Confirmatory Factor Analysis: Evaluation

Psychology 588: Covariance structure and factor models

Evaluation

- Evaluating a model is equivalent to testing a hypothesis of a set of constraints (that make $\hat{\boldsymbol{\Sigma}} \neq \mathbf{S}$ whether explicit or implicit), provided that all other assumptions fulfilled --- similar to the logic behind null hypothesis testing

- All overall fits quantify $\mathbf{S} - \hat{\boldsymbol{\Sigma}}$, whether they are statistical and/or standardized (i.e., bounded by [0,1])

- Overall fit indicates goodness (or badness) of the whole model, summarizing $\mathbf{S} - \hat{\boldsymbol{\Sigma}}$ into a scalar value

  Individual fits tell goodness of fit to particular manifest DVs (as indicated by R^2 or SMC) and standard error of parameters indicate how reliable parameter estimates are --- both overall and individual fits should be evaluated!

Residuals

- Residuals $s_{ij} - \hat{\sigma}_{ij}$ indicate how well the specified model's implied covariance matrix $\hat{\boldsymbol{\Sigma}}$ approximates the sample covariance matrix $\mathbf{S}$

- Sources of residuals:

  - $\boldsymbol{\Sigma} \neq \boldsymbol{\Sigma}(\boldsymbol{\theta})$, which we want to know by the residuals

  - sampling fluctuation --- with large N, smaller residuals are expected if the model is correct

  ** scale of observed variables determines the size of residuals --- if correlations are analyzed, residuals would not have scale dependency, varying [-2,2]

RMR (root mean-square of residual)

- RMR represents average residual, as SD indicates an average deviation of a variable around its mean

$$\mathrm{RMR} = \left(2 \sum_{i=1}^{q} \sum_{j=1}^{i} \frac{\left(s_{ij} - \hat{\sigma}_{ij} \right)^2}{q\left(q+1 \right)} \right)^{1/2}$$

  Note that the denominator counts variances as well

- As an alternative to $s_{ij} - \hat{\sigma}_{ij}$, fit to covariances may be compared by normalized residuals (useful to spot where the model predict poorly with an adjustment for sampling error):

$$\mathrm{N.R.} = \frac{\mathrm{residual}}{\mathrm{avar}\left(\mathrm{residual} \right)^{1/2}} = \frac{s_{ij} - \hat{\sigma}_{ij}}{\left(\left(\hat{\sigma}_{ii}\hat{\sigma}_{jj} + \hat{\sigma}_{ij}^2 \right) / N \right)^{1/2}}$$

Chi-square test (Browne, 74, 82, 84)

- The most common parametric statistical test for overall fit
- Chi-square test implies a null hypothesis of:

$$f\left(\mathbf{S}-\hat{\boldsymbol{\Sigma}}_{\text{tested}}\right)=0 \quad \text{or} \quad f\left(\mathbf{S}-\hat{\boldsymbol{\Sigma}}_{\text{nested}}\right)-f\left(\mathbf{S}-\hat{\boldsymbol{\Sigma}}_{\text{nestting}}\right)=0$$

  - Insignificant results suggest that the "tested" and "nested" models (with larger model df and $\hat{\chi}^2$) are not so bad
  - Which of $\alpha = 0.05$ vs. 0.25 would lead to more conservative test for a proposed model?
  - Re. the nested-nesting comparison, you may substantively prefer more parsimonious model sometime and more parameterized model other times --- what α to use?

