SEM 1: Confirmatory Factor Analysis
Week 4 - Advanced CFA topics

Sacha Epskamp

23-04-2019
I see myself as someone who is talkative

<table>
<thead>
<tr>
<th>Disagree strongly</th>
<th>Disagree a little</th>
<th>Neither agree nor disagree</th>
<th>Agree a little</th>
<th>Agree Strongly</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
</tr>
</tbody>
</table>

- If data is ordinal and consists of only a few levels of measurement data cannot be assumed normal
  - Roughly less than five categories (Rhemtulla, Brosseau-Liard, & Savalei, 2012).
- In this case a typical solution is to use threshold models
  - Polychoric correlation if both variables are ordinal
  - Polyserial correlation if one item is ordinal and the other is continuous
- Estimation not via maximum likelihood but typically via (diagonally) weighted least squares:
  \[ F_{\text{WLS}} = (s - \sigma)^\top W^{-1}(s - \sigma) \]
set.seed(1)
# Setup:
sampleSize <- 1000
cor <- 0.5
thresh1 <- c(-2,0,2)
thresh2 <- c(-1,0.5,1.6)

# Generate data:
library("mvtnorm")
corMat <- matrix(c(1,0.5,0.5,1),2,2)
Data <- as.data.frame(rmvnorm(sampleSize, sigma = corMat))
names(Data) <- c("y1","y2")

# Make categorical:
Data[,1] <- as.numeric(cut(Data[,1], breaks = c(-Inf,thresh1,Inf)))
Data[,2] <- as.numeric(cut(Data[,2], breaks = c(-Inf,thresh2,Inf)))
Note: zeroes in marginal crosstables might be problematic..
# Pearson correlation:
cor(Data[,1], Data[,2])

## [1] 0.4076942

# Polychoric correlation:
library("lavaan")
lavCor(Data, ordered = c("y1","y2"))

## y1  y2
## y1 1.000
## y2 0.499 1.000
No thresholds:

Model <- 'f1 =~ y1
f2 =~ y2'

fit <- cfa(Model, Data, std.lv = TRUE)
parameterEstimates(fit)

<table>
<thead>
<tr>
<th>##</th>
<th>lhs op rhs</th>
<th>est</th>
<th>se</th>
<th>z</th>
<th>pvalue</th>
<th>ci.lower</th>
<th>ci.upper</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>f1 =~ y1</td>
<td>0.613</td>
<td>0.014</td>
<td>44.721</td>
<td>0</td>
<td>0.586</td>
<td>0.640</td>
</tr>
<tr>
<td>2</td>
<td>f2 =~ y2</td>
<td>0.778</td>
<td>0.017</td>
<td>44.721</td>
<td>0</td>
<td>0.744</td>
<td>0.812</td>
</tr>
<tr>
<td>3</td>
<td>y1 ~~ y1</td>
<td>0.000</td>
<td>0.000</td>
<td>NA</td>
<td>NA</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>4</td>
<td>y2 ~~ y2</td>
<td>0.000</td>
<td>0.000</td>
<td>NA</td>
<td>NA</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>5</td>
<td>f1 ~~ f1</td>
<td>1.000</td>
<td>0.000</td>
<td>NA</td>
<td>NA</td>
<td>1.000</td>
<td>1.000</td>
</tr>
<tr>
<td>6</td>
<td>f2 ~~ f2</td>
<td>1.000</td>
<td>0.000</td>
<td>NA</td>
<td>NA</td>
<td>1.000</td>
<td>1.000</td>
</tr>
<tr>
<td>7</td>
<td>f1 ~~ f2</td>
<td>0.408</td>
<td>0.026</td>
<td>15.463</td>
<td>0</td>
<td>0.356</td>
<td>0.459</td>
</tr>
</tbody>
</table>
With thresholds:

```r
Model <- 'y1 ~~ y2'
fit <- cfa(Model, Data, std.lv = TRUE,
          ordered = c("y1","y2"))
parameterEstimates(fit)
```

