Grønneberg, S., & Foldnes, N. (2018). Testing Model Fit by Bootstrap Selection. *Structural Equation Modeling: A Multidisciplinary Journal*, 1-9.
doi:10.1080/10705511.2018.1503543

Testing Model Fit by Bootstrap Selection

Abstract

Over the last few decades, many robust statistics have been proposed in order to assess the fit of structural equation models. To date, however, no clear recommendations have emerged as to which test statistic performs best. It is likely that no single statistic will universally outperform all contenders across all conditions of data, sample size, and model characteristics. In a real-world situation, a researcher must choose which statistic to report. We propose a bootstrap selection mechanism that identifies the test statistic that exhibits the best performance under the given data and model conditions among any set of candidates. This mechanism eliminates the ambiguity of the current practice and offers a wide array of test statistics available for reporting. In a Monte Carlo study, the bootstrap selector demonstrated promising performance in controlling Type I errors compared to current test statistics.

Keywords: goodness-of-fit, robustness, structural equation modeling, bootstrapping

Testing Model Fit by Bootstrap Selection

Assessment of overall model fit is a central concern in structural equation modeling (SEM). Using a test statistic derived from the estimated model, researchers seek to evaluate whether the model exhibits good fit to the data. Such test statistics are also used to compare the fit of nested models, for example, in invariance testing of factor models. A general framework of model testing is based on the minimum discrepancy function used during parameter estimation (Browne, 1982). After estimating the model parameters using, for example, the method of normal theory maximum likelihood (ML), model fit is assessed by multiplying the minimum fit function value by the sample size. The resulting test statistic T_{ML} converges in distribution to a weighted sum of independent chi-square variables, each with one degree of freedom. Under ideal conditions (e.g., underlying normally distributed data in the case of ML estimation), each weight is equal to one, and the limiting distribution is a chi-square distribution. In a practical situation, however, the chosen discrepancy function will most likely be misspecified with respect to the underlying data, and the corresponding test statistic will not follow a chi-square distribution. Even in situations such as ML estimation in conjunction with underlying normality, where the test statistic asymptotically happens to follow a true chi-square distribution, the sample size will often be small or moderate, so that the test statistic will have a finite sampling distribution that does not match the nominal chi-square distribution.

Many attempts have been made to approximate the true asymptotic distribution using a more refined approximation than a nominal chi-square distribution. The first such approximation was proposed by Satorra and Bentler (1988), who replaced the weights in the limiting distribution by their mean value, which resulted in a mean-scaled statistic T_{SB}. Over the last decade, many more approximations have been suggested and evaluated using Monte Carlo methods. Asparouhov and Muthén (2010) proposed a scaled-and-shifted statistic, which we refer to as T_{SS}, whereas Wu and Lin (2016) introduced a scaled F test, here denoted by T_{CF}. Recently, Foldnes and Grønneberg (2017) proposed eigenvalue block

averaging (EBA), wherein the weights are estimated and replaced by mean values in blocks of increasing order. In the present study, we include two EBA test statistics: the full eigenvalue approximation T_{EBAF} and the two-block approximation T_{EBA2}. See Wu (2017) for a discussion and evaluation of other test statistics.

We remark that these statistics are based on an asymptotic theory and thus may underperform when the sampling distribution of the test statistic strays from the asymptotic distribution. For instance, in the ideal case of multivariate normal data, the normal-theory ML test statistic may produce inflated Type I error rates in small sample sizes. For further discussion on this issue, see Foldnes and Grønneberg (2017, p. 110). To evaluate the performance of the numerous approximations to test statistics, we must rely on Monte Carlo studies. Despite the large number of such studies, no clear advice on which test statistics to use has emerged. It is likely that no single statistic will universally outperform all contenders as seen in recent studies by Wu and Lin (2016), Wu (2017), and Foldnes and Grønneberg (2017).

Researchers frequently must evaluate model fit based on moderately-sized data that depart from multivariate normality. Given the large number of proposed test statistics designed to handle such situations, researchers face the challenge of choosing a statistic that serves as a basis for model-fit evaluation in terms of a p-value for correct model specification and as a basis for calculating fit indices. The goal of the present paper is to present and to evaluate a bootstrap-based selection procedure that will identify the most reliable test statistic for the given data and model. An additional benefit of this selector is its objectivity, which eliminates any potential temptation for researchers to report test statistics that favor their proposed model.

In the next section, we formally present the selection algorithm. Next, we illustrate the selector using a real-world example. In the section that follows, we report on the performance of the selector in a Monte Carlo study.

The final sections contain discussion and concluding remarks. A theoretical analysis

of the algorithm may be found in the Appendix.

The Bootstrap Selector

In the present article, we introduce a selection mechanism that will select a test statistic and its associated p-value from a set of potential test statistics. The pool of available test statistics should include current best-performing test statistics for SEM. The aim of our proposed method is to select the most well-behaved test statistic for any given situation among an array of available test statistics. As a welcome side effect, the method eliminates the need for researchers to select a test statistic based on assumptions alone—thereby promoting objectivity. Our approach is not based on approximating a limiting distribution but on resampling techniques.

Simply stated, in a given sample, we resample with replacement to obtain bootstrap samples. Each bootstrap sample is drawn from a transformed sample where the model fits perfectly. This procedure was suggested by Beran and Srivastava (1985) and consequently used by Bollen and Stine (1992) to produce the Bollen–Stine bootstrap test (here denoted by BOST). Under correct model specification, the ideal test statistic will produce p-values that are uniformly distributed on the unit interval. This guarantees that Type I error rates exactly match any chosen level of significance. For each of the available test statistics, we calculate the associated p-value. Next, we repeat this procedure over many bootstrap samples, which enables us to approximate the distribution of the p-values for each method. The test statistic with the p-values that most closely follow a uniform distribution is chosen for model fit evaluation. In other words, we choose among the available test statistics the one that best emulates an ideal test statistic. Our selector is inspired by the nonparametric focused information criterion of Jullum and Hjort (2017).

