國立政治大學「教育與心理研究」
2008 年 12 月，31 卷 4 期，頁 1-22

Modern Robust Methods for Covariance in Structural Equation Modeling: ADF, SCALED, and Bootstrapping

吳佩真*

摘　要

ML和GLS是結構方程模式分析最常使用的參數估計法，兩種方法是基於常態分配假設來進行估計，然而，真實資料卻時常違反常態性假設。在此情形下，基於這二種估計法所求得的參數是否可靠，值得商榷。本研究旨在比較不同非常態情形下，這二種方法與四種不受常態性假設影響的強韌統計方法（ADF, SCALED, bootstrap-M_o和bootstrap-M_A）第一類錯誤率控制情形。結果發現：ML與GLS在所有非常態模擬資料，即使樣本數高達5,000，二者的第一類錯誤率超過35%。而ADF容易受小樣本影響產生過高的第一類錯誤率。SCALED, bootstrap-M_o和bootstrap-M_A較不易受樣本數影響，且可降低非常態所造成的問題。最後，提出未來研究與實務的建議。

關鍵詞：ADF, SCALED, bootstrapping, covariance structure

---

* 吳佩真：國立屏東教育大學教育心理與輔導學系助理教授
誌謝：本研究係由國科會研究計畫（NSC 94-2413-H-153-006）所支持，兩名匿名審查者提供的寶貴意見，在此一併致謝。
電子郵件：pcwu@mail.npue.edu.tw
收件日期：2007.06.23；修改日期：2008.02.19；接受日期：2008.06.12

Journal of Education & Psychology
December, 2008, Vol. 31 No. 4, pp. 1-22

Modern Robust Methods for Covariance in Structural Equation Modeling: ADF, SCALED, and Bootstrapping

Pei-Chen Wu*

Abstract

Although the maximum likelihood estimator based on normality theory is default in most available programs in structural equation modeling, the majority of data investigated in behavioral and social sciences violate the assumption of multivariate normality. This study evaluated six covariance structure analysis techniques (i.e., ML, GLS, ADF, SCALED, bootstrap-M_o and bootstrap-M_A) under various conditions of nonnormality. Results clearly illustrated that the ML and GLS failed to provide a good control of Type I error rates in all conditions of nonnormality even with the sample size of 5000. The ADF was essentially unusable in small to intermediated sample sizes. The SCALED and two bootstrap methods provided promising advantages but they were confined by small sample sizes. Additionally, the minimum requirements of sample sizes and bootstrapped samples for bootstrapping procedures were identified. Finally, a few suggestions were provided in the hope of improving the current practice.

Keywords: ADF, SCALED, bootstrapping, covariance structure

---

* Pei-Chen Wu: Assistant Professor, Department of Educational Psychology and Counseling, National PingTung University of Education

Acknowledgements: This research was supported by National Science Council of Taiwan (NSC 94-2413-H-153-006).

E-mail: pcwu@mail.npue.edu.tw

Manuscript recieved: 2007.06.23; Revised; 2008.02.19; Accepted; 2008.06.12

Introduction

Structural equation modeling (SEM) is a statistical technique to estimate and test causal relationships embedded in the model (Bentler, 1988) and it has played a significant role in multivariate analysis, with extensive applications in the behavioral and social sciences. The well known advantage of SEM is that it permits one to simultaneously test and model measured variables, latent variables and measurement errors on the basis of theoretical framework (Bentler & Dudgeon, 1996; Bollen, 2002). The estimations of SEM mainly rely on covariance matrix. That is, $\Sigma(\theta)$, the population covariance matrix, is used as a test statistic for evaluating the quality of structural model. Under multivariate normality, $\Sigma(\theta)$ can approximate the population covariance matrix well. However, under nonnormality, $\Sigma(\theta)$ could be contaminated, which in turn produces the disastrously consequences such as a high Type I error rate, low power and the inflation of fit indices (Bentler & Dudgeon, 1996; Yuan & Benlter, 2001; Yuan, Bentler, & Chan, 2004). In reality, most real data sets in the behavioral and social sciences, especially those collected by self-reported questionnaires, potentially violate the assumption of normality. For example, Micceri (1989) investigated 440 large sample achievement and psychometric measures and found all data were significantly nonnormally distributed, with several classes of contamination. Additionally, in some important areas of studies such as depression, abnormality and psychopathology, the nature of data represents nonnormally. Under such situations, can test statistics in covariance structure analysis estimated based on normal theory be reliable? And can the inferential conclusions based on covariance structure analysis be trusted?

Furthermore, Breckler (1990) and Jöreskog (1993) have noted that most researchers applied normal theory-based maximum likelihood (ML) or generalized lease squares (GLS) without seriously considering whether the assumption of normality had been violated. This lack of attention on data quality is probably due to the facts that the effects of nonnormally distributed data on covariance structures estimated based on normality theory are not well understood by applied researchers; that there exist various results on the robustness of ML or GLS procedures with nonnormal data sets (e.g., Anderson & Amemiya, 1988; Browne,

1984; Hu, Bentler, & Kano, 1992; Satorra & Bentler, 1990, 1991); and that there are rare analytical results indicating that asymptotic robustness (i.e., the validity of normal theory-based methods with large-sample nonnormal data) is not enough when the data contain influential cases (e.g., outliers or heavy tails). Thus, such an investigation is one of the major focuses of the current study.

