Structural equation modeling with R (`lavaan` package)

Paolo Ghisletta

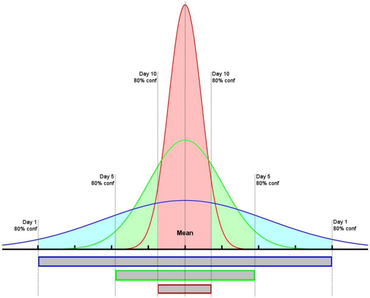
Faculty of Psychology and Educational Sciences, University of Geneva, Switzerland
Swiss Distance Learning University, Switzerland
LIVES–Overcoming vulnerability: Life course perspectives, Universities of Lausanne and Geneva, Switzerland

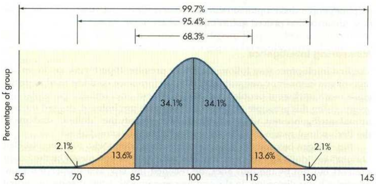
Nov. 1, 2016

Plan

Continuous data

The distinctive feature of continuous data is that (sometimes) we may assume **normality** (Gaussian distribution).

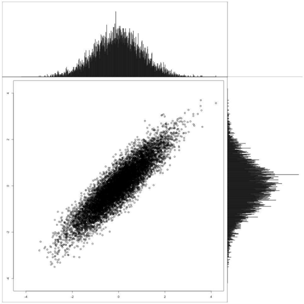




If a variable is normally distributed then its information can be summarized by two statistics : its mean and its variance ($x \sim \mathcal{N}(\mu, \sigma^2)$).

Continuous data (cont.)

If two variables are normally distributed and linearly associated they are said to be **bivariate normal**.



Bivariate normal variables can be summarized by 5 statistics : 2 means, 2 variances, and 1 covariance.
We can generalize bivariate normality to **multivariate normality** when additional variables are considered.

Plan

Definition

Structural Equation Modeling (SEM) is a statistical technique that allows :

- formally representing a multivariate theory about a large number of measured variables
- test the adequacy of such a theory to explain the structure of the data

The application of SEM was limited to multivariate normal variables, because the estimation technique required that the information about the variables could be summarized by means, variances, and covariances.

Recent advances in SEM, however, allow applying SEM to variables that are not multivariate normal.

Common research situations

SEMs are particularly often used with

- questionnaires with multiple items
- test batteries
- theories with multiple latent (unmeasured) variables

Maximum likelihood estimation

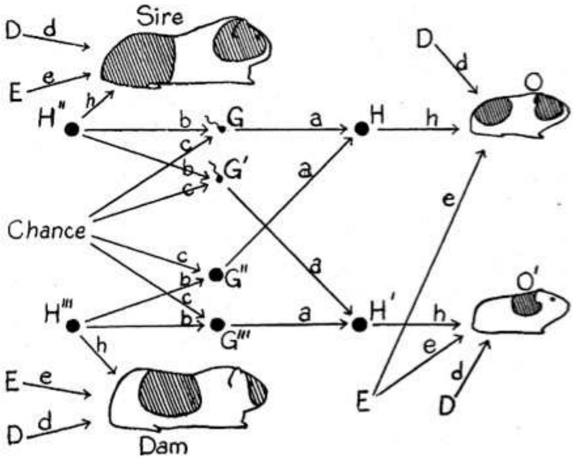
Maximum Likelihood Estimation (MLE) is a method to estimate the parameters of a statistical model. This method produces parameter estimates that make the observed results (your data) the most probable given that the model is correct.

