Topic 4: Overview of structural equation models

Structural equation modelling

Factor analysis (FA) is only one example of a new approach to data analysis which is **not based on individual observations**. We were not able to use the regression approach since the input factors were **latent** (not observable). There were too many unknowns. We went to analyse the covariance matrix Σ (and its estimator $\boldsymbol{S}$) which involved the actual parameters of interest—σ_i^2 and Λ. That is, we switched **from the level of individual observations** to analyse covariance matrices instead. There are a **series** of methods which are based on analysis of **covariances** rather than individual cases. Instead of minimising functions of observed and predicted **individual values**, we minimise the differences between **sample covariances and covariances predicted by the model**.

The fundamental hypothesis in these analyses is

$$H_0 : \Sigma = \Sigma(\boldsymbol{\theta}) \quad \text{against} \quad H_1 : \Sigma \neq \Sigma(\boldsymbol{\theta}).$$

Here Σ has $p(p+1)/2$ unknown elements (estimated by $\boldsymbol{S}$) **but these are assumed to be reproducible by just** $k = \dim(\boldsymbol{\theta}) < p(p+1)/2$ **parameters**. Note that more generally we could consider fitting **means and covariances, or means and covariances and higher moments** to a given structure. **Regression analysis with random inputs, simultaneous equations systems, confirmatory factor analysis, canonical correlations, (M)ANOVA** can be considered special cases.

Structural equation modelling is an important statistical tool in economics and behavioural sciences. Structural equations express relationships among several variables that can be either directly observed variables (manifest variables) or unobserved hypothetical variables (latent variables). In **structural models**, as opposed to **functional models**, all variables are taken to be **random** rather than having fixed levels. In addition, for maximum likelihood estimation and generalised least squares estimation (see next slide), the random variables are assumed to have an approximately multivariate normal distribution. Hence you are advised to remove outliers and consider transformations to normality before fitting.

General form of the model

$$\boldsymbol{\eta} = B\boldsymbol{\eta} + \Gamma\boldsymbol{\xi} + \boldsymbol{\zeta}. \tag{4.5}$$

Here,

$\boldsymbol{\eta} \in \mathbb{R}^m$ vector of output latent variables;
$\boldsymbol{\xi} \in \mathbb{R}^{n'}$ vector of input latent variables;
$B \in \mathcal{M}_{m,m}$, $\Gamma \in \mathcal{M}_{m,n'}$ coefficient matrices;

Note: $(I - B)$ is assumed to be nonsingular.

$\boldsymbol{\zeta} \in \mathbb{R}^m$ disturbance vector with $E\boldsymbol{\zeta} = 0$.
To this **modelling equation** (4.5) we attach two **measurement equations**:

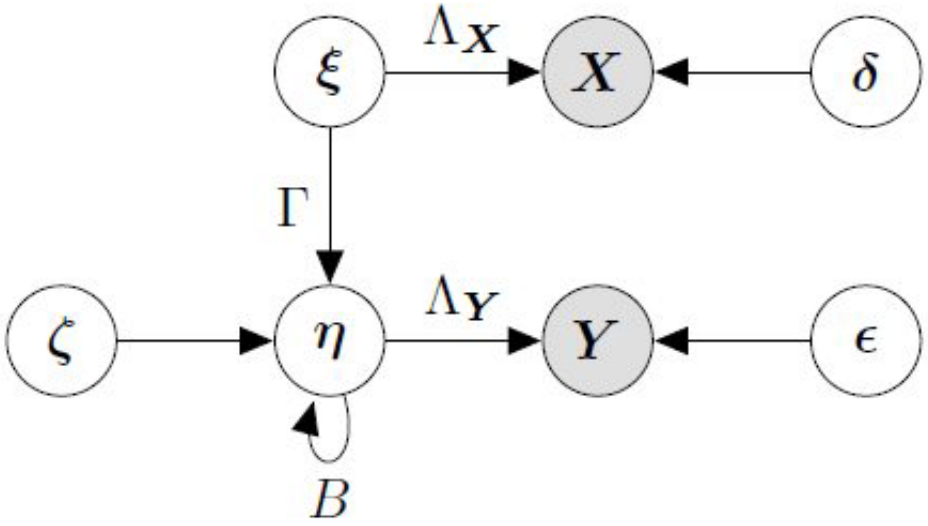
$$\boldsymbol{Y} = \Lambda_{\boldsymbol{Y}}\boldsymbol{\eta} + \boldsymbol{\epsilon}; \tag{4.6}$$

$$\boldsymbol{X} = \Lambda_{\boldsymbol{X}}\boldsymbol{\xi} + \boldsymbol{\delta}; \tag{4.7}$$

$$\boldsymbol{Y} \in \mathbb{R}^p,\ \boldsymbol{X} \in \mathbb{R}^q;\ \Lambda_{\boldsymbol{Y}} \in m_{p\times m}, \Lambda_{\boldsymbol{X}} \in m_{q\times n'}$$

with $\boldsymbol{\epsilon} \in \mathbb{R}^p,\ \boldsymbol{\delta} \in \mathbb{R}^q$ zero-mean measurement errors. These errors are assumed to be uncorrelated with $\boldsymbol{\xi}$ and $\boldsymbol{\zeta}$ and with each other.

