University of California
Los Angeles

Evaluation of the Robustness of Modified Covariance Structure Test Statistics

A thesis submitted in partial satisfaction
of the requirements for the degree
Master of Science in Statistics

by

Xiaoxiao Tong

2012

Abstract of the Thesis

Evaluation of the Robustness of Modified Covariance Structure Test Statistics

by

Xiaoxiao Tong

Master of Science in Statistics

University of California, Los Angeles, 2012

Professor Frederic R Paik Schoenberg, Chair

Problems about whether a hypothesized covariance structure model is an appropriate representation of the population covariance structure of multiple variables can be addressed using goodness-of-fit testing in structural equation modeling. Many test statistics and their extensions have been developed for various specific conditions and some of them have been extensively used in practice. However, their expected performances might break down under violations of multivariate normality or sufficiently large sample sizes. This paper evaluates the robustness of four modified goodness-of-fit test statistics $T_{SB}(new)$, T_{MV}, T_{YB} and T_F in SEM. Monte Carlo simulation demonstrates that the robustness of covariance structure statistics vary as a function of the correctness of the model as well as distributional characteristics of observed data. Suggestions for application of these modified test statistics are given after taking both the literature and current simulation result into account. A surprising result was the failure of T_{MV}, the Satorra-Bentler mean-scaled and variance-adjusted test statistic, to perform correctly even asymptotically in one condition.

The thesis of Xiaoxiao Tong is approved.

Peter M Bentler

Hongquan Xu

Nicolas Christou

Frederic R Paik Schoenberg, Committee Chair

University of California, Los Angeles

2012

To my mother and father . . .

who teach me how to count as a start

Table of Contents

List of Tables

Acknowledgments

First and foremost, I would like to show my deepest gratitude to my supervisor, Professor Peter M. Bentler, a respectable, knowledgable and responsible scholar, who has provided me with valuable guidances in every stage of writing this thesis. Without his instructions and help, I could not have completed this. His keen and vigorous academic observation inspires me not only in this thesis but also in my future development.

I would like to express my sincere gratitude to other professors in the department of statistics for all their kindness and help. They have taught me a lot during the past year and brought me great insights into statistics. I would also like to thank Glenda Jones, our lovely Student Affairs Officer, who has helped me out of many problems I encountered during my graduate life at UCLA. Last but not least, I wish to thank my parents, all my friends and classmates for their encouragement and support.

CHAPTER 1

Introduction

Covariance structure analysis in structural equation modeling has been used extensively in psychological, social and behavioral sciences. Goodness-of-fit test statistics by which to assess the adequacy of hypothesized covariance structure models have been studied over the decades, and their performances under various distributional conditions across different sample sizes have been examined.

Classical goodness-of-fit testing is based on the assumption that the test statistics employed are asymptotically chi-square distributed, but this property may not hold when the factors and errors and hence the observed variables are nonnormally distributed. Even when the factors and errors are normally distributed in the population, the performance of test statistics in small sample sizes may still be compromised (Hu, Bentler and Kano, 1992; Curran, West, & Finch, 1996). For example, the most widely utilized test statistic, the classical likelihood ratio statistic T_{ML} based on normal theory maximum likelihood (ML) estimation, has been verified in many simulation studies to yield quite distorted conclusions about model adequacy under violations of multivariate normality. The well-known Satorra-Bentler's (1994) scaled test statistics T_{SB}, as well the mean scaled and variance adjusted test statistics T_{MV} were thus developed to be robust to nonnormaity, and have been shown to perform well under such conditions (Yuan and Bentler, 2010; Tong and Bentler, in press). These two test statistics are derived from a linear combination of quadratic normal variates, whose coefficients are the eigenvalues of a product matrix involved in the calculations of model fitting. The comparative performance of T_{SB} and T_{MV} is mainly affected by these eigenvalues and their associated coefficient of variation (Yuan and Bentler, 2010). Tong and Bentler (in press) suggested to use T_{SB} when little is known about about the distribution of observed data, but preferably use T_{MV} at a small or moderate sample sizes when normality or asymptotic robustness assumptions hold. However, they also noted a failure of T_{MV} in one condition. A newly proposed extension to the normal theory statistic T_{ML} by Lin and Bentler (2012), the mean scaled and skewness adjusted test statistic T_{MS}, was developed to improve its robustness under small sample sizes, but failed to perform ideally as expected in a recent simulation study (Tong and Bentler, in press). It is suggested that T_{MS} could be considered when researchers want to be more conservative in confirming the fit of a model, but with limitation to normally distributed data.

An alternative approach to be applied under nonnormality is the classical asymptotically distribution free (ADF) method and its associated test statistic T_{ADF} proposed by Browne (1984). It is theoretically elegant but empirically unsatisfactory. Unlike the Satorra-Bentler scaled test statistics, which attempt to center the statistic so that its mean will be closer to that of a chi-square variate, T_{ADF} is precisely distributed as an asymptotic chi-square variate. However, unreasonably large sample sizes are required for ADF test statistic to exhibit such an advantage; otherwise it will break down spectacularly (Hu, Bentler and Kano, 1992; Curran, West, & Finch, 1996). A relatively unknown residual-based ADF test statistic T_B derived by Browne (1984) can be applied to any consistent estimators with no specific distribution assumptions of the observed data. However, Yuan and Bentler (1998) showed that the residual-based ADF test statistic, like the classical ADF statistic, requires a very large sample size to give reliable inference. The Yuan-Bentler residual-based test statistic T_{YB} (Yuan and Bentler, 1998) was then developed to improve the performance of T_B for small samples under general distributional conditions, and has shown remarkably better performance under such conditions (Bentler and Yuan, 1999). Another more radical modification of the residual-based ADF statistic, the Yuan-Bentler residual-based F-statistic T_F (Yuan and Bentler, 1998), was designed to take sample size into account more adequately. Different from the above test statistics, T_F is evaluated by reference to an F-distribution instead of a χ^2 distribution. Simulation studies have shown that the modified F-statistic outperforms various test statistics with asymptotic χ^2 distribution at the smallest sample sizes (Bentler and Yuan, 1999), yet the test statistic has not been employed as much as T_{ML} or the Satorra-Bentler scaled test statistic T_{SB} in practice.

