SEM with observed variables: estimation

Psychology 588: Covariance structure and factor models

Estimation

- Tries to find a solution that best approximates ideally population covariance matrix $\boldsymbol{\Sigma}$, but in reality a sample estimate $\mathbf{S}$

$$F = f\left(\boldsymbol{\Sigma}(\hat{\boldsymbol{\theta}}) \middle| \mathbf{S}\right)$$

- The “best” approximation is defined in various ways, leading to different fitting functions
- Our job is to find which is the best under what data conditions

Desirable fitting functions

- have the following properties:

  $f\left(\mathbf{S}, \boldsymbol{\Sigma}(\hat{\boldsymbol{\theta}})\right)$ is a scalar

  $f\left(\mathbf{S}, \boldsymbol{\Sigma}(\hat{\boldsymbol{\theta}})\right) \geq 0$

  $f\left(\mathbf{S}, \boldsymbol{\Sigma}(\hat{\boldsymbol{\theta}})\right) = 0$ if and only if $\boldsymbol{\Sigma}(\hat{\boldsymbol{\theta}}) = \mathbf{S}$

  $f\left(\mathbf{S}, \boldsymbol{\Sigma}(\hat{\boldsymbol{\theta}})\right)$ is continuous both in $\mathbf{S}$ and $\boldsymbol{\Sigma}(\hat{\boldsymbol{\theta}})$

- Minimizing such fitting functions provides a consistent estimator of $\boldsymbol{\theta}$ (e.g., ML, ULS and GLS) --- true for all functions to be considered

Desirable asymptotic properties of estimators

- Unbiased
- Consistent
- Efficient

Note: asymptotic means "$N \to \infty$" by definition but its practical meaning is "as N becomes sufficiently large"
--- "how large is sufficient" will depend on many things such as complexity of the model, size of measurement errors, etc.

$\boldsymbol{\theta}$: parameters in the population

$\hat{\boldsymbol{\theta}}_N$: estimate of $\boldsymbol{\theta}$ from a sample of size N

- If $E\left(\hat{\boldsymbol{\theta}}_N\right) = \boldsymbol{\theta}$, $\hat{\boldsymbol{\theta}}_N$ is unbiased
- If $E\left(\hat{\boldsymbol{\theta}}_N\right) = \boldsymbol{\theta}$ as $N \to \infty$, $\hat{\boldsymbol{\theta}}_N$ is asymptotically unbiased
- If $P\left(\left|\hat{\boldsymbol{\theta}}_N - \boldsymbol{\theta}\right| < \alpha\right) = 1$ as $N \to \infty$ for any $\alpha > 0$, $\hat{\boldsymbol{\theta}}_N$ is consistent --- or called $''\text{plim}\,\hat{\boldsymbol{\theta}} = \boldsymbol{\theta}''$
- $\hat{\boldsymbol{\theta}}_N$ is efficient if its asymptotic variance is the minimum of all consistent estimator of $\boldsymbol{\theta}$ --- see Appendix B for more details on asymptotic properties of estimators

Maximum likelihood

- ML assumes:
  - Satisfactorily large sample
  - All observed variables distributed multivariate normal --- we will consider later a relaxed alternative to this for exogenous $\mathbf{x}$
  - All observations independent and identically distributed
- Minimizing its fitting function F_{ML} maximizes joint log likelihood of the model parameters $\boldsymbol{\theta}$ given observed data $\mathbf{S}$

$$F_{\mathrm{ML}} = \log\left|\hat{\boldsymbol{\Sigma}}\right| + \mathrm{tr}\left(\mathbf{S}\hat{\boldsymbol{\Sigma}}^{-1}\right) - \log|\mathbf{S}| - (p+q)$$

Obviously, both $\mathbf{S}$ and $\hat{\boldsymbol{\Sigma}}$ must be nonsingular for F_{ML} to be defined

How to optimize a fitting function?