Nonnormal distributions contain influential observations. In general, influential observations are classified into two categories. In the first category, the influences observations are only some extreme cases called outliers. Under such situations, nonnormality is due to outliers. ***S***, the sample covariance matrix, is excessively affected by a small proportion of outliers, which in turn leads to the inflation of fit indices for the majority of the data (Bentler & Dudgeon, 1996; Yuan & Benlter, 2001). Additionally, ***S*** is very sensitive to outliers because it has unbounded influence function and zero breakdown point (Hample, Ronchetti, Rousseeuw, & Stahel, 1986; Huber, 1981; Wilcox, 1997, 2003; Yuan & Bentler, 1998a). Put differently, any individual case can cause ***S*** to be arbitrarily large. (The readers are referred to Staudte and Sheather (1990) and Wilcox (1997) for the detailed discussion of influence function and breakdown points). In the second category, the influential observations are due to the heavy tails of sampling distribution (but heavy tails are not created by outliers). In such cases, the influential observations are associated with the skewness and kurtosis. When sampling distribution has heavy tails, ***S*** is an inefficient estimator of the population covariance matrix (Tyler, 1983; Yuan et al., 2004) and the population of fourth-order moments, kurtoses, do not exist. Browne (1982, 1984) pointed out that the kurtosis is critical because it is a key term in the mathematical expression for the covariances of covariances. Importantly, when heavy tails of a data set are coming from outliers, the situation becomes worse (e.g., Devlin, Gnanadesikan, & Kettenring, 1981; Yuan & Bentler, 1998a).

In general, there are two alternatives dealing with nonnormality in SEM. One is to identify outliers by means of some analytical procedures (e.g., Bollen & Arminger, 1991; Chatterjee & Yilmaz, 1992; Lee & Wang, 1996) and subjectively decide whether to remove them. This procedure is not encouraging since the most influential observations may not

be real outliers (Huber, 1981). Another way is to apply a robust approach--downweighting the effects of outliers (e.g., Campbell, 1980; Rousseeuw & van Zomeren, 1990; Wilcox, 1997; Yuan & Bentler, 1998a, 1998b). As mentioned above, the sample covariance matrix has unbounded influence function and zero breakdown point. On the other hand, by giving a proper weight to each individual case, the robust covariances have bounded influence functions as well as nonzero breakdown points. It implies that robust covariances are less affected by any influential observation. The term robust covariance in SEM is "well defined as long as the structural model $\Sigma(\theta)$ is invariant under a constant scaling factor (ICSF). That is, for any parameter vector θ and positive constant *a*, there exists a parameter vector θ^* such that $\Sigma(\theta^*) = a\Sigma(\theta)$" (Yuan & Bentler, 1998a: 368-369).

Several robust covariance structures have been developed to address the problems associated with the ML or GLS under various nonnormal conditions (e.g. Browne, 1984; Campbell, 1980; Hu et al., 1992; Kano, Berkane, & Bentler, 1990; Satorra & Bentler, 1990, 1991; Yuan & Bentler, 1998a, 1998b). Among these robust approaches, the most widely used methods are asymptotically distribution free estimator, ADF (Browne, 1984) and scaling corrected, SCALED (Satorra & Bentler, 1988, 1990). Both of them have their strengths and weaknesses under various conditions. In addition to these methods, the bootstrapping recently has been regarded as a promising approach (e.g., Yung & Bentler, 1994, 1996; Yuan, et al., 2004; Yuan & Hayashi, 2003; Yuan, Hayashi, & Yanagihara, 2007). Yet, there still exists many unknown issues about the applications of these robust methods. For example, do these robust methods perform well under various conditions of nonnormality? Therefore, such an exploration is another major focus in the present study. Related literature of these robust methods is discussed as follows.

Asymptotically distribution free estimator (ADF)

Brown (1984) developed an "asymptotically distribution free" (ADF) estimator procedure that does not assume multivariate normality of the measured variables. The ADF is based on multivariate elliptical distributions, which are symmetric with tails that can be either heavier or lighter than those of a normal distribution

as well as identical. The key of the ADF estimation is to utilize an optimal weight matrix which consists of a combination of second-order and fourth-order terms. The major theoretical advantage of the ADF is to produce asymptotically (large sample) unbiased estimates of the χ^2 goodness-of-fit test, parameter estimates as well as standard errors. There are, however, two important limitations of the ADF. The first limitation is that the calculation of the matrix of fourth-order moments needs a large sample size to generate stable estimates. Sample sizes of 1000 are necessary with relative simple models under typical conditions of nonnormality (Curran, West, & Finch, 1996) and 5000 observations are necessary for more complicated models (Hu & Bentler, 1995; Hu et al., 1992). Briefly, at a large sample size ($n \geq 5000$), the ADF performs as expected, providing observed Type I error rates at the nominal level. However, with complicated models or small to moderate sample sizes, the ADF has been problematic in terms of high rates of nonconvergence as well as high Type I error rates (Curran et al., 1996; Hu et al., 1992; Muthén & Kaplan, 1992). Now, questions may arise: large sample sizes ($n \geq 5000$) are rare in psychological and behavioral research. The second limitation is that the ADF is computationally demanding due to the calculations which need the inversion of its optimal weight matrix. With more than 30 measured variables, implementation of the ADF becomes impractical (Bentler, 1995; Hu et al., 1992). Importantly, the ADF is biased under certain conditions (e.g., Chou, Bentler, & Satorra, 1991), but conflict results are reported by Muthén and Kaplan (1992).

Scaling Corrected (SCALED)

Satorra and Bentler (1988, 1990, 1994) developed two modifications of standard goodness-of-fit statistic test based on ML and heterogeneous kurtosis, HK (Kano et al., 1990). Here, only modification of ML was proposed. The SCALED is a rescaled test statistic when normality theory χ^2 statistic does not follow the expected χ^2 distribution under nonnormality, by applying a scaling constant to the covariance matrix of the parameter estimates. That is, the normal theory χ^2 is divided by a constant *k*, whose value is a function of the model residual weight matrix, the observed multivariate kurtosis and the degree of freedom for the model.

Compared to the ADF, the SCALED is less affected by the model complexity and sample size. Also, it appears to provide good estimates of χ^2 for sample size 200 and higher. However, it has a tendency to over-reject models at smaller sample sizes (Hu et al., 1992; Hu & Bentler, 1995). Previous studies (e.g., Curran et al., 1996; Hu et al., 1992; Satorra & Bentler, 1990) have shown that the SCALED is superior to the ADF in nonnormal conditions, but the reason why the SCALED is better than the ADF is not clear. One speculation may be that the SCALED uses a matrix generated from these moments directly, but the ADF applies the inverted matrix, which leads to the accuracy problems in intermediate-size samples and the sufficiency problems in small-size samples (Hu et al., 1992).

