Jan Štochl, Ph.D.
Department of Psychiatry
University of Cambridge
Email: js883@cam.ac.uk

Confirmatory factor analysis in MPlus

The Psychometrics Centre

UNIVERSITY OF
CAMBRIDGE
The Psychometrics Centre

This course

The course is funded by the ESRC RDI and hosted by The Psychometrics Centre

E·S·R·C
ECONOMIC
& SOCIAL
RESEARCH
COUNCIL

UNIVERSITY OF
CAMBRIDGE
The Psychometrics Centre

RDI
Researcher
Development
Initiative

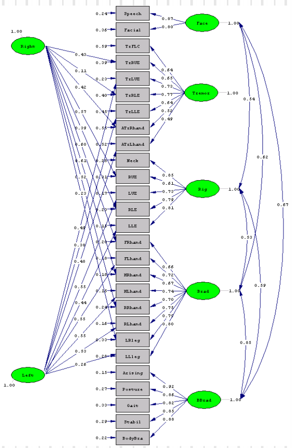
Tutors

Jan Stochl js883@cam.ac.uk

Anna Brown ab936@medschl.cam.ac.uk

University of Cambridge, Department of Psychiatry

http://www.psychiatry.cam.ac.uk



UNIVERSITY OF CAMBRIDGE

Agenda of day 1

General ideas and introduction to factor analysis

Introduction to Mplus, fitting basic CFA models

Introduction to „Goodness of fit“

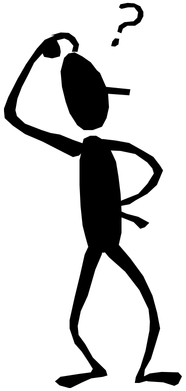
Differences between Principal Component Analysis, Exploratory Factor Analysis and Confirmatory Factor Analysis

General ideas and introduction to factor analysis

A bit of theory…….

What is factor analysis?

It is a statistical method used to find a small set of unobserved variables (also called latent variables, constructs or factors) which can account for the covariance (correlations) among a larger set of observed variables (also called manifest variables)



Factor analysis useful for:

- Assessment of dimensionality (i.e. how many latent variables underly your questionnaire, survey…)
- Assessment of validity of items in questionnaires and surveys, and therefore helps to eliminate less valid items
- Providing scores of respondent in latent variables
- Finding correlations among latent variables
- Answering specific scientific questions about relationship between observed and latent variables
- Helping to optimize length of questionnaires or surveys

And many others….

UNIVERSITY OF CAMBRIDGE

What is observed variable?

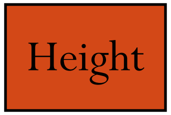
- Sometimes also called indicators, manifest variables
- It is something what we can measure directly on certain scale

Examples: Height, circumference of head, number of pushups, time necessary to answer question during intelligence testing, ….these are called continuous observed variables

Examples: Responses on disagree – agree scales, never – always scales, school grades,…. these are called categorical observed variables

Hereafter we denote observed variable as rectangular with its name inside,

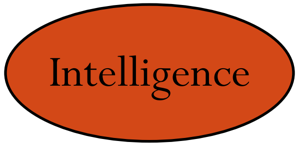
for example height will be denoted as


Height

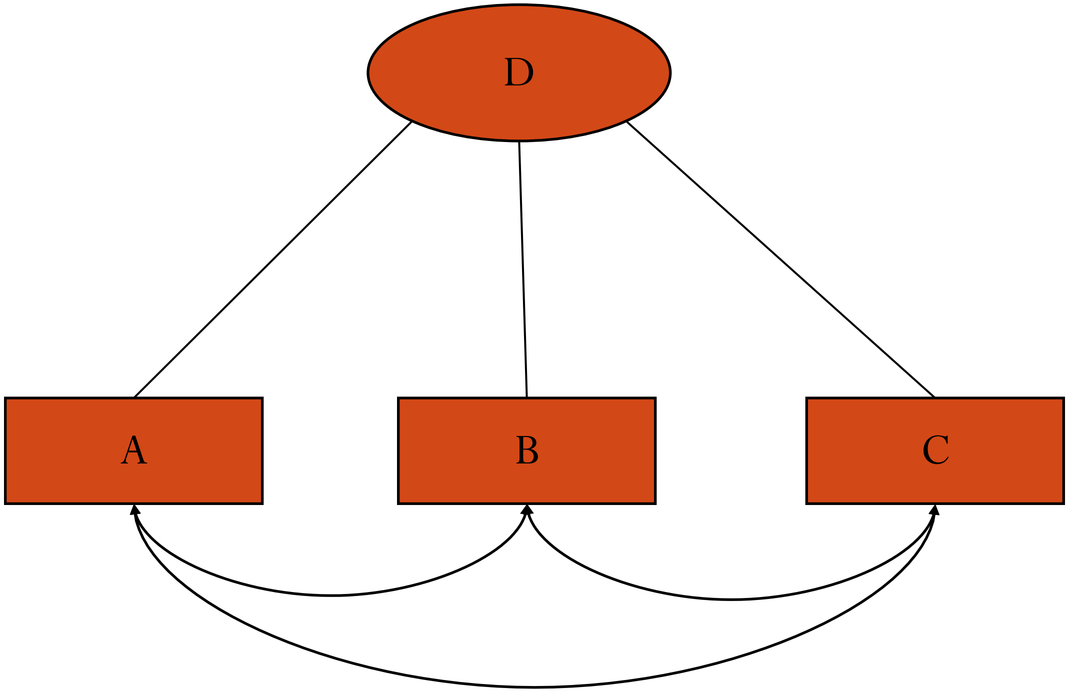
What is latent variable?