- Find the partial derivatives w.r.t. all free model parameters $\partial F(\boldsymbol{\theta})/\partial \boldsymbol{\theta}$ and solve $\partial F(\boldsymbol{\theta})/\partial \boldsymbol{\theta} = \mathbf{0}$ --- necessary for minimization
- The second-derivative matrix $\partial^2 F(\boldsymbol{\theta})/\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'$ is positive definite (nonsingular) at the $\hat{\boldsymbol{\theta}}$ that minimizes $f(\boldsymbol{\theta})$ --- sufficient for minimization
- Usually, there is no closed form solution to this problem; instead, the minimization attained by an iterative numerical method based on gradients $\mathbf{g}$ (i.e., the 1st derivatives), given
  1. starting values of $\hat{\boldsymbol{\theta}}$
  2. step length (i.e., how much to change $\hat{\boldsymbol{\theta}}$ per iteration)
  3. when to stop iteration

- Parameters are initialized at some rational (instead of random) values --- e.g., reports from prior research, OLS estimates for loadings and causal paths, etc.

- The iterative update can be written as:

$$\hat{\boldsymbol{\theta}}^{(i+1)} = \hat{\boldsymbol{\theta}}^{(i)} - \mathbf{C}^{(i)}\mathbf{g}^{(i)}$$

  - Steepest descent: $\mathbf{C} = \mathbf{I}$
  - Newton-Raphson: $\mathbf{C} = \mathbf{H}^{-1}$, $\mathbf{H} = 2^{\text{nd}}$ partial derivatives (so called Hessian matrix)

- Iteration stops when $\left| f\left(\hat{\boldsymbol{\theta}}\right)^{(i)} - f\left(\hat{\boldsymbol{\theta}}\right)^{(i+1)} \right| < \alpha$ or $\left| \hat{\theta}^{(i)} - \hat{\theta}^{(i+1)} \right| < \alpha$ for all parameters, with α at an arbitrarily small value (e.g., 10^{-7})

Properties of ML

- Asymptotically unbiased
- Consistent
- Efficient
- ML estimates are asymptotically normal --- while the 1st PDs = 0 provides the estimates, square-roots of the diagonal entries of the 2nd PDs give their standard errors, allowing for a z-test
- F_{ML} is scale invariant:

$$F_{\mathrm{ML}}\left(\mathbf{S},\hat{\mathbf{\Sigma}}\right)=F_{\mathrm{ML}}\left(\mathbf{DSD},\mathbf{D}\hat{\mathbf{\Sigma}}\mathbf{D}\right)$$

with a diagonal matrix $\mathbf{D}$, with diagonal entries all non-zero

- Scale freeness of ML estimates of $\boldsymbol{\theta}$: $\mathbf{y} = \mathbf{B}\mathbf{y} + \boldsymbol{\Gamma}\mathbf{x} + \boldsymbol{\zeta}$ is functionally equivalent to $\tilde{\mathbf{y}} = \tilde{\mathbf{B}}\tilde{\mathbf{y}} + \tilde{\boldsymbol{\Gamma}}\tilde{\mathbf{x}} + \tilde{\boldsymbol{\zeta}}$ with $\tilde{\mathbf{y}} = \mathbf{D}_y\mathbf{y}$, $\tilde{\mathbf{x}} = \mathbf{D}_x\mathbf{x}$ and so $\tilde{\mathbf{B}} = \mathbf{D}_y\mathbf{B}\mathbf{D}_y^{-1}$, $\tilde{\boldsymbol{\Gamma}} = \mathbf{D}_y\boldsymbol{\Gamma}\mathbf{D}_x^{-1}$, $\tilde{\boldsymbol{\zeta}} = \mathbf{D}_y\boldsymbol{\zeta}$

  e.g., with $\mathbf{D}_y = \mathrm{diag}\left(s_{y_1}^{-1},...,s_{y_p}^{-1}\right)$ and $\mathbf{D}_x = \mathrm{diag}\left(s_{x_1}^{-1},...,s_{x_q}^{-1}\right)$, $\tilde{\mathbf{y}}$ and $\tilde{\mathbf{x}}$ become standardized variables, and $\tilde{\mathbf{B}}$ and $\tilde{\boldsymbol{\Gamma}}$ become standardized estimates of parameters

- What happens to the invariance and freeness if some parameters are subject to non-zero constant or equality constraints?