The purpose of this paper is to compare the robustness of several above modified test statistics and address their relative applications. Since T_{ML}, T_{ADF} and T_B have been extensively studied and their performances are easy to break down conditionally, this paper will focus on the relatively unknown test statistics T_{MV}, T_{YB} and T_F. Since T_{SB} has been reported to perform stably and ideally under various conditions, it is selected as a benchmark in the following study. The performances of four goodness-of-fit test statistics, namely T_{SB}, T_{MV}, T_{YB} and T_F, are evaluated under violations of normality across various sample sizes. Their powers are examined under a correct structural model as well as under a misspecified model. Tong and Bentler (in press) found out that a simple modification to T_{SB} for the case of sample size smaller than degrees of freedom, $T_{SB(New)}$, performed better than the standard version of the scaled statistic in each of the conditions studied. Hence, T_{SB} will be replaced by $T_{SB(New)}$ when the degrees of freedom exceeds sample size in the following study. Headrick's (2002; Headrick & Swailowsky, 1999) relatively unstudied methodology for generating nonnormal data is used due to its ability to generate a wider range of skew and kurtosis as well as control higher order moments than the more standard Fleishman (1978) and Vale and Maurelli (1983) procedure. The test statistics are briefly reviewed in Chapter 2, and empirical performances of these test statistics will be studied in Chapter 3 and 4.

CHAPTER 2

Test Statistics

2.1 Covariance Structure Analysis

Suppose $X = (X_1, X_2, \cdots, X_p)$ is a stochastic p-vector of observed variables with population covariance matrix Σ. Let $X_i = (x_{i1}, x_{i2}, \cdots, x_{ip}) \quad i = 1, 2, \cdots, N = n + 1$ be a sample from X with sample covariance matrix S, an unbiased estimator of Σ. Covariance structure analysis techniques test the hypothesis that Σ, can be expressed as a matrix valued function, $\Sigma(\theta)$, of a $q-$dimensional parameter vector θ at some value θ_0. This can be written as $H_0 : \Sigma = \Sigma(\theta_0)$. The goodness-of-fit test statistics used in covariance structure analysis are generally formulated as a function of the discrepancy of the sample covariance matrix, S, from the structured covariance matrix based on a specified model, $\Sigma(\theta)$. Assume $F(S, \Sigma(\theta))$ is a scalar valued discrepancy function of S from $\Sigma(\theta)$, then parameter estimates are obtained by minimizing $F(S, \Sigma(\theta))$. Many goodness-of-fit test statistics can be expressed as $T = c(N - 1)\hat{F}$, where $\hat{F}$ is the minimum of $F(S, \Sigma(\theta))$, N is the number of samples, and c is a scaling factor. When the model assumptions hold, the test statistics are generally distributed as an asymptotic χ^2 with $p(p + 1)/2 - q$ degrees of freedom, where p is the number of variables and q is the number of free parameters. The residual-based test statistics do not take the usual form, but are computed based on the distribution of the residuals $(S - \Sigma(\hat{\theta}))$, where $\hat{\theta}$ is the value of θ that minimizes the discrepancy function $F(S, \Sigma(\theta))$. These test statistics do not require specific distributions to have an asymptotic χ^2 distribution, or a related F distribution.

2.2 Mean Scaled and Moment Adjusted Test Statistics

The discrepancy function $F(S, \Sigma(\theta))$ typically takes the form of normal-theory maximum-likelihood (ML) discrepancy function

$$F_{ML}(\theta) = \log|\Sigma(\theta)| + tr(S\Sigma^{-1}(\theta)) - \log|S| - p \tag{2.1}$$

and the generalized least squares function

$$F_{GLS}(\theta) = (s - \sigma(\theta))' V_n (s - \sigma(\theta)) \tag{2.2}$$

where p is the number of observed variables. Let $vech(\cdot)$ be an operator which transforms a symmetric matrix into a vector by stacking the nonduplicated elements of the matrix, $s = vech(S), \sigma(\theta) = vech[\Sigma(\theta)]$. Then s and $\sigma(\theta)$ are $p^* = p(p+1)/2$ dimensional vectors. Under general conditions it follows from the multivariate central limit theorem (Anderson, 2003) that

$$\sqrt{n}(s - \sigma(\theta)) \xrightarrow{d} N(0, \Gamma) \tag{2.3}$$

where Γ is the asymptotic covariance matrix of s. Typical elements of Γ are given by

$$\gamma_{ij,kl} = \sigma_{ijkl} - \sigma_{ij}\sigma_{kl} \tag{2.4}$$

where the multivariate product moment for four variables z_i, z_j, z_k and z_l is defined as

