Njål Foldnes, George A. Marcoulides & Ulf Henning Olsson (2019) Examining the Performance of the Modified ADF Goodness-of-fit Test Statistic in Structural Equation Models, Structural Equation Modeling: A Multidisciplinary Journal, 26:5, 778-789, DOI: 10.1080/10705511.2019.1586545

Examining the Performance of the Modified ADF Goodness-of-Fit Test Statistic in Structural Equation Models

Abstract

The asymptotically distribution-free (ADF) test statistic introduced in a landmark paper by Browne (1984) depends on very mild distributional assumptions and is theoretically superior to many other so-called robust tests available in structural equation modeling. The ADF test, however, often leads to model overrejection even at modest sample sizes. To overcome its poor small-sample performance, Chun, Browne, and Shapiro (2018) recently proposed a family of robust test statistics obtained by modifying the ADF statistic. This study investigates by simulation the performance of the new modified test statistics. The results revealed that although a few of the test statistics adequately controlled Type I error rates in each of the examined conditions, most performed quite poorly. This result underscores the importance of choosing a modified test statistic that performs well for specific examined conditions. A parametric bootstrap method is proposed for identifying such a best-performing modified test statistic. Through further simulation it is shown that the proposed bootstrap approach performs well.

Examining the Performance of the Modified ADF Goodness-of-Fit Test Statistic in Structural Equation Models

Introduction

Structural equation modeling (SEM) has for decades maintained an unprecedented level of popularity in social and behavioral science research. One reason for its pervasive use is that it offers researchers an opportunity to conduct detailed investigations of theoretical models. For example, in a theoretical model with observed and latent variables, specific relationships among the observed and the latent variables can be explicitly formulated and tested. With the help of popular software such as EQS (Bentler, 2008), Mplus (Muthén & Muthén, 2012), LISREL (Jöreskog & Sörbom, 2015) and lavaan (Rosseel, 2012), parameters in these models can then be readily estimated and assessed. Thus for any given model, a parameter vector θ, containing all the unknown parameters in the proposed model to be estimated can be stipulated, with the estimation based on a sample covariance matrix S assumed to converge to the population covariance matrix for large samples. To obtain the actual model parameter estimates, a fit function is then minimized, whereas to assess the appropriateness of a model the discrepancy between the model-implied and the sample covariance matrices is investigated.

As long as the sampled empirical data come from a population that has a multivariate normal distribution, model parameters and fit statistics may be efficiently estimated by a normal-theory based maximum likelihood (ML) procedure. The resulting test statistic, here denoted T_{ML}, equals the sample size times the minimum value of the discrepancy function (commonly denoted as $F_{\mathrm{ML}}(\theta)$), and will in the multivariate normal case approximate a central chi-square distribution as long as the model holds in the population. In practice, however, the assumption of multivariate normality seldom holds and this has led to the development of several robust statistics based on fourth-order sample moments. For example, Satorra and Bentler (1994) introduced two such robust fit statistics by modifying T_{ML}. Other robust approximations (e.g., Foldnes & Grønneberg, 2018; Wu & Lin, 2016) have also been presented in the literature. In addition, computer intensive methods have been proposed. For example, the approach proposed by Bollen and Stine (1992) utilizes a non-parametric bootstrap, whereas the approach introduced by Grønneberg and Foldnes (2018) combines the bootstrap and robust approximations. Despite the availability of these various approaches and the plethora of Monte Carlo studies examining their performance, no clear test statistic candidate among the bootstrap-based and the approximation-based methods has yet emerged as the best choice across model, sample size, and distributional conditions.

Preceding the above mentioned robust approximations and computer intensive methods, Browne (1984) proposed in a landmark paper the asymptotic distribution-free (ADF) statistic as a seemingly natural choice for situations with violations of the normality assumption. A key feature of this statistic, denoted as T_{ADF}, is that it is not based on approximating the distribution of T_{ML}, nor does it rely on the bootstrap. Instead, it is based on estimating the fourth-order moments of the underlying distribution and, as such, yields a test statistic which is asymptotically chi-square distributed under very general non-normal conditions. So the ADF statistic is theoretically superior to many of the above mentioned robust approximations later developed (for a more detailed overview concerning the widespread impact of the ADF test, see Cai (2012)).

Unfortunately, in practice T_{ADF} may not be a good choice for evaluating models when the observed sample is only moderately sized. To date, numerous simulation studies (e.g., Curran, West, & Finch, 1996; Hu, Bentler, & Kano, 1992) have demonstrated that unless the sample size is very large, T_{ADF} may even severely overreject correctly specified models. Indeed, in situations involving moderately sized sample, the above-mentioned bootstrap- and approximation-based test statistics will generally outperform T_{ADF} (e.g., Fouladi, 2000; Nevitt & Hancock, 2001).

As a remedy to the well-known poor small-sample performance of ADF, Chun et al. (2018) recently proposed some modifications to the original T_{ADF} test statistic.

Undoubtedly the potential significance of this recent contribution is immense, especially given the current lack of a convincingly best test statistic that can be used with typical sample sizes. As indicated by Chun et al. (2018), the proposed modification yields $d-1$ new modified test statistics, where d is the model degrees of freedom. That is, for each integer $m = 1, 2, \ldots, d-1$, there is a corresponding modified test statistic, denoted by $T_{\mathrm{M}}(m)$. As is the case with T_{ADF}, under quite general conditions each $T_{\mathrm{M}}(m)$ is asymptotically distribution-free in the sense that it converges in distribution to a chi-square distribution with m degrees of freedom.

In order to effectively utilize this newly proposed modified test statistic, a natural uncertainty for any applied researcher would be: Given a posited model and a set sample size, how should the number m of modified degrees of freedom be determined? Although Chun et al. (2018) did sketch a heuristic for determining m, they did not report details on the performance of the recommended heuristic. Instead, they evaluated the performance of the modified test statistic at fixed values of m. Across three models, Chun et al. (2018) tabulated the performance of $T_{\mathrm{M}}(m)$ only for limited ranges of m. For instance, Model 3 had $d = 189$ degrees of freedom, and hence m may range from $m = 1$ to $m = 188$. However, the authors present results only for $m = 77, \ldots, 85$ (Chun et al., 2018, Table 7).

In fact, to the best of our knowledge, no study has thus far systematically investigated how the small-sample performance of $T_{\mathrm{M}}(m)$ is affected by m. Additionally, no study has to date evaluated whether the proposed heuristic for determining m will in effect result in an acceptable procedure for testing model fit.

The purpose of this study is twofold: (i) to investigate how the choice of m actually affects the performance of $T_{\mathrm{M}}(m)$ under realistic sample size conditions, and (ii) to investigate how to expressly choose the value of m. To accomplish these goals, we empirically evaluate how the heuristic proposed by Chun et al. (2018) performs and subsequently propose and evaluate a new bootstrap-based approach for determining m.

The remainder of this article is organized as follows. We first review and illustrate the modified ADF test statistic. We then discuss heuristics for determining m, and present a new bootstrap-based method that can be applied for this task. Next we present three informative simulation studies. The first study evaluates the performance of $T_{\mathrm{M}}(m)$ as m varies. The second study evaluates the performance of several simple heuristics for determining m. The third simulation study evaluates the bootstrap approach for determining m. Finally, we present a general discussion, highlight limitations and remaining challenges, and end with some closing remarks.

