NEW TESTING PROCEDURES 
FOR STRUCTURAL EQUATION MODELING

STEFFEN GRØNNEBERG AND NJÁL FOLDNES

ABSTRACT. We introduce and evaluate a new class of hypothesis testing procedures for moment structures. The methods are valid under weak assumptions and includes the well-known Satorra-Bentler adjustment as a special case. The proposed procedures applies also to difference testing among nested models. We prove the consistency of our approach. We introduce a bootstrap selection mechanism to optimally choose a p-value approximation for a given sample. Also, we propose bootstrap procedures for assessing the asymptotic robustness (AR) of the normal-theory maximum likelihood test, and for the key assumption underlying the Satorra-Bentler adjustment (Satorra-Bentler consistency). Simulation studies indicate that our new p-value approximations performs well even under severe nonnormality and realistic sample sizes, but that our tests for AR and Satorra-Bentler consistency require very large sample sizes to work well. R code for implementing our methods is provided.

1. INTRODUCTION

In testing hypotheses in psychometrics, test statistics often converge in law to a mixture of independent chi squares, under the null hypothesis of correct model specification. This paper presents novel methods for calculating p-values based on such test statistics, and a novel selection procedure aimed at identifying the best p-value among any given set of candidate p-value procedures. Although the proposed methods can be used in the general setting of moment structure inference, we here focus on the framework of structural equation modeling (SEM).

As shown in Shapiro (1983) and Satorra (1989), a large class of p-values in the context of SEM originates from convergence in distribution results (derived under the null hypothesis) for a test statistic \( T_n \) based on \( n \) observations, of the form

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

where \( \lambda = (\lambda_1, \ldots, \lambda_d)' \) consists of unknown population parameters. If \( \lambda \) was known, eq. (1) motivates the “oracle” p-value

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

Date: October 10, 2016.
The above probability is with respect to $Z_1, \ldots, Z_d$, while $T_n$ is considered fixed. In a practical setting $\lambda$ is however unknown. Let $\hat{\lambda}$ be a consistent estimator of $\lambda$, i.e., $\hat{\lambda} \xrightarrow{P} \lambda$. In the present article we propose to estimate $p_n$ by

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

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

We show that for large samples, the error originating from replacing $\lambda$ with $\hat{\lambda}$ is vanishing, so that $\hat{p}_n - p_n$ converges in probability to $0$. The estimator $\hat{p}_n$ defined above is the canonical member of a new class of estimators, obtained by grouping the $\hat{\lambda}_j$ by magnitude and replacing them by group means in order to reduce variance in $\hat{p}_n$. Although the idea behind this new class of p-value approximations is simple, we are unaware that it is found previously in the literature.

Since we introduce a whole class of p-value approximations, where no member seems to be uniformly best in all conditions, we also introduce a selector to aid the user in choosing which p-value approximation to apply. The core idea of this selector is to choose the p-value approximation whose distribution is closest to the uniform, as measured by the supremum distance. This is achieved through the non-parametric bootstrap and is seen to work very well in our simulation experiments.

The paper is structured as follows. In Section 2 we review fit statistics of moment structure models with a special emphasis on the well-known Satorra-Bentler (SB) statistic. Section 3 proposes a class of new procedures, that incorporates the SB statistic as a special case, to evaluate model fit and parameter restrictions in covariance models. We give conditions under which the estimators are consistent, which implies the fundamental p-value property of converging in distribution to a uniform distribution. In Section 4 we introduce a bootstrap procedure that selects, for a given sample, a good candidate among a list of p-value approximations. Next, in Section 5 we introduce a bootstrap test for assessing whether the normal-theory maximum likelihood test statistic may be trusted, i.e., whether asymptotic robustness holds. Also, we introduce a test for the consistency of the SB statistic, which may help decide whether the SB statistic may be trusted. Monte Carlo results on the performance of the proposed new methods are presented in Section 6. In the final section we discuss our findings and point out further directions for research. Proofs of theoretical results are presented in the appendix.

2. Fit statistics for moment structure models

Consider a $p$-dimensional vector of population moments $\sigma^\circ$. In covariance modeling, $\sigma^\circ$ consists of second-order moments, but in more general structural equation models the means may also be included in $\sigma^\circ$. The corresponding sample moment vector $s$ is assumed to converge in probability to $\sigma^\circ$, i.e., $s \xrightarrow{P} \sigma^\circ$, and
be asymptotically normal, i.e., $\sqrt{n}(s - \sigma^0) \xrightarrow{D} N(0, \Gamma)$. A structural equation model implies a certain parametrization $\theta \mapsto \sigma(\theta)$ with $\theta$ varying in a set $\Theta$. Let the free parameters in the proposed model be contained in the $q$-vector $\theta$. The model has degrees of freedom given by $d = p - q$.

The model is said to be correctly specified if there is a $\theta^0 \in \Theta$ such that $\sigma(\theta^0) = \sigma^0$. A very general class of estimators for $\theta^0$ introduced by Brown (1982, 1984) is obtained by minimising discrepancy functions $F = F(s, \sigma)$ that obey the following three conditions:

1. $F(s, \sigma) \geq 0$ for all $s, \sigma$;
2. $F(s, \sigma) = 0$ if and only if $s = \sigma$;
3. $F$ is twice continuously differentiable jointly.

That is, we consider estimators obtained as

$$\hat{\theta} = \arg\min_{\theta \in \Theta} F(s, \sigma(\theta)).$$

It is well known that the widely used normal-theory maximum likelihood (NTML) estimator is such a minimal discrepancy estimator.

Similarly, we may define the least false parameter configuration, which we denote with $\theta^\circ$. That is,

$$\theta^\circ = \arg\min_{\theta \in \Theta} F(\sigma^\circ, \sigma(\theta)).$$

Irrespective of the correctness of the model, we have $\hat{\theta} \xrightarrow{P} \theta^\circ$ under mild regularity conditions.

One, out of several mainly asymptotically equivalent (see Satorra, 1989) ways of assessing the correctness of the model is to study $T_n = nF(s, \sigma(\hat{\theta}))$. If the model is misspecified, i.e. if $\sigma^0 \neq \sigma(\theta^0)$, then $T_n \to \infty$ since $s \xrightarrow{P} \sigma^0 \neq \sigma(\theta^0)$. Under correct model specification and other assumptions presented in Shapiro (1983) and Satorra (1989), we have $T_n = \sqrt{n}(s - \sigma^0)U\sqrt{n}(s - \sigma^0) + o_P(1)$. Assuming (for simplicity) that $\Delta V \Delta$ is non-singular (see comment immediately following eq. (9) in Satorra (1989)) where $\Delta$ is the $p \times q$ derivative matrix $\partial \sigma(\theta)/\partial \theta'$ evaluated at $\theta^0$, and $V = \frac{1}{2} \frac{\partial^2 F(s, \sigma)}{\partial s \partial \sigma}$, evaluated at $(\sigma^0, \sigma^0)$, we have

$$U = V - V\Delta \{\Delta'V\Delta\}^{-1} \Delta'V.$$  

Note that $U$ has rank $d$. Since we assume $\sqrt{n}(s - \sigma^0) \xrightarrow{D} Q \sim N(0, \Gamma)$, the continuous mapping theorem now implies that $T_n \xrightarrow{D} Q'UQ$. By Theorem 1 in Box (1954), we have

$$T_n \xrightarrow{D} Q'UQ = \sum_{j=1}^{d} \lambda_j Z_j^2, \quad Z_1, \ldots, Z_d \sim N(0,1) \text{ IID},$$

where $\lambda_1, \ldots, \lambda_d$ are the $d$ non-zero eigenvalues of $U\Gamma$ under the standard scaling of the eigenvectors. That is, the parameters $\lambda$ in eq. (4) are eigenvalues of a certain matrix that depends both on the underlying distribution and on the proposed model. Note that estimating $U$ and $\Gamma$ is a standard problem in moment models which we will not discuss in technical detail. The usual estimators are based on
replacing expectations with averages of the observed data, and the true least-false parameter $\theta^*$ by the estimator $\hat{\theta}$. This is the estimator readily available in software packages such as the R package lavaan (Rosseel, 2012). We here assume that consistent estimators $\hat{U}$, $\hat{\Gamma}$ and $\hat{\lambda}$ are given. We may use the plug-in method to form $\hat{\lambda}$, so that $\hat{\lambda}$ is the $d$ non-zero eigenvalues of $\hat{U}\hat{\Gamma}$ under the standard scaling of the eigenvectors.

