Introduction to CFA

PSY544 – Introduction to Factor Analysis

Week 11

Introduction

- So far, all we covered was exploratory factor analysis (EFA) – or “unrestricted” factor analysis

- We’ve covered a huge a chunk of stuff and you should be proud of yourself! You have all the knowledge you need to become master exploratory factor analysts.

- You’ve also managed to see me twice a week and not jump out of the window.

Introduction

- Today, we begin with the rest of the course, which will cover confirmatory factor analysis (CFA) – or “restricted” factor analysis.

- The difference between EFA and CFA lies in the incorporation of prior hypothesis about the factor structure into the model specification.

- In EFA, the analyst seeks to explore the number and nature of the major common factors. Rotation to simple structure is usually necessary.
- In CFA, the analyst has a specific prior hypothesis about the number and nature of the major common factors. This hypothesis is directly incorporated into model specification. No rotation is involved.

Introduction

- **In both EFA and CFA, we employ the same statistical model – the common factor model. So, all you have learned about the common factor model previously in this course doesn't change a bit.**

- **However, CFA requires additional assumptions concerning the number and positions of (typically) zero loadings in the factor loading matrix $\boldsymbol{\Lambda}$ to reflect the prior hypothesis.**

- **The methodology of CFA is also pretty different from that of EFA.**

Introduction

- It is no longer possible to obtain estimates of the factor loadings once the unique factor variances (or communalities) are estimated.

- All parameters in CFA have to be estimated simultaneously by numerically minimizing some discrepancy function.

- In effect, CFA tends to be slower than EFA, even though the number of estimated parameters is smaller.

Software

- Plethora of software exists for confirmatory factor analysis or – more generally – for structural equation modeling (of which FA is a special case)

- LISREL, EQS, Mplus, RAMONA, SePATH, Mx, AMOS…

- …meh. In this course, we will use R and the *lavaan* package. For all its quirks and a steep learning curve, it's a modern piece of software that allows for great flexibility. Oh, and you don't have to sell a kidney to work with it – it's free.

Exploratory (Unrestricted) Factor Analysis

- As you already know, in EFA, there are typically no (solid) prior ideas about the number of the common factors or their nature (the position of zero loadings)

- Sure, the analyst might have *some* ideas about the variables being analyzed, these don’t need to be expressed nor they need to be correct.
- If the analyst conducts a blind rotation (like Quartimax) of the estimated factors, they will never know if the failure to see non-zero loadings where expected is because their hypothesis is incorrect or whether the rotational criterion is inadequate for the given situation.

Exploratory (Unrestricted) Factor Analysis

- Also, in EFA, the decision when to stop is based heavily on the analyst’s judgement and the entire thing is largely data-driven rather than theory-driven.

- That’s fine, as long as it’s acknowledged as such.

Confirmatory (Restricted) Factor Analysis

- CFA should be used only when there is a solid prior hypothesis about the number and nature of the common factors.
- It’s totally fine (actually preferable in a lot of cases) to have several competing hypotheses.

- The analyst must be able to specify the number and position of zero loadings before the analysis. After that, the corresponding models are fit to data and the degree of model-data fit is assessed, which suggests the extent to which the prior hypothesis fits the empirical reality.

Confirmatory (Restricted) Factor Analysis

- CFA is not a data-driven enterprise. It’s theory-driven.

- CFA can seduce you to use it in a data-driven way. That’s dangerous, because it can lead to “confirmatory” models that are merely statistical artifacts.

- Confirmatory model is still a model. As such, it is nothing but an approximation. Make sure to be just as cautious in this regard as you would be while performing EFA.

Constraints

- In both EFA and CFA, the specification of the model involves certain kinds of constraints.
- Constraints are equations that reference some model parameters and the model must satisfy them. For example:

$$\lambda_{11} = 0 \qquad \varphi_{22} = 1 \qquad \lambda_{21} = \lambda_{41}$$

- There are two kinds of constraints – **identification conditions** and **restrictions.**

Constraints

- **Identification conditions** are imposed to select one particular solution from a class of infinitely many solutions that would produce the same model-implied correlation/covariance matrix.

- They do not affect the fit of the model, and hence are not testable.

Constraints

- **Restrictions** usually represent aspects of a prior hypothesis and serve to represent that hypothesis.

- They do affect the implied correlation/covariance matrix, hence they do affect the fit of the model, and so are testable.

Constraints

- In EFA, we only impose the identification conditions.

- It’s the restrictions that make CFA what it is. We only impose them in CFA.