The modified ADF test statistic

Let X be a random p-dimensional vector, with finite fourth order moments, and with population covariance matrix Σ. Let S be an unbiased estimator of Σ obtained from a sample of n independent observations. We denote the vector of all non-duplicated elements of S by $s = vech(S)$, which is a $p^* \times 1$ vector, where $p^* = \frac{1}{2}p(p+1)$. Similarly we define the population counterpart as $\sigma = vech(\Sigma)$. Let all the independent parameters in the model be contained in the q vector θ and let $\Sigma(\theta)$ denote the model implied covariance matrix. We assume that $\sigma(\theta) = vech(\Sigma(\theta))$ is differentiable and we denote its $p^* \times q$ Jacobian matrix by $\Delta(\theta) = \frac{\partial\sigma(\theta)}{\partial\theta}$. We declare that the model holds if there exists a parameter vector θ_0 so that $\Sigma(\theta_0) = \Sigma$. Asymptotically, $\sqrt{n}(s - \sigma)$ follows a multivariate normal distribution with zero mean vector and a covariance matrix Γ. By imposing mild assumptions on the employed estimator and the rank of Δ and Γ, Browne (1984) showed that the test statistic in Equation 1 will asymptotically follow a chi-square distribution with $p^* - q$ degrees of freedom:

$$T_{\mathrm{ADF}} = n(s - \hat{\sigma})'[\ \hat{\Gamma}^{-1} - \hat{\Gamma}^{-1}\hat{\Delta}(\hat{\Delta}'\hat{\Gamma}^{-1}\hat{\Delta})^{-1}\hat{\Delta}'\hat{\Gamma}^{-1}](s - \hat{\sigma}), \tag{1}$$

where $\hat{\Gamma}$ denotes a consistent estimate of Γ.

Despite the theoretical appeal of this general result, it is well known that the ADF test statistic in most cases has too slow of a convergence toward the chi-square distribution to be useful for making inferences that are based on small samples (e.g., Curran et al., 1996). The verified slow convergence rate has been attributed to the instability of estimating the fourth-order elements in Γ with small to moderate sample sizes. Related to this issue is also the tendency for Γ to be ill-conditioned (i.e., that the inverse of $\hat{\Gamma}$ in eq. (1) has high variability (Chun et al., 2018; Huang & Bentler, 2015). One measure of ill-conditioning is the condition number $\text{cond}(\Gamma) = \delta_{max}/\delta_{min}$, where δ_{max} and δ_{min} are the largest and smallest eigenvalues of Γ. Huang and Bentler (2015) postulated that this ill-conditioning aspect is the main reason behind the poor small-sample performance of the T_{ADF}.

To remedy the effect of ill-conditioning in Γ, Chun et al. (2018) proposed to transform Γ into a reduced dimensional matrix. This is achieved by replacing $\hat{\Gamma}$ by $\Upsilon'\hat{\Gamma}\Upsilon$ in eq. (1), where Υ is a $p^* \times r$ matrix with $q < r < p^*$ such that $\Upsilon'\hat{\Gamma}\Upsilon$ is non-singular. The modified ADF statistic now results from also substituting $\hat{\Delta}$ by $\Upsilon'\hat{\Delta}$ and $(s - \hat{\sigma})$ by $\Upsilon'(s - \hat{\sigma})$, which gives

$$\begin{aligned} T_{\text{M}}(m) = n(s-\hat{\sigma})'\Upsilon\Big[&(\Upsilon'\hat{\Gamma}\Upsilon)^{-1} \\ &- (\Upsilon'\hat{\Gamma}\Upsilon)^{-1}\Upsilon'\hat{\Delta}\left(\hat{\Delta}'\Upsilon(\Upsilon'\hat{\Gamma}\Upsilon)^{-1}\Upsilon'\hat{\Delta}\right)^{-1}\hat{\Delta}'\Upsilon(\Upsilon'\hat{\Gamma}\Upsilon)^{-1}\Big]\Upsilon'(s-\hat{\sigma}) \end{aligned} \quad (2)$$

Given that$\Upsilon'\Gamma\Upsilon$ is non-singular, $T_{\text{M}}(m)$ converges in distribution to a chi-square statistic with $m := r - q$ degrees of freedom, where $r = \text{rank}(\Upsilon'\Gamma\Upsilon)$ (Chun et al., 2018, Theorem 1). Note that in the present study we limit the discussion to a specific choice of Υ, namely the matrix whose jth column is the eigenvector of $\hat{\Gamma}$ that corresponds to the jth largest eigenvalue of $\hat{\Gamma}$ (Chun et al., 2018, p.55). Although it is possible that other choices of Υ may lead to different results than those reported in the present study, this topic is beyond the scope of the present article. Future studies may wish to examine whether this choice has any notable ramifications.

Illustration

We illustrate results obtained with the modified ADF test statistic using the political democracy model discussed in Bollen (1989), see Figure 1, where residual errors are not depicted for ease of presentation. As can be seen by examining the model, there are four indicators of political democracy measured twice (in 1960 and 1965), and three indicators of industrialization measured once (in 1960). The model has $q = 31$ free parameters, and $d = 35$ degrees of freedom. The sample consists of $n = 75$ countries. The model was estimated in lavaan (Rosseel, 2012) using normal-theory maximum likelihood estimation. There are 34 possible values for r, namely the integers ranging from 32 to 65. We calculated $T_{\text{M}}(m)$ for each corresponding degrees of freedom value m, running from 1 to 34. The p-values associated with $T_{\text{M}}(m)$, $m = 1, \ldots, 34$ were then computed and are plotted in Figure 2.

The most striking aspect of the results displayed in Figure 2 is the clear division of the modified test statistic p-values into two distinct clusters. Accordingly, all test statistics with a modified degrees of freedom $m < 14$ indicate a well-fitting model, whereas all test statistics with $m \geq 14$ indicate poor model fit. This clearly illustrates that the choice of the modified degrees of freedom m is indeed a crucial element in the proposed modified ADF procedure. In the present illustration there undeniably is a marked transition from $m = 13$ to $m = 14$, corresponding to the transition from $r = 44$ to $r = 45$. Having observed this result, it is then only natural to inquire whether this might somehow be reflected in the eigenvalues of $\hat{\Gamma}$. Examining these eigenvalues we detected no discernible shift between the 44th and 45th largest eigenvalues of $\hat{\Gamma}$. The condition number of $\hat{\Gamma}$ is large: $\text{cond}(\hat{\Gamma}) \simeq 1.92 \cdot 10^6$.

Methods for determining m

In order for $T_{\text{M}}(m)$ to improve the performance of T_{ADF} in small sample settings, Υ should be chosen so that instability in the estimation of Γ is reduced. To do so then requires that the recommendations of Chun et al. (2018) be followed to implement the modified statistic with the columns of Υ consisting of those eigenvectors of $\hat{\Gamma}$ that correspond to the $m + q$ largest eigenvalues of $\hat{\Gamma}$.

