UNIVERSITY OF CALIFORNIA

Los Angeles

Using Monte Carlo Normal Distributions
to Evaluate Structural Models with Nonnormal Data

A dissertation submitted in partial satisfaction

of the requirements for the degree

Doctor of Philosophy in Statistics

by

Siavash Jalal

2017

ABSTRACT OF THE DISSERTATION

Using Monte Carlo Normal Distributions
to Evaluate Structural Models with Nonnormal Data

by

Siavash Jalal
Doctor of Philosophy in Statistics
University of California, Los Angeles, 2017
Professor Peter M Bentler, Chair

One of the main problems of statistical inference in Structural Equation Modeling (SEM) is the overall goodness of fit test. Many statistical theories have been developed based on asymptotic distributions of test statistics. When the model includes a large number of variables or the population is not from the multivariate normal distribution, the rates of convergence of these asymptotic distributions are very slow, and thus in these situations the asymptotic distributions do not approximate the distribution of the test statistics very well. Modifications to theoretical models and also bootstrap methods have been developed by researchers to improve the accuracy of hypothesis testing, mainly accuracy of Type I error, but when the sample size is small or the number of variables is large those methods have their limitations. Here we propose a Monte Carlo test that is able to control Type I error with more accuracy and it overcomes some of the limitations in the bootstrapping and theoretical approaches. Our simulation study shows that the suggested Monte Carlo test has more accurate observed significance level, as compared to other tests. Problems that occur in the bootstrapping are highlighted and it is shown that the new Monte Carlo test can overcome those problems. A power analysis shows that the new test has a reasonable power.

The dissertation of Siavash Jalal is approved.

Craig Kyle Enders

Yingnian Wu

Arash Ali Amini

Peter M Bentler, Committee Chair

University of California, Los Angeles

2017

To my beloved parents who have raised me to be who I am.

To my wonderful wife Bahareh for her love, support, and encouragement.

To my beautiful daughters, Paneez and Parmida.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

First and foremost, I would like to express my gratitude to my advisor, Professor Peter Bentler, who has been instrumental in the completion of my PhD. He has truly been an exceptional advisor and patiently has provided me with countless hours of revisions and advice on my thesis and I cannot thank him enough for that. Additionally, I would like to thank the members of my committee, Dr. Craig Enders, Dr. Yingnian Wu, and Dr. Arash Amini, who provided me with important feedback and suggestions. I would also like to acknowledge the amazing faculty and staff of the UCLA Statistics Department.

This work used computational and storage services associated with the Hoffman2 Shared Cluster provided by UCLA Institute for Digital Research and Education's Research Technology Group. I would like to thank Raffaella D'Auria from UCLA Hoffman2 IDRE for technical support.

I would also like to acknowledge Dr. Mortaza Jamishidian from California State University, Fullerton, a wonderful teacher, mentor, and friend who has always believed in me and generously has given me his endless support and guidance.

Last but not the least, I would like to thank my family; my father for inspiring me to further my education, my mother for her unconditional love and support, my wife Bahareh who has been extremely supportive of me throughout this entire process and has made countless sacrifices to help me get to this point.

VITA

1995	B.S. (Applied Mathematics), Isfahan University of Technology, Isfahan.
1998	M.S. (Industrial Engineering), University of Tehran, Tehran.
2009	M.A. (Applied Mathematics), CSUF, Fullerton, California.
2016	C.Phil. (Statistics), UCLA, Los Angeles, California.

PUBLICATIONS

Jalal, S., & Azadeh, M.A., “Identifying the economic importance of industrial sectors by multivariate analysis”, *Journal of the Faculty of Engineering (University of Tehran, Iran),* 35(5), pp. 437 – 449, 2001.

Jamshidian, M., & Jalal, S. “Tests of Homoscedasticity, Normality and Missing Completely at Random for Incomplete Multivariate Data”, *Psychometrika,* 75(4), pp. 649 – 674, 2010.

Jamshidian, M. Jalal, S., & Jansen, C., “MissMech: An R Package for Testing Homoscedasticity, Multivariate Normality and Missing Completely at Random (MCAR)”, *Journal of Statistical Software,* 56(6), 2014.

INTRODUCTION

Structural Equation Modeling (SEM) has been used in the analysis of multivariate data with latent variables. Since the measurement of latent variables often arises in social and behavioral studies, SEM is very popular in those areas of research and in those studies, it is very common to have a situation that data is from a nonnormal population. A goodness of fit test in statistics is used to describe how well a statistical model fits the observed data. Finding an accurate evaluation of goodness of fit is one of the most challenging problems in SEM specially when the population is not multivariate normal and when access to a large sample size is not possible.

To evaluate the goodness of fit of a model, often the discrepancy between observations and their expected values under the model is measured. In SEM, an overall goodness of fit test describes whether a hypothesized structured model is appropriate to represent the relationships among observed and latent variables. More precisely, a goodness of fit test statistic is defined as a functional of the discrepancy between the sample covariance matrix and the covariance matrix estimated under the model.

Statistical theories for the goodness of fit test can be divided into two categories, when observed variables follow a multivariate normal distribution, and when they are from an unknown distribution. In both cases the appropriate test statistics mostly can be approximated accurately by a $\boldsymbol{\chi}^2$ distribution when the number of observations is very large and the number of observed variables is small. However, in practice the number of observations might not be sufficiently large enough and thus the theoretical approximation to the distribution of test statistics would not be accurate. To date, there has not been a test which adequately performs in all distributional conditions at small sample sizes. Specially, when the number of variables in the models gets larger almost all existing tests break down. In practice there are many studies which use small sample sizes and often the p-values that are reported for the goodness of fit test in those studies are doubtful. In this study we propose a new Monte Carlo test that we expect to improve an asymptotic based test statistics. In a simulation study with a wide range of distributional conditions and sample sizes, we show that the proposed Monte Carlo method performs well in controlling Type I error. We discuss the existing bootstrap method in SEM and we shed some light on the problems that occur with the bootstrap test.

This dissertation is organized as follows: Chapter 1 provides an overview on some of the more important developments in goodness of fit test statistics. Chapter 2 contains a review on the bootstrap method in SEM, and the new Monte Carlo approach is introduced as an alternative resampling method. In Chapter 3 we examine the performance of the proposed Monte Carlo method and we compare it with the performance of existing tests. Chapter 5 concludes the dissertation and gives some suggestions for future studies on this topic.

CHAPTER 1

Goodness of Fit Test in SEM

If we skip early ideas of SEM in the form of factor analysis (e.g. Spearman, 1904; Thurstone, 1944) and move fast-forward, developments of statistical theories for goodness of fit test in SEM started with development of Confirmatory Factor Analysis (CFA) by Jöreskog, (1969). Since then, researchers developed various theoretical methods to evaluate overall model fit in more general situations. When the sample size is small and the number of observed variables are large, existing test statistics often fail to give a reliable evaluation of overall model fit. This becomes more problematic when the assumption of normality of observed variables is not valid. Specifically, most of the test statistics are not able to control Type I error and they either over or under reject a true model at small sample sizes. To date, researchers have tried to develop new test statistics and methods to overcome this problem. Often, by ad hoc approaches they have improved performance of some of the classical tests, yet, there does not exist a method to perform well in all conditions. In this chapter we give a review on some of the existing test statistics in SEM.

Sections of this chapter are organized as follows: Section 1.1 covers Maximum Likelihood test based on a normality assumption of the population. In section 1.2, in brief, we talk about distribution free tests and Generalized Least Square (GLS) method even though, GLS in a special format can also seen as a normal theory based test. In section 1.3 we discuss some corrections to the ML test. The core of our implementation of the Monte Carlo test is based on the Satorra-Bentler scaled test statistics; therefore, separately, in section 1.4 we describe Satorra-Bentler scaled test statistics. Other theoretical methods such as elliptical based tests, Heterogeneous kurtosis tests, and residual based tests are skipped since they are not relevant to our main argument.

1.1 Maximum Likelihood test

Let $\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_n$ be a multivariate random sample of size n from a p-variate population with $\mathrm{E}(\mathbf{x}_i) = \boldsymbol{\mu}$ and $\mathrm{cov}(\mathbf{x}_i) = \boldsymbol{\Sigma}$ for $i = 1, \ldots, n$. A covariance structure model $\boldsymbol{\Sigma}(\boldsymbol{\theta})$ with a q-dimensional vector of unknown parameter $\boldsymbol{\theta}$ is proposed to fit the data. We are interested to test the null hypothesis

$$H_0 : \boldsymbol{\Sigma}(\boldsymbol{\theta}) = \boldsymbol{\Sigma}, \tag{1.1}$$

against alternative hypothesis that the population covariance matrix can be any arbitrary positive definitive matrix. If the model is true, we wish to estimate $\boldsymbol{\theta}$ by an estimator $\hat{\boldsymbol{\theta}}$ such that $\boldsymbol{\Sigma}(\hat{\boldsymbol{\theta}})$ gets as close as possible to $\boldsymbol{\Sigma}$. Since $\boldsymbol{\Sigma}$ is unknown, in practice $\boldsymbol{\Sigma}$ is replaced with the sample covariance matrix, as an unbiased estimator of $\boldsymbol{\Sigma}$, which is

$$\boldsymbol{S} = \frac{1}{n-1} \sum_{i=1}^{n} (\mathbf{x}_i - \bar{\mathbf{x}}) (\mathbf{x}_i - \bar{\mathbf{x}})^T ,$$

where $\bar{\mathbf{x}} = \frac{1}{n} \sum_{i=1}^{n} \mathbf{x}_i$ is the sample mean. In covariance structure analysis, a discrepancy function $F[\boldsymbol{\Sigma}(\boldsymbol{\theta}), \boldsymbol{S}]$ measures the discrepancy between $\boldsymbol{\Sigma}(\boldsymbol{\theta})$ and $\boldsymbol{S}$, (see e.g., Browne 1974, 1984; Jöreskog 1967, 1969). The most common and reliable approach to problems of parameter estimation and testing for the covariance structure model is the normal theory approach, in which it is assumed that observed variables have a multivariate normal distribution. When $\mathbf{x}_i$ is from a multivariate normal distribution, i.e. $N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, the Maximum Likelihood (ML) discrepancy function is

$$F_{ML}(\boldsymbol{\theta}) = \ln |\boldsymbol{\Sigma}(\boldsymbol{\theta})| - \ln |\boldsymbol{S}| + \mathrm{tr}(\boldsymbol{S}\boldsymbol{\Sigma}^{-1}(\boldsymbol{\theta})) - p, \tag{1.2}$$

(see e.g., Jöreskog, 1969). Thus, the maximum likelihood estimate $\hat{\boldsymbol{\theta}}_{ML}$ obtained by minimizing F_{ML} and

$$T_{ML} = (n-1) F_{ML}(\hat{\boldsymbol{\theta}}) \tag{1.3}$$

is called the Maximum Likelihood test statistic. Moreover, the asymptotic distribution of T_{ML} is central $\boldsymbol{\chi}^2$ with $df = p^* - q$ degrees of freedom, where $p^* = p(p+1)/2$. Thus, for large sample size we can test the null hypothesis (1.1) and reject it if T_{ML} exceeds the critical value of $\boldsymbol{\chi}^2$ at significance level α.

1.2 Distribution free test statistic

An alternative approach to the normal theory model is the generalized least squares (GLS) method studied by Browne (1974). The GLS discrepancy function is defined as

$$F_{GLS}(\boldsymbol{\theta}) = \frac{1}{2}[\mathrm{vec}(\boldsymbol{S} - \boldsymbol{\Sigma}(\boldsymbol{\theta}))^T(\boldsymbol{V} \otimes \boldsymbol{V})\mathrm{vec}(\boldsymbol{S} - \boldsymbol{\Sigma}(\boldsymbol{\theta}))] = \frac{1}{2}\mathrm{tr}[(\boldsymbol{S} - \boldsymbol{\Sigma}(\boldsymbol{\theta}))\boldsymbol{V}]^2, \qquad (1.4)$$

where $\boldsymbol{V}$ is a $p \times p$ constant positive definite matrix or a stochastic matrix that converges in probability to a constant positive definitive matrix $\boldsymbol{V}^*$, vec() is vectorization operator that transforms a matrix to a vector by staking rows of the matrix, and $\otimes$ is the Kronecker product. Browne (1974) showed that if $\boldsymbol{V}$ converges in probability to $\boldsymbol{\Sigma}^{-1}$, then the GLS estimator $\hat{\boldsymbol{\theta}}_{GLS}$ that minimizes $F_{GLS}(\boldsymbol{\theta})$ is asymptotically equivalent to $\hat{\boldsymbol{\theta}}_{ML}$ and $T_{GLS} = (n-1)F_{GLS}(\hat{\boldsymbol{\theta}}_{GLS})$ has a $\boldsymbol{\chi}^2$ distribution with $df = p^* - q$ degrees of freedom.

To deal with the situation in which the distribution of observed variables is not multivariate normal, the asymptotically distribution free (ADF) covariance structure method was introduced by Browne(1984). Let $\boldsymbol{s}$ and $\boldsymbol{\sigma}(\boldsymbol{\theta})$ be $p^* \times 1$ column vectors formed by stacking elements of the lower triangle of the sample covariance, $\boldsymbol{S}$, and true population covariance, $\boldsymbol{\Sigma}(\boldsymbol{\theta})$, matrices by their rows, respectively. By multivariate central limit theorem,

$$n^{\frac{1}{2}}(\boldsymbol{s} - \boldsymbol{\sigma}(\boldsymbol{\theta})) \xrightarrow{L} N(\mathbf{0}, \boldsymbol{\Gamma}), \qquad (1.5)$$

where $\boldsymbol{\Gamma}$ is the asymptotic covariance matrix of $\sqrt{n}(\boldsymbol{s} - \boldsymbol{\sigma}(\boldsymbol{\theta}))$ with its elements defined by $\boldsymbol{\Gamma}[ij, kl] = \sigma_{ijkl} - \sigma_{ij}\sigma_{kl}$, given $\sigma_{ijkl} = \mathrm{E}[(x_i - \mu_i)(x_j - \mu_j)(x_k - \mu_k)(x_l - \mu_l)]$ and σ_{ij} and σ_{kl} being elements of population covariance matrix. The limit notation "$\xrightarrow{L}$" stated for convergence in distribution. The asymptotic distribution given by (1.5) is crucial to justify asymptotic properties of ADF.

The ADF discrepancy function also known as arbitrarily distribution generalized least square (AGLS) function is defined by

$$F_{ADF}(\boldsymbol{\theta}) = (\boldsymbol{s} - \boldsymbol{\sigma}(\boldsymbol{\theta}))^T \mathbf{W}^{-1} (\boldsymbol{s} - \boldsymbol{\sigma}(\boldsymbol{\theta})), \tag{1.6}$$

where $\mathbf{W}$ is a $p^* \times p^*$ positive definite weighted matrix. The optimal weight matrix is the one in which $F_{ADF}(\boldsymbol{\theta})$ reaches to its minimum value. For instance, a $\mathbf{W}$ that converges in probability to $\mathbf{\Gamma}$ is optimal. In practice, one can use the weight matrix as a consistent estimator of $\mathbf{\Gamma}$ with its typical element obtained by

$$\hat{\mathbf{\Gamma}}[ij, kl] = s_{ijkl} - s_{ij} s_{kl},$$

where $s_{ijkl} = 1/n \sum_{t=1}^{n} (x_{ti} - \bar{x}_i)(x_{tj} - \bar{x}_j)(x_{tk} - \bar{x}_k)(x_{tl} - \bar{x}_l)$ is the multivariate sample fourth moment and s_{ij}s are elements of sample covariance matrix. The ADF estimator $\hat{\boldsymbol{\theta}}_{ADF}$ minimizes $F_{ADF}(\boldsymbol{\theta})$ and

$$(n-1) F_{ADF}(\hat{\boldsymbol{\theta}}_{ADF}) \xrightarrow{L} \boldsymbol{\chi}^2_{p^*-q}.$$

The ADF test statistics, T_{ADF} involves the sample fourth moments and requires a very large sample size to estimate the model. Due to the need for large sample sizes, and computational problems of the ADF method in smaller to moderate sample size, further discussions about this type of test statistic are omitted here.

1.3 Correction to ML test

It is known that the asymptotic distribution of the likelihood ratio test statistic for higher dimensional data is only valid for very large sample sizes. For smaller sample sizes, even when the data are from normal distribution, it overestimates nominal Type I errors (see e.g., Bentler & Chou, 1987; Boomsma, 1982; Moshagen, 2012). It is also known that the normal model is very sensitive to violation of normality of observed variables (see e.g., Bentler &

Yuan, 1999; Browne, 1987; Chou, Bentler, & Satorra, 1991; Hu, Bentler, & Kano, 1992; Muthén & Kaplan, 1992; Yuan & Bentler, 1998).

To improve the performance of the likelihood ratio test statistics under the normality assumption in order to get better control of Type I errors, various corrections to T_{ML} have been proposed. The basic idea of these correction methods is to take account of the dimension of the model and modify the test statistics by multiplying them by a scale factor that depends on sample size and number of observed and latent variables. This modification adjusts the mean of the test statistics for small sample size and, in most situations, decreases the rejection rate and therefore controls the Type I error. Bartlett (1950), as part of a series of corrections to likelihood test statistics in multivariate data analysis problems, introduced a correction for the likelihood test statistics in factor analysis by replacing $(n-1)$ in equation (1.3) with $B = n - (2p+11)/6 - 2n_f/3$, which is equivalent with multiplication of T_{ML} by

$$c_b = 1 - \frac{2p + 4n_f + 5}{6(n-1)}, \tag{1.7}$$

where p is number of indicators, n is sample size and n_f is the number of latent factors. A Study by Fouladi (2000) showed that the Bartlett correction as $T_{MLb} = c_b T_{ML}$ can be approximated more closely than T_{ML} by $\boldsymbol{\chi}^2$ with $p^* - q$ degrees of freedom, thus it can improve the performance of T_{ML} in SEM for small sample sizes. However, Monte Carlo simulation suggested that the Bartlett correction over-corrects the mean of T_{ML} in general SEM with higher model dimension, and it reduces the rejection rate below the nominal Type I error (see e.g., Herzog, Boomsma, & Reinecke, 2007; Nevitt & Hancock, 2004). Yuan (2005) suggested to approximate the test statistics with a linear transformation to chi-square in the form of $b\boldsymbol{\chi}_{(df)} + a$ and he also proposed an ad-hoc correction by replacing $(n-1)$ in equation (1.3) with $n - (2p+13)/6 - n_f/3$. Herzog and Boomsma (2009) found that the Yuan correction also rejects the correct models with a type one error less than the nominal level.

Another approach to adjust likelihood test statistics in the normal model was developed by Swain (1975). Swain's multiplication factor to T_{ML} is

$$c_s = 1 - \frac{p(2p^2+3p-1) - h(2h^2+3h-1)}{12d(n-1)}, \tag{1.8}$$

where,

$$h = (\sqrt{1+4p(p+1)-8d} - 1)/2,$$

and d is the degrees of freedom of the model. Swain's corrected test statistic is defined as $T_{MLs} = c_s T_{ML} \sim \boldsymbol{\chi}^2_{(df)}$ (for comparisons of Swain's corrected test statistic to Bartlett's correction see for example, Fouladi, 2000; Herzog et al., 2007; Nevitt & Hancock, 2004). Herzog et al. (2007) suggested that T_{MLs} should be applied when large structural equation models are analyzed and the observed variables have a multivariate normal distribution. A recent paper by Yuan, Tian, and Yanagihara (2015) followed up the idea of Yuan (2005), proposing a method to approximate a correction to T_{ML} empirically, so that the mean of the resulting statistic get approximately equal the degrees of freedom of the nominal chi-square distribution. They called their corrected test statistics $T_{MLe} = \hat{c}_e T_{ML}$ which they claimed can be approximated very accurately by $\boldsymbol{\chi}^2_{(df)}$. In their prediction method they estimate $\hat{c}_e$ using empirical and simulated results.

1.4 Satorra-Bentler scaled test statistics

Satorra & Bentler (1986, 1988, and 1994) developed test statistics for distribution free models to overcome the computational problems with the ADF methods at small sample sizes and to achieve more reliable test statistics. When the distribution of data is not from a multivariate normal, the asymptotic distribution of T_{ML} from equation (1.3) is not $\boldsymbol{\chi}^2_{(p^*-q)}$. Instead, Satorra and Bentler (1988) described the asymptotic distribution of T_{ML} and its variants such as T_{GLS} as weighted sum of $p^* - q$ independent χ^2 with 1 degree of freedom.