Below we provide a more detailed description of our procedure. Let $\hat{p}_n$ denote the p-value associated with a test of correct model specification based on an available test statistic T_n with a sample size n. We remark that $\hat{p}_n$ is a statistic in the same manner that

T_n is a statistic: It has a distribution under random sampling from the underlying population. We wish to select the test statistic for which the sampling distribution of $\hat{p}_n$ most closely follows 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. For each test statistic, we approximate

$$D_n = \sup_{0 \leq x \leq 1} |P_{H_0}(\hat{p}_n \leq x) - x| \tag{1}$$

and select the statistic with the smallest value of D_n. The probability P_{H_0} is the probability measure induced by the data-generating distribution of $\Sigma(\theta^\circ)^{1/2}\Sigma^{-1/2}X_i$, where Σ is the true covariance matrix and $\Sigma(\theta^\circ)$ is the population model-implied covariance matrix evaluated at the population parameters θ_0 that minimizes the discrepancy function. Under P_{H_0}, we know that p-values should be uniformly distributed, which necessitates studying the transformed sample (under which H_0 is true) instead of the original sample.

The approximation to D_n is accomplished via the nonparametric 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, which chooses among L available test statistics. Here, S_n and $\Sigma(\hat{\theta})$ denote the sample and model-implied covariance matrices, respectively. The supremum in Algorithm 1 is the Kolmogorov–Smirnov test statistic, which is implemented in most statistical software packages.

We use the empirical distribution function $\hat{P}_n$ of $(\tilde{X}_i)$ as an approximation to P_{H_0} and approximate this probability distribution through resampling. Next, we plug this approximation into D_n to generate $\hat{D}_n$ for each p-value approximation. The selector may be used among any test statistics available for hypothesis testing in moment structures. Also, note that D_n is one of many possible success criteria. A researcher 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 (e.g., $x = .05$).

In the Appendix, we provide an analytical overview of the proposed algorithm and

Algorithm 1 Selection algorithm

procedure SELECT(sample, B)
  $\tilde{X}_i = \Sigma(\hat{\theta})^{1/2} S_n^{-1/2} X_i$ for $i = 1, 2, \ldots, n$.
  for $k \leftarrow 1, \ldots, B$ **do**
    boot.sample $\leftarrow$ Draw with replacement from transformed sample $\tilde{X}_i$
    for $l \in 1, \ldots, L$ **do**
      $\hat{p}_{n,l} \leftarrow$ p-value based on boot.sample and test statistic $T_{n,l}$
    end for
  end for
  for $l \in 1, \ldots, L$ **do**
    $\hat{D}_{B,n,l} \leftarrow \sup_{0 \leq x \leq 1} |B^{-1} \sum_{k=1}^{B} I\{\hat{p}_{n,l} < x\} - x|$
  end for
  return $\operatorname{argmin}_{1 \leq l \leq L} \hat{D}_{B,n,l}$
end procedure

demonstrate that if there is a single consistent test statistic among the candidates, the selector will choose the statistic with probability approaching one as sample size increases. This means that the selector test is consistent as long as an asymptotically correct test statistic is included among the candidates. The asymptotic distribution free test developed by Browne (1984) or the full eigenvalue approximation (EBAF) proposed by Foldnes and Grønneberg (2017) are examples of consistent test statistics. We recommend including the EBAF among the available test statistics because it can work with any minimum discrepancy function.

Comparison With the Bollen–Stine Bootstrap

We here compare the proposed selection method and the classical bootstrap procedure by Bollen and Stine (1992). The two methods share some similarities: They both use the same data transformation and both are based on the nonparametric

bootstrap. Otherwise, the two methods are quite different. The Bollen–Stine bootstrap directly approximates the distribution of the test statistic, whereas the selection method uses the bootstrap to approximate D_n in (1), resulting in $\hat{D}_n$. Also, Bollen–Stine bootstrapping is a fixed procedure, whereas the selection method is more flexible, allowing for different sets of candidate test procedures and for different success criteria. Furthermore, the selection method allows researchers to chooses among test procedures, whereas the Bollen–Stine procedure is itself a test procedure. Whether our proposed bootstrap approach performs better than the Bollen–Stine is an empirical question that future researchers should investigate in Monte Carlo studies. We conducted a simple simulation study and found that neither our selection method nor the Bollen–Stine approach outperformed the other across all conditions. However, both approaches outperformed other test statistics that do not rely on bootstrapping, which emphasizes the promising performance of bootstrap techniques relative to more commonly used test procedures and necessitates further study in SEM literature.

Illustration

We considered 10 self-report items taken from the International Personality Item Pool (ipip.orig.org). From the original dataset of 2,800 subjects supplied in the R package psych (Revelle, 2017), we took the first 200 rows as our illustrative dataset. Our goal was to test a two-factor model, where latent factors Agreeableness and Conscientiousness each have five indicators. The model has 34 degrees of freedom. We considered six test statistics for evaluating model fit. The p-values associated with a test of correct model specification for each statistic may be seen in Table 1. A researcher must decide which of these statistics to use as a basis for evaluating model fit.

