Foldnes, N., & Grønneberg, S. (2018). Approximating test statistics using eigenvalue block averaging. *Structural Equation Modeling: A Multidisciplinary Journal*, *25*(1), 101-114 DOI: http://doi.org/10.1080/10705511.2017.1373021

Approximating test statistics using eigenvalue block averaging

Njål Foldnes and Steffen Grønneberg

Department of Economics

BI Norwegian Business School

Oslo, Norway 0484

Correspondence concerning this article should be sent to *njal.foldnes@bi.no*

Abstract

We introduce and evaluate a new class of approximations to common test statistics in structural equation modeling. Such test statistics asymptotically follow the distribution of a weighted sum of i.i.d. chi-square variates, where the weights are eigenvalues of a certain matrix. The proposed eigenvalue block averaging (EBA) method involves creating blocks of these eigenvalues and to replace them within each block with the block average. The Satorra-Bentler scaling procedure is a special case of this framework, using one single block. The proposed procedures applies also to difference testing among nested models. We investigate the EBA procedure both theoretically in the asymptotic case, and with simulation studies for the finite-sample case, under both ML and DWLS estimation. Comparison is made with three established approximations: Satorra-Bentler, the scaled and shifted, and the scaled F tests.

Keywords: Satorra-Bentler, fit statistics, non-normal data, structural equation modeling

Approximating test statistics using eigenvalue block averaging

In general, test statistics for moment structural models converge in law to the distribution of a weighted sum of independent chi squares, under the null hypothesis of correct model specification. More precisely, a test statistic T_n based on n observations will obey (Shapiro, 1983; Satorra, 1989)

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

where the weights $\lambda = (\lambda_1, \ldots, \lambda_d)'$ are the non-zero eigenvalues of an unknown population matrix. Under optimal conditions, in which the estimator is correctly specified for the data at hand, or under conditions of so-called asymptotic robustness (e.g., Shapiro, 1987; Browne & Shapiro, 1988), the weights λ_j are all equal to one, and T_n converges to a chi-square distribution. However, in most cases the weights are not equal to one, and T_n should not be referred to a nominal chi-square distribution.

One approach to this problem is to construct a distribution that approximates the distribution of the weighted sum in (1), and refer T_n to this approximating distribution. That is, using characteristics of the data and the model, a distribution is constructed that tries to emulate the distribution of $\sum \lambda_j Z_j^2$. Let X_{approx} be a random variable that follows this approximating distribution. Then the p-value of the test of correct model specification is obtained as $P(X_{\text{approx}} > T_n)$, where T_n is considered fixed and the probability is with respect to X_{approx}. For instance, the scaling of Satorra and Bentler (1988) approximates the weighted sum in (1) by setting all the weights equal to the average $\bar{\lambda} = \sum_{j=1}^{d} \hat{\lambda}_j / d$ of the estimated eigenvalues. That is, $X_{\text{approx}} = \sum_j \bar{\lambda} Z_j^2$, with p-value $P(\sum_j \bar{\lambda} Z_j^2 > T_n)$, which can be recasted in the more familiar form $P(\chi_d^2 > T_n/\bar{\lambda})$. Other recently proposed approximations to the distribution in (1) are the scaled F distribution (Wu & Lin, 2016) and the scaled and shifted χ_d^2 (Asparouhov & Muthén, 2010). The scaled-and-shifted test statistic is closely related to the Sattertwaithe type test statistic proposed by Satorra and Bentler (1994), and these two statistics have been reported to have similar performance

(Foldnes & Olsson, 2015).

If λ was known, eq. (1) motivates the "oracle" p-value

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

which would yield an asymptotically valid test of model fit. In a practical setting λ is unfortunately unknown, but consistent estimates $\hat{\lambda}$ may be obtained. This suggests the approximation $X_{\mathsf{approx}} = \sum_j \hat{\lambda}_j Z_j^2$ and the associated p-value

$$\hat{p}_n = P\left(\sum_{j=1}^{d} \hat{\lambda}_j Z_j^2 > T_n\right). \tag{3}$$

However, this consistency may come at a price, given the variability of the $\hat{\lambda}$. In practice, it may be better to replace the $\hat{\lambda}$ in eq.(3) with more stable weights $\tilde{\lambda}$, obtained through grouping the $\hat{\lambda}$ by magnitude in blocks and calculating block averages. We refer to this method as eigenvalue block averaging (EBA). As there are many ways to form blocks, the EBA method yields many new approximations to the limiting distribution in (1).

Although the EBA idea is simple, to the best of our knowledge it has not been discussed before. However, Wu and Lin (2016) investigated the full eigenvalue approximation in (3), which technically is an EBA procedure with singleton blocks. Also, at the other extremum, EBA with one single block is identical to the well-known Satorra-Bentler scaling procedure. We are not aware of any literature on EBA approximations between these two extremes. The goal of the present paper is to present the EBA framework, and to evaluate EBA tests both asymptotically and in finite samples, by comparing EBA to three establisthed test statistics for structural equation models.

This article is organized as follows. First, we review the literature on test statistics for moment structural moments, followed by a section formally introducing the EBA tests. We then illustrate the established and proposed tests on a real-world example, followed by asymptotic and finite-sample evaluations of the tests for single and nested model testing. The final section contains discussion and concluding remarks.

Test statistics

A structural equation model implies a parametrization $\theta \mapsto \sigma(\theta)$, where the free parameters in the proposed model are contained in the q-vector θ. The model has degrees of freedom given by $d = p^* - q$, where p^* denotes the dimension of $\sigma(\theta)$. In covariance structure models $\sigma(\theta)$ consists of second-order moments, but in more general structural equation models the means may also be included in $\sigma(\theta)$. The corresponding sample moment vector s is assumed to converge in probability to $\sigma_0 = \sigma(\theta_0)$, and be asymptotically normal, i.e., $\sqrt{n}(s - \sigma_0) \xrightarrow[n\to\infty]{D} N(0, \Gamma)$. Here Γ is the asymptotic covariance matrix of $\sqrt{n}s$. A very general class of estimators for θ_0 introduced by Browne (1982, 1984) is obtained by minimising discrepancy functions $F = F(s, \sigma)$ that obey the following three conditions: $F(s, \sigma) \geq 0$ for all s, σ; $F(s, \sigma) = 0$ if and only if $s = \sigma$; and F is twice continuously differentiable jointly. That is, we consider estimators obtained as

$$\hat{\theta} = \underset{\theta}{\operatorname{argmin}}\, F(s, \sigma(\theta)).$$

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

Minimum discrepancy estimation leads to the fit statistic $T_n = nF(s, \sigma(\hat{\theta}))$, which is asymptotically equivalent to several other tests for model fit (Satorra, 1989). Correct model specification and other assumptions (Shapiro, 1983) imply the convergence in eq. (1). The weights $\lambda_1, \ldots, \lambda_d$ are the non-zero eigenvalues of $U\Gamma$, where $U = V - V\Delta \left\{\Delta' V \Delta\right\}^{-1} \Delta' V$, Δ is the $p \times q$ derivative matrix $\partial\sigma(\theta)/\partial\theta'$ evaluated at θ_0, and $V = -\frac{1}{2}\frac{\partial^2 F(s,\sigma)}{\partial s \partial \sigma}$, evaluated at (σ_0, σ_0). Clearly, if all the λ are equal to one, then T_n converges to a chi-square distribution with d degrees of freedom, and we are in a so-called asymptotic robust situation. Conditions necessary for this have been been characterized (e.g., Shapiro, 1987; Browne & Shapiro, 1988; Amemiya, Anderson, et al., 1990; Satorra & Bentler, 1990; Mooijaart & Bentler, 1991). 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).