Note that the asymptotically distribution-free (ADF) estimator of Browne (1984), where the estimate is obtained by minimising a quadratic form whose weight matrix is the inverse of a distribution-free estimate of $\Gamma$, yields a test statistic $T_{\text{ADF}}$ whose population eigenvalues are all equal to one. Hence ADF estimation leads to consistent p-values for model fit. However, ADF estimation is unstable in small samples, and it is well-known that $T_{\text{ADF}}$ has unacceptably poor performance in small and medium samples sizes (Curran et al., 1996; Hu et al., 1992). Another test statistic with consistent p-value approximation is the residual-based test statistic (Browne, 1984, eq. 2.20), which is not of the form $T_n = nF(s, \sigma(\hat{\theta}))$ investigated in the present article. Unfortunately this statistic suffers from the same lack of acceptable finite-sample performance as $T_{\text{ADF}}$. Therefore, a more popular approach has been to use normal-theory based estimators, and to correct the test statistic for non-normality in the data. We now proceed to describe such methods.

Based on the convergence result in eq. (1), Satorra & Bentler (1994) proposed to rescale $T_n$ by dividing it by the mean value of the eigenvalues to form

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

where $\hat{c} = \frac{1}{d}\sum_{j=1}^{d} \hat{\lambda}_j$. Using $T_{\text{SB}}$ as a test statistic is a widely used SEM practice under conditions of non-normal data. Simulation studies report that $T_{\text{SB}}$ outperforms the NTML fit statistic $T_{\text{ML}}$ in such conditions, but that Type I error rates under $T_{\text{SB}}$ are seriously inflated under substantial excess kurtosis in the data (Bentler & Yuan, 1999; Nevitt & Hancock, 2004; Savalei, 2010; Foldnes & Olsson, 2015). Also, Yuan & Bentler (2010) theoretically demonstrated that $T_{\text{SB}}$ departs from a chi-square with increasing dispersion of the eigenvalues in (1).

Recently Asparouhov & Muthén (2010) proposed a test statistic that agrees with the reference chi-square distribution in both asymptotic mean and variance, obtained from $T_{\text{ML}}$ by scaling and shifting. This statistic, found to perform slightly better (Foldnes & Olsson, 2015) than a Satterthwaite type test statistic proposed by Satorra & Bentler (1994), is given by

$$T_{\text{SS}} = \sqrt{\frac{d}{\text{tr}((\hat{U}\hat{\Gamma})^2)} \cdot T_n + \frac{d(\text{tr}(\hat{U}\hat{\Gamma}))^2}{\text{tr}((\hat{U}\hat{\Gamma})^2)}}.$$
A quite different testing methodology is offered by the so-called Bollen-Stine bootstrap (Bollen & Stine 1992), which is based on the non-parametric bootstrap (Efron & Tibshirani 1994). Instead of starting with the fundamental result in eq. (1), one starts by transforming the sample observations $X_i$ into $\tilde{X}_i = \Sigma(\hat{\theta})^{-1/2} S_n^{-1/2} X_i$ for $i = 1, 2, \ldots, n$, where $S_n$ and $\Sigma(\hat{\theta})$ are the sample and model-implied covariance matrices, respectively. Noting that the model holds exactly in this transformed sample, we proceed by assuming that the transformed sample may serve as a proxy for the population from which the original sample was drawn. The Bollen-Stine p-value is now obtained by drawing bootstrap samples from the transformed sample, and calculating the proportion of bootstrap test statistics that exceed the test statistic obtained from the original sample. The validity of this approach is derived in Beran & Srivastava (1985). Nevitt & Hancock (2001) report that Bollen-Stine bootstrapping outperformed the SB scaling approach under correct model specification at realistic sample sizes, having Type I errors slightly below the nominal level. Despite the promising performance of the Bollen-Stine bootstrap, it seems to be relatively understudied. In fact, we are not aware of any later simulation study that systematically evaluates its performance relative to other robust test statistics.

Our upcoming selection methodology to be described in Section 4 will fuse the two ideas discussed above, trying to combine the strength of the convergence in eq. (1) with the power of the non-parametric bootstrap. Our tests for AR and Satorra-Bentler consistency described in Section 5 are also based on the non-parametric bootstrap. Before describing these bootstrap based methods, we return to the fundamental convergence result in eq. (1) and present new approximations for the oracle p-value.

### 3. A NEW CLASS OF P-VALUE APPROXIMATIONS

In this section we introduce and establish the consistency of a new computational technique for p-values. The proposed methodology applies as long as the null distribution of a test statistic is a weighted sum of independent chi squares and the weights can be estimated consistently. This means that the method may be used both for conventional goodness-of-fit testing of a single proposed model, and for nested model comparison tests. Consistency is established in Theorem 1 and the proof is found in Appendix A.

The convergence result in eq. (1) is only valid if the model is correctly specified. But we here note that $UT'$ is defined also when the model is misspecified, and that the number of non-zero eigenvalues is known to be $d$ from the model configuration. We may therefore speak of and estimate $\lambda = (\lambda_1, \lambda_2, \ldots, \lambda_d)'$ without knowing if the model is correctly specified. We refer to the p-value in (3) as the full p-value approximation. We will also shortly introduce other estimators by combining the
\( \hat{\lambda}_j \) in eq. (3) in ways that may reduce variability in \( \hat{p}_n \), although at the expense of consistency. This may be reasonable in situations where the full estimates are unstable, e.g. under small sample sizes and highly non-normal data. In fact, the familiar \( T_{SB} \) procedure may conceptualized as an (inconsistent) p-value approximation where the \( \lambda_j \) are replaced by the mean value of the canonical estimates, i.e.,

\[
\hat{\lambda}_{SB}^j = \frac{\sum_{d=1}^{d} \hat{\lambda}_j}{d}, \quad j = 1, \ldots, d,
\]

and clearly

\[
\hat{p}_{SB} = P \left( \sum_{j=1}^{d} \hat{\lambda}_{SB}^j Z_j^2 > T_n \right).
\]

We obtain a valid approximation as long as \( \hat{\lambda} \overset{P}{\longrightarrow} \lambda \), as the following theorem shows. Note that we make no assumptions on \( T_n \). That is, the approximation holds irrespective of the correctness of the model. Note also that we typically have

\[
\| \hat{\lambda} - \lambda \| = O_P(n^{-1/2}), \quad \text{i.e., } \sqrt{n}[\hat{p}_n - p_n] \text{ stays bounded in probability.}
\]

**Theorem 1.** 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 \( \hat{\lambda} \overset{P}{\longrightarrow} \lambda \) where \( \lambda \) only has positive elements, then \( \hat{p}_n - p_n = \| \hat{\lambda} - \lambda \| O_P(1) \), and hence, \( \hat{p}_n - p_n \overset{P}{\longrightarrow} 0 \).