To apply MLE to multivariate normal data we just have to apply it to the vector of means and the covariance matrix.
Let $\mathbf{m}$ be the vector of observed means, S the observed covariance matrix, $\boldsymbol{\mu}_{\hat{\theta}}$ the prediction (expectation) about the means, and $\boldsymbol{\Sigma}_{\hat{\theta}}$ the expectation about the covariance matrix, where $\hat{\theta}$ are the estimated parameters. The MLE is defined as

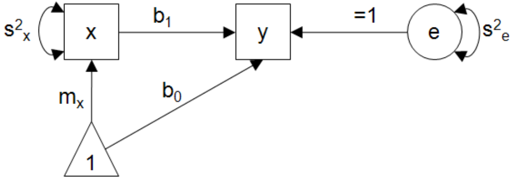
$$-2\ln\mathcal{L} = (N-1)\left[\ln|\Sigma_{\hat{\theta}}| - \ln|S| + \operatorname{trace}\left(S \times \Sigma_{\hat{\theta}}^{-1}\right) - k\right] + \frac{N}{N-1}\left(\mathbf{m}-\boldsymbol{\mu}_{\hat{\theta}}\right)' \Sigma_{\hat{\theta}}^{-1}\left(\mathbf{m}-\boldsymbol{\mu}_{\hat{\theta}}\right) + 1$$

Sewell Wright's tracing rules

Wright, S. (1918). On the nature of size factors. *Genetics, 3*, 367-374.
Wright, S. (1920). The relative importance of heredity and environment in determining the piebald pattern of guinea-pigs. *Proceedings of the National Academy of Sciences, 6*, 320-332.
Wright, S. (1934) The method of path coefficients. *Annals of Mathematical Statistics, 5*, 161-215.



Example : Simple Regression



$$\mathbf{m}^\mathsf{T} = \begin{bmatrix} m_x & m_y & 0 \end{bmatrix} \qquad \boldsymbol{\mu}_{\boldsymbol{\theta}}^\mathsf{T} = \begin{bmatrix} m_x & b_0 + b_1 m_x & 0 \end{bmatrix}$$

$$\mathbf{S} = \begin{bmatrix} s_x^2 & s_{x,y} & 0 \\ s_{y,x} & s_y^2 & s_{y,e} \\ 0 & s_{e,y} & s_e^2 \end{bmatrix} \qquad \boldsymbol{\Sigma}_{\theta} = \begin{bmatrix} s_x^2 & b_1 s_x^2 & 0 \\ b_1 s_x^2 & (b_1)^2 s_x^2 + s_e^2 & s_e^2 \\ 0 & s_e^2 & s_e^2 \end{bmatrix}$$

Example : Simple Regression (cont.)

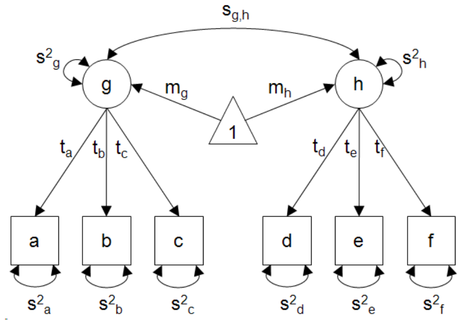
$$\mathbf{m}^{\intercal} = \begin{bmatrix} m_x & m_y & 0 \end{bmatrix} \qquad \boldsymbol{\mu}_{\boldsymbol{\theta}}^{\intercal} = \begin{bmatrix} m_x & b_0 + b_1 m_x & 0 \end{bmatrix}$$

$$\mathbf{S} = \begin{bmatrix} s_x^2 & s_{x,y} & 0 \\ s_{y,x} & s_y^2 & s_{y,e} \\ 0 & s_{e,y} & s_e^2 \end{bmatrix} \qquad \boldsymbol{\Sigma}_{\boldsymbol{\theta}} = \begin{bmatrix} s_x^2 & b_1 s_x^2 & 0 \\ b_1 s_x^2 & (b_1)^2 s_x^2 + s_e^2 & s_e^2 \\ 0 & s_e^2 & s_e^2 \end{bmatrix}$$

$$\begin{aligned} m_x &= m_x \\ m_y &= b_0 + b_1 m_x \\ m_e &= 0 \end{aligned} \qquad \begin{aligned} s_x^2 &= s_x^2 \\ s_{x,y} &= b_1 s_x^2 \\ s_{x,e} &= 0 \end{aligned} \qquad \begin{aligned} s_y^2 &= (b_1)^2 s_x^2 + s_e^2 \\ s_{y,e} &= s_e^2 \\ s_e^2 &= s_e^2 \end{aligned}$$

These equations are simultaneously solved for. Because there is a unique solution for each unknown, OLS estimation is appropriate.

