Lecture 10 - Latent variable models

Björn Andersson

University of Oslo

November 22 2021

The single-factor model

- We have so far talked mostly about the single factor model:

$$X_j = \mu_j + \lambda_j F + \epsilon_j$$

- The central assumption of this model is that a single factor can explain the variance of the item score
- Corresponds to a single attribute or construct - e.g. a mathematics test and mathematics ability
- If a single factor model holds we can say that the test is measuring only a single attribute - the test is unidimensional (homogenous)

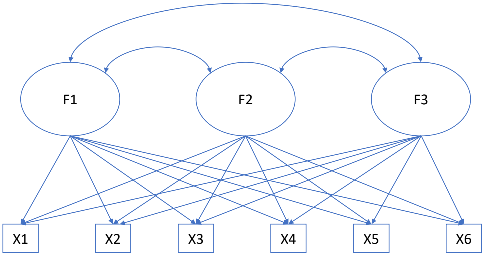
Today

- We will go through the general multiple-factor model
- Illustrate some special cases of the model that are common
- How to use the models and how to interpret them
- A few things about model comparisons and model fit
- A sketch of more complex models

Multiple factor model

- We can extend the single factor model to have additional factors
- These may correspond to different aspects of an attribute or to separate attributes
  - A mathematics test measuring algebra, calculus and geometry
  - A mathematics test measuring both reading comprehension and mathematics ability
- We can also imagine models where there is one general factor and several ”subfactors”, that are measured by clusters of items
  - An intelligence test that measures general intelligence along with specific subdomains connected to specific items
- The models can be very complex and we need to consider if the models we can think of are actually possible to apply (identifiability constraints)

Graphical representation of factor models



Confirmatory factor analysis

- In confirmatory factor analysis, a hypothesis regarding a certain factor structure in the data is to be tested.
- The number of factors and the relations between the factors and items are hypothesized before collecting the data.
- In this course we will focus on confirmatory models.
- Can you think of an example of a confirmatory model?

Exploratory factor analysis

- With exploratory factor analysis (EFA), the number of factors and their relationship to each item is not specified beforehand.
- The object is to find the number of factors and the factor structure which underlies the observed data.
- Hypothesis generating procedure.

Factor analysis with item and subtest scores

- Like for the single factor model, we will consider the case of item scores, subtest scores or scores from multiple different tests
- Strictly speaking, the model we will use applies only when there is a linear relationship between the factor and the item score
- The linearity assumption is questionable with binary and ordinal data
    - At this point we will just note that there are factor models which take the ordinality into account - we will discuss this in more detail in the item response theory (IRT) course early next semester (factor analysis with ordinal data is basically equivalent to a type of IRT model)
- We will proceed with assuming that the regular factor model is OK to use

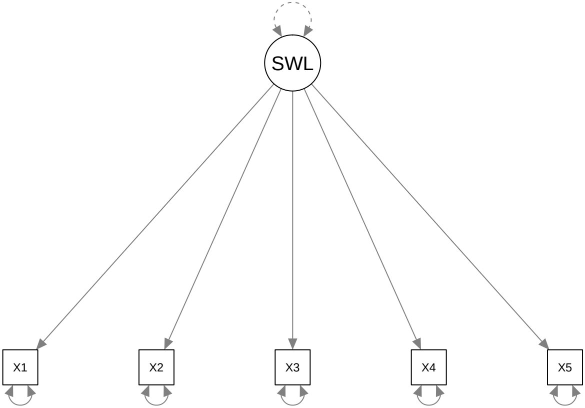
Satisfaction With Life Scale

- Recall the satisfaction with life scale that had five items
- A single factor model for the items on the scale is then

$$X_j = \mu_j + \lambda_j F + \epsilon_j$$

- We have so far specified that F has mean 0 and variance 1 in the population
- The unknown parameters are μ_j, the difficulty of the item, λ_j, the factor loading, and Ψ_j^2, the variance of the error term.
- We assume that the factor F is uncorrelated with the error term ϵ_j and that the error terms of different items are uncorrelated (But note that we **can** add residual correlations between items - needs motivation from a substantive perspective.)