**Proof.** See Appendix A. \( \square \)

We see that the \( T_{SB} \) procedure is a valid large sample approximation to \( p_n \) if \( \lambda = c \cdot (1, 1, \ldots, 1) \). If this is true in the population, Theorem 1 implies the consistency of the \( T_{SB} \) procedure. The only crucial assumption of the theorem is that each \( \lambda_j > 0 \). Recall that in goodness of fit testing in SEM, we are guaranteed \( d \) non-zero eigenvectors by the Box Theorem, see the discussion near eq. (5). Hence this assumption is innocuous.

A direct consequence of Theorem 1 is that \( \hat{p}_n \) fulfills the following property considered fundamental to p-values.

**Corollary 1.** Suppose the conditions of Theorem 1 holds. If \( T_n \overset{D}{\longrightarrow} \sum_{j=1}^{d} \lambda_j Z_j^2 \) for \( Z_1, \ldots, Z_d \sim N(0, 1) \) IID, then \( \hat{p}_n \overset{D}{\longrightarrow} U[0, 1] \).

**Proof.** Since (1) holds, it follows that \( p_n \overset{D}{\longrightarrow} U[0, 1] \). Then the corollary follows from the standard asymptotic result that if \( X_n - Y_n = o_P(1) \) and \( X_n \overset{D}{\longrightarrow} Z \) then also \( Y_n \overset{D}{\longrightarrow} Z \). \( \square \)

From our perspective of aiming at consistent p-values, the \( T_{SB} \) 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 consistent – and hence asymptotically unbiased. However, in finite samples the variability of $\hat{\lambda}$ may lead to excessive variability in $\hat{p}_n$. We therefore wish to find middle-grounds between the SB approximation and $\hat{p}_n$. This amounts to using the consistent estimates $\hat{\lambda}_j$ to calculate new weights that may reduce the sample variability of $p_n$, and at the same time reduce the effect of inconsistency in SB. Consider for instance the following split-half approximation, where the lower half of the eigenvalues are replaced by their mean value, and likewise for the upper half of the eigenvalues.

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

where

$$\hat{\lambda}_1 = \cdots = \hat{\lambda}_{[d/2]} = \frac{1}{[d/2]} \sum_{j=1}^{[d/2]} \lambda_j$$

and

$$\hat{\lambda}_{[d/2]+1} = \cdots = \hat{\lambda}_d = \frac{1}{d-[d/2]} \sum_{j=[d/2]+1}^{d} \lambda_j.$$  

This procedure allows the p-value approximation an additional degree of freedom compared to the SB statistic, where all eigenvalues are estimated to be equal to each other. A whole class of middle-grounds between the full $\hat{p}_n$ and $\hat{p}_{n,SB}$ can be defined as follows. 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

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

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

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

An extension of the above framework is tests that assess nested hypotheses in SEM. Due to its great practical importance, we here include a short discussion on this special case. We again focus on the statistic $T_n$, since this statistic is typically asymptotically equivalent to other tests of interests, as described in Satorra (1989).

Following Satorra (1989), let $H : \sigma = \sigma(\theta), \theta \in \Theta$ and $H_0 : \sigma = \sigma(\theta), \theta \in \Theta_0$ where $\Theta_0 = \{\theta \in \Theta : a(\theta) = 0\}$ for some continuously differentiable function $a$. We assume that the matrix $\frac{\partial a(\theta)}{\partial \theta}$ has full row rank, say $m$. We let

$$\hat{\theta} = \arg\min_{\theta \in \Theta} F(s, \sigma(\theta)), \quad \tilde{\theta} = \arg\min_{\theta \in \Theta_0} F(s, \sigma(\theta))$$
and $T_n = nF(s, \sigma(\theta))$ and $\tilde{T}_n = nF(s, \sigma(\tilde{\theta}))$. Under $H_0$ and the conditions of Lemma 1 (iv) in Satorra (1989) we have

$$T_n = \sqrt{n}(s - \sigma^0)'U\sqrt{n}(s - \sigma^0) + o_P(1)$$

and

$$\tilde{T}_n = \sqrt{n}(s - \sigma^0)'\tilde{U}\sqrt{n}(s - \sigma^0) + o_P(1),$$

for matrices $U$ and $\tilde{U}$ following the formula of eq. (4) under $H$ and $H_0$, respectively. Using the basic algebraic fact that $x'(A + B)x = x'Ax + x'Bx$ we conclude that the difference statistic is of the form

$$\tilde{T}_n - T_n = \sqrt{n}(s - \sigma^0)'U_d\sqrt{n}(s - \sigma^0) + o_P(1),$$

where $U_d = \tilde{U} - U$ has rank $m$.

By the continuous mapping theorem, the convergence $\sqrt{n}(s - \sigma^0) \xrightarrow{D} N(0, \Gamma)$, and Theorem 1 in Box (1954), we therefore have that

$$(7) \tilde{T}_n - T_n \xrightarrow{D} \sum_{j=1}^{m} \alpha_j Z_j^2, \quad Z_1, \ldots, Z_m \sim N(0,1) \text{ IID},$$

where $\alpha_1, \ldots, \alpha_m$ are the $m$ non-zero eigenvalues of $U_d\Gamma$.

Distribution-free consistent estimators $\hat{U}_d$ and $\hat{\Gamma}$ for $U_d$ and $\Gamma$ are found and discussed in Satorra & Bentler (2001), and we do not review them here. Again the standard estimators can be found in software such as the R package lavaan. One then forms $\hat{\alpha} = (\hat{\alpha}_1, \ldots, \hat{\alpha}_m)'$ equal to the $m$ largest eigenvectors of $\hat{U}_d\hat{\Gamma}$ and calculates the full p-value approximation

$$\hat{p}_n = P \left( \sum_{j=1}^{m} \hat{\alpha}_j Z_j^2 > \tilde{T}_n - T_n \right).$$

We remark that a single equality constraint, say, $\beta_{i,j} = 0$, can be treated as special case of the above framework. In this case, the number of restrictions is 1, and hence the limiting distribution in eq. (7) is a scaled $\chi^2_1$. The SB and the proposed p-value approximations then coincide exactly. Note that Theorem 1 implies that these procedures are consistent.

4. A SELECTION ALGORITHM FOR P-VALUE APPROXIMATIONS

The framework of the last section leads to several competing p-value approximations, and we next introduce a way of selecting among these. Our selector is inspired by Beran & Srivastava (1983), the Bollen-Stine bootstrap (Bollen & Stine, 1992), and the non-parametric focused information criterion of Julium & Hjort (2016, forthcoming).

We wish to select the p-value approximation $\hat{p}_n$ whose distribution is closest to the uniform distribution under the null hypothesis. We formalize this by estimating
the supremum distance between the cumulative distribution function of \( \hat{p}_n \) under
the null hypothesis and the uniform distribution, i.e. we approximate
\[
D_n = \sup_{0 \leq x \leq 1} |P_{H_0}(\hat{p}_n \leq x) - x|
\]
for each p-value approximation, and select the method with the least value of \( D_n \).
The probability \( P_{H_0} \) is the probability measure induced by the data-generating
distribution that is closest to fulfilling the null hypothesis compared to the true
data-generating mechanism, which we let be the data generating distribution of
\( \Sigma(\theta^o)^{1/2} \Sigma^{-1/2} X_i \), where \( \Sigma \) is the true covariance matrix. Under \( P_{H_0} \), we know
that p-values should be uniformly distributed. If we consider asymptotically consis-
tent p-values, minimizing \( D_n \) will mean that we choose the approximation whose
convergence has been best achieved at our sample-size \( n \).