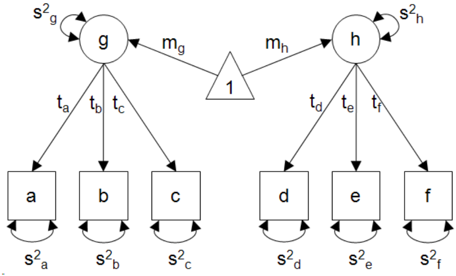
2-correlated-factor model with simple structure



$$\boldsymbol{\mu}_{\boldsymbol{\theta}}^{\mathsf{T}} = \begin{bmatrix} m_g t_a & m_g t_b & m_g t_c & m_h t_d & m_h t_e & m_h t_f \end{bmatrix}$$

$$\boldsymbol{\Sigma}_{\boldsymbol{\theta}} = \begin{bmatrix} (t_a)^2 s_g^2 + s_a^2 & t_a s_g^2 t_b & t_a s_g^2 t_c & t_a s_{g,h} t_d & t_a s_{g,h} t_e & t_a s_{g,h} t_f \\ t_b s_g^2 t_a & (t_b)^2 s_g^2 + s_b^2 & t_b s_g^2 t_c & t_b s_{g,h} t_d & t_b s_{g,h} t_e & t_b s_{g,h} t_f \\ t_c s_g^2 t_a & t_c s_g^2 t_b & (t_c)^2 s_g^2 + s_c^2 & t_c s_{g,h} t_d & t_c s_{g,h} t_e & t_c s_{g,h} t_f \\ t_d s_{g,h} t_a & t_d s_{g,h} t_b & t_d s_{g,h} t_c & (t_d)^2 s_h^2 + s_d^2 & t_d s_h^2 t_e & t_d s_h^2 t_f \\ t_e s_{g,h} t_a & t_e s_{g,h} t_b & t_e s_{g,h} t_c & t_e s_h^2 t_d & (t_e)^2 s_h^2 + s_e^2 & t_e s_h^2 t_f \\ t_f s_{g,h} t_a & t_f s_{g,h} t_b & t_f s_{g,h} t_c & t_f s_h^2 t_d & t_f s_h^2 t_e & (t_f)^2 s_h^2 + s_f^2 \end{bmatrix}$$

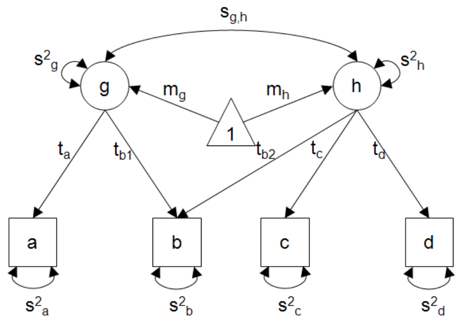
2-uncorrelated-factor model with simple structure



$$\boldsymbol{\mu}_{\boldsymbol{\theta}}^{\mathsf{T}} = \begin{bmatrix} m_g t_a & m_g t_b & m_g t_c & m_h t_d & m_h t_e & m_h t_f \end{bmatrix}$$

$$\boldsymbol{\Sigma}_{\boldsymbol{\theta}} = \begin{bmatrix} (t_a)^2 s_g^2 + s_a^2 & t_a s_g^2 t_b & t_a s_g^2 t_c & 0 & 0 & 0 \\ t_b s_g^2 t_a & (t_b)^2 s_g^2 + s_b^2 & t_b s_g^2 t_c & 0 & 0 & 0 \\ t_c s_g^2 t_a & t_c s_g^2 t_b & (t_c)^2 s_g^2 + s_c^2 & 0 & 0 & 0 \\ 0 & 0 & 0 & (t_d)^2 s_h^2 + s_d^2 & t_d s_h^2 t_e & t_d s_h^2 t_f \\ 0 & 0 & 0 & t_e s_h^2 t_d & (t_e)^2 s_h^2 + s_e^2 & t_e s_h^2 t_f \\ 0 & 0 & 0 & t_f s_h^2 t_d & t_f s_h^2 t_e & (t_f)^2 s_h^2 + s_f^2 \end{bmatrix}$$

