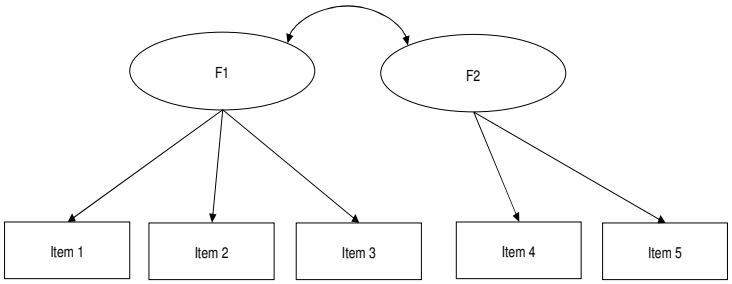
$F_m = \text{factor } m, m = 1, \cdots, M, M \leq p$

$\lambda_{im} = \text{factor loading for item } i \text{ and factor } m$

$e_i = \text{error term for item } i, i = 1, \cdots, p$

$e_i \sim N(0, \Psi_i), \Psi_i \text{ unique variance item } i$

Path diagram – a simple example



EFA - purpose

- A statistical method that investigates the covariance or correlation matrix and seeks to identify groups of items satisfying:
  - High correlation between items in same group
  - Low correlation between items from one group and items outside this group
  - The maximum number of groups equal the number of items

Correlations – example 1

item	1	2	3	4	5	6
1	1					
2	0.68	1				
3	0.76	0.69	1			
4	0.72	0.74	0.64	1		
5	0.81	0.85	0.67	0.73	1	
6	0.79	0.71	0.83	0.82	0.66	1

Correlations – example 2

item	1	2	3	4	5	6
1	1					
2	0.84	1				
3	0.23	0.39	1			
4	0.78	0.69	0.29	1		
5	0.65	0.72	0.32	0.81	1	
6	0.37	0.31	0.86	0.36	0.27	1

More on the purpose of EFA

How does the identification of groups of items from the correlation matrix relate to the FA model?

$$x_i = \mu_i + \lambda_{i1}F_1 + \lambda_{i2}F_2 + \cdots + \lambda_{iM}F_M + e_i$$

To see the connection we have to look closer at the factor loading.

Factor loadings – example 2

	F1	F2	F3	F4	F5
item 1	1	0	0	0	0
item 2	1	0	0	0	0
item 3	0	1	0	0	0
item 4	1	0	0	0	0
item 5	0	1	0	0	0

Factor loadings – example 1

$$x_i = 3.75 + 0.97 * F_1 + 0.05 * F_2 + e_i$$
$$x_j = 3.98 + 0.02 * F_1 + 0.91 * F_2 + e_j$$

$$x_i = 3.7 + 0.48 * F_1 + 0.53 * F_2 + e_i$$
$$x_j = 2.6 + 0.36 * F_1 + 0.49 * F_2 + e_j$$

EFA in Stata version 15 an example

```
factor f32*, pf factors(4) blanks(0.1)
screeplot
rotate, varimax
```

Method of estimation

- `factor f32*, pf factors(4) blanks(0.1)`
  - pf: principal factor (default)
  - pcf: principal component factor
  - ipf: iterated principal factor
  - ml: maximum likelihood

  All of these methods of estimation seek to maximize different measures of goodness-of-fit.

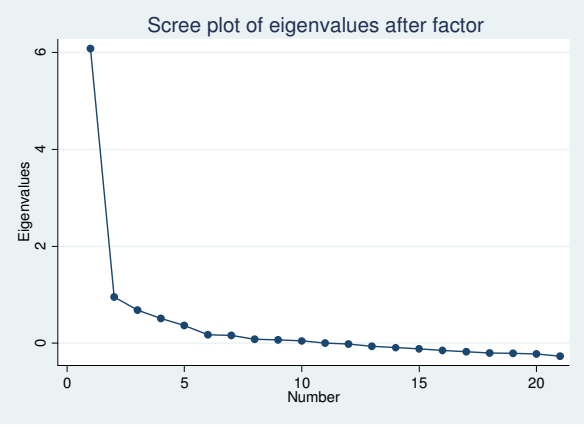
Select number of factors M

- `factor f32*, pf factors(4) blanks(0.1)`
  - State the number: factor(#)
  - Follow from the estimation method:
    - Eigenvalue > 1 (default at pcf)
    - Eigenvalue > 0 (default at pf)
  - Examine scree plot
  - Factor providing a non-significant log-likelihood ratio test (at ml)
  - Factor with minimum value of information criteria AIC or BIC (at ml)

More on eigenvalues

- Eigenvalues (and eigenvectors) are important parts in the estimation solution no matter which estimation method used.
- A rough interpretation is that an eigenvalue represents a measure of the amount of variance explained by the attending factor – thereby the importance of the factor.