Next, we transformed the $n = 200$ sample so that the transformed sample shared a covariance matrix identical to the model-implied covariance matrix obtained from the original sample. Hence, the model fits perfectly for the transformed sample. We drew 5,000

Test statistic	T_{ML}	T_{SB}	T_{SS}	T_{CF}	T_{EBAF}	T_{EBA2}
p-value	.010	.037	.063	.065	.066	.055

Table 1

p-values for each of six model fit test statistics. ML = Normal-theory maximum likelihood. SB = Satorra–Bentler. SS = Scaled-and-shifted. CF = Scaled F test. EBAF = Full eigenvalue approximation. EBA2 = Two-block eigenvalue approximation.

bootstrap samples with replacement from the transformed sample and calculate the p-value in each bootstrap sample for each of the six test statistics in Table 1.

The panels in Figure 1 present the distribution of these p-values. Under ideal conditions, the p-values should be uniformly distributed. However, all the available test statistics seem to produce skewed p-values, with p-values appearing more frequently in the lower half of the unit interval. T_{ML} clearly produces too many low p-values in the current condition, but this is partly alleviated by the mean-scaling in T_{SB}. T_{SS}, T_{EBAF}, T_{EBA2}, and T_{CF} seem less likely to produce small p-values compared to T_{ML} and T_{SB}. In order to choose a test statistic, we calculated $\hat{D}$, which is a measure of the distance between the observed distribution of p-values and the ideal uniform distribution. In Figure 2, we present QQ plots with the uniform distribution for each of the six candidates, where the distances $\hat{D}$ have been indicated by vertical line segments. In Figure 2, T_{ML} departs substantially from the uniform distribution for all quantiles. T_{SB} generally displays a closer fit but still differs from the nominal distribution in the lower quantiles. For the remaining test statistics, the p-value distribution is close to uniform for low quantiles (normally the area of most practical concern in hypothesis testing) but strays from the uniform distribution at higher quantiles.

The values of $\hat{D}$ are presented in Table 2, which reveals that the smallest $\hat{D}$ was obtained under T_{EBA2}. Therefore, we conclude that T_{EBA2} is the most reliable among the candidates in the current condition, and we report the p-value of correct model

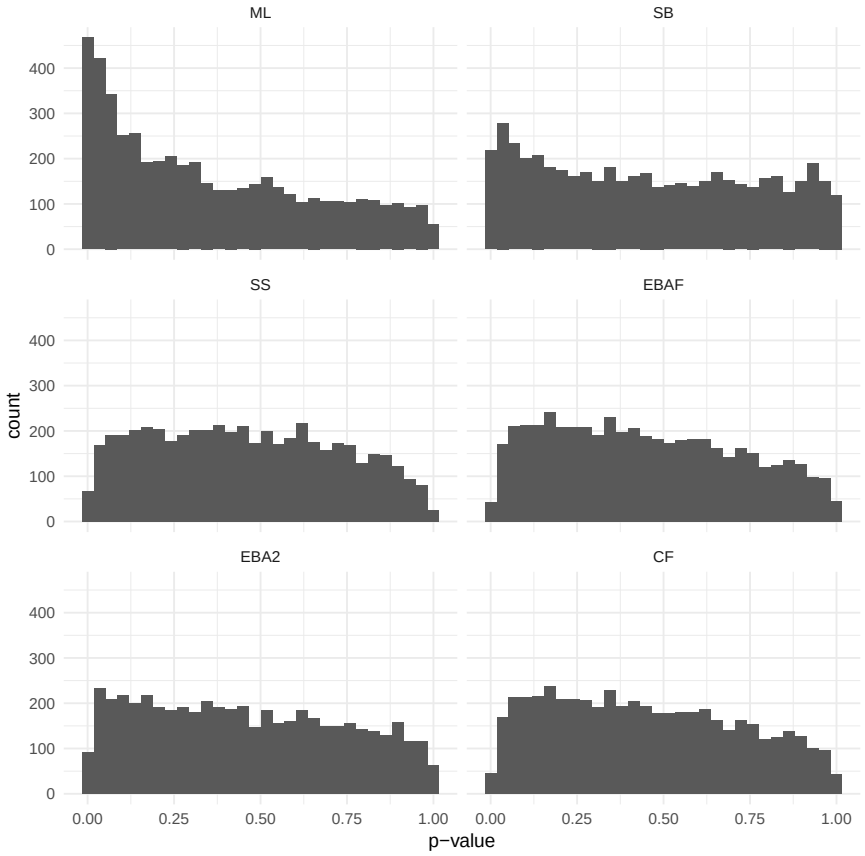


Figure 1. Histograms of *p*-value distribution for each of six candidate test statistics. ML = Normal-theory maximum likelihood. SB = Satorra–Bentler. SS = Scaled-and-shifted. CF = Scaled F test. EBAF = Full eigenvalue approximation. EBA2 = Two-block eigenvalue approximation.

specification to be .055.