- Approximation to a chi-square distribution will depend on assumptions made on observed DVs (e.g., not excessive kurtosis)

- Given non-normal data:
  - Data conditions where chi-square testing robust (Anderson & Gerbing, 1984; Bentler & Bonett, 1980)
  - Robust chi-square statistics (by Satora, Bentler, etc.)
  - Alternative estimators (e.g., ADF) with relaxed conditions (by Browne)
  - Bootstrapping

- What's an optimal N? --- dilemma between power and substantive meaning of misfit (see p. 269 for a state of the art in interpreting a chi-square value)

Incremental fit indices: Normed FI

$$\Delta_1 = \frac{F_b - F_m}{F_b} = \frac{\chi_b^2 - \chi_m^2}{\chi_b^2}$$

where F_b and F_m indicate the fitting function (e.g., F_{ML} or F_{GLS}) for the baseline (independence) and the hypothesized models, respectively --- one example for the baseline model is

$$n = q, \quad \boldsymbol{\xi} = \mathbf{x}, \quad \boldsymbol{\Lambda}_x = \mathbf{I}, \quad \boldsymbol{\Theta}_\delta = \mathbf{0}, \quad \boldsymbol{\Phi} = \mathrm{diag}\left(\phi_{11},...,\phi_{nn}\right)$$

- NFI (Bentler and Bonett, 1980) represents relative fit improvement compared to the upper bound F_b, due to the parameters that explain the data covariances
- Standardized to $0 \le \Delta_1 \le 1$

- Model df not taken into account, and so more parameterized models will be always preferred (if Δ_1 values are compared literally)

- "Sensitive" to change of N --- consequence of using a larger N (i.e., smaller sampling error) is confounded with improved fit due to better specification of the model

Δ_2: Incremental FI (Bollen, 1988)

$$\Delta_2 = \frac{F_b - F_m}{F_b - df_m \big/ \left(N-1\right)} = \frac{\chi_b^2 - \chi_m^2}{\chi_b^2 - df_m}$$

$$\approx \frac{\chi_b^2 - \chi_m^2}{\overline{\chi}_b^2 - df_m} = \frac{\text{observed improvement}}{\text{expectation under correct model}}$$

- $E(\Delta_1) = 1$ when the tested model is correct
- Δ_2 incorporates N so as to lessen the unaccounted dependency of Δ_1 on N
- Adjusts for df_m so that more parsimonious model is preferred for the same fit

- Not standardized since $\Delta_2 > 1$ for overfitting models that fit better than what's expected, i.e., $\chi^2_m < df_m$

- With large N, $\Delta_2 - \Delta_1$ becomes trivial since

$$\frac{F_b - F_m}{F_b - df_m / (N-1)} \approx \frac{F_b - F_m}{F_b}$$

ρ_1 : Relative FI (Bollen, 1986)

$$\rho_1 = \frac{\frac{F_b}{df_b} - \frac{F_m}{df_m}}{\frac{F_b}{df_b}} = \frac{\frac{\chi_b^2}{df_b} - \frac{\chi_m^2}{df_m}}{\frac{\chi_b^2}{df_b}}$$

- Comparison of fit made per df --- fit improvement evaluated as “efficiency” per df
- N not used in calculation, and so different levels of sampling error unaccounted
- Not lower bounded by 0 --- occurs when the test model’s relative improvement per df is worse than the baseline model’s

ρ_2 : Tucker-Lewis Index (1973)

$$\rho_2 = \frac{\frac{F_b}{df_b} - \frac{F_m}{df_m}}{\frac{F_b}{df_b} - \frac{1}{N-1}} = \frac{\frac{\chi_b^2}{df_b} - \frac{\chi_m^2}{df_m}}{\frac{\chi_b^2}{df_b} - 1}$$

- Parallel to the adjustment of Δ_1 for df
- ρ_1 defines the best fitting model as $F_m = \chi_m^2 = 0,$ while ρ_2 defines it as $\chi_m^2 = df_m,$ the expectation with a correct "null" model
- Not standardized
- ρ_2 value substantially less than 1 indicates the model misspecified and larger than 1 an overfitting

Comparative Fit Index

$$\mathrm{CFI} = 1 - \frac{\max\left(\chi_m^2 - df_m, 0\right)}{\max\left(\chi_b^2 - df_b, 0\right)} \qquad cf.\ \ \mathrm{RNI} = 1 - \frac{\chi_m^2 - df_m}{\chi_b^2 - df_b}$$