- However, we still need the identification conditions in CFA. Because we impose more constraints (identification conditions + restrictions), CFA will usually fit data more poorly than EFA.

The CFA model

- **Both CFA and EFA employ the same model – the common factor model.**

- **The data model:**

$$\boldsymbol{x} = \boldsymbol{\mu} + \boldsymbol{\Lambda z} + \boldsymbol{u}$$

- **The covariance structure:**

$$\boldsymbol{\Sigma} = \boldsymbol{\Lambda\Phi\Lambda'} + \boldsymbol{D_{\psi}}$$

The CFA model

- **The matrices $\boldsymbol{\Lambda}$, $\boldsymbol{\Phi}$ and $\boldsymbol{D_\psi}$ are the "parameter matrices".**

- **A prior hypothesis about the number and nature of the common factors is incorporated into the analysis through specifying the elements of these three matrices.**

- **There are three possible kinds of parameters:**
  - **Free parameters (unknown and to be estimated)**
  - **Fixed parameters (fixed a priori to a certain value)**
  - **Constrained parameters (tied to the value of another parameter(s))**

The CFA model

- **Consider a situation where you have six MVs (x_1 through x_6) and two common factors (z_1 and z_2).**

- **Your hypothesis is that:**
  - **The factor loadings of the first three MVs on z_1 are substantial and those of the other three MVs are essentially zero.**
  - **The factor loadings of the first three MVs on z_2 are essentially zero and those of the other three MVs are substantial.**
  - **The two common factors z_1 and z_2 are correlated.**

The CFA model

- The first step is to write down the relevant dimensions. We have p = 6 MVs and m = 2 common factors.

- As such, we know that $\boldsymbol{\Lambda}$ is a 6 x 2 matrix, $\boldsymbol{\Phi}$ is a 2 x 2 matrix and $\boldsymbol{D}_{\psi}$ is a 6 x 6 matrix.

- We can build the matrices and impose the restrictions that follow from our hypothesis.

The CFA model

$$\boldsymbol{\Lambda} = \begin{bmatrix} \lambda_{11} & 0 \\ \lambda_{21} & 0 \\ \lambda_{31} & 0 \\ 0 & \lambda_{42} \\ 0 & \lambda_{52} \\ 0 & \lambda_{62} \end{bmatrix} ;\ \boldsymbol{\Phi} = \begin{bmatrix} 1 & \\ \phi_{21} & 1 \end{bmatrix}$$

The CFA model

$$\boldsymbol{D}_{\psi} = \begin{bmatrix} \psi_{11} & & & & & \\ & \psi_{22} & & & & \\ & & \psi_{33} & & & \\ & & & \psi_{44} & & \\ & & & & \psi_{55} & \\ & & & & & \psi_{66} \end{bmatrix}$$

The CFA model

- How many degrees of freedom does our model have?

- Well, first of all, let's count the number of parameters we are freely estimating. That's six factor loadings, six unique variances, and one correlation between factors, a total of 13 parameters to estimate.

- Our data is a 6 x 6 correlation / covariance matrix, which has [6 * (6+1)]/2 = 21 unique elements – the number of degrees of freedom for the null model.
- In our case, the DF number is 21 – 13 = 8 degrees of freedom.

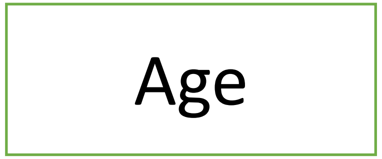
Path diagrams

- Path diagrams are a standard way to communicate a CFA model

- Let’s spend some time on the basics.