Generative model for X and Y



The above quite general model (4.5)–(4.6)–(4.7) is called **Keesling–Wiley–Jöreskog** model. Its interpretation is that the input and output latent variables $\boldsymbol{\xi}$ and $\boldsymbol{\eta}$ are connected by a system of linear equations (the structural model (4.5) with coefficient matrices B and Γ and an error vector $\boldsymbol{\zeta}$. The random vectors $\boldsymbol{Y}$ and $\boldsymbol{X}$ represent the observable vectors (measurements).

The implied covariance matrix for this model can be obtained. Let

$$\mathrm{Var}(\boldsymbol{\xi}) = \Phi; \mathrm{Var}(\boldsymbol{\zeta}) = \Psi; \mathrm{Var}(\boldsymbol{\epsilon}) = \boldsymbol{\theta}_{\epsilon}; \mathrm{Var}(\boldsymbol{\delta}) = \boldsymbol{\theta}_{\boldsymbol{\delta}}.$$

Then,

$$\begin{aligned}\Sigma = \Sigma(\boldsymbol{\theta}) &= \begin{pmatrix} \sum_{\boldsymbol{YY}}(\boldsymbol{\theta}) & \Sigma_{\boldsymbol{YX}}(\boldsymbol{\theta}) \\ \Sigma_{\boldsymbol{XY}}(\boldsymbol{\theta}) & \Sigma_{\boldsymbol{XX}}(\boldsymbol{\theta}) \end{pmatrix} \\ &= \begin{pmatrix} \Lambda_{\boldsymbol{Y}}(I-B)^{-1}\left(\Gamma\Phi\Gamma^{\top}+\Psi\right)\left[(I-B)^{-1}\right]^{\top}\Lambda_{\boldsymbol{Y}}^{\top} + \boldsymbol{\theta}_{\boldsymbol{\epsilon}} & \Lambda_{\boldsymbol{Y}}(I-B)^{-1}\Gamma\Phi_{\boldsymbol{X}}^{\top} \\ \Lambda_{\boldsymbol{X}}\Phi\Gamma^{\top}\left[(I-B)^{-1}\right]^{\top}\Lambda_{\boldsymbol{Y}}^{\top} & \Lambda_{\boldsymbol{X}}\Phi\Lambda_{\boldsymbol{X}}^{\top} + \boldsymbol{\theta}_{\delta} \end{pmatrix}. \end{aligned} \tag{4.8}$$

Estimation

Under the normality assumption, we can use the MLE. Since the "data" is the estimated covariance matrix $\boldsymbol{S} = \frac{1}{n-1}\sum_{i=1}^{n}\left\{ \begin{pmatrix} \boldsymbol{Y}_i - \hat{\boldsymbol{Y}} \\ \boldsymbol{X}_i - \hat{\boldsymbol{X}} \end{pmatrix} \begin{pmatrix} \boldsymbol{Y}_i - \hat{\boldsymbol{Y}} \\ \boldsymbol{X}_i - \hat{\boldsymbol{X}} \end{pmatrix}^{\top} \right\}$, and since it is known that $(n-1)\boldsymbol{S} \sim W_{p+q}(n-1, \Sigma)$, we can utilise the form of the Wishart density to derive that

$$\log L(\boldsymbol{S}, \Sigma(\boldsymbol{\theta})) = \text{constant} - \frac{n-1}{2}\{\log|\Sigma(\boldsymbol{\theta})| + \text{tr}[\boldsymbol{S}\Sigma^{-1}(\boldsymbol{\theta})]\}$$

This is the function that has to be maximised. Hence, to find MLE, we minimise

$$F_{\text{ML}}(\boldsymbol{\theta}) = \log|\Sigma(\boldsymbol{\theta})| + \text{tr}[\boldsymbol{S}\Sigma^{-1}(\boldsymbol{\theta})] - \log|\boldsymbol{S}| - (p+q). \tag{4.9}$$

The function (4.9) has the advantage that F_{ML} would be zero for the "saturated model" (with $\hat{\Sigma} = \boldsymbol{S}$). I.e., a perfect fit is indicated by zero (and any non-perfect fit gives rise to > 0 value of F_{ML}).

Model evaluation

Under normality, model adequacy is mostly tested by an asymptotic χ^2-test.

Under $H_0 : \Sigma = \Sigma(\boldsymbol{\theta})$ versus $H_1 : \Sigma \neq \Sigma(\boldsymbol{\theta})$, the statistic to be used is $T = (n-1)F_{\mathrm{ML}}(\hat{\boldsymbol{\theta}}_{\mathrm{ML}})$ and under H_0, its asymptotic distribution is χ^2 with $\mathrm{df} = \frac{(p+q)(p+q+1)}{2} - \dim(\boldsymbol{\theta})$.