- Overall (badness of) fit of identifiable models:

$$(N-1)F_{\mathrm{ML}} \sim \chi^2$$

with $H_0: \mathbf{S} = \hat{\boldsymbol{\Sigma}}, \quad df = (p+q)(p+q+1)/2 - t$

  - Practical dilemma: asymptotic theory requires sufficiently large N and the chi-square statistic proportionally increases with N while its model df doesn't change

  - Relaxed condition for exogenous $\mathbf{x}$: observed exogenous $\mathbf{x}$ don't have to be multinormal; instead, if $\mathbf{y}$ conditional to $\mathbf{x}$ is multinormal and $\mathbf{x}$ is independent of $\boldsymbol{\theta}$, then all ML properties hold

Unweighted least squares

- ULS assumes nothing
- Consistent but not efficient
- Not scale invariant or scale free
- No statistical testing available; bootstrapping may be used for the overall fit

$$F_{\text{ULS}} = \frac{1}{2}\text{tr}\left[\left(\mathbf{S} - \hat{\boldsymbol{\Sigma}}\right)^2\right]$$

$$cf.\ \ \text{tr}\left[\left(\mathbf{z} - \hat{\mathbf{z}}\right)\left(\mathbf{z} - \hat{\mathbf{z}}\right)'\right], \quad \mathbf{z} = \begin{bmatrix}\mathbf{y} \\ \mathbf{x}\end{bmatrix}$$

Generalized least squares

- GLS adjusts the residual matrix, $\mathbf{S} - \hat{\boldsymbol{\Sigma}}$ for unequal variances and covariances

$$F_{\text{GLS}} = \frac{1}{2}\text{tr}\left[\left(\left[\mathbf{S} - \hat{\boldsymbol{\Sigma}}\right]\mathbf{W}^{-1}\right)^2\right]$$

Analogous to weighted least-squares for OLS --- typically used to deal with heterogeneous variance over observations (vertically written):

$$\hat{\mathbf{b}}_{\text{OLS}} = \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{y}$$

$$\hat{\mathbf{b}}_{\text{WLS}} = \left(\mathbf{X}'\mathbf{W}^{-1}\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{W}^{-1}\mathbf{y}, \quad \mathbf{W} = \text{diag}\left(\sigma_1^2,...,\sigma_N^2\right)$$

- Assuming that $\mathbf{W}$ is positive-definite (asymptotically or given so), GLS estimates $\hat{\boldsymbol{\theta}}$ are
  - consistent
  - distributed multinormal with known ACOV, and so a z-test available
  - but not all $\mathbf{W}^{-1}$ lead to efficient estimators
- Two more conditions needed for efficiency:
  - s_{ij} is unbiased estimator of σ_{ij}
  - Entries of $\mathbf{S}$ are asymptotically distributed multinormal, with mean of σ_{ij} and asymptotic covariance of

$$\operatorname{acov}\left(s_{ij}, s_{gh}\right) = N^{-1}\left(\sigma_{ig}\sigma_{jh} + \sigma_{ih}\sigma_{jg}\right)$$

- The ACOV condition is satisfied if $\mathbf{x}$ and $\mathbf{y}$ are multinormal, or if "tail" of the distribution is not excessively thick or thin

- Most common choice of $\mathbf{W}$ is $\mathbf{S}$, and so $\operatorname{plim}\mathbf{W} = \boldsymbol{\Sigma}$

$$F_{\text{GLS}} = \frac{1}{2}\operatorname{tr}\left[\left(\mathbf{I} - \boldsymbol{\Sigma}(\boldsymbol{\theta})\mathbf{S}^{-1}\right)^2\right]$$

  which is scale invariant and its estimates are scale free

- Like F_{ML}, $(N-1)F_{\text{GLS}}$ is asymptotically chi-square distributed; consequently, with large N, F_{ML} and F_{GLS} should be close to each other if H_0 is true (i.e., the model correctly specified) --- see, e.g., Table 4.3, p. 121

