SEM 1: Confirmatory Factor Analysis
Week 1 - Common Cause Modeling

Sacha Epskamp

03-04-2017
Factor Analysis

- Many research questions in psychology relate to *psychological constructs*
  - Intelligence, ability, personality traits, disorders, types, etcetera
- To use these in research, they must first be *measured*
  - Questionnaires, tests
- Factor analysis allows us to evaluate how well a test measures presumed underlying traits
  - Exploratory Factor Analysis (EFA): no hypothesis on underlying causal structure
  - Confirmatory Factor Analysis (CFA): Able to test hypotheses
- Allows also to test for *measurement invariance*
  - Does the test measure the same thing in multiple groups?
**Structural Equation Modeling**

- Factor analysis is part of the more general framework of Structural Equation Modeling (SEM)
- SEM allows for testing of causal relationships between both observed and latent variables
- We will discuss SEM in SEM 2!
Today

- Course outline
- Short stats recap
- CFA modeling
  - Very technical, but covers the most important concepts of CFA
  - Next weeks will be more practical and conceptual!
Introduction Round

- Who are you?
- What is your specialization?
- Why did you choose SEM 1?
- What do you expect from SEM 1?
- Will you do SEM 2?
The SEM 1 team

- Sacha Epskamp (sacha.epskamp@gmail.com)
- Gaby Lunansky (G.Lunansky@uva.nl)
- Eiko Fried (eiko.fried@gmail.com)
Schedule

Week 1 – Introduction to common cause modeling
  • Monday April 3: Lecture
  • Wednesday April 5 - Practical

Week 2 – Fitting and modifying CFA models
  • Monday April 10 - Lecture
  • Wednesday April 12 Practical

Week 3 – Measurement invariance, assumptions and power
  • Wednesday April 19 - Lecture

Week 4 – Advanced CFA models
  • Monday April 24 - Lecture
  • Wednesday April 26 Practical

Week 5 – Presentations
  • Monday May 1
Syllabus on Blackboard!

- Individual assignments: 40%
  - Every week one assignment worth 10%
- Group project: 30%
  - Pre-data report, post-data report and presentation (final week)
- Individual project: 30%
  - Report due in final week

NO EXAM!
Individual Assignments

Each week, the assignment will be made available 11:00 on Wednesday, and will be due 11:00 the next Wednesday. Each assignment will contribute to 10% of your grade.

- Work on the assignments alone.
- Hand in a PDF file and an .R file (in case R was used). If you use Jasp, hand in the Jasp object as well as a screenshot of the options used.
- Make sure your PDF report is as standalone readable as possible. E.g., if you are asked to report a factor loading matrix, then report it in the PDF and not just say “look at .R file”.
- Assignments are due before 11:00. If you do not hand in an assignment before 11:00, you will get a 1.
- If you encounter any problems, or have any feedback, please let me know before the deadline, as then I can take it into account or help you.
Group Project

In this project, you will be asked to go through the entire empirical cycle of performing a confirmatory factor analysis. The group project will contribute to 30% of your final grade. Form groups of 4-5 people (ideally at most 6 groups).

- Step 1: Research setup
- Step 2: Research hypothesis
  - Pre-data report (5%)
- Step 3: Data collection
- Step 4: Data analysis
- Step 5: Present your findings
  - Presentation (5%)
  - Post-data report (20%)

You have to justify each member’s contributions in the final report. See syllabus for details!
Individual Project

• Find a paper that estimated a CFA model and reported the data
  • Analysis must not have originally been performed in Lavaan, Jasp or Onyx
• Ideally a slightly complicated analysis (many factors, higher order model, bifactor model, latent growth model or measurement invariance tests)
• Replicate the analysis in Lavaan, Onyx or Jasp
• Write a report on your findings (30% of your grade), critically discussing the original paper. Did you replicate their findings? Did the original authors fail to mention important details? Try to include as many steps as possible as recommended by Brown

See syllabus for details!
Practicals

The practicals will have three goals:

- Work on group project
- Work on individual project
- Discuss assignments

Your attendance is not required for lectures or practicals (your group might think differently about this though). I recommend you to start forming groups as soon as possible, and use the first practical to setup your questionnaire and work on the pre-data report. Start collecting data as soon as possible.
Course roughly follows the book