- Also called constructs (especially in psychology), factors, unobserved variables
- We cannot measure them directly
- We assume they are underlying abilities causing respondents to score high or low on observed variables

  Example: Testee performs poorly items of mathematical test (these items are observed variables) because his/her mathematical ability (latent variable) is low
- Examples of latent variables: Intelligence, well-being, mathematical ability, Parkinson′s disease….
- We denote latent variable as oval with its name inside, that is, e.g.


Intelligence

Principle of factor analysis



Covariances

- Describe bivariate relationships

$$\mathrm{cov}(X,Y)=\sum_{i=1}^{N}\frac{(x_i-\overline{x})(y_i-\overline{y})}{N}$$

- May range from -∞ to +∞ (in theory)
- Covariance equals 0 for unrelated variables
- Difficult to say how „strong“ is the relationship without knowing the variances

Covariance matrix

- Suppose we have *k* variables: $X_1, X_2, \ldots, X_k$

$$\text{covariance matrix} = \begin{bmatrix} \sum_{i=1}^{N} \frac{(x_{1i}-\bar{x}_1)(x_{1i}-\bar{x}_1)}{N} & \sum_{i=1}^{N} \frac{(x_{1i}-\bar{x}_1)(x_{2i}-\bar{x}_2)}{N} & \cdots & \sum_{i=1}^{N} \frac{(x_{1i}-\bar{x}_1)(x_{ki}-\bar{x}_k)}{N} \\ \sum_{i=1}^{N} \frac{(x_{2i}-\bar{x}_2)(x_{1i}-\bar{x}_1)}{N} & \ddots & & \vdots \\ \vdots & & & \\ \sum_{i=1}^{N} \frac{(x_{ki}-\bar{x}_k)(x_{1i}-\bar{x}_1)}{N} & \cdots & & \sum_{i=1}^{N} \frac{(x_{ki}-\bar{x}_k)(x_{ki}-\bar{x}_k)}{N} \end{bmatrix}$$

Some properties of covariances and example of covariance matrix

- For any constant *a*: $\text{cov}(X, a) = 0$
- $\text{cov}(X, X) = \text{var}(X)$
- Covariance is symmetrical, i.e. $\text{cov}(X, Y) = \text{cov}(Y, X)$

	X1	X2	X3	X4	X5	X6
X1	47.471					
X2	9.943	19.622				
X3	25.620	15.310	68.695			
X4	7.918	3.397	9.143	11.315		
X5	9.867	3.273	11.015	11.200	21.467	
X6	17.305	6.829	22.796	19.034	25.146	63.163

Correlation

- Standardized covariance, i.e.