The scaling procedure proposed by Satorra and Bentler (1988) is defined as $T_{\text{SB}} = T_n/\hat{c}$, where $\hat{c} = \text{trace}(\hat{U}\hat{\Gamma})/d$. Asymptotically T_{SB} converges to a distribution whose expectation equals d, the expectation of the nominal chi-square distribution. In conditions where all eigenvalues are equal, $\lambda_1 = \ldots = \lambda_d$, T_{SB} will converge in distribution to a chi-square distribution. Using T_{SB} as a test statistic is a widely used SEM practice under conditions of non-normal data. Simulation studies report that T_{SB} outperforms the ML fit statistic T_{ML} in such conditions, but that Type I error rates under T_{SB} may become inflated under substantial excess kurtosis in the data (Bentler & Yuan, 1999; Nevitt & Hancock, 2004; Foldnes & Olsson, 2015). Also, Yuan and Bentler (2010) demonstrated that T_{SB} departs from a chi-square with increasing dispersion of the eigenvalues λ_j, $j = 1, \ldots, d$.

Recently Asparouhov and Muthén (2010) proposed a test statistic that agrees with the reference chi-square distribution in both asymptotic mean and variance, obtained from T_{ML} by scaling and shifting. This statistic is given by $T_{\text{SS}} = a \cdot T_n + d - b$, where $a = \sqrt{d/\text{trace}\left((\hat{U}\hat{\Gamma})^2\right)}$ and $b = \sqrt{d\left(\text{trace}(\hat{U}\hat{\Gamma})\right)^2/\text{trace}\left((\hat{U}\hat{\Gamma})^2\right)}$. In a simulation study, Foldnes and Olsson (2015) found that T_{SB} and T_{SS} tended to respectively overreject and underreject correctly specified models.

Very recently, Wu and Lin (2016) proposed a scaled F distribution that matches the mean, variance and skewness of $\sum_{j=1}^{d} \hat{\lambda}_j Z_j^2$, where the $\hat{\lambda}_j$ are the eigenvalues of $\hat{U}\hat{\Gamma}$. The scaling, and the two degrees of freedom of the F distribution, are functions of $\sum_j \hat{\lambda}_j, \sum_j \hat{\lambda}_j^2$ and $\sum_j \hat{\lambda}_j^3$. In a simulation study, Wu and Lin (2016) found the scaled F test to perform similarly to the Sattertwaithe type test statistic proposed by Satorra and Bentler (1994).

EBA test statistics

In this section we introduce new tests for model fit, based on the asymptotic result in (1). 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. Also, the tests may be applied in a context more general than the prototypical case of T_{ML}, for instance with diagonally weighted least squares (DWLS) estimation and model testing.

Note that the p-value $\hat{p}_n$ in (3) is theoretically optimal when the sample size goes to infinity. That is, since $\hat{\lambda}$ converges to λ in probability, the difference between $\hat{p}_n$ and the oracle p-value p_n in eq. (2) goes to zero in probability, meaning that it has zero asymptotic bias. However, in situations with small sample sizes and highly non-normal data the estimates $\hat{\lambda}_j$ become unstable and highly variable. Since $\hat{p}_n$ directly employs each individual estimate $\hat{\lambda}_j$ it may inherit this instability, leading to poor finite-sample performance.

One established way of overcoming this instability is offered by the SB-test. As previously discussed, this test estimates each λ_j by the grand average of all estimated eigenvalues. Clearly, unless all the population eigenvalues are identical, this method is inconsistent. However, the averaging process may result in less variability at the cost of some bias.

In the present study our perspective is that of a bias-variance tradeoff, in which the SB test and the full use of estimated eigenvalues in $\hat{p}_n$ are viewed as extreme end points on a spectrum. At one end of the spectrum, importance is given to stabilizing the eigenvalues, as done in the SB test. At the other end, importance is given to asymptotic bias. We propose intermediate solutions, referred to as EBA tests, between these two extremes. EBA testing involves grouping the $\hat{\lambda}_j$ in blocks by magnitude, and replacing them by group averages, as we will shortly formalize mathematically below. The resulting EBA tests may be viewed as middle-grounds between the 1-block EBA (the SB test) and the d-block EBA in (3).

Consider first the following split-half approximation, where the lower half of the eigenvalues constitute one block, and are replaced by their mean value, and likewise for the

block containing upper half of the eigenvalues:

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

where

$$\tilde{\lambda}_1 = \cdots = \tilde{\lambda}_{\lceil d/2 \rceil} = \frac{1}{\lceil d/2 \rceil} \sum_{j=1}^{\lceil d/2 \rceil} \hat{\lambda}_j, \quad \text{and} \quad \tilde{\lambda}_{\lceil d/2 \rceil + 1} = \cdots = \tilde{\lambda}_d = \frac{1}{d - \lceil d/2 \rceil} \sum_{j=\lceil d/2 \rceil + 1}^{d} \hat{\lambda}_j.$$

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. In general, a class of middle-grounds between 1-block and d-block EBA can be defined as follows. Choose cut-off integers $1 < \tau_1 < \tau_2 < \cdots < \tau_k < d$ with $1 \leq k < d$. Also let $\tau_0 = 1$. Then, for $\tau_{l-1} \leq r < \tau_l$ let

$$\tilde{\lambda}_r = \frac{1}{\tau_l - \tau_{l-1}} \sum_{\tau_{l-1} \leq j < \tau_l} \hat{\lambda}_j, \tag{4}$$

and for $\tau_k \leq r \leq d$,

$$\tilde{\lambda}_r = \frac{1}{d - \tau_k} \sum_{\tau_k \leq j \leq d} \hat{\lambda}_j.$$

Let us denote this choice by $\tilde{\lambda}(\tau) = (\tilde{\lambda}_1(\tau), \ldots, \tilde{\lambda}_r(\tau))'$. The proposed p-value estimator is then

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

The cut-offs τ defining the blocks may appear with (approximately) equal distance, such that $\hat{p}_{n,3}$ is obtained from three (approximately) equally-sized blocks, and $\hat{p}_{n,4}$ from four (approximately) equally-sized blocks. For instance, with $d = 35$ and four blocks, the block sizes are $9, 9, 9$ and 8. In the current study we investigated four EBA tests obtained from equally-sized blocks: At one extreme is the 1-block SB test, and at the other extreme is singleton blocks, i.e. the full use of all estimated eigenvalues. We refer to this latter test as EBAF, whose p-value is given by (3). In between these two extremes we considered two midde-ground tests, namely the split-half, denoted by EBA2, and the use of four blocks, denoted by EBA4.

Instead of insisting that the blocks should have equal sizes, another strategy is to use a clustering algorithm. Such algorithms iteratively form blocks of eigenvalues of possibly unequal sizes, in order to minimize the variability within each block while maximising the between-block variance. They start with some set of blocks and then adjusts these blocks iteratively to reduce the sum of squared deviations in each class. In the present study we employed both the natural breaks classification of Jenks (1967), where the number of blocks is pre-specified by the user, and a clustering method proposed by Wang and Song (2011) where the number of blocks is chosen by an optimization algorithm. The output of the Jenks algorithm is then the grouping of eigenvalues into the pre-specified number of blocks. In the current study we investigated the Jenks method with 2 or 4 blocks. Replacing the eigenvalues in each block by the block average yields the tests EBA2J and EBA4J, respectively. The output of the Wang and Song (2011) method is the grouping of eigenvalues into the optimal number of blocks. We denote by EBAA the test obtained by replacing eigenvalues in each block by the block average in each of the automatically chosen blocks.