Bootstrapping

The bootstrap, introduced by Efron (1979), is a computationally-intensive statistical tool which allows researchers free from the theoretical distributions of classical test statistics. The bootstrapping procedures involve resampling the data with replacement many times to generate an empirical estimate of the entire sampling distribution of a statistic. Recently, studies using bootstrapping on covariance structures have been on the increase (e.g., Bollen & Stine, 1990, 1992; Stine, 1989; Yuan & Hayashi, 2003, 2006; Yuan et al., 2007; Yung & Bentler, 1994; 1996). The bootstrap approach on covariance structures was initially developed by Beran and Srivastava (1985). Further, Bollen and Stine proposed a bootstrap method for adjusting the *p* value associated with T_{ML} (the likelihood ration statistic based on normal theory). Their results indicated that naive bootstrapping of T_{ML} for SEM models was inaccurate due to the distribution of bootstrapped model test statistics followed a noncentral chi-square distribution rather than a central one. To address this problem, they applied a transformation on the original data to make the model-implied covariance matrix become the true underlying covariance matrix in the population. Extending the study of Bollen and Stine, Yung and Bentler proposed two bootstrap methods, symbolized as bootstrap-M_o and bootstrap-M_A. Due to the limitations of the number of repeated samples (i.e., only 10 repeated samples) as well as conditions of nonnormality, their results showed that both bootstrap-M_o and bootstrap-M_A were not reliable enough. Recently, Yuan and

Hayashi used three bootstrap (T_{ML}, T_{SB}, T_B) to estimate Type I error, power and sample-size determination. Their results illustrated that T_{ML} was asymptotically pivotal, when data were normally distributed. T_{SB} (a rescaled version of the likelihood ratio statistic proposed by Satorra and Bentler) was not asymptotically pivotal for nonnormal distributions but for elliptical distributions. T_B (an asymptotically distribution-free statistic proposed by Browne) was asymptotically pivotal for sampling distributions with finite fourth-order moments (kurtoses). Furthermore, they indicated that combining downweighting with bootstrapping provided a theoretical justification for using bootstrapping to the data with heavy tails.

Relative to the classical approaches (e.g., ML and GLS), the robust approaches (e.g, ADF, SCALED and bootstrapping) proposed here have following promising advantages. First, by downweighting the effects of influential cases, robust methods result in smaller chi-square statistics that provide more support to a theoretical model. Second, robust methods give more reasonable solutions with problematic data (e.g., outliers and heavy tails), yet classical methods cause improper solutions such as Heywood cases. Third, both classical and robust methods lead to the same conclusions with approximately normal data sets (Bollen & Stine, 1992; Yuan & Bentler, 1998a, 1998b). Although the proposed robust methods have shown promise for covariance structure analyses, unknown issues related to them still exist. For instance, relative to the ADF, the SCALED is less affected by the model complexity and sample sizes. What is the minimum sample size for the SCALED to get a good control of Type I error rates? Under what conditions, the SCALED outperforms the ADF? There are various variants of bootstrapping, and which bootstrap method should be applied? Based on the study by Bollen and Stine, it suggested that the bootstrap-M_o is used to estimate the bootstrap distribution of the test statistic. Here, both bootstrap-M_o and bootstrap-M_A derived from Yung and Bentler (1996) are applied, but are they equivalent under various conditions of nonnormality? Additionally, as is well known, when using bootstrap methods, the minimum sample size is required. However, minimum sample size requirements for the original sample are rather vague. Studies only suggest that the bootstrap may be inappropriate with rela-

tively small sample sizes (e.g., Ichikawa & Konishi, 1995; Yung & Bentler, 1996). For practical data sets potentially containing outliers or heavy tails, how could we determine the sample size needed to achieve a good control of Type I error rate in covariance structure models? Finally, the minimum of B (i.e., number of bootstrapping) required to yield accurate estimates of *p* values in SEM is also unclear. Yung and Bentler suggested ideal $B = n^n$, but this collection is impractical for practical experiments.

Based on previous arguments, the objective of this study is to address the issues (mentioned above) associated with robust covariance in SEM. Monte Carlo simulations are utilized to evaluate two classical methods based on normal theory (ML and GLS) and four robust methods (ADF, SCALED , bootstrap-M_o and bootstrap-M_A) under different conditions of nonnormality (e.g., various sample sizes, skewness and kurtosis) in terms of their Type I error rates.

Method

Procedure

This study examined the relative performance of two classical methods relying on normal theory (ML, GLS) and four robust methods for covariance structure models (ADF, SCALED, bootstrap-M_o and bootstrap-M_A) under various conditions of multivariate nonnormality, with a particular emphasis on properly specified models. Monte Carlo simulations were designed based on two conditions manipulated here: distribution type of population (five distribution types) and sample size (n=100, 200, 500, 1000, 5000). When investigating both bootstrap methods, the number of bootstrapping B (B=200, 300, 500, 1000, 1500, 2000, 5000) was considered in addition to the conditions of distribution type and sample size. Three hundred random samples were analyzed in each of the population conditions. The outcomes were compared in terms of their Type I error rates.

