New methods for calculating p-values for moment structure models

Steffen Grønneberg and Njål Foldnes

SEM meeting in Ghent, March, 2017

BI Norwegian Business School

Table of contents

Review of estimation and goodness-of-fit tests in SEM

Minimum discrepancy estimation

- Covariance models imply a parametric covariance structure

$$\theta \mapsto \Sigma(\theta),$$

  where, if the model is correct, there is a θ° with $\Sigma = \Sigma(\theta^\circ)$.

- We estimate θ° via

$$\hat{\theta} = \underset{\theta}{\operatorname{argmin}}\, F(s, \sigma(\theta)), \quad s = \operatorname{vech}(S),\ \sigma(\theta) = \operatorname{vech}(\Sigma(\theta))$$

  where S is the empirical covariance matrix, and where $F(\cdot, \cdot)$ is a *discrepancy function*, i.e., $F(s, \sigma) \geq 0$ for all s, σ; $F(s, \sigma) = 0$ if and only if $s = \sigma$; and F is twice continuously differentiable jointly.

- Shapiro (1983) derives expansions that give $\sqrt{n}(\hat{\theta} - \theta^\circ) \xrightarrow[n\to\infty]{D} N(0, \Sigma_\theta)$ under weak assumptions.

Testing the model

- Let $T_n = nF(s, \sigma(\hat{\theta}))$. Shapiro (1983) gives

$$T_n = \sqrt{n}(s - \sigma^\circ)' U \sqrt{n}(s - \sigma^\circ) + o_P(1).$$

  where

  1. Δ is the $p \times q$ derivative matrix $\partial\sigma(\theta)/\partial\theta'$
  2. $V = \frac{1}{2} \frac{\partial^2 F(s,\sigma)}{\partial s \partial \sigma}$
  3. $U = V - V\Delta \left\{\Delta' V \Delta\right\}^{-1} \Delta' V$

- Under model misspecification we have $T_n \to \infty$.
- If the model is correct, Shapiro (1983) showed that

$$T_n \xrightarrow[n\to\infty]{D} \sum_{j=1}^{d} \lambda_j Z_j^2, \qquad Z_1, \ldots, Z_d \sim N(0,1) \text{ IID}$$

  where $\lambda_1, \ldots, \lambda_d$ are the d non-zero eigenvalues of $U\Gamma$, where Γ is the asymptotic covariance matrix of $\sqrt{n}(s - \sigma)$.

- This induces a goodness of fit test for covariance models

Testing nested models

- Nested model comparison is done similarly.
- Let $\sigma = \sigma(\theta), \theta \in \Theta$ nest the restricted model $\sigma = \sigma(\theta), \theta \in \Theta_0$ where $\Theta_0 = \{\theta \in \Theta : a(\theta) = 0\}$ for some function a.
- Let

$$\hat{\theta} = \operatorname*{argmin}_{\theta \in \Theta} F(s, \sigma(\theta)), \qquad \tilde{\theta} = \operatorname*{argmin}_{\theta \in \Theta_0} F(s, \sigma(\theta))$$

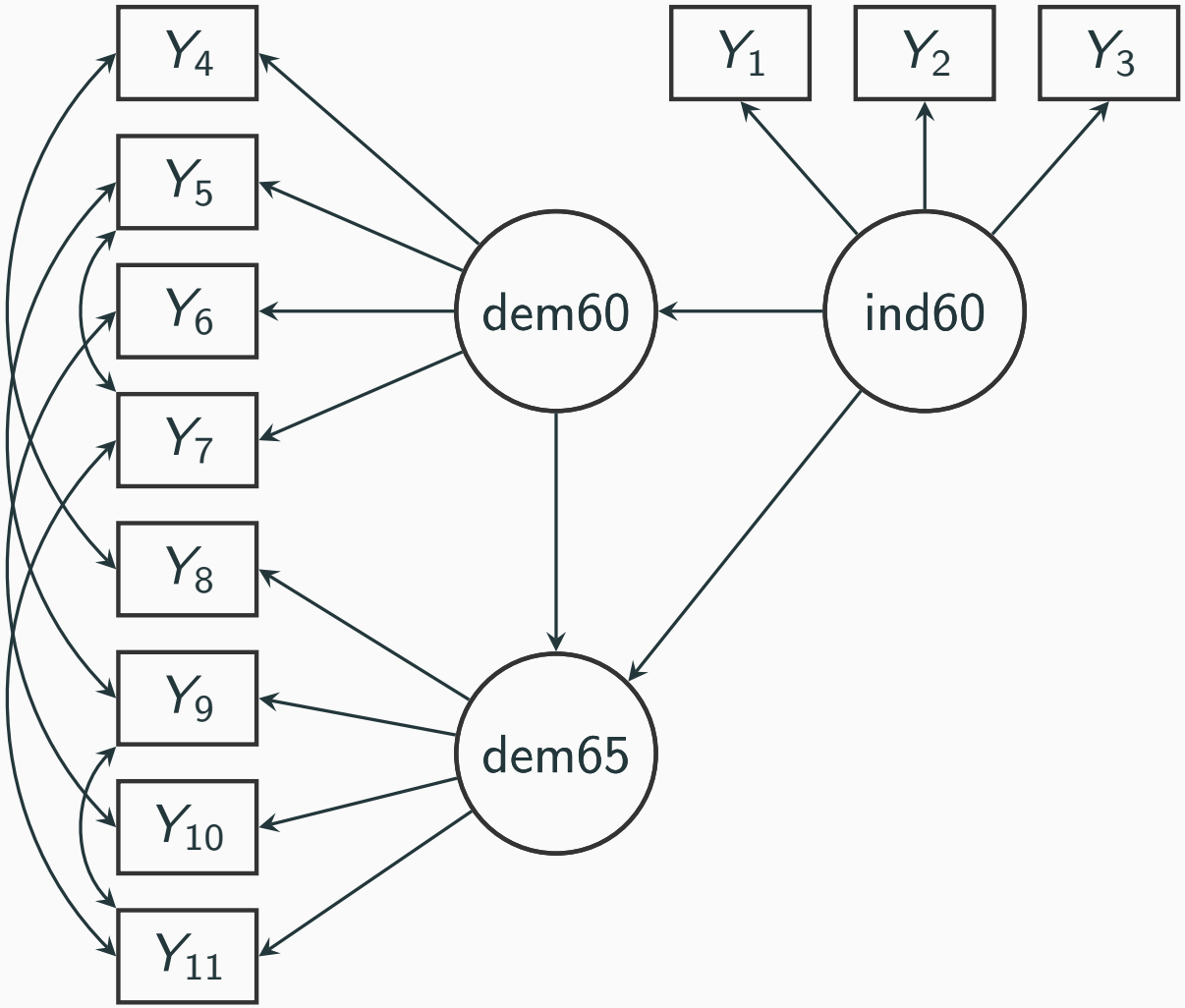
- A likelihood ratio type test is based on $\tilde{T}_n - T_n$ where $T_n = nF(s, \sigma(\hat{\theta}))$ and $\tilde{T}_n = nF(s, \sigma(\tilde{\theta}))$.
- Under the null hypothesis, we again have numbers $\alpha_1, \ldots, \alpha_m$

