Accounting for Population Uncertainty in Covariance Structure Analysis

Hao Wu

Boston College

May 21, 2013

Joint work with:

Michael W. Browne
The Ohio State University

Outline

Covariance Structures

- Covariance matrix among observed variables are usually implied by their relations with theoretically assumed latent variables.
- For example, the factor analysis model:

$$\begin{aligned} \mathbf{x} &= \boldsymbol{\Lambda}\mathbf{z} + \mathbf{u} \\ \boldsymbol{\Omega}(\boldsymbol{\xi}) &= \boldsymbol{\Lambda}\boldsymbol{\Phi}\boldsymbol{\Lambda}' + \boldsymbol{\Psi} \end{aligned}$$

  where $\boldsymbol{\Omega}$, $\boldsymbol{\Phi}$ and $\boldsymbol{\Psi}$ are the covariance matrices for $\mathbf{x}$, $\mathbf{z}$ and $\mathbf{u}$, respectively, and $\boldsymbol{\xi}$ is a vector of all parameters including the unknown loadings, factor correlations and unique variances.

parameter estimation and test statistic

- Parameter estimates are obtained by minimizing a discrepancy function F:

$$\hat{\xi} = \arg\min F(\mathbf{S}, \boldsymbol{\Omega}(\xi))$$

- We focus on the maximum Wishart likelihood (MWL) discrepancy function

$$F = -\ln |\boldsymbol{\Omega}^{-1}\mathbf{S}| + \mathrm{tr}(\boldsymbol{\Omega}^{-1}\mathbf{S}) - p$$

- If the true covariance matrix $\boldsymbol{\Sigma} = \boldsymbol{\Omega}(\xi_0)$ for some ξ_0,

$$\begin{aligned} \sqrt{n}(\hat{\xi} - \xi_0) & \xrightarrow{\mathcal{L}} \mathrm{N}(0, 2(\boldsymbol{\Delta}^{*\prime}\mathbf{V}^*\boldsymbol{\Delta}^*)^{-1}) \\ n\hat{F} & \xrightarrow{\mathcal{L}} \chi^2_{df} \end{aligned}$$

where $\boldsymbol{\Delta}^* = \partial\omega(\xi_0)/\partial\xi'$, $\mathbf{V}^* = (\boldsymbol{\Sigma} \otimes \boldsymbol{\Sigma})^{-1}$ and df is the degrees of freedom of the covariance structure.

when the model is not correct: $\boldsymbol{\Sigma} \neq \boldsymbol{\Omega}$

- Models are only approximations to the reality.
- Covariance structures rarely hold exactly in the population.
- All models will be rejected for exact fit when sample size is large enough.
- Misspecification must be accounted.

Parameter estimates

- $\hat{\boldsymbol{\xi}} \xrightarrow{\mathbf{p}} \boldsymbol{\xi}^{\#}$ as $n \to \infty$, where $\boldsymbol{\xi}^{\#} = \arg\min F(\boldsymbol{\Sigma}, \boldsymbol{\Omega}(\boldsymbol{\xi}))$ is a function of $\boldsymbol{\Sigma}$.
- Asymptotic normality with variance involving second derivatives of $\boldsymbol{\Omega}(\boldsymbol{\xi})$.

Test Statistics and RMSEA

$$n\hat{F} \xrightarrow{\mathcal{L}} \chi^2_{df,\delta} \quad \text{as} \quad n \to \infty$$

under the Pitman drift assumption:

- $nF^{\#} \to \delta$
- where $F^{\#} = F(\boldsymbol{\Sigma}, \boldsymbol{\Omega}^{\#})$ and $\boldsymbol{\Omega}^{\#} = \boldsymbol{\Omega}(\boldsymbol{\xi}^{\#})$.
- Misspecification diminishes with increasing sample size.
- Without this assumption, $n\hat{F} \to \infty$.