When we use the normality assumption to determine a discrepancy function, the optimal weight matrix is the one which converges to a simplified form of $\boldsymbol{\Gamma}$. For instance, in practice in the normal theory ML we can set weight matrix as $\mathbf{W} = 2\mathbf{K}_p^T(\hat{\mathbf{\Sigma}} \otimes \hat{\mathbf{\Sigma}})\mathbf{K}_p$, where $\hat{\mathbf{\Sigma}}$ is a consistent estimate of $\mathbf{\Sigma}$ and $\mathbf{K}_p$ is $p^2 \times p^*$ duplication matrix. Let $\dot{\boldsymbol{\sigma}}(\boldsymbol{\theta}) = \partial\boldsymbol{\sigma}(\boldsymbol{\theta})/\partial\boldsymbol{\theta}^T$ be the Jacobian matrix in numerator layout notation evaluated at true parameters of the model. If for simplicity we set $\dot{\boldsymbol{\sigma}} = \dot{\boldsymbol{\sigma}}(\boldsymbol{\theta})$, the residual weight matrix under the model is given by

$$\mathbf{U} = \mathbf{W}^{-1} - \mathbf{W}^{-1}\dot{\boldsymbol{\sigma}}(\dot{\boldsymbol{\sigma}}^T\mathbf{W}^{-1}\dot{\boldsymbol{\sigma}})^{-1}\dot{\boldsymbol{\sigma}}^T\mathbf{W}^{-1},$$

where $\mathbf{W}$ is the weight matrix which is used to estimate parameters (see e.g. Bentler & Dudgeon, 1996). The asymptotic distribution of T_{ML} without a normality assumption was obtained by Satorra and Bentler (1988) as

$$T_{ML} \xrightarrow{L} \sum_{i=1}^{p^*-q} \alpha_i \tau_i \tag{1.9}$$

where τ_is are independent and have χ^2 distribution with 1 degree of freedom and α_is are non zero eigenvalues of the matrix $\mathbf{U\Gamma}$, where $\mathbf{\Gamma}$ is the the asymptotic covariance matrix of $\sqrt{n}(\boldsymbol{s} - \boldsymbol{\sigma}(\boldsymbol{\theta}))$ defined in (1.5).

Since the distribution in the right hand side of (1.9) is not known, Satorra and Bentler (1988) first introduced a correction factor to T_{ML} that re-scales its mean to the degrees of freedom of the asymptotic χ^2 distribution. It is easy to see that the mean of the asymptotic distribution of (1.9) is $\sum_{i=1}^{p^*-q} \alpha_i = \mathrm{tr}(\mathbf{U\Gamma})$. The SB scaled test statistic is defined as

$$T_{SB} = \frac{T_{ML}}{c}, \tag{1.10}$$

where $c = \mathrm{tr}(\hat{\mathbf{U}}\hat{\mathbf{\Gamma}})/d$, and where $\mathbf{U}$ and $\mathbf{\Gamma}$ being replaced by their consistent estimators $\hat{\mathbf{U}}$ and $\hat{\mathbf{\Gamma}}$, respectively. Satorra and Bentler showed that if observations are from an elliptical distribution then T_{SB} asymptotically distributed as $\boldsymbol{\chi}^2$ with $d = p^* - q$ degrees of freedom and even in non elliptical distributions the empirical distribution of T_{SB} can be approximated fairly well by $\boldsymbol{\chi}^2_{(d)}$ distribution (see e.g. Satorra & Bentler, 1988, 1994; Yuan & Bentler, 1998).

Secondly, inspired by the Satterthwaite (1941) variance correction to a linear combination of chi-square variates, Satorra and Bentler (1988) proposed a correction factor that adjusts both mean and variance of the test statistic. The variance of asymptotic distribution of T_{ML}, driven from (1.9), is $2\text{tr}[(\mathbf{U\Gamma})^2]$. Suppose a correction of the form aT_{ML} can be approximated by a χ^2 distribution with a given degrees of freedom, by setting its mean and variance equal to mean and variance of the χ^2 distribution, the Satorra-Bentler mean and variance adjusted test statistic is obtained by

$$T_{MVA} = \frac{d'}{\text{tr}(\hat{\mathbf{U}}\hat{\mathbf{\Gamma}})} T_{ML}, \tag{1.11}$$

where

$$d' = \frac{[\text{tr}(\hat{\mathbf{U}}\hat{\mathbf{\Gamma}})]^2}{\text{tr}[(\hat{\mathbf{U}}\hat{\mathbf{\Gamma}})^2]}$$

and T_{MVA}, asymptotically, has a $\boldsymbol{\chi}^2$ distribution with adjusted d' degrees of freedom. Note that the degrees of freedom, d', is not integer and to find the cutoff value, a $\boldsymbol{\chi}^2$ with noninteger degrees of freedom needs to be evaluated.

Asparouhov and Muthén (2010) introduced another implementation to the mean and variance adjusted test statistic that uses the usual degrees of freedom d. They assumed a correction of the form $aT_{ML}+b$ that is assumed to have a χ^2 distribution with usual $d = p^* - q$ degrees of freedom. The expected value and variance of the corrected test statistic is given by

$$\text{E}(aT_{ML} + b) = a\text{E}(T_{ML}) + b = a\text{tr}(\mathbf{U\Gamma}) + b = d$$

and

$$\text{Var}(aT_{ML} + b) = a^2\text{Var}(T_{ML}) = 2a^2\text{tr}[(\mathbf{U\Gamma})^2] = 2d.$$

From the second equation we can obtain $a = \sqrt{d/\text{tr}[(\mathbf{U\Gamma})^2]}$ and with replacing it to the first equation $b = d - \sqrt{d \times [\text{tr}(\mathbf{U\Gamma})]^2/\text{tr}[(\mathbf{U\Gamma})^2]}$. Thus, the Asparouhov-Muthén test statistic is given by

$$T_{AM} = T_{ML}\sqrt{\frac{d}{\text{tr}[(\hat{\mathbf{U}}\hat{\mathbf{\Gamma}})^2]}} + d - \sqrt{\frac{d \times \text{tr}(\hat{\mathbf{U}}\hat{\mathbf{\Gamma}})]^2}{\text{tr}[(\hat{\mathbf{U}}\hat{\mathbf{\Gamma}})^2]}}, \tag{1.12}$$

which asymptotically has a χ^2 distribution with d degrees of freedom. The simulation study by Asparouhov and Muthén (2010) and our simulation study here in chapter 3 shows T_{MVA} and T_{AM} perform similarly.

Even thought T_{SB} and T_{MVA} can give a promising results when the observed variables are not from a multivariate normal distribution, simulation studies have shown in covariance models with small sample size T_{SB} test also over-rejects the null hypothesis when it is true, and T_{MVA} seems to over-correct the test statistics and rejects the null hypothesis less than nominal Type I error (see e.g., Fouladi, 2000; Herzog et al. 2007; Nevitt & Hancock 2004). Despite the lack of statistical justification Bartlett and Swain corrections has been applied to T_{SB} and T_{MVA} by researchers. Results from simulation studies showed some improvement to T_{SB} but since those correction factors are often less than one, T_{MVA} rejected the null hypothesis way less than the nominal Type I error (see e.g., Fouladi, 2000; Herzog et al. 2007; Nevitt & Hancock 2004). Lin and Bentler (2012) proposed a third moment adjusted test based on scaling the mean and adjusting for the skewness of the test statistic and they showed their new test performs better than previous corrections by Satorra and Bentler in some condition for very small sample size. However, expanded evaluation of their test suggested that the mean scaled and skewness adjusted test performs well only under normal distributions and the Satorra-Bentler scaled test performs overall better than other competitors (see e.g., Tong & Bentler, 2013). Wu and Lin (2016) also proposed a new scaled F distribution approximation to the test statistics which they found to perform similar to Satorra-Bentler mean and variance adjusted test in controlling type one errors. Recently, Jiang and Yuan (2017) introduced four new corrected statistics and they showed their proposed test statistics control Type I error more accurately than existing tests at small sample sizes in nonnormal data. Their first three tests involve the rank of matrix $\hat{\mathbf{U}}\hat{\mathbf{\Gamma}}$ in situations that the rank of $\hat{\mathbf{U}}\hat{\mathbf{\Gamma}}$ is not equal to $p^* - q$. Bentler and Yuan (1999) previously noted that in order to use T_{SB} legitimately, the rank of $\hat{\mathbf{U}}\hat{\mathbf{\Gamma}}$ needs to be equal to $p^* - q$ which in practice is not always the case.

CHAPTER 2

Resampling Methods

In this section we describe use of bootstrapping in SEM and propose a new Monte Carlo approach that we expect to improve evaluation of overall model fit. Monte Carlo tests existed before introduction of bootstrapping however, the generality of bootstrap applications in statistics overshadowed the use of 'Monte Carlo test'. Specially in parametric problems, despite the differences between two methods, the term 'parametric bootstrap' is often used instead. According to Hall and Titterington (1989), "However, there are important differences between bootstrap methods and Monte Carlo tests; the latter are specifically designed to exploit the advantages of 'blurring' in a simulation study (Marriott, 1979)" (p. 460). In statistical inference, bootstrap methods have been introduced to estimate a test statistic which either does not have a known asymptotic distribution or the sample sizes are not large enough that the statistic converges to its asymptotic distribution. In SEM bootstrapping also have been used and implemented in almost all SEM software such as LISREL (Jöreskog & Sörbom, 1996) and EQS (Bentler, 2006). The Monte Carlo approach is general in the sense that it can be implemented to any pivotal test to improve its level of accuracy at small sample sizes. In section 2.1 we describe bootstrap methods in SEM. Section 2.2 contains the proposed Monte Carlo method.

2.1 Bootstrap method in SEM

When the asymptotic distribution of the test statistics is not available or when the assumptions for the asymptotic theories are not valid, the bootstrap method as an alternative to the theoretical statistical testing methods is useful. The bootstrap method introduced by Efron (1979) is a resampling procedure. The basic idea of bootstrapping is simple. First we compute the observed test statistics based on the observations. Then, we draw B samples of the same size as the original observations with replacement from the original observations and compute the test statistics for each new sample to create the empirical distribution of the test statistic. From this empirical distribution we can find the rejection cut-off point or p-value. Despite the huge impact of the bootstrap method in many areas of statistics, the usefulness of the bootstrap method in SEM was considered cautiously by researchers. Most focus of using bootstrap in SEM has been in parameter standard error estimation and the overall goodness of fit test did not get the same attention among structural equation modelers. However, in some cases due to the lack of existence of an appropriate theoretical method, bootstrapping might be the only way to attack the problem (see e.g., Sharma & Kim, 2013). Bollen and Stine (1992) showed that the naive bootstrap, when bootstrap samples are drawn directly from the original data, will be inaccurate. This is simply because the bootstrap samples are not generated from a population that supports the null hypothesis. They introduced a modified parametric bootstrap method by adjusting the original data in a way that the resampling are taken from a set of data that ensure the null hypothesis. They adjusted the original data matrix $\boldsymbol{X}$ by

$$\mathbf{Z} = \boldsymbol{X}\boldsymbol{S}^{-1/2}\boldsymbol{\Sigma}^{1/2}(\hat{\boldsymbol{\theta}}), \tag{2.1}$$

where the power $\frac{1}{2}$ represents the matrix square root. Since the covariance matrix of $\mathbf{Z}$ is $\boldsymbol{\Sigma}(\hat{\boldsymbol{\theta}})$ they suggest to draw bootstrap samples from $\mathbf{Z}$ and not the original sample. Application of bootstrapping in factor analysis was studied by Ichikawa and Konishi (1995). Later, Yung and Bentler (1996) studied applications and usefulness of the bootstrap method in SEM. Those studies suggest that the bootstrap is effective in moderate to large sample size and when the number of observations is less than 150 the bootstrap results are not accurate. Yung and Bentler (1994) used the bootstrap method to correct ADF test statistics and their results suggested the improvement of corrected ADF test statistics for sample size as low as 500. Nevitt and Hancoock (2001) in a Monte Carlo study evaluated the performance of bootstrap in different scenarios and they also concluded that in order to use bootstrap effectively, sample size needs to be moderate or large and for very small sample size the bootstrap cannot be trusted.

Although bootstrapping is a very effective method in statistics, there have been many doubts about using the bootstrap method in higher dimensional data (see e.g., El Karoui & Purdom 2015). However, researchers still trying to develop new varieties of bootstrap methods in special situations (see e.g., Cornea-Madeira & Davidson 2014).

Inspired by the fact that the correction factors by Bartlett and Swain do not depend on unknown parameters of the model, in the next section a Monte Carlo approximation to T_{ML} is suggested to approximate the distribution of the test statistic empirically when the sample size is very small. The proposed method is very similar to the parametric bootstrapping procedure but instead of resampling from the original data, the Monte Carlo approach uses artificially generated random samples from a multivariate normal distribution. This approach contrasts to finding an empirical correction factor as proposed by Yuan et al. (2015). The Monte Carlo approximation also extends to T_{SB} for problems in which the distribution of observations is not multivariate normal.

2.2 Monte Carlo test

Monte Carlo simulation has been used for years to evaluate adequacy of a test statistic. In the Monte Carlo study of a test statistic T, a known structured model is assumed to be true in the population. Many, e.g $k = 1, \ldots, 1,000$, random samples based on a specific distribution are replicated by computer under the true model and for each replication $T_{(k)}$ statistic is computed. Then the empirical distribution of those k test statistics is compared to the theoretical distribution (for example, χ^2). While Monte Carlo simulation has been practiced for years in SEM, the use of Monte Carlo simulation to estimate the distribution of the test statistic when evaluating a given structured model in a specific sample has not been studied in the literature.

The use of Monte Carlo tests dates before the bootstrap methods introduced by Efron (1979). In a discussion on the spectral analysis of point processes (Bartlett, 1963), Bernard proposed a Monte Carlo method for the first time in a parametric context. Later, in a more theoretical approach, the power of Monte Carlo tests was studied by Hope (1968) and compared with uniformly most powerful tests. However, after more attraction and applicability of Efron's resampling methods (1979), in both parametric and nonparametric forms, Monte Carlo tests have been dominated by bootstrapping and did not get enough attention by practitioners. Hall and Titterington (1989) discussed the level accuracy and the power of Monte Carlo tests and showed if a Monte Carlo test is based on an asymptotically pivotal statistic, then it is more accurate than the corresponding asymptotic test in terms of Type I error (for more technical detail in the context of bootstrapping see also, Beran, 1988; Hall, 1992). A test statistic is asymptotically pivotal if the asymptotic distribution of the test statistic does not depend on any unknown quantity.

Let π denote the unknown population which generated the original data matrix $\boldsymbol{X}$ with sample size n. In the parametric Monte Carlo tests the population π depends on the parameters of interest and also on some nuisance parameters which are accounted for through the analysis. Suppose $T(\boldsymbol{X})$ is an asymptotically pivotal test statistic which is used for our judgment about the null hypothesis. Let $\hat{\pi}$ be the estimated population by replacing parameters of population by their estimated values. We draw M samples $\boldsymbol{X}_1^*, \boldsymbol{X}_2^*, \ldots, \boldsymbol{X}_M^*$ of size n from $\hat{\pi}$ and compute $T_1^*, T_2^*, \ldots, T_M^*$ from $\boldsymbol{X}_1^*, \boldsymbol{X}_2^*, \ldots, \boldsymbol{X}_M^*$ in the same manner as we computed T from $\boldsymbol{X}$. We rank $T_1^*, T_2^*, \ldots, T_M^*$ as $T_{(1)}^* \leq T_{(2)}^* \leq \ldots \leq T_{(M)}^*$. Based on the level of the test we determine m as integer part of αM and we reject the null hypothesis if $T > T_{(m)}^*$.

Test statistics in SEM are mostly based on asymptotic distributions that do not involve parameters (i.e. $\boldsymbol{\theta}$ and other unknown characteristics of the population), therefore the pivotalness condition of the test statistic is met. Then, we can claim that the desired distribution of those test statistics can be obtained under the null hypothesis using different sets of data that agree with the null hypothesis. This assumes, of course, that we do not violate the regularity conditions for that specific test statistic (i.e. conditions of the asymptotics are valid).

For example in ML test statistics, let us assume that there exist a "best" correction factor c^* for a specific model with specific number of observed and latent variables that depends on sample size n. By the best correction factor we mean that when the null hypothesis (1.1) is true then the distribution of $c^* T_{ML}(\hat{\boldsymbol{\theta}})$ is closest to $\boldsymbol{\chi}^2_{(d)}$ than it would be by applying any other correction factors, e.g. c_b, c_s, or $\hat{c}_e$. Since the degrees of freedom of chi-square distribution and also c^* do not depend on the true population parameter $\boldsymbol{\theta} \in \Omega_0$; where Ω_0 is the parameter space under the null hypothesis, then the distribution of $\boldsymbol{\chi}^2_{(d)}/c^*$ does not change for any other choice of parameter $\boldsymbol{\theta}^* \in \Omega_0$ and can be approximated by $T_{ML}(\hat{\boldsymbol{\theta}^*})$. Therefore, to approximate the null distribution of T_{ML} we can generate many independent multivariate normal random data sets with the same size and dimension as the original sample under the same structured model in the null hypothesis and use any arbitrarily parameters, e.g. $\hat{\boldsymbol{\theta}}_{ML}$ to evaluate $T^{(k)}_{ML}$ for each kth simulated data set. We call this approach a Monte Carlo approximation method. It is clear that the asymptotic distribution of each $T^{(k)}_{ML}$ is also $\boldsymbol{\chi}^2$ with the same degrees of freedom as the observed test statistics. However, when the sample size is small, the empirical distribution obtained from $T^{(k)}_{ML}$ is expected to be a better approximation than any correction method that involves a correction factor such as c_b, c_s, and $\hat{c}_e$. Moreover, the Monte Carlo approximation method can carry on any other correction to T_{ML} that might be approximated by a distribution different than the form of $\boldsymbol{\chi}^2_{(d)}/c^*$ that is obtained from a specific structured model.

If the distribution of observation is unknown and cannot assumed to be normal, we can use the same Monte Carlo algorithm with a distribution free test statistic, e.g T_{SB}. Let $\mathbf{x}_i$, for $i = 1, \ldots, n$ be a multivariate random sample of size n from a p-variate population with unknown distribution and the population structured covariance model with model parameter $\boldsymbol{\theta} \in \Omega_0$. Satorra & Bentler (1988) showed that if x_i is from an elliptical distribution then $T_{SB}(\hat{\boldsymbol{\theta}}, \boldsymbol{X})$, where $\boldsymbol{X}$ is the matrix of observation with rows of x_i', asymptotically has a $\boldsymbol{\chi}^2_{(df)}$ with $df = p^* - q$. They also indicated that even if the data are not elliptical, $\boldsymbol{\chi}^2$ can give a valid approximation to T_{SB}. Also let $\mathbf{y}_i$, for $i = 1, \ldots, n$ be a Monte Carlo multivariate random sample of size n from a p-variate normal population with the same structured covariance model with parameter $\boldsymbol{\theta}^* \in \Omega_0$. Similarly $T_{SB}(\hat{\boldsymbol{\theta}}^*, \mathbf{Y})$, where $\mathbf{Y}$ is the matrix with rows of y_i', can be approximated by the same $\boldsymbol{\chi}^2$ distribution. With the similar analogy as the normal model we can argue that the approximation for $T_{SB}(\hat{\boldsymbol{\theta}}, \boldsymbol{X})$ with $T_{SB}(\hat{\boldsymbol{\theta}^*}, \mathbf{Y})$ should be better than any kind of Bartlett correction factor method in term of controlling Type I error. Note that any specific form of correction factor for T_{SB} does not exist in the literature, and researchers often use the same correction factors as those used to correct T_{ML}. In the next section, the effect of the model parameter $\boldsymbol{\theta}$ on the distribution of test statistics, i.e. $T_{SB}(\hat{\boldsymbol{\theta}})$ and $T_{ML}(\hat{\boldsymbol{\theta}})$ for small sample size is evaluated.

The Monte Carlo algorithm to test hypothesis (1.1) can be described with the following steps:

1. Assuming that the null hypothesis model is true, estimate $\hat{\boldsymbol{\theta}}_{ML}$ and, if normality of observation can be assumed, evaluate observed T_{ML}; Otherwise evaluate T_{SB}.