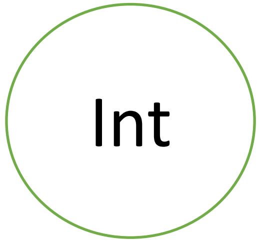
Path diagrams

- Rectangles denote manifest variables



Age

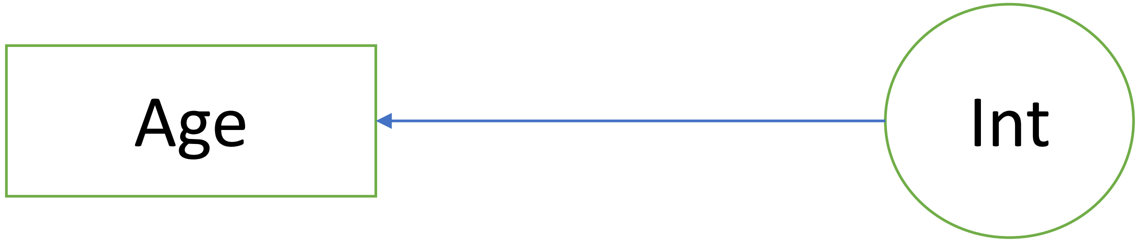
- Circles denote latent variables



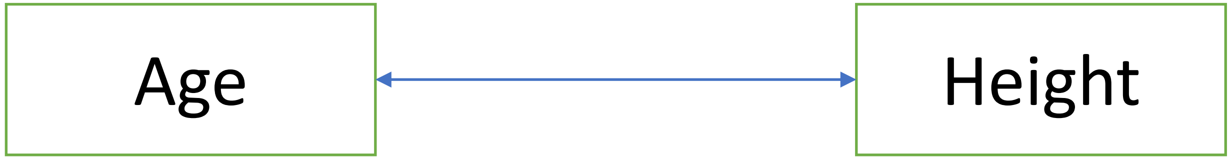
Int

Path diagrams

- One-sided, linear arrows denote a regression path



- Double-sided, curved arrows denote a correlation / covariance (pretend like the line is curved, OK?)