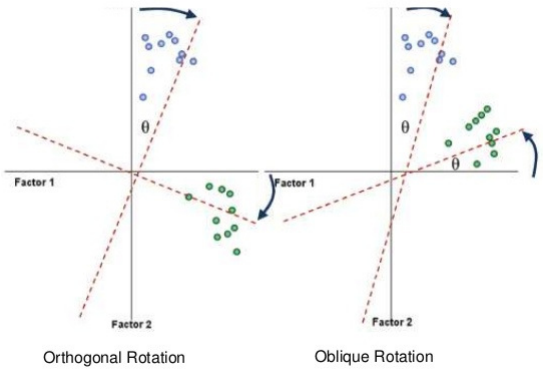
EFA: scree plot



Rotation of loadings

- There is a non-uniqueness to the FA model and that has caused some constraints to the model for it to be estimated. These have lead to the solution of factor loadings to possible be less interpretable.
- In order to improve interpretability rotation of the factor loadings is introduced.
- Rotation methods strives to find a ”simple” pattern of factor loadings, where the methods differ in their definition of ”simple”.
- Methods being orthogonal or oblique.
  - Orthogonal:Varimax
  - Oblique: Promax
  - Many other

Rotation – a simple example



Orthogonal Rotation Oblique Rotation

37

More theory of EFA

The connection is reflected in the covariance estimator derived from the FA model:

$$\Sigma = \Lambda\Lambda' + \Psi$$

$$\Sigma = \text{covariance matrix}$$

$$\Lambda = \text{matrix of factorloadings}$$

$$\Psi = \text{unique variance diagonal matrix}$$

$$\Lambda\Lambda' = (L\sqrt{V})(\sqrt{V}L)'$$

$$L = \text{matrix consisting of eigenvectors i coloms}$$

$$V = \text{diagonal matrix with eigenvalues in the diagonal}$$

$$\Lambda\Lambda' + \Psi = \Lambda^{*}\Lambda^{*'} + \Psi$$

$$\Lambda = \Lambda^{*}G,\ G = \text{orthogonal matrix}$$

Excercise EFA

- Handouts:
  - The wording of the exercise
  - Output from Stata
    - A:Eigenvalues
    - A:Factor loadings
    - B: Rotation of factor loadings
  - The original questionnaire with the items used in Danish
  - A translation of the items to English

Take home message from exercise

- Given the same data different researchers will likely expound different results of a EFA. This often originates in:
  - different view on research purpose
  - different background knowledge
- When you read an article about an EFA, it is inevitable that more models, than the one presented, have been considered.

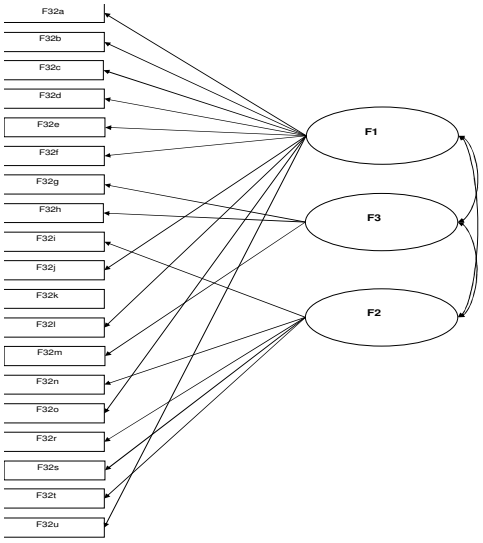
CFA purpose

- The starting point is a theory of how observed variables (items) and latent variables (factors) interact often illustrated by means of a path diagram
- The CFA is a test of how well the given theory/model is able to account for the variation in the given data
- There are several ways to accomplish this both by means of traditional test and through different fit indices

CFA: assessment of fit

- Chi-square test of model fit χ^2- non-significant p>0.05 or p>0.01
- Goodness-of-fit index above: >acceptable (>good)
  - Global fit index (GFI) >0.90 (>0.95)
  - Non-normed fit index (NNFI) >0.90 (>0.95)
  - Comparative fit index (CFI) >0.90 (>0.95)
  - Tucker-Lewis fit index (TLI) >0.90 (>0.95)
- Residual considerations:
  - Root mean square error of approximation (RMSEA) <0.08 (>0.05)
  - Standardised root mean square residual (SRMR) <0.05
  - Root mean square residual (RMR) <0.05
- When comparing several models minimum value of information criteria AIC, BIC, ss BIC among others.