$$\sigma_{ijkl} = E(z_i - \mu_i)(z_j - \mu_j)(z_k - \mu_k)(z_l - \mu_l) \tag{2.5}$$

and σ_{ij} is the usual sample covariance. Let $\dot{\sigma}(\theta) = \partial\sigma(\theta)/\partial\theta$ denote the $p^* \times q$ Jacobian matrix. Then there exists a full column rank $p^* \times (p^* - q)$ matrix $\dot{\sigma}_c(\theta)$ whose columns are orthogonal to those of $\dot{\sigma}(\theta)$. To ensure that the model is identified at $\hat{\theta}$, we assume that $\dot{\sigma}(\theta)$ has full rank in a neighborhood of $\hat{\theta}$, and denote $\dot{\sigma} = \dot{\sigma}(\hat{\theta})$. Under multivariate normality, let $W = 2^{-1}D_p'(\Sigma^{-1} \otimes \Sigma^{-1})D_p$, where D_p is a $p^2 \times p^*$ duplication matrix (Magnus and Neudecker, 1988) and

$$U = W - W\dot{\sigma}(\dot{\sigma}'W\dot{\sigma})^{-1}\dot{\sigma}'W \tag{2.6}$$

Then the goodness-of-fit chi-square statistic is given as:

$$T_{ML} = n\hat{F}_{ML} \tag{2.7}$$

where $\hat{F}_{ML}$ is the minimum of (2.1) evaluated at the maximum likelihood estimate of parameters. Under the assumption of multivariate normality and the null hypothesis, T_{ML} has a χ^2 distribution with degrees of freedom $d = p^* - q$. This also holds asymptotically under specific nonnormal conditions (see e.g., Savalei, 2008). For example, in a confirmatory factor analysis, when all factors are independently distributed and the elements of the covariance matrices of common factors are free parameters, T_{ML} can be insensitive to violations of the normality assumption. More generally, the distribution of T_{ML} can be characterized by a linear combination of independent chi-square variates, each with one degree of freedom:

$$T_{ML} \xrightarrow{d} \sum_{i=1}^{d} \lambda_i z_i^2 \tag{2.8}$$

where $z_i \sim N(0, 1)$ independently and λ_i are the non-zero eigenvalues of $U\Gamma$. Since

$$E[\sum_{i=1}^{d} \lambda_i z_i^2] = \sum_{i=1}^{d} \lambda_i = trace(U\Gamma) \tag{2.9}$$

Satorra and Bentler (1988) proposed a scaled chi-square statistic:

$$T_{SB} = T_{ML}/k \tag{2.10}$$

where $k = trace(U\Gamma)/d$ is a scaling constant that corrects T_{ML} so that the sampling distribution of T_{SB} at least matches the first moment of the nominal chi-square distribution. The scaling constant k is an estimate of the average of the nonzero eigenvalues of $U\Gamma$, and $U\Gamma$ should be replaced by their consistent estimators $\hat{U}$ and $\hat{\Gamma}$ for calculation. For normal theory based maximum likelihood estimation, a consistent estimator of Γ is given by S_Y, the sample covariance matrix of $Y_i = vech[(X_i - \bar{X})(X_i - \bar{X})']$. However, when the sample size is smaller than the degrees of freedom ($N < d$), (2.10) is not the correct formula since there will not be d nonzero eigenvalues. Hence, when $N < d$, Tong and Bentler (in press) proposed the use of $k = trace(U\Gamma)/N$ instead. This new Satorra-Bentler scaled chi-square statistic is thus given by:

$$T_{SB(New)} = T_{ML}/k \tag{2.11}$$

where $k = trace(U\Gamma)/\min(d, N)$, and $T_{SB(New)}$ is referred to a χ^2 distribution with $\min(d, N)$ degrees of freedom. A more sophisticated correction, the Satorra-Bentler mean scaled and variance adjusted statistic is given as:

$$T_{MV} = vT_{ML}/trace(U\Gamma) \tag{2.12}$$

where $v = [trace(U\Gamma)]^2/trace[(U\Gamma)^2]$. T_{MV} involves both scaling the mean and a Saitterwarthe second moment adjustment of the degrees of freedom (Saitterwarthe, 1941), and the new reference distribution is a central χ^2 with degrees of freedom v. The newly proposed mean scaled and skewness adjusted statistic by Lin and Bentler (2012) is defined as:

$$T_{MS} = v^* T_{ML}/trace(U\Gamma) \tag{2.13}$$

where $v^* = trace[(U\Gamma)^2]^3/trace[(U\Gamma)^3]^2$ is a function of the skewness of T_{ML}. In addition to scaling the mean as in T_{SB} and T_{MV}, T_{MS} adjusts the degrees of freedom such that asymptotically, the quadratic form of T as in (2.8) has the same skewness with a new reference distribution $\chi^2(v^*)$. Simulation study by Tong and Bentler (in press) on T_{MS} indicates that T_{MS} may downwardly overcorrect T_{ML} and cannot be trusted in model testing when data is non normally distributed. The potential of T_{MS} under multivariate normality in small samples needs to be further studied.