2. Using parameters estimated in step one, draw M independent multivariate normal random samples of size n with dimension p and structured covariance matrix $\boldsymbol{\Sigma}(\hat{\boldsymbol{\theta}}_{ML})$.

3. For all $m = 1, \ldots, M$, random samples in step two, if the normality assumption is valid for observations compute a new $\boldsymbol{S}_{(m)}$ and fit the model to get a new $\boldsymbol{\Sigma}_{(m)}(\hat{\boldsymbol{\theta}})$ then compute $T_{ML}^{(m)}$ or compute $T_{SB}^{(m)}$ if normality cannot be assumed.

4. At the significance level α if the observed test statistics in part one is greater than $100(1-\alpha)\%$ of simulated test statistics in step three, reject the null hypothesis (1.1).

Or alternatively, find the p-value of the hypothesis test from the number of simulated test statistics in step three that are greater than the observed test statistics in step one, divided by M.

Monte Carlo tests do not give a well defined rejection region compared to conventional test statistics. Since the rejection region of the Monte Carlo tests is based on a random sample it cause varying probabilities of rejection. This is called blurring of the rejection region in Monte Carlo tests (see e.g. Marriott, 1979). We can reduce the blurring by taking the number of Monte Carlo samples sufficiently large. Marriott (1979) suggested 99 Monte Carlo samples as an appropriate number of samples in a Monte Carlo test. However, to get more accurate results, in the simulation study in the next chapter we used 1000 samples for our Monte Carlo test.

CHAPTER 3

Simulation Study

In this chapter we study the behavior of the proposed Monte Carlo method in comparison to other goodness of fit tests in populations with different distribution conditions. Since the aim of our study is to introduce a test that can evaluate higher dimensional models at small sample sizes, in our simulation study we are considering data with larger number of variables and smaller samples compared to previous Monte Carlo studies. Although, in the concept of hypothesis testing the tail behavior of the distribution of a test statistic is crucial, we also measure and report the overall difference between empirical distribution of a test statistic and its reference distribution.

Simulation studies in this chapter show that the newly proposed Monte Carlo approach generally outperforms its existing competitors in terms of Type I error rejection rates. Also, a power study of the proposed test indicates satisfying power. Moreover, we found that the Bollen–Stine bootstrap has critical problems when the dimension of data gets larger. Previous Monte Carlo studies of bootstrapping were limited to the tail distribution of the test statistic. Our study shows that the bootstrap method does not perform well to evaluate the overall distribution of the test statistic. Errors associated with sampling methods, that are either caused by limited number of bootstrap samples or the test statistic not being pivotal, were also demonstrated. It is shown that the variation of p-values of the Monte Carlo test is exceptionally less than that of the Bollen–Stine bootstrap. We will also shed some light on the effect of estimated parameters on the distribution of test statistics. Materials covered in this chapter are organized as follows:

In section 3.1 we describe the conditions of the simulation study including model and distributions used to generate data and methods. In section 3.2 we study the behavior of test statistics in controlling level of Type I error. In section 3.3 we take a closer look at traditional bootstrap test in SEM and we discuss problems that arise using the Bollen-Stine bootstrap approach. In section 3.4 we study the effect of values of estimated parameters in test statistics in small sample sizes. Since the pivotalness of the test statistic is an essential assumption in our Monte Carlo approach, in section 3.5 we demonstrate the convergence of Satorra-Bentler scaled statistic. Finally, power is discussed in section 3.6.

3.1 Models and data generation designs

In this study two data generation schema are considered to draw the data from a population with specific structured covariance model (For similar data generation designs see e.g. Hu et al. 1992 and Yuan & Bentler 1998). Consider a confirmatory factor model defined by

$$\mathbf{x} = \mathbf{\Lambda f} + \boldsymbol{e} \tag{3.1}$$

with 3 common factors and 5 indicators per factor. The covariance matrix between factors is given by

$$\mathbf{\Phi} = \begin{pmatrix} 1 & 0.3 & 0.4 \\ 0.3 & 1 & 0.5 \\ 0.4 & 0.5 & 1 \end{pmatrix}.$$

The factor loading matrix is defined as a 15 by 3 matrix

$$\mathbf{\Lambda} = \begin{pmatrix} \boldsymbol{\lambda} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\lambda} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \boldsymbol{\lambda} \end{pmatrix}$$

where $\boldsymbol{\lambda} = (0.70, 0.70, 0.75, 0.80, 0.80)^T$ and $\mathbf{0}$ is a vector of zeros. Unique factors $\boldsymbol{e}$ are assumed to be uncorrelated with diagonal covariance matrix $\mathbf{\Psi}$ set to make variances of $\mathbf{x}$

equal to 1. Thus, the population structured covariance matrix is

$$\boldsymbol{\Sigma}(\boldsymbol{\theta}) = \boldsymbol{\Lambda}\boldsymbol{\Phi}\boldsymbol{\Lambda}^T + \boldsymbol{\Psi}. \tag{3.2}$$

This model has 33 free parameters with degrees of freedom of 87. In the first data generation schema we generate common factors, $\mathbf{f}$, and unique factors, $\boldsymbol{e}$, and we use equation (3.1) to draw our samples. Four different distribution conditions have been applied as follow:

- Condition A_1: Common factors $\mathbf{f}$ are drawn from multivariate normal distribution with zero mean and covariance $\boldsymbol{\Phi}$. Unique factors $\boldsymbol{e}$ are drawn from a multivariate normal distribution with zero mean and covariance $\boldsymbol{\Psi}$. Condition A_1 generates multivariate normal observations with zero mean and covariance $\boldsymbol{\Sigma}(\boldsymbol{\theta})$.

- Condition A_2: First we draw $\mathbf{f}_1$ from $N(\mathbf{0}, \boldsymbol{\Phi})$ and $\boldsymbol{e}_1$ from $N(\mathbf{0}, \boldsymbol{\Psi})$ and a single variable R from $\sqrt{\chi^2_5/3}$. Then Common factors $\mathbf{f}$ and unique factors $\boldsymbol{e}$ are calculated by $\mathbf{f}_1/R$ and $\boldsymbol{e}_1/R$ respectively.

- Condition A_3: This condition is similar to A_2 but $\boldsymbol{e}_1$ is drawn from a multivariate log-normal distribution with mean vector $\mathbf{0}$ and covariance matrix $\boldsymbol{\Psi}$.

- Condition A_4: This condition is also similar to A_2 with the difference that $\mathbf{f}_1 \sim \mathit{Lognormal}(\mathbf{0}, \boldsymbol{\Phi})$ and $\boldsymbol{e}_1 \sim \mathit{Lognormal}(\mathbf{0}, \boldsymbol{\Psi})$.

To generate data from a multivariate log-normal, first we draw independent random variables from a standard normal distribution and after using an exponential transformation we standardize them according to their mean and variance. With multiplying the resulting multivariate random variable by the square root of the covariance matrix, we achieve a multivariate random variable with the given covariance matrix. Readers should notice that the marginal distribution of the result will not have log-normal distribution but here for convenience we call it multivariate log-normal.

For the second data generation method, we use the structured covariance matrix defined in the equation (3.2). First we generate a 15 dimensional multivariate sample, $\boldsymbol{\xi}$, with an identity covariance matrix. Then we set

$$\mathbf{x} = \boldsymbol{\Sigma}(\boldsymbol{\theta})^{1/2}\boldsymbol{\xi}. \tag{3.3}$$

Therefore, the confirmatory factor model (3.1) should fit the data that is drawn from equation (3.3). Following two distribution conditions are considered for the data generation schema in the equation (3.3).

- Condition A_5: $\boldsymbol{\xi}$ is drawn from a multivariate log-normal distribution with the mean vector $\mathbf{0}$ and identity covariance matrix $\mathbf{I}$.
- Condition A_6: $\boldsymbol{\xi} = \boldsymbol{\xi}_1/R$ where $\boldsymbol{\xi}_1$ is drawn from a multivariate log-normal distribution with mean vector $\mathbf{0}$ and identity covariance matrix $\mathbf{I}$ and R is a single variable from $\sqrt{\chi^2_5/3}$.

The difference between conditions 5 and 6 to conditions 1 to 4 is that the fourth order moment matrices of the observed variables are different (see e.g. Yuan & Bentler 1998). To study the change in the behavior of the goodness of fit test statistics in higher dimensional models, we also expand the above 6 conditions to a 30 dimensional model that have 3 factors and 10 dependent variables for each factor. The factor covariance matrix is similar to the model in (3.1). The factor loading matrix is a 30 by 3 matrix with loading for each factor defined as $\boldsymbol{\lambda} = (0.70, 0.70, 0.75, 0.80, 0.80, 0.70, 0.70, 0.75, 0.80, 0.80)^T$. The factor covariance matrix $\boldsymbol{\Phi}$ remains the same. The diagonal covariance matrix $\boldsymbol{\Psi}$ will be a 30 by 30 matrix that sets to make the variance of observed variables equal to 1. This model has 63 free parameters and the degrees of freedom is 402.

We draw data from 6 conditions similar to conditions A_1, A_2, A_3, A_4, A_5, and A_6 that previously discussed and we call them B_1, B_2, B_3, B_4, B_5, and B_6 respectively. Here, the letter A means that the model is 15 dimensional and letter B means that the model is a 30 dimensional model. For example, B_1 is the same data generation schema as A_1 with 30 observations instead of 15 observations. For each of those 12 models and conditions, 7 different sample sizes included in this study indicating 50, 100, and 150 as small, 300 and 500 as moderate, and 2000 and 5000 as large samples. This makes $12 \times 7 = 84$ cases in total.

The simulation study in this chapter has been done with an R program. Estimation of parameters and test statistics were computed using the package lavaan, (Rosseel, 2012) with the "EQS" option for optimization. The results from parameter estimation and test statistics has been also verified by commercial EQS software version 6.2, Bentler (2006).

For each case we generated a sample based on the true hypothesized model (in the power analysis in section 3.6 from a misspecified model) and for each sample we fit the the data using maximum likelihood estimation. We compute related test statistics for each sample (i.e., T_{ML}, T_{SB}, T_{MVA}). The process is replicated 1,000 times. For each replication we draw 1,000 bootstrap samples to obtain the bootstrap test and 1,000 artificial normal samples for the proposed Monte Carlo method.

This process computationally is very time consuming. For each one of 84 cases we need to solve 2,001,000 factor models. In some cases the optimization does not converge or we will get negative variances that makes the number of computations even larger (those cases were omitted until the necessarily number of samples with a converged solution is reached). To expedite the computation process a cluster computation from the Hoffman2 shared cluster[1] has been used. 1,000 replications were distributed within 100 to 200 computing nodes, and each computing node ran 5 to 10 replications independently.

[1] The Hoffman2 cluster is a campus computing resource at UCLA, maintain by the Institute for Digital Research and Education (IDRE).

3.2 Type I error analysis

This section reports on a comprehensive simulation study that examines the performance of the proposed Monte Carlo approach in controlling the Type I error compare to some of theoretical test statistics as well as bootstrapping. The empirical rejection rate is evaluated from $k = 1, \ldots, 1,000$ replicated random samples from the true model. Theoretical test statistics covered in this simulation study are maximum likelihood, T_{ML}, Satorra-Bentler mean correction, T_{SB}, Satorra-Bentler mean correction and variance adjusted, T_{MAV}, and Asparouhov-Muthén implementation of mean and variance adjusted test, T_{AM}. The effect of Bartlett's correction factor and Swain's correction to the maximum likelihood and to the Satorra-Bentler scaled test statistics are also studied. $T_{MLb} = c_b T_{ML}$ and $T_{SBb} = c_b T_{SB}$ represent Bartlett corrections to the ML and SB tests respectively. Similarly T_{MLs} and T_{SBs} represent modified T_{ML} and T_{SB} with the Swain's correction factor c_s.

To draw a Bollen–Stine bootstrap sample, first we use the transformation (2.1) to make the covariance structure consistent with the hypothesized model. We then draw a sample with replacement with the same size as the original data and fit the true hypothesized model in the bootstrap sample and compute T_{ML}. Here, $B = 1,000$ bootstrap samples are drawn from each original data and thus we have 1,000 bootstrap test statistics to estimate the empirical distribution of the ML test statistic under the true model. Since not every bootstrap sample results in a converged solution we eliminate those non-converged samples and keep drawing bootstrap samples to achieve 1,000 samples. This bootstrap empirical distribution is used to determine p-value of each original T_{ML} statistics. This is accomplished by computing the proportion of the 1,000 bootstrap test statistics values that exceeded that of the original sample. T_{MLB} is the notation used for the bootstrap test.

For the Monte Carlo approximation of the test statistics for each sample we generate $M = 1,000$ independent samples $\mathbf{Z}_m$, $m = 1, 2, \cdots, M$ from a multivariate normal distribution with the same size and dimension as the original data and with the mean equal to zero and identity covariance matrix. Then we set $Y_m = Z_m \boldsymbol{\Sigma}^{\frac{1}{2}}(\hat{\boldsymbol{\theta}})$, where $\hat{\boldsymbol{\theta}}$ is the maximum likelihood estimator of parameters from the original sample. Therefore Y_m has covariance structure that is consistent with the hypothesized model (i.e., the null is true). Now we fit the hypothesized model to each of the Monte Carlo samples Y_m and compute the desired test statistic. By computing the proportion of the 1,000 Monte Carlo test statistics values that is greater than the original test statistic we determine the p-value of the test. In this simulation study we implement the proposed Monte Carlo method to estimate distribution of T_{ML}, T_{SB}, and T_{MVA}. T_{MLMC} is the notation used for the Monte Carlo approximation of the ML test, and T_{SBMC} is used for the Monte Carlo approximation of the SB test. Implementation of the Monte Carlo method to approximate the mean and variance adjusted test statistic, T_{MVA}, is not as straightforward as other test statistics. The asymptotic distribution of T_{ML} and T_{SB} when the null hypothesis is true under some regularity conditions (e.g. normality for T_{ML} or elliptical distributions assumption for T_{SB}) is a central $\boldsymbol{\chi}^2$ with the degrees of freedom that depends on the model and not the data. The asymptotic distribution of T_{MVA} is not constant from sample to sample. In this case we used the Asparouhov-Muthén version of scaled mean and variance adjusted test statistic, T_{AM}, since the degrees of freedom of asymptotic distribution is a constant value that depends only on sample size and the hypothesized model. Another way to overcome this problem is to use the cumulative distribution of each Monte Carlo test statistic based on its asymptotic $\boldsymbol{\chi}^2$ distribution with the degrees of freedom that is computed from the Monte Carlo sample. Let $T_{MVA}^{(m)}$ be the scaled mean and variance adjusted test statistic from m-th Monte Carlo normal sample that is computed from the equation (1.8) and d'_m is the adjusted degrees of freedom. The Distribution of $F_{MVA}^{(m)} = P(T_{MVA}^{(m)} > \boldsymbol{\chi}^2_{d'_m})$ does not depend on the sample and if the null is true, it is from a uniform (0, 1) distribution. For the original sample, similarly we compute $F_{MVA} = P(T_{MVA} > \boldsymbol{\chi}^2_{d'})$, where T_{MVA} and d' are the test statistic and the adjusted degrees of freedom of the sample. The p-value of the Monte Carlo test is determine by computing the proportion of the 1,000 Monte Carlo value $F_{MVA}^{(m)}$ values that is smaller than F_{MVA} of the original sample. Basically F_{MVA} and $F_{MVA}^{(m)}$

are theoretical p-values of the original and the Monte Carlo samples. Since the p-value of a test statistic is the probability of observing data that are more extreme against the null hypothesis, the proportion of Monte Carlo samples that have p-values (i.e. $F_{MVA}^{(m)}$) smaller than of the original sample is an empirical estimation of p-value of the test statistic. Table 3.1 summarizes test statistics that are covered in simulations of this chapter.

Similar to bootstrapping, in Monte Carlo normal samples we also get some non-converged samples that we eliminate to get 1,000 Monte Carlo samples. In our simulations non-convergence cases occurred mostly in small samples. The number of non-convergences to obtain 1,000 converged bootstrap samples got larger in the model B which has higher dimension. For example the average number of non-converged bootstrap samples in 1,000 replications of model A_1 with sample size equal to 50 was 0.524 compare to 0.366 for Monte Carlo samples. For model B_1 the average number of non-converged cases was 247.811 for bootstrap and 0.870 for Monte Carlo samples. When sample size increased to 100, the average number of non-converged cases at most was 1.518 for bootstrap and 0.830 for Monte Carlo for condition B_4. For sample size more than 100 the number of non-converged cases was almost zero for both bootstrap and Monte Carlo samples.

Table 3.1: List of test statistics used in the simulation study

T_{ML}	Maximum likelihood test statistic
T_{MLb}	Bartlett's corrected test
T_{MLs}	Swain's corrected test
T_{SB}	Satorra-Bentler scaled test
T_{SBb}	Satorra-Bentler scaled test with Bartlett's correction
T_{SBs}	Satorra-Bentler scaled test with Swain's correction
T_{MVA}	Satorra-Bentler mean and variance adjusted test
T_{AM}	Asparouhov-Muthen implementation of mean and variance adjusted test
T_{MLB}	Maximum likelihood bootstrap (Bollen-Stine)
T_{MLMC}	Monte Carlo re-sampling method for T_{ML}
T_{SBMC}	Monte Carlo re-sampling method for T_{SB}
T_{MVAMC}	Monte Carlo re-sampling method for T_{MVA}
T_{AMMC}	Monte Carlo re-sampling method for T_{AM}

Table 3.2 consists observed rejection percentages, R, for nominal Type I error $\alpha = 0.05$ for condition A_1. It also includes one sample Kolmogorov-Smirnov test statistics noted as D_{KS}. For those theoretical test statistics, we calculate D_{KS} as the supremum of the absolute difference between empirical distribution functions of the test statistic and cumulative distribution function of $\boldsymbol{\chi}^2_{df}$. In the model A_1 the degrees of freedom is 87. Since for the scaled mean and variance adjusted test statistic, T_{MVA}, degrees of freedom varies from sample to sample to determine D_{KS} statistic we use the p-values of the test, instead of empirical distribution function, and compare it with a uniform distribution. This is because the p-value of a test statistic and the value of the test statistic can be determined from each other and D_{KS} will be equivalent using either the statistic or the p-value. Here, the p-value of the test is $1 - F(T_{MVA})$ where F is the cumulative distribution function of $\boldsymbol{\chi}^2_{d'}$. In the bootstrap and Monte Carlo estimations, their empirical distributions from re-sampling also change for each replicated sample, therefore we use p-values to determine D_{KS}. The Kolmogorov-Smirnov test statistic is a number between 0 and 1 where 0 means the empirical distribution is exactly identical to the theoretical CDF, with the 95 percentile about 0.043 and the 99 percentile 0.051 for sample size equal to 1000 (i.e. number of replicated samples in here).

The distribution of data in condition A_1 is normal. When the sample size is small, T_{ML} performs poorly and over rejects the correct model. T_{SB} performs even worse when sample size is as small as 50. T_{MVA} and the Asparouhov-Muthén version of it perform very similar to each other for every condition and sample sizes and we only mention results of one of them throughout this chapter. T_{MVA} performs better than T_{ML} and T_{SB} on the tail but its performance of the overall distribution function for small samples is as bad as the other two tests. Rejection rates for those three test statistics are acceptable for moderate and large sample sizes. They eventually converge for sample size as large as 2000 and KS distance also gets within the acceptable range.