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. 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

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

and $T_n = nF(s, \sigma(\hat{\theta}))$ and $\tilde{T}_n = nF(s, \sigma(\tilde{\theta}))$. Under H_0 and the conditions of Lemma 1 (iv) in Satorra (1989) the difference statistic converges as

$$\tilde{T}_n - T_n \xrightarrow[n\to\infty]{D} \sum_{j=1}^{m} \alpha_j Z_j^2, \qquad Z_1, \ldots, Z_m \sim N(0,1) \text{ IID}, \tag{5}$$

where $\alpha_1, \ldots, \alpha_m$ are the m non-zero eigenvalues of $U_d\Gamma$, where $U_d = \tilde{U} - U$ has rank m. Distribution-free consistent estimators $\hat{U}_d$ and $\hat{\Gamma}$ for U_d and Γ are found and discussed in

Satorra and Bentler (2001).

In the next section we illustrate the EBA procedures on a real-world data sample, followed by a section where we evaluate, both asympotically and in finite samples, the performance of EBA. Probabilities of the type (3) were calculated using the R package CompQuadForm (Duchesne & De Micheaux, 2010), while model estimation and eigenvalue extraction were done with lavaan (Rosseel, 2012).

Block-formation by clustering methods were done using R packages BAMMtools (Rabosky et al., 2014) for the Jenks method and Ckmeans.1d.dp for the method of Wang and Song (2011). R code demonstrating the use of these packages may be found in the appendix.

Example

We consider data from a study (Foldnes, 2017) conducted among $n = 98$ students at a business school, where items from the shortened version of the Attitudes Toward Mathematics Inventory (Lim & Chapman, 2013) were used to model the correlation between enjoyment of mathematics (ENJ) and self-confidence (SC) in mathematics. The model depicted is depicted in Figure 1, which has 13 degrees of freedom.

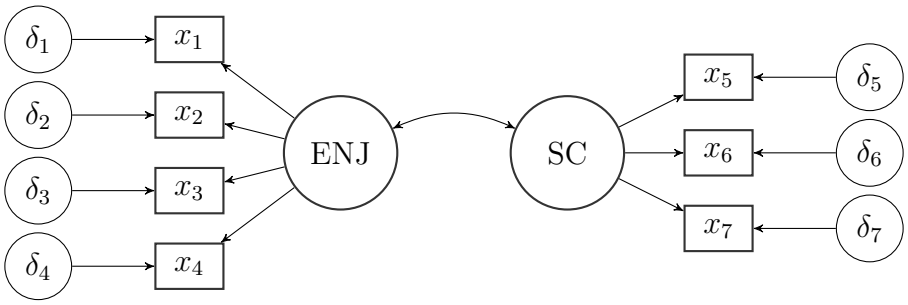


Figure 1. Modeling enjoyment of mathematics and self-confidence in mathematics.

Two estimation methods, DWLS and ML, were considered, with test statistics $T_{\text{DWLS}} = 7.90$ and $T_{\text{ML}} = 25.26$. In each case we extracted the 13 estimated eigenvalues from $\hat{U}\hat{\Gamma}$. These are the weights used in EBAF, and are given in row 1 and 9 of Table 1,

which also contains the weights used by T_n, ML, SB, EBA2, EBA2J, EBA4, EBA4J and EBAA. For both estimation methods, SB p-values are smaller than the p-values for the other robust tests, which is unsurprising, given the reported tendency of SB to overreject correct models under non-normality (e.g., Foldnes & Olsson, 2015). Also, under DWLS, the automatic EBAA test yields only one cluster, so that EBAA in that condition coincided with SB, while under ML, EBAA has two clusters and is equivalent to EBA2. Overall, the p-values vary moderately among the tests.

For the ML case, we also plotted the probability density function of X_{approx} for SB, SS, CF and three EBA tests in Figure 2. The p-values associated with SS and CF were 0.223 and 0.195, respectively. We see that in this real-world situation, the distributions of CF and the three EBA tests are quite similar to each other. The SB and SS tests are seen to be based on distributions that differ quite a lot from those of the CF and the EBA tests.

In summary, Table 1 and Figure 2 indicate that there is some variability among the established and newly proposed tests. For a practitioner, the question remains about which of these tests should be used for evaluating the model. As shown in the next sections, there is unfortunately no single robust test that performs best under all possible conditions of sample size and underlying distribution. A possible way to select a test in a given situation is to simulate data whose distribution is close to that of the observed data. The flexible data-generating method recently proposed by Grønneberg and Foldnes (2017) may be used to emulate the characteristics of the observed sample. One can then observe which of the test candidates performs best on average on the simulated data. However we consider this idea outside the scope of the present study.

In the next section we proceed by evaluating the performance of the EBA tests and the established robust tests by Monte Carlo, in order to gain some insight into the systematic differences with respect to empirical Type I error control.

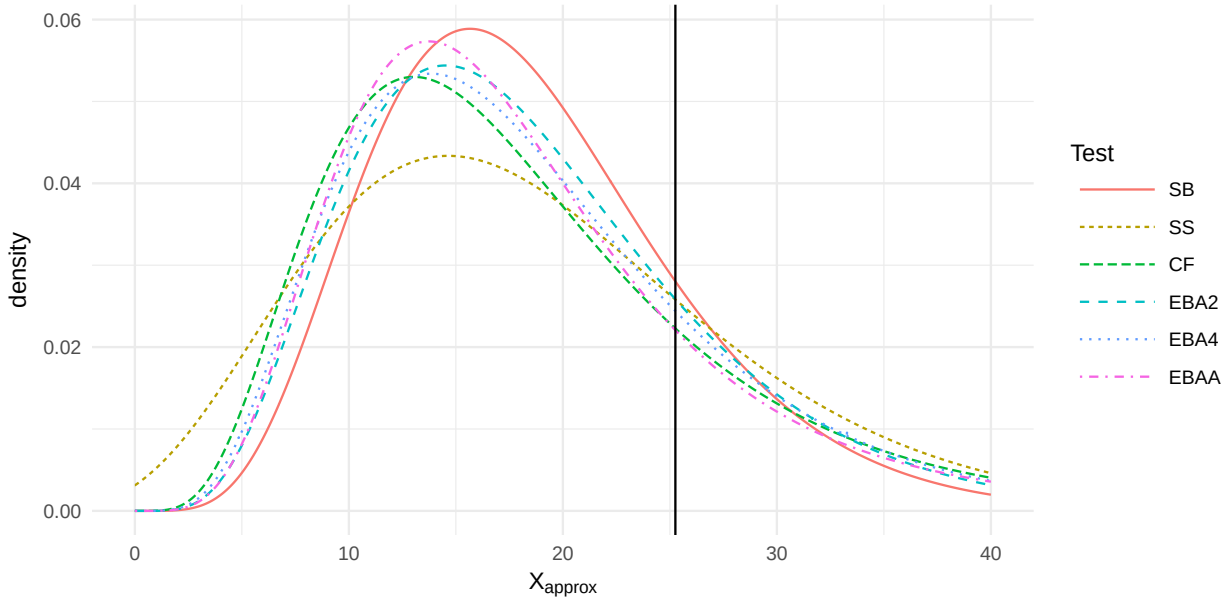


Figure 2. Probability density curves of X_{approx} for the case of testing a two-factor model based on $n = 98$ observations with the ML estimator. Vertical line represents $T_{\text{ML}} = 25.26$. The areas below curves to the right of this line correspond to p-values. SB=Satorra-Bentler. SS=scaled and shifted. CF=scaled F. EBA2 and EBA4= eigenvalue block approximation with 2 and 4 equally-sized blocks. EBAA= automatic eigenvalue clustering.