2.3 Residual-Based Test Statistics

The original residual-based test statistics T_B developed by Browne (1982, 1984) enjoys a theoretical advantage: if the sample size is large enough, its distribution is fully known. As denoted above, the statistic is defined as following for the estimate $\hat{\theta}$:

$$T_B(\hat{\theta}) = n\hat{e}'\dot{\sigma}_c(\hat{\theta})[\dot{\sigma}_c'(\hat{\theta})S_Y\dot{\sigma}_c(\hat{\theta})]^{-1}\dot{\sigma}_c'(\hat{\theta})\hat{e} \tag{2.14}$$

where $\hat{e} = s - \sigma(\hat{\theta})$ is the discrepancy between the data and the model estimated by any consistent estimator. $T_B(\hat{\theta})$ is asymptotically distributed as the χ^2 distribution with $(p^* - q)$ degrees of freedom, regardless of the distributional characteristics of observed variables as well the estimation method employed. It is also worth noticing that the value of $T_B(\hat{\theta})$ does not depend on the choice of $\dot{\sigma}_c(\hat{\theta})$, even though the orthogonal complement matrix $\dot{\sigma}_c(\theta)$ is not unique. ML estimator will be used for T_B in this paper. Since T_B requires extremely large sample size to be reliable, Yuan and Bentler (1998) proposed the modified residual-based test statistics T_{YB}. The idea originated from regression literature, where the cross-products of model residuals are used for estimating asymptotic covariances and standard errors (Bentler and Yuan, 1999). For a consistent estimate $\hat{\theta}$, Γ can be estimated, except for S_Y, through the following decomposition:

$$\hat{\Gamma} = \frac{1}{n}\sum_{i=1}^{N}[Y_i - \sigma(\hat{\theta})][Y_i - \sigma(\hat{\theta})]' = S_Y + \frac{N}{n}[\bar{Y} - \sigma(\hat{\theta})][\bar{Y} - \sigma(\hat{\theta})]' \qquad (2.15)$$

Replacing S_Y in (2.14) by $\hat{\Gamma}$, the Yuan-Bentler residual-based statistic is given by:

$$T_{YB}(\hat{\theta}) = T_B(\hat{\theta})/[1 + NT_B(\hat{\theta})/n^2)] \qquad (2.16)$$

T_{YB} also asymptotically follows the χ^2 distribution with $(p^* - q)$ degrees of freedom. But as $T_{YB}(\hat{\theta}) < T_B(\hat{\theta})$ for any consistent estimate $\hat{\theta}$, the problem of over rejection with T_B is expected to be improved by T_{YB}. Simulation study by Fouladi (2000) has shown that the Yuan-Bentler residual-based test statistics dramatically outperforms other distribution-free test statistics in covariance structure analysis, but is considered to be consistently conservative when compared with the almost equally powerful Satorra-Bentler scaled test statistic. No current studies have compared the performance of T_{YB} and T_{MV} under violations of normality, and this will be covered in the following sections. Inspired by the well-known Hotelling's T^2 statistic, Yuan and Bentler (1998) further proposed to use the Hotelling's T^2 distribution to approximate that of T_B instead of a chi-square. This leads to the Yuan-Bentler residual-based F-statistic:

$$T_F(\hat{\theta}) = [N - (p^* - q)]T_B(\hat{\theta})/[n(p^* - q)] \qquad (2.17)$$

which is referred to an F-distribution with degrees of freedom $(p^* - q, N - (p^* - q))$. T_F is also asymptotically equivalent with T_B, but its performance is very likely to differ from that of T_B for finite samples. One common limitation of the above residual-based test statistics T_B, T_{YB} and T_F is that they all require a sample size as large as $p^* - q + 1$. This is due to the fact that the $p^* - q$ square matrix $[\dot{\sigma}_c'(\hat{\theta}) S_Y \dot{\sigma}_c(\hat{\theta})]$ has to be invertible in order to compute T_B, and consequently T_{YB} and T_F.

In Section 3 and 4, four goodness-of-fit test statistics, the Satorra-Bentler scaled test statistic T_{SB}, the Satorra-Bentler mean scaled and variance adjusted test statistic T_{MV}, the Yuan-Bentler residual-based test statistic T_{YB} and the Yuan-Bentler residual-based F-statistic T_F, are examined under violations of multivariate normality across small to large sample sizes through Monte Carlo simulations. Their performances are judged by the statistical mean, variance (standard error), Type I error control and empirical power in rejecting a misspecified model.

CHAPTER 3

Simulation Method

3.1 Confirmatory Factor Analysis

The confirmatory factor model is specified as

$$X = \Lambda\eta + \epsilon \tag{3.1}$$

where X is a vector of observed indicators that depends on Λ, a common factor loading matrix, η is a vector of latent factor scores (common factors) and ϵ is a vector of unique errors (unique factors). Typically, we assume that η is normally distributed and uncorrelated with ϵ. Hence, the restricted covariance structure of X is:

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

where Φ is the covariance matrix of the latent factors and Ψ is a diagonal matrix of variances of errors. Since the observed indicators are a function of parameters in the factor analytic model, nonnormality in observed indicators is an implied consequences of nonnormality in the distributions of factors and errors.