Model Specifications and Distributional Conditions

An oblique three-factor model with three indicators per factor was examined. This confirmatory factor model was a correctly specified model, in which the population parameters included that all factor ladings were set to .70, uniqueness were set to .51, interfactor correlations were set to .30 and factor variances were set to 1.0. Five population distributions were con-

sidered through the manipulation of univariate skewness and kurtosis. Distribution 1 was multivariate normal with univariate skewness and kurtoses equal to 0. Distribution 2 and 3 were moderately nonnormal. The former distributed symmetrically with univariate skewness of 0 and kurtoses of 7 and the latter represented nonsymmetrically with univariate skewness of 2 and kurtoses of 7. Distribution 4 and 5 were severely nonnormal, with the former distributing symmetrically with univariate skewness of 0 and kurtoses of 21 and the latter representing nonsymmetrically with univariate skewness of 3 and kurtoses of 21. Model specifications were partially referred in Curran et al. (1996). Only asymmetrical distributions were analyzed in the study of Curran et al., however, both symmetrical and asymmetrical distributions were taken into consideration in the present study. Simulated raw data were generated in EQS (Bentler, 1995) to reach the desired conditions. The method developed by Vale and Maurelli (1983) was implemented to generate independent observations from specific nonnormal distributions. The programming accuracy check was done with SAS PROC UNIVARIATE.

Test Statistics

Maximum-likelihood (ML)

Jöreskog and Goldberger (1972) applied Maximum-likelihood in SEM. Let ***S*** represent the unbiased estimator that is based on a sample size n of $p \times p$ population covariance matrix Σ, whose elements are functions of a $q \times 1$ parameter vector $\theta : \Sigma = \Sigma(\theta)$. ML function is

$$F_{ML} = \log|\Sigma| - \log|S| + tr(S\Sigma^{-1}) - p, \quad (1)$$

where "|.|" is the determinant of a matrix, "tr" is the trace, and p is the total number of manifest variables (x and y) in the model.

Generalized least squares (GLS)

Jöreskog and Goldberger (1972) extended Generalized least squares method in the path analysis. GLS functions is

$$F_{GLS} = \frac{1}{2} tr\left\{\left[S - \Sigma(\hat{\theta})\right]W^{-1}\right\}^2, \quad (2)$$

where S is the observed covariance matrix, $\Sigma(\hat{\theta})$ is the covariance matrix implied by the hypothesized model and W^{-1} is a weight matrix.

Asymptotic distribution free (ADF)

Asymptotic distribution free was proposed by Brown (1984). ADF function is

$$F_{ADF} = \frac{1}{2}(k+1)^{-1} tr\left\{\left[S - \Sigma(\theta)\right]W^{-1}\right\}^2$$

$$-\delta\left\{tr\left[S - \Sigma(\theta)\right]W^{-1}\right\}^2, \quad (3)$$

where k is common kurtosis parameter of a distribution, W is any consistent estimator of Σ and $\delta = k/[4(k+1)^2 + 2pk(k+1)]$.

Scaling Corrected (SCALED)

Satorra and Bentler (1988) developed two modifications of the standard goodness-of-fit test (T_{ML}, T_{HK}). In this study only the modification of T_{ML} is used. The scaled ML function is

$$\bar{T} = T_{ML} / k, \quad (4)$$

where k is the scaling estimate and $k = tr(\hat{U}\hat{V}_{ss})/df$, $\hat{U}$ is a consistent estimator of U on the basis of θ. $\hat{V}_{ss}$ is the distribution-free estimator with $p^* \times p^*$ positive definite weight matrix.

Bootstrap

Yung and Bentler (1996) extended the study of Bollen and Stine (1992) and proposed two bootstrap methods, symbolized as bootstrap-M_o and bootstrap-M_A. Two methods are achieved by transforming the observed sample so that its first and second moments completely satisfy the hypothesized mean and covariance structure. The bootstrap-M_o is

$$R(M_o) \equiv \{y_i = \hat{\Sigma}_o^{1/2} S_n^{-1/2}(x_i - \bar{x}_n) + \hat{\mu}_o,$$

$$i = 1,\ 2,\ \}, \quad (5)$$

where $\hat{\Sigma}_o = \Sigma(\hat{\theta})$ and $\hat{\mu}_o = \mu(\hat{\theta})$ are estimated under the null hypothesis.

When the structure equations and the parameter values (θ_A) are hypothesized without any specified parameter values of θ_A, Σ_A and μ_A are just vectors with known value. The bootstrap-M_A is

$$R(M_A) \equiv \{y_i = \hat{\Sigma}_A^{1/2} S_n^{-1/2}(x_i - \bar{x}_n) + \hat{\mu}_A,$$

$$i = 1,\ 2,\ ....n\}. \quad (6)$$

The bootstrap distribution of T^* is used as an estimator for sampling distribution of the original test statistic T. The bootstrap estimate of the p value of T is p^*, which is defined by

$$p^* = \frac{1}{B}\sum_{j=1}^{B} I\left\{T_j^* > T\right\}. \quad (7)$$

In this study, let B= 200, 300, 500, 1000, 1500, 2000 and 5000.

Results

To evaluate the Type I error rates of the model test statistics under various conditions of nonnormality, the criterion of robustness proposed by Bradley (1978) was applied. According to Bradley's liberal criterion, when one estimator provides an empirical alpha within the interval $[.5\alpha,\ 1.5\alpha]$, it is regarded robust. For his stringent criterion, one estimator is considered robust if it provides an empiri-

cal alpha within the interval $[.9\alpha, 1.1\alpha]$. Here, Bradley's liberal criterion was utilized. Using $\alpha = .05$, the intervals for a robust estimator were [.025, .075]. Table 1 details the empirical Type I error rates across different conditions of nonnormality.