$$\operatorname{cor}(X,Y)=\frac{\operatorname{cov}(X,Y)}{\sigma_X\sigma_Y}$$

- Covariance divided by product of standard deviations of variables

- Ranges from -1 to +1

- If correlation equals 0 then there is no (linear) relationship

- Allows easy comparison of how „strong“ is the relationship between 2 variables

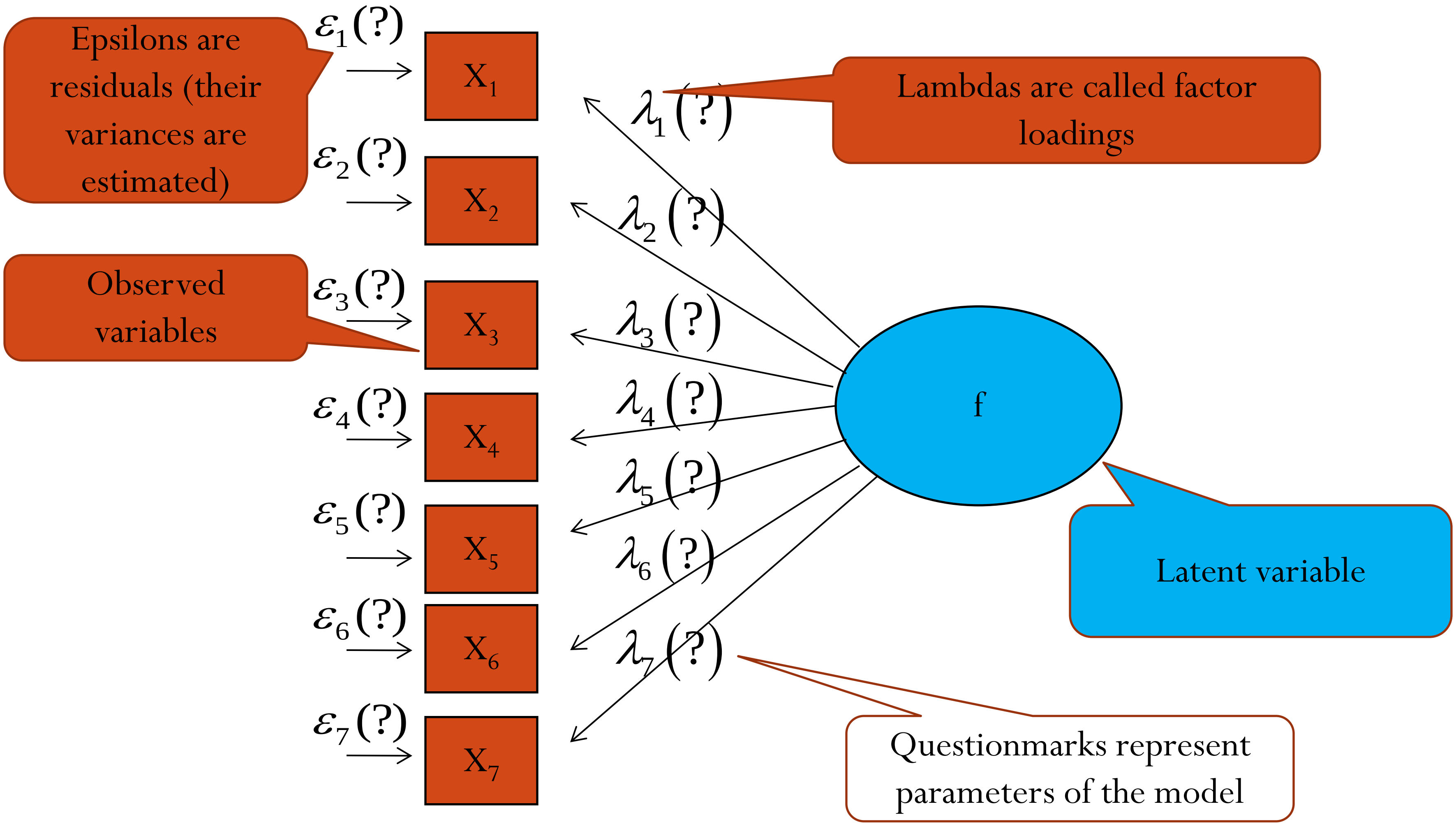
VisualizingCorrelations.nbp

UNIVERSITY OF CAMBRIDGE

Example of correlation matrix

	X1	*X2*	*X3*	*X4*	*X5*	*X6*	*X7*
X1	1						
X2	0.57	1					
X3	0.51	0.82	1				
X4	0.42	0.82	0.88	1			
X5	0.59	0.79	0.69	0.74	1		
X6	0.48	0.87	0.79	0.83	0.8	1	
X7	0.52	0.89	0.82	0.82	0.81	0.91	1

One factor is underlying this correlation matrix?



Epsilons are residuals (their variances are estimated)

Observed variables

Lambdas are called factor loadings

Latent variable

Questionmarks represent parameters of the model

Formally (for the 1-factor model)

$$x_i = \alpha_i + \lambda_i f + \varepsilon_i \qquad (i = 1, 2, ..., 7)$$

α_i represent mean of each observed variable. It vanishes if the variables are centered around mean

- Usuall regression assumptions apply

- Now we can choose scale and origin of *f* (it does not affect the form of the regression equation), so we choose mean *f* =0 and standard deviation *f*= 1, so now

$$\text{Cov}(X_i, X_k) = \lambda_i \lambda_k \qquad (i, k = 1, 2, ..., 7;\ i \neq k)$$

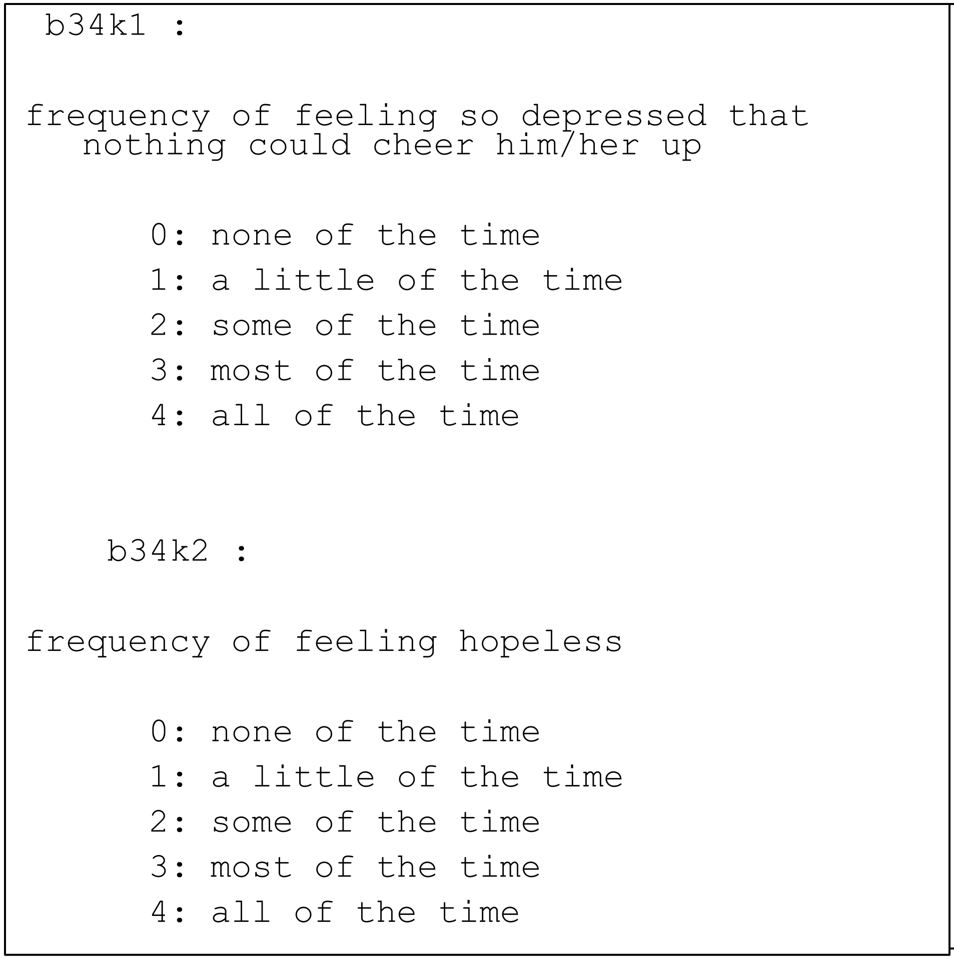
We can use this property to make deductions about λ_is and ε_is.