- Yet, another popular comparative index by Bentler (1990, *Psychological Bulletin*, *107*, 238–246)
- $\chi^2 - df$ represents noncentrality of a model --- expected deviation from the center of a (central) chi-square distribution for any wrong model
- Identical to RNI (Relative Noncentrality Index, McDonald & Marsh, 1990, *Psychological Bulletin, 107*, 247-255), except CFI bounded by [0,1]
- All comparative indices depend on choice of baseline

Goodness of Fit Index (Jöreskog & Sörbom, 1986)

$$\mathrm{GFI}_{\mathrm{ML}} = 1 - \frac{\mathrm{tr}\left(\left(\boldsymbol{\Sigma}^{-1}\mathbf{S} - \mathbf{I}\right)^2\right)}{\mathrm{tr}\left(\left(\boldsymbol{\Sigma}^{-1}\mathbf{S}\right)^2\right)}, \qquad \mathrm{AGFI}_{\mathrm{ML}} = 1 - \frac{q(q+1)}{2df}(1 - \mathrm{GFI})$$

- Measures relative fit of entries in $\mathbf{S}$ --- conceptually similar to R^2
- Absolute fit index in that the misfit was not compared to a worst situation or anything
- AGFI adjusts for df, preferring simpler models
- Similar indices can be defined for ULS and GLS (p. 277)

Root Mean Square Error of Approximation

$$\text{RMSEA} = \sqrt{\frac{\hat{F}_0}{df}} \hat{=} \sqrt{\max\left(\frac{\hat{F}}{df} - \frac{1}{N-1}, 0\right)}$$

$$E(F) = F_0 + \frac{df}{N-1}, \quad E\left(\chi^2_{\text{non-cent}}\right) = \chi^2_{\text{cent}} + df$$

- F_0 and F are, respectively, fit functions in the population and a sample for a hypothesized model; and when they are parametrically defined (e.g., ML, GLS), RMSEA provides statistical information on misfit due to misspecification
- Given 90% CI of an RMSEA, if its lower bound is 0 we don't reject the null hypothesis of exact fit by the considered model

- Quantifies discrepancy per df like Δ_2 and ρ_2
- RMSEA due to Steiger & Lind (1980, paper presented in the Psychometric Society meeting)
- Browne and Cudeck (1993, In *Testing structural equation models*, Ed. by Bollen & Long, pp. 136-162) suggests RMSEA < 0.05 for well-fitting model, < 0.08 for reasonable approximation, and > 0.1 unacceptable

And more…

- Parsimony indices adjusted for parsimony ratio, df_m / df_b
- Hoelter's CN estimates N to reject a model with a specific F value at α

$$\mathrm{CN} = \chi^2_\alpha / F + 1$$

- Akaike's and Bayes Information Criteria (AIC, BCC, BIC) useful for comparison of non-nested models as far as the same data are analyzed
- All fit indices can be cross-validated, jackknifed, and bootstrapped (if handled by more able computational environment, e.g., R or Matlab)
- Further reading: chapters 2, 3, 5, 6 and 8 in *testing structural equation models*, eds. Bollen & Long, 1993, Sage

Individual fit measures

- Estimates proper & “reasonable”?
- Roots of asymptotic variance of parameters are sampling errors and so we can statistically tell if a particular parameter differs from 0 by taking $\hat{\theta}/\mathrm{SE}(\hat{\theta})$ as a Z statistic
- R^2 for observed DVs tells how much the model explains of the variance of each observed DV --- coefficient of determination do the same thing but collectively for all observed variables based on the “generalized variance”, $\mathrm{CD} = 1 - \left|\hat{\boldsymbol{\Theta}}_{\delta}\right| \Big/ \left|\hat{\boldsymbol{\Sigma}}\right|$