	Method	1	2	3	4	5	6	7	8	9	10	11	12	13	p
DWLS	EBAF	0.81	0.56	0.49	0.40	0.32	0.23	0.21	0.16	0.12	0.11	0.09	0.08	0.05	0.029
	T	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.850
	SB	0.28	0.28	0.28	0.28	0.28	0.28	0.28	0.28	0.28	0.28	0.28	0.28	0.28	0.009
	EBA2	0.43	0.43	0.43	0.43	0.43	0.43	0.43	0.10	0.10	0.10	0.10	0.10	0.10	0.019
	EBA4	0.56	0.56	0.56	0.56	0.26	0.26	0.26	0.13	0.13	0.13	0.08	0.08	0.08	0.025
	EBA2J	0.56	0.56	0.56	0.56	0.15	0.15	0.15	0.15	0.15	0.15	0.15	0.15	0.15	0.024
	EBA4J	0.81	0.48	0.48	0.48	0.26	0.26	0.26	0.10	0.10	0.10	0.10	0.10	0.10	0.028
	EBAA	0.28	0.28	0.28	0.28	0.28	0.28	0.28	0.28	0.28	0.28	0.28	0.28	0.28	0.009
ML	EBAF	5.46	2.38	2.01	1.52	1.40	1.12	1.08	0.95	0.67	0.61	0.53	0.42	0.36	0.193
	T	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.021
	SB	1.42	1.42	1.42	1.42	1.42	1.42	1.42	1.42	1.42	1.42	1.42	1.42	1.42	0.167
	EBA2	2.14	2.14	2.14	2.14	2.14	2.14	2.14	0.59	0.59	0.59	0.59	0.59	0.59	0.186
	EBA4	2.84	2.84	2.84	2.84	1.20	1.20	1.20	0.74	0.74	0.74	0.43	0.43	0.43	0.192
	EBA2J	5.46	1.09	1.09	1.09	1.09	1.09	1.09	1.09	1.09	1.09	1.09	1.09	1.09	0.181
	EBA4J	5.46	2.20	2.20	1.21	1.21	1.21	1.21	1.21	0.52	0.52	0.52	0.52	0.52	0.192
	EBAA	5.46	1.09	1.09	1.09	1.09	1.09	1.09	1.09	1.09	1.09	1.09	1.09	1.09	0.181

Table 1

Estimated λ_j, $j = 1, \ldots, 13$, in first row (EBAF), together with $\tilde{\lambda}_j$ for other methods. DWLS= diagonally weighted least squares estimator. ML= maximum likelihood estimator. EBAF= Full eigenvalue estimation; T = χ^2 test; SB=Satorra-Bentler; EBAi=i-block equal-size eigenvalue blocks; EBAiJ= i-block Jenks eigenvalue blocks; EBAA = automatic eigenvalue clustering. p= p-value.

Method

The performance of six EBA procedures and three established test statistics were assessed, both theoretically and empirically. The EBA procedures investigated are EBAF, EBA2, EBA4, EBA2J, EBA4J and EBAA, while the established test statistics are SB, SS and CF. In addition we included the oracle test in (2), here denoted by OR. These test procedures are not specifically linked to ML estimation and its associated test statistic T_{ML}. In each evaluation case we therefore included a second estimator, namely DWLS with its associated test statistic T_{DWLS}.

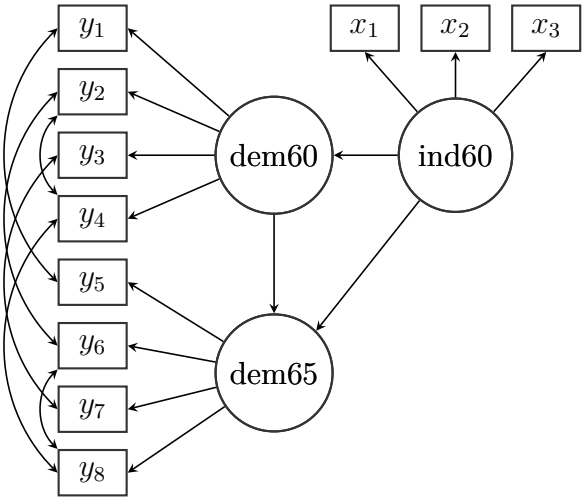
Theoretically, asymptotic rejection rates were computed based on eigenvalues extracted from the population matrix $U\Gamma$. This is possible due to a recently proposed method (Foldnes & Grønneberg, 2017) that allows the exact calculation of Γ, and consequently, of λ_j. For the EBA tests we solved the equation $P(\sum \lambda_j Z_j^2 > c) = 0.05$ numerically for c, and then the asymptotic rejection rate was calculated as $P(\sum \tilde{\lambda}_j Z_j^2 > c)$, where the $\tilde{\lambda}_j$ depend on the block-formation strategy.

Empirically, we conducted two simulation studies. Study 1 involves the testing of a single correctly specified model, while Study 2 involves the testing of two correctly specified nested models. The asymptotic and empirical rejection rates reported in the present article were computed at the $\alpha = 0.05$ level of significance.

Models

Our model is the political democracy model discussed by Bollen in his textbook (Bollen, 1989), see Figure 3, 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 model, referred to as $\mathcal{M}_1$, has $d = 35$ degrees of freedom. Study 1 involves tests of correct model specification based on $\mathcal{M}_1$. In Study 2, we considered testing a constrained model $\mathcal{M}_0$ against $\mathcal{M}_1$. $\mathcal{M}_0(d = 45)$ is nested within $\mathcal{M}_1$, and imposes ten correctly specified

Figure 3. Bollen's political democracy model.



equalities on unique and residual covariances.

Data generation

In order to theoretically evaluate the performance of the test statistics, and to evaluate the finite-sample performance of the oracle OR, the population values λ in (1) must be exactly calculated. Recently, Foldnes and Grønneberg (2017) presented an algorithm for obtaining Γ under distributions produced by the Vale-Maurelli (VM) transform (Vale & Maurelli, 1983). We therefore used the VM transform in the present study. We calculated Γ and U (for both ML and DWLS) and obtained population eigenvalues λ under each distributional condition. Data generation was achieved by fixing the parameters in the model, and using the model-implied covariance matrix as the target covariance matrix for the VM transform. Two nonnormal distributional conditions, denoted by D_1 and D_2, were specified by vectors containing heterogeneous skewness s and kurtosis k for the 11 univariate marginals as follows. For D_1, $s = (1, 1, 1, 1, 1, 2, 2, 2, 2, 2, 2)$ and $k = (5, 5, 5, 5, 5, 10, 10, 10, 10, 10, 10)$. For D_2, $s = (2, 2, 2, 2, 2, 3, 3, 3, 3, 3, 3)$ and $k = (7, 7, 7, 7, 7, 21, 21, 21, 21, 21, 21)$. With the terminology used by Curran, West, and Finch (1996), distributions D_1 and D_2 might be said to represent moderate and severe

nonnormality, respectively. For the simulation studies, replications leading to nonconvergence or improper solutions were removed from further analysis. In each cell we simulated 10^4 replications with proper solutions, resulting in a standard error of 0.0022 for the empirical rejection rate, given that the true Type I error rate was 0.05.

RESULTS

Asymptotic performance

Study 1. In each of six conditions (two estimators × three distributions), the 35 non-zero population eigenvalues were calculated. In Figure 4, violin plots give the distribution of these eigenvalues in each condition. With estimator ML, eigenvalues tend to get larger and span a larger range when moving from normality (N) via moderate nonnormality (D_1) to the severe nonnormality (D_2). With estimator DWLS, the eigenvalues are not much affected by the underlying distribution, indicating that the large-sample distribution of T_{DWLS} is not sensitive to the underlying distribution.