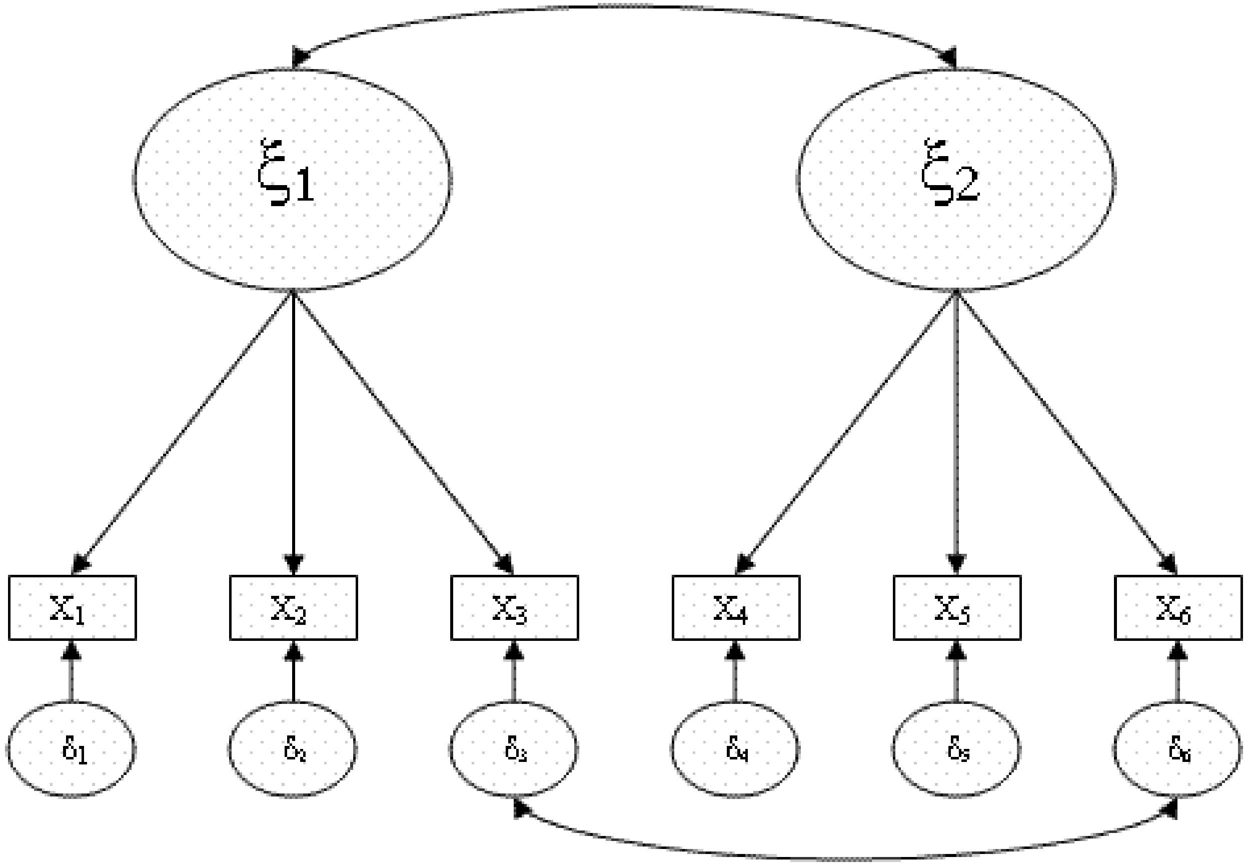


Path diagrams

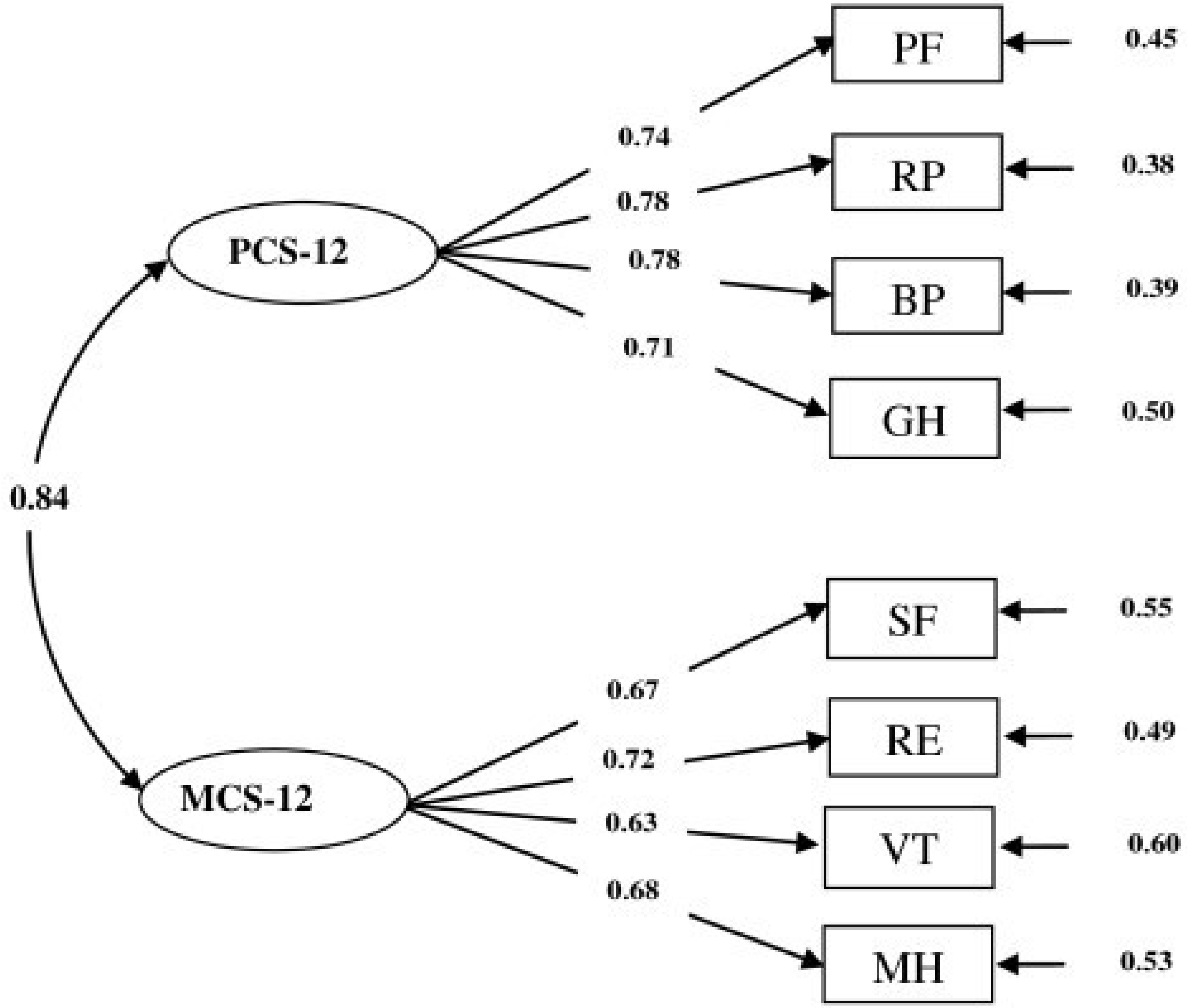
- A path diagram should contain as much model-related information as possible, ideally **all of it**

- Each arrow stands for a parameter, and so should be labeled with the value of that particular parameter