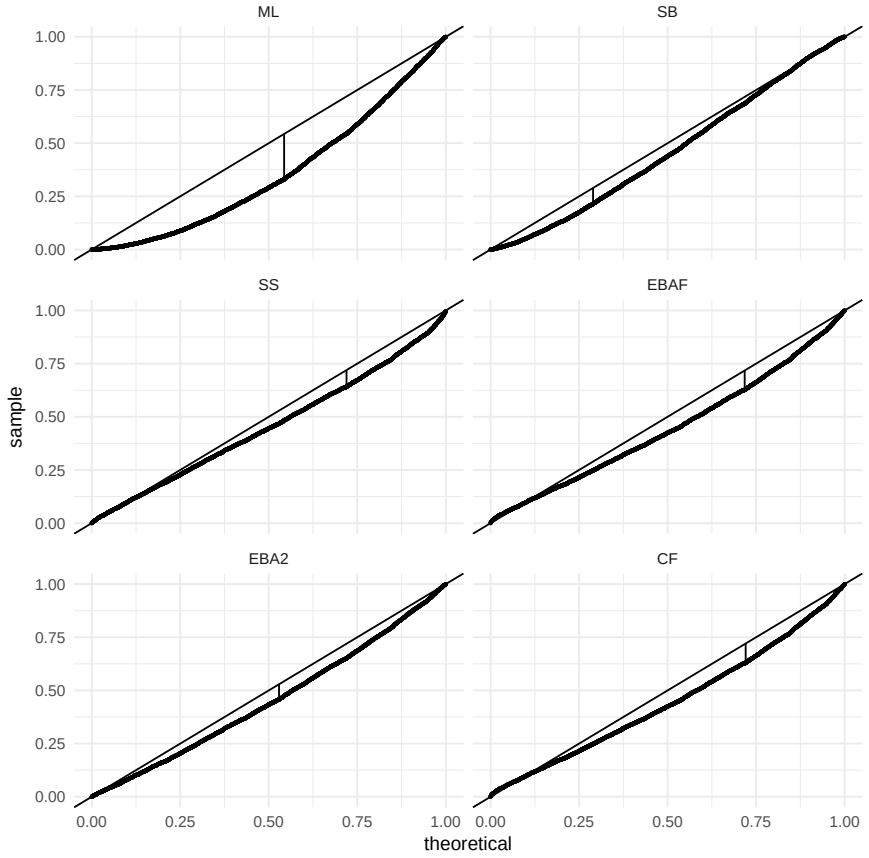


Figure 2. QQ-plot of p-values with the uniform distribution for each of the six candidate test statistics. The vertical line segments indicate $\hat{D}$. ML = Normal-theory maximum likelihood. SB = Satorra–Bentler. SS = Scaled-and-shifted. CF = Scaled F test. EBAF = Full eigenvalue approximation. EBA2 = Two-block eigenvalue approximation.

Test statistic	T_{ML}	T_{SB}	T_{SS}	T_{CF}	T_{EBAF}	T_{EBA2}
$\hat{D}$	.214	.076	.079	.091	.072	.091

Table 2

$\hat{D}$ for each of six candidate test statistics. ML = Normal-theory maximum likelihood. SB = Satorra–Bentler. SS = Scaled-and-shifted. CF = Scaled F test. EBAF = Full eigenvalue approximation. EBA2 = Two-block eigenvalue approximation.

Monte Carlo Evaluation

In this section, we evaluate the selection procedure in Monte Carlo studies. In the first study, we considered the setting of goodness-of-fit testing for a single model. The second study concerned chi-square difference testing for two nested models. In the third study, we investigated in a concrete small-sample case how well the estimates $\hat{D}_n$ approximated their population counterpart D_n in (1).

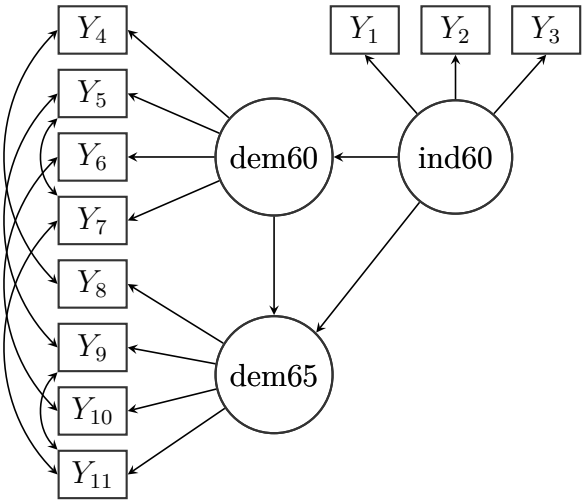
In the first two studies, we evaluated the selection procedure of an Algorithm in terms of Type I error control 1 using three candidate statistics: SB, EBA2, and EBAF. In addition to reporting the performance of ML, SB, EBA2, EBAF, and the selector, we also included the somewhat understudied Bollen–Stine (BOST) bootstrap test statistic. We used the political democracy model discussed by Bollen in his textbook (Bollen, 1989), depicted in Figure 3, where the residual errors are omitted for ease of presentation. Four measures of political democracy were measured twice (in 1960 and 1965), and three measures of industrialization were measured once (in 1960). The unconstrained model has $d = 35$ degrees of freedom. For nested model testing, we also considered a constrained model with $d = 46$ degrees of freedom, which impose 10 equalities among unique variances and residual covariances and one equality between two factor loadings.

Three population distributions were considered. Distribution 1 was a multivariate normal distribution. The nonnormal distributions were generated using the transformation algorithm of (Vale & Maurelli, 1983); Distribution 2 had univariate skewness 1 and kurtosis 7, and Distribution 3 had skewness 2 and kurtosis 21.

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 ML, SB, EBA2, EBAF, and BOST. In the selection algorithm (SEL), the p-value was calculated using a candidate set containing SB, EBA2, and EBAF and using $\hat{D}_n$ as a criterion function.

Model estimation was done using the R package lavaan (Rosseel, 2012), and the

Figure 3. Bollen's political democracy model. dem60: Democracy in 1960. dem65: Democracy in 1965. ind60: Industrialization in 1960.



p-values from EBA2 and EBAF were calculated using the imhof procedure in the CompQuadForm (Duchesne & De Micheaux, 2010) package. In each simulation cell, we generated $2,000$ samples. For each sample, $1,000$ bootstrap samples were drawn.