In this study, a confirmatory factor model with 15 observed variables and 3 common factors is used to generate a model-based simulation. A simple structure of Λ is used where each set of five observed variables load onto a single factor with loadings of $\lambda = (0.7, 0.7, 0.75, 0.8, 0.8)$ respectively, as shown in (3.3). Under each condition, the common and unique factors are generated using Headrick's fifth-order transformation (Headrick, 2002), and then the 15 observed variables are generated by a linear combination of these factors.

$$\Lambda^T = \begin{pmatrix} \lambda & 0 & 0 \\ 0 & \lambda & 0 \\ 0 & 0 & \lambda \end{pmatrix} \quad (3.3)$$

After generation of the population covariance matrix Σ, random samples of a given size from the population are taken. In each sample, the parameters of the model are estimated and the above four test statistics are computed by calling EQS using the REQS function in R (Mair, Wu, & Bentler, 2010) and specifying METHOD = ML, ROBUST in EQS. In estimation, the factor loading of the last indicator of each factor is fixed for identification at 0.8, and all the remaining nonzero parameters are free to be estimated. In this case, $p^* = 15 \times 16/2 = 120, q = 33$ (free parameters include 12 coefficients, 15 variances of the unique factors, 3 variances of the common factors and 3 corresponding covariances) and thus the degrees of freedom $d = p^* - q = 87$. The behavior of T_{SB}, T_{MV}, T_{YB} and T_F are observed at sample sizes of 50, 100, 250, 500, 1,000, 2,500 and 5,000. Particularly, when $N = 50 < d = 87$, the behavior of $T_{SB(New)}$ is also observed while T_{YB} and T_F can not be computed as indicated in Section 2.3. At each sample size, 1,500 replications are drawn from the population. A statistical summary of the mean value and standard error of T under the confirmatory factor analysis model across the 1,500 replications, and the empirical rejection rate (Type I Error) at significance levels of $\alpha = 0.05$ on the basis of the assumed χ^2 or F distribution, are reported in Tables 4.1- 4.3. An ideal type I error rate should approach 5% rejection of the null hypothesis, with a deviation of less than $2[(.05)(.95)/1500]^{0.5} =$ .01125, resulting in an acceptable 95% confidence interval $[0.0387, 0.0613]$.

To measure the empirical power of these test statistics, a misspecified model with an additional path from η_1 to y_6 is used for hypothesis testing. The loading of this path is fixed at 0.8 in estimation. The observed variables are still generated under the correct model, but are then analyzed under the incorrectly specified model. The empirical power, reported in the fourth row for each cell in Tables 4.1- 4.3, is defined as the proportion of rejections of the null hypothesis for convergent simulated trials. A high rejection rate typically implies ideal performance of the test statistic, but this is not the case when simultaneously a high type I error rate exists (e.g., larger than 0.0613).

3.2 Data Generation

Three different conditions of distributions of factors and errors are simulated to examine the robustness of the above test statistics, and are identical to those used in Tong and Bentler (in press). In Condition 1, both common and unique factors are identically independently distributed as $N(0, 1)$, resulting in a multivariate normal distribution of the observed variables. This Condition is designed to perform as a benchmark to see whether these test statistics can behave as expected at least under multivariate normality.

Condition 2 is designed to be consistent with asymptotic robustness theory, where the common and unique factors are independently generated nonnormal distributions. The common factors are correlated with specified first six moments and intercorrelations as in Table 3.1, while the unique factors are independent with arbitrarily chosen first six moments. As noted in Tong and Bentler (in press), T_{ML} performs at least as well as T_{SB} under Conditions 1 and 2, and gives a slightly better Type I error rate at small and moderate sample sizes. Furthermore, their simulation study has shown that under the first two conditions, T_{MV} significantly outperforms T_{ML} and T_{SB} at small and moderate sample sizes, in terms of the frequency of rejecting the null hypothesis under the correct model. Therefore, Condition 2 is kept in this paper to evaluate the performances of residual-based test statistic under the asymptotic robustness theory.

In Condition 3, based on the distributions in Condition 2, the factors and error variates are divided by a random variable $Z = [\chi^2(5)]^{1/2}/\sqrt{3}$ that is distributed independently of the original factors and errors. This division results in the dependence of factors and errors, even though they remain uncorrelated. Because of the dependence, asymptotic robustness of normal-theory statistics is not to be expected under Condition 3. This is designed to examine the robustness of the test statistics under general violations of multivariate normality. Under the model $\Sigma(\theta)$, the degrees of freedom is

Table 3.1: Specified Distributions of Factors ($\mu = 0, \sigma^2 = 1$)

	Skew	Kurtosis	Fifth	Sixth	Correlations		
η_1	0	-1	0	28	1.0	0.3	0.4
η_2	1	2	4	24		1.0	0.5
η_3	2	6	24	120			1.0

$d = p^* - q = 87$. According to asymptotic robustness theory, we expect the normal-theory based test statistics to be valid for nonnormal data in Condition 2, in addition to the standard normal data in Condition 1. Regardless of the three types of distributions and conditions considered, the anticipated means of T_{SB}, T_{MV} and T_{YB} are 87 since they are asymptotically distributed as the χ^2 with degrees of freedom 87. Particularly, when $N < d$, the expected mean of $T_{SB(New)}$ is corrected to N. The predicted mean of T_F is $[N - (p^* - q)]/[N - (p^* - q) - 2]$, which will vary across all sample sizes and approach 1 with increasing sample sizes.

CHAPTER 4

Results and Analysis

The simulation results under each condition are reported in Table 4.1 - 4.3, one table per condition. The columns of each table give the sample size used for a particular set of 1,500 replications from the population. At each sample size, a sample was drawn, and each of the four modified test statistics shown in the rows of the table was computed; the process was replicated 1,500 times. Then the resulting T statistics were used to compute (a) the mean of the 1,500 statistics, (b) the standard deviation of the 1,500 statistics, (c) the frequency of rejecting the null hypothesis at the 0.05 level under the correct model, i.e., the type I error, and (d) the frequency of rejecting the null hypothesis at the 0.05 level under the incorrect (misspecified) model, i.e, the empirical power. These are the four entries in each cell of each table.