This of course now raises the issue of how to choose m. Chun et al. (2018) proposed a simple way to determine m: Given a proportion $0 < \beta < 1$, let k be the number of eigenvalues of $\hat{\Gamma}$ that are less than $\beta\delta_{max}$, where δ_{max} is the maximum eigenvalue of $\hat{\Gamma}$. Then m is determined as $m = \max(d - k, 1)$, and Υ will consist of the $m + q$ eigenvectors of Γ that correspond to the largest eigenvalues. Chun et al. (2018) proposed that β be in the range $0.0005 - 0.012$, but did not provide a statistical justification for recommending this range. Even within this range different β will generally determine different values of m. For instance, for the illustrative example in the preceding section $\delta_{max} = 1836.4$, and 23 and 49 of the eigenvalues of $\hat{\Gamma}$ are smaller than $0.0005 \cdot \delta_{max}$ and $0.012 \cdot \delta_{max}$, respectively. So using the lower bound of $\beta = 0.0005$ yields $m = \max(35 - 23, 1) = 12$, while the upper bound of $\beta = 0.012$ yields $m = \max(35 - 49, 1) = 1$. This means that $T_M(1), \ldots, T_M(12)$ are all admissible test statistics according to the heuristic range. Chun et al. (2018) do not provide recommendations on how to choose among the values of m prescribed by their heuristic. We will refer to these methods as *eigenvalue heuristics*, although how to choose the specific proportion level β for these heuristics is yet undetermined. The best choice of β may depend on both the sample and model characteristics.

We next propose a new method for determining m, based on the bootstrap selection in Grønneberg and Foldnes (2018). The method simulates B bootstrap samples from a multivariate normal distribution where the model fits perfectly, and calculates $T_{\mathrm{M}}(m)$ for $m = 1, \ldots d - 1$ in each bootstrap sample. Then m is determined by inspecting how close the rejection rate of $T_{\mathrm{M}}(m)$ approaches the 0.05 nominal level of significance. As outlined in Algorithm 1, this is done by minimizing the absolute value of the difference between the observed bootstrap rejection rates and 0.05.

Algorithm 1 Bootstrap selection of m

```
procedure SELECT(sample, model, B)
    Calculate the model-implied covariance matrix $\hat{\Sigma}$
    for $k \leftarrow 1, \ldots, B$ do
        boot.sample $\leftarrow$ A random sample drawn from $\mathrm{N}(0, \hat{\Sigma})$
        for $m \in 1, \ldots, d-1$ do
            Calculate $T_{\mathrm{M}}(m)$ from fitting the model to boot.sample
            Rejection$(m, k) \leftarrow 1$ if $T_{\mathrm{M}}(m)$ leads rejection of the model, 0 otherwise
        end for
    end for
    for $m \in 1, \ldots, d-1$ do
        RejectionRate$(m) \leftarrow \frac{\sum_{k=1}^{B} \mathrm{Rejection}(m,k)}{B}$
    end for
    return $\arg\min_{1 \le m \le d-1} |\mathrm{RejectionRate}(m) - 0.05|$
end procedure
```

Method

This section provides a detailed description of the proposed models examined in this simulation study, the analyzed sample sizes, and the distributional characteristics evaluated in the simulations in terms of data generation and program implementation. The selected conditions examined in this study were based on a detailed review of the literature on past simulation studies. Accordingly, a number of study features were selected to be fixed across conditions while others were varied. Features fixed in the simulations included the models examined, while those that were varied included the sample size and the underlying distributions.

Models

The simulation studies employed the same two confirmatory factor analytic models. We denote by $\mathcal{M}_1$ as a 3-factor model $x = \Lambda\xi + \delta$ where each factor ξ_1, ξ_2 and ξ_3 has five indicators, resulting in $d = 87$ degrees of freedom. Data generation for $\mathcal{M}_1$ was done using the following factor loadings and covariance matrix Φ of ξ:

$$\Lambda = \begin{bmatrix} 1.00 & & \\ 0.80 & & \\ 0.80 & & \\ 0.80 & & \\ 0.80 & & \\ & 1.00 & \\ & 0.50 & \\ & 0.50 & \\ & 0.50 & \\ & 0.50 & \\ & & 1.00 \\ & & 0.30 \\ & & 0.30 \\ & & 0.30 \\ & & 0.30 \end{bmatrix}, \qquad \Phi = \begin{bmatrix} 1 & & \\ 0.50 & 1 & \\ 0.50 & 0.50 & 1 \end{bmatrix},$$

while the residuals δ have unit variances. The second model, denoted by $\mathcal{M}_2$, is a five-factor model with 265 degrees of freedom, where each factor has five indicator variables:

$$\Lambda = \begin{bmatrix}
1.00 & & & & \\
0.80 & & & & \\
0.80 & & & & \\
0.80 & & & & \\
0.80 & & & & \\
 & 1.00 & & & \\
 & 0.50 & & & \\
 & 0.50 & & & \\
 & 0.50 & & & \\
 & 0.50 & & & \\
 & & 1.00 & & \\
 & & 0.30 & & \\
 & & 0.30 & & \\
 & & 0.30 & & \\
 & & 0.30 & & \\
 & & & 1.00 & \\
 & & & 0.30 & \\
 & & & 0.30 & \\
 & & & 0.30 & \\
 & & & 0.30 & \\
 & & & & 1.00 \\
 & & & & 0.30 \\
 & & & & 0.30 \\
 & & & & 0.30 \\
 & & & & 0.30
\end{bmatrix}, \qquad \Phi = \begin{bmatrix}
1 & & & & \\
0.50 & 1 & & & \\
0.50 & 0.50 & 1 & & \\
0.10 & 0.10 & 0.10 & 1 & \\
0.10 & 0.10 & 0.10 & .50 & 1
\end{bmatrix},$$

and the residuals have unit variance.

Sample sizes

Three different sample sizes were selected to reflect small, medium, and large sample sizes. The selected sample sizes for $\mathcal{M}_1$ were $n = 150, n = 300$ and $n = 1000$, while for the larger model $\mathcal{M}_2$ samples of size $n = 350, n = 700$ and $n = 1500$ were used.

Distributions

Data from two distributional conditions were employed in this simulation study, normal and non-normal data. Given that the ADF approach is theoretically and empirically less sensitive to the underlying distribution than other robust statistics (Hoogland & Boomsma, 1998, p. 263), only two distributions were considered in this study. Non-normal data were generated by the Vale-Maurelli (VM) transform (Vale & Maurelli, 1983), so that each marginal distribution had skewness 2 and excess kurtosis 10. To investigate the claim (Chun et al., 2018; Huang & Bentler, 2015) that the condition number of Γ affects the finite-sample performance of T_{ADF}, we calculated Γ under both multivariate normal and non-normal distributions (Foldnes & Grønneberg, 2017). For model $\mathcal{M}_1$ the condition number under normal and VM distributions were 38.3 and 111.9, respectively. For model $\mathcal{M}_2$ the condition numbers were 39.0 and 120.8, for the multivariate normal and VM distribution, respectively. Hence, if the condition number in fact predicts performance, then we would expect the performance of T_{ADF} to be markedly worse in the VM distributional condition compared to the multivariate normal condition.

Another way to compare the two distributions is to calculate the asymptotic standard errors of the model estimates. Model $\mathcal{M}_1$ and $\mathcal{M}_2$ have 18 and 30 free parameters, respectively. We entered the calculated Γ into a sandwich-type formula (Browne, 1984, eq. (2.12a)) and obtained the asymptotic covariance matrix of $\sqrt{n}\hat{\theta}$. For simplicity, we only consider the diagonal of this matrix, that is, we focus on the variances of the 18 estimates for $\mathcal{M}_1$ and the 30 estimates for $\mathcal{M}_2$, see Figure 3. It is evident that the parameter estimates vary much more under the VM distribution than under the normal distribution, as many points lie far above the $x = y$ line. This reflects the larger variability of $\hat{\Gamma}$ under the VM distribution compared to the multivariate normal distribution.