Book is not required for the course

Reference material

Some differences in notation and computation as to book

We use different software as in the book
Presumed Prior Knowledge

- Undergraduate level statistics
  - Linear regression
  - Correlations / covariances
  - Normal distribution
- Matrix algebra
  - Matrix multiplication!
- R
  - Although you are allowed to work in Jasp (essentially GUI around the R package we will use) or Onyx (a graphical SEM program), knowing R is highly recommended

If you lack prior knowledge, you are expected to try and catch up on these topics! I can send you some materials if needed
Scalars, Vectors and Matrices

Normal faced letters indicate scalars (just a single number), **boldfaced** letters indicate *column vectors*, and **UPPERCASE BOLDFACED** letters indicate *matrices*. If a scalar is an element of a matrix, the first subscript will denote the *row* and the second subscript the *column*. If a vector is part of a matrix, the subscript will denote either which row or column (the vector is always transformed to a column-vector).

\[
Y = \begin{bmatrix}
1 & 2 & 3 \\
4 & 5 & 6 \\
7 & 8 & 9 \\
10 & 11 & 12
\end{bmatrix}
\]

\[
y_2 = \begin{bmatrix}
4 \\
5 \\
6
\end{bmatrix}
\]

\[y_{23} = 6\]
Roman and Greek letters

A Roman letter indicates something that is *observed* (e.g., observed scores $y_i$ of subject $i$) or *specified* (e.g., sample size $n$). A Greek letter indicates something that is *not observed*, e.g., latent variable scores $\eta_i$ or factor loadings matrix $\Lambda$. 
<table>
<thead>
<tr>
<th>Greek Letter</th>
<th>English Name</th>
<th>Pronunciation</th>
<th>Pronunciation (Greek)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Αα</td>
<td>Alpha</td>
<td>[a]</td>
<td>αλφα</td>
</tr>
<tr>
<td>Ββ</td>
<td>Beta</td>
<td>[b]</td>
<td>βήτα</td>
</tr>
<tr>
<td>Γγ</td>
<td>Gamma</td>
<td>[g]</td>
<td>γάμμα</td>
</tr>
<tr>
<td>Δδ</td>
<td>Delta</td>
<td>[d]</td>
<td>δέλτα</td>
</tr>
<tr>
<td>Εε</td>
<td>Epsilon</td>
<td>[e]</td>
<td>εψιλόν</td>
</tr>
<tr>
<td>Ζζ</td>
<td>Zeta</td>
<td>[dz]</td>
<td>ζήτα</td>
</tr>
<tr>
<td>Ηη</td>
<td>Eta</td>
<td>[ɛ]</td>
<td>ητα</td>
</tr>
<tr>
<td>Θθ</td>
<td>Theta</td>
<td>[θ]</td>
<td>θέτα</td>
</tr>
<tr>
<td>Ιι</td>
<td>Iota</td>
<td>[i]</td>
<td>ιωτα</td>
</tr>
<tr>
<td>Κκ</td>
<td>Kappa</td>
<td>[k]</td>
<td>κατσια</td>
</tr>
<tr>
<td>Λλ</td>
<td>Lambda</td>
<td>[l]</td>
<td>λαμβδα</td>
</tr>
<tr>
<td>Μμ</td>
<td>Mu</td>
<td>[m]</td>
<td>μο</td>
</tr>
<tr>
<td>Νν</td>
<td>Nu</td>
<td>[n]</td>
<td>νο</td>
</tr>
<tr>
<td>Ξξ</td>
<td>Xi</td>
<td>[ks]</td>
<td>ξει</td>
</tr>
<tr>
<td>Οο</td>
<td>Omicron</td>
<td>[o]</td>
<td>ομικρον</td>
</tr>
<tr>
<td>Ππ</td>
<td>Pi</td>
<td>[p]</td>
<td>πει</td>
</tr>
<tr>
<td>Ρρ</td>
<td>Rho</td>
<td>[r]</td>
<td>ρω</td>
</tr>
<tr>
<td>Σσζζ</td>
<td>Sigma</td>
<td>[s]</td>
<td>σιμα</td>
</tr>
<tr>
<td>Ττ</td>
<td>Tau</td>
<td>[t]</td>
<td>ταυ</td>
</tr>
<tr>
<td>Υυ</td>
<td>Upsilon</td>
<td>[u]</td>
<td>υπιλον</td>
</tr>
<tr>
<td>Φφ</td>
<td>Phi</td>
<td>[φ]</td>
<td>φει</td>
</tr>
<tr>
<td>Χχ</td>
<td>Chi</td>
<td>[χ]</td>
<td>χει</td>
</tr>
<tr>
<td>Ψψ</td>
<td>Psi</td>
<td>[ps]</td>
<td>ψει</td>
</tr>
<tr>
<td>Ωω</td>
<td>Omega</td>
<td>[o]</td>
<td>ω μεγα</td>
</tr>
</tbody>
</table>
Random variables and parameters