<table>
<thead>
<tr>
<th></th>
<th>lhs</th>
<th>op</th>
<th>rhs</th>
<th>est</th>
<th>se</th>
<th>z</th>
<th>pvalue</th>
<th>ci.lower</th>
<th>ci.upper</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>y1</td>
<td>~</td>
<td>y2</td>
<td>0.499</td>
<td>0.029</td>
<td>17.143</td>
<td>0.000</td>
<td>0.442</td>
<td>0.556</td>
</tr>
<tr>
<td>2</td>
<td>y1</td>
<td>-</td>
<td>t1</td>
<td>-1.812</td>
<td>0.075</td>
<td>-24.079</td>
<td>0.000</td>
<td>-1.959</td>
<td>-1.664</td>
</tr>
<tr>
<td>3</td>
<td>y1</td>
<td>-</td>
<td>t2</td>
<td>0.023</td>
<td>0.040</td>
<td>0.569</td>
<td>0.569</td>
<td>-0.055</td>
<td>0.100</td>
</tr>
<tr>
<td>4</td>
<td>y1</td>
<td>-</td>
<td>t3</td>
<td>1.911</td>
<td>0.081</td>
<td>23.524</td>
<td>0.000</td>
<td>1.752</td>
<td>2.070</td>
</tr>
<tr>
<td>5</td>
<td>y2</td>
<td>-</td>
<td>t1</td>
<td>-0.986</td>
<td>0.048</td>
<td>-20.753</td>
<td>0.000</td>
<td>-1.079</td>
<td>-0.893</td>
</tr>
<tr>
<td>6</td>
<td>y2</td>
<td>-</td>
<td>t2</td>
<td>0.476</td>
<td>0.041</td>
<td>11.519</td>
<td>0.000</td>
<td>0.395</td>
<td>0.557</td>
</tr>
<tr>
<td>7</td>
<td>y2</td>
<td>-</td>
<td>t3</td>
<td>1.580</td>
<td>0.064</td>
<td>24.653</td>
<td>0.000</td>
<td>1.455</td>
<td>1.706</td>
</tr>
<tr>
<td>8</td>
<td>y1</td>
<td>~</td>
<td>y1</td>
<td>1.000</td>
<td>0.000</td>
<td>NA</td>
<td>NA</td>
<td>1.000</td>
<td>1.000</td>
</tr>
<tr>
<td>9</td>
<td>y2</td>
<td>~</td>
<td>y2</td>
<td>1.000</td>
<td>0.000</td>
<td>NA</td>
<td>NA</td>
<td>1.000</td>
<td>1.000</td>
</tr>
<tr>
<td>10</td>
<td>y1</td>
<td>~*</td>
<td>y1</td>
<td>1.000</td>
<td>0.000</td>
<td>NA</td>
<td>NA</td>
<td>1.000</td>
<td>1.000</td>
</tr>
<tr>
<td>11</td>
<td>y2</td>
<td>~*</td>
<td>y2</td>
<td>1.000</td>
<td>0.000</td>
<td>NA</td>
<td>NA</td>
<td>1.000</td>
<td>1.000</td>
</tr>
<tr>
<td>12</td>
<td>y1</td>
<td>~1</td>
<td>y1</td>
<td>0.000</td>
<td>0.000</td>
<td>NA</td>
<td>NA</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>13</td>
<td>y2</td>
<td>~1</td>
<td>y2</td>
<td>0.000</td>
<td>0.000</td>
<td>NA</td>
<td>NA</td>
<td>0.000</td>
<td>0.000</td>
</tr>
</tbody>
</table>
Why are data missing? In a general $X$ predicts $Y$ case:

- **Missing completely at random** (MCAR)
  - Missingness is independent of $Y$ or $X$
  - Everything is fine!

- **Missing at random** (MAR)
  - Missingness is independent of $Y$, but not of $X$
  - Example: Men less willing to respond to mental health questionnaire
  - Not a big problem

- **Missing not at random** (MNAR)
  - Missingness depends on $Y$
  - Example: People with severe mental health problems fill in less questionnaires
  - This is bad :( 

Unfortunately, there is no way to know exactly how your data is missing.
Missing completely at random (MCAR)

<table>
<thead>
<tr>
<th>X</th>
<th>Y</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>6</td>
<td>5</td>
</tr>
<tr>
<td>5</td>
<td>6</td>
</tr>
<tr>
<td>8</td>
<td>5</td>
</tr>
<tr>
<td>6</td>
<td>7</td>
</tr>
<tr>
<td>7</td>
<td>7</td>
</tr>
<tr>
<td>6</td>
<td>9</td>
</tr>
<tr>
<td>9</td>
<td>8</td>
</tr>
<tr>
<td>9</td>
<td>9</td>
</tr>
<tr>
<td>12</td>
<td>9</td>
</tr>
</tbody>
</table>

![Scatter plot showing regression lines](image)

- Blue line: Regression based on observed data
- Black line: Regression based on all data
Missing at random (MAR)