$$\tilde{T}_n - T_n \xrightarrow[n \to \infty]{D} \sum_{j=1}^{m} \alpha_j Z_j^2, \qquad Z_1, \ldots, Z_m \sim N(0,1) \text{ IID},$$

- Nested model comparison of the "likelihood ratio type" in general GMM also has this type of limit.

An example calculated at the population level

Figure 1: Bollen's political democracy model. dem60: Democracy in 1960. dem65: Democracy in 1965. ind60: Industrialisation in 1960.



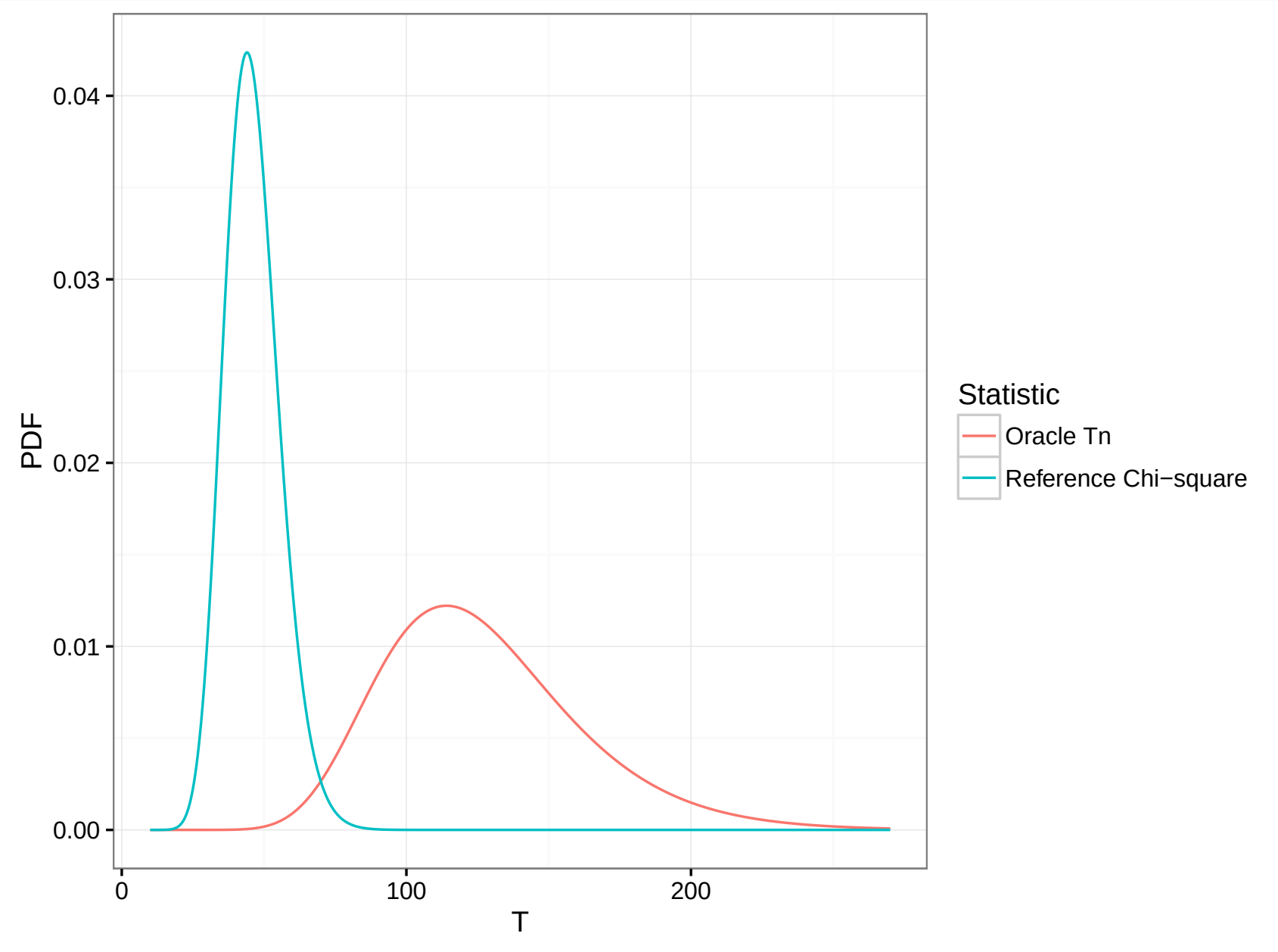
Population eigenvalues from NTML

- We fix the parameters in the SEM model and obtain σ°
- We generate highly non-normal data, with univariate kurtosis 21
- The model has (correctly specified) constraints on the unique variances, with 46 df.

The λ

Oracle: 11.0 8.8 8.2 7.6 7.4 7.1 7.0 f 4.3 4.1 3.3 3.2 3.0 2.7 2.5 2.5 2.4 2.2 2.2 2.2 2.1 2.0 2.0 1.8 1.8 1.4 1.3 1.3 1.3 1.3 1.3 1.2 1.2 1.2 1.2 1.2 1.2 1.1 1.1 1.1 1.1 1.1 1.0 1.0 1.0 1.0 1.0

Non-normality inflates the NTML chi-square



The ADF test

- To our knowledge, no one in psychometrics has actually used the result $T_n \xrightarrow[n\to\infty]{D} \sum_{j=1}^{d} \lambda_j Z_j^2$ directly.
- The problem is that $\lambda_1, \ldots, \lambda_d$ are unknown.
- The "asymptotically distribution free" solution to this problem: Choose F to simplify the limit using a GLS-type idea that also induces some *asymptotic* optimality properties.
- Asymptotics depends on $V = V_n = \frac{1}{2}\frac{\partial^2 F(s,\sigma)}{\partial s \partial \sigma}$
- F is so-called well-specified, if $V_n^{-1} \xrightarrow[n\to\infty]{P} \Gamma$, which implies $T_n = nF(s, \sigma(\hat{\theta})) \xrightarrow[n\to\infty]{D} \chi_d^2$.
- This is the so-called ADF test. It requires thousands of observations to have good finite sample properties (!) and is practically useless.

The Satorra-Bentler adjustment

- The well-used Satorra-Bentler adjustment instead tries to adjust the distribution of T_n to a χ^2_d.
- Instead of using T_n as a test statistic, they use