What if parametric assumptions not met

- Give up statistical inferences
- Use goodness (or badness) of fit indices (e.g., GFI, CFI, TLI, RMSEA, AIC, BIC, etc.) --- these non-statistical measures assume nothing and, in consequence, parametric testing unavailable
- Use non-parametric testing (e.g., bootstrap, randomization, etc.) --- to be elaborated in the following
- If nonnormality is the only major issue, data might be transformed to approximate normality --- consequently, the modeled linear relationships are for the transformed variables, which should be substantively meaningful

The bootstrap

- The idea of bootstrapping evolved from jackknifing (due to Tukey) by introducing random selection of observational units from the sample, with replacement

- The bootstrap is one way of "empirically" obtaining a sampling distribution of a statistic (e.g., F_{ML}, F_{GLS}, direct & indirect effects, etc.)

- The bootstrap distribution is asymptotically normal (as original $N \rightarrow \infty$)

- When only covariance or correlation data are provided, resampling is not possible --- pseudo raw data may be sampled from a parametric distribution, e.g., $N(\mathbf{0}, \mathbf{S})$ --- called "parametric bootstrapping"

“Bollen-Stine” bootstrap method

- Transforms the data $\mathbf{X}$ so that the resulting covariance matrix equals to model-implied cov matrix $\hat{\boldsymbol{\Sigma}}$ as (vertically written):

$$\mathbf{Z} = \mathbf{X}\mathbf{S}^{-0.5}\hat{\boldsymbol{\Sigma}}^{0.5}$$

  which makes H_0 true in the bootstrap population $\mathbf{Z}$, and sampling error of the statistic becomes smaller (“efficient”) leading to a more confident statistical inference

- The idea of transforming $\mathbf{X}$ to $\mathbf{Z}$ with which H_0 is true was given by Beran and Srivastava (1985) more generally for the eigen structure of $\mathrm{cov}(\mathbf{X})$

- Bollen, K.A., & Stine, R.A. (1993). Bootstrapping goodness-of-fit measures in structural equation models. In K.A. Bollen & J.S. Long (Eds), Testing structural equation models (pp. 111-135). Newbury Park, CA: Sage.
- Beran, R., & Srivastava, M.S. (1985). Bootstrap tests and confidence regions for functions of a covariance matrix. Annals of Statsitics, 13, 95-115.

- Is the Bollen-Stine method always good? --- depends on whether the specified model is correct:

  The transformation "forces" H_0 to be true in the bootstrap population $\boldsymbol{Z}$ so as to produce (narrower) sampling distribution of statistics (overall fits) for more powerful testing against H_{A} (typically a nested model, $df_0 < df_{\mathrm{A}}$)

  However, when H_0 is substantially false, a test for H_{A} based on a "wrong" transformation may be misleading

Scale and standardization

$$Y = -3 + 0.5X_1 + 2X_2 + 1.8X_3 + \zeta$$

$$\tilde{Y} = 0.67\tilde{X}_1 + 0.83\tilde{X}_2 + 0.75\tilde{X}_3 + \tilde{\zeta} \quad \text{(standardized)}$$

Y: job satisfaction (1 = least, 7 = most satisfied; s_y = 1.2)
X_1: annual salary in \$10K (s_{x1} = 1.6)
X_2: annual bonus in \$10K (s_{x2} = 0.5)
X_3: gender (1 = female, 0 = male, s_{x3} = 0.5)

- How should we compare effect of the salary and bonus?
- What is meant by 1.8 of gender effect? Would it be comparable to other effects if standardized?
- If alternative equations considered only with X_1 and X_2, separately for each gender, any concern for comparability?