Data generation

All data generation was conducted in the R computing environment. Model estimation using normal-theory based maximum likelihood was computed using the package lavaan (Rosseel, 2012). Finally, the modified ADF statistics were computed using auxiliary functions from lavaan, see the Appendix for R code. Interested readers can access this material in order to inspect, run, and modify our code. Because simulation study 3 was much more computationally intensive than study 1 and 2 (as for each simulated dataset in each condition a bootstrap procedure needed to be employed), the simulations were all performed on the Abel computer cluster, owned by the University of Oslo and Uninett/Sigma2.

The three studies

All three studies assumed a correctly specified model. Given the sparse literature on modified ADF statistics, the present article is restricted to the primary concern of adequate Type I error control. The outcome variable in all three studies was the rejection rate of the test statistic calculated at the conventional $\alpha = 0.05$ level.

The first study investigated the empirical Type I error control of $T_{\mathrm{M}}(m)$ across the range of $m = 1, 2, \ldots, d-1$, as well as of T_{ADF}. In each condition, 2000 samples were generated and the empirical rejection rates was calculated as the proportion of $T_{\mathrm{M}}(m)$ values that exceeded the 0.95 quantile of the chi-square distribution with m degrees of freedom.

The second study investigated the performance of the eigenvalue heuristic for determining the modified degrees of freedom number m. It operates by determining r as the number of eigenvalues of $\hat{\Gamma}$ that is larger than some given proportion β of the maximum eigenvalue δ_{max}. This heuristic was proposed, but not evaluated, by Chun et al. (2018), who suggested that β take some value between 0.0005 and 0.012. Accordingly, in the present study we included $\beta = 0.0005, 0.001, 0.005, 0.012$. In the case that $r < q$, we set r equal to $q + 1$, so that $T_M(1)$ is picked for model evaluation. Given the insensitivity of the modified test statistic to underlying distribution (as will be demonstrated in the reported findings from study 1), we limited ourselves to the non-normal data condition for study 2.

The third study evaluated the performance of the proposed bootstrap approach for determining m. For each original simulated sample, we simulated $B = 1000$ bootstrap samples based on the model-implied covariance matrix. In each bootstrap sample we calculated $T_{\mathrm{M}}(m)$ for $m = 1, \ldots, d - 1$. Finally, m was determined so that the rejection rate was closest to 0.05. The p-value was then calculated with respect to $T_{\mathrm{M}}(m)$ based on the original simulated sample. This was replicated 1000 times to yield the rejection rate associated with the bootstrap procedure. See Algorithm 1 for further details.

Results

Study 1

The rejection rates of $T_{\mathrm{M}}(m)$ for $m = 1, 2, \ldots, 86$ for the examined model $\mathcal{M}_1$ are plotted in Figure 4. The rejection rate of T_{ADF} for $m = 87$ is also plotted in the same figure. We observe that the rejection rate of T_{ADF} expectedly approaches the nominal 0.05 as sample size increases, although very slowly. To supplement Figures 4 and 5 we give in Table 1 for each condition the value of m whose associated rejection rate comes the closest to the nominal 5% level. Also listed are the range of m whose associated rejection rates are larger than 0.025 and smaller than 0.075, which we deem acceptable (Bradley, 1978). These results demonstrate that m is much more affected by sample size than by the underlying distribution of the data. Rejection rates increase almost monotonically with increasing degrees of freedom. Values of m below 50 lead to exceptionally poor Type I error control for all distributions and sample sizes. For the largest values of m, $T_{\mathrm{M}}(m)$

performs somewhat similarly to T_{ADF} in the sense that it severely overrejects the correctly specified $\mathcal{M}_1$. Indeed, the "window" of acceptable m values in Table 1 are quite narrow in each condition in that there are only a few m values in each condition that yield adequate Type I error control.

Figure 5 presents plots of the empirical rejection rates for model $\mathcal{M}_2$, for $T_{\mathrm{M}}(m)$ for $m = 1, 2, \ldots, 264$, as well as for T_{ADF} at $m = 265$. We observe a similar pattern of results with model $\mathcal{M}_2$. Specifically, increasing rejection rates with m and a narrow interval of m values where $T_{\mathrm{M}}(m)$ adequately controls Type I error rates. The optimal values of m are found in Table 1. As can be seen from these findings, it is again evident that sample size has a strong effect on the optimal m value, while the underlying distribution has a much smaller impact on the optimal m value. The window of acceptable m values is quite narrow also for $\mathcal{M}_2$. For instance, under normality at $n = 700$ there are only seven acceptable values across the whole range of possible m values ($m = 1$ to $m = 264$). It is also important to note that in both examined models and at all sample sizes, T_{ADF} is found to be unaffected by the underlying distribution.

Study 2

Figure 6 displays the rejection rates of the eigenvalue heuristics for model $\mathcal{M}_1$ across three sample sizes, in the upper three panels. In the lower three panels is seen the eigenvalue heuristic rejection rates for model $\mathcal{M}_2$ across three sample sizes. It appears that the appropriate 5% rejection rate is attained within the range proposed by Chun et al. (2018). However, none of the four values within this range evaluated in this study come close to attaining adequate Type I error control. The rejection rates are zero for $\beta = 0.005$ and $\beta = 0.012$, while for $\beta = 0.0005$ and $\beta = 0.001$ the rejection rates are generally too high. The optimal value for β for the conditions in Study 2 seems to lie between 0.001 and 0.005, but currently no method exists for determining this optimal value. We remark that the results are in accordance with the results from Study 1, where it was found that high values of m were associated with high rejection rates, while low values of m corresponded to close-to-zero rejection rates. The lower β is, the larger value of m will be chosen by the heuristic, and consequently the larger the rejection rate will be. Similarly, increasing β entails that the m chosen by the heuristic decreases, resulting in lower rejection rates.

Study 3

Table 2 presents the rejection rates obtained by using the parametric bootstrap approach to determine m. Examining Table 2, it is evident that the rejection rates are unsurprisingly better with an underlying normal distribution compared to a non-normal distribution. These results were expected since the multivariate normal distribution was used to generate the bootstrap samples in Algorithm 1. Nevertheless, it is also apparent that even under non-normality the bootstrap demonstrates acceptable Type I error control, except at the smallest sample sizes under the large model, where the bootstrap tends to overreject. However the bootstrap rejection rates (11.1% and 8.2%) in these conditions are much closer to the nominal level than any of the eigenvalue heuristic rejection rates depicted in the two leftmost panels in the lower row of Figure 6.

Without a doubt, the bootstrap clearly outperforms the four eigenvalue heuristics investigated in study 2. If we deem a rejection rate acceptable if it falls in the $0.025 - 0.075$ interval, the bootstrap is acceptable in 10 of the 12 conditions in Table 2. Compare this to the four eigenvalue heuristics, whose rejection rates under non-normality are depicted in Figure 6. Three of the four heuristics have unacceptable rejection rates in every condition, while one heuristic has unacceptable rejection rates in five of six conditions.

The last column contains the mean value of m across the 1000 replications used for study 3. Using the optimal m values from Table 1 as benchmarks, for both models and all sample sizes the bootstrap moderately overestimates m under normality and underestimates m under non-normality.

Discussion

The first simulation study systematically investigated the performance of each modified test statistic across the full range of permissible modified degrees of freedom. This differs from Chun et al. (2018), which reported results for only a limited range of m.