SWLS single-factor model graphically



Alternate model for SWLS

- To illustrate some alternative models we can consider the SWLS again
- The subject matter of the items indicates that two items pertain to past satisfaction while three items pertain to current satisfaction
- We can imagine these two clusters of items as measuring two separate attributes F_1 and F_2
- We call such a model an independent-cluster model (sometimes referred to as a simple structure model)

Alternate model for SWLS

- The model for items $j \in \{1, 2, 3\}$ is then:

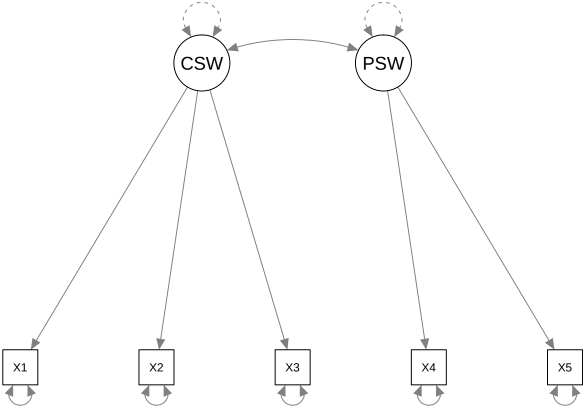
$$X_j = \mu_j + \lambda_{j1} F_1 + (0 \times F_2) + \epsilon_j$$

  and the model for items $k \in \{4, 5\}$ is

$$X_k = \mu_k + (0 \times F_1) + \lambda_{k2} F_2 + \epsilon_k$$

- We also assume that the two factors are correlated, i.e. $\text{Cov}(F_1, F_2) \neq 0$.
- Question: How many parameters do we estimate compared to the single factor model?

SWLS two-factor model graphically



The general case

- We can specify a fully general model where each item is measuring r number of attributes
- We thus have a general model:

$$X_j = \mu_j + \lambda_{j1} F_1 + \cdots + \lambda_{jr} F_r + \epsilon_j$$

- As with the other factor models, ϵ_j is uncorrelated with the factors and with other error terms
- The factors $F_1, \ldots, F_r$ can be correlated or uncorrelated
- This model may be true for all items - but it is not possible to estimate such a model
- We need to add constraints to make all the unknown parameters identified.

Parameter identification

- A general identification rule is highly complex to describe
- For independent cluster items we have the following rules:
  - The model parameters are identified if there are at least three items for each factor
  - If we assume that the factors are correlated, we need only two items for each correlated factor
- For further discussion, see Chapter 9 in the textbook.

Parameter identification

- We can distinguish between factorially simple and factorially complex items
- Simple: The item indicates only a single factor. Measures only one attribute.
- Complex: The item indicates two or more factors. Measures more than one attribute.

Correlations and covariances

- We have so far used the covariances to estimate the models
- Using the correlations doesn't change the model fit
- In fact, the obtained parameter estimates from using the covariance and correlation matrices are possible to directly transform between each other
- Factor loading from correlation matrix = factor loading from the covariance matrix / standard deviation of the observed variable, i.e.

$$\hat{\lambda}_j^{\text{Cor}} = \frac{\hat{\lambda}_j^{\text{Cov}}}{\hat{\sigma}_{X_j}}$$

Estimation methods

- In the previous lectures with the single factor model, we used the ULS estimator
- Typically a technique called maximum likelihood is used with factor analysis, particularly with more complex models
- The maximum likelihood estimator is often statistically efficient, meaning that there is not a better estimator as the sample size tends to infinity
- For now, we will not discuss the estimation methods in detail
- We can note that there are several different approaches and that different methods can give slightly different results for given data

Model selection and model fit

- If we have different models that we want to compare, we can use hypothesis testing to identify if there are statistically significant differences between two models
- When using maximum likelihood, we can use a likelihood ratio test between nested models or we can use an information criterion such as the Bayesian Information Criterion - BIC (works also when models are not nested)
- There are several ways to evaluate model fit
  - Goodness of fit index (GFI)
  - Root Mean Square Error of Approximation (RMSEA)
  - Differences between the observed and model-implied covariance/correlation matrices - i.e. discrepancies (another word is residuals)

Guidelines for model fit

- GFI - higher than 0.90 is ”acceptable”, higher than 0.95 is ”good”
- RMSEA - less than 0.05 is a ”good” fit, but 0.06 has also been suggested in simulation studies (Hu and Bentler, 1999)

  Reference: Hu, L.-t., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling, 6(1)*, 1-55. `https://doi.org/10.1080/10705519909540118`
- Mean residuals - should be 0
- Note that there is not a consensus on which model fit index is the best

SWLS single-factor model results

We fitted the single-factor model with ULS earlier and now we use the ML estimator instead.