Results for Single Model Testing

In Table 3, we present Type I error rates for single model testing at the 5% significance level. As expected, ML becomes inflated when data are nonnormally distributed. The mean-scaling of SB reduces inflation, but with nonnormal data and small sample sizes, Type I error rates are still higher than 10%. These findings match those of previous studies (Foldnes & Olsson, 2015). BOST performs better than SB, coming close to the nominal level even for highly nonnormal data and a medium sample size. Among the eigenvalue-based approximations, EBA2 performs the best, whereas EBAF yields far too low rejection rates with nonnormal data. The selection algorithm SEL also performs generally well—on par with EBA2 and BOST. It is noteworthy that SEL outperforms ML for normal data.

Table 4 presents the selection proportions for SEL in each of the nine conditions and

Distribution	n	ML	SB	BOST	EBAF	EBA2	SEL
Normal	100	.077	.086	.023	.036	.050	.051
	300	.055	.053	.037	.037	.043	.045
	900	.068	.067	.059	.063	.064	.065
Distribution 2	100	.215	.108	.035	.021	.048	.042
	300	.197	.070	.053	.024	.045	.045
	900	.219	.063	.033	.037	.051	.051
Distribution 3	100	.488	.164	.038	.009	.072	.031
	300	.591	.094	.068	.013	.050	.038
	900	.685	.076	.059	.015	.042	.038

Table 3

Type I error rates for single model testing. Normal: multivariate normal distribution, Distribution 2: skewness 1 *and kurtosis* 7*. Distribution 3: skewness* 2 *and kurtosis* 7*. ML = Normal-theory likelihood ratio test. SB = Satorra–Bentler. BOST = Bollen–Stine bootstrap. EBAF = Full eigenvalue approximation. EBA2 = Two-block eigenvalue approximation. SEL = bootstrap selector.*

shows that the selection algorithm wisely chose EBA2 in the majority of conditions. However, it is unexpected that SEL chose EBAF in 55% of the samples under Distribution 3 and $n = 100$, given the poor performance of EBAF in that condition, which had a 1% rejection rate.

Results for Nested Model Testing

Rejection rates for the chi-square difference test using a 5% level of significance are reported in Table 5. Again, the ML statistic was inflated by nonnormality in the data—a tendency only partially corrected for by SB. For instance, under the most misspecified condition (i.e., Distribution 3 and $n = 100$), SB rejection rates were 22%—far better than

Distribution	n	SB	EBA2	EBAF
Normal	100	.054	.931	.015
	300	.448	.516	.036
	900	.507	.263	.231
Distribution 3	100	.001	.865	.135
	300	.050	.894	.055
	900	.153	.783	.063
Distribution 3	100	.000	.449	.551
	300	.001	.733	.267
	900	.004	.846	.150

Table 4

Choice proportions for selection algorithm, single model testing. SB = Satorra–Bentler. EBAF = Full eigenvalue approximation. EBA2 = Two-block eigenvalue approximation

the 91% obtained with ML. But in this condition, as in all conditions, BOST performed better than SB, with a rejection rate of 13%. However, EBAF performed even better in this condition, whereas the selection algorithm was only slightly worse than BOST. Overall, EBAF outperformed the other test statistics, including SB and BOST. EBA2, which was found to perform best in the nonnested case, did not perform as well as EBAF in the nested case. The selection algorithm, SEL, also performed well, with better performance than SB and BOST in most conditions, and only slightly worse than EBAF. The selection proportions are presented in Table 6, where we unexpectedly found EBA2 to be the most chosen procedure, despite the slightly better performance of EBAF in most conditions.

How Well Does $\hat{D}_n$ Approximate D_n?

In this section, we present results for the third Monte Carlo study; our aim was to investigate how closely the bootstrap estimates $\hat{D}_n$ approximate the population value D_n.

Distribution	n	ML	SB	BOST	EBAF	EBA2	SEL
Normal	100	.068	.080	.037	.062	.069	.075
	300	.054	.059	.046	.053	.055	.058
	900	.051	.053	.051	.051	.052	.053
Distribution 2	100	.582	.137	.096	.076	.099	.096
	300	.659	.088	.081	.052	.066	.062
	900	.702	.059	.053	.035	.043	.045
Distribution 3	100	.911	.221	.129	.115	.159	.135
	300	.961	.126	.118	.062	.089	.082
	900	.976	.087	.089	.044	.064	.061

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*. ML = Normal-theory likelihood ratio test. SB = Satorra–Bentler. BOST = Bollen–Stine bootstrap. EBAF = Full eigenvalue approximation. EBA2 = Half eigenvalue approximation. SEL = p-value obtained from selection algorithm*

Because this is a Monte Carlo study, where the underlying distribution is known, we were able to calculate D_n and make this comparison. We revisited the two-factor model described in the Illustration section, which has five indicators for each factor. We drew 300 samples, each of size $n = 150$, from a multivariate normal distribution whose covariance matrix is the model-implied covariance matrix when the two-factor model was fitted to the original large dataset. For each simulated sample, $B = 1,000$ bootstrap samples were drawn and $\hat{D}_n$ was calculated according to Algorithm 1 for each of the following test statistics: $T_{\text{ML}}, T_{\text{SB}}, T_{\text{SS}}, T_{\text{EBAF}}, T_{\text{EBA2}}$, and T_{CF}. Using 4,000 simulated $n = 150$ samples, we approximated D_n with high precision. We present our results in Figure 4, where the D_n are represented by triangles and the $\hat{D}_n$ by boxplots. Because the data were normal, p-values