The Bartlett correction to the ML test improves on T_{ML} for small samples. For example the rejection rate for T_{ML} for sample size 50 is 29.8% and the KS test statistic is 0.376. The

Table 3.2: Performance of different test statistics for model A, condition 1

		Sample size						
		n = 50	n = 100	n = 150	n = 300	n = 500	n = 2000	n = 5000
T_{ML}	R	29.8	12.3	8.1	6.9	5.9	4.6	5.2
	D_{KS}	0.376	0.173	0.137	0.06	0.043	0.021	0.029
T_{MLb}	R	3.9	3.2	3.6	4.4	5.2	4.3	5.0
	D_{KS}	0.08	0.059	0.036	0.036	0.02	0.021	0.025
T_{MLs}	R	7.2	4.6	4.6	4.6	5.4	4.3	5.1
	D_{KS}	0.068	0.013	0.028	0.021	0.017	0.019	0.026
T_{SB}	R	43.7	16.7	11.9	7.6	6.6	4.8	5.0
	D_{KS}	0.507	0.249	0.196	0.085	0.055	0.025	0.032
T_{SBb}	R	9.3	5.0	5.2	5.3	5.2	4.6	4.8
	D_{KS}	0.11	0.037	0.057	0.024	0.023	0.017	0.028
T_{SBs}	R	16.3	7.0	5.9	5.8	5.3	4.7	4.9
	D_{KS}	0.203	0.085	0.089	0.036	0.029	0.017	0.029
T_{MVA}	R	8.3	2.8	3.7	3.9	4.8	4.3	4.8
	D_{KS}	0.436	0.262	0.203	0.117	0.073	0.027	0.034
T_{AM}	R	9.9	3.4	4.0	4.0	4.8	4.3	4.8
	D_{KS}	0.418	0.253	0.196	0.114	0.071	0.026	0.033
T_{MLB}	R	0.0	1.0	1.4	2.9	4.3	4.5	4.7
	D_{KS}	0.19	0.092	0.08	0.057	0.038	0.018	0.029
T_{MLMC}	R	5.6	4.5	4.3	4.9	5.3	4.5	4.7
	D_{KS}	0.027	0.027	0.018	0.032	0.019	0.02	0.028
T_{SBMC}	R	6.3	3.8	4.4	4.5	5.3	4.8	4.6
	D_{KS}	0.031	0.029	0.026	0.031	0.017	0.02	0.026
T_{MVAMC}	R	6.1	3.8	4.6	4.6	5.2	4.8	4.6
	D_{KS}	0.032	0.031	0.026	0.03	0.017	0.02	0.026
T_{AMMC}	R	6.1	3.8	4.6	4.6	5.2	4.8	4.6
	D_{KS}	0.033	0.031	0.026	0.03	0.017	0.02	0.026

R is observed rejection percentages for $\alpha = 5\%$

D_{KS} is Kolmogorov-Smirnov distance statistic

Table 3.3: Performance of different test statistics for model B, condition 1

		Sample size						
		n = 50	n = 100	n = 150	n = 300	n = 500	n = 2000	n = 5000
T_{ML}	R	99.5	56.7	36.1	14.6	10.6	5.5	6.0
	D_{KS}	0.972	0.642	0.443	0.255	0.158	0.06	0.043
T_{MLb}	R	5.2	2.9	4.4	5.1	5.5	4.8	5.6
	D_{KS}	0.032	0.078	0.049	0.046	0.019	0.034	0.029
T_{MLs}	R	17.6	6.2	6.8	5.7	6.2	5.2	5.6
	D_{KS}	0.254	0.054	0.059	0.043	0.035	0.032	0.03
T_{SB}	R	100.0	73.7	46.0	17.0	12.3	6.0	6.0
	D_{KS}	0.992	0.742	0.528	0.306	0.189	0.063	0.045
T_{SBb}	R	23.7	7.3	8.0	6.6	6.6	5.2	5.5
	D_{KS}	0.334	0.113	0.095	0.062	0.053	0.032	0.031
T_{SBs}	R	46.8	13.0	11.5	7.9	6.9	5.3	5.5
	D_{KS}	0.539	0.215	0.164	0.094	0.07	0.036	0.033
T_{MVA}	R	57.9	4.8	3.9	3.4	2.9	4.0	5.0
	D_{KS}	0.835	0.572	0.441	0.286	0.216	0.056	0.045
T_{AM}	R	65.4	5.9	4.8	3.5	3.1	4.0	5.0
	D_{KS}	0.836	0.561	0.431	0.281	0.212	0.056	0.044
T_{MLB}	R	0.0	0.0	0.0	0.8	1.4	3.3	4.6
	D_{KS}	0.764	0.321	0.179	0.118	0.104	0.037	0.035
T_{MLMC}	R	4.6	4.1	5.7	5.6	6.0	4.7	5.5
	D_{KS}	0.032	0.034	0.031	0.027	0.029	0.037	0.028
T_{SBMC}	R	4.1	3.9	5.3	5.4	5.9	4.9	5.6
	D_{KS}	0.031	0.029	0.03	0.025	0.033	0.034	0.029
T_{MVAMC}	R	4.8	3.6	4.9	5.4	5.9	4.9	5.6
	D_{KS}	0.028	0.029	0.031	0.024	0.034	0.033	0.028
T_{AMMC}	R	4.7	3.6	4.9	5.4	5.9	4.9	5.6
	D_{KS}	0.029	0.029	0.031	0.024	0.033	0.033	0.028

R is observed rejection percentages for $\alpha = 5\%$

D_{KS} is Kolmogorov-Smirnov distance statistic

Bartlett correction adjusts the rejection rate to 3.9% and the KS distance drops to 0.08. The rejection rate for Swain's correction test in this case is 7.2% and its whole distribution performance, which is evaluated by KS test, is equal to 0.068. Applying Bartlett and Swain's correction to T_{SB} also improves its performance, and the rejection rate for sample size 50 decreases from 43.7% to 9.3% for Bartlett's correction and to 16.3% for Swain's correction. We did not apply those type of corrections to T_{MVA} because as we discuss later this test under rejects the correct model when the data are not from a multivariate normal distribution and using those corrections makes it even worse.

The bootstrap method in condition A_1 tends to under-reject the correct model until the sample size exceeds 300 cases. The overall performance of the bootstrap previously has not been studied in literature and, as we noted earlier, we measure it by the KS distance test statistic for p-values and uniform (0, 1). In condition A_1, the KS distance statistics for bootstrap are getting to reasonable values for the sample size equal or greater than 500.

All four Monte Carlo tests have equivalently good performance among other test statistics in this condition. This is expected because condition A_1 has exactly multivariate normal distribution and the Monte Carlo method uses sampling from a normal distribution. The only difference is that each Monte Carlo sample uses the estimated parameters and not the true population parameters. In section 3.4 we will study the effect of values of parameters to goodness of fit test statistics and we will see that the value of parameters has negligible effect on the test statistic.

Table 3.3 reports on the performance of test statistics for condition B_1 which has a similar distribution condition for a model with the dimension equal to 30 and degrees of freedom equal to 402. We observe that the larger model condition has a big effect on the theoretical test statistics for small samples. When the sample size is small (i.e. 50, 100, 150), T_{ML} totally breaks down and also T_{SB} does not show any better results. The rejection rate for T_{MVA} interestingly recovers when sample size is 100 and jumps down from 57.9% to 4.8%. This might be misleading but when we look at the overall distribution of the test we observe that the KS distance does not recover until the sample size reaches 2000. The Bartlett and Swain's corrections in this case give a big improvement to the results in both rejection rate and KS distance. T_{MLb} has reasonable results in all sample sizes. Similarly, the result of Bartlett's correction is better than Swain's correction. Note that the model studied here is based on the model defined in the equation (3.1) and in more general SEM, there might be situations that Swain's correction out performs Bartlett's corrected test. However, an extended comparison of those two correction methods is not of interest in this study.

The most interesting finding in the Table 3.3 is that the bootstrap method for small samples when the dimension of the model is large fails even when the data is from a normal distribution. For sample size equal to 50 we do not reject any of the replicated data and the KS distance is 0.764. Even when sample size is 500 the rejection rate is 1.4 and the KS distance is 0.104. For sample size 2000 and 5000 the bootstrap method gives acceptable results in this condition. The results of Monte Carlo tests stay acceptable for all sample sizes and a higher model dimension does not seems to have any effect in the normal case.

Tables 3.4 and 3.5 contains the results of the condition of elliptically distributed data (i.e. condition A_2 and B_2). In this condition, the asymptotically robustness of T_{ML} is not valid and we expect that T_{ML} will fail even for large sample sizes. On the other hand T_{SB} gives promising results for moderate to large sample sizes. For condition A_2, T_{SB} recovers completely for sample size equal and greater than 500 as both rejection rates and KS distance statistic are acceptable. For smaller sample sizes, Bartlett and Swain's correction improve the performance of T_{SB}. For sample size equal to 50, the rejection rate for T_{SBb} is 6.6% and the KS distance is 0.152. T_{MVA} in this condition also under-rejects the correct model and it does not converge for sample size as large as 5000. The bootstrap method shows some good results at the tail. The rejection percentage for bootstrap is acceptable in A_2 for sample size equal to 50, 2000, and 5000 but it over rejects for sample sizes 100 to 500 and it gets worse for higher dimension model B. The overall distribution of p-values for the bootstrap does not perform well. The minimum KS distance for bootstrap in condition A_2 among all sample sizes is 0.194 for sample size equal to 2000, and for model B_2 is 0.374. This results puts a flag on using the bootstrap for higher dimensional models.

As can be seen in Table 3.4, because of problems with the asymptotic robustness of T_{ML}, applying the Monte Carlo method for the ML test statistic does not work in this condition. But the results of T_{SBMC} are acceptable for model and condition A_2 for all sample sizes. Using the Monte Carlo approach we get the rejection rate equal to 3.8 and KS distance reduces to 0.06. For model and condition B_2, shown in Table 3.5, the Monte Carlo approach gives a better result for very small sample sizes but the Bartlett corrected test outperforms the Monte Carlo method. This is in fact the only situation that the Monte Carlo method is outperformed by any other test we have covered in this study. An explanation is that the Monte Carlo method uses a normal distribution to estimate T_{SB}. In higher dimensional models, when the data are from an elliptical distribution, the distribution of T_{SB} does not converge fast enough to its counterpart when the data is from a multivariate normal distribution. Eventually, in both cases, the SB scaled test statistic converges to a $\boldsymbol{\chi}^2$ distribution but the sample size needs to be very large. This has not mentioned in previous studies (e.g. Yuan & Bentler 1998) since their study was limited to dimension as largest as 15 and as we see here, this problem is not detectable for in model A_2. The Monte Carlo approach to T_{MVA} improves the performance of T_{MVA} test for small samples but it still suffers from the over-rejection of the true model.

The rest of the results are included in Tables 3.6 to 3.13 as follows: Tables 3.6 and 3.7 report on results of condition 3 for model A and B. Tables 3.8 and 3.9 are results for condition 4. Tables 3.10 and 3.11 are results for condition 5 that uses the second data generation method. Finally, Tables 3.12 and 3.13 contain results for condition 6. In conditions 3 to 6 data are from variables with nonzero skewness and kurtosis. The result in these tables are very similar to condition A_2 and B_2 respectively. The only noticeable difference is that the Bartlett corrected test statistic, T_{SBb}, perform worse and T_{SBMC} performs better than those in condition 2 in both 15 and 30 dimensional data and for all sample sizes. In fact T_{SBMC}

Table 3.4: Performance of different test statistics for model A, condition 2

		Sample size						
		n = 50	n = 100	n = 150	n = 300	n = 500	n = 2000	n = 5000
T_{ML}	R	95.3	96.0	96.7	98.2	98.7	99.8	100.0
	D_{KS}	0.91	0.917	0.924	0.958	0.964	0.988	0.997
T_{MLb}	R	78.2	90.9	94.2	98.1	98.7	99.8	100.0
	D_{KS}	0.75	0.865	0.895	0.95	0.96	0.987	0.997
T_{MLs}	R	83.2	92.6	94.5	98.1	98.7	99.8	100.0
	D_{KS}	0.802	0.879	0.902	0.952	0.961	0.987	0.997
T_{SB}	R	45.0	13.3	9.8	5.8	4.0	3.4	4.4
	D_{KS}	0.567	0.241	0.174	0.083	0.046	0.055	0.031
T_{SBb}	R	6.6	4.6	3.7	3.5	3.3	3.0	4.4
	D_{KS}	0.152	0.049	0.049	0.027	0.062	0.065	0.028
T_{SBs}	R	13.0	5.3	4.8	4.1	3.4	3.0	4.4
	D_{KS}	0.259	0.082	0.075	0.037	0.054	0.062	0.029
T_{MVA}	R	1.0	0.4	0.0	0.4	0.0	1.0	0.7
	D_{KS}	0.515	0.384	0.337	0.275	0.23	0.143	0.163
T_{AM}	R	1.1	0.6	0.1	0.6	0.0	1.0	0.7
	D_{KS}	0.471	0.351	0.305	0.255	0.218	0.149	0.153
T_{MLB}	R	2.9	10.3	12.5	12.1	10.2	4.8	2.7
	D_{KS}	0.292	0.365	0.378	0.354	0.334	0.194	0.218
T_{MLMC}	R	80.8	92.1	94.4	98.1	98.7	99.7	100
	D_{KS}	0.78	0.875	0.899	0.95	0.961	0.986	0.996
T_{SBMC}	R	3.8	3.2	3.3	3.0	2.9	3.1	4.2
	D_{KS}	0.06	0.061	0.045	0.034	0.072	0.066	0.03
T_{MVAMC}	R	0.4	0.5	0.2	0.5	0.0	1.0	0.7
	D_{KS}	0.225	0.228	0.194	0.169	0.187	0.14	0.15
T_{AMMC}	R	0.4	0.6	0.3	0.6	0.0	1.0	0.9
	D_{KS}	0.241	0.254	0.21	0.178	0.206	0.15	0.143

R is observed rejection percentages for $\alpha = 5\%$

D_{KS} is Kolmogorov-Smirnov distance statistic

Table 3.5: Performance of different test statistics for model B, condition 2

		Sample size						
		n = 50	n = 100	n = 150	n = 300	n = 500	n = 2000	n = 5000
T_{ML}	R	100	100	100	100	100	100	100
	D_{KS}	1	1	1	1	1	1	1
T_{MLb}	R	98.9	100	100	100	100	100	100
	D_{KS}	0.955	0.988	0.997	1	1	1	1
T_{MLs}	R	99.7	100	100	100	100	100	100
	D_{KS}	0.979	0.991	0.998	1	1	1	1
T_{SB}	R	96.1	39.3	14.5	3.9	1.7	2.6	2.5
	D_{KS}	0.922	0.461	0.227	0.077	0.141	0.153	0.11
T_{SBb}	R	4.7	0.7	0.6	0.7	0.9	2.4	2.3
	D_{KS}	0.097	0.248	0.279	0.297	0.286	0.185	0.125
T_{SBs}	R	14.6	1.8	0.8	0.9	1.0	2.5	2.4
	D_{KS}	0.183	0.153	0.216	0.266	0.268	0.181	0.123
T_{MVA}	R	1.1	0.0	0.0	0.0	0.0	0.1	0.0
	D_{KS}	0.617	0.487	0.442	0.375	0.323	0.261	0.221
T_{AM}	R	2.0	0.0	0.0	0.0	0.0	0.1	0.0
	D_{KS}	0.549	0.413	0.386	0.345	0.304	0.262	0.229
T_{MLB}	R	0.0	11.9	24.3	31.3	26.3	8.2	2.6
	D_{KS}	0.433	0.513	0.601	0.618	0.578	0.456	0.374
T_{MLMC}	R	98.7	100.0	100.0	100.0	100.0	100.0	100.0
	D_{KS}	0.95	0.989	0.998	1.0	1.0	1.0	1.0
T_{SBMC}	R	1.0	0.1	0.5	0.5	1.1	2.2	2.4
	D_{KS}	0.397	0.355	0.354	0.329	0.305	0.192	0.126
T_{MVAMC}	R	0.0	0.0	0.0	0.0	0.0	0.1	0.1
	D_{KS}	0.804	0.651	0.562	0.441	0.384	0.27	0.232
T_{AMMC}	R	0.0	0.0	0.0	0.0	0.0	0.1	0.1
	D_{KS}	0.799	0.674	0.593	0.471	0.406	0.291	0.243

R is observed rejection percentages for $\alpha = 5\%$

D_{KS} is Kolmogorov-Smirnov distance statistic

has the best performance for all other conditions and in most cases it gives very reasonable results for sample sizes as small as 50.

In summary, T_{SBMC} has the best performance among all conditions and models and gives promising results at small sample sizes. The minimum rejection rate for model A among all conditions is 2.9 and the maximum is 11.2. The largest KS distance is 0.257 which is for condition 6. The minimum rejection rate for model B among all condition is 0.1 and the maximum is 9.4. The largest KS distance is 0.397 which is in condition 2 and if we exclude this condition, the maximum KS distance is 0.207 and the rejection rates range between 1.5 and 9.4. Our results of those theoretical methods and bootstrapping agree with results in previous studies (see e.g. Moshagen 2012, Nevitt & Hancoock 2001 and 2004, Ichikawa & Konishi 1995, Jiang & Yuan 2017). It should also be noted that we study the performance of these test statistics in higher dimensional models that make the problems with Type I rejection rate more noticeable. For example, from this simulation study, we observe that the bootstrap method cannot be trusted for models with many variables even for large sample sizes. In the next section we take a closer look at the bootstrap method and we try to study the root of the problem with bootstrapping in SEM.

Table 3.6: Performance of different test statistics for model A, condition 3

		Sample size						
		n = 50	n = 100	n = 150	n = 300	n = 500	n = 2000	n = 5000
T_{ML}	R	83.1	79.5	81.0	89.8	91.3	98.1	99.7
	D_{KS}	0.793	0.758	0.779	0.849	0.867	0.947	0.979
T_{MLb}	R	58.2	68.8	75.9	87.3	90.4	98.1	99.7
	D_{KS}	0.574	0.661	0.716	0.827	0.858	0.945	0.978
T_{MLs}	R	64.8	71.3	77.3	87.8	90.4	98.1	99.7
	D_{KS}	0.634	0.685	0.728	0.833	0.86	0.946	0.978
T_{SB}	R	56.5	18.8	10.4	7.0	5.5	3.8	4.2
	D_{KS}	0.664	0.353	0.244	0.14	0.084	0.033	0.038
T_{SBb}	R	11.8	6.5	4.8	4.8	4.6	3.7	3.9
	D_{KS}	0.266	0.139	0.109	0.07	0.05	0.028	0.034
T_{SBs}	R	18.6	8.3	5.9	5.2	4.9	3.7	3.9
	D_{KS}	0.374	0.188	0.138	0.086	0.056	0.027	0.035
T_{MVA}	R	0.5	0.1	0.0	0.1	0.1	0.0	0.6
	D_{KS}	0.552	0.421	0.386	0.319	0.286	0.224	0.202
T_{AM}	R	0.9	0.1	0.1	0.1	0.2	0.0	0.6
	D_{KS}	0.499	0.385	0.358	0.299	0.267	0.213	0.193
T_{MLB}	R	0.7	2.8	4.0	4.9	6.5	3.7	2.1
	D_{KS}	0.227	0.289	0.317	0.335	0.313	0.275	0.24
T_{MLMC}	R	62.6	70.2	76.7	87.4	90.3	97.8	99.7
	D_{KS}	0.609	0.676	0.726	0.828	0.855	0.944	0.977
T_{SBMC}	R	6.4	5.0	4.0	4.5	4.1	3.6	3.9
	D_{KS}	0.172	0.094	0.083	0.059	0.042	0.029	0.033
T_{MVAMC}	R	0.3	0.1	0.2	0.1	0.1	0.1	0.4
	D_{KS}	0.214	0.219	0.213	0.216	0.22	0.204	0.194
T_{AMMC}	R	0.3	0.1	0.2	0.1	0.2	0.1	0.5
	D_{KS}	0.228	0.244	0.24	0.207	0.206	0.19	0.186

R is observed rejection percentages for $\alpha = 5\%$

D_{KS} is Kolmogorov-Smirnov distance statistic

Table 3.7: Performance of different test statistics for model B, condition 3

		Sample size						
		n = 50	n = 100	n = 150	n = 300	n = 500	n = 2000	n = 5000
T_{ML}	R	100.0	99.8	99.7	99.8	99.9	100.0	100.0
	D_{KS}	1	0.987	0.985	0.992	0.998	1	1
T_{MLb}	R	91.0	96.2	98.3	99.6	99.9	100	100
	D_{KS}	0.861	0.921	0.954	0.985	0.996	1	1
T_{MLs}	R	96.1	97.9	98.4	99.6	99.9	100.0	100.0
	D_{KS}	0.921	0.938	0.962	0.987	0.997	1	1
T_{SB}	R	98.9	70.3	38.3	11.0	4.8	3.3	2.9
	D_{KS}	0.976	0.721	0.522	0.224	0.118	0.054	0.053
T_{SBb}	R	20.5	6.5	3.9	3.1	1.7	2.6	2.7
	D_{KS}	0.305	0.086	0.052	0.095	0.097	0.087	0.066
T_{SBs}	R	41.4	10.9	6.2	3.5	2.0	2.6	2.7
	D_{KS}	0.53	0.19	0.117	0.066	0.08	0.083	0.064
T_{MVA}	R	2.5	0.0	0.0	0.0	0.0	0.0	0.0
	D_{KS}	0.681	0.547	0.496	0.43	0.392	0.314	0.265
T_{AM}	R	4.2	0.0	0.0	0.0	0.0	0.0	0.0
	D_{KS}	0.63	0.481	0.435	0.387	0.358	0.298	0.256
T_{MLB}	R	0.0	2.4	6.5	13.3	11.5	5.2	2.1
	D_{KS}	0.505	0.368	0.465	0.546	0.551	0.47	0.413
T_{MLMC}	R	89.4	97	98.4	99.7	99.9	100	100
	D_{KS}	0.852	0.927	0.96	0.987	0.996	1	1
T_{SBMC}	R	4.0	2.7	2.5	2.6	1.5	2.4	2.4
	D_{KS}	0.065	0.072	0.072	0.129	0.113	0.094	0.071
T_{MVAMC}	R	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	D_{KS}	0.709	0.544	0.48	0.377	0.342	0.263	0.247
T_{AMMC}	R	0.0	0.0	0.1	0.0	0.0	0.0	0.0
	D_{KS}	0.698	0.57	0.514	0.414	0.37	0.279	0.26