Population eigenvalues were then used to compute asymptotic rejection rates for each test statistic, see Table 2. Since T_{DWLS} is not distributed as a chi-square under any distribution, T_{DWLS} rejection rates are far off the nominal level. T_{ML} is correctly specified for normal data, and hence has the optimal rejection rate of 0.05 under N, but has highly inflated rejection rates under non-normality. The SB scaling yields too high rejection rates under DWLS, but close to nominal rates under ML, even with highly non-normal data. The SS test performs much better than SB with DWLS, and is also preferrable to SB under ML. CF reaches almost perfect rejection rates under both DWLS and ML. The increasing asymptotic performance in the sequence SB, SS and CF reflects that SB only matches the first, SS the two first, and CF the three first moments of the weighted sum in (1). The full eigenvalue approximation EBAF yields perfect asymptotic Type I error control in all conditions, which is in line with the theory, since EBAF is a consistent test. The other EBA methods generally lead to underrejection, especially under DWLS, with

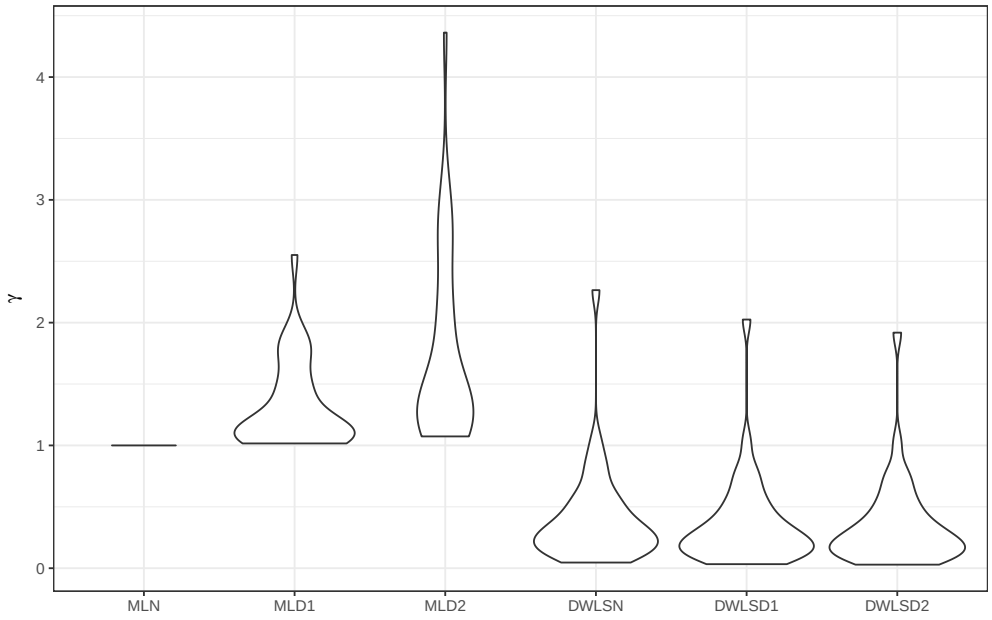


Figure 4. Study 1: Violin plots for the distribution of 35 population eigenvalues. MLN, MLD1 and MLD2 refer to ML estimation under multivariate normality, moderate and severe nonnormality, respectively. DWLSN, DWLSD1 and DWLSD2 refer to DWLS estimation under multivariate normality, moderate and severe nonnormality, respectively.

EBA2 performing the worst, while EBA4J attains almost perfect Type I error control.

Study 2. The chi-square difference test has ten degrees of freedom, and the corresponding oracle eigenvalues are presented in Table 3. Similar to the pattern in Figure 4, the eigenvalues are much more sensitive to the underlying distribution under the ML estimator, compared to the DWLS estimator. With ML, the eigenvalues become larger and more varied with increasing nonnormality.

Table 4 contains asymptotic rejection rates for nested model testing. SB overrejects in all conditons except for ML under normality. SS overrejects very slightly, while the F test achieves perfect rejection rates. The EBA approximations tend to underreject the null, but less so compared to Study 1, with EBA4J achieving almost perfect Type I error control in all conditions.

	dist	T	SB	SS	CF	EBAF	EBA2	EBA4	EBA2J	EBA4J	EBAA
DWLS	N	0.000	0.103	0.055	0.049	0.050	0.025	0.035	0.039	0.049	0.047
	D1	0.000	0.106	0.055	0.050	0.050	0.025	0.035	0.037	0.048	0.037
	D2	0.000	0.106	0.055	0.050	0.050	0.026	0.036	0.037	0.048	0.037
ML	N	0.050	0.050	0.050	0.050	0.050	0.050	0.050	0.050	0.050	0.050
	D1	0.347	0.056	0.051	0.050	0.050	0.047	0.049	0.049	0.050	0.050
	D2	0.748	0.065	0.052	0.050	0.050	0.043	0.047	0.047	0.050	0.048

Table 2

Study 1: Asymptotic rejection rates. T=χ^2 test. SB=Satorra-Bentler. SS=scaled and shifted. CF=scaled F test. EBAF= Full eigenvalue estimation; EBAi=i-block equal-size eigenvalue blocks; EBAiJ= i-block Jenks eigenvalue blocks; EBAA = automatic eigenvalue clustering.

DWLS	N	0.620	0.364	0.315	0.276	0.253	0.234	0.191	0.183	0.128	0.073
	D_1	0.503	0.326	0.288	0.244	0.184	0.174	0.157	0.133	0.102	0.058
	D_2	0.511	0.340	0.294	0.246	0.181	0.175	0.151	0.125	0.099	0.055
ML	N	1	1	1	1	1	1	1	1	1	1
	D_1	5.854	4.090	2.875	2.679	2.436	2.275	2.133	1.865	1.755	1.490
	D_2	11.280	7.607	4.724	3.990	3.702	3.564	3.276	2.784	2.629	2.093

Table 3

Study 2. The population eigenvalues of $U_d\Gamma$, rounded to three decimal places. N= normal, D_1=moderate non-normality, D_2=severe non-normality. DWLS=diagonally weigthed least squares estimator. ML=normal-theory maximum likelihood estimator.

	dist	T	SB	SS	CF	EBAF	EBA2	EBA4	EBA2J	EBA4J	EBAA
DWLS	N	0.000	0.068	0.052	0.050	0.050	0.043	0.046	0.045	0.049	0.033
	D_1	0.000	0.070	0.052	0.050	0.050	0.043	0.047	0.045	0.049	0.032
	D_2	0.000	0.071	0.052	0.050	0.050	0.043	0.047	0.045	0.049	0.031
ML	N	0.050	0.050	0.050	0.050	0.050	0.050	0.050	0.050	0.050	0.050
	D_1	0.732	0.063	0.051	0.050	0.050	0.044	0.047	0.048	0.050	0.048
	D_2	0.928	0.070	0.052	0.050	0.050	0.040	0.045	0.048	0.050	0.048

Table 4

Study 2: Asymptotic rejection rates. N= normal, D_1=moderate non-normality, D_2=severe non-normality. DWLS=diagonally weigthed least squares estimator. ML=normal-theory maximum likelihood estimator. T= χ^2 test. SB=Satorra-Bentler. SS=scaled and shifted. CF=scaled F test. EBAF= full eigenvalue approximation. EBAi= i-block eigenvalue approximation. EBAiJ= i-block Jenks eigenvalue approximation. EBAA = automatic eigenvalue clustering.