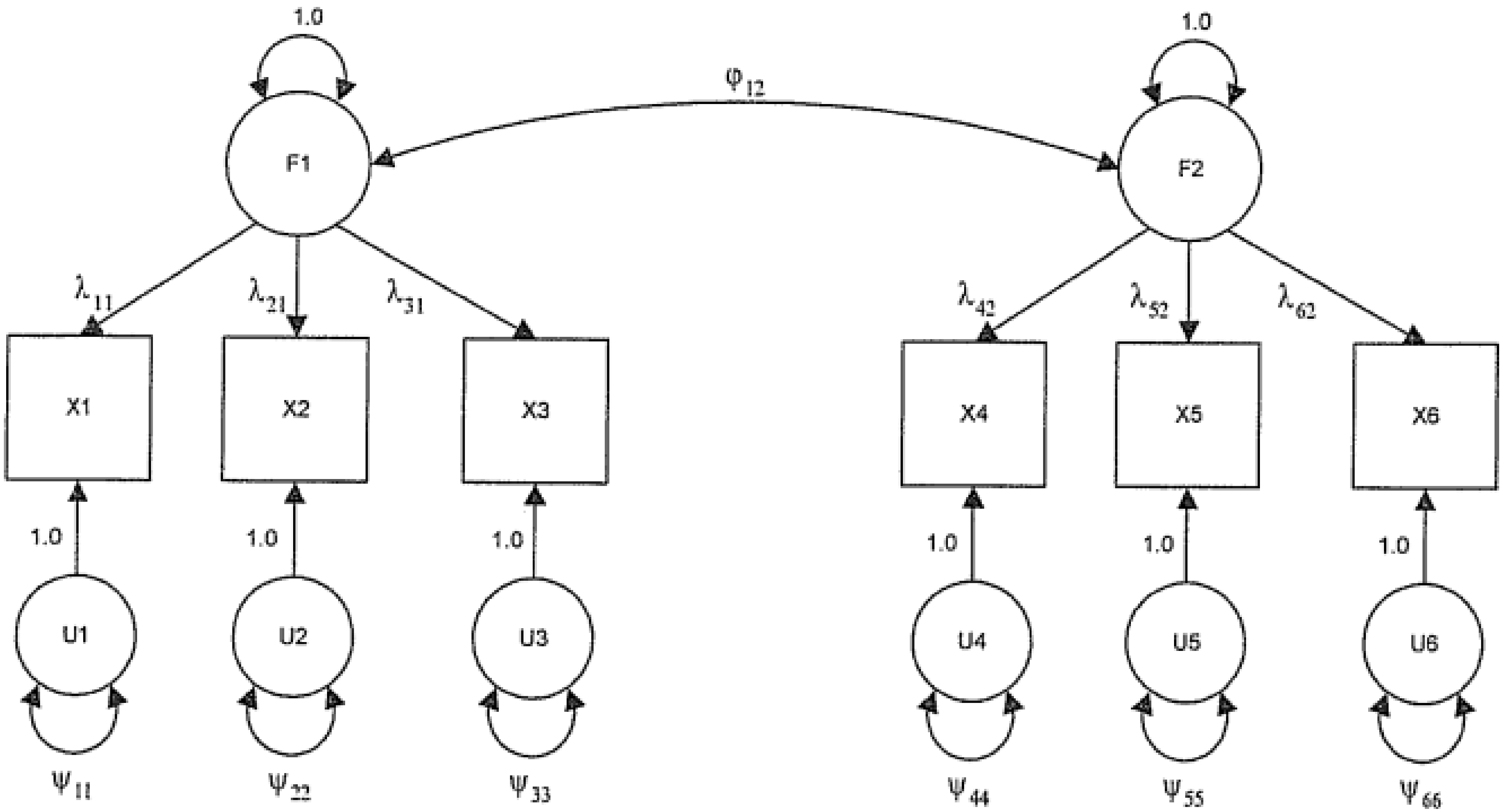
Path diagrams



Path diagrams



Path diagrams



Path diagrams

- Certain kinds of software will allow you to specify model using only path diagrams (LISREL, for instance), while for some software, path diagrams are the only way to specify a model (AMOS, I think) – however, even in AMOS, the software will “translate” the information contained in the path diagram into the model matrices.

Estimation

- As previously with EFA, estimation of parameters in CFA follows the basic principles of minimum discrepancy estimation.
- We are looking for a vector of parameters for which the following is true: *the model-implied covariance / correlation matrix has minimum “distance” from the observed covariance / correlation matrix*

  (in other words, the discrepancy function value is at a minimum)

Estimation

- Conceptually speaking:

- Ordinary least squares (OLS) – simple summed differences between the observed and the model-implied matrices

- Generalized least squares (GLS) – the differences between the observed and the model-implied matrices are weighted by corresponding elements in the observed matrix (discrepancy in a larger element is penalized less than discrepancy in a smaller element)

Estimation

- Of course, the constrained parameters are not being estimated.