The second simulation study evaluated the performance of eigenvalue heuristics proposed by Chun et al. (2018) for the determination of m, and the third study evaluated a new bootstrap based procedure for determining m.

The findings from study 1 showed that the finite-sample performance of the modified test statistics was highly dependent on m, the number of modified degrees of freedom. For both models, all sample sizes and distributional conditions, the same pattern of results was observed: For low values of m the modified test statistics exhibited poor performance, almost always failing to reject the model. Then, as m increased, there came a small interval of m values where the modified test statistics exhibit acceptable Type I error control, rejecting the model in about 5% of the simulated samples. Then, as m increased beyond this narrow interval, the modified test statistics resembled the original ADF test in severely overrejecting the model. We remark that this observed pattern of rejection rates increasing monotonously with increased modified degrees of freedom was not observed in the simulations reported by (Chun et al., 2018, Table 3), where the rejection rates dropped from above 5% to below 5%, before increasing again to above 5%, with increasing m. In accordance with the findings of Chun et al. (2018), results from study 1 revealed that only a small number of the modified test statistics were able to control Type I error much better than the original ADF test. Also, the m associated with the optimal modified test statistic was shown to depend upon sample size. However, although the condition number of Γ was much larger under the non-normal distribution than under multivariate normality, the optimal m value did not change much between the two distributional conditions. Additionally, the asymptotic variances of the estimated parameters were larger under non-normality, compared to normality. Therefore, it was surprising that the effect of non-normality on ADF and the modified test statistics was only modest. These findings do not support the acknowledged claim (Chun et al., 2018; Huang & Bentler, 2015) that the condition number of Γ might explain the poor performance of the ADF test. We found that ADF performed just as poorly under multivariate normality as it did under the non-normal distribution, despite a much larger condition number of Γ under the latter distribution.

Study 1 revealed that the values of m that are associated with acceptable performance of the modified test statistics were few. It was also determined that the optimal degrees of freedom m varied with sample size. The larger the sample size, the larger the value of m that is needed to achieve adequate Type I error control. These results clearly indicated the vital importance of investigating whether there are methods that may reliably identify the correct value of m, which in turn would result in acceptable Type I error control.

Study 2 was purposely designed to evaluate methods based on eigenvalues of $\hat{\Gamma}$, as proposed by Chun et al. (2018). However, none of these eigenvalue heuristics performed satisfactorily. The performance of the eigenvalue heuristic was very sensitive to the chosen cut-off value β. For instance, in most conditions, setting $\beta = 0.001$ in the heuristic produced far too high rejection rates, while setting $\beta = 0.005$ resulted in far too low rejection rates (see Figure 6). It therefore seems difficult to find a method that estimates β from the specific data and model at hand so that adequate type I error control are maintained. This might be seen in relation to the above-mentioned lack of association between the condition number of Γ and the performance of the modified test statistics. We could find no support for an association between the eigenvalues of Γ or $\hat{\Gamma}$ and the performance of modified test statistics.

In the third study, we evaluated a parametric bootstrap procedure to determine m, which was found to outperform the eigenvalue heuristics. In most conditions the bootstrap was able to determine a m value close to the optimal value, and thus resulted in acceptable Type I error control. Only in the large model, for the smallest sample sizes and under non-normality conditions did the rejection rates wander too far from the 5% level to be deemed adequate. A weakness of the proposed parametric bootstrap is that the bootstrap samples are generated based on multivariate normality. This is in contrast to the conditions that the original ADF test was designed to handle, namely moderate to severe non-normality. However, as observed in study 1, the performance of ADF is not very sensitive to the underlying distribution. This may help explain why the parametric bootstrap might offer acceptable performance even under non-normality. While preparing to conduct this study we also experimented with the non-parametric bootstrap for determining m, but we found that it consistently underestimated m, resulting in too low rejection rates. For this reason, we abandoned any further work on its application. Of course, an added limitation with the bootstrap procedure, common to all such procedures, is the amount of time needed to identify m. In each bootstrap sample, many modified test statistics must be computed, so that the bootstrap procedure needs several minutes in total running time. However, the bootstrap may be implemented with parallel computing on a modern multiple-core computer, thereby reducing considerably the required running time.

As with any simulation study, obtained results are strictly speaking only valid for the conditions investigated and one must be cautious in overgeneralizing these findings. In this study, only two factor models and two distributional conditions were examined. Although we found that the modified test statistics depended little on the underlying distribution, it is possible that more extreme conditions of non-normality than those considered in this study and extra complex models may influence the performance of the modified test statistics in a more pronounced manner.

Conclusion

We have investigated by Monte Carlo simulation the performance of newly proposed test statistics based on modifying the ADF test of Browne (1984). These tests exhibited rather large variability in performance, as a function of the degrees of freedom chosen. Only a small range of degrees of freedom resulted in acceptable performance for the modified test statistics. This range was found to depend on sample size, and to a lesser extent on underlying normality. Earlier proposed heuristics for determining the degrees of freedom were also found to perform poorly. In contrast, the proposed bootstrap procedure was better able to determine the optimal value for the degrees of freedom to use in the modified test statistic, except under conditions of small sample size and non-normal data.

References

Bentler, P. (2008). *Eqs 6 structural equations program manual.* Encino, CA: Multivariate Software.

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*, 205–229.

Bradley, J. V. (1978). Robustness? *British Journal of Mathematical and Statistical Psychology*, *31*(2), 144–152.

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

Cai, L. (2012). Three cheers for the asymptotically distribution free theory of estimation and inference: Some recent applications in linear and nonlinear latent variable modeling. In M. C. Edwards & R. C. MacCallum (Eds.), *Current topics in the theory and application of latent variable models* (pp. 119–131). Routledge.

Chun, S. Y., Browne, M. W., & Shapiro, A. (2018). Modified distribution-free goodness-of-fit test statistic. *Psychometrika*, *83*, 48–66.

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

Foldnes, N., & Grønneberg, S. (2017). The asymptotic covariance matrix and its use in simulation studies. *Structural Equation Modeling*, *24*, 881–896.

Foldnes, N., & Grønneberg, S. (2018). Approximating test statistics using eigenvalue block averaging. *Structural Equation Modeling*, *25*, 101–114.

Fouladi, R. T. (2000). Performance of modified test statistics in covariance and correlation structure analysis under conditions of multivariate nonnormality. *Structural Equation*

Modeling, *7*, 356–410.

Grønneberg, S., & Foldnes, N. (2018). Testing model fit by bootstrap selection. *Structural Equation Modeling*, 1–9. doi: 10.1080/10705511.2018.1503543

Hoogland, J. J., & Boomsma, A. (1998). Robustness studies in covariance structure modeling: An overview and a meta-analysis. *Sociological Methods & Research*, *26*, 329–367.

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

Huang, Y., & Bentler, P. M. (2015). Behavior of asymptotically distribution free test statistics in covariance versus correlation structure analysis. *Structural Equation Modeling*, 489–503.

Jöreskog, K., & Sörbom, D. (2015). Lisrel 9.20 for windows [computer software]. *Skokie, IL: Scientific Software International.*

Muthén, B., & Muthén, L. (2012). Mplus version 7: User's guide. *Los Angeles, CA: Muthén & Muthén.*

Nevitt, J., & Hancock, G. R. (2001). Performance of bootstrapping approaches to model test statistics and parameter standard error estimation in structural equation modeling. *Structural equation modeling*, *8*, 353–377.

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

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.

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

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*, *23*, 409–421.