Example from the exercise – path diagram



Example continued: estimation in Stata

```
sem (F1-> f32a f32b f32c f32d f32e f32f f32j f32l f32o f32u) ///
    (F2-> f32e f32g f32h f32m)  ///
    (F3-> f32i f32n f32r f32s f32t), standardized
estat gof, stats(all)
estat residuals
```


Example continued: estimation in Stata

```
. estat gof, stats(all)
```

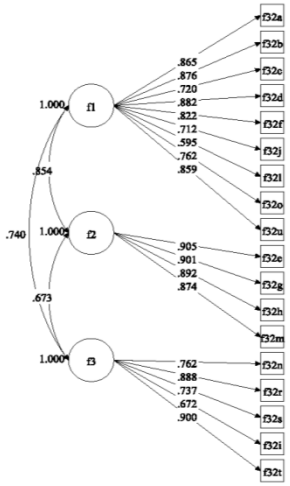
Fit statistic	Value	Description
Likelihood ratio		
chi2_ms(132)	493.154	model vs. saturated
p > chi2	0.000	
chi2_bs(153)	2828.180	baseline vs. saturated
p > chi2	0.000	
Population error		
RMSEA	0.075	Root mean squared error of approximation
90% CI, lower bound	0.068	
upper bound	0.083	
pclose	0.000	Probability RMSEA <= 0.05
Information criteria		
AIC	1486.365	Akaike's information criterion
BIC	1724.508	Bayesian information criterion
Baseline comparison		
CFI	0.865	Comparative fit index
TLI	0.844	Tucker-Lewis index
Size of residuals		
SRMR	0.060	Standardized root mean squared residual
CD	0.978	Coefficient of determination

Example continued: estimation in M*plus*

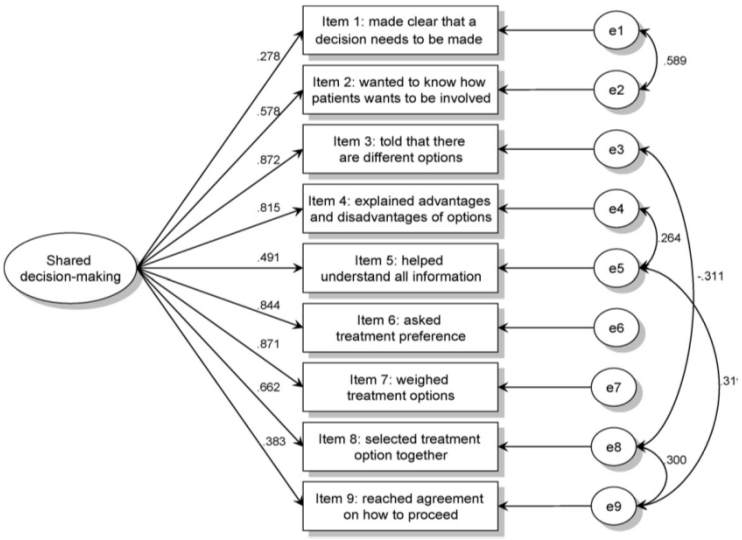
TESTS OF MODEL FIT

	Continuous		Categorical
Chi-Square Test of Model Fit			
Value	478.695		174.108*
Degrees of Freedom	132		132
P-Value	0.0000		0.0082
CFI/TLI			
CFI	0.863		0.988
TLI	0.841		0.987
Information Criteria			
Number of Free Parameters	57		
Akaike (AIC)	1641.746		
Bayesian (BIC)	1879.889		
Sample-Size Adjusted BIC (n* = (n + 2) / 24)	1698.976		
RMSEA (Root Mean Square Error Of Approximation)			
Estimate	0.074		0.026
90 Percent C.I.	0.067	0.081	0.014 0.036
Probability RMSEA <= .05	0.000		1.000
SRMR (Standardized Root Mean Square Residual)	0.056		0.078

CFA reporting results – example 1



CFA reporting results – example 2



A few essential assumptions FA

- Reflective items
  - If normative items are present this can be incorporated in a model form the broader frame of models of Structural Equation Models (SEM)
- Continuous response scale.
  - As this is highly unlikely to be true for many questionnaire data there are several ways to overcome this

Ordered response categories

- If the number of response categories is a minimum of 5 studies have found it to be an acceptable approximation
- A more appropriate estimate of the correlation matrix can be calculated and subsequently used in the FA analysis – a tetrachoric or polychoric correlation matrix
  - e.g. in Stata polychoric followed by factormat commands
- The FA model fit into the broader model frame of SEM and therein an estimation method that allows for ordered or dichotome response categories can be applied (WLSM , WLSMV)
  - SEM programs Lisrel, M*plus*, Amos etc.
  - Stata version 15+16 GSEM