Subscript $i$ will denote the case, the row in your data set. Often this will correspond to a particular subject, and thus we will use the term subject often to denote cases.

- A variable with subscript $i$ can differ per case. These are called random variables.
  - e.g., observed responses $y_i$, latent variables $\eta_i$ or residual $\varepsilon_i$
- I will omit subscript $i$ from random variables when I describe their distribution.
  - e.g., $y \sim N(...)$, indicating that the distribution of all subjects response patterns follow a normal distribution
- Letters without subscript $i$ to begin with do not differ per case. These are called parameters. They have no distribution (no variance across people)
  - e.g., factor loadings $\Lambda$, variance–covariance matrices $\Sigma$, $\Psi$ and $\Theta$
The transpose operator, $\top$, switches rows and columns:

\[
Y = \begin{bmatrix}
1 & 2 & 3 \\
4 & 5 & 6 \\
7 & 8 & 9 \\
10 & 11 & 12
\end{bmatrix}
\]

\[
Y^\top = \begin{bmatrix}
1 & 4 & 7 & 10 \\
2 & 5 & 8 & 11 \\
3 & 6 & 9 & 12
\end{bmatrix}
\]

\[
y_2^\top = \begin{bmatrix}
4 & 5 & 6
\end{bmatrix}
\]
Matrix Multiplication

If $AB = C$, with $A$ being a $x \times y$ matrix and $B$ being a $y \times z$ matrix, then $C$ is a $x \times z$ matrix with following elements:

$$c_{ij} = \sum_{k=1}^{y} a_{ik} b_{kj}$$

For example:

$$\begin{bmatrix} 4 & 5 & 8 \\ 2 & 1 & 7 \\ 3 & 6 & 9 \end{bmatrix} \begin{bmatrix} 5 & 1 & 9 \\ 4 & 3 & 6 \\ 2 & 8 & 7 \end{bmatrix} = \begin{bmatrix} 56 & 83 & 122 \\ 28 & 61 & 73 \\ 57 & 93 & 126 \end{bmatrix}$$

In R, use `%*%` to compute this!
Matrix Multiplication

Multivariate Normal

We assume data are multivariate normally distributed:

\[ y_i \sim N(\mu, \Sigma) \]

- \( y_i \) is a vector of item responses (e.g., answers to “are you fatigued?”, “are you concentrating well?”) of person \( i \)
- \( \mu \) is a vector of means
- \( \Sigma \) is a variance–covariance matrix, standardizable to a correlation matrix.
Example data of $n = 10$ subjects:

$$
Y = \begin{bmatrix}
\mathbf{y}_1^\top \\
\mathbf{y}_2^\top \\
\mathbf{y}_3^\top \\
\mathbf{y}_4^\top \\
\mathbf{y}_5^\top \\
\mathbf{y}_6^\top \\
\mathbf{y}_7^\top \\
\mathbf{y}_8^\top \\
\mathbf{y}_9^\top \\
\mathbf{y}_{10}^\top
\end{bmatrix} = \begin{bmatrix}
4.09 & 4.94 & 4.58 \\
2.48 & 4.27 & 5.44 \\
4.93 & 4.24 & 6.24 \\
2.17 & 5.36 & 5.79 \\
3.39 & 4.62 & 4.81 \\
1.41 & 3.74 & 5.21 \\
2.87 & 2.60 & 5.17 \\
2.62 & 2.37 & 4.55 \\
3.49 & 3.74 & 6.75 \\
4.10 & 5.69 & 3.00
\end{bmatrix}
$$
Means