Model	Distribution	Sample size	Optimal m	Acceptable m
$\mathcal{M}_1$	Normal	150	52	50-53
		300	66	64-68
		1000	80	76-81
	Non-normal	150	56	54-58
		300	69	66-70
		1000	80	77-82
$\mathcal{M}_2$	Normal	350	153	150-156
		700	190	185-193
		1500	221	214-224
	Non-normal	350	163	158-165
		700	196	190-199
		1500	222	214-227

Table 1

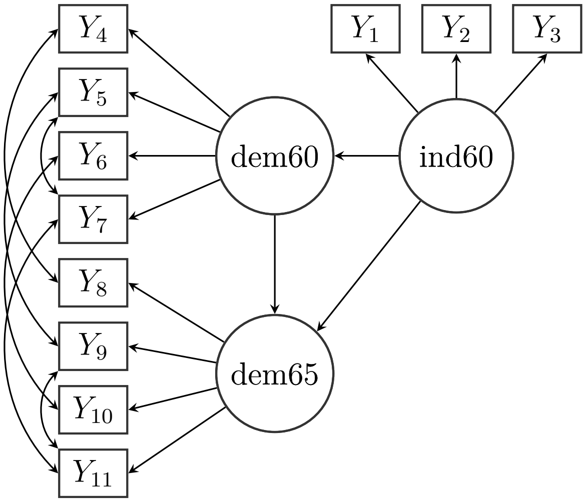
Study 1: Optimal value of m in terms of Type I error control, together with range of m whose rejection rates are in $(0.025, 0.075)$.

Model	Distribution	Sample size	Rejection rate	Mean of m
$\mathcal{M}_1$	Normal	150	0.052	55.7
		300	0.041	67.8
		1000	0.057	79.9
	Non-normal	150	0.033	54.2
		300	0.033	66.5
		1000	0.049	79.2
$\mathcal{M}_2$	Normal	350	0.038	161.6
		700	0.049	195.5
		1500	0.053	222.6
	Non-normal	350	0.111	159.5
		700	0.082	193.5
		1500	0.049	220.9

Table 2

Study 3: Rejection rates obtained when determining m by the bootstrap procedure in Algorithm 1.

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



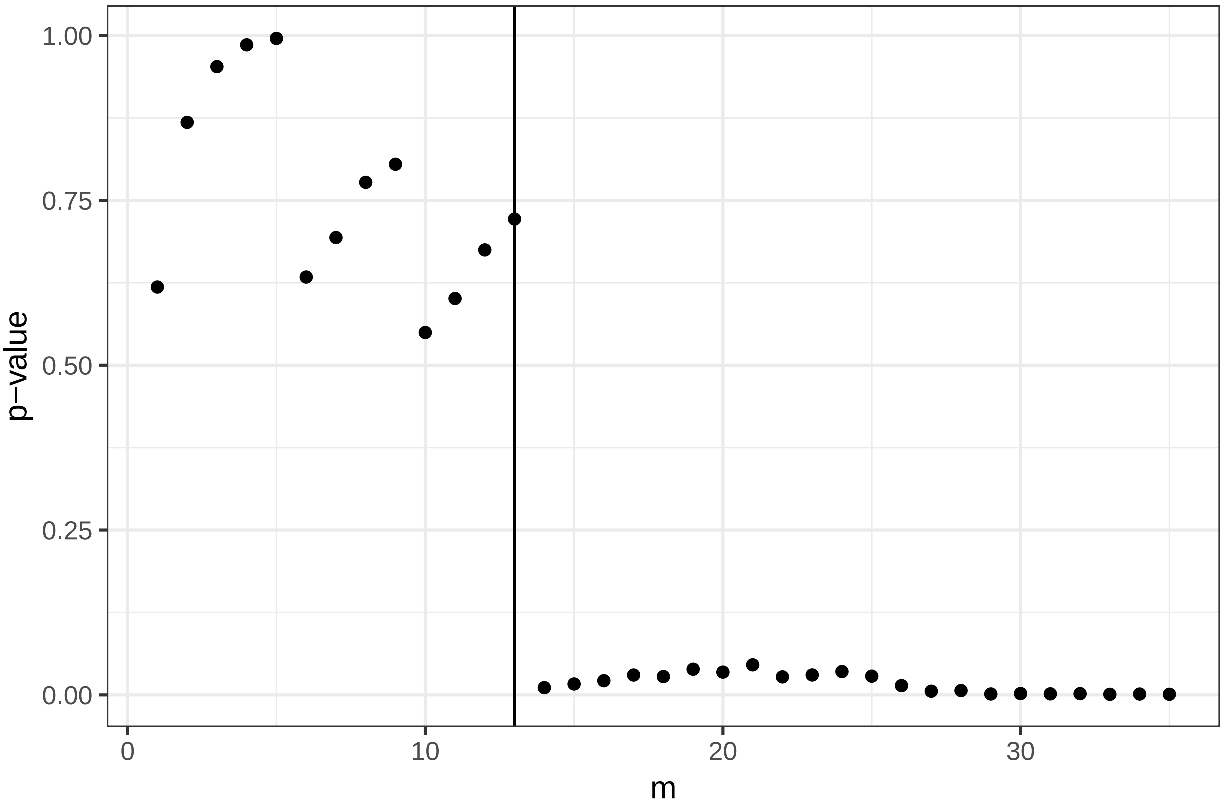


Figure 2. p-values associated with the modified ADF test statistic. Vertical line at $m = 13$.

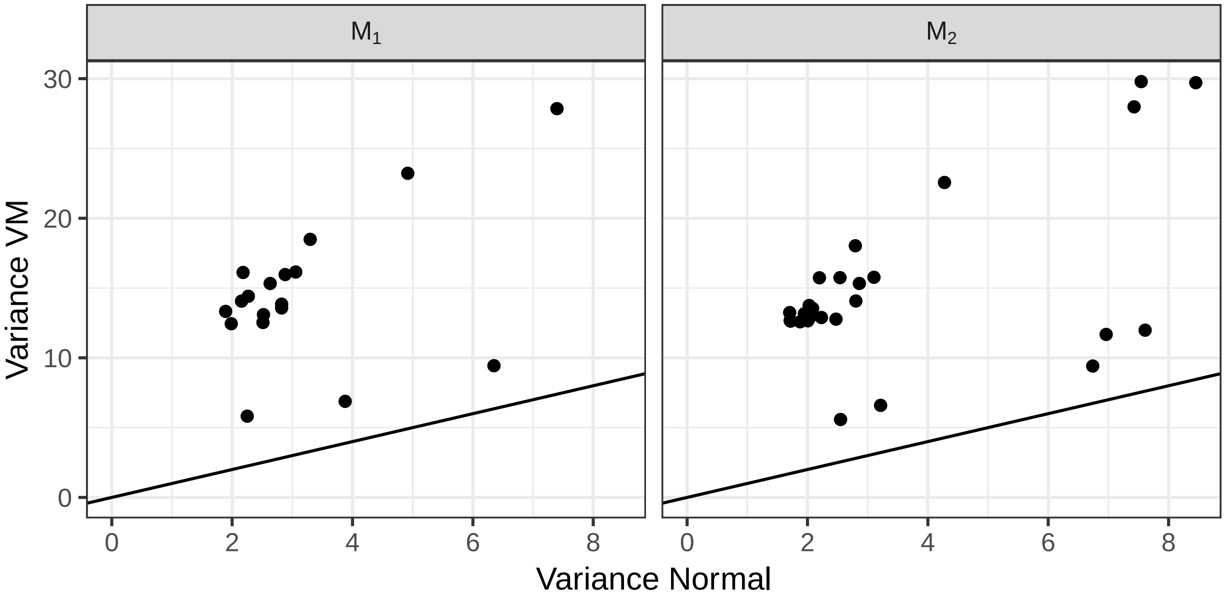


Figure 3. Scatter plot of the asymptotic variances of 18 estimates for $\mathcal{M}_1$, left panel, and 30 estimates for $\mathcal{M}_2$, right panel. y-axis refers to the variances under the VM distribution, x-axis to the variances under the multivariate normald distribution. The line $y = x$ is also included.