The approximation to \( D_n \) is done via the non-parametric bootstrap, based on
the transformed sample \( \tilde{X}_i = \Sigma(\hat{\theta})^{1/2} S_n^{-1/2} X_i \) for \( i = 1, 2, \ldots, n \), as described in
Algorithm 1. The supremum in Algorithm 1 is the test statistic of the Kolmogorov-
Smirnov test, which is implemented in most statistical software packages. Formally,
what we do is to use the empirical distribution function \( \tilde{P}_n \) of \( (\tilde{X}_i) \) as an approxi-
imation to \( P_{H_0} \), and approximate this probability through re-sampling. We then
plug this approximation into \( D_n \) to generate \( \hat{D}_n \) for each p-value approximation.

We note that we may use this selector among any p-value approximation for
hypothesis testing in moment structures, and not just the suggestions in Section 3.
Also, \( D_n \) is only one out of many possible success criteria. One could also investigate
the mean square error of the approximation, or the distance from \( P_{H_0}(\hat{p}_n \leq x) \) to
\( x \) at a particular point \( x \). In our simulations, the performance of \( D_n \) as a selection
criterion was overall satisfactory.

**Algorithm 1** Selection algorithm
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: \( \text{boot.sample} \leftarrow \text{Draw with replacement from transformed sample } \tilde{X}_i \)
5: for \( l \in 1, \ldots, L \) do
6: \( \hat{p}_{n,l} \leftarrow \text{based on boot.sample} \)
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 \( \arg\min_{1 \leq l \leq L} \hat{D}_{B,n,l} \)
13: end procedure
5. Hypothesis tests for Satorra-Bentler Consistency and Asymptotic Robustness

In this section we propose a bootstrap procedure for testing Satorra-Bentler consistency, that is, that all non-zero eigenvalues are equal. This also leads naturally to a test for asymptotic robustness (AR), that is, that all non-zero eigenvalues are equal to 1. Such tests may help a practitioner to decide whether it is advisable to apply the NTML test, the SB test, or to instead use the Bollen-Stine bootstrap or the new procedures proposed in the present article.

There is a substantial body of theoretical literature on AR (Shapiro, 1987; Browne & Shapiro, 1988; Amemiya & Anderson, 1990; Satorra & Bentler, 1990), where exact conditions are given that involve certain relationships between $\Gamma$ and $\Delta$ that must hold for $T_{ML}$ to retain its asymptotic chi-square distribution under non-normality. However, these conditions are hard to check in practice, and currently no practical procedure exist for verifying asymptotic robustness in a real-world setting (Yuan, 2005, p. 118). Similarly, we unaware of the existence of tests for SB consistency. This lack of tests might be due to the fact that testing statements concerning the eigenvalues of $\Gamma$ involves testing statements about high moment properties of a distribution. Without detailed parametric assumptions on the data it seems very difficult to construct tests that perform well in small-sample situations. It is therefore expected that our proposed procedures will require a large sample size to attain Type I error rates close to the nominal level. This is confirmed to be the case in the simulation experiment in Section 6.3.

The proposed bootstrap test is summarized in Algorithm 2 and is inspired by Section 4.2 in Beran & Srivastava (1985). A proof of its consistency, which we do not provide, seems to require a non-trivial extension of the theory contained in Beran (1984); Beran & Srivastava (1985). We note that an important difference between our suggested test and the procedures in Section 4.2 in Beran & Srivastava (1985), who work with eigenvalues of empirical covariance (i.e., symmetric) matrices, is that $\Gamma$ is typically not symmetrical.

Let $E$ be the matrix of normalised (complex) eigenvectors sorted by descending values of the eigenvalues $\lambda$ of $\Gamma$. We have $\Gamma = E \Delta E^{-1}$ (Meyer, 2000, p.514) where $\Delta = \begin{pmatrix} \Delta_d & 0 \\ 0 & 0 \end{pmatrix}$, and where $\Delta_d$ is the diagonal matrix with elements $\lambda_1, \ldots, \lambda_d$. Define

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

where $c$ denotes the mean value of the eigenvalues $\lambda_1, \ldots, \lambda_d$. We propose the following bootstrap procedure. Let $\hat{A}$ be estimated from the original sample, by replacing $E, \Delta, c$ with $\hat{E}, \hat{\Delta}, \hat{c}$. For each bootstrap sample drawn from the original
sample, we calculate $\hat{U}_{\text{boot}}^\cdot \hat{\Gamma}_{\text{boot}}^\cdot$ and form the matrix

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

The crucial observation is now that $W_n^*$ converges to a matrix for which the null-hypothesis is true, that is, whose non-zero eigenvalues are all equal. To see this, note that

$$W_n^* \xrightarrow{P \to \infty} cE \begin{pmatrix} \Delta_d^{-1/2} & 0 \\ 0 & 0 \end{pmatrix} E^{-1} \cdot U \Gamma \cdot E \begin{pmatrix} \Delta_d^{-1/2} & 0 \\ 0 & 0 \end{pmatrix} E^{-1}$$

$$= cE \begin{pmatrix} \Delta_d^{-1/2} & 0 \\ 0 & 0 \end{pmatrix} E^{-1} E \Delta E^{-1} E \begin{pmatrix} \Delta_d^{-1/2} & 0 \\ 0 & 0 \end{pmatrix} E^{-1}$$

$$= cE \begin{pmatrix} \Delta_d^{-1/2} & 0 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} \Delta_d & 0 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} \Delta_d^{-1/2} & 0 \\ 0 & 0 \end{pmatrix} E^{-1} = E \begin{pmatrix} cI_d & 0 \\ 0 & 0 \end{pmatrix} E^{-1},$$

where the last matrix has $d$ non-zero eigenvalues equal to $d$. In the bootstrap sample, the $d$ largest eigenvalues of $W_n^*$ is then computed as $\hat{\lambda}_{\text{boot}}$. This process is repeated many times, and we get realizations $\hat{\lambda}_k, \hat{\lambda}_{\text{boot}}$, giving us information about the sampling variability of the estimated eigenvalues under the null hypothesis of identical eigenvalues.

The above procedure may also be adapted to test for asymptotic robustness of the NTML statistic $T_{\text{ML}}$, that is, whether $\lambda_j = 1$ for all $j = 1, \ldots, d$. By setting $c = 1$ in eq. [8], Algorithm [2] then produces a p-value for the test of consistency of $T_{\text{ML}}$ based testing. Since this test does not need to estimate $c$, it should converge to the correct level I error rate slightly faster than the general case. The test statistic that is bootstrapped is then

$$W_n^* = \hat{A}_1 U_{\text{boot}}^\cdot \hat{\Gamma}_{\text{boot}}^\cdot \hat{A}_1.$$  

where $\hat{A}_1$ is the estimator of $A_1 = E \begin{pmatrix} \Delta_d^{-1/2} & 0 \\ 0 & 0 \end{pmatrix} E^{-1}$.