The center of the normal distribution of variable $y_j$ is controlled by its \textit{mean} $\mu_j$:
Means

Means per column:

\[ \bar{y}_j = \frac{1}{n} \sum_{p=1}^{n} y_{ij}, \]

in which \( y_{ij} \) indicates row \( i \), column \( j \) of data matrix \( Y \). Combining these values in a vector \( \bar{y} \) we get:

\[ \bar{y}^\top = [3.16 \quad 4.16 \quad 5.16] \]

In R: `colMeans()`. This is an unbiased estimator of \( \mu \):

\[ \mathcal{E} (\bar{y}) = \mu \]
Diagonal elements of $\Sigma$ correspond to variances:

$$\text{Var}(y_j) = \sigma_j^2 = \sigma_{jj}$$

It’s square root equals the standard deviation, which usually is interpreted:

$$\text{SD}(y_j) = \sqrt{\text{Var}(y_j)} = \sigma_j$$

This parameter controls the width of the normal distribution. E.g., 95% of the probability mass is contained in $(\mu_j - 1.96\sigma_j, \mu_j + 1.96\sigma_j)$
Covariances encode association, a positive covariance encodes that when one variable is above (below) its mean, the other variable probably also is above (below) its mean. Negative covariances encode that whenever one variable is above (below) its mean, the other variable is likely below (above) its mean:
Covariances can be standardized to **correlations**, which are always between $-1$ and $1$. 
Variance–covariance matrix

Computing the variance–covariance matrix:

\[ s_{jk}^* = \frac{\sum_{i=1}^{n} (y_{ij} - \bar{y}_j)(y_{ik} - \bar{y}_k)}{n - 1}, \]

Which becomes:

\[ S^* = \begin{bmatrix}
1.11 & 0.35 & -0.09 \\
0.35 & 1.18 & -0.29 \\
-0.09 & -0.29 & 1.07
\end{bmatrix} \]

In R: `cov()`. This is an unbiased estimator of \( \Sigma \):

\[ \mathcal{E} (S^*) = \Sigma \]
Variance–covariance matrix

Dividing by $n - 1$ leads to an unbiased estimator, but not the maximum likelihood estimate! This is obtained by dividing by $n$:

$$s_{jk} = \frac{\sum_{i=1}^{n}(y_{ij} - \bar{y}_j)(y_{ik} - \bar{y}_k)}{n},$$

The R function `cov()` returns the unbiased estimator (divided by $n - 1$), to obtain the ML solution:

$$S = \frac{n - 1}{n} S^*$$

Lavaan, the software we use, automatically applies this correction! So you use $S^*$ as input, but $S$ in computations.
Correlation matrix

Standardizing to correlations:

\[ r_{jk} = \frac{S_{jk}}{\sqrt{S_{jj}} \sqrt{S_{kk}}}, \]

which becomes

\[
R = \begin{bmatrix}
1.00 & 0.31 & -0.08 \\
0.31 & 1.00 & -0.26 \\
-0.08 & -0.26 & 1.00
\end{bmatrix}.
\]

In R: `cov2cor()` (or simply use `cor()` on data)
Psychometrics

The field of science concerned with measurement of psychological constructs. For example:

• Personality traits
• Cognitive skills
• Ability
• Psychopathological disorders
• Types
Psychometrics

The field of science concerned with measurement of psychological constructs.
Psychometrics

The field of science concerned with measurement of psychological constructs.
For example:

- Personality traits
- Cognitive skills
- Ability
- Psychopathological disorders
- Types
How do you measure temperature?
How do you measure temperature?