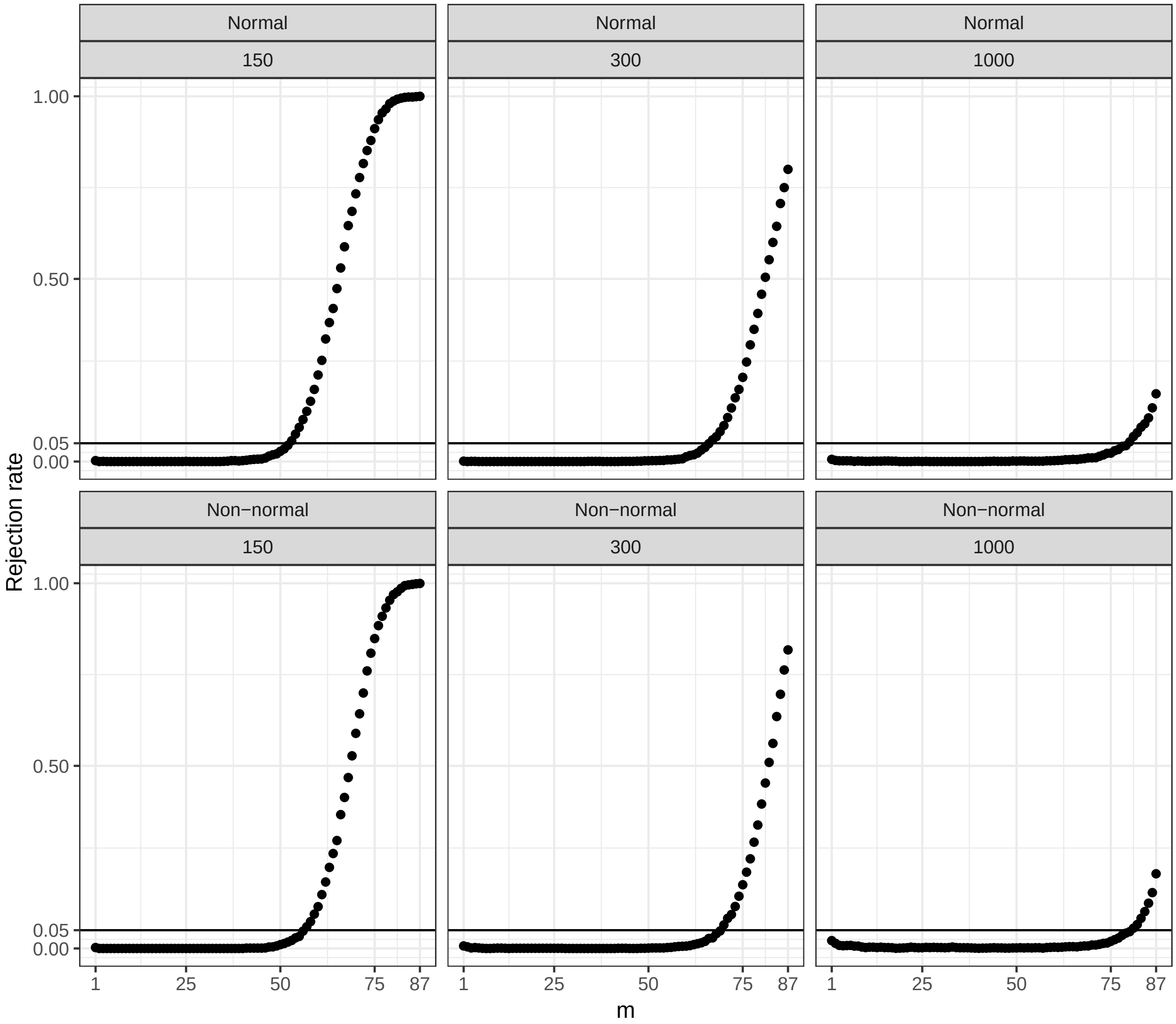


Figure 4. Study 1: Model $\mathcal{M}_1$, rejection rates for $T_{\mathrm{M}}(m)$, $m = 1, \ldots, 86$. At $m = 87$ the rejection rate of T_{ADF} is also plotted. Each panel shows a combination of distribution and sample size, with a horizontal line at $\alpha = 0.05$.