	Item 1	Item 2	Item 3	Item 4	Item 5
λ_j	1.308	1.133	1.141	0.922	1.148
Ψ_j^2	0.843	1.198	1.150	1.907	2.023

The GFI was 0.981, mean residuals <0.01 and the RMSEA was 0.070.

Alternate model for SWLS

- Fitting this model for the SWLS gives the following results:

Item	CSWL	PSWL	Ψ^2
1	1.323	0	0.803
2	1.143	0	1.174
3	1.138	0	1.156
4	0	1.022	1.712
5	0	1.289	1.680

  with Cor(CSWL, PSWL) $=$ 0.861.
- We can see that the factor loadings are largely the same as before
- The correlation between the factors is high, indicating that the factors may not be distinct
- The GFI was 0.990, mean residuals $<$0.01 and the RMSEA was 0.041.

Three factor model for SWLS

- We can also imagine that each item is measuring a general factor and that there are clusters of items that measure specific factors in addition to the general factor
- An additional assumption typically made is that the specific factors are uncorrelated with the general factor and that they are uncorrelated with each other (the second assumption is strictly speaking not necessary)
- This type of model is sometimes referred to as a bifactor model
- Question: How do we graphically represent this model?

Evaluating the single factor assumption

- Speaking more generally, consider a case of a general factor along with multiple subfactors that can be interpreted substantively
    - We can have general mathematics ability and specific abilities in algebra and calculus, for example
- Often it is desirable to have a single factor model - but is it justified?
- One way to evaluate a single factor assumption is to fit a more complex model and see how much of the explained variance that is accounted for by the general factor
- If the general factor is dominating, the application of a single factor model can be considered appropriate

Evaluating the single factor assumption - some guidelines

- If the standardized factor loadings of the specific factor are lower than 0.3, this is an indication that the specific factor does not have a major impact
- The product of two standardized factor loadings is the amount of correlation they ”explain”
  - Two loadings less than 0.3 => ”explains” less than 0.1 correlation
- We can look at how much of the common variance is explained by the general factor - higher than 0.7 means a high level of unidimensionality

Classical Test Theory and multiple factors

Consider classical test theory, but instead of a single latent factor we have two latent factors, one of which is not directly related to the factor we want to measure.

$$X = T + \xi + E.$$

What happens?

Classical Test Theory and multiple factors

More specifically, let's say we have two items and two factors F_1 (denoting the factor we want to measure) and F_2 (denoting a factor we do not want to measure) and our model is

$$X_1 = 0.5F_1 + 0.8F_2 + \sqrt{1 - 0.89}\epsilon_1$$
$$X_2 = 0.8F_1 + 0.4F_2 + \sqrt{1 - 0.8}\epsilon_2$$

where $F_1, F_2, \epsilon_1, \epsilon_2$ are $N(0, 1)$ and independent. This implies that $\sigma^2_{X_1} = \sigma^2_{X_2} = 1$.
Note that this model is not identified, so we can not estimate the parameters. However, we can still see what happens to the reliability and to the true score correlation.

Classical Test Theory and multiple factors

The reliabilities of item X_1 and X_2 are

$$\rho_{X_1,X_1'} = \frac{\mathrm{Cov}(X_1, X_1')}{\mathrm{Var}(X_1)} = \mathrm{Cov}(0.5F_1, 0.5F_1) + \mathrm{Cov}(0.8F_2, 0.8F_2)$$
$$= 0.25 + 0.64 = 0.89.$$

$$\rho_{X_2,X_2'} = \frac{\mathrm{Cov}(X_2, X_2')}{\mathrm{Var}(X_2)} = \mathrm{Cov}(0.8F_1, 0.8F_1) + \mathrm{Cov}(0.4F_2, 0.4F_2)$$
$$= 0.64 + 0.16 = 0.8.$$

So item X_1 is more reliable.