Practical 1: Fitting 1-factor model

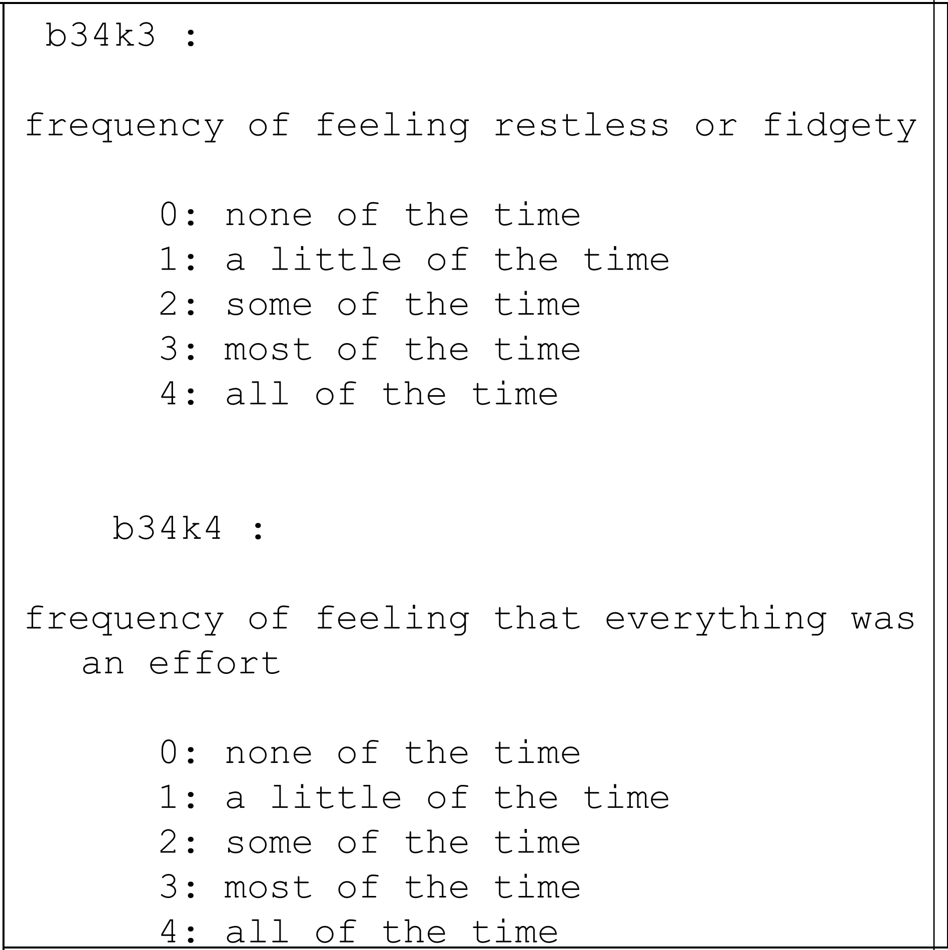
Data: Correlation matrix 4 by 4

	X1	X2	X3	X4
X1	1			
X2	0.821	1		
X3	0.588	0.622	1	
X4	0.661	0.695	0.661	1

Kessler items (K4)



```
 b34k1 :

frequency of feeling so depressed that
   nothing could cheer him/her up

      0: none of the time
      1: a little of the time
      2: some of the time
      3: most of the time
      4: all of the time


   b34k2 :

frequency of feeling hopeless

      0: none of the time
      1: a little of the time
      2: some of the time
      3: most of the time
      4: all of the time
```



```
 b34k3 :

frequency of feeling restless or fidgety

      0: none of the time
      1: a little of the time
      2: some of the time
      3: most of the time
      4: all of the time


   b34k4 :

frequency of feeling that everything was
   an effort

      0: none of the time
      1: a little of the time
      2: some of the time
      3: most of the time
      4: all of the time
```

Determining scale of latent variable

- 2 possibilities:

  - fix one loading for every factor to 1 (Mplus default). The latent variable then have the same scale as the item with this fixed loading.

  - fix the variance of latent variable to 1. The scale of the latent factor is then the same as for z-scores.

    Example: MODEL: f1 BY X1* X2 X3 X4; f1@1

- Selection of the method is somehow arbitrary and may depend on the research question.

- Selection of the method does not influence model fit

Which estimator should be used?