2-factor model with complex structure



$$\mu_\theta^\intercal = \begin{bmatrix} m_g t_a & m_g t_{b1} + m_h t_{b2} & m_h t_c & m_h t_d \end{bmatrix}$$

$$\Sigma_\theta = \begin{bmatrix} (t_a)^2 s_g^2 + s_a^2 & t_a s_g^2 t_{b1} + t_a s_{g,h} t_{b2} & t_a s_{g,h} t_c & t_a s_{g,h} t_d \\ t_{b1} s_g^2 t_a + t_{b2} s_{g,h} t_a & s_{b-tot}^2 & t_{b1} s_{g,h} t_c + t_{b2} s_h^2 t_c & t_{b1} s_{g,h} t_d + t_{b2} s_h^2 t_d \\ t_c s_{g,h} t_a & t_c s_{g,h} t_{b1} + t_c s_h^2 t_{b2} & (t_c)^2 s_h^2 + s_c^2 & t_c s_h^2 t_d \\ t_d s_{g,h} t_a & t_d s_{g,h} t_{b1} + t_d s_h^2 t_{b2} & t_d s_h^2 t_c & (t_d)^2 s_h^2 + s_d^2 \end{bmatrix}$$

where $s_{b-tot}^2 = (t_{b1})^2 s_g^2 + (t_{b2})^2 s_h^2 + 2 t_{b1} s_{g,h} t_{b2} + s_b^2$

What SEM information do we care about ?

- Overall adjustment : How well does the model reproduce the structure of the data ? i.e., how close are $\mathbf{\Sigma}_{\hat{\theta}}$ and S ?
- Null hypothesis H_0 of each parameter in $\hat{\theta}$. Is each parameter needed (i.e., different from 0) ?
- Can we omit certain parameters ? (i.e., trimming strategy)
- Should we estimate additional parameters ? (i.e., building strategy)
- What is the effect size of each parameter ?

How can and should we test SEMs ?

- *Strictly confirmatory strategy* : Test only the model representing one's theory ; Limiting and misleading, thus discouraged
- *Model modification strategy* : Adjust one's theory so that the final model describes better the data. Model loses status of hypothesis, thus discouraged ; Must cross-validate. Attn : Loss of Type I error rate !
- *Model comparison strategy* : Alternative models are postulated and tested against preferred model ; Preferred method

Special cases of SEM

- General linear model (simple and multiple linear regression, t-test, ANOVA, etc.)
- Path analysis
- Confirmatory factor analysis
- Latent variable paths
- Multi-trait multi-method analysis
- Latent (growth) curve models
- Moderation and mediation models

Plan

SEM packages in R

- `sem` : By John Fox. One of the first packages, does not have many advanced options. http://socserv.mcmaster.ca/jfox/Misc/sem/index.html
- `OpenMx` : By Michael Neale, Steve Boker, et al. Revived the Mx freeware. Offers most advanced options. Has no default options. http://openmx.psyc.virginia.edu/
- `lavaan` : By Yves Rosseel. Very didactic and easy to use for most common models. http://lavaan.ugent.be/. Yves Rosseel (2012). `lavaan` : An R Package for Structural Equation Modeling. *Journal of Statistical Software, 48(2)*, 1-36. http://www.jstatsoft.org/v48/i02/

For more information, see https://pairach.com/2011/08/13/r-packages-for-structural-equation-model/.