- By looking at a thermometer
- For this to make sense, we need to assume that:
  - Temperature *causes* the level shown in a thermometer
  - The thermometer features relatively little measurement error
We can summarize a causal hypothesis in a **path diagram**:

- Circular nodes (or suns): latent (unobserved) variables
We can summarize a causal hypothesis in a path diagram:

- Circular nodes (or suns): latent (unobserved) variables
- Square nodes: observed variables
We can summarize a causal hypothesis in a path diagram:

- Circular nodes (or suns): latent (unobserved) variables
- Square nodes: observed variables
  - Also termed *indicators* of a latent variable
We can summarize a causal hypothesis in a path diagram:

- Circular nodes (or suns): latent (unobserved) variables
- Square nodes: observed variables
  - Also termed *indicators* of a latent variable
- Unidirectional links: causal effects
We can summarize a causal hypothesis in a path diagram:

- Circular nodes (or suns): latent (unobserved) variables
- Square nodes: observed variables
  - Also termed *indicators* of a latent variable
- Unidirectional links: causal effects
- Bidirectional links: (co)variances
Assume all observed and latent variables are normally distributed and all causal effects are linear. Without loss of information, we can center data and assume all means are 0 (we don’t use means until week 3). Now, the path diagram encodes a causal equation:

\[
\psi_{11} \quad \eta_1 \quad \lambda_{11} \quad y_1 \quad \varepsilon_1 \quad \theta_{11}
\]

\[
y_{i1} = \lambda_{11}\eta_{i1} + \varepsilon_{i1}
\]

\[
\eta_1 \sim N(0, \sqrt{\psi_{11}})
\]

\[
\varepsilon_1 \sim N(0, \sqrt{\theta_{11}})
\]

\(\lambda_{11}\) is called a factor loading, \(\varepsilon_{i1}\) the residual variance and \(\psi_{11}\) the factor variance.
\( y_{i1} = \lambda_{11} \eta_{i1} + \varepsilon_{i1} \)
\( \eta_1 \sim N(0, \sqrt{\psi_{11}}) \)
\( \varepsilon_1 \sim N(0, \sqrt{\theta_{11}}) \)

Variance of \( y_1 \):
\[
\text{Var}(y_1) = \lambda_{11}^2 \psi_{11} + \theta_{11}
\]
How much variance does the latent variable explain?
\[
\frac{\lambda_{11}^2 \psi_{11}}{\lambda_{11}^2 \psi_{11} + \theta_{11}}
\]
Multiplying $\lambda_{11}$ by constant $c$ and dividing $\psi_{11}$ by $c^2$ leads to the same variance. Thus, these parameters are not identified. We need to scale the latent variable by fixing one of these parameters. For instance, let $\lambda_{11} = 1$

\[ y_{i1} = \eta_{i1} + \varepsilon_{i1} \]

\[ \eta_1 \sim N(0, \sqrt{\psi_{11}}) \]

\[ \varepsilon_1 \sim N(0, \sqrt{\theta_{11}}) \]

\[ \text{Var}(y_1) = \psi_{11} + \theta_{11} \]
Problems in Estimating Latent Variables

- Well known problem in Psychometrics: you can not jointly estimate the *parameters* (parameters stable over people, such as $\lambda_{11}$) and *latent variables* (e.g., $\eta_{i1}$)
  - We need to be able to add data to get better estimates (e.g., smaller standard error), but adding an item adds parameters and adding a person adds latent variables!

- Paradoxically, the first step in solving this problem is getting rid of the latent variable $\eta_{i1}$
  - In item-response theory (binary items), the latent variable is integrated out
  - In factor analysis, we make use of *covariance modeling*

- After estimating parameters, the latent variable can be extrapolated. Although as we will see in SEM 2, often this is not needed and we can use covariance modeling to test all our hypotheses!
We match **observed variances and covariances** to parameters so that we do not need to estimate latent variables themselves. If the parameters reproduce the observed variances and covariances, we explain the data!

\[ y_{i1} = \eta_{i1} + \varepsilon_{i1} \]

\[ \eta_1 \sim N(0, \sqrt{\psi_{11}}) \]

\[ \varepsilon_1 \sim N(0, \sqrt{\theta_{11}}) \]

\[ \text{Var}(y_1) = \psi_{11} + \theta_{11} \]
We match observed variances and covariances to parameters so that we do not need to estimate latent variables themselves. If the parameters reproduce the observed variances and covariances, we explain the data!