- 2 basic types of estimators: Maximum likelihood based or least squares based

- Default estimator depends on type of analysis and measurement level of observed variables

- Default estimator can be changed ANALYSIS command

  Example: ANALYSIS: Estimator = ML;

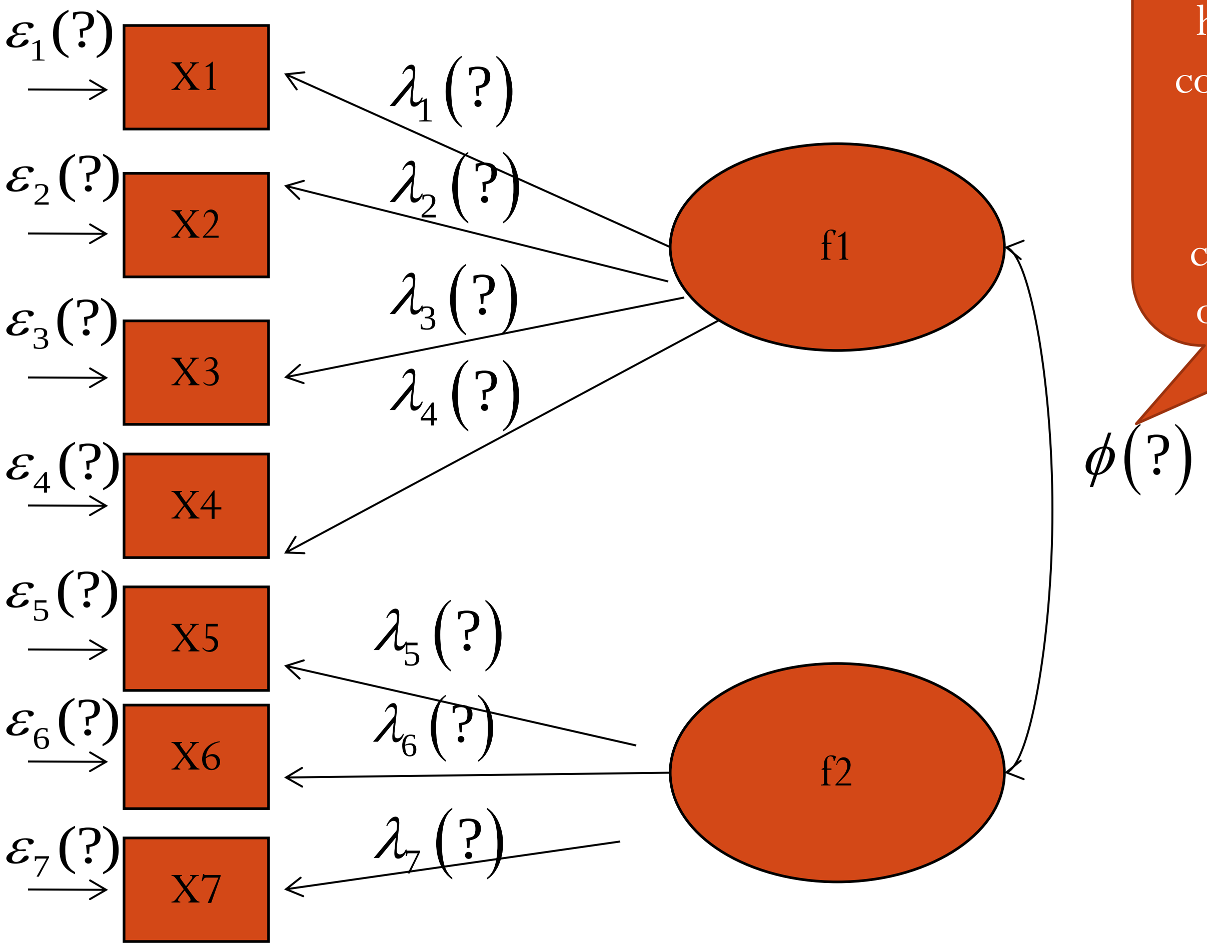
- Available estimators: ML, MLM, MLMV, MLR, MLF, MUML, WLS, WLSM, WLSMV, ULS, ULSMV, GLS

- More in estimator selection in Mplus user guide
  Muthén L. K., & Muthén, B. O. (1998-2010). *Mplus User's Guide. Sixth Edition. Los Angeles, CA. Muthén & Muthén.*

Example 2 of correlation matrix

	X1	X2	X3	X4	X5	X6	X7
X1	1						
X2	0.75	1					
X3	0.66	0.72	1				
X4	-0.73	-0.78	-0.31	1			
X5	0.06	0.12	0.03	0.14	1		
X6	0.11	0.17	-0.04	0.05	0.68	1	
X7	0.18	0.02	0.1	0.11	0.67	0.8	1

Two factors underlying



New parameter here representing correlation between factors (will be close to zero for correlation matrix on previous slide)

Practical 2: Fitting 2-factor model

grant data

grant.dat

grant_cfa_ml.inp

grant_cfa_ml.out

How the confirmatory factor analysis works in practice?

1. You provide Mplus (or other software) with correlation matrix (or covariance or raw data) from your sample
2. You specify your hypothesis about underlying structure (how many factors and which items load on which factor). This is how you create model.
3. Mplus will create correlation (or covariance) matrix that conforms to your hypothesis and at the same time maximizes likelihood of your data. Also factor loadings and residual variances are estimated.
4. Your sample correlation (or covariance) matrix (real sample correlation matrix) is compared to the correlation (covariance) matrix created by computer (artificial population correlation matrix which fits your hypothesis).
5. If the difference is small enough your data fits the model.

But what is small enough?