As expected, both ML and GLS had acceptable Type I error rates under conditions of normality even at the smallest sample size (e.g., n=100). With the departures from multivariate normality, however, the ML and GLS were not robust even at the largest sample sizes (e.g., n=5000), with percentages of model rejections ranging from about .20 to .45. Apparently, the more severe the nonnormality, the greater the corresponding Type I error rates. For example, under the severely nonnormal conditions (e.g., distribution 5: skewness=3, kurtosis=21), the percentages of model rejections of both ML and GLS reached more than .35 even with the largest sample sizes. Regarding the ADF, under the multivariate normal distribution models, rejection rates were within the robustness interval of [.025, .075], with an exception of $n \le 200$. Under nonnormal conditions, the ADF yielded observed Type I error rates nearly at the upper bound of .075, given n= 5000. Yet, with small to moderate sample sizes, the ADF appeared to be disastrously problematic; that is, it resulted in rejection rates as high as .48 under certain condition (e.g., distribution 5 with n=100). Under normal condition models, the SCALED operated a good control of Type I error rates, with only marginally above the .075 upper bound at n=100. Nevertheless, it was confined to small sample sizes. That is, it appeared to be robust at $n \ge 200$ under a moderate departure from nornormality (e.g., distribution 2 and 3) and at $n \ge 500$ under the severely nonnormal conditions (e.g., distribution 4 and 5). Compared with the ADF, the SCALED was less affected by the degree of nonnormality and sample size. The ADF required $n \ge 5000$ to maintain a good control of Type I error, but the SCALED only needs $n \ge 200$ with samples drawn from moderately nonnormal population and $n \ge 500$ with the data departure from severe nonnormality. Additionally, the SCALED provided rejection rates within the robustness interval given $n \ge 500$, regardless of distribution types.

With respect to the bootstrap methods (e.g., bootstrap-M_o and bootstrap-M_A), both of them suffer from small

Table 1 Summary Statistics on the Empirical Type I Error Rates Across Different Conditions of Non-normality

						bootstrap- M_o							bootstrap- M_A						
Distribution type	n	ML	GLS	ADF	SCALED	200	300	500	1000	1500	2000	5000	200	300	500	1000	1500	2000	5000
1	100	.062	.055	**.310**	**.078**	**.080**	.075	**.085**	**.078**	.066	.068	.075	.074	**.085**	**.078**	.065	.075	**.089**	.060
(Skewness=0	200	.058	.050	**.215**	.060	.065	.053	.050	.048	.042	.045	.048	.061	.065	.046	.050	.045	.045	.042
Kurtoses=0)	500	.060	.062	.075	.058	.068	.050	.045	.045	.042	.045	.045	.060	.065	.045	.047	.046	.050	.048
	1000	.060	.068	.070	.065	.062	.055	.045	.045	.040	.042	.043	.058	.062	.048	.050	.049	.050	.040
	5000	.065	.060	.062	.060	.068	.058	.042	.040	.045	.045	.036	.065	.065	.050	.049	.050	.041	.035
2	100	**.275**	**.285**	**.330**	**.100**	**.078**	**.076**	.055	.075	**.076**	**.085**	.060	**.077**	**.076**	**.078**	.075	**.076**	**.077**	.074
(Skewness=0	200	**.242**	**.258**	**.285**	.070	.066	.056	.050	.048	.040	.049	.045	.067	.062	.049	.049	.045	.045	.050
Kurtoses=7)	500	**.200**	**.243**	**.215**	.069	.050	.058	.049	.045	.045	.045	.043	.060	.053	.040	.045	.040	.041	.045
	1000	**.201**	**.215**	**.175**	.063	.045	.053	.045	.045	.048	.040	.039	.065	.050	.042	.040	.041	.047	.040
	5000	**.195**	**.200**	.070	.064	.050	.055	.040	.045	.040	.039	.038	.060	.055	.040	.035	.040	.039	.039
3	100	**.280**	**.305**	**.342**	**.105**	.062	**.078**	**.080**	**.077**	.056	**.076**	.075	**.081**	**.076**	**.077**	.066	**.076**	.075	.074
(Skewness=2	200	**.355**	**.300**	**.305**	.070	.060	.058	.049	.046	.048	.046	.046	.075	.068	.050	.046	.044	.045	.045
Kurtoses=7)	500	**.370**	**.250**	**.180**	.068	.050	.055	.042	.035	.038	.040	.040	.063	.058	.040	.036	.035	.048	.035
	1000	**.310**	**.220**	**.088**	.065	.055	.054	.040	.045	.042	.038	.039	.058	.050	.043	.043	.045	.040	.038
	5000	**.200**	**.210**	.070	.064	.052	.050	.045	.040	.040	.039	.038	.055	.050	.041	.042	.045	.039	.035
4	100	**.418**	**.420**	**.350**	**.114**	**.084**	**.078**	.075	**.076**	**.079**	**.078**	**.078**	.071	**.082**	.074	.055	**.083**	.050	**.080**
(Skewness=0	200	**.415**	**.388**	**.305**	**.095**	.071	.061	.049	.048	.045	.044	.045	.066	.061	.049	.048	.046	.047	.051
Kurtoses=21)	500	**.305**	**.420**	**.185**	.068	.050	.053	.041	.045	.040	.035	.038	.065	.051	.045	.040	.041	.043	.040
	1000	**.310**	**.380**	**.095**	.061	.058	.050	.041	.035	.037	.040	.036	.061	.052	.040	.038	.039	.045	.037
	5000	**.300**	**.350**	.071	.058	.051	.052	.039	.038	.038	.036	.035	.060	.050	.045	.041	.045	.040	.039
5	100	**.450**	**.380**	**.480**	**.128**	**.085**	**.078**	.075	**.078**	**.076**	.060	.075	**.090**	**.076**	.066	.063	**.078**	.068	.055
(Skewness=3	200	**.420**	**.450**	**.345**	**.105**	.068	.066	.046	.046	.040	.045	.040	.071	.067	.046	.040	.040	.047	.043
Kurtoses=21)	500	**.440**	**.420**	**.220**	.072	.062	.050	.045	.042	.045	.039	.037	.063	.058	.045	.041	.038	.035	.036
	1000	**.380**	**.410**	**.105**	.070	.059	.055	.045	.040	.041	.038	.043	.061	.052	.048	.048	.049	.040	.041
	5000	**.350**	**.450**	.075	.062	.054	.059	.037	.042	.035	.047	.035	.060	.051	.040	.040	.038	.037	.035