\[ y_{i1} = \eta_{i1} + \varepsilon_{i1} \]

\[ \eta_1 \sim N(0, \sqrt{\psi_{11}}) \]

\[ \varepsilon_1 \sim N(0, \sqrt{\theta_{11}}) \]

\[ \text{Var}(y_1) = \psi_{11} + \theta_{11} \]

Problem: two parameters and only one observed variance, model is under-identified!
Covariance modeling

In general, we aim to estimate parameters that lead to a model implied variance–covariance matrix $\Sigma$ by minimizing (note that the Brown book makes a mistake here):

$$F_{ML} = \text{trace} \left( S \Sigma^{-1} \right) - \ln | S \Sigma^{-1} | - p,$$

in which $S$ is the observed variance–covariance matrix, the trace operator takes the sum of diagonal values, the $| \ldots |$ notation indicates the determinant and $p$ is the number of observed variables. Optimizing this expression is called maximum likelihood estimation. In principle, this expression is optimized if $\Sigma$ resembles $S$ as much as possible! When $S = \Sigma$, $F_{ML} = 0$. 
Identification

Two main rules:

• Because the unit of the latent variable is unknown, we need to **scale** the latent variable by fixing its variance to 1 or fixing one (usually the first) factor loading to 1.

• We need at least as many observations (sample variances and covariances) as the number or parameters; we require non-negative **degrees of freedom** (DF)
  
  - DF = a − b
  - a: number of observations: $a = \frac{p(p + 1)}{2}$ variances and covariances.
  - b: number of parameters we need to estimate (do not count parameters we fixed for scaling)

• In general, we need 3 indicators for a single latent variable model, or 2 per factor for models with multiple (correlated) latent variables.
We need at least as many observed variances and covariances as parameters!

\[ y_{i1} = \eta_{i1} + \varepsilon_{i1} \]
\[ \eta_1 \sim N(0, \sqrt{\psi_{11}}) \]
\[ \varepsilon_1 \sim N(0, \sqrt{\theta_{11}}) \]
\[ \text{Var}(y_1) = \psi_{11} + \theta_{11} \]

Two parameters, one observation. DF = −1. Under-identified!
$\eta_{i1} \sim N(0, \sqrt{\psi_{11}})$
$\varepsilon_1 \sim N(0, \sqrt{\theta_{11}})$
$\varepsilon_2 \sim N(0, \sqrt{\theta_{22}})$

$y_{i1} = \eta_{i1} + \varepsilon_{i1}$
$y_{i2} = \lambda_{21} \eta_{i1} + \varepsilon_{i2}$

$\eta_{i1}$ is now a common cause. $y_{i1}$ and $y_{i2}$ are assumed independent after controlling for $\eta_{i1}$: local independence. Number of observations: 2 variances ($\operatorname{Var}(y_{1})$ and $\operatorname{Var}(y_{2})$) + 1 covariance ($\operatorname{Cov}(y_{1}, y_{2})$) = 3.

Model is under-identified!
General factor analysis framework:

\[ y_i = \Lambda \eta_i + \varepsilon_i \]
\[ y \sim N(0, \Sigma) \]
\[ \eta \sim N(0, \Psi) \]
\[ \varepsilon \sim N(0, \Theta), \]

in which:

- \( y_i \) is a \( p \)-length vector of item responses
- \( \eta_i \) an \( m \)-length vector of latent variables
- \( \varepsilon_i \) an \( p \)-length vector of residuals
- \( \Lambda \) a \( p \times m \) matrix of factor loadings
- \( \Psi \) an \( m \times m \) symmetric variance–covariance matrix (assume always all latent variables are correlated)
- \( \Theta \) is a \( p \times p \) symmetric variance–covariance matrix, mostly diagonal (unless you explicitly expect violations of local independence)
The general framework:

\[ y_i = \Lambda \eta_i + \varepsilon_i \]
\[ y \sim N(0, \Sigma) \]
\[ \eta \sim N(0, \Psi) \]
\[ \varepsilon \sim N(0, \Theta), \]