We suppose that an extension of Corollary 4 in Beran & Srivastava (1985) holds also in our setting. That corollary requires the test statistic $h(\lambda)$ to be non-negative and zero under the null hypothesis, and that it has partial derivatives that are zero under the null hypothesis and that its double derivative matrix is positive definite under the null hypothesis. The additional restriction that also the partial derivatives vanish under the null-hypothesis means we must consider two different test statistics, adapted from the examples in Section 4.3 in Beran & Srivastava (1985). For asymptotic robustness, this holds for $h_{AR}(\lambda) = d \log[d^{-1} \sum_{j=1}^d \lambda_j] - \log[\prod_{j=1}^d \lambda_j]$. For Satorra-Bentler consistency, this holds for $h_{SB}(\lambda) = \log[(\lambda_1 + \lambda_d)^2] - \log[4 \lambda_1 \lambda_d]$. Algorithm [2] summarizes this discussion.
Algorithm 2 Bootstrap testing for Satorra-Bentler consistency and Asymptotic Robustness

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

6. Monte Carlo evaluations

In this section we evaluate the proposed procedures by Monte Carlo methods. We first evaluate our new class of p-value approximations in the setting of goodness-of-fit testing for a single model. Specifically, two members of this class are evaluated, $\hat{p}_n$ and $\hat{p}_{n,\text{half}}$, referred to as the full and half eigenvalue approximations, respectively. We then consider chi-square difference testing for two nested models. In both cases we evaluate the selection procedure in Algorithm 1 where the candidates for selection are SB and the full and half eigenvalue approximations. Finally, we evaluate the empirical performance of the proposed bootstrap test for SB consistency and for AR. We remark that we here limit ourselves to study the empirical performance of the procedures when it comes to controlling type I error rates, leaving the topic of power for future studies.

Our model is the political democracy model discussed by Bollen in his textbook (Bollen, 1989), see Figure 1 where the residual errors are not depicted for ease of presentation. There are four measures of political democracy measured twice (in 1960 and 1965), and three measures of industrialization measured once (in 1960). The unconstrained model $M_1$ has $d = 35$ degrees of freedom. For nested model testing, we also consider a constrained model $M_0$, nested within $M_1$, with $d = 46$ degrees of freedom, which impose ten equalities among unique variances and residual covariances, and one equality between two factor loadings.
Model estimation and eigenvalues were computed using the R package lavaan \cite{Rosseel2012}, while the p-values of type $\hat{p}$ were calculated with the imhof procedure in the package CompQuadForm \cite{Duchesne2010}. In each simulation cell we generated 2000 samples. For each sample, 1000 bootstrap samples were drawn.

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

|              | Y_1 | Y_2 | Y_3 | Y_4 | Y_5 | Y_6 | Y_7 | Y_8 | Y_9 | Y_10 | Y_11 |
|--------------|-----|-----|-----|-----|-----|-----|-----|-----|-----|------|------|
| Distribution 2 | 1.87| 1.59| 1.49| 1.44| 1.43| 1.42| 1.38| 1.36| 1.35| 1.34| 1.31 |
| Distribution 3 | 4.16| 3.24| 2.88| 2.82| 2.70| 2.67| 2.51| 2.41| 2.35| 2.31| 2.16 |

**Table 1.** Eigenvalues $\lambda_i$, 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.
Table 2. Type I error rates for testing model $M_1$. 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 $p_n$.

| Distribution | n  | NTML | SB  | SS  | BOST | EFULL | EHALF | SEL | ORAC |
|--------------|----|------|-----|-----|------|-------|-------|-----|------|
| Normal       | 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|
| Distribution 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|
| Distribution 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 3. Choice proportions for selection algorithm, testing model $M_1$. SB=Satorra-Bentler. EHALF=half eigenvalue approximation, $\hat{p}_{n,\text{half}}$. EFULL= Full eigenvalue approximation, $\hat{p}_n$.

| Distribution | n  | SB  | EHALF | EFULL |
|--------------|----|-----|-------|-------|
| Normal       | 100| 0.054| 0.931 | 0.015 |
|              | 300| 0.448| 0.516 | 0.036 |
|              | 900| 0.507| 0.263 | 0.231 |
| Distribution 3| 100| 0.001| 0.865 | 0.135 |
|              | 300| 0.050| 0.894 | 0.055 |
|              | 900| 0.153| 0.783 | 0.063 |
| Distribution 3| 100| 0.000| 0.449 | 0.551 |
|              | 300| 0.001| 0.733 | 0.267 |
|              | 900| 0.004| 0.846 | 0.150 |

6.1. **Goodness-of-fit testing for $M_1$.** Three population distributions were considered. Distribution 1 was a multivariate normal distribution. The non-normal distributions were generated using the transform of Vale & Maurelli (1983), with Distribution 2 having univariate skewness 1 and kurtosis 7, and Distribution 3 having skewness 2 and kurtosis 21. These distributional characteristics are the same.
as those used in the influential study by Curran et al. (1996), and replicated in the bootstrap study by Nevitt & Hancock (2001). The "oracle" eigenvalues associated with Distribution 2 and 3 are given in Table 1 numerically calculated from very large samples, where we clearly see that the values are quite spread out, and that the spread increases when we move from Distribution 2 to Distribution 3. Note that under the Distribution 1, we have $\lambda = (1,1,\ldots,1)'$.

Three sample sizes $n$ were used: 100, 300 and 900. Hence the resulting full factorial design has nine conditions. In each sample we calculated p-values associated with the established test statistics associated with normal-theory maximum likelihood ratio (NTML), Satorra-Bentler scaling (SB), the scale-and-shifted statistic (SS) and the Bollen-Stine (BOST) test. Also, we calculated in each sample the full eigenvalue approximation $\hat{p}_n$ (EFULL) and the split-half eigenvalue estimation $\hat{p}_{n,\text{half}}$ (EHALF). The selection algorithm (SEL) p-value was calculated using a candidate set with members SB, EHALF and EFULL, and using $\hat{D}_n$ as criterion function. Finally, the oracle (ORAC) p-value $p_n$ was calculated, using the values in Table 1. This allows us to evaluate how well the asymptotic result in eq.(1) applies in finite-sample conditions.

In Table 2 we present Type I error rates at the the 5% significance level. As expected, NTML becomes inflated when data is non-normal. The mean-scaling of SB reduces the inflation, but with non-normal data and small sample sizes, Type I error rates are still higher than 10%. The scaled-and-shifted statistic on the other hand, leads to rejection rates much lower than the nominal 5%. These findings are in accord with Foldnes & Olsson (2015). The Bollen-Stine bootstrap test performs better than SB and SS, coming close to the nominal level even for highly non-normal data and medium sample size. Among the new p-value approximations, it is the middle-ground approximation EHALF that performs the best. While EFULL yields far too low rejection rates with non-normal data. EHALF as well as BOST with non-normal data. The selection algorithm SEL also performs generally well, on par with EHALF and BOST. It is notable that for normal data, SEL outperforms NTML. Table 3 presents the selection proportions for SEL in each of the nine conditions. It is seen that the selection algorithm wisely chooses EHALF in the majority of conditions. It is however unexpected that SEL chooses EFULL in 55% of the samples under Distribution 3 and $n = 100$, given the poor performance of EFULL in that condition, with a 1% rejection rate. The final column in Table 2 gives the oracle solution, and demonstrates that the asymptotic result in eq.(1) is far from realized at $n = 100$ under Distribution 3.

6.2. Testing nested models. The chi-square difference test has 11 degrees of freedom, and the corresponding 11 oracle eigenvalues for Distribution 2 and 3 are given in Table 4. The spread in eigenvalues is substantial, especially for Distribution 3.
Table 4. 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.

| Distribution | $n$ | ML  | SB  | BOST | EFULL | EHALF | SEL  | ORAC |
|--------------|-----|-----|-----|------|-------|-------|------|------|
| Normal       | 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 |
| Distribution 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 |
| Distribution 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 5. 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 $p_n$.