Sample correlation matrix and corresponding residual matrix

UNIVERSITY OF CAMBRIDGE

Sample correlation matrix

	X1	*X2*	*X3*	*X4*	*X5*	*X6*	*X7*
X1	1						
X2	0.57	1					
X3	0.51	0.82	1				
X4	0.42	0.82	0.88	1			
X5	0.59	0.79	0.69	0.74	1		
X6	0.48	0.87	0.79	0.83	0.8	1	
X7	0.52	0.89	0.82	0.82	0.81	0.91	1

Residual correlation matrix

	X1	**X2**	**X3**	**X4**	**X5**	**X6**	**X7**
X1	0.00						
X2	0.05	0.00					
X3	0.02	0.00	0.00				
X4	-0.08	-0.01	0.11	0.00			
X5	0.12	0.00	-0.05	-0.01	0.00		
X6	-0.04	-0.01	-0.03	0.00	0.01	0.00	
X7	-0.01	0.00	-0.01	-0.02	0.01	0.02	0.00

Goodness of fit

- Each software provides different goodness of fit statistics
- They are based on different ideas (e.g. summarizing elements in residual matrix, information theory, etc.)
- Some of them are known to favour certain types of model
- Fortunately Mplus provides only few of them and the ones that are known to provide good information about model fit

```
Degrees of Freedom = 14
Minimum Fit Function Chi-Square = 283.78 (P = 0.0)
Normal Theory Weighted Least Squares Chi-Square = 243.60 (P = 0.0)
Satorra-Bentler Scaled Chi-Square = 29.05 (P = 0.010)
Chi-Square Corrected for Non-Normality = 35.79 (P = 0.0011)
Estimated Non-centrality Parameter (NCP) = 15.05
90 Percent Confidence Interval for NCP = (3.33 ; 34.52)
Minimum Fit Function Value = 0.60
Population Discrepancy Function Value (F0) = 0.032
90 Percent Confidence Interval for F0 = (0.0071 ; 0.073)
Root Mean Square Error of Approximation (RMSEA) = 0.048
90 Percent Confidence Interval for RMSEA = (0.022 ; 0.072)
P-Value for Test of Close Fit (RMSEA < 0.05) = 0.52
Expected Cross-Validation Index (ECVI) = 0.12
90 Percent Confidence Interval for ECVI = (0.096 ; 0.16)
ECVI for Saturated Model = 0.12
ECVI for Independence Model = 11.74
Chi-Square for Independence Model with 21 Degrees of Freedom = 5514.61
 Independence AIC = 5528.61
Model AIC = 57.05
Saturated AIC = 56.00
Independence CAIC = 5564.71
Model CAIC = 129.25
Saturated CAIC = 200.40
Normed Fit Index (NFI) = 0.99
Non-Normed Fit Index (NNFI) = 1.00
Parsimony Normed Fit Index (PNFI) = 0.66
Comparative Fit Index (CFI) = 1.00
Incremental Fit Index (IFI) = 1.00
Relative Fit Index (RFI) = 0.99
Critical N (CN) = 473.52
Root Mean Square Residual (RMR) = 0.038
Standardized RMR = 0.038
Goodness of Fit Index (GFI) = 0.87
Adjusted Goodness of Fit Index (AGFI) = 0.74
Parsimony Goodness of Fit Index (PGFI) = 0.44
```

Goodness of fit
Chi-square and log-likelihood

- Testing hypothesis that the population correlation (covariance) matrix is equal to correlation (covariance) matrix estimated in Mplus.

- Chi-square is widely use as model fit, although it has certain undesired properties
  - Sensitive to sample size (the larger sample size the more likely is the rejection of the model)
  - Sensitive to model complexity (the more complex model the more likely is the rejection of the model)

- Log-likelihood value can be used to compare nested models (those models in which where the more constrained model has all parameters of less constraint one + applied to same data)

- -2 x loglikelihood follows chi-square distribution with df equal to difference in number of estimated parameters

Goodness of fit
TLI, CFI

- So-called comparative fit indices or incremental fit indices (measure improvement of fit)

- Compare your model with baseline model (model, where all observed variables are mutually uncorrelated)

- Tucker-Lewis index (TLI) – the higher the better (can exceed 1), at least 0.95 is recommended as cutoff value.

- TLI is underestimated for small sample sizes (say less than 100)′and has large sampling variability. Therefore CFI is preferred.

- Comparative Fit Index (CFI) – range 0-1, the higher the better, recommended cutoff also 0.95

Goodness of fit
Error of Approximation Indices

- Root-mean-square Error of Approximation (RMSEA)

- RMSEA has known distribution and therefore confidence intervals can be computed

- Recommended cutoffs: > 0.1 poor fit
  0.05-0.08 fair fit
  < 0.05 close fit

Goodness of fit
Residual Based Fit Indices

- Measure average differences between sample and estimated population covariance (correlation) matrix.

- Standardised root mean square residual (SRMR) – range 0-1, the smaller the better, recommended cutoff 0.08

Goodness of fit - recommendations

- Always check residual matrix – you can find source of poor fit
- Do not make your decisions on the basis of one fit index. (As Rod McDonald says „There is always at least one fit index that shows good fit of your model“).
- Chi-square based goodness of fit statistics tend to reject model when the sample size is big or when the model is complex. In that case use other statistics.

Degrees of freedom and identifiability of the model

- This adresses problem of model identification (will be covered in detail tomorrow)

- Model is underidentified (model parameters have infinite number of solutions) if the number of model parameters (*q*) exceeds $p(p+1)/2$, where *p* is the number of observed parameters, that is $p(p+1)/2-q<0$.

- Model is just identified if $p(p+1)/2-q=0$. Such model has just one solution. This model, however, cannot be statistically tested (fit is always perfect).

- We aim at overidentified models, that is $p(p+1)/2-q>0$.