Classical Test Theory and multiple factors

However, the true score correlations of X_1 and X_2 with F_1, equal to the factor loadings for F_1, are

$$\rho_{X_1,F_1} = 0.5$$

and

$$\rho_{X_2,F_1} = 0.8.$$

So X_2 is a "better" item for measuring F_1, even though it has lower reliability.

Restricted factor analysis

- A multiple-factor model with restrictions such that some factor loadings are set to zero is called a restricted factor model
- A synonym is a confirmatory factor model
- We need to be mindful of the identification constraints necessary even with restricted models

Unrestricted factor analysis

- If we have no pre-specified factor structure, we can do an unrestricted factor analysis with a pre-specified number of factors
- Such a factor model is however not uniquely identified and we need to choose something called a factor rotation for identification
- Generally speaking, this is an exploratory approach and should not be used to draw definite conclusions about the factor structure of set of items
- We can use this as a hypothesis generating procedure

Adding covariates to the model

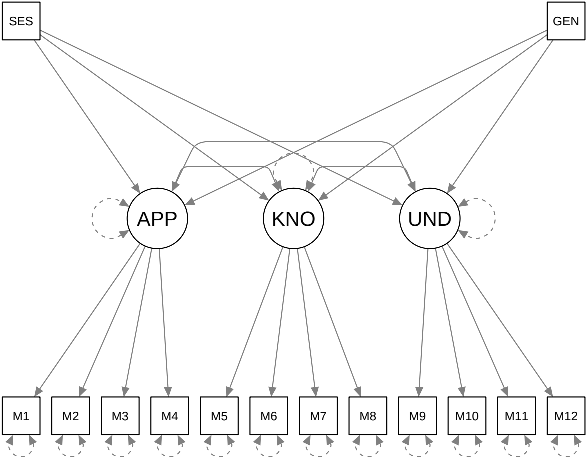
- We can extend the factor model by including a column vector of fixed covariates **c**.
- We obtain a regression model, which relates the covariates to the mean of the factor:

$$X_j = \mu_j + \lambda_j F + \epsilon_j,$$

where the mean of F is $\boldsymbol{\beta}'\mathbf{c}$

- We interpret the regression coefficients just as for a regular linear regression model: a regression coefficient β means that a one unit increase in the covariate c is associated with a β increase in the mean of the factor F
- This allows us to infer relationships between the factor and demographic variables such as gender and socio-economic status

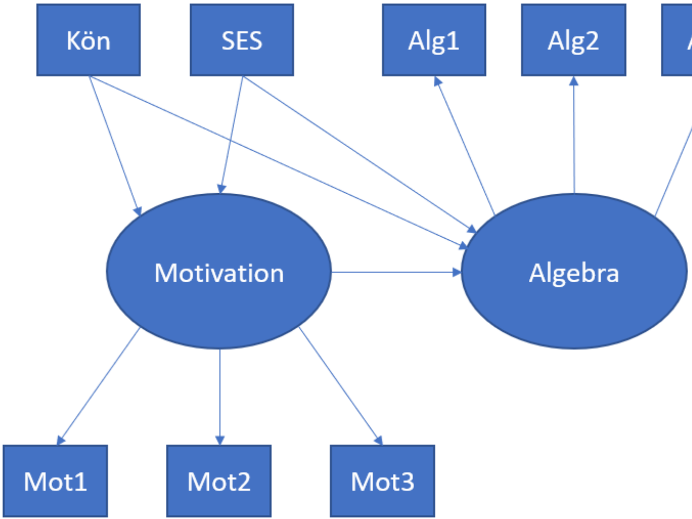
Models with covariates graphically



Structural equation models

- We can also consider an even more complex model with multiple latent variables that are related through regression equations.
- We can include both latent variables and covariates in the same model
- Hence we have three types of variables:
  - Item responses - observed indicators of a latent variable
  - Covariates - observed variables not viewed as having measurement error
  - Latent variables - representations of attributes, that are measured by the item responses

Structural equation models



Computer implementation

- We will be using the R package `lavaan` - `http://lavaan.ugent.be/`
- `lavaan` can estimate all the models we have talked about so far in the course
- During the lab we will go through the basic features of `lavaan`
- To make neat model plots from `lavaan` objects you can use the package `semPlot` (not required)