Note: ML=maximum likelihood, GLS= generalized least squares, ADF= asymptotic distribution free, SCALED= Satorra-Bentler rescaled. Bold entries indicate the Type I error rates are not in the interval (.025, .075).

sample sizes (e.g., $n=100$) under both normal and nonnormal conditions. The rejection rates arbitrarily varied at $n=100$, implying that bootstrapping was unstable with small sample sizes. Nonetheless, both bootstrap methods maintained a good control of Type I error rates given adequate sample size (e.g., $n \geq 200$). To better understand the patterns of rejection rates of bootstrapping, two separate three-way ANOVA analyses for the bootstrap-M_o and bootstrap-M_A were conducted with an exclusion of $n=100$ (because bootstrapping with $n=100$ was unstable). In each univariate factorial analysis, there were 300 replications for five distributions crossed with four sample sizes crossed with seven bootstrapping replications ($300\times5\times4\times7=42000$ cases). Table 2 summaries Eta-squared values associated with each factorial analysis. Results illustrated that both first-order and second-order interactions were nonsignificant. Only a significant main effect of the number of bootstrapping (B) was identified. These results signify that the empirical Type I error rates of both bootstrap methods mainly depend on the number of bootstrapping, but distribution type and sample size were not influential factors for rejection rates. Furthermore, the number of bootstrapping $B \geq 500$ produced nonsignificant differences of rejections from a posteriori. With a closer inspection, most rejection rates were below .05 given $B \geq 500$; that is, $B \geq 500$ seemed to be irrelevant in terms of model rejections. Additionally, the bootstrap-M_o performed comparably with bootstrap-M_A in terms of their Type I error rates. Comparing both bootstrap methods with the SCALED, it appeared that all were less affected by distribution types. To achieve acceptable rejection rates, $n \geq 200$ was required for both bootstrap methods, but $n \geq 500$ was needed under severely nonnormal conditions for the SCALED.

To assess whether Type I error control was a function of distribution type (e.g., distribution1-5) and sample size (e.g., n=200, 300, 500, 1000, 5000), factorial analyses of variance were conducted for each method (ML, GLS, ADF, SACLED). Because the performance of both bootstrap methods greatly relied on the bootstrapped samples and they were not significantly influenced by distribution type as mentioned above, both bootstrap methods were excluded. In each factorial analysis, there were 300 replications for five distributions crossed with

Table 2 Partial Eta-squared of factorial analyses of variance for the bootstrap methods

	bootstrap- M_o	bootstrap- M_A
distribution	.019	.021
sample size	.011	.008
number of bootstrapping (B)	.287***	.235***
distribution × sample size	.012	.020
distribution × (B)	.016	.014
sample size × (B)	.011	.013
distribution × sample size × (B)	.015	.009
R^2 corrected model	.371***	.301***

*** p<.001

five sample sizes (300×5×5=7500 cases). In Table 3, results showed that about 70% of the variance in the empirical rejection rates was explained by first and second order effects for the ML and GLS. The explained variance for the ADF and SCALED was 47% and 29.5%, respectively. In addition, main effects and interaction effects were significant for the ML, GLS and ADF. It is suggested that the empirical Type I error rates varied as a function of distribution type and sample size for the ML, GLS and ADF. For the SCALED, however, only sample size had a significant effect on rejection rates.

Conclusions

This study was to investigate the performance of general and robust covariance structure analysis techniques under various conditions of multivariate nonnormalilty, with an emphasis on properly specified models. As is well-known, normal-theory methods (e.g., ML and GLS) are assumed that the fourth-order moments are equal to 0, which in turn fails to reflect the quality of a covariance structure model under conditions of nonnormality. The more severe the nonnormality, the worse control of Type I error rates. The present Monte Carlo simulation clearly demonstrates that both ML and GLS have Type I error rates as high as .35 and .45, respectively, under severely nonnormal distributions even with the largest sample size (n=5000). In practice, the applied researchers may use the asymptotic robust theory, such as ADF, to justify the use of normal-theory with nonnormal data (Browne, 1982). Theoretically, the ADF depends on computing the fourth-order moments of the measure variables. The

Table3 Partial Eta-squared of factorial analyses of variancefor the ML, GLS, ADF, SCALED

	ML	GLS	ADF	SCALED
distribution	.621***	.523***	.231***	.013
sample size	.218***	.225***	.318***	.211***
distribution × sample size	.255***	.214***	.212***	.019
R^2 corrected model	.715***	.698***	.470***	.295***

*** p<.001

fourth-order moments reflect the heavy or light tails. If data contain heavy tails, it provides unstable estimators especially with small sample sizes. This study, corresponding to the previous studies (e.g., Chou et al., 1991; Curran et al., 1996; Hu et al., 1992), reveals that the ADF suffers from small and moderate sample sizes, regardless of distribution types. For example, with sample sizes smaller than 200, it provides model rejection rates over .215 under normal conditions and model rejections ranging from .285 to .480 under nonnormal conditions. Even for a sample size of 1000, it still fails to have a good control of Type I error rates for all conditions of nonnormality. Only until $n = 5000$, acceptable rejection rates could be achieved. Compared with the ADF, the SCALED is computed on the basis of the model, estimation methods and the fourth-order moments. The SACLED, therefore, performs well overall, with exceptions of its tendency to overreject models at smaller sample sizes. Results signify that minimum sample size of the SCALED to get a good control of Type I error rates is 200 for moderately nonnormal conditions and 500 for severely nonnormal conditions. The reason for the superior performance of the SCALED over ADF may be due to the sample fourth-order moments. The SCALED uses a matrix computed from the sample fourth-order moments directly but the ADF utilizes the relevant matrix to be inverted. When the sample size is small, this inverse may not exist (Hu et al, 1992).