$$T_{n,\text{SB}} := \frac{T_n}{\hat{c}},$$

and use

$$\text{p-value} = P(\chi^2_d > T_{n,\text{SB}}).$$

- Thousands of applied papers use this technique (8500+ mentions of "Satorra-Bentler" on Google Scholar).
- It is only well-motivated when $\lambda_1 = \lambda_2 = \cdots = \lambda_d$, and this is not a reasonable assumption in practice. This is not well-known among practitioners.

The oracle p-value

The oracle distribution

We refer to the distribution of $\sum_{j=1}^{d} \lambda_j Z_j^2$ as the *oracle distribution*

Oracle p-value

Consider a fixed test value T_n, the oracle p-value is

$$p_n = P\left(\sum_{j=1}^{d} \lambda_j Z_j^2 > T_n\right). \tag{1}$$

where the probability is with respect to $Z_1, \ldots, Z_d$.

A new p-value approximation

$$\text{Oracle:} \qquad p_n = P\left(\sum_{j=1}^{d} \lambda_j Z_j^2 > T_n\right).$$

Limitation: We do not know the $\lambda_1, \ldots, \lambda_d$ and the sample size is finite.

In practice, we may estimate the λ_j:

$$\hat{p}_n = P\left(\sum_{j=1}^{d} \hat{\lambda}_j Z_j^2 > T_n\right), \tag{2}$$

and hope that the asymptotics "kick in" at moderate sample sizes!

Consistent tests

- An obvious idea: estimate $\lambda_1, \ldots, \lambda_d$ then use $T_n \xrightarrow[n\to\infty]{D} \sum_{j=1}^{d} \lambda_j Z_j^2$ to get p-values. This has not been done before in psychometrics, and we find this very strange.
- It is easy find estimators of $\lambda_1, \ldots, \lambda_d$, and we show this leads to consistent testing procedures.

Theorem

Let (T_n) be a sequence of random variables, and let $p_n = 1 - H(T_n; \lambda)$ and $\hat{p}_n = 1 - H(T_n; \hat{\lambda})$ where $H(q; \lambda_1, \ldots, \lambda_r) = P(\sum_{j=1}^{d} \lambda_j Z_j^2 \leq q)$. If $\hat{\lambda} \xrightarrow[n\to\infty]{P} \lambda$ where λ only has positive elements, then $\hat{p}_n - p_n = \|\hat{\lambda} - \lambda\| O_P(1)$, and hence, $\hat{p}_n - p_n \xrightarrow[n\to\infty]{P} 0$.

