General structural model – Part 1: Covariance structure and identification

Latent variables

- Interchangeably used:
  - constructs --- substantively defined concepts
  - common factors or components --- something uncorrelated with measurement errors (cf. case of cause indicators)
  - true scores in measurement theory
- Generality vs. specificity --- latent variables are operationally defined to imply a set of indicators (but the reverse is not true), i.e., more generalizable than what's intended by each indicator
- Good indicators
  - represent distinctive aspects of a defined concept
  - reasonably inter-correlated (i.e., internally reliable) since they possess some common characteristics
  - factorially simple; ideally uni-factorial

General structural models

- General structural equation models combine:
    - measurement models that operationally define a set of theoretical concepts, allowing for fallible measurement, with
    - a path model that explains complex relationships among a set of (mostly) latent variables
- All forms of models considered so far are special cases of GM (e.g., path models only with observed variables, CFA, MIMIC, etc.) --- rules and principles considered so far apply only to a part of GM, and so we will consider ways of combining them as well

- The general SE model subsumes most multivariate (causal) models, including MANOVA, discriminant function analysis, multivariate regression, canonical correlation analysis (CCA)
- The equivalence is about the model form, not about what optimization function is used
  - The above listed models can be considered as special cases of CCA, which uses the OLS like loss function (i.e., maximum accounted variance of manifest variables)
  - This type of variance maximization tends to fit more of variances than covariances --- cf. principal component analysis vs. common factor analysis
  - Component model type of approach to SEM --- Partial Least Squares (Wold, 1974, European Economic Review, 67-86) and Generalized Structured Component Analysis (Hwang & Takane, 2004, Psychometrika, 81-99)

The model form

- By combining measurement models and a path model, we mean:

$$\boldsymbol{\eta} = \mathbf{B}\boldsymbol{\eta} + \boldsymbol{\Gamma}\boldsymbol{\xi} + \boldsymbol{\zeta} = \left(\mathbf{I} - \mathbf{B}\right)^{-1}\left(\boldsymbol{\Gamma}\boldsymbol{\xi} + \boldsymbol{\zeta}\right)$$

$$\mathbf{y} = \boldsymbol{\Lambda}_y \boldsymbol{\eta} + \boldsymbol{\varepsilon} = \boldsymbol{\Lambda}_y \left(\mathbf{I} - \mathbf{B}\right)^{-1}\left(\boldsymbol{\Gamma}\boldsymbol{\xi} + \boldsymbol{\zeta}\right) + \boldsymbol{\varepsilon}$$

$$\mathbf{x} = \boldsymbol{\Lambda}_x \boldsymbol{\xi} + \boldsymbol{\delta}$$

where we have 4 parameter matrices of "regression weights" $(\mathbf{B}, \boldsymbol{\Gamma}, \boldsymbol{\Lambda}_y, \boldsymbol{\Lambda}_x)$ and 4 covariance matrices for exogenous latent variables $(\boldsymbol{\Phi}, \boldsymbol{\Psi}, \boldsymbol{\Theta}_{\varepsilon}, \boldsymbol{\Theta}_{\delta})$ --- any cause indicators included in $\boldsymbol{\xi}$

- Based on this structural representation, we build covariance structure of observed variables $(\mathbf{y}$ and $\mathbf{x})$

Implied covariance matrix

$$\boldsymbol{\Sigma}(\boldsymbol{\theta}) = \left[\begin{array}{c:c} \boldsymbol{\Sigma}_{yy}(\boldsymbol{\theta}) & \boldsymbol{\Sigma}_{yx}(\boldsymbol{\theta}) \\ \hdashline \boldsymbol{\Sigma}_{xy}(\boldsymbol{\theta}) & \boldsymbol{\Sigma}_{xx}(\boldsymbol{\theta}) \end{array}\right]$$

$$= \left[\begin{array}{c:c} \boldsymbol{\Lambda}_y (\mathbf{I}-\mathbf{B})^{-1} (\boldsymbol{\Gamma}\boldsymbol{\Phi}\boldsymbol{\Gamma}' + \boldsymbol{\Psi})(\mathbf{I}-\mathbf{B})^{-1\prime} \boldsymbol{\Lambda}_y' + \boldsymbol{\Theta}_\varepsilon & \boldsymbol{\Lambda}_y (\mathbf{I}-\mathbf{B})^{-1} \boldsymbol{\Gamma}\boldsymbol{\Phi}\boldsymbol{\Lambda}_x' \\ \hdashline \boldsymbol{\Lambda}_x \boldsymbol{\Phi}\boldsymbol{\Gamma}'(\mathbf{I}-\mathbf{B})^{-1\prime} \boldsymbol{\Lambda}_y' & \boldsymbol{\Lambda}_x \boldsymbol{\Phi}\boldsymbol{\Lambda}_x' + \boldsymbol{\Theta}_\delta \end{array}\right]$$