UNIVERSITY OF CAMBRIDGE

Number of observed variables per latent variable

- For 1-factor model, 3 observed variables result in just identified model (0 degrees of freedom) 🛑, 4 observed variables result in model with 2 degrees of freedom ✓

- If your model involves many items the situation is more complicated – you can have only 2 items to represent factor but then question of domain coverage arises

- Have this in mind when you design your research tool (survey, questionnaire, …) – rather include more items

General recommendations for CFA analysis

- If the input sample correlation matrix consists of low correlations (say below 0.3), do not perform factor analysis. There is not much to model!
- Estimated only models that have substantial meaning
- Remember that parsimonous models are more appreciated
- Check standard errors of parameter estimates
- Check significance of factor loadings for model modification

Factor analysis useful for:

- Assessment of dimensionality (i.e. how many latent variables underly your questionnaire, survey…)
- Assessment of validity of items in questionnaires and surveys, and therefore helps to eliminate less valid items
- Providing scores of respondent in latent variables
- Finding correlations among latent variables
- Answering specific scientific questions about relationship between observed and latent variables
- Helping to optimize length of questionnaires or surveys

And many others….

Other types of factor analysis

Comparison of CFA, EFA and PCA

Types of factor analysis

- Principal component analysis (PCA) – sometimes considered as one type of factor analysis, but PCA is conceptually different from FA!!!
- Exploratory factor analysis (EFA) – data driven automated searching engine for finding underlying factors
- Confirmatory factor analysis (CFA) – theory driven, more parsimonous and scientifically more sound methodology for finding underlying factors

We did CFA until now!!

Principal component analysis (PCA)

- It is data reduction technique

- Reduces the number of observed variables to a smaller number of principal components which account for most of the variance of the observed variables

- Not model based, treat observed variables as measured without error, components cannot be interpreted as latent variables, not recommended for understanding latent structure of the data

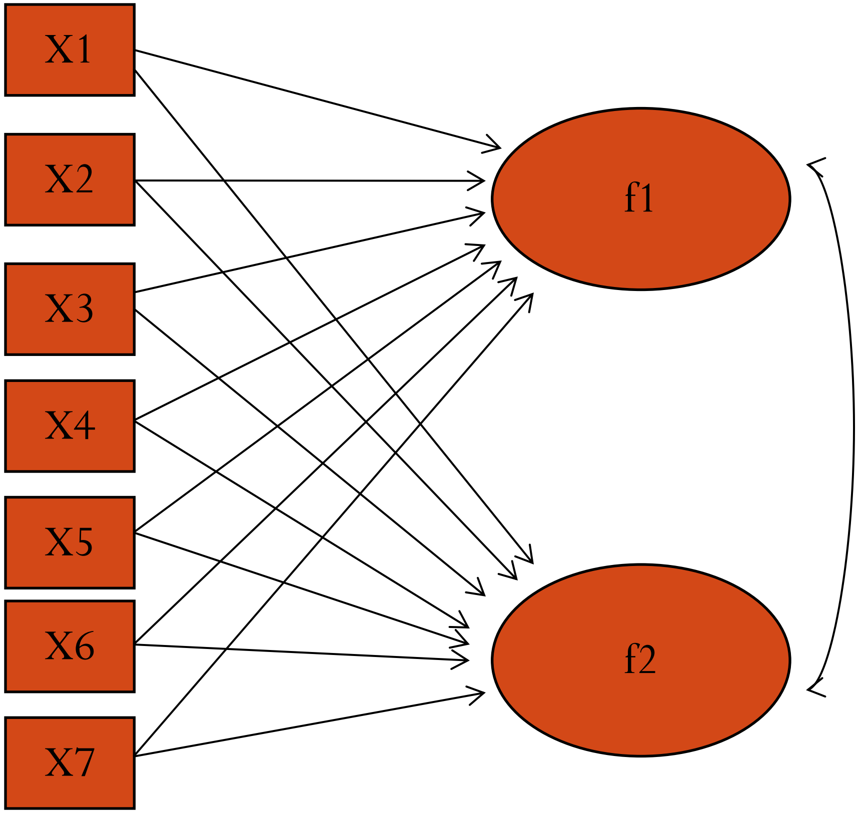
- Useful e.g. for treatment of collinearity in multiple regression

UNIVERSITY OF CAMBRIDGE

Principal component analysis (PCA)

- The direction of the effect is different from EFA and CFA

- Unique variances are missing (thus it does not account for measurement error)