Allows you to derive the model-implied variance–covariance matrix:

\[ \Sigma = \Lambda \Psi \Lambda^\top + \Theta \]
\[ \Lambda = [1], \Psi = [\psi_{11}], \Theta = [\theta_{11}] \]

\[ \Sigma = [\psi_{11} + \theta_{11}] \]
Upper triangular elements in symmetric matrices not shown.

\[ \Lambda = \begin{bmatrix} 1 & \lambda_{21} \end{bmatrix}, \quad \Psi = [\psi_{11}], \quad \Theta = \begin{bmatrix} \theta_{11} & 0 \\ 0 & \theta_{22} \end{bmatrix} \]

\[ \Sigma = \begin{bmatrix} \psi_{11} + \theta_{11} & \lambda_{21}^2 \psi_{11} + \theta_{22} \\ \psi_{11} \lambda_{21} & \lambda_{21}^2 \psi_{11} + \theta_{22} \end{bmatrix} \]
\[ \Lambda = \begin{bmatrix} 1 & \\ \lambda_{21} & \lambda_{31} \end{bmatrix}, \quad \Psi = [\psi_{11}], \quad \Theta = \begin{bmatrix} \theta_{11} & \theta_{22} \\ 0 & 0 & \theta_{33} \end{bmatrix} \]

\[ \Sigma = \begin{bmatrix} \psi_{11} + \theta_{11} & \psi_{11} \lambda_{21} & \psi_{11} \lambda_{31} \\ \psi_{11} \lambda_{21} & \lambda_{21}^2 \psi_{11} + \theta_{22} & \psi_{11} \lambda_{21} \lambda_{31} \\ \psi_{11} \lambda_{31} & \psi_{11} \lambda_{21} \lambda_{31} & \lambda_{31}^2 \psi_{11} + \theta_{33} \end{bmatrix} \]

DF = 0, just identified (but saturated, will explain the data perfectly)
\[ \Lambda = \begin{bmatrix} 1 & 0 \\ \lambda_{21} & 0 \\ \lambda_{31} & 0 \\ 0 & 1 \\ 0 & \lambda_{52} \\ 0 & \lambda_{62} \end{bmatrix}, \quad \Psi = \begin{bmatrix} \psi_{11} \\ \psi_{21} \\ \psi_{22} \end{bmatrix}, \quad \Theta = \begin{bmatrix} \theta_{11} & 0 & 0 & 0 & \theta_{44} & \theta_{55} & \theta_{66} \\ 0 & \theta_{22} & 0 & 0 & 0 \\ 0 & 0 & \theta_{33} & 0 & 0 \\ 0 & 0 & 0 & \theta_{44} & 0 \\ 0 & 0 & 0 & 0 & \theta_{55} \\ 0 & 0 & 0 & 0 & 0 & \theta_{66} \end{bmatrix} \]
\[ \Lambda = \begin{bmatrix} 1 & 0 \\ \lambda_{21} & 0 \\ \lambda_{31} & 0 \\ 0 & 1 \\ 0 & \lambda_{52} \\ 0 & \lambda_{62} \end{bmatrix} , \quad \Psi = \begin{bmatrix} \psi_{11} \\ \psi_{21} \\ \psi_{22} \end{bmatrix} , \quad \Theta = \begin{bmatrix} \theta_{11} & 0 & \theta_{22} \\ 0 & \theta_{33} \\ 0 & \theta_{43} & \theta_{44} \\ 0 & 0 & 0 & \theta_{55} \\ 0 & 0 & 0 & 0 & \theta_{66} \end{bmatrix} \]

Residual covariance / correlation
Cross-loading
Conclusion

- Factor analysis allows one to test a measurement model
  - Does a test or questionnaire measure a single latent variable?
- Done via the common cause model
  - one or more latent variables cause responses on observed indicators
  - Indicators assumed to be mostly locally independent
- Estimation done by estimating elements of $\Lambda$, $\Psi$ and $\Theta$
- These matrices result in model implied variance–covariance matrix $\Sigma$, which should resemble observed variance–covariance matrix $S$
Practical

- Form groups and start working on group project!
- Aim to have a questionnaire designed by the end of the first practical
- You can ask on the blackboard forum if can not find a group
- First assignment will be available on Wednesday
Thank you for your attention!