Condition 1 in Table 4.1 is the baseline condition in which the factors and errors, and hence the observed variables, are multivariate normally distributed. Asymptotically, T_{SB} and T_{YB} yield a mean test statistic T of about 87, and the standard deviation is around 13.19. T_{SB} seems to approach the mean of 87 a little faster than T_{YB}, while T_{YB} shows a relatively smaller deviation than that of T_{SB} across all sample sizes except for 5,000. Both the mean and the standard deviation of T_{MV} increase as the sample size get larger, but still shows an overcorrection to the standard χ^2 distribution with a degree of 87 at the largest sample size. The mean of T_F converges to 1 as sample size gets larger as predicted. An ideal type I error rate, as indicated in previous chapter, should stay within 95% confidence interval $[0.0387, 0.0613]$. T_{SB} and T_{YB} yield ideal type I error rates at a sample size as small as 500, followed by T_{MV} at 1,000, and T_F

when the sample size reaches 2,500. Under small sample sizes, T_{MV} outperforms the others, followed by T_{SB}. While T_{YB} tends to accept the null hypothesis too readily at small samples, T_F rejects the correct model too frequently. Both T_{YB} and T_F are not applicable in the case $N < d$, and thus they can not be trusted at small samples. Under moderate and large sample sizes, T_{YB} and T_{SB} perform almost on par, followed by T_{MV}, while T_F still frequently rejects the model except for the largest samples. The empirical power of all the test statistics reaches almost 100% when sample size is as large as 500. At smaller sample sizes, T_{SB} performs best in rejecting the misspecified model, while T_{MV} loses its advantage. T_{YB} and T_F accept the wrong model too frequently and yield very low rejection rates at small sample sizes. A closer examination of the type I error rate and the corresponding empirical power reveals a contradiction, and this indicates that any test statistic with an ideal type I error rate is not necessarily reliable unless it is empirically powerful in rejecting a wrong model.

Condition 2 is designed to be consistent with asymptotic robustness theory. As we can see from Table 4.2, the behavior of the four test statistics is very similar to that in Condition 1. All four test statistics exhibit robustness to some extent. T_{SB} and T_{YB} behave like a χ^2 variate with 87 degrees of freedom asymptotically, while T_{MV} approaches this limit quite slowly. In terms of type I error control, T_{MV} still outperforms the other statistics at small samples, but even T_{MV} does not yield quite ideal type I error. Under moderate and large sample sizes over 500, both T_{SB} and T_{YB} perform stably well, followed by T_{MV} and T_F. The performance of T_F, even though it still rejects the correct model too often, has slightly improved compared to that under Condition 1. The behavior of T_{YB} and T_F are expected to vary little across three different Conditions since they should not depend on any specific distributions of the observed variables. The empirical power repeats the pattern we have observed in Condition 1, with T_{SB} performing the best, followed by T_{MV}, while T_{YB} and T_F still performing badly under small and moderate sample sizes.

Condition 3 simulated a situation when the asymptotic robustness of normal-theory based test statistic is no longer valid. As expected, T_{SB}, T_{YB} and T_F demonstrate their robustness under multivariate nonnormality. T_{MV} completely breaks down in this case and tends to always accept the null hypothesis. Its outstanding performance at small samples in previous conditions disappears in this case. Under the smallest sample size, T_{SB} is the only test statistic to be applied in this study. When sample size reaches 100, T_F gives a very promising type I error rate, but a second thought on its empirical power, which is only 0.074, will likely lead us nowhere but to trust T_{SB} again. Under moderate sample sizes, T_{SB} still performs the best, followed by T_F and T_{YB}; however, T_F enjoys an advantage over T_{SB} and T_{YB} in terms of the empirical power. Under large samples, T_{YB} demonstrates its excellent robustness, followed by T_F and T_{SB}. T_F tends to slightly over reject while T_{SB} tends to under reject the null hypothesis, however we should not jump into a hasty conclusion in one simulation study.

In conclusion, there is no simple winner in this study. T_{SB}, T_{YB} and T_F all show strong robustness across the three conditions simulated, T_{MV} also demonstrate obvious advantage under certain conditions. For practical applications, following suggestions are proposed. When we have little information about the distributional characteristics of the observed data, it may be beneficial to examine T_{SB}, T_{YB} and T_F simultaneously for hypothesis testing. Particularly, when the sample size is small or moderate, T_{SB} should be more reliable; and when the sample size is large enough, 1,000 for instance, T_{YB} and T_F are more likely to give a reliable inference. However, when we have sufficient confidence in the assumptions of normality or asymptotic robustness with a small or moderate size of observations, T_{MV} is highly recommended as an addition to T_{SB}.