Middle-grounds

- Although our suggested approach is consistent, and does work better than the ADF test, we noticed that Satorra-Bentler was sometimes better – even though it was not consistent.
- λ is the eigenvalues of a large matrix depending on estimated high order moments, and is difficult to estimate well in realistic samples
- We then realized Satorra-Bentler can be re-interpreted as estimating λ under the restriction $\lambda_1 = \lambda_2 = \cdots = \lambda_d$. Indeed,

$$\hat{p}_{SB} = P\left(\sum_{j=1}^{d} \overline{\hat{\lambda}} Z_j^2 > T_n\right),$$

  where $\overline{\hat{\lambda}}$ is the mean of $\hat{\lambda}_1, \ldots, \hat{\lambda}_d$.
- While inconsistent, i.e., biased, it has low variability.
- This led us to suggest middle-ground estimators of λ balancing bias and variance.

Middle-ground p-value approximations

Between the "full" and the SB p-value approximations, many middle-ground approximations are possible. Here we only consider one:

$$\hat{p}_{n,half} = P\left(\sum_{j=1}^{d} \tilde{\lambda}_j Z_j^2 > T_n\right)$$

$$\tilde{\lambda}_1 = \cdots = \tilde{\lambda}_{\lceil d/2 \rceil} = \frac{1}{\lceil d/2 \rceil} \sum_{j=1}^{\lceil d/2 \rceil} \hat{\lambda}_j$$

and

$$\tilde{\lambda}_{\lceil d/2 \rceil + 1} = \cdots = \tilde{\lambda}_d = \frac{1}{d - \lceil d/2 \rceil} \sum_{j=\lceil d/2 \rceil + 1}^{d} \hat{\lambda}_j.$$

The λ

Oracle: 11.0 8.8 8.2 7.6 7.4 7.1 7.0 f 4.3 4.1 3.3 3.2 3.0 2.7 2.5 2.5 2.4 2.2 2.2 2.2 2.1 2.0 2.0 1.8 1.8 1.4 1.3 1.3 1.3 1.3 1.3 1.2 1.2 1.2 1.2 1.2 1.2 1.1 1.1 1.1 1.1 1.1 1.0 1.0 1.0 1.0 1.0

Half: 4.3 4.3 4.3 4.3 4.3 4.3 4.3 4.3 4.3 4.3 4.3 4.3 4.3 4.3 4.3 4.3 4.3 4.3 4.3 4.3 4.3 4.3 4.3 4.3 4.3 4.3 4.3 4.3 4.3 4.3 1.2 1.2 1.2 1.2 1.2 1.2 1.2 1.2 1.2 1.2 1.2 1.2 1.2 1.2 1.2 1.2 1.2 1.2 1.2 1.2 1.2 1.2 1.2 1.2 1.2 1.2 1.2 1.2

SB: 2.8 2.8 2.8 2.8 2.8 2.8 2.8 2.8 2.8 2.8 2.8 2.8 2.8 2.8 2.8 2.8 2.8 2.8 2.8 2.8 2.8 2.8 2.8 2.8 2.8 2.8 2.8 2.8 2.8 2.8 2.8 2.8 2.8 2.8 2.8 2.8 2.8 2.8 2.8 2.8 2.8 2.8 2.8 2.8 2.8 2.8 2.8 2.8 2.8 2.8 2.8 2.8 2.8 2.8 2.8 2.8 2.8 2.8

Computational tools

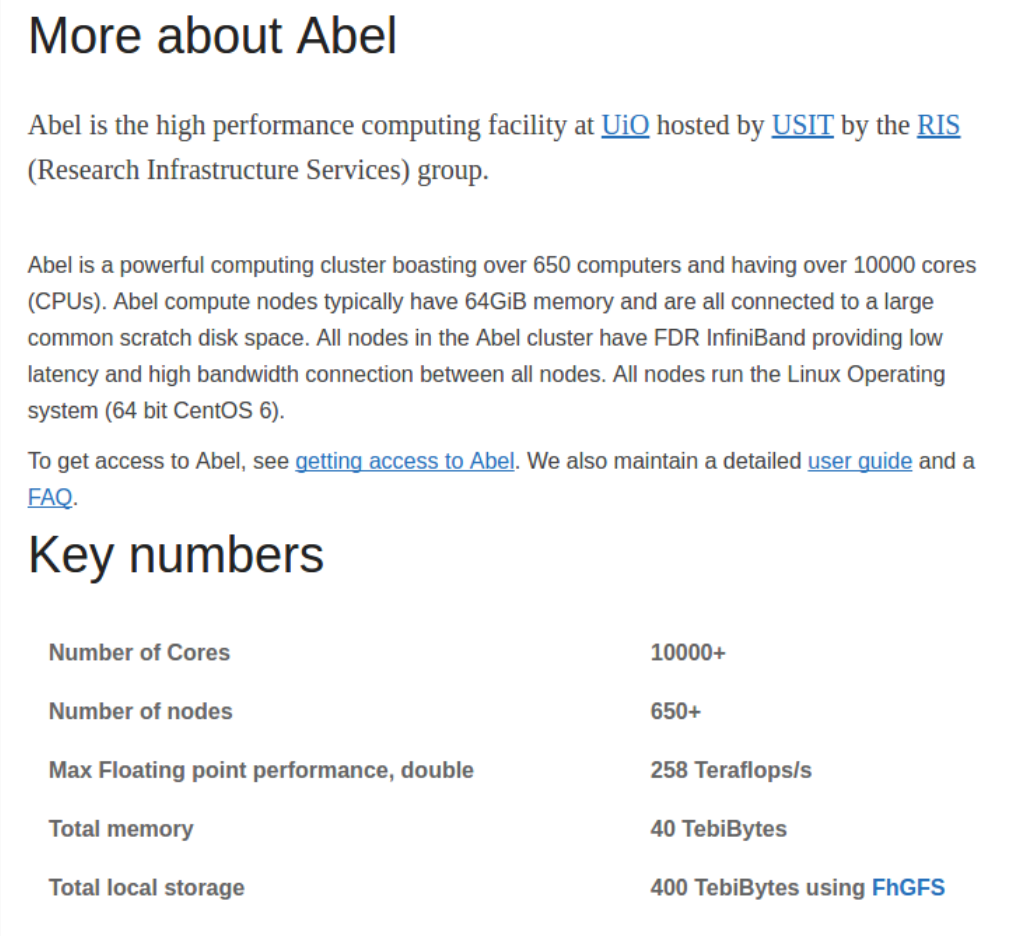
All computations implemented in R.