RMSEA: $\varepsilon = \sqrt{F^{\#}/df}$

- CI can be obtained from the non-central χ^2 sampling distribution of $n\hat{F}$.
- Test of close fit: $\varepsilon \leq 0.05$ vs. $\varepsilon > 0.05$

problems of the traditional approach

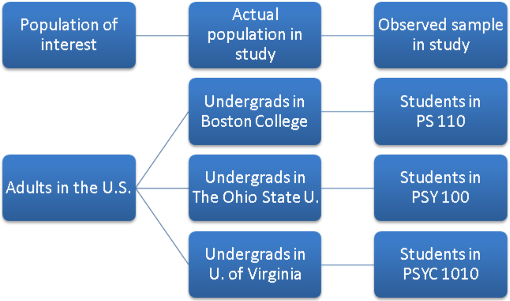
- Misspecification is not modeled or explained.
  - The same procedure is used as estimating a correctly specified model.
  - Misspecification is acknowledged in a post hoc manner.
- Misspecification is assumed a fixed effect in the population.
  - Fixed effect: does not change over replications;
  - Random effect: changes over replications.
  - Does misspecification change when the measurement is replicated?
- The Pitman drift assumption is impractical:
  - misspecification is a property of the population and should not be related to sample size.

How does misspecification arise?

Example: In a study, investigators have students in PS 110 fill out a questionnaire and find that the desired structure of the questionnaire is rejected.

- The study is more likely targeting an ideal population
  - such as “all adults in the U.S.”;
  - a more general/standard measurement environment
- The sample is not representative of the ideal population.
  - all observations from people of very similar age group
  - possibly measured under the same incidental effect of unknown sources
- Misspecification may come from the difference between the ideal and operational populations.

- If this study were to be replicated, would it be replicated in the same university?
  - Very unlikely.
  - Incidental effects may also change: different time of year (e.g. semester) etc.
  - The operational population has uncertainty, so misspecification is a random effect (model error).



A new approach

In the new approach to modeling misspecification, we assume

- Misspecification arises due to stochastic model error.
- Different replications of the same measurement involve different population covariance matrix $\boldsymbol{\Sigma}$.
- These unstructured $\boldsymbol{\Sigma}$ is a random effect centered on a structure $\boldsymbol{\Omega}(\xi_0)$.

The Model

- Sampling error is modeled in the traditional way:

$$\mathbf{S}|\boldsymbol{\Sigma} \sim \mathrm{W}_p(n^{-1}\boldsymbol{\Sigma}, n)$$

- Model error is also modeled randomly with a conjugate distribution:

$$\boldsymbol{\Sigma}|m, \boldsymbol{\Omega} \sim \mathrm{IW}_p(m\boldsymbol{\Omega}, m)$$

- Parameters of interest:
  - ξ as in covariance structure $\boldsymbol{\Omega}(\xi)$
  - $v = 1/m$: misspecification parameter (dispersion parameter)

$$
\begin{tikzcd}
\boldsymbol{\Omega}(\boldsymbol{\theta}) \arrow[d, "\boldsymbol{\Sigma} \mid m, \boldsymbol{\Omega} \sim \mathrm{IW}_p(m\boldsymbol{\Omega}, m)"] \\
\boldsymbol{\Sigma} \arrow[d, "\mathbf{S} \mid \boldsymbol{\Sigma} \sim \mathrm{W}_p(n^{-1}\boldsymbol{\Sigma}, n)"] \\
\mathbf{S}
\end{tikzcd}
$$

Parameter Estimation

- The marginal distribution is a matrix-variate Beta distribution:

$$\mathbf{S}|m, \boldsymbol{\Omega} \sim \mathrm{B}_p^{\mathrm{II}}(\frac{m}{n}\boldsymbol{\Omega}, \frac{n}{2}, \frac{m}{2})$$

- Parameters can be estimated by maximizing an adjusted marginal likelihood - Maximum Beta Likelihood (MBL).
- The marginal likelihood function is adjusted to correct for a downwards bias in estimating $v = 1/m$.