Finite-sample performance

Study 1. Finite-sample rejection rates for testing $\mathcal{M}_1$ are given in Table 5. We discuss the DWLS case first, where, generally, all test statistics are quite robust to the underlying distribution. SB produces consistently too high error rates, at about 0.1. SS error rates are consistently below the nominal level $\alpha = 0.05$, but approaches α with increasing sample size. Under conditions of small sample size and non-normality, SS has rejections rates below 0.03. CF rejection rates are close to those of SS, but are consistently lower. EBAF and CF have almost identical rejection rates across all conditions. In contrast to SS/CF/EBAF, the two-block EBA2 consistently has rejection rates well above α, and has poor Type I error control. EBA4 performs better than EBA2, with rejection rates lying generally between those of SS/CF/EBAF on one hand, and EBA2 on the other hand. EBA2J lies slightly below EBA4 in terms or rejection rates, while EBA4J performs quite poorly with rejection rates below those of SS. EBAA has higher rejection rates than SS, and lower than EBA2J. To sum up, the procedures with best Type I error control across all DWLS conditions are SS, EBA4, EBA2J and EBAA.

Next we consider ML estimation, where the test statistics are more sensitive to the underlying distribution than was the case for DWLS. Note that T yields exactly the same rejection rates as the oracle OR, under multivariate normal data N. Under non-normality, however, rejection rates of T become very large. SB again has inflated rejection rates, especially under non-normality and small sample size. SS has too low rejection rates, with especially poor performance under non-normality. Again, CF and EBAF have near identical rejection rates across all conditions, slightly below those of SS. EBA2 has very good Type I error control in all conditions. EBA4 outperforms SS/CF/EBAF, but still has poorer error control than EBA2. Under non-normality, the clustering procedures EBA2J, EBA4J and EBAA have rejection rates lower than those of EBA4. To sum up, across all ML conditions, EBA2 by far had the best Type I error control, with EBA4 as a runner-up.

Study 2. Finite-sample rejection rates for nested model testing are given in Table 6.

We discuss the DWLS case first. The tests are sensitive to the underlying distribution. All tests produce rejection rates above the nominal level, especially under non-normality. SB has the highest rejection rates. EBA2 and EBAA have lower rejection rates. However, the group of tests SS, F, EBAF, EBA4, EBA2J and EBA4J has equal performance across all conditions, and attains better Type I error than SB, EBA2 and EBAA.

Under ML estimation, the situation is similar to the DWLS case, with a pattern of high rejection rates, decreasing toward the nominal level with increasing sample size. SB has the highest rejection rates. EBAA and EBA2 have lower rejection rates, but these are higher than the rather similar rejection rates in the group of SS, F, EBAF, EBA4, EBA2J and EBA4J. This group achieves the lowest rejection rates, and so represent the best-performing tests in terms of Type I error control.

	n	Distr	T	SB	SS	CF	EBAF	EBA2	EBA4	EBA2J	EBA4J	EBAA	OR
DWLS	100	N	.000	.104	.039	.035	.034	.064	.047	.044	.036	.042	.059
		D_1	.000	.109	.030	.025	.024	.064	.041	.034	.026	.030	.099
		D_2	.000	.110	.023	.020	.019	.057	.032	.025	.020	.021	.133
	300	N	.000	.107	.049	.044	.044	.072	.057	.054	.046	.052	.057
		D_1	.000	.113	.045	.039	.039	.073	.055	.051	.041	.046	.082
		D_2	.000	.118	.039	.033	.032	.071	.049	.043	.034	.039	.106
	1000	N	.000	.099	.050	.045	.045	.071	.060	.058	.046	.054	.049
		D_1	.000	.107	.050	.045	.045	.074	.061	.058	.047	.055	.064
		D_2	.000	.117	.048	.042	.042	.076	.057	.055	.044	.052	.077
ML	100	N	.078	.085	.044	.039	.039	.053	.044	.048	.040	.062	.078
		D_1	.277	.105	.021	.017	.016	.053	.029	.024	.018	.022	.041
		D_2	.516	.125	.014	.010	.009	.055	.026	.016	.010	.013	.018
	300	N	.059	.062	.046	.044	.044	.048	.045	.048	.045	.062	.059
		D_1	.317	.072	.024	.020	.020	.044	.031	.026	.022	.025	.051
		D_2	.617	.081	.015	.011	.011	.042	.024	.017	.012	.015	.039
	1000	N	.049	.050	.046	.045	.045	.047	.045	.047	.045	.050	.049
		D_1	.324	.059	.031	.028	.028	.043	.036	.034	.029	.034	.050
		D_2	.689	.066	.023	.019	.019	.043	.031	.025	.019	.024	.046

Table 5

Study 1: Rejection rates. N=normality; D_1=moderate nonnormality. D_2=severe nonnormality. DWLS=diagonally weigthed least squares. ML= maximum likelihood. T= χ^2 test. SB=Satorra-Bentler. SS=scaled and shifted. CF=scaled F. EBAF= full eigenvalue approximation. EBAi= i-block clustering. EBAiJ= i-block Jenks clustering. EBAA = Automatic clustering. OR=oracle.

	n	Distr	T	SB	SS	CF	EBAF	EBA2	EBA4	EBA2J	EBA4J	EBAA	OR
DWLS	100	N	.000	.096	.067	.065	.063	.075	.068	.069	.064	.091	.093
		D_1	.004	.236	.176	.171	.169	.194	.179	.179	.170	.209	.293
		D_2	.022	.313	.236	.231	.229	.260	.240	.237	.230	.270	.382
	300	N	.000	.076	.054	.052	.052	.062	.057	.058	.052	.074	.061
		D_1	.000	.151	.107	.104	.102	.121	.110	.111	.104	.136	.170
		D_2	.002	.191	.139	.136	.134	.155	.144	.141	.135	.166	.221
	1000	N	.000	.066	.051	.049	.049	.055	.052	.053	.049	.066	.051
		D_1	.000	.106	.075	.074	.073	.083	.077	.078	.074	.100	.094
		D_2	.000	.126	.092	.089	.088	.100	.093	.094	.089	.117	.125
ML	100	N	.071	.082	.068	.066	.066	.070	.068	.069	.067	.082	.071
		D_1	.627	.198	.131	.128	.126	.154	.139	.135	.127	.164	.018
		D_2	.855	.276	.180	.176	.172	.212	.188	.184	.173	.220	.008
	300	N	.054	.058	.054	.054	.054	.054	.054	.054	.054	.058	.054
		D_1	.682	.124	.084	.082	.080	.098	.088	.086	.082	.102	.035
		D_2	.886	.165	.110	.107	.105	.129	.115	.112	.105	.133	.027
	1000	N	.050	.052	.051	.051	.051	.051	.051	.051	.051	.052	.050
		D_1	.711	.084	.062	.060	.059	.069	.064	.063	.060	.073	.048
		D_2	.911	.111	.073	.071	.070	.087	.076	.075	.071	.087	.043

Table 6

Study 2: Rejection rates. N=normality; D_1=moderate nonnormality. D_2=severe nonnormality. DWLS=diagonally weigthed least squares. ML= maximum likelihood. T= χ^2 test. SB=Satorra-Bentler. SS=scaled and shifted. CF=scaled F. EBAF= full eigenvalue approximation. EBAi= i-block clustering. EBAiJ= i-block Jenks clustering. EBAA = Automatic clustering. OR=oracle.