- The lavaan package of Yves Rosseel gives access to $\hat{U}$ and $\hat{\Gamma}$; and also to certain matrices needed for nested model testing
- The CompQuadForm package, Duchesne, Pierre, and Pierre Lafaye De Micheaux. "Computing the distribution of quadratic forms: Further comparisons between the LiuTangZhang approximation and exact methods." Computational Statistics and Data Analysis 54.4 (2010).

Also, we used the Abel Cluster at UiO.

Simulation results

We used UiO/NOTUR's Abel computing cluster for simulations. Our simulations used 3000 hours of computation time (about four months). The cluster allowed us to "time-travel"



More about Abel

Abel is the high performance computing facility at UiO hosted by USIT by the RIS (Research Infrastructure Services) group.

Abel is a powerful computing cluster boasting over 650 computers and having over 10000 cores (CPUs). Abel compute nodes typically have 64GiB memory and are all connected to a large common scratch disk space. All nodes in the Abel cluster have FDR InfiniBand providing low latency and high bandwidth connection between all nodes. All nodes run the Linux Operating system (64 bit CentOS 6).

To get access to Abel, see getting access to Abel. We also maintain a detailed user guide and a FAQ.

Key numbers

Number of Cores	10000+
Number of nodes	650+
Max Floating point performance, double	258 Teraflops/s
Total memory	40 TebiBytes
Total local storage	400 TebiBytes using FhGFS

Simulation results

We generate data from the Multivariate Normal distribution (where $\lambda_1 = \cdots = \lambda_d = 1$), and two non-normal distributions with the following λ_i:

	1.87	1.59	1.49	1.44	1.43	1.42	1.38
	1.36	1.35	1.34	1.31	1.29	1.26	1.13
Distribution 2	1.12	1.11	1.11	1.10	1.10	1.09	1.09
	1.08	1.08	1.07	1.07	1.07	1.06	1.05
	1.04	1.03	1.03	1.03	1.02	1.02	1.01
	4.16	3.24	2.88	2.82	2.70	2.67	2.51
	2.41	2.35	2.31	2.16	2.12	2.03	1.52
Distribution 3	1.50	1.47	1.43	1.40	1.38	1.36	1.35
	1.33	1.32	1.29	1.27	1.25	1.21	1.20
	1.13	1.13	1.11	1.09	1.08	1.08	1.06

Table 1: Eigenvalues λ_i, for $i = 1, \ldots, 35$, for Bollen's political democracy model, assuming correct model specification. Distribution 2 and 3 have univariate skewness and kurtosis $s = 1, k = 7$ and $s = 2, k = 21$, respectively.

Simulation results

Dist	n	NTML	SB	SS	BOST	EFULL	EHALF	SEL	ORAC
Norm	100	0.077	0.086	0.050	0.023	0.036	0.050	0.051	0.077
	300	0.055	0.053	0.052	0.037	0.037	0.043	0.045	0.055
	900	0.068	0.067	0.050	0.059	0.063	0.064	0.065	0.068
Dist 2	100	0.215	0.108	0.019	0.035	0.021	0.048	0.042	0.057
	300	0.197	0.070	0.018	0.053	0.024	0.045	0.045	0.057
	900	0.219	0.063	0.033	0.054	0.037	0.051	0.051	0.059
Dist 3	100	0.488	0.164	0.017	0.038	0.009	0.072	0.031	0.024
	300	0.591	0.094	0.013	0.068	0.013	0.050	0.038	0.045
	900	0.685	0.076	0.017	0.059	0.015	0.042	0.038	0.046

Table 2: Type I error rates for testing model $\mathcal{M}_1$. Normal: multivariate normal distribution, Distribution 2: skewness 1 and kurtosis 7. Distribution 3: skewness 2 and kurtosis 7. NTML=normal-theory likelihood ratio test. SB=Satorra-Bentler. SS=Scaled and shifted. BOST=Bollen-Stine bootstrap. EFULL= Full eigenvalue approximation, $\hat{p}_n$. EHALF= half eigenvalue approximation, $\hat{p}_{n,\text{half}}$. SEL = p-value obtained from selection algorithm. ORAC= oracle p-value p_n.