Relationship to RMSEA

When v is small,

$$\hat{v}^{\mathrm{IW}} \overset{\mathbf{a}}{=} \varepsilon^2$$

- Our measure of misspecification is asymptotically equivalent to (the square of) RMSEA.
- Note $\hat{v}^{\mathrm{IW}}$ is an estimate from the population model of inverted Wishart distribution.

Replication framework and consistency

- Both model error $(\mathbf{\Sigma}|\mathbf{\Omega}, m)$ and sampling error $(\mathbf{S}|\mathbf{\Sigma}, n)$ are sources of randomness.
- $\mathbf{S}$ has a Beta distribution.
- $\hat{\xi} = \xi_0 + o_p(1)$ and $\hat{v} = v_0 + o_p(v_0 + 1/n)$ as both $v_0 \to 0$ and $n \to \infty$.

A Weak Pitman's Drift

The sampling distributions are derived under the assumption of $n \to \infty$ and $v_0 = 1/m_0 \to 0$. This assumption is similar to the Pitman's drift assumption

- the sample size is assumed large, and
- misspecification is assumed small.

However, there is one major difference:

- We do not restrict the bivariate limit to a particular rate, while Pitman drift assumes $nF^{\#} \to \delta$, which would be equivalent to $m_0 = O(n)$ in our system.

This assumption is weaker and is more plausible in practice than Pitman drift assumption.

ξ

Under some regularity conditions, the MBLE $\hat{\xi}$ satisfy

$$\frac{\hat{\xi} - \xi_0}{\sqrt{v_0 + \epsilon}} \xrightarrow{\mathcal{L}} \mathrm{N}(0, 2(\boldsymbol{\Delta}^{*\prime}\mathbf{V}^*\boldsymbol{\Delta}^*)^{-1})$$

or

$$\hat{\xi} \overset{\mathbf{d}}{\approx} \mathrm{N}(\xi_0, 2(v_0 + \epsilon)(\boldsymbol{\Delta}^{*\prime}\mathbf{V}^*\boldsymbol{\Delta}^*)^{-1})$$

where $\epsilon = 1/n$ and $\mathbf{V}^* = (\boldsymbol{\Omega}_0 \otimes \boldsymbol{\Omega}_0)^{-1}$. Note the additive effects of the sampling and model error on the dispersion of the parameter estimate.

V

As $n \to \infty$ and $v_0 \to 0$, the MBLE $\hat{v}$ has sampling distribution

$$\frac{\hat{v}_0 + \epsilon}{v_0 + \epsilon} \xrightarrow{\mathcal{L}} \chi^2_{df}/df$$

where $\hat{v}_0$ is defined as

$$\hat{v}_0 = \begin{cases} \hat{v} & \hat{v} > 0 \\ (2df)^{-1}\mathrm{tr}\left\{(\hat{\mathbf{\Omega}} - \mathbf{S})\hat{\mathbf{\Omega}}^{-1}\right\}^2 - \frac{1}{n} & \hat{v} = 0 \end{cases}$$

Confidence bounds or intervals can be obtained by inverting this sampling distribution.

A t distribution

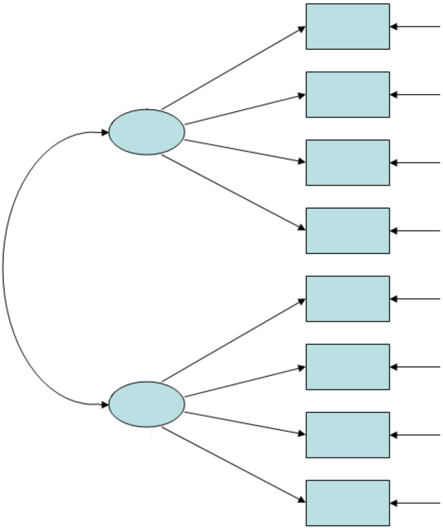
$\hat{\boldsymbol{\xi}}$ and $\hat{v}_0$ are asymptotically independent. As a result,

$$\frac{\hat{\xi}_i - \xi_{i0}}{\sqrt{2(\hat{v}_0 + \epsilon)[\hat{\boldsymbol{\Delta}}'\hat{\mathbf{V}}\hat{\boldsymbol{\Delta}}]^{ii}}} \xrightarrow{\mathcal{L}} \mathrm{t}_{df}$$

and a CI on $\boldsymbol{\xi}_{i0}$ is given by

$$\hat{\xi}_i \pm t_{df,1-\alpha/2}\sqrt{2(\hat{v}_0 + \epsilon)[\hat{\boldsymbol{\Delta}}'\hat{\mathbf{V}}\hat{\boldsymbol{\Delta}}]^{ii}}.$$

Model Being Used



A factor analysis model for correlation structure

$$\mathbf{\Omega} = \mathbf{D}(\mathbf{\Lambda}\mathbf{\Phi}\mathbf{\Lambda}' + \mathbf{D}_{\psi})\mathbf{D}$$

- Factor loadings
$$\mathbf{\Lambda} = \begin{pmatrix} 0.5 & 0.5 & 0.6 & 0.6 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0.7 & 0.7 & 0.8 & 0.8 \end{pmatrix}'$$
- Factor correlation $\rho = 0.5$
- Unique variances

$$\psi = [0.75, 0.75, 0.64, 0.64, 0.51, 0.51, 0.36, 0.36]'$$

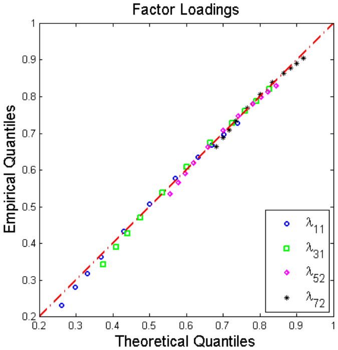
Design

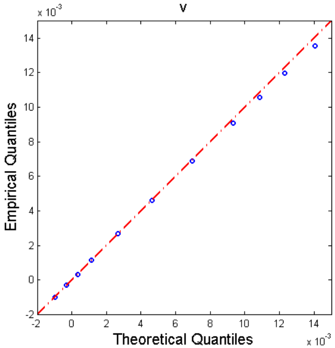
n and m chosen to be 200, 500, 1000 and ∞, Monto Carlo sample size $50,000$.

- Study I: 15 combinations of $(n,\ m_0)$, exam of the MBL estimates.
- Study II: $n = m = 1000$, comparison of MBL and MWL (traditional approach).

Study I

Selected results from $n = m = 200$ condition:





Study II

The missing rates (%) of the MBL and MWL 95% CIs with $n = m = 1000$.

	MBL	MWL
λ_{11}	5.49	32.09
λ_{31}	5.30	32.05
λ_{52}	5.34	31.79
λ_{72}	5.39	31.83
ρ	5.59	32.22
ψ_1	5.63	32.24
ψ_3	5.43	32.15
ψ_5	5.31	31.86
ψ_7	5.37	31.87

Summary

A new model for misspecified covariance structures

- It assumes misspecification arises from a random model error due to population uncertainty, and
- uses a weak version of the Pitman drift assumption in asymptotics.

We found

- $\hat{v}^{\mathrm{IW}} \approx \varepsilon^2$.
- the asymptotic variance of $\hat{\xi}$ has two additive parts due to sampling and model error respectively.
- the test statistic has a central χ^2 distribution

Simulation results show

- The asymptotic distributions work well.
- Assuming a random model error, MWL CIs have poor coverage as it fails to account for the extra source of variability.

Thank you!

Questions?