Discussion

The performance of the established and proposed new statistics have been studied, both asymptotically and in finite samples. The most important case for a practitioner is of course finite-sample performance. Consistent patterns among the test procedures were found across sample sizes and underlying distribution in both Study 1 and Study 2. For Study 1, the results in Table 5 suggest the following grouping of tests that perform similarly, ranked according to increasing rejection rates:

$$\text{Study 1: CF/EBAF/EBA4J} < \text{SS/EBAA/EBA2J} < \text{EBA4} < \text{EBA2} < \text{SB},$$

while the results in Table 6 suggest the following grouping, ranked according to increasing rejection rates:

$$\text{Study 2: SS/CF/EBAF/EBA4/EBA2J/EBA4J} < \text{EBA2} < \text{EBAA} < \text{SB}.$$

Also, some general observations holding across sample size, distributions, estimators and models might be made: SB has the highest rejection rates. CF consistently has slightly lower rejection rates than SS. Remarkably, CF and EBAF have almost identical rejection rates in both models, for all sample sizes, distributions and estimators. This echoes the findings of (Wu & Lin, 2016). In general, the EBA procedures perform similarly to SS and CF, with the exception of EBAA and EBA2, which tend to have somewhat higher rejection rates than SS/CF, but lower than SB.

Comparing the performance of EBA2 and EBAF across the two studies, it is noticeable that EBAF performed best in Study 2 (10 eigenvalues), while EBA2 performed best in Study 1 (35 eigenvalues). A possible explanation for this pattern is that in Study 2 there are more sample observations for each estimated eigenvalue. Intuitively, the eigenvalues are therefore estimated with higher precision in Study 2 compared to Study 1. The full use of the individual eigenvalues in Study 2 is more warranted than under conditions such as in Study 1, where there are far fewer observations per eigenvalue. In this

latter condition it is therefore not surprising that the 2-block method is found superior to EBAF.

We now turn to the question of evaluation. It is important to notice that there are two, sometimes conflicting, ways of evaluating test statistics. From a practical point of view, the important question is: How well does the test control Type I error rates? This is the evaluation criterion in most simulation studies. However, the tests under consideration in the present study were designed to emulate the oracle distribution in (2). So theoretically, the important question is: How well does the test approximate the oracle? Of course, it is hoped that these two evaluation criterions merge, and they certainly will for very large sample sizes. However, Tables 5 and 6 demonstrate that under realistic sample sizes, the oracle does not always achieve acceptable Type I error control. In some conditions it might therefore happen that a test statistic does a poor job approximating the oracle OR, but by some coincidence achieves good Type I error control. Consider for instance the condition in Study 1 of ML estimation under severe non-normality and the smallest sample size. Here EBA2 outperforms all the other tests by a large margin, with a rejection rate of 0.055. However, the oracle has not yet reached its asymptotic limit of $\alpha = 0.05$, having a Type I rejection rate of only 0.018. Hence EBA2 does a very good job of controlling Type I error rates, while failing to achieve its theoretical aim of emulating the oracle. On the other hand, EBA2J matches the oracle rejection rate closely with a rejection rate of 0.016, but in terms of Type I error control this is unacceptably low.

Broadly speaking, evaluation in terms of Type I error control gave the following results. In Study 1, with $d = 35$ single model testing, the group SS, EBA4, EBA2J and EBAA performed similarly, and attained the best Type I error control under DWLS, while EBA2 clearly outperformed all other tests under ML. In Study 2, with $d = 10$ nested model testing, the tests SS, CF, EBAF, EBA4, EBA2J, EBA4J performed equally well, and better than SB, EBA2 and EBA4.

The second evaluation criterion considers how well the tests emulate the oracle. In

Study 1, EBA2 performed the best, with the exception of DWLS under normality, where EBA4, EBA2J and EBAA more closely matched the oracle rejection rates. In Study 2, the oracle was best approached by EBAA under DWLS, and by CF, EBAF, EBA4J under ML.

Conclusion

Recently two test procedures, the scaled and shifted test SS (Asparouhov & Muthén, 2010), and the scaled F test CF (Wu & Lin, 2016) have been proposed based on approximating the asymptotic distribution of a weighted sum of chi-square variates in (1). The SS and CF procedures match, respectively, the first two and the first three moments of the asympotic distribution, and are hence theoretically superior to the original scaling procedure of Satorra and Bentler (1988). In the present paper we have theoretically and empirically demonstrated, in the context of a specific model, that SS and CF outperforms the SB procedure both for single and nested model testing. In accord with earlier simulation studies, we therefore recommend SS and CF over the SB procedure, although SS and CF both tend to underreject correct models. Note that this recommendation still holds under normally distributed data.

We have also proposed new approximations to the weighted sum of i.i.d. chi-square variates, based on arranging eigenvalues in blocks and replacing them by average values. This introduces a whole new class of approximations to the asymptotic distribution of test statistics in structural equation modeling. We have compared six members of this class to the existing procedures CF and SB, both theoretically and empirically. In terms of correct Type I error control, the new procedures perform as well as SS and CF, and in some cases better. For instance, in the important case of ML testing a single model under non-normality, a two-block eigenvalue approximation was found to outperform all other statistics.

Given the established SS and CF tests, and several well-performing EBA tests, the question then remains how one might perform model testing in a practical situation. In

most cases, tests like CF, SS and the EBA variants, seem to result in similar model fit evaluations, as was the case for the illustrative example in the present study. However, in some situations the tests might assess model fit differently. In such cases it would be recommended to report several test statistics. We suggest reporting one fixed-block and one dynamic-block EBA procedure. For instance SS, EBA2 and EBA4J could be reported.

References

Amemiya, Y., Anderson, T. W., et al. (1990). Asymptotic chi-square tests for a large class of factor analysis models. *The Annals of Statistics*, *18*(3), 1453–1463.

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

Bentler, P. M., & Yuan, K.-H. (1999). Structural equation modeling with small samples: Test statistics. *Multivariate Behavioral Research*, *34*(2), 181–197.

Bollen, K. A. (1989). *Structural equations with latent variables*. New York: Wiley. doi: 10.1002/9781118619179

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.

Browne, M. W., & Shapiro, A. (1988). Robustness of normal theory methods in the analysis of linear latent variate models. *British Journal of Mathematical and Statistical Psychology*, *41*(2), 193–208.

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*(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 & Data Analysis*, *54*(4), 858–862.

Foldnes, N. (2017). The impact of class attendance on student learning in a flipped classroom. *Nordic Journal of Digital Literacy*, *12*(1-2), 8–18.

Foldnes, N., & Grønneberg, S. (2017). The asymptotic covariance matrix and its use in

simulation studies. *Structural Equation Modeling: A Multidisciplinary Journal*, 1–16.

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.

Grønneberg, S., & Foldnes, N. (2017). Covariance model simulation using regular vines. *Psychometrika*, 1–17.

Jenks, G. F. (1967). The data model concept in statistical mapping. *International yearbook of cartography*, *7*(1), 186–190.

Lim, S. Y., & Chapman, E. (2013). Development of a short form of the attitudes toward mathematics inventory. *Educational Studies in Mathematics*, *82*(1), 145–164.

Mooijaart, A., & Bentler, P. M. (1991). Robustness of normal theory statistics in structural equation models. *Statistica Neerlandica*, *45*(2), 159–171.

Nevitt, J., & Hancock, G. R. (2004). Evaluating small sample approaches for model test statistics in structural equation modeling. *Multivariate Behavioral Research*, *39*(3), 439–478.

Rabosky, D., Grundler, M., Anderson, C., Title, P., Shi, J., Brown, J., ... Larson, J. (2014). Bammtools: an R package for the analysis of evolutionary dynamics on phylogenetic trees. *Methods in Ecology and Evolution*, *5*, 701-707.

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

Satorra, A. (1989). Alternative test criteria in covariance structure analysis: A unified approach. *Psychometrika*, *54*(1), 131–151.

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*.

Satorra, A., & Bentler, P. (1994). Corrections to test statistics and standard errors in covariance structure analysis. In A. Von Eye & C. Clogg (Eds.), *Latent variable analysis: applications for developmental research* (chap. 16). Sage.

Satorra, A., & Bentler, P. M. (1990). Model conditions for asymptotic robustness in the analysis of linear relations. *Computational Statistics & Data Analysis*, *10*(3), 235–249.