Simulation results nested models

We also did simulations for nested model comparison

D2	3.92	3.49	3.19	2.99	2.94	2.78	2.72	1.85	1.56	1.54	1.30
D3	10.64	8.79	8.06	7.58	7.37	6.94	6.76	4.09	3.16	3.10	2.04

Table 3: Eigenvalues of $U_d\Gamma$ for nested model testing. Distribution 2 has skewness 1 and kurtosis 7; Distribution 3 has skewness 2 and kurtosis 21. Rounded to two decimal places.

We now had even larger spread in the eigenvalues.

Interestingly, EFULL is now better than EHALF.

Dist	n	NTML	SB	BOST	EFULL	EHALF	SEL	ORAC
Norm	100	0.068	0.080	0.037	0.062	0.069	0.075	0.068
	300	0.054	0.059	0.046	0.053	0.055	0.058	0.054
	900	0.051	0.053	0.051	0.051	0.052	0.053	0.051
Dist 2	100	0.582	0.137	0.096	0.076	0.099	0.096	0.028
	300	0.659	0.088	0.081	0.052	0.066	0.062	0.035
	900	0.702	0.059	0.053	0.035	0.043	0.045	0.046
Dist 3	100	0.911	0.221	0.129	0.115	0.159	0.135	0.005
	300	0.961	0.126	0.118	0.062	0.089	0.082	0.018
	900	0.976	0.087	0.089	0.044	0.064	0.061	0.043

Table 4: Type I error rates for nested model testing. Normal: multivariate normal distribution, Distribution 2: skewness 1 and kurtosis 7. Distribution 3: skewness 2 and kurtosis 7. NTML=normal-theory likelihood ratio test. SB=Satorra-Bentler. BOST=Bollen-Stine bootstrap. EFULL= Full eigenvalue approximation, $\hat{p}_n$. EHALF= half eigenvalue approximation, $\hat{p}_{n,\text{half}}$. SEL = p-value obtained from selection algorithm. ORAC= oracle p-value p_n.

A new selection algorithm

Many competing approximations

- The SB procedure is well motivated under an equality constraint among all eigenvalues. But if the eigenvalues differ considerably in the population, this restriction may lead to poor estimates due to a high bias.
- In contrast, $\hat{p}_n$ is always a valid approximation for p_n in that it is asymptotically unbiased. However, in finite samples the variability of $\hat{\lambda}$ may lead to excessive variability in $\hat{p}_n$.
- The half p-value approximation lies somewhat in-between these two extremes. If the eigenvalues differ considerably, we split the eigenvalues into two groups: The highest half and the lowest half are estimated by their means, in order to reduce finite sample variability.
- Many more are possible to construct...

How do we choose the best p-value approximation, in a given real-world setting?

Notation for middle-ground approximations

Choose cut-off integers $1 < \tau_1 < \tau_2 < \cdots < \tau_k < d$ with $1 \leq k < d$. For $\tau_{l-1} \leq k < \tau_l$ let

$$\tilde{\lambda}_k = \frac{1}{\tau_l - \tau_{l-1}} \sum_{j=\tau_{l-1}}^{\tau_l - 1} \hat{\lambda}_j \tag{3}$$

where $\tau_0 = 1$ and $\tau_{k+1} = d$. Let us denote this choice by $\tilde{\lambda}(\tau) = (\tilde{\lambda}_1(\tau), \ldots, \tilde{\lambda}_r(\tau))'$. The proposed p-value estimator is then

$$\hat{p}_n(\tau) = P\left(\sum_{j=1}^{d} \tilde{\lambda}_j(\tau) Z_j^2 > T_n \right).$$

The selector

- We now have a whole class of p-value approximations. Which should we use? We introduce the following selection procedure:
- Suppose we could transform data via

$$\tilde{X}_i := \Sigma(\theta^\circ)^{1/2}\Sigma^{-1/2}X_i$$

- The population covariance matrix of $(\tilde{X}_i)$ is $\Sigma(\theta^\circ)$, where $\theta^\circ = \text{argmin}_\theta F(\sigma, \sigma(\theta))$ is the least false parameter configuration. Hence the transformed data follows the null hypothesis.
- Crucially, if the null hypothesis is true, $\Sigma(\theta^\circ)^{1/2}\Sigma^{-1/2} = I$, so $\tilde{X}_i = X_i$.
- Under the null hypothesis, an exact p-value is exactly uniform on $[0, 1]$.
- We therefore know that the best p-value approximation $\hat{p}_n$ minimizes

$$D_n = \sup_{0 \leq x \leq 1} |P_{H_0}(\hat{p}_n \leq x) - x|, \qquad \tilde{X}_i \sim P_{H_0}$$

- Via a non-parametric bootstrap procedure, we estimate D_n, and select the p-value approximation that minimizes $\hat{D}_n$.