Rejection rates observed at the nominal 5% level of significance are reported in Table 5. Again, the NTML statistic is inflated by non-normality in the data, a tendency only partially corrected for by SB. For instance, under the most harsh condition, with Distribution 3 and $n = 100$, SB rejection rates are 22%, far better than the 91% obtained with NTML. But in this condition, as in all conditions, BOST performs better, with a rejection rate of 13%. However, the new procedure EFULL performs still better in this condition, while the selection algorithm is only slightly worse than BOST. Overall EFULL outperforms the other test statistics, including SB and BOST. EHALF, which was found to have best performance in the non-nested case, does not perform as well as EFULL in the nested case. The selection algorithm SEL also performs well, with better performance than SB and BOST in most conditions, and only slightly worse then EFULL. The selection proportions are given in Table 6, where EHALF is unexpectedly found to be the
most chosen procedure, despite the slightly better performance of EFULL in most conditions.

| Distribution | n  | SB  | EHALF | EFULL |
|--------------|----|-----|-------|-------|
| Normal       | 100| 0.601 | 0.357  | 0.042 |
|              | 300| 0.672 | 0.205  | 0.122 |
|              | 900| 0.593 | 0.091  | 0.316 |
| Distribution 3| 100| 0.116 | 0.714  | 0.170 |
|              | 300| 0.209 | 0.662  | 0.128 |
|              | 900| 0.263 | 0.595  | 0.142 |
| Distribution 3| 100| 0.012 | 0.663  | 0.325 |
|              | 300| 0.059 | 0.725  | 0.215 |
|              | 900| 0.104 | 0.714  | 0.182 |

Table 6. Choice proportions for selection algorithm, nested models. SB=Satorra-Bentler. EHALF=half eigenvalue approximation, \( \hat{p}_{n, \text{half}} \). EFULL= Full eigenvalue approximation, \( \hat{p}_n \).

6.3. **Tests for AR and for SB consistency.** To evaluate Type I error rates of the SB consistency and AR tests proposed in Algorithm 2, we simulated multivariate normal data for the Bollen model. Under normal data both AR and SB consistency holds. We simulated 2000 samples for sample sizes \( n = 200, 400, 800 \) and 2000. For each sample 1000 bootstrap samples were drawn. The rejection rates are given in Table 7 and clearly demonstrates that 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.

7. **Discussion**

This paper deals with the fundamental problem of hypothesis testing in moment structure models. We present new insight and practically applicable statistical methodology for SEM and related models.

Some of our conclusions may seem surprising, as they go against what is often taught in standard courses on SEM. For example, the simulation summarized in Table 2 shows that our selector can have better finite sample performance than
the NTML test also when data are exactly normal. Since this paper have focused
exclusively on Type I error, “better” here means having a rejection rate closer to
the nominal one.

Since this conclusion may seem counterintuitive, it is worth pausing and consid-
ering what the NTML does. Firstly, we must keep in mind that the NTML is a
test based on asymptotic theory, also when data are exactly normally distributed.
That is, the Type I error rate of an NTML test at level $\alpha$ converges to $\alpha$ under
normality, and for the model considered in Table 2 convergence is still not quite
achieved for $n = 100$. Secondly, we note that under normality, NTML calculates
the oracle p-value exactly. That is, it is the ultimate approximation to the oracle
test, which has a rejection rate of 7.7%. Hence, the NTML has only one source of
approximation error: the validity of the fundamental convergence of the oracle.

All methods considered in this paper – with the important exception of the
Bollen-Stine bootstrap and the selector – tries to approximate the oracle, and
thereby introducing another source of approximation error. Let us say they are
oracle-based. Except the NTML, which calculates the oracle perfectly – but only
under exact normality, the oracle-based methods have varying degrees of success in
their approximation. Strictly speaking, oracle-based methods should be judged on
whether they manage to achieve what they set out to do: approximate the oracle.
But that is not the success criteria of interests to the user: When a level $\alpha$ test
is employed, the Type I error rate should be very close to $\alpha$. As is clear from our
simulation studies, this may not be the case even when using the actual oracle.
When oracle-based tests have Type I error rate considerably closer to $\alpha$ than the
oracle, it is tempting to say that they are performing well. This temptation should
be avoided, as the deviation in Type I error compared to the oracle is then solely
due to chance variations caused by the estimation of $\lambda$.

The selector overcomes this hurdle by being anchored not in the fundamental
convergence of the oracle, but by transforming the data to a setting where the null
hypothesis holds. It is then known that a correct p-value is uniformly distributed,
i.e., the Type I error rate of a test with level $\alpha$ is to be exactly $\alpha$. It is this
anchoring that allows us to search for the procedure which best achieves this goal,
without having to compare our methods to the finite sample performance of the
oracle. And so when the selector has a Type I error rate close to the nominal, it
is by design, and not solely due to chance variations. This is a property shared
with the Bollen-Stine bootstrap procedure, but the Bollen-Stine procedure rests on
the quality of the approximation of the empirical distribution function compared to
the data’s actual distribution function. So do we, since we use the non-parametric
bootstrap in our selector, but we are able to combine the fundamental convergence
of the oracle with the non-parametric bootstrap. We have seen that this allows us
to combine the strengths of both methods.
Let us return to the NTML, and look at the proposed methods from a slightly different perspective that elaborates on the above. It is well-known that the NTML usually has a much too high Type I error rate under non-normality. The major source of the mismatch between nominal and actual Type I error rate is that the NTML need not be a consistent approximation to the oracle. The NTML can be seen as estimating $\lambda$ always by the constant $(1, \ldots, 1)'$. When $\lambda$ is far away from $(1, \ldots, 1)'$, NTML performs poorly. And for a user, it typically performs poorly in a particularly bad way: even when a hypothesized theory holds, the NTML will most likely reject it.

The Satorra-Bentler test has previously been reported to have inflated Type I error rates under non-normality, and this behaviour is also observed in simulations in the present paper. This is mainly due to two reasons: firstly, it may be that the Satorra-Bentler procedure is inconsistent, i.e. $\lambda$ has variation among its elements. While inconsistency is an asymptotic property, which may seem irrelevant in small samples, it does mean that the procedure does not aim to estimate what the user wants, and may therefore be reflected also in small-sample situations when the procedure is used uncritically. Secondly, the Satorra-Bentler procedure estimates $\lambda$, and the variability of the resulting p-value approximation may give inflated Type I errors even when the procedure is consistent.

These two problems, consistency and finite sample approximation error, are shared also by our suggested p-value approximations. However, the contextual framework presented in the present paper allows us to argue about balancing these issues, and selecting among competing approximations. This perspective may lead to further insight in future research, and has already led to our proposed selector.

We note that while $\hat{p}_n$ and $\hat{p}_{n,\text{half}}$ can be computed just as fast as the Satorra-Bentler test statistics, both the selector and the Bollen-Stine bootstrap procedure takes considerable more computation time. Our simulation experiments indicate that the selector and the Bollen-Stine bootstrap are comparable in performance, but that the selector works slightly better, especially in small sample situations. Our recommendations to practitioners are therefore clear: use the selector in small sample situations, and use the selector or the Bollen-Stine bootstrap in medium sample situations. In large sample situations, consistent p-value approximation gives similar answers. Since the assumptions underlying asymptotic robustness and Satorra-Bentler consistency rests on delicate properties of high order moments that can only be properly tested in large sample situations, we do not recommend using the NTML nor the Satorra-Bentler statistic without assessing its performance with the selector. In many cases, the Satorra-Bentler statistic will be the superior test, but it is difficult to know this without using techniques such as the re-sampling methods underlying the selector.
With current and future computers containing multiple units that can perform computation simultaneously (multi-core central processing units and multi-core graphical processing units supporting general purpose calculations), using the selector does not take much time to run. In our prototype implementation in the scripting language R (which means our code is not compiled, and therefore slow), it takes a few additional minutes compared to standard p-value approximations that we have seen often performs considerably worse. Considering the enormous amount of time and effort many researchers use in gathering and analyzing data, the extra time spent on using the selector is vanishing in comparison.