Distribution	n	SB	EBA2	EBAF
Normal	100	.601	.357	.042
	300	.672	.205	.122
	900	.593	.091	.316
Distribution 3	100	.116	.714	.170
	300	.209	.662	.128
	900	.263	.595	.142
Distribution 3	100	.012	.663	.325
	300	.059	.725	.215
	900	.104	.714	.182

Table 6

Choice proportions for selection algorithm, nested model testing. SB = Satorra–Bentler. EBA2 = Half eigenvalue approximation. EBAF = Full eigenvalue approximation.

from T_{ML} most closely followed the uniform distribution. Ideally, the selector should therefore select T_{ML} every time, but in a real-world setting, we do not know D_n, only its bootstrap approximation $\hat{D}_n$. The selector chose T_{ML} as the preferred test statistic in 206 of the 300 samples, so it was able to identify the best statistic in the majority of samples. However, the second best statistic (in terms of D_n), T_{SB}, was only chosen seven times, whereas T_{SS} and T_{EBA2} were chosen 38 and 49 times, respectively. In other words, the selector was not able to detect that T_{SB} was the second best statistic in terms of D_n and generally preferred T_{SS} and T_{EBA2} to T_{SB}. As seen in Figure 4, the reason is that $\hat{D}_n$ tended to overestimate D_n for T_{SB} and to a lesser degree also for T_{ML}. For the remaining four test statistics, $\hat{D}_n$ offered good approximations to D_n. In this condition with multivariate normal samples of size $n = 150$, $\hat{D}_n$ was usually larger than its population counterpart in the T_{SB} case.

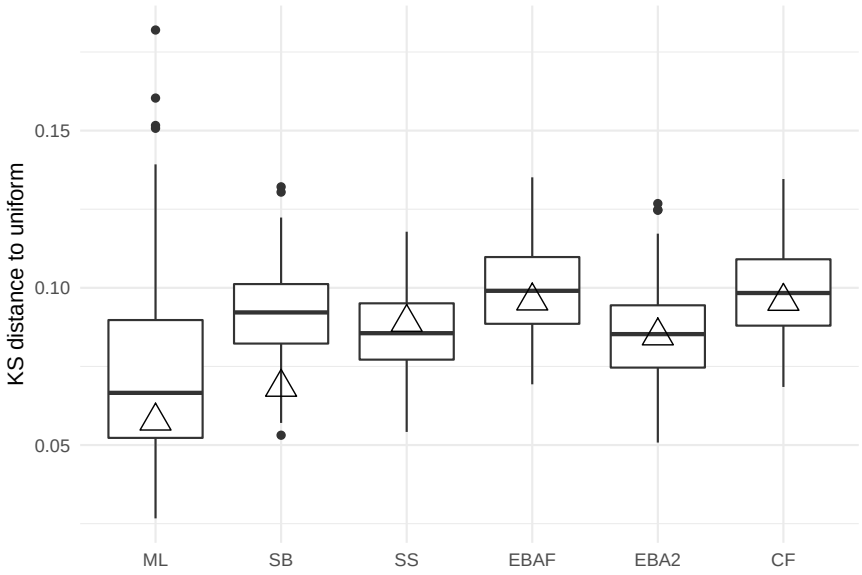


Figure 4. Boxplots of $\hat{D}_n$ for six test statistics. Triangles represent D_n. ML = Normal-theory maximum likelihood. SB = Satorra–Bentler. SS = scaled-and-shifted. CF = Scaled F test. EBAF = Full eigenvalue approximation. EBA2 = Two-block eigenvalue approximation.

Discussion

In the first two Monte Carlo studies, we evaluated the bootstrap selector in two scenarios: the test of a single model and nested-model testing. The selector was given three candidate test statistics to choose from: the well-established SB mean-adjusted test and two newly proposed tests, EBAF and EBA2, based on eigenvalue approximation.

Overall, EBA2 performed best when testing single models. In the scenario of nested model testing, however, EBAF maintained the best Type I error control. This illustrates the central problem addressed in the present paper: It is difficult to identify a priori which test statistic will perform best. The proposed selection method is an attempt to remedy this problem, based on choosing the statistic whose bootstrap distribution best matches a particular criterion (e.g., that p-values should be uniformly distributed).

In the third Monte Carlo study, we investigated a single normal data condition with

sample size $n = 150$, wherein we sought to evaluate how well the bootstrap estimates $\hat{D}_n$ approximated the true target value D_n. We generally found that, despite the small sample size, $\hat{D}_n$ offered good approximations to D_n. In most samples, the selector was able to identify the best test statistic. However, we also found that for some samples, the selector chose a test statistic not among the best performing statistics. This illustrates that a different sample from the same population might lead the selector to choose a different test statistic.

The present study had some limitations. We considered only one selection criterion, namely the minimization of the Kolmogorov–Smirnov distance between the bootstrap distribution and the theoretical uniform distribution. Other criteria may be considered, for example, matching the observed rejection rates at a specific significance level (i.e., 5%) to the nominal level. In other words, we may choose the test whose rejection rate under bootstrap resampling most closely echoes a specific nominal level. We consider this topic important in future research. Another limitation was that we considered only six test candidates in our Monte Carlo studies. Adding more test statistics to the pool may improve the performance of the selector because some of the added test statistics may outperform the ones already included. Our Monte Carlo study was limited to Type I error control and did not investigate power. A practical limitation is the running time needed for bootstrap-based procedures. The selector and the Bollen–Stine had similar running times—much longer than calculating test statistics such as SB and the EBA variants. 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 couple of additional minutes compared to standard p-value approximations, which often perform 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 may be worthwhile.