Selection procedure

1: **procedure** SELECT(sample, B)
2: $\tilde{X}_i = \Sigma(\hat{\theta})^{1/2} S_n^{-1/2} X_i$ for $i = 1, 2, \ldots, n$.
3: **for** $k \leftarrow 1, \ldots, B$ **do**
4: boot.sample $\leftarrow$ Draw with replacement from transformed sample $\tilde{X}_i$
5: **for** $l \in 1, \ldots, L$ **do**
6: $\hat{p}_{n,l} \leftarrow$ based on boot.sample
7: **end for**
8: **end for**
9: **for** $l \in 1, \ldots, L$ **do**
10: $\hat{D}_{B,n,l} \leftarrow \sup_{0 \leq x \leq 1} |B^{-1} \sum_{k=1}^{B} I\{\hat{p}_{n,l} < x\} - x|$
11: **end for**
12: **return** $\operatorname{argmin}_{1 \leq l \leq L} \hat{D}_{B,n,l}$
13: **end procedure**

Selector works well for non-nested test

Dist	n	NTML	SB	SS	BOST	EFULL	EHALF	SEL	ORAC
Norm	100	0.077	0.086	0.050	0.023	0.036	0.050	0.051	0.077
	300	0.055	0.053	0.052	0.037	0.037	0.043	0.045	0.055
	900	0.068	0.067	0.050	0.059	0.063	0.064	0.065	0.068
Dist 2	100	0.215	0.108	0.019	0.035	0.021	0.048	0.042	0.057
	300	0.197	0.070	0.018	0.053	0.024	0.045	0.045	0.057
	900	0.219	0.063	0.033	0.054	0.037	0.051	0.051	0.059
Dist 3	100	0.488	0.164	0.017	0.038	0.009	0.072	0.031	0.024
	300	0.591	0.094	0.013	0.068	0.013	0.050	0.038	0.045
	900	0.685	0.076	0.017	0.059	0.015	0.042	0.038	0.046

Table 5: Type I error rates for testing model $\mathcal{M}_1$. Normal: multivariate normal distribution, Distribution 2: skewness 1 and kurtosis 7. Distribution 3: skewness 2 and kurtosis 7. NTML=normal-theory likelihood ratio test. SB=Satorra-Bentler. SS=Scaled and shifted. BOST=Bollen-Stine bootstrap. EFULL= Full eigenvalue approximation, $\hat{p}_n$. EHALF= half eigenvalue approximation, $\hat{p}_{n,\text{half}}$. SEL = p-value obtained from selection algorithm. ORAC= oracle p-value p_n.

Selector works well for nested models

Dist	n	NTML	SB	BOST	EFULL	EHALF	SEL	ORAC
	100	0.068	0.080	0.037	0.062	0.069	0.075	0.068
Norm	300	0.054	0.059	0.046	0.053	0.055	0.058	0.054
	900	0.051	0.053	0.051	0.051	0.052	0.053	0.051
	100	0.582	0.137	0.096	0.076	0.099	0.096	0.028
Dist 2	300	0.659	0.088	0.081	0.052	0.066	0.062	0.035
	900	0.702	0.059	0.053	0.035	0.043	0.045	0.046
	100	0.911	0.221	0.129	0.115	0.159	0.135	0.005
Dist 3	300	0.961	0.126	0.118	0.062	0.089	0.082	0.018
	900	0.976	0.087	0.089	0.044	0.064	0.061	0.043

Table 6: Type I error rates for nested model testing. Normal: multivariate normal distribution, Distribution 2: skewness 1 and kurtosis 7. Distribution 3: skewness 2 and kurtosis 7. NTML=normal-theory likelihood ratio test. SB=Satorra-Bentler. BOST=Bollen-Stine bootstrap. EFULL= Full eigenvalue approximation, $\hat{p}_n$. EHALF= half eigenvalue approximation, $\hat{p}_{n,\text{half}}$. SEL = p-value obtained from selection algorithm. ORAC= oracle p-value p_n.

New bootstrap tests for asymptotic robustness and SB consistency

When can we trust the NTML and the SB statistics?

According to asymptotic theory

- SB procedure is consistent if non-zero eigenvalues of $U\Gamma$ are all equal.
- NTML based procedure is consistent if all non-zero eigenvalues are equal to 1, which is known as asymptotic robustness.

It is of practical interest to develop procedures that may help assess if these assumptions hold in a given real-world setting. Such procedures might help the practitioner decide whether it is advisable to apply the NTML test, the SB test, or instead maybe some of the new p-value approximations...

Bootstrapping eigenvalues

Let E be the matrix of normalised (complex) eigenvectors of $U\Gamma$. Then $U\Gamma = E\Lambda E^{-1}$ where $\Lambda = \begin{pmatrix} \Lambda_d & 0 \\ 0 & 0 \end{pmatrix}$, and where Λ_d is the diagonal matrix with elements $\lambda_1, \ldots, \lambda_d$. Define

$$A = c^{1/2} \cdot E \begin{pmatrix} \Lambda_d^{-1/2} & 0 \\ 0 & 0 \end{pmatrix} E^{-1}, \tag{4}$$

where c denotes the mean value of the eigenvalues $\lambda_1, \ldots, \lambda_d$. In each bootstrap sample, we calculate $\hat{A}$ and $\hat{U}_{\text{boot}}\hat{\Gamma}_{\text{boot}}$ and form the matrix

$$W_n^* = \hat{A}\hat{U}_{\text{boot}}\hat{\Gamma}_{\text{boot}}\hat{A}.$$

Crucial observation: W_n^* converges to a matrix whose non-zero eigenvalues are all equal to c

Bootstrapping eigenvalues

In each bootstrap sample drawn from the original sample, we calculate

$$W_n^* = \hat{A}\hat{U}_{\text{boot}}\hat{\Gamma}_{\text{boot}}\hat{A}.$$

and the eigenvalues of W_n^* are then computed as $\hat{\lambda}_{\text{boot}}$. This process is repeated many times, and we get realizations $\hat{\lambda}_{k,\text{boot}}$, giving us information about the sampling variability of the estimated eigenvalues under the null hypothesis of identical eigenvalues.