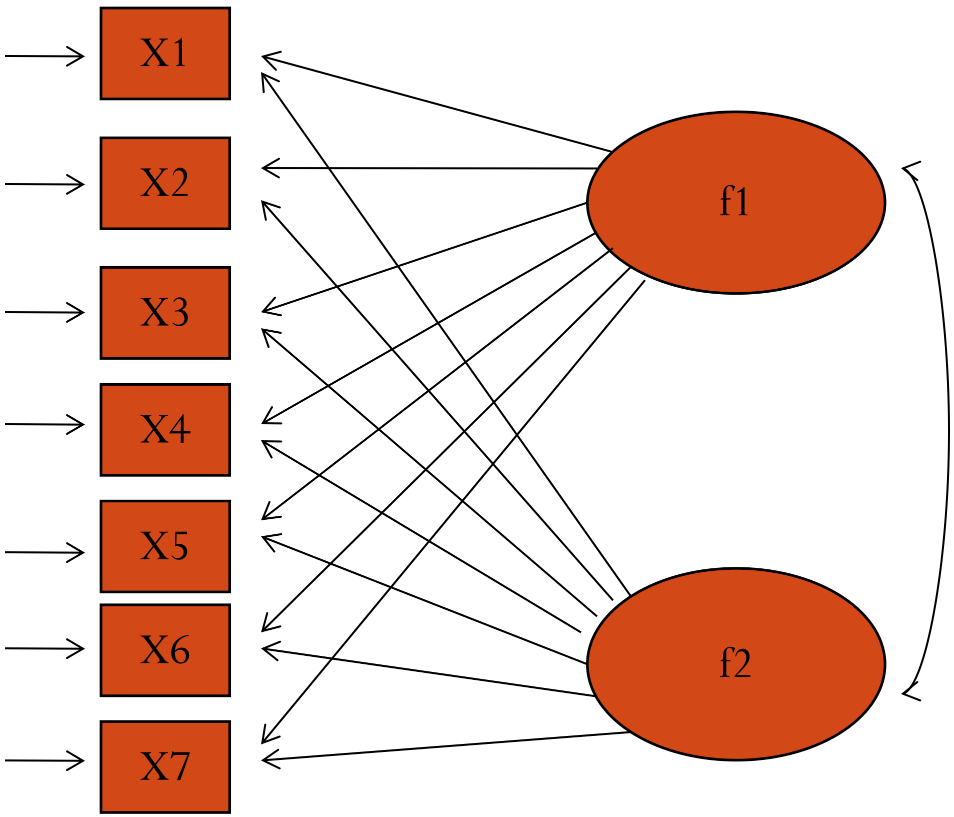


Exploratory factor analysis (EFA)

- Is a statistical technique which identifies the number of latent variables and the underlying factor structure of a set of observed variables
- It is model based (advantage over PCA), „automated“ searching procedure for underlying structure, post-hoc interpretations of latent variables (disadvantage comparing to CFA)
- Traditionally has been used to explore the possible underlying factor structure of a set of measured variables without imposing any preconceived structure on the outcome

Exploratory factor analysis (EFA)

- EFA models are generally underidentified (infinite number of solutions)
- We therefore „rotate" the solution (actually we rotate coordinates) according to some criteria (e.g. until the variance of squared loadings is maximal – in Varimax rotation)
- Number of factors is determined using eigenvalues (usually number of factors=number of eigenvalues over 1)



Exploratory factor analysis (EFA)

- Correlations between factors can be controlled by rotation method (orthogonal versus oblique)
- Usually rotation is necessary to interpret factor loadings.
- Traditionally has been used to explore the possible underlying factor structure of a set of measured variables without imposing any preconceived structure on the outcome

Similarities and differences between PCA and EFA

- PCA and EFA may look similar and in practice may look like giving similar results. But the principal components (from PCA analysis) and factors (from EFA analysis) have very different interpretations

- Use EFA when you are interested in making statements about the factors that are responsible for a set of observed responses

- Use PCA when you are simply interested in performing data reduction.

UNIVERSITY OF CAMBRIDGE

Further reading

General literature on factor analysis

- McDonald, R. P. (1999). *Test theory: A unified treatment. Mahwah: Lawrence Erlbaum Associates, Inc.*
- Bartholomew, D. J., et al. (2008). *Analysis of Multivariate Social Science Data. London: CRC Press.*
- Dunteman, G. H. (1989). *Principal component analysis. Newbury Park: Sage.*
- Bollen, K. A. (1989). *Structural equations with latent variables. New York: John Wiley&sons.*
- Kaplan, D. (2000). *Structural equation modeling: Foundations and extensions. Thousand Oaks: Sage Publications.*
- Maruyama, G. M. (1998). *Basics of Structural Equation Modeling. Thousand Oaks: Sage Publications.*
- McDonald, R. P. (1985). *Factor analysis and related methods. Hillsdale NJ: Lawrence Erlbaum Associates.*

MPlus

- Muthén, L. K., & Muthén, B. O. (1998-2010). *Mplus User's Guide. Sixth Edition. Los Angeles, CA: Muthén & Muthén.*
- *Visit www.statmodel.com for many papers and discussion on Mplus*

Goodness of fit

- Hu, L., & Bentler, M. P. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal, 6(1), 1-55.*

Exercise 1

Decathlon data

1. Open data file „decathlon.xls“. Prepare dataset for analysis in Mplus (in Excel, Notepad or N2Mplus)
2. Fit 1-factor model for items 1-5. Use ML estimator and subsequently change it to WLS
3. Fit 1-factor model for items 6-10. Change the default scaling of latent variable and compare factor validities of items
4. Compare fit indices of 1-factor model for all items and 2-factor model with uncorrelated factors as specified
5. Compare 2-factor model with uncorrelated factors with the one with correlated factor. Which of the models would you prefer?
6. Perform exploratory factor analysis with 1 to 2 factors. Use some ortogonal and oblique rotation. Which solution would you prefer?
7. Try to estimate other models that are theoretically meaningful.

Exercise 2

Prediction of side of onset from premorbid handedness in Parkinson‘s disease

1. Open data file „Handedness.xls“. Data consist of 7 items measuring premorbid handedness of patients with PD + 1 item measuring side of PD onset (onsetside). All items are categorical (onsetside has 3 categories, other items 5 categories)
2. Prepare dataset for analysis in Mplus (in Excel, Notepad or N2Mplus)
3. Look at correlation matrix and assess how much is side of onset is correlated with other items.
4. Introduce factor handedness that loads to all items. Can you predict side of PD onset from premorbid handedness?
5. Exclude item „onsetside“ from model. Change the scale of factor (fix the variance of factor „handedness“ to 1) and asees the validities of items. Which is the least valid item?