Table 4.1: Summary of Simulation Results for Condition 1 (Factors and errors are independently distributed normal variates)

	Sample Size						
Test Statistics	50	100	250	500	1,000	2,500	5,000
SB scaled							
Mean	61.58	96.179	90.306	87.944	88.656	87.969	86.983
SD	9.441	14.439	13.628	13.279	13.006	13.251	12.879
Type I Error	0.261	0.173	0.092	0.053	0.054	0.058	0.047
Empirical Power	0.44	0.577	0.915	1.00	1.00	1.00	1.00
MV							
Mean	30.221	42.08	59.397	69.763	78.441	83.593	84.768
SD	4.801	6.196	8.642	10.245	11.314	12.496	12.503
Type I Error	0.069	0.043	0.039	0.036	0.043	0.053	0.042
Empirical Power	0.177	0.316	0.837	1.00	1.00	1.00	1.00
YBRES							
Mean	NA	87.054	90.88	89.247	89.487	88.107	87.216
SD	NA	4.128	10.799	12.513	12.735	13.057	12.903
Type I Error	NA	0.00	0.045	0.053	0.061	0.057	0.049
Empirical Power	NA	0.00	0.439	0.976	1.00	1.00	1.00
YBRESF							
Mean	NA	1.395	1.094	1.04	1.035	1.014	1.003
SD	NA	0.691	0.208	0.179	0.162	0.156	0.1512
Type I Error	NA	0.098	0.111	0.074	0.071	0.061	0.049
Empirical Power	NA	0.157	0.626	0.987	1.00	1.00	1.00

Table 4.2: Summary of Simulation Results for Condition 2 (Factors and errors are independently distributed non-normal variates)

	Sample Size						
Test Statistics	50	100	250	500	1,000	2,500	5,000
SB scaled/new							
Mean	61.44	95.768	90.247	88.906	87.475	87.575	86.991
SD	9.40	14.492	13.262	13.337	13.251	13.727	12.818
Type I Error	0.272	0.167	0.075	0.063	0.048	0.061	0.043
Empirical Power	0.435	0.595	0.91	1.00	1.00	1.00	1.00
MV							
Mean	25.832	35.454	51.883	64.214	73.184	81.139	83.646
SD	5.104	6.335	8.065	9.868	10.994	12.638	12.726
Type I Error	0.038	0.029	0.038	0.034	0.042	0.053	0.037
Empirical Power	0.126	0.241	0.815	1.00	1.00	1.00	1.00
YBRES							
Mean	NA	86.297	90.297	89.548	88.23	88.048	87.243
SD	NA	3.952	10.227	12.189	12.650	13.472	12.821
Type I Error	NA	0.00	0.034	0.049	0.049	0.067	0.047
Empirical Power	NA	0.00	0.558	0.992	1.00	1.00	1.00
YBRESF							
Mean	NA	1.358	1.081	1.044	1.019	1.014	1.003
SD	NA	0.644	0.196	0.174	0.161	0.161	0.15
Type I Error	NA	0.08	0.093	0.066	0.058	0.068	0.048
Empirical Power	NA	0.125	0.755	0.996	1.00	1.00	1.00

Table 4.3: Summary of Simulation Results for Condition 3 (Factors and errors are dependently distributed non-normal variates)

	Sample Size						
Test Statistics	50	100	250	500	1,000	2,500	5,000
SB scaled/new							
Mean	63.69	97.765	90.584	88.562	86.669	86.62	87.027
SD	8.858	14.729	12.338	13.748	12.152	12.312	12.802
Type I Error	0.312	0.174	0.066	0.052	0.043	0.043	0.045
Empirical Power	0.461	0.432	0.579	0.842	0.977	0.997	0.999
MV							
Mean	13.322	14.335	17.013	20.106	23.681	29.506	35.712
SD	5.114	6.268	8.647	10.734	12.546	15.605	17.753
Type I Error	0.013	0.005	0.002	0.002	0.003	0.003	0.01
Empirical Power	0.021	0.026	0.097	0.358	0.772	0.951	0.981
YBRES							
Mean	NA	86.663	89.928	89.954	88.743	88.506	88.079
SD	NA	3.883	9.473	11.254	11.647	12.080	12.825
Type I Error	NA	0.00	0.018	0.033	0.054	0.047	0.05
Empirical Power	NA	0.00	0.294	0.913	1.00	1.00	1.00
YBRESF							
Mean	NA	1.301	1.072	1.049	1.025	1.019	1.013
SD	NA	0.571	0.179	0.165	0.148	0.144	0.145
Type I Error	NA	0.061	0.069	0.057	0.061	0.053	0.053
Empirical Power	NA	0.074	0.529	0.946	1.00	1.00	1.00

CHAPTER 5

Discussion

The behavior of four modified goodness-of-fit test statistics was evaluated through a Monte Carlo study. The well-known and extensively applied test statistic Satorra-Bentler scaled test statistic was used as a benchmark, which has been considered to work quite reliably under a wide variety of conditions (e.g., Hu et al., 1992; Curran et al., 1996). A relatively unknown mean scaled and variance adjusted test statistic T_{MV} was shown to outperform T_{SB} under certain conditions, but also to break down completely in one condition. In fact this was the only statistic to not perform adequately at $N = 5,000$. Although the failure of T_{MV} has been observed previously (Tong and Bentler, in press), a theoretical explanation is unclear. As we can see from equations (2.9) - (2.12), the comparative performance of T_{SB} and T_{MV} will mainly be affected by the eigenvalues of the product matrix $U\Gamma$. This problem has been addressed by Yuan and Bentler (2010). They evaluated the type I error and mean-square error of T_{MV} and T_{SB} under different coefficients of variation in the eigenvalues of $U\Gamma$, and found that T_{MV} will perform better than T_{SB} when the disparity of eigenvalues is large. This might lead to the situations we observed at small and moderate sample size under Condition 1 and 2. However, as Yuan and Bentler (2010) noted, it is currently not easy to test the level of disparity of the eigenvalues (measured by coefficient of variation). It is also not clear how such a disparity could explain the good performance of T_{SB} and bad performance of T_{MV}. There seems to be no effective way of determining which test statistic should be applied given any datasets under a specified model, although T_{SB} never fails completely and thus will be preferred over T_{MV} due to empirical simulation results.