Asymptotic robustness

The above procedure may also be adapted to test for asymptotic robustness of the NTML statistic T_{ML}, that is, whether $\lambda_j = 1$ for all $j = 1, \ldots, d$. We just set $c = 1$, i.e., use

$$A_1 = E \begin{pmatrix} \Lambda_d^{-1/2} & 0 \\ 0 & 0 \end{pmatrix} E^{-1},$$

Beran, Rudolf, and Muni S. Srivastava. "Bootstrap tests and confidence regions for functions of a covariance matrix." The Annals of Statistics (1985): 95-115.

Test statistics for the bootstrapping procedure

We need to define test statistics based on the bootstrap replications. We follow Section 4.3 in the Beran and Srivastava 1985 paper, and propose h that is non-negative and zero under the null hypothesis, and also with partial derivatives equal to zero under the null hypothesis.

- For SB consistency

$$h_{SB}(\lambda) = \log[(\lambda_1 + \lambda_d)^2] - \log[4\lambda_1\lambda_d].$$

- For AR

$$h_{AR}(\lambda) = d\log[d^{-1}\sum_{j=1}^{d}\lambda_j] - \log[\prod_{j=1}^{d}\lambda_j].$$

Testing for SB consistency and for AR

1: **procedure** BOOTSTRAP(sample, B)
2: Calculate $\hat{U}, \hat{\Gamma}, \hat{A}, \hat{A}_1$ from sample
3: $\hat{\lambda} \leftarrow$ The d largest eigenvalues of $\hat{U}\hat{\Gamma}$
4: $T_{n,\text{SB}} = h_{SB}(\hat{\lambda})$
5: $T_{n,\text{AR}} = h_{AR}(\hat{\lambda})$
6: **for** $k \leftarrow 1, \ldots, B$ **do**
7: boot.sample $\leftarrow$ Draw with replacement from sample
8: $\hat{U}_{\text{boot}}\hat{\Gamma}_{\text{boot}} \leftarrow$ Based on boot.sample
9: $W^*_{n,\text{SB}} \leftarrow \hat{A}\hat{U}_{\text{boot}}\hat{\Gamma}_{\text{boot}}\hat{A}$
10: $\hat{\lambda}_{k,\text{boot}} = (\hat{\lambda}_{k,1,\text{boot}}, \ldots, \hat{\lambda}_{k,d,\text{boot}})' \leftarrow$ the d largest eigenvalues of $W^*_{n,\text{SB}}$
11: $T_{n,k,\text{SB}} \leftarrow h_{SB}(\hat{\lambda}_{k,\text{boot}})$
12: $W^*_{n,\text{AR}} \leftarrow \hat{A}\hat{U}_{\text{boot}}\hat{\Gamma}_{\text{boot}}\hat{A}$
13: $\hat{\lambda}_{k,\text{boot}} = (\hat{\lambda}_{k,1,\text{boot}}, \ldots, \hat{\lambda}_{k,d,\text{boot}})' \leftarrow$ the d largest eigenvalues of $W^*_{n,\text{AR}}$
14: $T_{n,k,\text{AR}} \leftarrow h_{AR}(\hat{\lambda}_{k,\text{boot}})$
15: **end for**
16: **return** $B^{-1}\sum_{k=1}^{B} I\{T_{n,k,\text{SB}} > T_{n,\text{SB}}\}$ and $B^{-1}\sum_{k=1}^{B} I\{T_{n,k,\text{AR}} > T_{n,\text{AR}}\}$
17: **end procedure**

Bootstrap tests for AR and SB consistency

Unfortunately, these procedures need large sample sizes in order to reach acceptable Type I error rates.

Test	$n = 200$	$n = 400$	$n = 800$	$n = 2000$
AR	0.354	0.203	0.081	0.035
SB	0.369	0.195	0.070	0.033

Table 7: Type I error rates for tests of asymptotic robustness (AR) and Satorra-Bentler (SB) consistency.

Conclusions

We have introduced the following p-value for goodness-of-fit in SEM

$$\hat{p}_n = P\left(\sum_{j=1}^{d} \hat{\lambda}_j Z_j^2 > T_n\right).$$

where T_n is the typical NTML test statistic. The $\hat{\lambda}_j$ are constructed from the eigenvalues of $\hat{U}\hat{\Gamma}$. This framework encompasses many approaches:

- If we set $\hat{\lambda}_j = 1$ we get the NTML p-value
- If we set $\hat{\lambda}_j = \overline{\hat{\lambda}}$ we get the SB p-value
- If we split the eigenvalues of $\hat{U}\hat{\Gamma}$ into groups and use mean values within these groups, we get middle-ground approximations, e.g. $\hat{p}_{n,\text{half}}$.
- If we simply use the eigenvalues of $\hat{U}\hat{\Gamma}$, we get the "full" p-value $\hat{p}_n$, which is always consistent

Conclusions contd.

The new p-value approximations seem to perform better than SB in some non-normal conditions. In addition:

- A selection algorithm tries to pick the best p-value approximation for a given sample. Performs well.
- Bootstrap tests for AR and SB consistency, but these are presently of limited practical value