Applied researchers are often personally interested in controlling Type I error as well as possible, as their research hypothesis is often the null hypothesis. If they use testing procedures, such as the NTML with non-normal data, where the Type I error is seriously inflated, this is to their disadvantage. This point is also connected to the use of pragmatic fit indices available in the literature. The p-value approximations discussed in this paper are all based on solid statistical theory. The ad-hoc nature of some of these fit indices, with somewhat arbitrary cut-off points being interpreted in various ways, are not based on statistical theory.

Finally, we mention that the ideas contained in this paper can be generalized in several directions, including SEM with ordinal variables and in multi-group settings. Also, additional simulation experiments should be performed on the proposed methods, such as power studies, allowing the selector more options, and experimenting with different selection criterias.

References

Amemiya, Y. & Anderson, T. (1990). Asymptotic chi-square tests for a large class of factor analysis models. *The Annals of Statistics*, 1453–1463.

Asparouhov, T. & Muthén, B. (2010). Simple second order chi-square correction. Retrieved from Mplus website: http://www.statmodel.com/download/WLSMV_new_chisq1.

Bentler, P. M. & Yuan, K.-H. (1999). Structural equation modeling with small samples: Test statistics. *Multivariate Behavioral Research* 34, 181–197.

Beran, R. (1984). Bootstrap methods in statistics. *Jahresbericht der Deutschen Mathematiker-Vereinigung*, 14–30.

Beran, R. & Srivastava, M. S. (1985). Bootstrap tests and confidence regions for functions of a covariance matrix. *The Annals of Statistics*, 95–115.

Bollen, K. A. (1989). *Structural equations with latent variables*. New York: Wiley.

Bollen, K. A. & Stine, R. A. (1992). Bootstrapping goodness-of-fit measures in structural equation models. *Sociological Methods & Research* 21, 205–229.
Box, G. (1954). Some theorems on quadratic forms applied in the study of analysis of variance problems. 1. effect of inequality of variance in the one-way classification. *The Annals of Mathematical Statistics* **25**, 290–302.

Browne, M. & Shapiro, A. (1988). Robustness of normal theory methods in the analysis of linear latent variable models. *British Journal of Mathematical and Statistical Psychology* **41**, 193–208.

Browne, M. W. (1982). Covariance structures. *Topics in applied multivariate analysis*, 72–141.

Browne, M. W. (1984). Asymptotically distribution-free methods for the analysis of covariance structures. *British Journal of Mathematical and Statistical Psychology* **37**, 62–83.

Curran, P. J., West, S. G. & Finch, J. F. (1996). The robustness of test statistics to nonnormality and specification error in confirmatory factor analysis. *Psychological Methods* **1**, 16–29.

Duchesne, P. & de Micheaux, P. L. (2010). Computing the distribution of quadratic forms: Further comparisons between the liu-tang-zhang approximation and exact methods. *Computational Statistics and Data Analysis* **54**, 858–862.

Efron, B. & Tibshirani, R. J. (1994). *An introduction to the bootstrap*. CRC press.

Foldnes, N. & Olsson, U. H. (2015). Correcting too much or too little? the performance of three chi-square corrections. *Multivariate Behavioral Research* **50**, 533–543.

Hu, L.-T., Bentler, P. M. & Kano, Y. (1992). Can test statistics in covariance structure analysis be trusted? *Psychological Bulletin* **112**, 351–62.

Jullum, M. & Hjort, N. L. (2016, forthcoming). Parametric or nonparametric: The fic approach. *Statistica Sinica* .

Laury-Micoulaut, C. (1976). The n-th centered moment of a multiple convolution and its applications to an intercloud gas model. *Astronomy and Astrophysics* **51**, 343–346.

Meyer, C. D., ed. (2000). *Matrix Analysis and Applied Linear Algebra*. Philadelphia, PA, USA: Society for Industrial and Applied Mathematics.

Nevitt, J. & Hancock, G. (2001). Performance of Bootstrapping Approaches to Model Test Statistics and Parameter Standard Error Estimation in Structural Equation Modeling. *Structural Equation Modeling: A Multidisciplinary Journal* **8**, 353–377.

Nevitt, J. & Hancock, G. (2004). Evaluating small sample approaches for model test statistics in structural equation modeling. *Multivariate Behavioral Research* **39**, 439–478.

Rosseel, Y. (2012). lavaan: An r package for structural equation modeling. *Journal of Statistical Software* **48**, 1–36.
Satorra, A. (1989). Alternative test criteria in covariance structure analysis: A unified approach. *Psychometrika* **54**, 131–151.

Satorra, A. & Bentler, P. M. (1990). Model conditions for asymptotic robustness in the analysis of linear relations. *Computational Statistics & Data Analysis* **10**, 235–249.

Satorra, A. & Bentler, P. M. (1994). Corrections to test statistics and standard errors in covariance structure analysis. In *Latent variable analysis: applications for developmental research*, A. V. Eye & C. Clogg, eds., chap. 16. Newbury Park, CA: Sage, pp. 399–419.

Satorra, A. & Bentler, P. M. (2001). A scaled difference chi-square test statistic for moment structure analysis. *Psychometrika* **66**, 507–514.

Savalei, V. (2010). Small sample statistics for incomplete nonnormal data: extensions of complete data formulae and a monte carlo comparison. *Structural Equation Modeling: A Multidisciplinary Journal* **17**, 241–264.

Shapiro, A. (1983). Asymptotic distribution theory in the analysis of covariance structures - a unified approach. *South African Statistical Journal* **17**, 33–81.

Shapiro, A. (1987). Robustness properties of the mdf analysis of moment structures. *South African Statistical Journal* **21**, 39–62.

Vale, C. & Maurelli, V. (1983). Simulating nonnormal distributions. *Psychometrika* **48**, 465–471.

Yuan, K. (2005). Fit indices versus test statistics. *Multivariate Behavioral Research* **40**, 115–148.

Yuan, K.-H. & Bentler, P. M. (2010). Two simple approximations to the distributions of quadratic forms. *British Journal of Mathematical and Statistical Psychology* **63**, 273–291.

**Appendix A. Proof of Theorem**

*Proof of Theorem 1.* By the mean value theorem, there is a sequence of random variables \(0 \leq r_n \leq 1\) so that 
\[
\tilde{p}_n = 1 - H(T_n; \tilde{\lambda}) = 1 - H(T_n; \lambda) + R_n = p_n + R_n
\]
where \(R_n = \sum_{j=1}^d (\hat{\lambda}_j - \lambda_j)H_j(T_n; \lambda + r_n(\hat{\lambda} - \lambda))\) and 
\[
H_j(q, (l_1, \ldots, l_d)') = \frac{\partial H_j(q, (l_1, \ldots, l_d)')}{\partial l_j}.
\]
The statement of the theorem therefore holds if we show that 
\[
H_j(T_n; \lambda + r_n(\hat{\lambda} - \lambda)) = O_p(1).
\]
To show this, we calculate \(H_j\). The cumulative distribution function of 
\[
S = \sum_{j=1}^d \lambda_j Z_j^2
\]
is 
\[
H_S(q) = \int_0^q h_S(s) \, ds
\]
where \(h_S\) is the density of \(S\). Denote the density of \(\lambda_j Z_j^2\) by \(h_j(z)\). Since \((Z_j)_{j=1}^d\) contains independent variables, so does \((\lambda_j Z_j^2)_{j=1}^d\). Hence \(h_S\) is given by \(d\)-times convolution, i.e. apply the well-known convolution formula iteratively, and see that

\[
g_S(s) = \int_{\mathbb{R}} \cdots \int_{\mathbb{R}} \left[ \prod_{j=2}^d h_j(x_{j-1}) \right] h_1 \left( s - \sum_{j=1}^{d-1} x_j \right) \, dx_1 \cdots dx_{r-1},
\]

(9)
We wish to calculate
\begin{equation}
\frac{\partial}{\partial \lambda_j} h_S(q) = \int_0^q \frac{\partial}{\partial \lambda_j} h_S(s) \, ds.
\end{equation}

It turns out that \(\frac{\partial}{\partial \lambda_j} h_S\) is a weighted sum of densities, which implies that \(\frac{\partial}{\partial \lambda_j} h_S\) is a weighted sum of cumulative distribution functions that is easy to bound uniformly.