New findings from this study focus on the performance of bootstrap-M_A and bootstrap-M_o. Results illustrate that both bootstrap methods work well in various conditions of nonnormality, with unstable performance at $n = 100$. It means that the minimum sample size of 200 is required for bootstrapping approaches. The minimum sample size needed here was

slightly less than that (e.g., n=300) in the study of Ichikawa and Konishi (1995). Furthermore, neither distribution type nor sample size affects the rejection rates for both bootstrap methods given $n \geq 200$. The rejection rates mainly depend on the number of bootstrapped samples (B). When $B \geq 500$, the negligible effects on rejection rates are identified. As such, it may suggest that the minimum B required to provide a good control of Type I error rates may be $B = 500$. Although the bootstrap- M_A and bootstrap- M_o are equivalent in terms of their Type I error rates, there still exists other puzzles concerned with both bootstrapping approaches. For instance, are they equivalent in terms of power or standard error? Are they equivalent under more complicated models (e.g., models with many parameters or different parameters values)? Would the minimum B and sample size greatly depend on model complexity? Studies for such a practice are further needed.

When considering the overall performance of all methods, it is evidently that two bootstrap methods beat the ML, GLS and ADF under properly specified models, regardless of distribution types of nonnormality. The SCALED, with slightly high but acceptable rejection rates, performs as well as bootstrap methods given $n \geq 500$. It implies that the resampling-based methods (i.e., bootstrapping) may be more conservative in its control of Type I error rates. Importantly, the empirical Type I error rates perform as a function of distribution type and sample size for the ML, GLS and ADF. However, for the SCALED, the rejection rates are merely influenced by sample sizes, and for bootstrap method, the rejection rates are only affected by the number of bootstrapped samples.

These findings yield some important implications for the practitioners. First, the majority of data investigated in psychological or behavioral research conceivably fail to follow normal distributions, especially for several significant areas of studies such as depression, abnormality and psychopathology. Therefore, the data should be checked for potential violations of multivariate normality assumption. The EQS computer program is especially useful in this regard. When handling the data containing outliers or heavy tails, the researcher is advised to use one of the distribution-free methods for parameter estimation. Second, many researchers seldom suspect

the quality of their data due to a false sense of generalizability of the ML method. They believe the validity of normal theory-based methods with large-sample nonnormal data (i.e., asymptotic robustness theory). Regretfully, the situations for asymptotic robustness rely on the data as well as the model and it is not known how to verify these conditions in practice (Yuan, Bentler, & Zhang, 2005). As such, it is inappropriate to heedlessly trust that the data and model satisfy these conditions. Third, compared with ML or GLS, although the robust methods proposed here minimize the effects of nonnormlaity, one suffers from its weaknesses. The ADF is not recommended since it is essentially unusable in small to intermediated sized samples. The SCALED and bootstrap methods are regarded as better alternatives. Thus, it is suggested that the programs of choices for multivariate nonnormal data are EQS, MPLUS and AMOS, with their implementations of the SCALED or bootstrap methods (EQS and MPLUS provide the SCALED procedures; EQS and AMOS yield the bootstrap procedures). Importantly, when we apply bootstrap methods, three cautions should be noted, including (a) lacking independent and identical distributed property of observations, (b) suffering from statistical property of efficiency and (c) failing with small sample sizes (Yung & Bentler, 1996). Furthermore, the present study demonstrates the robustness of the SCALED, bootstrap-M_A and bootstrap-M_o only for complete data. However, many real data sets are nonnormal and incomplete. The degree of nonnormality and proportion of missing data could ruin the robustness properties (e.g., Savalei, 2008) and as a result the researchers should be more careful to handle nonnormal as well as incomplete datasets.

Any Monte Carlo simulation study could be criticized for its limited generalizability since it is infeasible to design models as complex as the real world. The present study evaluates the performance of six methods under various conditions of nonnormality in terms of their Type I error control rates, with a particular emphasis on properly specified models. The dependent variable investigated here only focuses on the Type I error rates. Other issues such as bias, efficiency, power, and standard error could be further considered. Additionally, the degree of model misspecification, model complexity (i.e., number of parameter or parameter values)

and proportion of missing data can be critical concerns in further studies.

Reference

Anderson, T. W., & Amemiya, Y. (1988). The asymptotic normal distribution of estimators in factor analysis under general conditions. *Annuals of Statistics, 16*, 759-771.

Bentler, P. M. (1988). Causal modeling via structural equation systems. In J. R. Nesselroade & R. B. Cattell (Eds.), *Handbook of multivariate experimental psychology* (pp. 317-335). New York: Plenum.

Bentler, P. M. (1995). *EQS Structural equations program manual*. Encino, CA: Multivariate Software.

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

Beran, R., & Srivastava, M. S. (1985). Bootstrap tests and confidence region for functions of a covariance matrix. *Annuals of Statistics, 13*, 95-115.

Bollen, K. A. (2002). Latent variables in psychology and the social sciences. *Annual Review of Psychology, 53*, 605-634.

Bollen, K. A., & Arminger, G. (1991). Observational residuals in factor analysis and structural equation models. In P. V. Marsden (Ed.), *Sociological Methodology* (pp. 235-262). Oxford, England: Blackwell.

Bollen, K. A., & Stine, R. A. (1990). Direct and indirect effects: Classical and bootstrap estimates of variability. In C. C. Clogg (Ed.), *Sociological Methodology* (pp. 115-140). Oxford, England: Blackwell.

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

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

Breckler, S. J. (1990). Applications of covariance structure modeling in psychology: Cause for concern? *Psychological Bulletin, 107*(2), 260-273.

Browne, M. W. (1982). Covariance structures. In D. M. Hawkins (Ed.), *Topics in multivariate analysis* (pp. 72-141). Cambridge, England: Cambridge University.

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

Campbell, N. A. (1980). Robust procedures in multivariate analysis I: Robust covariance estimation. *Applied Statistics, 29*, 231-237.

Chatterjee, S., & Yilmaz, M. (1992). A review of regression diagnostics for behavioral research. *Applied Psychological Measurement, 16*, 209-227.