Conclusion

This paper deals with the fundamental problem of hypothesis testing in moment structure models. We present a new bootstrap-based selector which, based on the model and the available sample, identifies the most reliable test statistic. This means that researchers no longer must choose one candidate among the many available test statistics on which to base their assessment of model fit. The objectivity of the selector also addresses the concern that researchers may be tempted to use the test statistics that favor their model. Monte Carlo studies indicate that the selector performs relatively well in terms of controlling Type I errors.

The bootstrap method contained in this paper can be generalized in several directions, including SEM with ordinal variables and in multi-group settings. Additional simulation experiments should be performed on the bootstrap selector, such as power studies, allowing the selector to choose among more candidate test statistics, and experimenting with different selection criteria.

Appendix

Mathematical derivations

We here derive the consistency of the bootstrap selector assuming that one—and only one—of the candidate procedures is consistent and hence produces an asymptotically uniform p-value. This places the bootstrap selector as the fourth consistent testing procedure for SEM, in addition to the Bollen–Stine bootstrap, the ADF, and the EBAF. From such an asymptotic perspective, these four procedures are equally good. However, the bootstrap selector continuously seeks the best possible finite sample performance. Consistency is a desirable property, but it is not the motivation for using the method.

The proof of consistency we present here is an elaboration of an observation of Hannan and Quinn (1979, p. 191) and is well-known in the model selection literature. Hannan and Quinn (1979) noted that the model chosen by a consistent model selection method can be used as if the correct model was known in terms of the asymptotic behaviour of the model. We here show that such a conclusion also holds for the bootstrap selector under mild assumptions. The main assumption is an asymptotic uniqueness assumption. Going outside this assumption increases the complexity of the required arguments considerably, and we consider this outside the scope of the paper, especially because uniqueness is reasonable in the current context.

Suppose we have L competing methods for computing a certain p-value. Denote the p-value of method j based on n observations by $U_{n,j}$. A population parameter measuring the quality of the j'th method is denoted by $D_{n,j}$. An empirical estimate of $D_{n,j}$ is denoted by $\hat{D}_{n,j}$.

Assumption 1. *1. We assume that* $\hat{D}_{n,j} \geq 0$ *and* $D_{n,j} \geq 0$.

2. We assume that $\hat{D}_{n,j} - D_{n,j} \xrightarrow[n\to\infty]{P} 0$ *for* $j = 1, 2, \ldots, L$

3. We assume that the function $j \mapsto \hat{D}_{n,j}$ *over* $j = 1, 2, \ldots, L$ *has only one minimizer.*

Assumption 1 (2) is fulfilled by the bootstrap approximation of the

Kolmogorov–Smirnov type procedure suggested in the main part of the paper using arguments such as Beran and Srivastava (1985), see also Efron and Tibshirani (1994).

Under Assumption 1 (3) we denote

$$\hat{\kappa}_n = \underset{1\leq j\leq L}{\operatorname{argmin}} \hat{D}_{n,j}.$$

The following assumption means that $D_{n,\kappa}$ is asymptotically closest to zero and is an asymptotic uniqueness assumption. If the Kolmogorov–Smirnov procedure is used, the assumption means that there is a specific testing procedure which has the least uniform distance between the asymptotic distribution of the p-value and the uniform distribution. If a consistent method is included, this minimum is zero, because consistent methods have asymptotically uniform p-values.

Assumption 2. *For a* $\kappa \in \{1, \ldots, L\}$*, we suppose that* $\lim_{n\to\infty}(D_{n,\kappa} - D_{n,j}) < 0$ *for* $j \in \{1, \ldots, L\} \setminus \{\kappa\}$.

The asymptotically best procedure is therefore κ. We are interested in identifying when the asymptotic distribution of $U_{n,\hat{\kappa}_n}$ is the same as $U_{n,\kappa}$. In the case when a consistent procedure is in the set of considered methods, this means $U_{n,\hat{\kappa}_n}$ is asymptotically uniform on $[0, 1]$ as shown in the following theorem. We note that stronger conclusions outside the scope of the current paper, such as $\lim_{n\to\infty} |P(A_n \cap \{U_{n,\hat{\kappa}_n} = U_{n,\kappa}\}) - P(A_n)| = 0$ for any sequence of events $(A_n)_{n=1}^{\infty}$, follow through a simple extension of the proof of Lemma 2.

Theorem 1. *Suppose* $U_{n,\kappa} \xrightarrow[n\to\infty]{D} U[0, 1]$*. Under Assumption 1 and 2 we have that* $U_{n,\hat{\kappa}} \xrightarrow[n\to\infty]{D} U[0, 1]$.

Proof. Let $0 < x < 1$. We have
$|P(U_{n,\hat{\kappa}} \leq x) - x| = |P(U_{n,\hat{\kappa}} \leq x) - P(U_{n,\kappa} \leq x) - (x - P(U_{n,\kappa} \leq x))|$. By the triangle inequality, this is bounded by $|P(U_{n,\hat{\kappa}} \leq x) - P(U_{n,\kappa} \leq x)| + |x - P(U_{n,\kappa} \leq x)|$. The second term goes to zero by assumption. The first term goes to zero by the conclusion of Lemma 2. □

Lemma 1. *Under Assumption 1 and 2, we have that* $\lim_{n\to\infty} P(\hat{\kappa}_n = \kappa) = 1$.