R is observed rejection percentages for $\alpha = 5\%$

D_{KS} is Kolmogorov-Smirnov distance statistic

Table 3.8: Performance of different test statistics for model A, condition 4

		Sample size						
		n = 50	n = 100	n = 150	n = 300	n = 500	n = 2000	n = 5000
T_{ML}	R	81.0	79.8	82.0	86.6	89.1	97.1	99.2
	D_{KS}	0.767	0.756	0.778	0.819	0.843	0.933	0.969
T_{MLb}	R	55.6	70.2	75.9	84.1	88.0	97.1	99.2
	D_{KS}	0.546	0.672	0.721	0.797	0.831	0.931	0.969
T_{MLs}	R	62.1	72.6	77.2	84.5	88.2	97.1	99.2
	D_{KS}	0.605	0.693	0.736	0.802	0.834	0.932	0.969
T_{SB}	R	55.6	19.3	12.0	5.8	3.9	4.9	4.9
	D_{KS}	0.635	0.372	0.278	0.158	0.096	0.05	0.051
T_{SBb}	R	11.4	6.9	5.9	4.1	2.9	4.7	4.9
	D_{KS}	0.239	0.158	0.137	0.09	0.056	0.041	0.048
T_{SBs}	R	19.1	9.4	7.1	4.5	3.0	4.7	4.9
	D_{KS}	0.341	0.207	0.169	0.105	0.065	0.043	0.049
T_{MVA}	R	0.5	0.1	0.1	0.1	0.1	0.2	0.3
	D_{KS}	0.554	0.431	0.391	0.333	0.294	0.234	0.209
T_{AM}	R	0.8	0.1	0.1	0.2	0.2	0.2	0.3
	D_{KS}	0.5	0.389	0.358	0.309	0.269	0.216	0.195
T_{MLB}	R	0.9	3.0	5.0	5.3	5.3	3.8	2.1
	D_{KS}	0.229	0.31	0.344	0.352	0.339	0.303	0.267
T_{MLMC}	R	59.6	72.0	77.0	84.6	88.4	97.1	99.2
	D_{KS}	0.58	0.684	0.731	0.801	0.835	0.932	0.967
T_{SBMC}	R	7.2	5.3	5.1	3.9	2.4	4.4	4.7
	D_{KS}	0.147	0.119	0.106	0.082	0.051	0.038	0.047
T_{MVAMC}	R	0.4	0.1	0.1	0.2	0.2	0.3	0.2
	D_{KS}	0.219	0.212	0.205	0.235	0.228	0.216	0.203
T_{AMMC}	R	0.5	0.1	0.1	0.3	0.2	0.3	0.2
	D_{KS}	0.238	0.233	0.216	0.22	0.224	0.199	0.192

R is observed rejection percentages for $\alpha = 5\%$

D_{KS} is Kolmogorov-Smirnov distance statistic

Table 3.9: Performance of different test statistics for model B, condition 4

		Sample size						
		n = 50	n = 100	n = 150	n = 300	n = 500	n = 2000	n = 5000
T_{ML}	R	100.0	100.0	99.9	100.0	99.9	100.0	100.0
	D_{KS}	0.997	0.992	0.987	0.993	0.993	1	1
T_{MLb}	R	90.9	96.2	98.6	99.9	99.9	100.0	100.0
	D_{KS}	0.862	0.918	0.954	0.985	0.99	1	1
T_{MLs}	R	94.9	96.9	98.8	99.9	99.9	100.0	100.0
	D_{KS}	0.912	0.937	0.963	0.986	0.991	1	1
T_{SB}	R	99.2	69.9	35.8	12.0	4.6	4.0	2.5
	D_{KS}	0.975	0.715	0.504	0.236	0.062	0.05	0.068
T_{SBb}	R	22.1	5.5	3.8	2.6	2.1	3.2	2.1
	D_{KS}	0.333	0.099	0.03	0.064	0.112	0.085	0.081
T_{SBs}	R	44.7	9.6	5.7	3.4	2.4	3.2	2.1
	D_{KS}	0.534	0.204	0.09	0.038	0.092	0.081	0.079
T_{MVA}	R	2.2	0.0	0.0	0.0	0.0	0.0	0.0
	D_{KS}	0.678	0.546	0.506	0.433	0.379	0.312	0.277
T_{AM}	R	5.0	0.0	0.0	0.0	0.0	0.0	0.0
	D_{KS}	0.632	0.484	0.441	0.388	0.35	0.298	0.264
T_{MLB}	R	0.0	2.8	8.5	11.4	13.4	7	2.7
	D_{KS}	0.515	0.362	0.466	0.53	0.537	0.49	0.417
T_{MLMC}	R	89.4	96.5	98.5	99.9	99.9	100	100
	D_{KS}	0.852	0.928	0.959	0.984	0.991	1	1
T_{SBMC}	R	3.9	3.1	2.4	2.4	1.5	3.3	2.3
	D_{KS}	0.037	0.058	0.094	0.09	0.13	0.09	0.078
T_{MVAMC}	R	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	D_{KS}	0.7	0.564	0.478	0.376	0.344	0.263	0.254
T_{AMMC}	R	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	D_{KS}	0.686	0.584	0.51	0.407	0.374	0.284	0.269

R is observed rejection percentages for $\alpha = 5\%$

D_{KS} is Kolmogorov-Smirnov distance statistic

Table 3.10: Performance of different test statistics for model A, condition 5

		Sample size						
		n = 50	n = 100	n = 150	n = 300	n = 500	n = 2000	n = 5000
T_{ML}	R	38.1	22.4	21.6	16.4	14.3	12.3	9.0
	D_{KS}	0.395	0.213	0.194	0.147	0.095	0.096	0.09
T_{MLb}	R	12.7	13.2	14.6	13.1	12.7	12.0	8.8
	D_{KS}	0.091	0.104	0.119	0.105	0.082	0.087	0.087
T_{MLs}	R	17.5	14.9	16.5	13.6	13.1	12.1	8.9
	D_{KS}	0.153	0.114	0.132	0.114	0.085	0.089	0.088
T_{SB}	R	57.1	19.4	16.5	9.4	5.0	6.3	5.0
	D_{KS}	0.644	0.348	0.287	0.149	0.105	0.056	0.019
T_{SBb}	R	15.3	8.2	8.1	6.3	3.8	5.5	5.0
	D_{KS}	0.249	0.138	0.146	0.086	0.065	0.046	0.02
T_{SBs}	R	22.2	9.9	9.5	7.2	4.0	5.7	5.0
	D_{KS}	0.347	0.188	0.178	0.099	0.074	0.049	0.019
T_{MVA}	R	1.3	0.4	0.2	0.1	0.4	1.6	2.4
	D_{KS}	0.541	0.385	0.346	0.265	0.217	0.13	0.09
T_{AM}	R	2.0	0.7	0.3	0.3	0.5	2.1	2.4
	D_{KS}	0.51	0.36	0.326	0.249	0.207	0.126	0.087
T_{MLB}	R	0.0	0.1	0.3	1.0	1.7	2.9	3.0
	D_{KS}	0.223	0.172	0.202	0.196	0.173	0.133	0.09
T_{MLMC}	R	15.3	14.4	15.8	13.8	13.5	11.8	8.8
	D_{KS}	0.119	0.104	0.129	0.113	0.086	0.09	0.088
T_{SBMC}	R	8.6	6.9	7.0	5.8	3.7	5.3	5.0
	D_{KS}	0.156	0.095	0.114	0.073	0.052	0.042	0.018
T_{MVAMC}	R	0.6	0.5	0.3	0.3	0.5	1.6	2.2
	D_{KS}	0.11	0.142	0.155	0.165	0.152	0.115	0.079
T_{AMMC}	R	0.6	0.6	0.4	0.3	0.5	1.8	2.2
	D_{KS}	0.114	0.152	0.141	0.15	0.144	0.109	0.077

R is observed rejection percentages for $\alpha = 5\%$

D_{KS} is Kolmogorov-Smirnov distance statistic

Table 3.11: Performance of different test statistics for model B, condition 5

		Sample size						
		n = 50	n = 100	n = 150	n = 300	n = 500	n = 2000	n = 5000
T_{ML}	R	99.3	67.3	45.2	24.6	22.5	12.4	10.8
	D_{KS}	0.965	0.671	0.471	0.25	0.201	0.082	0.108
T_{MLb}	R	20.5	15.2	15.8	13.5	15.4	11.3	10.3
	D_{KS}	0.212	0.159	0.121	0.099	0.122	0.066	0.099
T_{MLs}	R	36.9	22.0	18.0	14.4	16.1	11.5	10.3
	D_{KS}	0.389	0.224	0.161	0.113	0.131	0.068	0.1
T_{SB}	R	100.0	88.3	55.1	18.3	10.6	5.7	4.9
	D_{KS}	0.999	0.844	0.637	0.333	0.22	0.063	0.033
T_{SBb}	R	41.7	16.1	9.4	5.9	6.3	5.1	4.9
	D_{KS}	0.536	0.296	0.203	0.09	0.074	0.029	0.019
T_{SBs}	R	67.1	23.3	13.4	6.8	6.7	5.2	4.9
	D_{KS}	0.713	0.402	0.271	0.122	0.093	0.033	0.021
T_{MVA}	R	28.2	0.1	0.2	0	0.1	0.1	0.5
	D_{KS}	0.811	0.611	0.502	0.373	0.32	0.186	0.122
T_{AM}	R	37.8	0.2	0.2	0.0	0.1	0.1	0.7
	D_{KS}	0.808	0.589	0.476	0.357	0.305	0.182	0.121
T_{MLB}	R	0.0	0.0	0.0	0.3	0.7	1.1	1.7
	D_{KS}	0.737	0.278	0.218	0.247	0.263	0.191	0.135
T_{MLMC}	R	18.7	17.9	17.1	13.4	16.6	11.3	10.3
	D_{KS}	0.195	0.184	0.142	0.107	0.133	0.068	0.102
T_{SBMC}	R	9.3	9.4	6.5	4.9	5.6	4.8	4.9
	D_{KS}	0.207	0.174	0.127	0.054	0.054	0.027	0.014
T_{MVAMC}	R	0.2	0.1	0.2	0.1	0.1	0.4	0.7
	D_{KS}	0.334	0.26	0.244	0.229	0.205	0.129	0.1
T_{AMMC}	R	0.4	0.1	0.2	0.1	0.1	0.4	0.7
	D_{KS}	0.298	0.266	0.251	0.239	0.218	0.134	0.098

R is observed rejection percentages for $\alpha = 5\%$

D_{KS} is Kolmogorov-Smirnov distance statistic

Table 3.12: Performance of different test statistics for model A, condition 6

		Sample size						
		n = 50	n = 100	n = 150	n = 300	n = 500	n = 2000	n = 5000
T_{ML}	R	84.0	78.2	85.0	86.4	90.1	98.0	99.2
	D_{KS}	0.797	0.744	0.802	0.827	0.856	0.95	0.971
T_{MLb}	R	60.0	67.0	79.0	84.2	89.4	97.8	99.2
	D_{KS}	0.575	0.642	0.752	0.8	0.846	0.948	0.97
T_{MLs}	R	65.9	70.0	81.0	85.1	89.4	97.8	99.2
	D_{KS}	0.634	0.667	0.762	0.807	0.848	0.949	0.97
T_{SB}	R	67.5	26.6	18.8	9.9	6.2	5.9	5.2
	D_{KS}	0.716	0.433	0.357	0.229	0.132	0.098	0.057
T_{SBb}	R	17.5	10.2	9.2	6.3	4.8	5.6	4.9
	D_{KS}	0.349	0.223	0.216	0.161	0.092	0.088	0.053
T_{SBs}	R	26.7	12.7	10.5	6.9	5.0	5.6	4.9
	D_{KS}	0.453	0.273	0.248	0.176	0.101	0.091	0.054
T_{MVA}	R	1.3	0.2	0.1	0.1	0.0	0.1	0.6
	D_{KS}	0.583	0.451	0.404	0.358	0.293	0.254	0.204
T_{AM}	R	2.0	0.4	0.1	0.1	0.1	0.2	0.6
	D_{KS}	0.541	0.412	0.372	0.329	0.276	0.239	0.191
T_{MLB}	R	0.7	1.9	4.6	5.0	4.6	3.4	2.4
	D_{KS}	0.212	0.28	0.32	0.35	0.311	0.299	0.262
T_{MLMC}	R	63.5	69.0	80.6	84.9	89.4	97.8	99.3
	D_{KS}	0.612	0.657	0.759	0.802	0.848	0.949	0.97
T_{SBMC}	R	11.2	8.7	7.9	6.0	4.2	5.5	5.1
	D_{KS}	0.257	0.178	0.187	0.148	0.083	0.086	0.053
T_{MVAMC}	R	0.7	0.3	0.1	0.2	0.1	0.1	0.5
	D_{KS}	0.16	0.188	0.229	0.256	0.232	0.232	0.197
T_{AMMC}	R	1.0	0.3	0.2	0.3	0.1	0.1	0.7
	D_{KS}	0.164	0.174	0.195	0.23	0.212	0.218	0.184

R is observed rejection percentages for $\alpha = 5\%$

D_{KS} is Kolmogorov-Smirnov distance statistic

Table 3.13: Performance of different test statistics for model B, condition 6

		Sample size						
		n = 50	n = 100	n = 150	n = 300	n = 500	n = 2000	n = 5000
T_{ML}	R	100.0	99.7	99.7	99.7	100.0	100.0	100.0
	D_{KS}	1	0.983	0.987	0.986	0.998	1	1
T_{MLb}	R	89.4	96.3	98.3	99.3	100.0	100.0	100.0
	D_{KS}	0.846	0.921	0.958	0.978	0.996	1	1
T_{MLs}	R	95.1	97	98.6	99.4	100.0	100.0	100.0
	D_{KS}	0.904	0.938	0.963	0.979	0.997	1	1
T_{SB}	R	99.5	80.7	48.6	18.8	8.0	4.0	3.4
	D_{KS}	0.987	0.787	0.598	0.322	0.15	0.027	0.026
T_{SBb}	R	32.9	12.7	6.5	5.3	3.3	3.1	3.3
	D_{KS}	0.439	0.221	0.141	0.079	0.034	0.059	0.039
T_{SBs}	R	57.7	18.6	9.9	6.7	3.6	3.2	3.3
	D_{KS}	0.638	0.324	0.209	0.112	0.035	0.055	0.038
T_{MVA}	R	5.1	0.1	0.0	0.0	0.0	0.0	0.0
	D_{KS}	0.71	0.565	0.51	0.435	0.393	0.313	0.27
T_{AM}	R	9.0	0.1	0.0	0.0	0.0	0.0	0.0
	D_{KS}	0.673	0.509	0.454	0.398	0.368	0.296	0.258
T_{MLB}	R	0.0	2.1	7.0	11.5	10.8	6.8	3.3
	D_{KS}	0.551	0.334	0.438	0.504	0.517	0.472	0.42
T_{MLMC}	R	88.5	96.8	98.5	99.3	100.0	100.0	100.0
	D_{KS}	0.837	0.928	0.962	0.979	0.995	1	1
T_{SBMC}	R	7.8	7.3	5.0	4.8	2.8	3.0	3.6
	D_{KS}	0.117	0.1	0.065	0.043	0.041	0.06	0.046
T_{MVAMC}	R	0.0	0.1	0.0	0.0	0.0	0.0	0.0
	D_{KS}	0.636	0.491	0.428	0.334	0.315	0.276	0.248
T_{AMMC}	R	0.0	0.1	0.0	0.0	0.0	0.0	0.0
	D_{KS}	0.614	0.512	0.456	0.367	0.344	0.301	0.25

R is observed rejection percentages for $\alpha = 5\%$

D_{KS} is Kolmogorov-Smirnov distance statistic

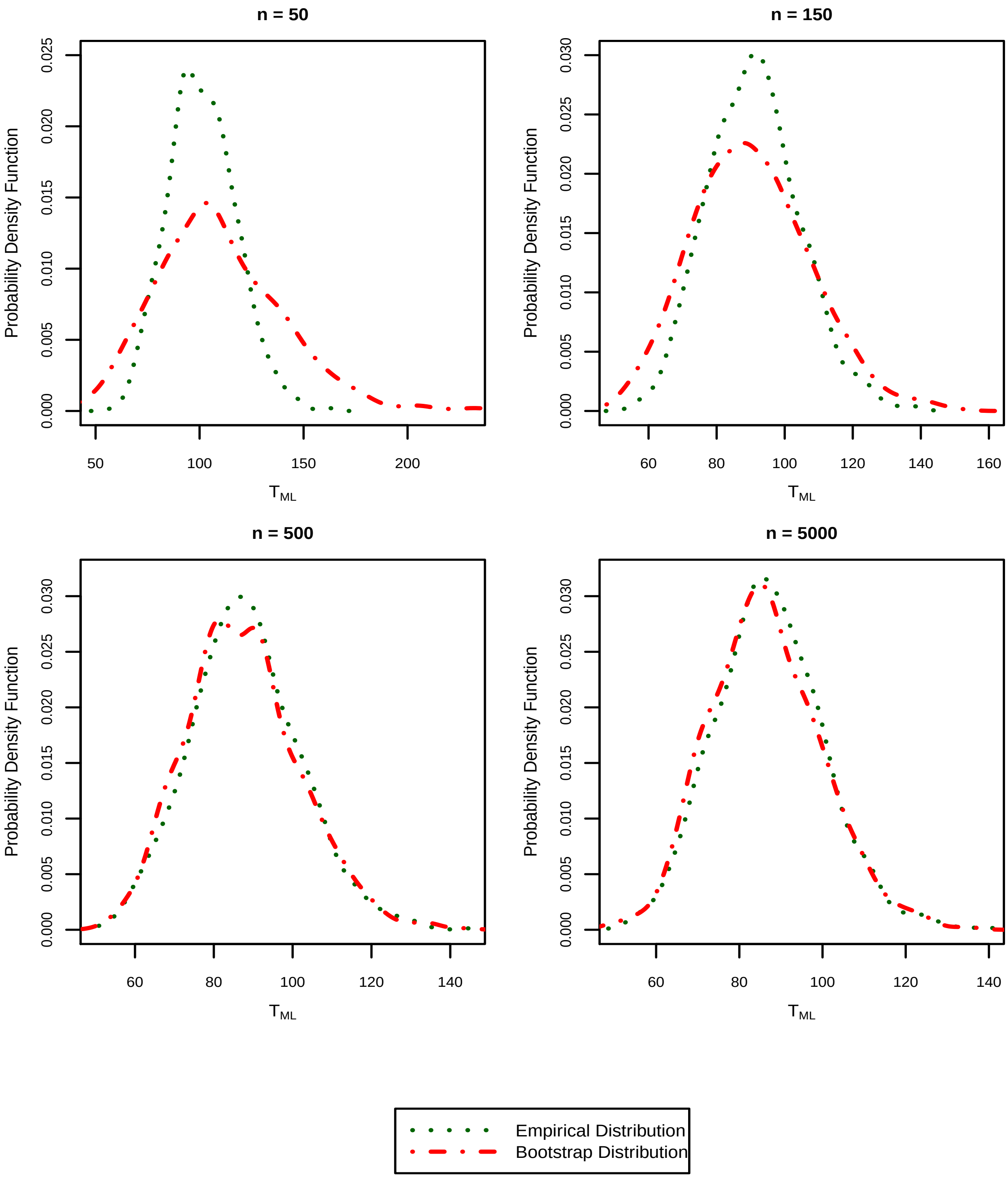
3.3 On the problems with the bootstrap in SEM

As mentioned previously, more caution is necessary when applying the bootstrap method in higher dimensional models. Previous Monte Carlo studies in the literature evaluating the performance of the bootstrap in SEM was limited to models with at most 9 indicator variables (see e.g., Enders, 2005; Ichikawa, & Konishi, 1995; Nevitt & Hancoock, 2001; Sharma & Kim, 2013). Except for a few cases, those studies do not consider very small sample sizes relative to the number of variables. From those studies it is not clear how the bootstrap method performs when the ratio of p/n is not small. In the last section we observed that when the dimension of models is large, the bootstrap method breaks down.