Satorra, A., & Bentler, P. M. (2001). A scaled difference chi-square test statistic for moment structure analysis. *Psychometrika*, *66*(4), 507–514.

Shapiro, A. (1983). Asymptotic distribution theory in the analysis of covariance structures. *South African Statistical Journal*, *17*(1), 33–81.

Shapiro, A. (1987). Robustness properties of the mdf analysis of moment structures. *South African Statistical Journal*, *21*(1), 39–62.

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

Wang, H., & Song, M. (2011). Ckmeans.1d.dp: Optimal k-means clustering in one dimension by dynamic programming. *The R Journal*, *3*(2), 29–33. Retrieved from `https://journal.r-project.org/archive/2011-2/RJournal_2011-2_Wang+Song.pdf`

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.

Yuan, K.-H. (2005). Fit indices versus test statistics. *Multivariate behavioral research*, *40*(1), 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*(2), 273–291.

Appendix

R code

```
# R version 3.3.1
library(lavaan) # Version 0.5-22
library(CompQuadForm) # Version 1.4.2
library(BAMMtools) # Version 2.1.6
library(Ckmeans.1d.dp) # Version 4.0.1

#sample size, skewness and kurtosis
n=300L
skewness=2L
kurtosis=10L
seed=1


#Illustration based on a two-factor model
# specify population model
population.model <-
    "f1 =~ x1 + 0.8*x2 + 1.2*x3+0.2*x4;
     f2 =~ x5 + 0.8*x6 + 1.2*x7+0.2*x8; f1 ~~0.5*f2"

#model to be estimated, has equality constraints on three residual variances
my.model <- "f1 =~ x1+x2+x3+x4; f2 =~ x5+x6+x7+x8; "

#simulate non-normal dataset
set.seed(seed)
my.dat = simulateData(population.model, sample.nobs=n,
                      skewness=rep(skewness,8), kurtosis=rep(kurtosis, 8))

#ML estimation
f = sem(my.model, data=my.dat)

#pvalues for default NTML and SB tests:
sem(my.model, data=my.dat, test="SB")

#extract test statistic T
T = fitmeasures(f, "chisq")
#Extract U*Gamma
UG <- inspect(f, "UGamma")
#The estimated eigenvalues
df = fitmeasures(f,"DF")
eig.hat <- Re(eigen(UG)$values[1:df])

########
## p-values for various tests, based on eigenvalues
########

#NTML
pNTML <- imhof(T, rep(1, df))$Qq

#SB
pSB <-  imhof(T, rep(mean(eig.hat), df))$Qq

#EBAF
```

```
pEBAF <- imhof(T, eig.hat)$Qq

#EBA2
eigs <- c(rep(mean(eig.hat[1:ceiling(df/2)]),ceiling(df/2)),
          rep(mean(eig.hat[(ceiling(df/2)+1):df]),df-ceiling(df/2)))
pEBA2 <- imhof(T, eigs)$Qq

#Jenks EBA2
breaks <- getJenksBreaks(eig.hat, k=3)
block1 <- eig.hat[eig.hat <= breaks[2]]; block2=eig.hat[eig.hat > breaks[2]]
eigs <- c(rep(mean(block1), length(block1)), rep(mean(block2), length(block2)))
pEBA2J <- imhof(T, eigs)$Qq

#EBAA
t = Ckmeans.1d.dp(eig.hat)
means <- t$centers
clusters <- t$cluster
eigs <- sapply(clusters, function(x) means[x])
pEBAA <- imhof(T, eigs)$Qq

cat("Simple model testing: \n")
print(round(data.frame(pNTML, pSB, pEBAF, pEBA2, pEBA2J, pEBAA),4))


######
## Nested Model Testing
######

#help function. From lavaan source code.

eigenvalues_diff <- function(m1, m0,  A.method = "exact") { #or delta. Note that shell command lavTestLRT
    has exact, while lav_test_diff_Satorra2000 has default delta.

    # extract information from m1 and m2
    T1 <- m1@test[[1]]$stat
    r1 <- m1@test[[1]]$df

    T0 <- m0@test[[1]]$stat
    r0 <- m0@test[[1]]$df

    # m = difference between the df's'
    m <- r0 - r1
    Gamma <- lavTech(m1, "Gamma") # the same for m1 and m0
    WLS.V <- lavTech(m1, "WLS.V")
    PI <- lavaan:::computeDelta(m1@Model)
    P <- lavTech(m1, "information")
    # needed? (yes, if H1 already has eq constraints)
    P.inv <- lavaan:::lav_model_information_augment_invert(m1@Model,
                                                           information = P,
                                                           inverted = TRUE)
    if(inherits(P.inv, "try-error")) {
        cat("Error! in P.inv \n")
        return(NA)
    }

    A <- lavaan:::lav_test_diff_A(m1, m0, method = A.method, reference = "H1")
```

```
    APA <- A %*% P.inv %*% t(A)
    cSums <- colSums(APA)
    rSums <- rowSums(APA)
    empty.idx <- which( abs(cSums) < .Machine$double.eps^0.5 &
                            abs(rSums) < .Machine$double.eps ^0.5 )
    if(length(empty.idx) > 0) {
        A <- A[-empty.idx,, drop = FALSE]
    }

    # PAAPAAP
    PAAPAAP <- P.inv %*% t(A) %*% solve(A %*% P.inv %*% t(A)) %*% A %*% P.inv

    g = 1
    UG.group <- WLS.V[[g]] %*% Gamma[[g]] %*% WLS.V[[g]] %*%
        PI[[g]] %*% PAAPAAP %*% t(PI[[g]])

    return(Re(eigen(UG.group)$values)[1:m])
}


my.model.restricted <- "f1 =~ x1 + b*x2 + c*x3+d*x4;
                        f2 =~ x5 + b*x6 + c*x7+d*x8;
                        x1~~a*x1;x2~~a*x2;x3~~a*x3;x4~~a*x4;
                        x5~~a*x5;x6~~a*x6;x7~~a*x7;x8~~a*x8;"

f.restricted = sem(my.model.restricted, my.dat)

#the estimated eigenvalues for difference test.  10 df.
eig.hat =eigenvalues_diff(f,f.restricted)

#chisquare difference
T <- fitmeasures(f.restricted, "chisq")-fitmeasures(f, "chisq")
df <- fitmeasures(f.restricted, "DF")-fitmeasures(f, "DF")
#NTML
pNTML <- imhof(T, rep(1, df))$Qq

#SB
pSB <-  imhof(T, rep(mean(eig.hat), df))$Qq

#EBAF
pEBAF <- imhof(T, eig.hat)$Qq

#EBA2
eigs <- c(rep(mean(eig.hat[1:ceiling(df/2)]),ceiling(df/2)),
          rep(mean(eig.hat[(ceiling(df/2)+1):df]),df-ceiling(df/2)))
pEBA2 <- imhof(T, eigs)$Qq

#Jenks EBA2
breaks <- getJenksBreaks(eig.hat, k=3)
block1 <- eig.hat[eig.hat <= breaks[2]]; block2=eig.hat[eig.hat > breaks[2]]
eigs <- c(rep(mean(block1), length(block1)), rep(mean(block2), length(block2)))
pEBA2J <- imhof(T, eigs)$Qq

#EBAA
```

```
t = Ckmeans.1d.dp(eig.hat)
means <- t$centers
clusters <- t$cluster
eigs <- sapply(clusters, function(x) means[x])
pEBAA <- imhof(T, eigs)$Qq

cat("\nNested model testing: \n")
print(round(data.frame(pNTML, pSB, pEBAF, pEBA2, pEBA2J, pEBAA),4))
```