Proof. We have

$$\begin{aligned} P(\hat{\kappa}_n = \kappa) &= P(\cap_{1\le j\le L, j\neq\kappa}\{\hat{D}_{n,\kappa} < \hat{D}_{n,j}\}) \\ &= P(\cap_{1\le j\le L, j\neq\kappa}\{\hat{D}_{n,\kappa} - D_{n,\kappa} + D_{n,\kappa} < \hat{D}_{n,j} - D_{n,j} + D_{n,j}\}) \\ &= P(\cap_{1\le j\le L, j\neq\kappa}\{(\hat{D}_{n,\kappa} - D_{n,\kappa}) - (\hat{D}_{n,j} - D_{n,j}) + D_{n,\kappa} - D_{n,j} < 0\}) \end{aligned}$$

By Assumption 1 (2), we have that

$$(\hat{D}_{n,\kappa} - D_{n,\kappa}) - (\hat{D}_{n,j} - D_{n,j}) + D_{n,\kappa} - D_{n,j} = o_P(1) + D_{n,\kappa} - D_{n,j}.$$

Because $\lim_{n\to\infty} D_{n,\kappa} - D_{n,j} < 0$ by Assumption 2, the conclusion follows. □

Lemma 2. *Under Assumption 1 and 2, we have that*

$$\lim_{n\to\infty} |P(U_{n,\hat{\kappa}} \le x) - P(U_{n,\kappa} \le x)| = 0$$

for all x.

Proof. Recall that if $A_1, A_2, \ldots, A_L$ are disjoint and $\cup_{j=1}^{L} A_j$ is the whole probability space Ω, we have for an event B that $P(B) = P(B \cap \Omega) = P(B \cap \cup_{j=1}^{L} A_j) = \sum_{j=1}^{L} P(B \cap A_j)$. Therefore,

$$\begin{aligned} P(U_{n,\hat{\kappa}} \le x) &= \sum_{j=1}^{L} P(U_{n,\hat{\kappa}} \le x, \hat{\kappa} = j) \\ &= \sum_{j=1}^{L} P(U_{n,j} \le x, \hat{\kappa} = j) \\ &= P(U_{n,\kappa} \le x, \hat{\kappa} = \kappa) + \sum_{1\le j\le L, j\neq\kappa} P(U_{n,j} \le x, \hat{\kappa} = j) \end{aligned}$$

For $j \neq \kappa$, we have $0 \le P(U_{n,j} \le x, \hat{\kappa} = j) \le P(\hat{\kappa} = j) \le P(\hat{\kappa} \neq \kappa) \to 0$ by Lemma 1. Further, we have $P(U_{n,\kappa} \le x) = P(U_{n,\kappa} \le x, \hat{\kappa} = \kappa) + P(U_{n,\kappa} \le x, \hat{\kappa} \neq \kappa)$ so that $P(U_{n,\kappa} \le x, \hat{\kappa} = \kappa) = P(U_{n,\kappa} \le x) - P(U_{n,\kappa} \le x, \hat{\kappa} \neq \kappa)$. Because $0 \le P(U_{n,\kappa} \le x, \hat{\kappa} \neq \kappa) \le P(\hat{\kappa} \neq \kappa) \to 0$, we have $P(U_{n,\kappa} \le x, \hat{\kappa} = \kappa) = P(U_{n,\kappa} \le x) - o(1)$, proving the result. □

References

Asparouhov, T., & Muthén, B. (2010). Simple second order chi-square correction. *Unpublished manuscript*. Retrieved from `www.statmodel.com/download/WLSMV_new_chi21.pdf`

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. doi: 10.1002/9781118619179

Bollen, K. A., & Stine, R. A. (1992). Bootstrapping goodness-of-fit measures in structural equation models. *Sociological Methods & Research*, *21*(2), 205–229.

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*(1), 62–83.

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 & Data Analysis*, *54*(4), 858–862.

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

Foldnes, N., & Grønneberg, S. (2017). Approximating test statistics using eigenvalue block averaging. *Structural Equation Modeling: A Multidisciplinary Journal*, 1–14.

Foldnes, N., & Olsson, U. H. (2015). Correcting too much or too little? The performance of three chi-square corrections. *Multivariate behavioral research*, *50*(5), 533–543.

Hannan, E. J., & Quinn, B. G. (1979). The determination of the order of an autoregression. *Journal of the Royal Statistical Society. Series B (Methodological)*, 190–195.

Jullum, M., & Hjort, N. L. (2017). Parametric or nonparametric: The fic approach. *Statistica Sinica*, *27*, 951–981.

Revelle, W. (2017). psych: Procedures for psychological, psychometric, and personality research [Computer software manual]. Evanston, Illinois. Retrieved from `https://CRAN.R-project.org/package=psych` (R package version 1.7.8)

Rosseel, Y. (2012). lavaan: An R package for structural equation modeling. *Journal of Statistical Software*, *48*(2), 1–36.

Satorra, A., & Bentler, P. (1988). Scaling corrections for statistics in covariance structure analysis (UCLA statistics series 2). *Los Angeles: University of California at Los Angeles, Department of Psychology*.

Vale, C. D., & Maurelli, V. A. (1983). Simulating multivariate nonnormal distributions. *Psychometrika*, *48*(3), 465–471.

Wu, H. (2017). Approximations to the distribution of a test statistic in covariance structure analysis: A comprehensive study. *British Journal of Mathematical and Statistical Psychology*.

Wu, H., & Lin, J. (2016). A scaled F distribution as an approximation to the distribution of test statistics in covariance structure analysis. *Structural Equation Modeling: A Multidisciplinary Journal*, *23*(3), 409–421.