The Yuan-Bentler residual-based test statistic and the extended F-statistic demonstrated promising robustness under three conditions simulated in Section 3. Bentler and Yuan (1999) examined the relative performance of T_{SB}, T_{YB} and T_F under small samples, and the result can be confirmed at sample size of 50 and 100 in this study as well. They found that T_{SB} break down with small sample sizes between 60 to 120 under various conditions, whether the assumptions of multivariate normality is violated or not. T_{YB} essentially always accepts the true model when it should be at least occasionally rejecting this model by chance, and this problem was also observed at sample size of 100 under all three conditions in this paper. T_F statistic performed remarkably well at all small sample sizes in their simulations, although it had some over rejections under conditions of normality. As observed again in this paper, T_F continued to outperform T_{SB} and T_{YB} at small sample size; but T_{MV} performed even better under multivariate normality and asymptotic robustness conditions. Another problem worth noticing is that they focused on evaluating the rejection rates under correct model and didn't address the empirical power of these test statistics. This problem is addressed in this paper, and as shown in Section 4, the empirical power of T_{YB} is attenuated greater than that of T_F at small and moderate sample sizes across all conditions. Since a good statistic possess the property of a controllable type I error while achieving a maximum power, T_F may not be exactly ideal for general hypothesis testing under small samples as Bentler and Yuan (1999) proposed. It is known that power decreases with increasing kurtosis (Foldnes, Olsson, & Foss, 2012; Foss, Jöreskog, & Olsson, 2011; Olsson, Foss, & Troye, 2003), so some lack of power can be expected. Thus, a more suitable test statistic for small samples under general distributions remains to be developed in the future.

It is clear in this study that T_{SB} and T_{YB} have tail behavior consistent with the asymptotic chi-square distribution under three conditions. T_{MV} approaches the χ^2 distribution much slower but still gives satisfactory rejection rates under specific conditions. The tail behavior of T_F shows characteristics of F distributions under all conditions, but with no fixed standard since its distribution depends on sample size. These indicate that at sufficient large sample sizes, all but T_{MV} could be used for hypothesis testings under general distribution of observed variables. However, as Yuan and Bentler (1998) pointed out, when sample sizes are greater than 200, the statistic T_{SB} gives very good and reliable performances on the condition that all the eigenvalues of $U\Gamma$ are equal or nearly equal; otherwise it tends to perform worse as sample sizes increase. This problem has not been observed in this paper, but we should certainly take that into consideration before giving any general suggestions. Based on this simulation study alone, the suggestions are already given at the end of Section 4.

The most worthwhile theoretical issues to be considered in the future are the following: 1. Develop a robust test statistic with controllable type I error rate as well maximum empirical power at very small sample sizes, especially when the sample sizes are smaller than the degrees of freedom; 2. Develop a direct way to compute the coefficient of variation of the eigenvalues of $U\Gamma$ in order to determine which of T_{SB} and T_{MV} should be employed; 3. Modify T_F to a larger extent so that it will be equipped with more empirical power at small samples. Success at this would also solve the first point.

References

[1] Anderson, T. (2003). *An introduction to multivariate statistical analysis (3rd ed.)*. New York, NY: Wiley-Interscience.

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

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

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

[5] Fleishman, A. I. (1978). A method of simulating non-normal distributions. *Psychometrika, 43,* 521-532.

[6] Foldnes, N., Olsson, U. H., & Foss, T. (2012). The effect of kurtosis on the power of two test statistics in covariance structure analysis. *British Journal of Mathematical and Statistical Psychology, 65,* 1-18.

[7] Foss, T., Jöreskog, K.G., & Olsson, U.H. (2011). Testing structural equation models: The effect of kurtosis. *Computational Statistics and Data Analysis, 55,* 2263-3375.

[8] Headrick, T. C. (2002). Fast fifth-order polynomial transforms for generating univariate and multivariate nonnormal distributions. *Computational Statistics and Data Analysis, 40,* 685-711.

[9] Headrick, T. C. & Swailowsky, S. S. (1999). Simulating correlated multivariate nonnormal distributions: Extending the fleishman power method. *Psychometrika, 64,* 25-34.

[10] Hu, L. and Bentler, P.M. & Kano,Y. (1992). Can test statistics in covariance structure analysis be trusted. *Psychological Bulletin, Vol. 112, No.2,* 351-362.

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

[12] Mair, P., Wu, E., & Bentler, P. M. (2010). EQS goes R: Simulations for SEM using the package REQS. *Structural Equation Modeling, 17,* 333-349.

[13] Olsson U.H., Foss, T., & Troye, S.V. (2003). Does the ADF fit function decease when the kurtosis increases. *British Journal of Mathematical and Statistical Psychology, 56,* 289-303.

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

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

[16] Tong, X. & Bentler, P.M. (in press). Evaluation of a new scaled and moment adjusted test statistic for SEM. *Structural Equation Modeling.*

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

[18] Yuan, K.H., & Bentler, P.M. (2010). Two simple approximations to the distribution of quadratic forms. *British Journal of Mathematical and Statistical Psychology, 63,* 273-291.