In this section we investigate the performance of the Bollen–Stine bootstrap method in more detail and we try to understand the root of the problem with bootstrapping in higher dimension models.

In any inferential statistics problem, if it were possible to access to the true population distribution that the data are from, and we could generate many samples from that population under the hypothesized model, by keeping all other aspects of the population constant, we could reproduce the exact empirical distribution of the subject test statistic. Bootstrapping is basically an imitation of this process. The difference is that in bootstrapping we are limited to the data and therefore, bootstrap samples contain repeated cases. We suspect these repeated cases in bootstrap samples can damage the distribution of the ML test statistic in SEM when we deal with a large model with a small sample. Since each sample has its own estimated empirical bootstrap distribution of the test statistic, it is not possible to compare the bootstrap distribution to the actual empirical distribution of the test statistic directly. Here, to investigate the effect of repetition of data in test statistics we use independent samples with repeated cases. First a sample based on the true hypothesized model is generated and the transformation (2.1) is applied to the sample based on the true model parameters (in bootstrapping we use estimated parameters). Then a sample with

Figure 3.1: Kernel density graphs of empirical distribution of T_{ML} vs. empirical bootstrap distribution of T_{ML} for model A_1



replacement is drawn from the transformed data. This way we achieve a random sample that is drawn from the correct model but with repeated cases. The desired test statistics can be calculated from this sample. We replicate this process 1,000 times and the empirical distribution of this test statistic (for simplicity we call it a bootstrap distribution) is compared with the empirical distribution of the test statistic that is computed based on samples without repeated cases.

Figure 3.1 shows the kernel density graphs of the empirical distribution of bootstrap T_{ML} versus the empirical distribution of T_{ML} for model A_1 (when the distribution of the data is normal) and for different sample sizes. Due to repeated cases that occur in the sample, when the sample size is small, the bootstrap distribution appears to have larger variance. The effect of repeated cases in the sample disappears as sample size gets larger. In a normal model, repeated cases have almost no effect when the sample size is 500 or larger. Figure 3.2 shows the kernel density graphs of empirical distribution of T_{ML} with and without repeated cases for model A_4 where the distribution of the data is not normal. We can observe that in this case also the variance of the bootstrap distribution at small sample sizes is larger than the empirical distribution of the test statistic. As sample size increases the variance of the bootstrap distribution is getting closer to the variance of the empirical distribution but the mean of the bootstrap distribution does not converge to the mean of the empirical distribution even for sample sizes as large as 5000.

The over-estimation of the empirical variance of the test statistic due to repeated cases gets worse when the dimension of the data increases. In Figure 3.3 this is shown by comparing the bootstrap distribution to the empirical distribution of the ML test statistics for the model B_1 which has 30 indicator variables. In this case the bootstrap distribution has larger variance than the empirical distribution of T_{ML} even for data with sample size equal to 500. In Figure 3.4 we observe the smaller mean in the bootstrap distribution for model B_4 which has a non-normal distribution. Since for large sample sizes the mean of the bootstrap is smaller than the mean of empirical distribution of T_{ML} it is suspected that this change of

Figure 3.2: Kernel density graphs of empirical distribution of T_{ML} vs. empirical bootstrap distribution of T_{ML} for model A_4

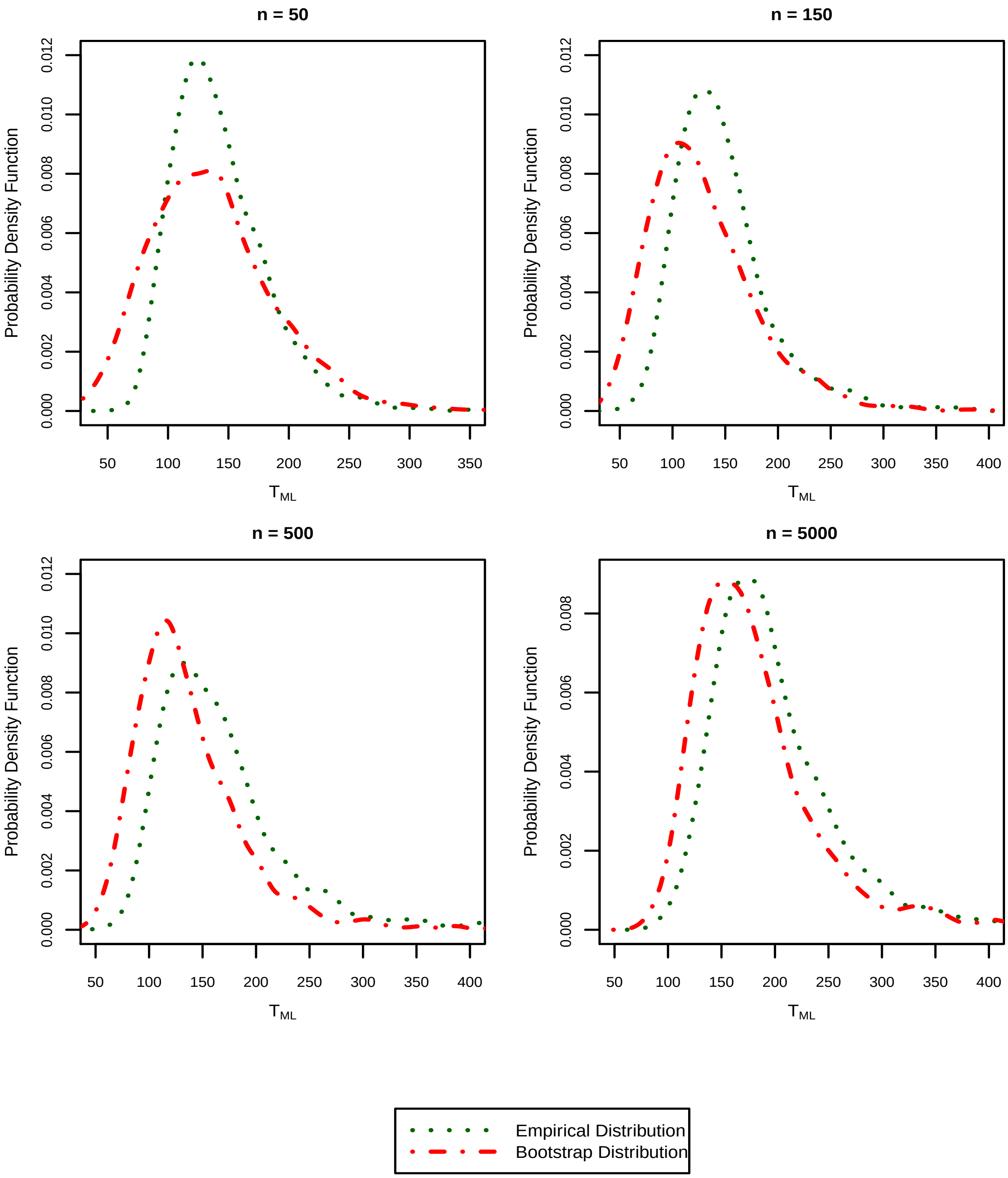


Figure 3.3: Kernel density graphs of empirical distribution of T_{ML} vs. empirical bootstrap distribution of T_{ML} for model B_1

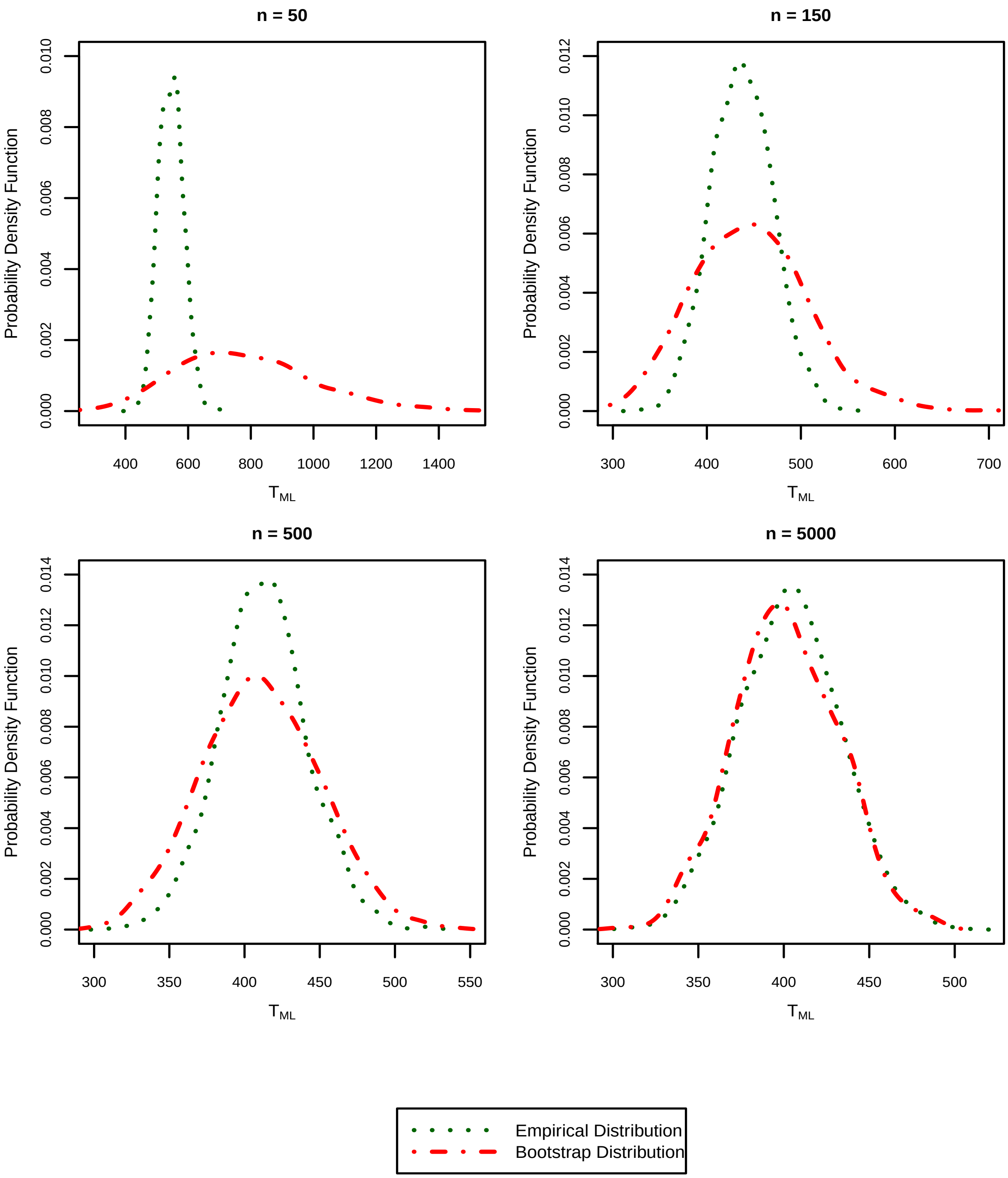
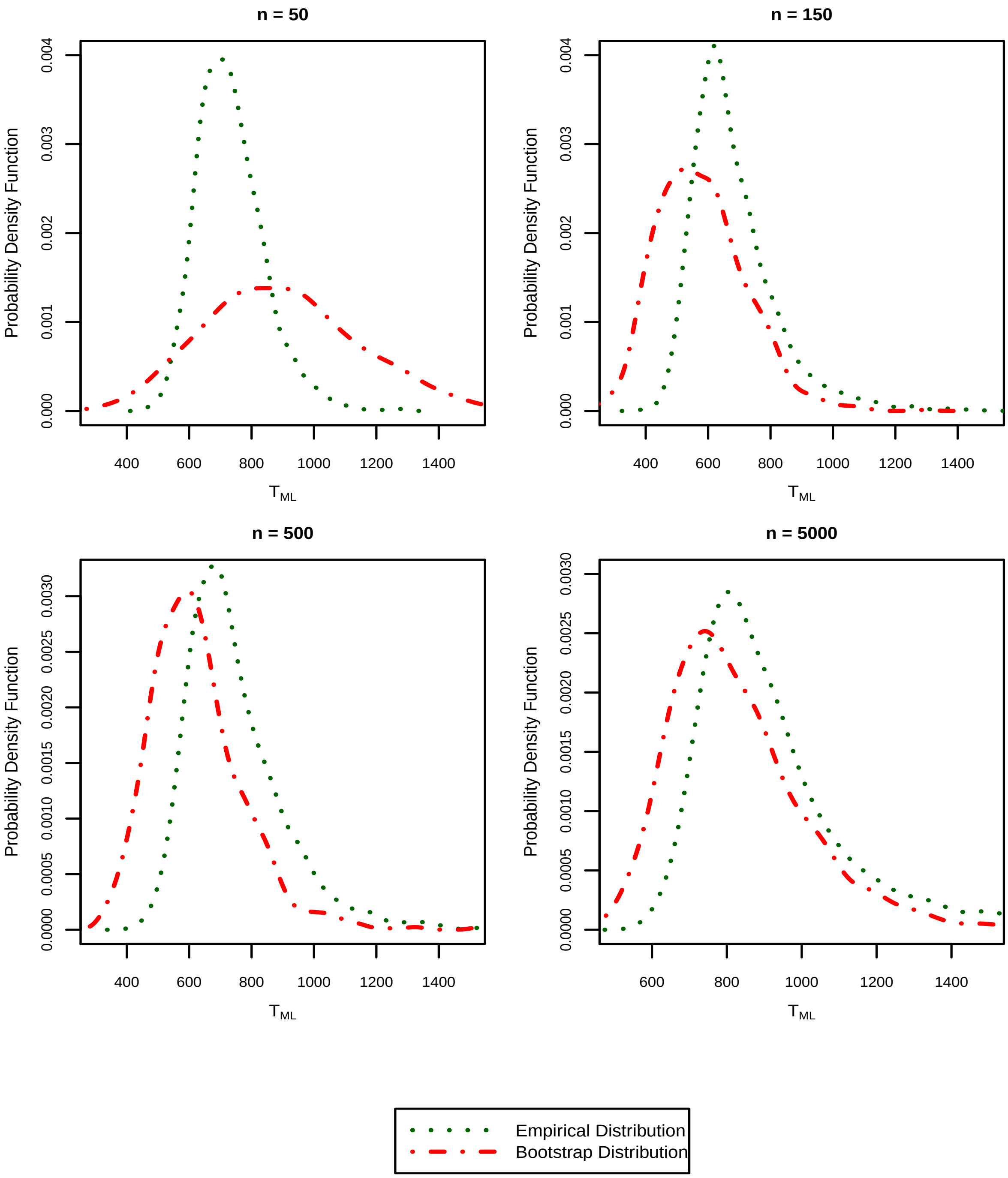


Figure 3.4: Kernel density graphs of empirical distribution of T_{ML} vs. empirical bootstrap distribution of T_{ML} for model B_4



the mean cannot be caused only by repeated cases.

One problem with the Bollen–Stine transformation is that if the data is not normally distributed then the transformed data does not necessarily have the same distribution as the original data. In fact the marginal distribution of variables in the transformed data driven by a linear combination of variables in the original data and this combination does not necessarily hold the marginal distributions of the original data. We do not have this problem when the original data is distributed normally. This has been mostly ignored in bootstrapping procedures in SEM as we hope the transformation does not change the overall distribution dramatically. In an unsuccessful attempt to eliminate the effect of this distributional change, instead of T_{ML} we have used T_{SB} as a robust test statistic in the bootstrap procedure. Using T_{SB} is as straight forward as T_{ML}. We simply compute T_{SB} for each bootstrap sample to estimate empirical distribution of SB test statistic and then we used it to determine the p-value of the original T_{SB} statistic. Figure 3.5 shows kernel density graphs of the empirical distribution of bootstrap T_{SB} (i.e. samples with repeated cases) verses the empirical distribution of T_{SB} for model A_4. As we see, even if we may overcome the problem with mean differences due to the change of distribution of the data in the transformed sample, the variance of empirical bootstrap distribution gets larger because of repeated cases in bootstrap samples and it is worse compare to T_{ML}.

Since the distribution of T_{ML} is very sensitive to the normality assumption of the data, the empirical distribution of the test statistic generated by Bollen–Stine bootstrap samples very much depends on the distribution of the original sample. This raises another problem with using bootstrapping in SEM. We may have data with T_{ML} greater than the test statistic from another sample but the p-value of the bootstrap test is smaller. This can be seen in Figure 3.6. We generated 1,000 replicated samples from the true model and we plot p-values of the bootstrap test against T_{ML}. The dispersion of points in the plot shows that we have cases with the greater test statistic and larger p-value. On the other hand the Monte Carlo test does not have this problem and the p-value is a strictly decreasing function of the test

Figure 3.5: Kernel density graphs of empirical distribution of T_{SB} vs. empirical bootstrap distribution of T_{SB} for model A_4