Chou, C. P., Bentler, P. M., & Satorra, A. (1991). Scaled test statistics and robust standard errors for nonnormal data in covariance structure analysis: A monte carlo study. *British Journal of Mathematical and Statistical Psychology, 44*, 347-357.

Curran, P. J., West, S. G., & Finch, J. F. (1996). The robustness of test statistics

to non-normality and specification error in confirmatory factor analysis. *Psychological Methods, 1*, 16-29.

Devlin, S. J., Gnanadesikan, R., & Kettenring, J. R. (1981). Robust estimation of dispersion matrices and principal components. *Journal of the American Statistical Association, 76*, 354-362.

Efron, B. (1979). Bootstrap methods: Another look at the jackknife. *The Annals of Statistics, 7*(1), 1-26.

Hampel, F. R., Ronchetti, E. M., Rousseeuw, P. J., & Stahel, W. A. (1986). *Robust statistics: The approach based on influence functions*. New York: Wiley.

Hu, L.-T., & Bentler, P. M. (1995). Evaluating model fit. In R. Hoyle (Ed.), *Structural equation modeling: Issue, concepts and applications* (pp. 76-99). Newbury Park, CA: Sage.

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

Huber, P. J. (1981). *Robust statistics*. New York: Wiley.

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

Jöreskog, K. G. (1993). Test structural equation models. In K. A. Bollen & J. S. Long (Eds.), *Testing structural equation models* (pp. 294-316). Newbury Park, CA: Sage.

Jöreskog, K. G., & Goldberger, A. S. (1972). Factor analysis by generalized lease squares. *Psychometrika, 37*, 243-260.

Kano, Y., Berkane, M., & Bentler, P. M. (1990). Covariance structure analysis with heterogenous kurtosis parameters. *Biometrika, 77*, 575-585.

Lee, S. Y., & Wang, S. J. (1996). Sensitivity analysis of structural equation models. *Psychometrika, 61*, 93-108.

Micceri, T. (1989). The unicorn, the normal curve and other improbable creatures. *Psychological Bulletin, 105*(1), 156-166.

Muthén, B., & 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.

Rousseeuw, P. J., & van Zomeren, B. C. (1990). Unmasking multivariate outliers and leverage points. *Journal of the American Statistical Association, 85*, 633-639.

Satorra, A., & Bentler, P. M. (1988). Scaling corrections for chi-square statistics in covariance structure analysis. In *Proceedings of the Business and Economics Sections* (pp. 308-313). Alexandria, VA: American Statistical Association.

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

Satorra, A., & Bentler, P. M. (1991). Goodness-of-fit test under IV estimation: Asymptotic robustness of a NT test statistic. In R. Gutierrez & M. J. Valderrama (Eds.), *Applied stochastic models and data analysis* (pp. 555-567). Singapore: World Scientific.

Satorra, A., & Bentler, P. M. (1994). Corrections to test statistic and standard errors

in covariance structure analysis. In A. Von Eye & C. C. Clogg (Eds.), *Analysis of latent variables in development research* (pp. 399-419). Newbury Park. CA: Sage.

Savalei, V. (2008). Is the ML Chi-Square ever robust to nonnormality? A cautionary note with missing data. *Structural Equation Modeling, 15*, 1-22.

Staudte, R. G., & Sheather, S. J. (1990). *Robust estimation and testing*. New York: Wiley.

Stine, R. A. (1989). An introduction to bootstrap methods: Examples and ideas. *Sociological Methods and Research, 8*, 243-291.

Tyler, D. E. (1983). Robust and efficiency properties of scatter matrices. *Biometrika, 70*, 411-420.

Vale, C. D., & Maurelli, V. A. (1983). Simulation multivariate non-normal distributions. *Psychometrika, 48*, 465-471.

Wilcox, R. R. (1997). *Introduction to robust estimation and hypothesis testing*. San Diego, CA: Academic Press.

Wilcox, R. R. (2003). *Applied contemporary statistical methods*. New York: Academic Press.

Yuan, K.-H., & Bentler, P. M. (1998a). Structural equation modeling with robust covariances. In A. E. Raftery (Ed.), *Sociological methodology* (pp. 363-396). Boston: Blackwell.

Yuan, K.-H., & Bentler, P. M. (1998b). Robust mean and covariance structure analysis. *British Journal of Mathematical and Statistical Psychology, 51*, 63-88.

Yuan, K.-H., & Bentler, P. M. (2001). Effect of outliers on estimators and tests in covariance structure analysis. *British Journal of Mathematical and Statistical Psychology, 54*, 161-175.

Yuan, K.-H., Bentler, P. M., & Chan, W. (2004). Structural equation modeling with heavy tailed distributions. *Psychometrika, 69*(3), 421-436.

Yuan, K.-H., Bentler, P. M., Zhang, W. (2005). The effect of skewness and kurtosis on mean and covariance structure analysis. *Sociological Methods & Research, 34*, 240-258.

Yuan, K.-H., & Hayashi, K. (2003). Bootstrap approach to inference and power analysis based on three test statistics for covariance structure models. *British Journal of Mathematical and Statistical Psychology, 53*, 93-110.

Yuan, K.-H., & Hayashi, K. (2006). Standard errors in covariance structure models: Asymptotics versus bootstrap. *British Journal of Mathematical and Statistical Psychology, 59*, 397-417.

Yuan, K.-H., Hayashi, K., & Yanagihara, H. (2007). A class of population covariance matrices in the bootstrap approach to covariance structure analysis. *Multivariate Behavioral Research, 42*, 261-281.

Yung, K.-H., & Bentler, P. M. (1994). Bootstrap-corrected ADF test statistics in covariance structure analysis. *British Journal of Mathematical and Statistical Psychology, 47*, 63-84.

Yung, K.-H., & 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: Issue and techniques* (pp. 195-226).

Mahwah, NJ: Lawrence Erlbaum Associates, Inc.