- Compare to the case only with observed variables:

$$\left[\begin{array}{c:c} (\mathbf{I}-\mathbf{B})^{-1} (\boldsymbol{\Gamma}\boldsymbol{\Phi}\boldsymbol{\Gamma}' + \boldsymbol{\Psi})(\mathbf{I}-\mathbf{B})^{-1\prime} & (\mathbf{I}-\mathbf{B})^{-1} \boldsymbol{\Gamma}\boldsymbol{\Phi} \\ \hdashline \boldsymbol{\Phi}\boldsymbol{\Gamma}'(\mathbf{I}-\mathbf{B})^{-1\prime} & \boldsymbol{\Phi} \end{array}\right]$$

- What if some ε are correlated with δ?

Identification

- $\boldsymbol{\theta}$ is <u>globally</u> identified if no vector $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$ exist such that:

$$\boldsymbol{\Sigma}\left(\boldsymbol{\theta}_1\right) = \boldsymbol{\Sigma}\left(\boldsymbol{\theta}_2\right), \quad \boldsymbol{\theta}_1 \neq \boldsymbol{\theta}_2$$

- If any element of $\boldsymbol{\theta}$ is unidentifiable, estimates of all others, if any, shouldn’t be interpreted

- Identifiability, as before, means parameters “can be” identified by algebraic form but doesn’t mean “will be” numerically

  E.g., γ_{11} is <u>algebraically</u> solved to be $\mathrm{cov}\left(x_2, y_1\right) / \mathrm{cov}\left(x_2, x_1\right)$ and sample estimate of $\mathrm{cov}(x_2, x_1)$ is very close to 0

- What we’ve learned so far about identification equally applies here but only for parts of GM, not as a whole

t-rule

- Number of free parameters in $\boldsymbol{\theta}$ does not exceed distinctive elements in $\mathbf{S}$:

$$t \le \frac{1}{2}\left(p+q\right)\left(p+q+1\right)$$

- Necessary, not sufficient

- How many distinctive elements in $\mathbf{S} = \begin{bmatrix} 1.0 & & \\ 0.5 & 1.0 & \\ 0.5 & 0.5 & 1.0 \end{bmatrix}$

Two-step rule

- Step 1 --- ignore all paths between latent variables, treat them as simply mutually correlated factors, and use any feasible identification rule for CFA (with fully unconstrained $\mathbf{\Phi}$)

- Step 2 --- if passes step 1, ignore all measurement relationships and use any feasible identification rule for SEM with observed variables

  Any cause indicators can be treated as latent variables with loading of 1 and no measurement error --- equivalent to simply including them in ξ

- Sufficient, not necessary; so, there will be identifiable models that do not pass two-step rule --- local identification will be useful in such cases, though it's empirical and fallible

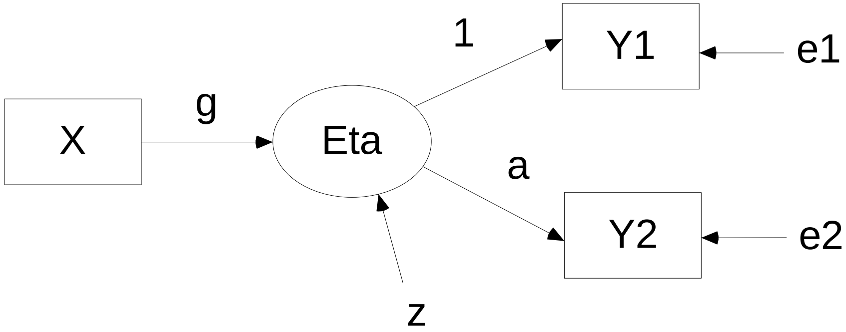
- See Fig. 8.3 (p. 329), 8.4 (p. 330) and 8.5 (p. 333)

MIMIC rule

- Multiple Indicators and MultIple Causes (MIMIC) model --- a latent variable is measured with multiple (reflective) indicators and caused by multiple observed variables (formative or cause indicators, or covariates)

- A sufficient condition per η for identification of a MIMIC model:

$$p \geq 2, \quad q \geq 1$$



$$\mathbf{S} = \begin{bmatrix} \phi & & \\ \gamma\phi & \gamma^2\phi + \psi + \theta_{11} & \\ \lambda\gamma\phi & \lambda\gamma^2\phi + \lambda\psi & \lambda^2\gamma^2\phi + \lambda^2\psi + \theta_{22} \end{bmatrix}$$

- For cases when $m > 1$, if all cause indicators affect all η's and each η is measured with at least 2 indicators, the 2-step rule applies; how?

- In general, too limited model condition in that no path modeling allowed between latent variables