<table>
<thead>
<tr>
<th>X</th>
<th>Y</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>6</td>
<td>5</td>
</tr>
<tr>
<td>5</td>
<td>6</td>
</tr>
<tr>
<td>8</td>
<td>5</td>
</tr>
<tr>
<td>6</td>
<td>7</td>
</tr>
<tr>
<td>7</td>
<td>7</td>
</tr>
<tr>
<td>6</td>
<td>9</td>
</tr>
<tr>
<td>9</td>
<td>8</td>
</tr>
<tr>
<td>9</td>
<td>9</td>
</tr>
<tr>
<td>12</td>
<td>9</td>
</tr>
</tbody>
</table>

Regression based on observed data
Regression based on all data
Missing not at random (MNAR)

<table>
<thead>
<tr>
<th>X</th>
<th>Y</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>6</td>
<td>5</td>
</tr>
<tr>
<td>5</td>
<td>6</td>
</tr>
<tr>
<td>8</td>
<td>5</td>
</tr>
<tr>
<td>6</td>
<td>7</td>
</tr>
<tr>
<td>7</td>
<td>7</td>
</tr>
<tr>
<td>6</td>
<td>9</td>
</tr>
<tr>
<td>9</td>
<td>8</td>
</tr>
<tr>
<td>9</td>
<td>9</td>
</tr>
<tr>
<td>12</td>
<td>9</td>
</tr>
</tbody>
</table>

Regression based on observed data
Regression based on all data
“Older” methods of handling missing data:
“Older” methods of handling missing data:

- Compute $S$ using **listwise** deletion
  - Assumes MCAR
  - Delete all rows with a missing value
  - Downside: deletes observed data
“Older” methods of handling missing data:

- Compute $S$ using **listwise** deletion
  - Assumes MCAR
  - Delete all rows with a missing value
  - Downside: deletes observed data

- Compute $S$ using **pairwise** estimation
  - Assumes MAR
  - Estimate each element of $S$ using all available data
  - Downside: Each covariance is based on different $n$ and variance–covariance matrix might not be positive definite
“Older” methods of handling missing data:

- **Compute \( S \)** using **listwise** deletion
  - Assumes MCAR
  - Delete all rows with a missing value
  - Downside: deletes observed data

- **Compute \( S \)** using **pairwise** estimation
  - Assumes MAR
  - Estimate each element of \( S \) using all available data
  - Downside: Each covariance is based on different \( n \) and variance–covariance matrix might not be positive definite

- **(Multiple) inputations**
  - Assumes MAR
  - Inpute missingness using mean scores or regression models
  - Downside: complicated, can increase bias if MNAR
Full-information maximum likelihood (FIML):

- Compute likelihood for each person or each data subset with the same missingness pattern
- Assumes MAR
- Uses the full data set and all observations
- Downside: full data needed (analysis can not be done using covariance matrix)
- Implemented in most software (e.g., lavaan, Mplus, psychonetrics)
  - Use missing = "FIML" in lavaan or estimator = "FIML" in psychonetrics
FIML estimator in *psychometrics* for every subset of data $i$ with same missingness pattern:

$$ F_{\text{FIML}} = \frac{1}{n} \sum_i n_i \left( \text{trace} \left( S_i \Sigma_i^{-1} \right) + (\bar{y}_i - \mu_i) \Sigma_i^{-1} (\bar{y}_i - \mu_i) - \ln |\Sigma_i^{-1}| \right) $$

- $n_i$: sample size of subset $i$
- $S_i$: sample covariances (ML) of subset $i$ (note, $S_i = O$ if $n_i = 1$)
- $\bar{y}_i$: sample means of of subset $i$ (note, same as the observed score if $n_i = 1$)
- $\Sigma_i$: Subset of $\Sigma$ containing only elements of observed data in subset $i$
- $\mu_i$: Subset of $\mu$ containing only elements of observed data in subset $i$

Downside: saturated model needs to be computed as well.
Assumptions of maximum likelihood estimation

1. Independence: Observations are a simple random sample from some population
   - Consequence: underestimated standard errors, inflated Type-I error rates
   - Solution 1: use SE correction for dependence structure
   - Solution 2: multilevel SEM

2. Multivariate Normality: Variables are univariate normally distributed at levels of all other variables, residuals are normal and homoscedastic, latent variables are normal, bivariate relations are linear
   - Consequence: standard errors are incorrect (probably too low), $\chi^2$ test value is not accurate (probably too high)
   - Solution 1: use “robust” standard errors (estimator = 'MLM', with complete data; estimator = 'MLR' with incomplete data)
   - Solution 2: use bootstrapped standard errors & test statistic
Higher order models
Higher order models

Mathematically, simply a second factor model on the latent variable variance–covariance matrix:

$$\Psi = \Lambda^* \Psi \Lambda^*^T + \Theta^*$$

Same rules of identification apply:

- The higher order factor must be scaled (one factor loading or the variance fixed to 1)
- The number of variances and covariances in $\Psi$ must be at least as much as the number of parameters used to model $\Psi$
Figure 1: Major stages in the evolution of psychometric theories from Spearman’s g to Cattell-Horn-Carroll (CHC) theory
Bi-factor models
Schmid-Leiman transformed higher order models are useful for assessing explained variance of the general factor.