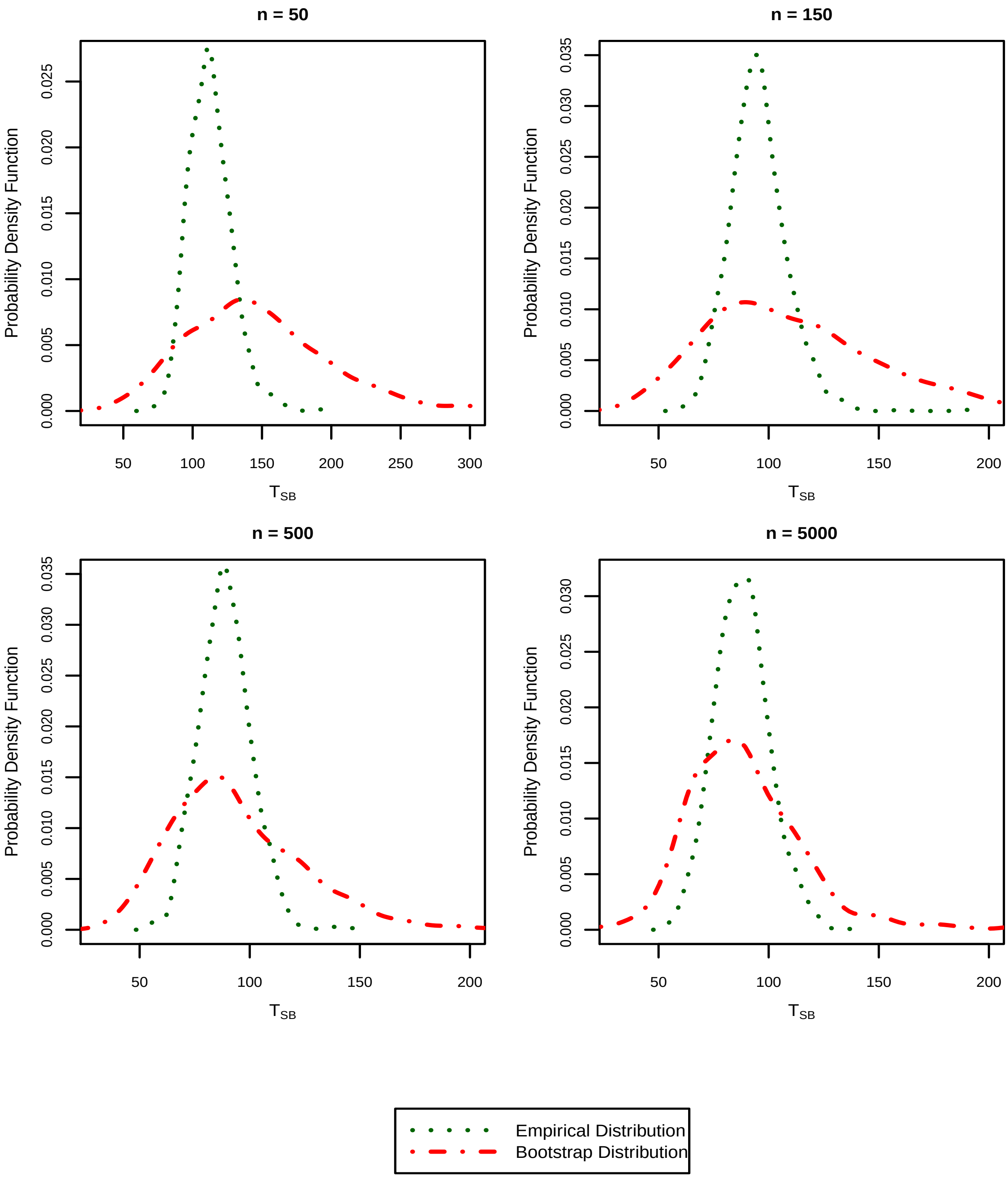
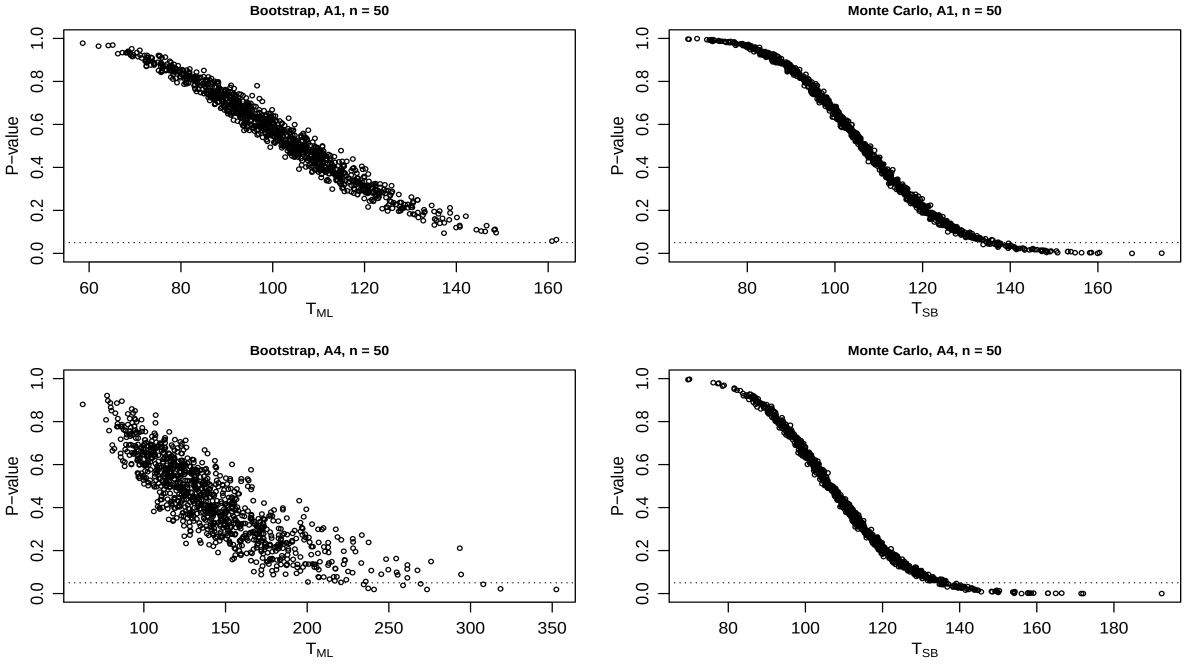


Figure 3.6: The p-value of the bootstrap test vs. the value of T_{ML} in Comparison to the p-value of Monte Carlo method vs. the value of T_{SB}



statistic. The dispersion of points in the bootstrap test is greater when the data is not from a multivariate normal distribution.

3.4 Illustration of the effect of model parameters on test statistics

In general, goodness of fit test statistics in SEM are a function of estimated parameter $\hat{\boldsymbol{\theta}}$ and therefore they depend on the true values of population parameters. When the asymptotic theories exist the distribution of those test statistics asymptotically do not depend on the values of model parameters however, in small samples the exact distribution of test statistics (e.g. T_{ML}) depends on the true population parameters. In Monte Carlo estimation of T_{SB} we use the estimated parameters of the model to generate Monte Carlo samples under the hypothesized model. In this section we study the effect of changes in model parameters to distribution of T_{SB} and we show that even for a very small samples the effect of parameters is negligible.

Figure 3.7: Q-Q plots of observed null distributions of T_{SB} with two model parameters $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$ when data are from normal distributions, condition A_1

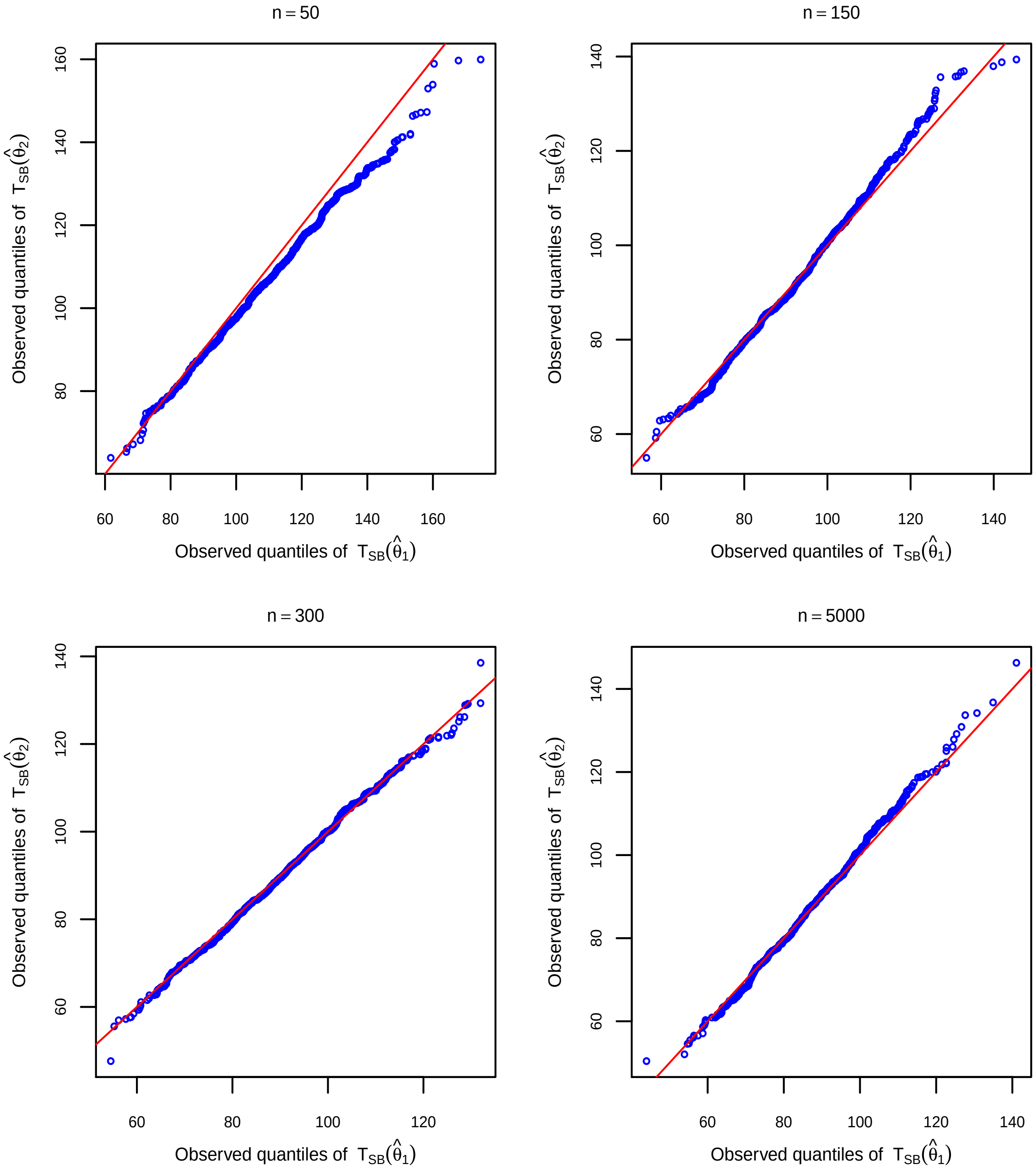
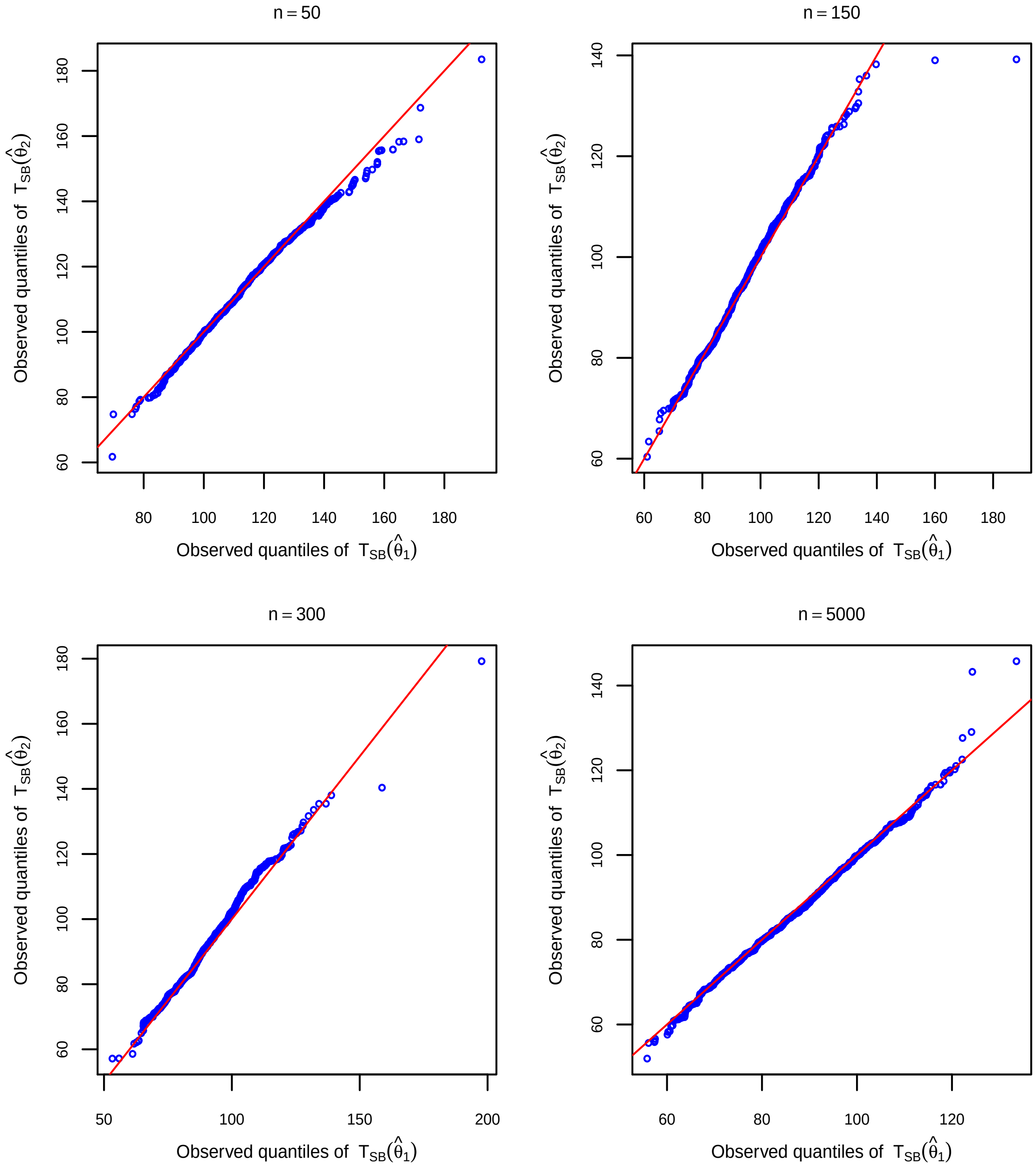


Figure 3.8: Q-Q plots of observed null distributions of T_{SB} with two model parameters $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$ when data are from non-normal distributions, condition A_4



We consider conditions A_1 for multivariate normal and A_4 to represent non-normal distributions. For each condition the observed distribution of T_{SB} is compared between two different parameter sets, $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2 \in \Omega_0$. For parameter set $\boldsymbol{\theta}_1$ we use the parameters defined in section 3.1. For parameters in $\boldsymbol{\theta}_2$ factor covariances have been changed to 0.1, all factor loadings set to 0.5, and uncorrelated covariance matrix of unique factors computed accordingly to make variances of $\mathbf{x}$ equal to 1.

Figure 3.7 shows quantile plots of observed distribution of $T_{SB}(\hat{\boldsymbol{\theta}}_1)$ verses that of $T_{SB}(\hat{\boldsymbol{\theta}}_2)$ for 1,000 replicated samples of size 50, 150, 300, and 5000 from two normally distributed population, one with structured covariance of $\boldsymbol{\Sigma}(\boldsymbol{\theta}_1)$ and the other one with $\boldsymbol{\Sigma}(\boldsymbol{\theta}_2)$. In all sample sizes, quantiles of two empirical distributions are fairly aligned to the line $y = x$. This means that changing parameters does not make a significant difference in the distribution of T_{SB} with sample size as low as 50. In Figure 3.8 samples have been drawn from a multivariate population with a non-normal distribution and the results are similar to those of the normal. In section 2.2 it has been suggested to use $\hat{\boldsymbol{\theta}}$ estimated from the original observation to generate Monte Carlo samples. The ideal situation is to actually use the true population parameters but in practice this is impossible. Another suggestion is to draw parameters randomly for each Monte Carlo random sample to reduce the effect of values of parameters, for example a uniform random number from $\hat{\boldsymbol{\theta}} \mp 2SD(\hat{\boldsymbol{\theta}})$. However, values of parameters do not have significant impact on the performance of the Monte Carlo test and one can even choose those parameters arbitrarily as long as they are in the null space, Ω_0.

3.5 On convergence of the Satorra-Bentler scaled statistic

In section 3.2 we see that Type I error rates of the Monte Carlo method for the SB scaled statistic, T_{SBMC}, are close to the nominal level across all sample sizes and all conditions. The performance of T_{SBMC} very much depends on how T_{SB} converges for different distributions. Particularly, how T_{SB} adjusts for violation of normality and whether its distribution is stable at small sample sizes. In this section we study the behavior of T_{SB} for different underlying distributional conditions.

Figure 3.9 shows kernel density graphs of empirical null distributions of T_{SB} for different distribution conditions and sample sizes for factor model A. Conditions are distinguished by different colors, for example red is used for multivariate normal. In the last section we saw that values of estimated parameters do not have a noticeable effect on the distribution of T_{SB} thus it is proper to assume the normal condition, red curve, represents the Monte Carlo test. The vertical dashed line shows the 95-th percentile of T_{SB} for condition A_1, which is the normal condition. Roughly speaking, it can be considered as the critical value of the Monte Carlo test for nominal level of $\alpha = 0.05$. We can see that lines are close to each other and are getting closer as sample size increases and for sample size equal to 5000 the difference of distribution of T_{SB} between different conditions almost disappears. Lines related to conditions A_5 and A_6, shown by brown and purple, stay close consistently for all sample sizes and are over the red curve. It is noticeable that the condition A_2, green curve, is lower than the normal condition for small sample size and get closer as sample size increases. The dotted curve shows the probability density function of the asymptotic χ^2 distribution with 87 degrees of freedom. When sample size is 50 the empirical distribution of T_{SB} has considerably greater mean than asymptotic χ^2 distribution in all conditions. Up to sample size equal to 300 the Monte Carlo normal curve shows better approximation to T_{SB} than the asymptotic χ^2 distribution.

Figure 3.10 shows kernel density graphs for model B, which has 30 random variables. Here, when the dimension of model gets larger, it is more recognizable that the condition 2, shown by green, has a smaller mean than other conditions at small samples sizes. The asymptotic χ^2 distribution in this model has 402 degrees of freedom and for small sample sizes the mean of T_{SB} in all conditions is greater than 402. By coincidence the dotted curve gets very close to the green curve of condition 2. This is because the specific distribution of data in condition 2 in this specific dimension of the model recovers the over-rejection of T_{SB} at small samples.

Figure 3.9: kernel density graphs of empirical null distributions of T_{SB} for model A, $p = 15$

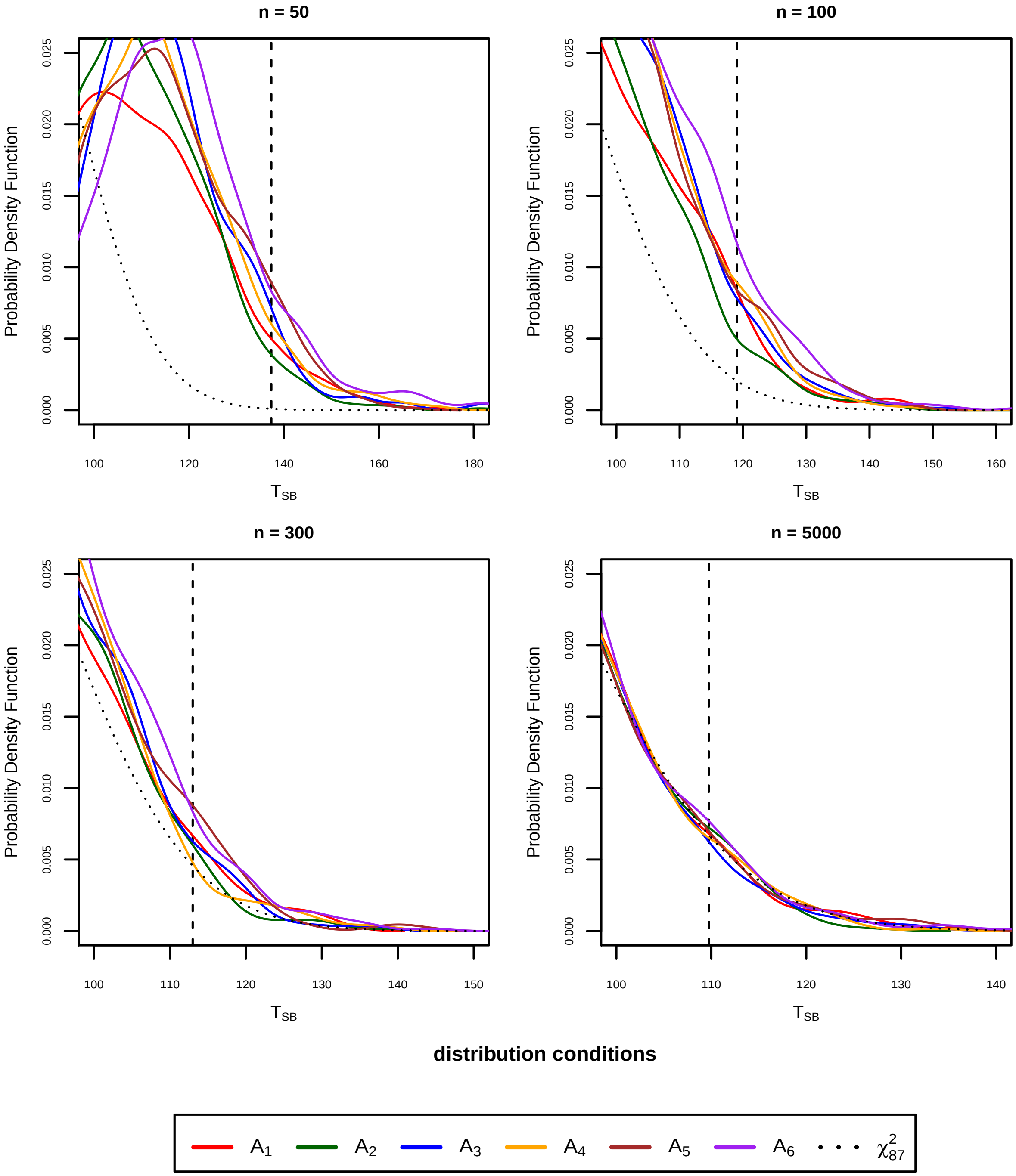
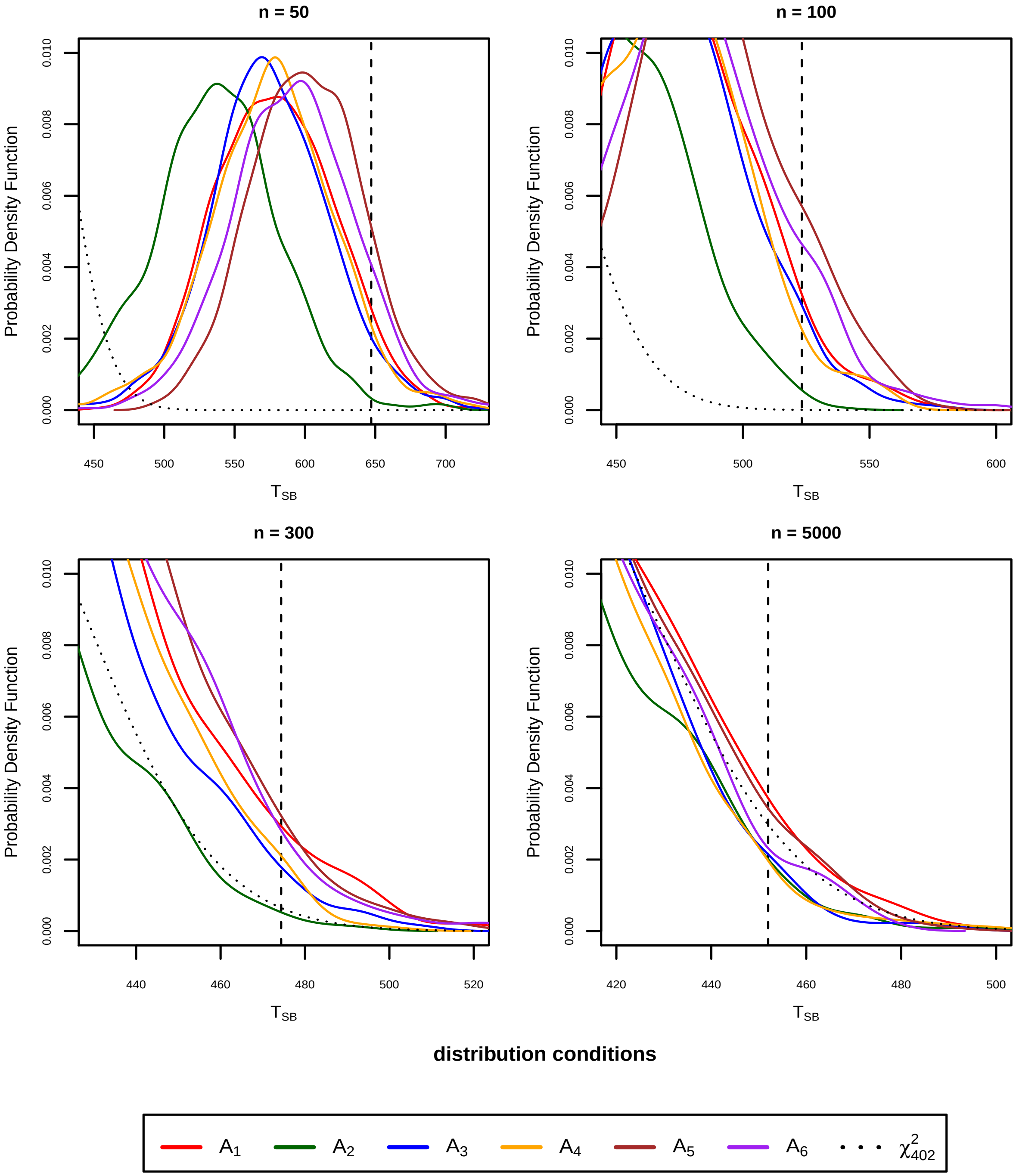