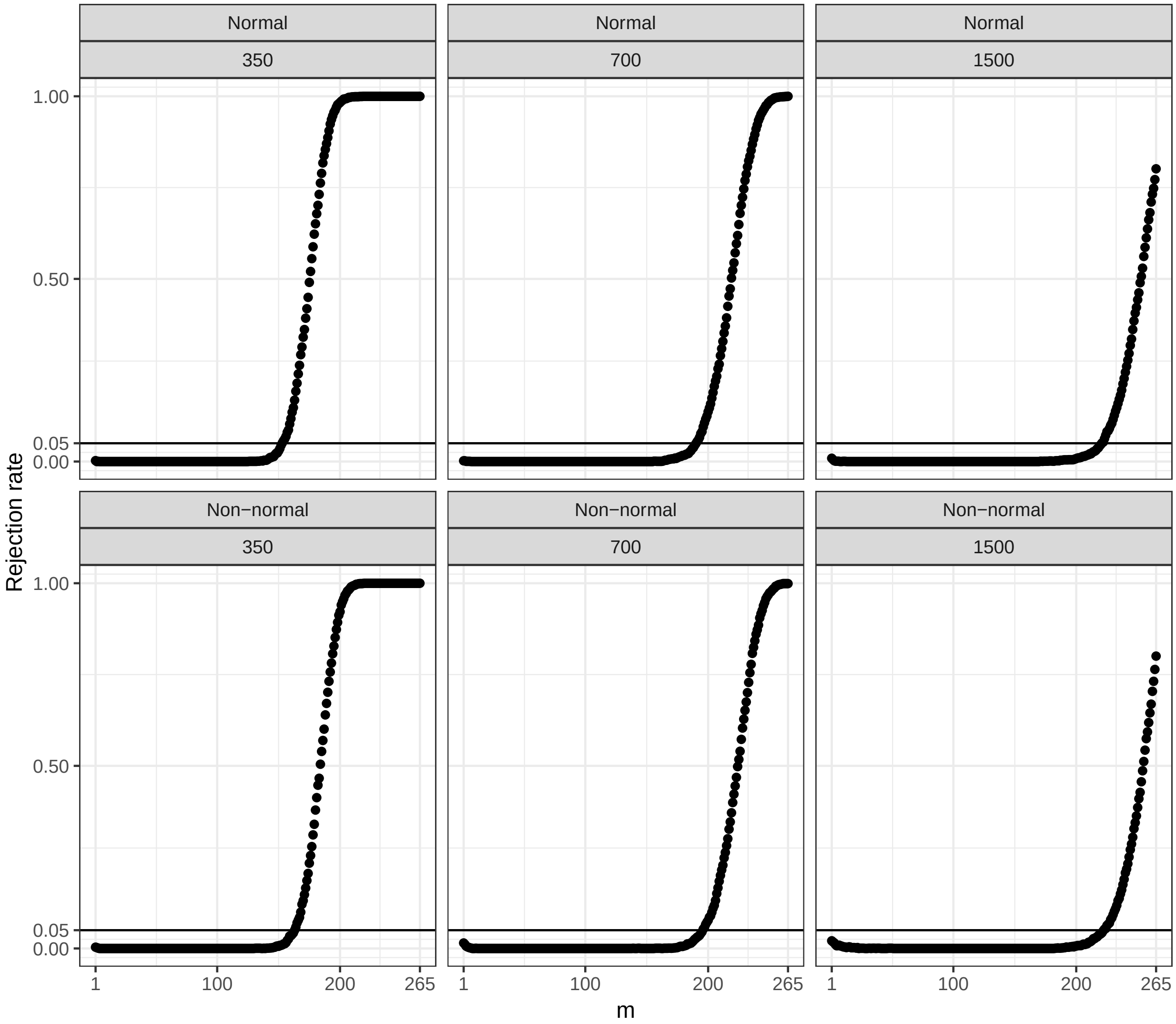


Figure 5. Study 1: Model $\mathcal{M}_2$, rejection rates for $T_{\mathrm{M}}(m)$, $m = 1, \ldots, 264$. At $m = 265$ the rejection rate of T_{ADF} is also plotted. Each panel shows a combination of distribution and sample size, with a horizontal line at $\alpha = 0.05$.

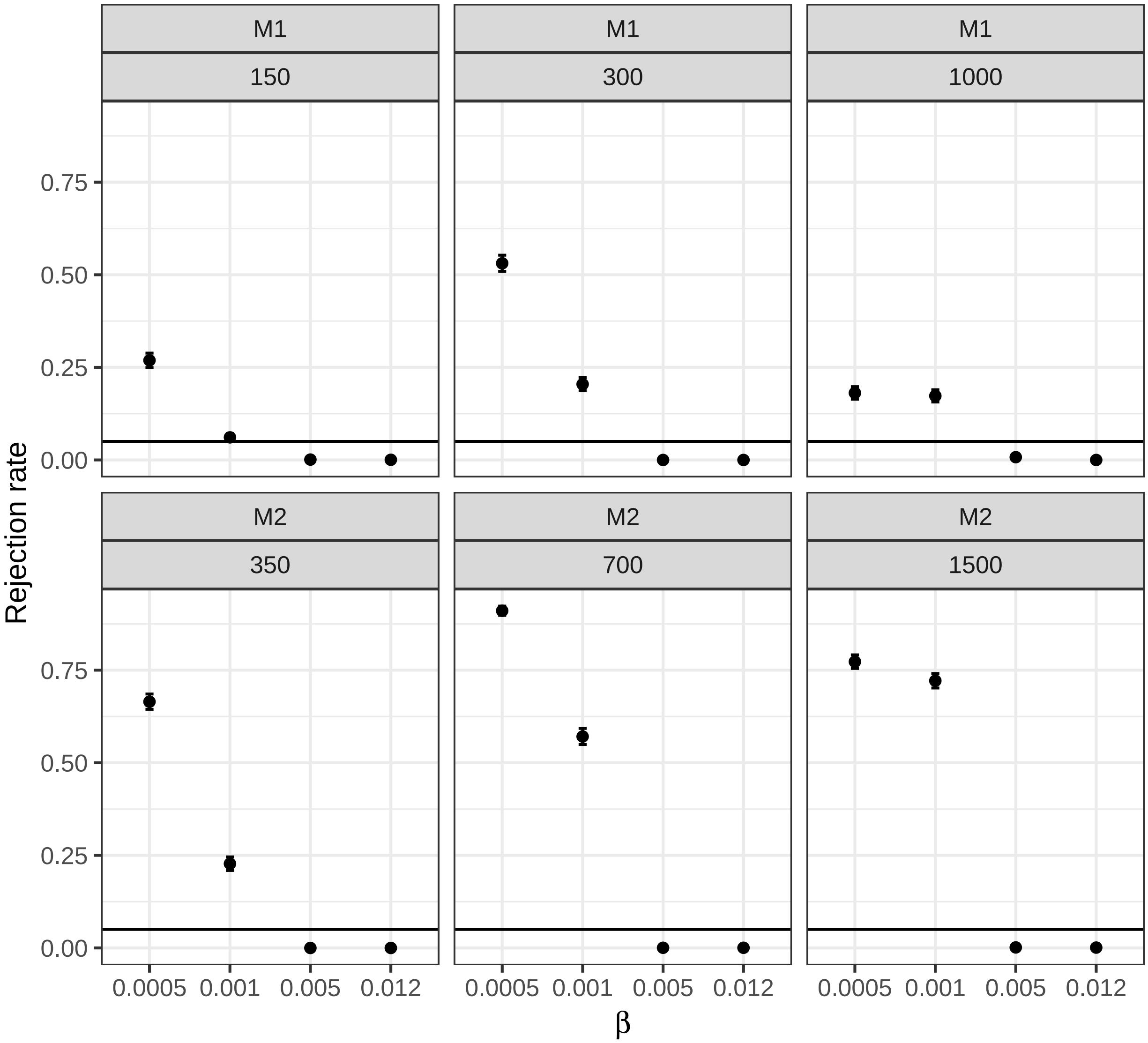


Figure 6. Study 2: Type I error rates for eigenvalue heuristic, applied to models $\mathcal{M}_1$ (upper panel) and $\mathcal{M}_2$ (lower panel). The horizontal line refers to the nominal $\alpha = 0.05$ Type I error rate.

Appendix

R code for Illustration

```
library(lavaan)

bollen.model <-

  "# measurement model

ind60 =~ x1 + x2 + x3

dem60 =~ y1 + y2 + y3 + y4

dem65 =~ y5 + y6 + y7 + y8

# regressions

dem60 ~ start(0.8)*ind60

dem65 ~ start(0.2)*ind60 + start(0.5)*dem60

# residual correlations

y1 ~~ start(0.2)*y5

y2 ~~ start(0.2)*y4 + start(0.2)*y6

y3 ~~ start(0.2)*y7

y4 ~~ start(0.2)*y8

y6 ~~ start(0.2)*y8"


fit= sem(bollen.model, data=PoliticalDemocracy)

Gamma <- lavTech(fit, "gamma")[[1]]

Delta <- lavaan:::computeDelta(lavmodel = fit@Model)[[1]]

resid= matrix(lav_matrix_vech(residuals(fit)$cov), ncol=1)

eigenvectors = eigen(Gamma)$vectors


mod.df = 10 # for example

r = ncol(Delta)+mod.df

ypsilon <- eigenvectors[, 1:r]
```

```
A = t(ypsilon)%*%Gamma%*%ypsilon

A=solve(A)

B=solve(t(Delta)%*%ypsilon%*%A%*%t(ypsilon)%*%Delta)


res=ypsilon %*% (A-A%*%t(ypsilon)%*%Delta%*%B%*%t(Delta)%*%ypsilon%*%A)%*% t(ypsilon)

n=nrow(PoliticalDemocracy)

chisquare = n * t(resid)%*%res%*%resid

pval = 1 - pchisq(chisquare, df = mod.df)
```