Many complicated nesting and equivalence relations, even though causal interpretations are vastly different!

What is the \( p \)-factor of psychopathology? Some risks of general factor modeling

Riet van Bork, Sacha Epskamp, Mijke Rhemtulla, Denny Borsboom, Han L. J. van der Maas

First Published November 23, 2017  |  Research Article

https://doi.org/10.1177/0959354317737185

Abstract

Recent research has suggested that a range of psychological disorders may stem from a single underlying common factor, which has been dubbed the \( p \)-factor. This finding may spur a line of research in psychopathology very similar to the history of factor modeling in intelligence and, more recently, personality research, in which similar general factors have been proposed. We point out some of the risks of modeling and interpreting general factors, derived from the fields of intelligence and personality research. We argue that: (a) factor-analytic resolution, i.e., convergence of the literature on a particular factor structure, should not be expected in the presence of multiple highly similar models; and (b) the true underlying model may not be a factor model at all, because alternative explanations can account for the correlational structure of psychopathology.
Exploratory Factor Analysis (EFA)

Exploratorily estimate $\Lambda$ (no free elements in $\Lambda$): \[
\Sigma = \Lambda \Psi \Lambda^\top + \Theta
\]

Very close, but not the same (!!) as principal component analysis (PCA):
\[
\Sigma = \Lambda \Psi \Lambda^\top
\]

Very different interpretation. EFA measures latents (there is measurement error) and captures common variance, PCA only summarizes the data and captures total variance.
Formative (principal component analysis)

Reflective (factor analysis)
Exploratory Factor Analysis (EFA)

If $\Lambda$ is not somehow constrained, latent variable variance is not identified. We can arbitrarily add rotation matrices $R$ and not change the decomposition:

$$
\Sigma = \Lambda R R^{-1} \Psi R^{-1}^T R^T \Lambda^T + \Theta
$$

Can be seen as a different factor model with $\Lambda^* = \Lambda R$ and $\Psi^* = R^{-1} \Psi R^{-1}^T$. To this end, in estimation one can assume uncorrelated factors, $\Psi = I$. Afterwards, rotation methods can be used to obtain simple structure for $\Lambda$ while possibly allowing factors to correlate:

- orthogonal (varimax): axes remain orthogonal, independent
- oblique (promax/oblimin): axes become correlated

I often use promax rotation.
Choosing the number of Factors is a bit more involved than PCA

- One method involves checking how many eigenvalues in $S - \hat{\Theta}$ are above 0
  - $\hat{\Theta}$ is then estimated using a 1-factor model
- Parallel analysis takes sampling variation into account, and checks how many eigenvalues are statistically above what can be expected given an independence model
Big 5 example

```r
library("psych")

# Load data:
data(bfi)
bfiSub <- bfi[,1:25]

# Correlations:
corMat <- cor(bfiSub, use = "pairwise.complete.obs")
N <- nrow(bfiSub)
```
fa.parallel(corMat, N, fa = "fa")

## Parallel analysis suggests that the number of factors = 6 and the number of components = NA
Exploratory versus confirmatory research

The picture of the Texas sharpshooter is taken from an illustration by Dirk-Jan Hoek (CC-BY).
Exploratory versus confirmatory research factor analysis

- Perhaps poorly chosen terms!
- Exploratory factor analysis allows for confirmatory tests!
  - E.g., test if a 5-factor structure generally fits better than a 4-factor structure
- Confirmatory factor analysis on the other hand is often exploratory!
  - E.g., cherry picking a model, many different tests, model modifications with modification indices, etcetera
  - In fact, you could argue CFA suffers from many arbitrary researcher degrees of freedom
- EFA and CFA are generally just very different, and EFA should not always be followed by CFA
  - E.g., personality questionnaires usually show very poor performance with CFA
  - An alternative is also exploratory SEM (ESEM), which puts EFA in a SEM framework
Conclusion

- When data are ordinal, polychoric and polyserial correlations can be computed
- Missing data needs assumption of missing at random (MAR), can be handled best with full-information maximum likelihood (FIML)
- General factor models:
  - Higher-order models
  - Bi-factor models
- Exploratory factor analysis is an alternative to confirmatory factor analysis