Figure 3.10: kernel density graphs of empirical null distributions of T_{SB} for model B, $p = 30$



3.6 Power analysis

The emphasis of this study was to introduce a method to control Type I error. If a test statistic over-rejects the correct hypothesized model then the power of that test is meaningless. However, in this section a small power analysis is reported to give us confidence that the proposed method is able to reject a misspecified model with acceptable power. The power study is done only on the model A and for all 6 distribution conditions defined in section 3.2. For each condition 1,000 replicated samples are drawn from population with structural covariance in equation (3.2) but this time with factor loading matrix as defined below:

$$\mathbf{\Lambda}^T = \begin{pmatrix} 0.70 & 0.70 & 0.75 & 0.80 & 0.80 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0.30 & 0 & 0 & 0 & 0 & 0.70 & 0.70 & 0.75 & 0.80 & 0.80 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0.35 & 0 & 0 & 0 & 0 & 0.70 & 0.70 & 0.75 & 0.80 & 0.80 \end{pmatrix}.$$

The difference between this model and the previous model defined in section 3.1 is that we added two extra paths to the model. One path from factor 2 to the first variable that before had only one direct path from factor 1, and other path is from third factor to variable 6, which had only one direct path from second factor. The misspecified hypothesized model is the one that we defined in section 3.1. Since the hypothesized model is not correct we expect to reject it and the proportion of rejections gives an estimate to power of the test under this model misspecification. The p-value for each replicated sample is computed and percentage of p-values that are less than nominal level of $alpha = 0.05$ are reported in Table 3.14. The cases where observed Type I error is exceptionally greater than nominal level are not reported in Table 3.14. For sample size greater than 300 for all conditions all test statistics reject the null hypothesis almost 100%. For condition 1 Monte Carlo method has competitive power compared to Bartlett correction to T_{ML} and T_{SB} at small sample sizes. The bootstrap test gives a very small rejection percentage when sample size is equal to 50. Overall T_{SBMC} gives reasonable rejection rates in all conditions. The lowest rejection percentage for T_{SBMC} is 19% for condition 2 at sample size equal to 50.

Table 3.14: Power of different test statistics for model A

		Sample size						
	Test Statistics	n = 50	n = 100	n = 150	n = 300	n = 500	n = 2000	n = 5000
condition 1	T_{ML}	72.7	86.7	98.5	100	100	100	100
	T_{MLb}	27.0	72.9	96.1	100	100	100	100
	T_{MLs}	36.6	77.3	97.0	100	100	100	100
	T_{SB}	84.8	90.4	99.0	100	100	100	100
	T_{SBb}	41.5	79.2	97.2	100	100	100	100
	T_{SBs}	52.8	82.0	97.6	100	100	100	100
	T_{MLB}	1.4	52.1	90.6	100	100	100	100
	T_{MLMC}	32.1	75.0	96.6	100	100	100	100
	T_{SBMC}	32.0	75.1	96.5	100	100	100	100
	T_{MVAMC}	31.4	74.8	96.4	100	100	100	100
	T_{AMMC}	31.6	74.8	96.4	100	100	100	100
condition 2	T_{SB}	76.8	71.4	79.3	95.7	98.9	99.9	100
	T_{SBb}	27.7	48.8	68.8	94.4	98.6	99.9	100
	T_{SBs}	37.7	53.6	71.6	94.7	98.7	99.9	100
	T_{MLB}	11.6	54.7	78.4	97.4	100	100	100
	T_{SBMC}	19.0	43.8	66.0	94.0	98.6	99.9	100
condition 3	T_{SB}	91.1	89.2	94.2	98.1	99.7	100	100
	T_{SBb}	57.3	77.1	90.7	97.9	99.7	100	100
	T_{SBs}	68.9	80.2	91.3	98.0	99.7	100	100
	T_{MLB}	6.5	48.2	76.0	96.0	99.7	100	100
	T_{SBMC}	46.6	73.2	89.2	97.8	99.7	100	100
condition 4	T_{SB}	83.0	74.6	83.2	94.7	98.9	100	100
	T_{SBb}	38.6	55.1	73.6	93.6	98.9	100	100
	T_{SBs}	49.2	61.3	76.2	93.8	98.9	100	100
	T_{MLB}	4.5	33.8	59.0	91.2	98.4	100	100
	T_{SBMC}	29.3	51.6	70.9	93.2	98.7	100	100
condition 5	T_{SB}	92.1	95.9	99.0	100	100	100	100
	T_{SBb}	57.0	85.6	97.5	100	100	100	100
	T_{SBs}	68.0	89.1	98.1	100	100	100	100
	T_{MLB}	1.3	40.0	83.0	100	100	100	100
	T_{SBMC}	44.4	82.8	96.8	100	100	100	100
condition 6	T_{SB}	90.7	89.5	90.9	99.7	100	100	100
	T_{SBb}	51.1	72.5	84.8	99.4	99.8	100	100
	T_{SBs}	63.9	77.2	86.8	99.5	99.8	100	100
	T_{MLB}	2.9	33.5	64.8	98.1	99.8	100	100
	T_{SBMC}	41.1	67.0	82.5	99.4	99.9	100	100

CHAPTER 4

Conclusions

Assessing an accurate goodness of fit test has always been a challenging concept among SEM practitioners. When the distribution of observations is not from a multivariate normal, typical ADF test statistics are not appropriate in the case of limited access to very large sample sizes. We introduced a new model based Monte Carlo test to evaluate overall goodness of fit in SEM. In a comprehensive simulation study we showed that the proposed method performs well in controlling Type I error when observations are not from a population with a multivariate normal distribution at small sample sizes. We also compared the new Monte Carlo method to those of existing statistics. In a variety of distributional conditions and sample sizes the proposed method is shown to outperform its competitors. A simulation study on the power of the Monte Carlo test also shows that the new test has an acceptable power in rejection of misspecified models.

We also discussed the classical bootstrap test in SEM introduced by Bollen and Stine (1992) in contrast to our Monte Carlo test. It was shown that the bootstrap method has critical problems when the dimension of the model gets larger. It appears that repeated cases as well as lack of robustness of T_{ML} in the Bollen-Stine bootstrap method is problematic at small samples in models with a larger number of variables. When the normality assumption of observations is not met, even for larger sample sizes, the bootstrap method performs poorly in a model with large number of observed variables. Those problems do not exist in our proposed Monte Carlo test.

In a new study by Jiang and Yuan (2017) four new test statistics were proposed to improve the performance of the goodness of fit test. Their conditions 1 and 3 are similar to conditions A_1 and A_2 in our study and results for similar test statistics reported in both studies such as T_{ML}, T_{SB}, and T_{MVA} statistics agree with one another. For their fourth corrected statistic, noted as P_{cor4}, they used the average of p-values from T_{SB} and T_{MVA}. In condition A_2, the P_{cor4} performs best among all other tests at small sample sizes. The rejection rate for the true model at sample sizes equal to 50, 100, 500, and 5000 is 6.4%, 1.4%, 0.6%, and 3.8%, respectfully in comparison our T_{SBMC} rejection rates in similar situations in the same order are 3.8%, 3.2%, 2.9%, and 4.2% which all are within the accepted criteria of rejection rate used in Jiang and Yuan (2017). In addition our test performs well in estimation of the overall distribution of the target test statistic. Jiang and Yuan (2017) did not report on the performance of their test in approximating the overall empirical distribution of the statistic and only reported the mean and variances of test statistics; therefore, we are not able to comment in this regard. They also did not study the performance of their proposed tests in models with a larger number of variables and it is not clear how those new tests perform in higher dimensional models.

Despite the fact that the proposed Monte Carlo test performs better compared to existed test statistics, it can be easily implemented to other asymptotic test statistics as long as those statistics are pivotal. For example, it will be interesting to see how the Monte Carlo approach performs in those statistics described in Jiang and Yuan (2017).

Another improvement to the proposed Monte Carlo test can be done by using an approximation to the distribution of the population based on the observations, instead of use of a normal distribution. For example, one can estimate skewness or kurtosis or both of marginals, and use them to generate Monte Carlo samples from a population with the estimated marginal skewness or kurtosis or both.

The Monte Carlo test can alternatively be implemented to situations in which the bootstrap method traditionally has been used. For example, in testing models with missing data, the Monte Carlo test can be used as an alternative to the bootstrap (s.g. Enders, 2005).

The aim of this study was to show the usefulness of the Monte Carlo method to improve performance of existing goodness of fit test statistics in SEM. The results of this study can open a debate on using the Monte Carlo test in situations where the existing asymptotic test is not reliable with small samples or a closed form for the asymptotic statistic does not exist but the test statistic still is asymptotically pivotal.

As noted previously, the term Monte Carlo test is often confused with the bootstrap test, specifically, in the parametric form. Here, we emphasized using the term 'Monte Carlo test' for two reasons: First, in the Monte Carlo test we did not resample from observations and instead we artificially generated Monte Carlo samples. Secondly, we wanted to make a distinction between the Monte Carlo test to the existing bootstrap test in the SEM literature. After all, although estimated parameters were used in the Monte Carlo samples, we have shown the effect of parameters in the test statistics are negligible, and the Monte Carlo test is rather more model-based than parametric.

REFERENCES

[1] Asparouhov, T., & Muthén, B. O. (2010). Simple second order chi-square correction. Technical report. Retrieved from Mplus website: www.statmodel.com

[2] Barnard, G. A. (1963). Discussion on The spectral analysis of point processes (by M. S. Bartlett). *Journal of the Royal Statistical Society, B, 25,* 294.

[3] Bartlett, M. S. (1950). Tests of significance in factor analysis. *British Journal of Psychology, Statistical Section Part 2, 3,* 77–85.

[4] Bentler, P. M. (2006). *EQS 6 structural equations program manual.* Encino, CA: Multivariate Software, Inc.

[5] Bentler, P. M., & Chou, C. P. (1987). Practical issues in structural modeling. *Sociological Methods & Research, 16,* 78–117.

[6] Bentler, P. M., & Dudgeon, P. (1996). Covariance structure analysis: Statistical practice, theory, and directions. *Annual Review of Psychology, 47,* 563–592.

[7] Bentler, P. M., & Yuan, K. H. (1999). Structural Equation Modeling with Small Samples: Test Statistics. *Multivariate Behavioral Research, 34:2,* 181-197.

[8] Beran, R. (1988). Prepivoting test statistics: A bootstrap view of asymptotic refinements. *Journal of the American Statistical Association, 83,* 687-697.

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

[10] Boomsma, A. (1982). The robustness of LISREL against small sample sizes in factor analysis models. In K.G. Jöreskog & H. Wold (Eds.), *Systems Under Indirect Observation: Causality, Structure, Prediction, Part I* 149–173.

[11] Browne, M. W. (1974). Generalized least squares estimators in the analysis of covariance structures. *South African Statistical Journal, 8,* 1–24.

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

[13] Browne, M. W. (1987). Robustness of statistical inference in factor analysis and related models. *Biometrika, 74,* 375-384.

[14] Chou, C. P., Bentler, P. M., & Satorra, A. (1991). Scaled test statistics and robust standard errors for non-normal data in covariance structure analysis: a Monte Carlo study. *British Journal of Mathematical and Statistical Psychology, 44,* 347-357.

[15] Cornea-Madeira A., & Davidson R. (2014). A Parametric Bootstrap for Heavy Tailed Distributions. *Econometric Theory, 31,* 1-22.

[16] Efron, B. (1979). Bootstrap methods: Another look at the jackknife. *Annals of Statistics, 7,* 1–26.

[17] El Karoui, N., & Purdom, E. (2015), Can We Trust the Bootstrap in High-Dimension? Technical Report 824, Department of Statistics, University of California, Berkeley.

[18] Enders, C. K. (2001). An SAS Macro for Implementing the Modified Bollen–Stine Bootstrap for Missing Data: Implementing the Bootstrap Using Existing Structural Equation Modeling Software. *Structural Equation Modeling, 12, 4,* 620–641.

[19] 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.

[20] Hall, P., & Titterington D. M. (1989). The effect of simulation order on level of accuracy and power of Monte Carlo tests. *Journal of the Royal Statistical Society, B, 51,* 459-467.

[21] Hall, P. (1992). *The Bootstrap and Edgeworth Expansion,* New York: Springer.

[22] Herzog, W., & Boomsma, A., (2009). Small-Sample robust estimators of noncentrality based and incremental model fit. *Structural Equation Modeling, 16,* 1–27.

[23] Herzog, W., Boomsma, A., & Reinecke, S. (2007). The model-size effect on traditional and modified tests of covariance structures. *Structural Equation Modeling, 14,* 361–390.

[24] Hope, A. C. A. (1968) A simplified Monte Carlo test procedure. *Journal of the Royal Statistical Society, B, 30,* 582-598.

[25] Hu, L., Bentler, P. M., & Kano, Y. (1992). Can test statistics in covariance structure analysis be trusted? *Psychological Bulletin, 112,* 351–362.

[26] Ichikawa, M., & Konishi, S. (1995). Application of the bootstrap methods in factor analysis. *Psychometrika, 60,* 77–93.

[27] Jiang, G., & Yuan, K. H. (2017). Four New Corrected Statistics for SEM With Small Samples and Nonnormally Distributed Data. *Structural Equation Modeling: A Multidisciplinary Journal, 00,* 1-16, DOI: 10.1080/10705511.2016.1277726.

[28] Jöreskog, K. G. (1967). Some contributions to maximum likelihood factor analysis. *Psychometrika, 32(4),* 443–482.

[29] Jöreskog, K. G. (1969). A general approach to confirmatory maximum likelihood factor analysis. *Psychometrika, 34,* 183-202.

[30] Lin, J., & Bentler, P. M. (2012). A third moment adjusted test statistic for small sample factor analysis. *Multivariate Behavioral Research, 47,* 448–462.

[31] Moshagen, M (2012). The model size effect in SEM: Inflated goodness-of-fit statistics are due to the size of the covariance matrix. *Structural Equation Modeling: A Multidisciplinary Journal, 19,* 86-98.

[32] Muthén, B.O., & Kaplan, D. (1992). A comparison of some methodologies for the factor analysis of non-normal Likert variables: A note on the size of the model. *British Journal of Mathematical and Statistical Psychology, 45,* 19–30.

[33] 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.

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

[35] R Development Core Team. (2016). *R: A language and environment for statistical computing.* [Computer software]. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from www.Rproject.org

[36] Rosseel, Y. (2012). lavaan: An R package for Structural Equation Modeling. *Journal of Statistical Software, 48(2)*, 1-36. http://www.jstatsoft.org/v48/i02/

[37] Sharma, P. N., & Kim, K. H. (2013). A comparison of PLS and ML bootstrapping techniques in SEM: A Monte Carlo study. *In New perspectives in partial least squares and related methods,* (pp. 201-208).

[38] Satorra, A., & Bentler, P. M. (1986). Some robustness properties of goodness of fit statistics in covariance structure analysis. *1986 ASA Proceedings of the Business and Economic Statistics Section,* 549 - 554.

[39] Satorra, A., & Bentler, P. M. (1988). Scaling corrections for chi-square statistics in covariance structure analysis. *American Statistical Association 1988 Proceedings of the Business and Economics Sections,* 308–313.

[40] Satorra, A., & Bentler, P. M. (1994). Corrections to test statistics and standard errors in covariance structure analysis. In A. von Eye and C. C. Clogg (Eds.), *Latent variables analysis: Applications for developmental research,* 399-419.

[41] Satterthwaite, F. E. (1941). Synthesis of variance. *Psychometrika, 6,* 309–316.

[42] Spearman, C. (1904). “General intelligence” objectively determined and measured. *American Journal of Psychology, 15,* 201-293.

[43] Swain, A.J. (1975). Analysis of parametric structures for variance matrices. Unpublished doctoral dissertation, Department of Statistics, University of Adelaide, Australia.

[44] Tong, X., & Bentler, P. M. (2013). Evaluation of a new mean scaled and moment adjusted test statistic for SEM. *Structural Equation Modeling, 20,* 148–156.

[45] Wu, H., & Lin J. (2016) A Scaled F-distribution as Approximation to the Distribution of Test Statistic in Covariance Structure Analysis. *Structural Equation Modeling 23,* 409-421

[46] Yuan, K. H. (2005). Fit Indices Versus Test Statistics. *Multivariate Behavioral Research, 40,* 115-148.

[47] Yuan, K. H. , & Bentler, P. M. (1998). Normal theory based test statistics in structure equation modeling. *British Journal of Mathematical and Statistical Psychology, 51*, 289-309.

[48] Yuan, K. H., Tian, Y., & Yanagihara, H. (2015). Empirical correction to the likelihood ratio statistic for structural equation modeling with many variables. *Psychometrika, 80,* 379–405.

[49] Yung, Y.-F.,& Bentler, P. M. (1994). Bootstrap-corrected ADF test statistics in covariance structure analysis. *British Journal of Mathematical and Statistical Psychology, 47,* 63–84.

[50] Yung, Y.-F., & Bentler, P. M. (1996). Bootstrapping techniques in analysis of mean and covariance structures. In G. A. Marcoulides & R. E. Schumacker (Eds.), *Advanced structural equation modeling: Issues and techniques* (pp. 195–226). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.