We now show this by calculating \(\frac{\partial}{\partial \lambda_j} h_S\). Since summation is commutative, the distribution of \(\sum_{j=1}^r \lambda_{\pi(j)} Z_{\pi(j)}^2\) is the same for any permutation \(\pi(1), \ldots, \pi(r)\) of \(\{1, \ldots, r\}\). We may therefore, without loss of generality, assume that \(j = d\). Using eq. (9), we have
\begin{equation}
\frac{\partial}{\partial \lambda_d} h_S(s) = \int_\mathbb{R} \cdots \int_\mathbb{R} \left\{ \frac{\partial}{\partial \lambda_d} h_d(x_d-1) \right\} \left( \prod_{j=2}^{d-1} h_j(x_j-1) \right) h_1(s - \sum_{j=1}^{d-1} x_j) \, dx_1 \cdots dx_{d-1}
\end{equation}

We hence need to calculate \(\frac{\partial}{\partial \lambda_d} h_d(x_d-1)\). Since \(\lambda_j Z_j^2\) is a linear transformation of \(Z_j^2 \sim \chi^2_1\), we have \(h_j(z) = h_{\chi^2}(z/\lambda_j)/\lambda_j\) where \(h_{\chi^2}(z) = \frac{z^{\frac{1}{2}}}{{\sqrt{\pi}}} e^{-z/2} I\{z \geq 0\}\) is the density of \(Z_j^2 \sim \chi^2_1\).

We have \(\frac{\partial}{\partial \lambda_d} h_d(z) = \frac{\partial}{\partial \lambda_d} \lambda_1^{-1} h_{\chi^2}(z/\lambda_d) = -\lambda_1^{-2} h_{\chi^2}(z/\lambda_d) + \lambda_1^{-1} h'_{\chi^2}(z/\lambda_d) \left( \frac{\partial}{\partial \lambda_d} \lambda_1^{-1} \right) = -\lambda_1^{-2} h_{\chi^2}(z/\lambda_d) - \lambda_1^{-3} z h'_{\chi^2}(z/\lambda_d)\). For \(z < 0\) then \(h_d(z) = 0\) and so \(\frac{\partial}{\partial \lambda_d} h_d(z) = 0\).

The event \(z = 0\) has probability zero and can be ignored. For \(z > 0\) we have
\[\sqrt{2\pi} h'_{\chi^2}(z) = \sqrt{2\pi} \frac{dz^{1/2}}{dz} z^{-1/2} e^{-z/2} = \frac{dz}{dz} z^{1/2} e^{-z/2} = \frac{1}{2} z^{-1/2} e^{-z/2} + (-\frac{1}{2}) z^{1/2} e^{-z/2} \]
so that \(h'_{\chi^2}(z/\lambda_d) = \frac{1}{2} \lambda_1^{-1} z^{-1/2} e^{-z/(2\lambda_d)} - \frac{1}{2} \lambda_1^{-1} z^{1/2} e^{-z/(2\lambda_d)}\). Inserting this into the expression obtained for \(\frac{\partial}{\partial \lambda_d} h_r(z)\) gives \(\frac{\partial}{\partial \lambda_d} h_d(z) = -\lambda_1^{-2} f_{\chi^2}(z/\lambda_d) - \frac{1}{2} \lambda_1^{-1} \lambda_1^{-2} z^{-1/2} e^{-z/(2\lambda_d)} - \frac{1}{2} \lambda_1^{-1} \lambda_1^{-2} z^{1/2} e^{-z/(2\lambda_d)}\).

We now note that \(z \mapsto -\lambda_1^{-2} h_{\chi^2}(z/\lambda_d)\) is a density, since it is the density of \(\lambda_j Z_j^2\).

Also, \(z^{1/2} e^{-z/(2\lambda_d)}\) and \(z^{1/2} e^{-z/(2\lambda_d)}\) are proportional to Gamma-distributions. Recall that the Gamma(\(\alpha, \beta\)) density for \(\alpha > 0, \beta > 0\) is \(h_{G(\alpha, \beta)}(z) = \beta^{\alpha} z^{\alpha-1} e^{-\beta z} / \Gamma(\alpha) I\{z \geq 0\}\) in which \(\Gamma(z) = \int_0^\infty u^{z-1} e^{-u} \, du\). In conclusion, we see that \(\frac{\partial}{\partial \lambda_d} h_d(z) = -\lambda_1^{-2} h_{\chi^2}(z/\lambda_d) - \lambda_1^{-3} \Gamma(3/2) h_{G(3/2,1/(2\lambda_d))}(z) + 2 \lambda_1^{-7/2} \Gamma(5/2) h_{G(5/2,1/(2\lambda_d))}(z)\).

By the linearity of integration \(x_{d-1} \mapsto h_{\chi^2}(x_{d-1}/\lambda_d)\), and \(x_{d-1} \mapsto h_{G(5/2,1/(2\lambda_d))}(x_{d-1})\) are densities, eq. (11) is a weighted sum of convolutions of densities that result in new densities \(h_A, h_B, h_C\).

That is, \(\frac{\partial}{\partial \lambda_d} h_S(s) = -\lambda_1^{-2} h_A(z) - 2^{-5/2} \lambda_1^{-4} \Gamma(3/2) h_B(z) + 2^{-7/2} \lambda_1^{-6} \Gamma(5/2) h_C(z)\).

Returning to eq. (10) we therefore see that \(H_S(q) = \int_0^q -\lambda_1^{-1} h_A(s) - 2^{-5/2} \lambda_1^{-4} \Gamma(3/2) h_B(s) + 2^{-7/2} \lambda_1^{-6} \Gamma(5/2) h_C(s) \, ds = -\lambda_1^{-1} H_A(q) - \lambda_1^{-4} \Gamma(3/2) H_B(q) + 2^{-7/2} \lambda_1^{-6} \Gamma(5/2) H_C(q)\) where \(H_A, H_B, H_C\) are the cumulative distribution functions of \(h_A, h_B, h_C\).

Recalling that cumulative distribution functions are probabilities, and hence has absolute values bounded by 1, we see that \(|H_j(T_n;x + r_n h_n)| \leq |\lambda_j + r_n (\lambda_j - \lambda_1^{-1} H_A)| + \lambda_1^{-4} \Gamma(3/2) |H_B(q)| + 2^{-7/2} \lambda_1^{-6} \Gamma(5/2) |H_C(q)|\).
\[ \lambda_j|^{-1} + 2^{-5/2}|\lambda_j + r_n(\hat{\lambda}_j - \lambda_j)|^{-4}\Gamma(3/2) + 2^{-7/2}|\lambda_j + r_n(\hat{\lambda}_j - \lambda_j)|^{-6}\Gamma(5/2). \]

Since \(0 \leq r_n \leq 1\) and \(\hat{\lambda}_j \xrightarrow{P_{n \to \infty}} \lambda_j > 0\), we see that \(|H_j(T_n; x + r_nh_n)| = O_P(1)\). \(\square\)