Reason:

$$\begin{aligned}\log L_0 =& \log L(\boldsymbol{S}, \hat{\Sigma}_{\mathrm{MLE}}) = \log L(\boldsymbol{S}, \Sigma(\hat{\theta}_{\mathrm{ML}})) \\ =& -\frac{n-1}{2}\left\{\log|\hat{\Sigma}_{\mathrm{MLE}}| + \mathrm{tr}[\boldsymbol{S}\hat{\Sigma}_{\mathrm{MLE}}^{-1}]\right\} + \mathrm{constant};\end{aligned}$$

$$\log L_1 = \log L(\boldsymbol{S}, \boldsymbol{S}) = -\frac{n-1}{2}\left\{\log|\boldsymbol{S}| + (p+q)\right\} + \mathrm{constant}.$$

Then,

$$\begin{aligned}-2\log\frac{L_0}{L_1} =& (n-1)\left\{\log|\hat{\Sigma}_{\mathrm{MLE}}| + \mathrm{tr}(\boldsymbol{S}\hat{\Sigma}_{\mathrm{MLE}}^{-1}) - \log|\boldsymbol{S}| - (p+q)\right\} \\ =& (n-1)F_{\mathrm{ML}}(\hat{\boldsymbol{\theta}}_{\mathrm{ML}}).\end{aligned}$$

Some particular SEM

From the general model (4.5)–(4.6)–(4.7), we can obtain the following particular models:

A) $\Lambda_{\boldsymbol{Y}} = I_m$, $\Lambda_{\boldsymbol{X}} = I_{n'}$; $p = m$; $q = n'$; $\boldsymbol{\theta_\epsilon} = 0$; $\boldsymbol{\theta_\delta} = 0 \implies \boldsymbol{Y} = B\boldsymbol{Y} + \Gamma\boldsymbol{X} + \boldsymbol{\zeta}$ is the classical econometric model: if $\boldsymbol{Y}$ represents dependent variables, and $\boldsymbol{X}$ independent, and B represents the effect of the dependent variables on each other.

B) $\Lambda_{\boldsymbol{Y}} = I_p$, $\Lambda_{\boldsymbol{X}} = I_q \implies$ The measurement error model:

- $\boldsymbol{\eta} = B\boldsymbol{\eta} + \Gamma\boldsymbol{\xi} + \boldsymbol{\zeta}$
- $\boldsymbol{Y} = \boldsymbol{\eta} + \boldsymbol{\epsilon}$, so $\boldsymbol{\epsilon}$ is the (random) error in the output (dependent variable / response).
- $\boldsymbol{X} = \boldsymbol{\xi} + \boldsymbol{\delta}$, so $\boldsymbol{\delta}$ is the (random) error in the input (independent variable / predictor).

C) Factor Analysis Models: Just take the measurement part $\boldsymbol{X} = \Lambda_{\boldsymbol{X}}\boldsymbol{\xi} + \boldsymbol{\delta}$.

Relationship between exploratory and confirmatory FA

In EFA the number of latent variables is not determined in advance; further, the measurement errors are assumed to be uncorrelated. In CFA a model is constructed to a great extent **in advance**, the number of latent variables $\boldsymbol{\xi}$ is set by the analyst, whether a latent variable influences an observed variable is specified, some direct effects of latent on observed values are fixed to zero or some other constant (e.g., one), measurement errors $\boldsymbol{\delta}$ may correlate, the covariance of latent variables can be either estimated or set to any value. In practice, distinction between EFA and CFA is more blurred. For instance, researchers using traditional EFA procedures may restrict their analysis to a group of indicators that they believe are influenced by one factor. Or, researchers with poorly fitting models in CFA often modify their model in an exploratory way with the goal of improving fit.

Software available for fitting structural equation models

There are two packages for SEM in R: `lavaan` and `sem`. `sem` is an older package, whereas `lavaan` aims to provide an extensible framework for SEMs and their extensions:

- can mimic commercial packages (including those below)
- provides convenience functions for specifying simple special cases (such as CFA) but also a more flexible interface for advanced users
- mean structures and multiple groups
- different estimators and standard errors (including robust)
- handling of missing data
- linear and nonlinear equality and inequality constraints
- categorical data support
- multilevel SEMs
- package `blavaan` for Bayesian estimation

Demonstration: Structural equation

Please watch this optional video:

The following video discusses the contents of this demonstration.

Transcript

This demonstration can be completed using the provided RStudio environment or your own RStudio.

To complete this task select the 'SEM_Example.demo.Rmd' in the 'Files' section of RStudio. Follow the demonstration contained within the RMD file.

If you choose to complete the example in your own RStudio, upload the following file:

SEM_Example.demo.Rmd

The output of the RMD file is also displayed below: