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

Traditional Approach to Covariance Structures

The New Approach

Sampling Distributions

Simulation Studies

Summary
Covariance Structures

- Covariance matrix among observed variables are usually implied by their relations with theoretically assumed latent variables.

- For example, the factor analysis model:

\[
x = \Lambda z + u
\]
\[
\Omega(\xi) = \Lambda \Phi \Lambda' + \Psi
\]

where \( \Omega, \Phi \) and \( \Psi \) are the covariance matrices for \( x, z \) and \( u \), respectively, and \( \xi \) 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 $F$:

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

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

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

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

$$
\sqrt{n}(\hat{\xi} - \xi_0) \xrightarrow{\mathcal{L}} N(0, 2(\Delta^*/V^*\Delta^*)^{-1})
$$

$$
n\hat{F} \xrightarrow{\mathcal{L}} \chi^2_{df}
$$

where $\Delta^* = \partial \omega(\xi_0)/\partial \xi'$, $V^* = (\Sigma \otimes \Sigma)^{-1}$ and $df$ is the degrees of freedom of the covariance structure.
when the model is not correct: $\Sigma \neq \Omega$

- 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

- $\hat{\xi} \xrightarrow{p} \xi^\#$ as $n \to \infty$, where $\xi^\# = \arg\min F(\Sigma, \Omega(\xi))$ is a function of $\Sigma$.
- Asymptotic normality with variance involving second derivatives of $\Omega(\xi)$. 
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 \( F\# = F(\Sigma, \Omega\#) \) and \( \Omega\# = \Omega(\xi\#) \).
- 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 \( \chi^2 \) 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 $\Sigma$.
- These unstructured $\Sigma$ is a random effect centered on a structure $\Omega(\xi_0)$. 
The Model

- Sampling error is modeled in the traditional way:

\[ S|\Sigma \sim W_p(n^{-1}\Sigma, n) \]

- Model error is also modeled \textit{randomly} with a conjugate distribution:

\[ \Sigma|m, \Omega \sim IW_p(m\Omega, m) \]

- Parameters of interest:
  - \( \xi \) as in covariance structure \( \Omega(\xi) \)
  - \( \nu = 1/m \): misspecification parameter (dispersion parameter)
\[ \Omega (\theta) \]

\[ \Sigma \mid m, \Omega \sim \text{IW}_p (m\Omega, m) \]

\[ \Sigma \]

\[ S \mid \Sigma \sim W_p (n^{-1}\Sigma, n) \]

\[ S \]
Parameter Estimation

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

\[ S|m, \Omega \sim B_p^\Pi \left( \frac{m}{n} \Omega, \frac{n}{2}, \frac{m}{2} \right) \]

- 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 \( \nu = 1/m \).
Relationship to RMSEA

When $\nu$ is small,

$$\hat{\nu}^\text{IW} \overset{a}{=} \varepsilon^2$$

- Our measure of misspecification is asymptotically equivalent to (the square of) RMSEA.
- Note $\hat{\nu}^\text{IW}$ is an estimate from the population model of inverted Wishart distribution.
Replication framework and consistency

- Both model error \((\Sigma | \Omega, m)\) and sampling error \((S | \Sigma, n)\) are sources of randomness.
- \(S\) 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 \( \nu_0 = 1/m_0 \to 0 \). 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.
Under some regularity conditions, the MBLE $\hat{\xi}$ satisfy

$$\frac{\hat{\xi} - \xi_0}{\sqrt{\nu_0 + \epsilon}} \xrightarrow{d} N(0, 2(\Delta^*\mathbf{V}^*\Delta^*)^{-1})$$

or

$$\hat{\xi} \overset{d}{\approx} N(\xi_0, 2(\nu_0 + \epsilon)(\Delta^*\mathbf{V}^*\Delta^*)^{-1})$$

where $\epsilon = 1/n$ and $\mathbf{V}^* = (\Omega_0 \otimes \Omega_0)^{-1}$. Note the additive effects of the sampling and model error on the dispersion of the parameter estimate.
As $n \to \infty$ and $v_0 \to 0$, the MBLE $\hat{\nu}$ has sampling distribution
\[
\frac{\hat{\nu}_0 + \epsilon}{\nu_0 + \epsilon} \xrightarrow{\mathcal{L}} \chi^2_{df}/df
\]
where $\hat{\nu}_0$ is defined as
\[
\hat{\nu}_0 = \begin{cases} 
\hat{\nu} & \hat{\nu} > 0 \\
(2df)^{-1} \text{tr} \left\{ (\hat{\Omega} - S)\hat{\Omega}^{-1} \right\}^2 - \frac{1}{n} & \hat{\nu} = 0
\end{cases}
\]
Confidence bounds or intervals can be obtained by inverting this sampling distribution.
A $t$ distribution

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

$$\frac{\hat{\xi}_i - \xi_{i0}}{\sqrt{2(\hat{\nu}_0 + \epsilon)[\hat{\Delta}'\hat{\nu}\hat{\Delta}]^i}} \overset{\mathcal{L}}{\longrightarrow} t_{df}$$

and a CI on $\xi_{i0}$ is given by

$$\hat{\xi}_i \pm t_{df,1-\alpha/2} \sqrt{2(\hat{\nu}_0 + \epsilon)[\hat{\Delta}'\hat{\nu}\hat{\Delta}]^i}.$$
Model Being Used
A factor analysis model for correlation structure

\[ \Omega = D(\Lambda \Phi \Lambda' + D\psi)D \]

- Factor loadings
  \[ \Lambda = \begin{pmatrix} 0.5 & 0.5 & 0.6 & 0.6 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0.7 & 0.7 & 0.8 & 0.8 \end{pmatrix}' \]
- Factor correlation \( \rho = 0.5 \)
- Unique variances
  \[ \psi = \begin{bmatrix} 0.75 & 0.75 & 0.64 & 0.64 & 0.51 & 0.51 & 0.36 & 0.36 \end{bmatrix}' \]
Design

$n$ and $m$ chosen to be 200, 500, 1000 and $\infty$, 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 $n = m = 1000$.

<table>
<thead>
<tr>
<th></th>
<th>MBL</th>
<th>MWL</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\lambda_{11}$</td>
<td>5.49</td>
<td>32.09</td>
</tr>
<tr>
<td>$\lambda_{31}$</td>
<td>5.30</td>
<td>32.05</td>
</tr>
<tr>
<td>$\lambda_{52}$</td>
<td>5.34</td>
<td>31.79</td>
</tr>
<tr>
<td>$\lambda_{72}$</td>
<td>5.39</td>
<td>31.83</td>
</tr>
<tr>
<td>$\rho$</td>
<td>5.59</td>
<td>32.22</td>
</tr>
<tr>
<td>$\psi_1$</td>
<td>5.63</td>
<td>32.24</td>
</tr>
<tr>
<td>$\psi_3$</td>
<td>5.43</td>
<td>32.15</td>
</tr>
<tr>
<td>$\psi_5$</td>
<td>5.31</td>
<td>31.86</td>
</tr>
<tr>
<td>$\psi_7$</td>
<td>5.37</td>
<td>31.87</td>
</tr>
</tbody>
</table>
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

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