Bootstrap approach to inference and power analysis based on three test statistics for covariance structure models

Ke-Hai Yuan\textsuperscript{1,*} and Kentaro Hayashi\textsuperscript{2}

\textsuperscript{1}University of Notre Dame, USA
\textsuperscript{2}Georgia State University, USA

We study several aspects of bootstrap inference for covariance structure models based on three test statistics, including Type I error, power and sample-size determination. Specifically, we discuss conditions for a test statistic to achieve a more accurate level of Type I error, both in theory and in practice. Details on power analysis and sample-size determination are given. For data sets with heavy tails, we propose applying a bootstrap methodology to a transformed sample by a downweighting procedure. One of the key conditions for safe bootstrap inference is generally satisfied by the transformed sample but may not be satisfied by the original sample with heavy tails. Several data sets illustrate that, by combining downweighting and bootstrapping, a researcher may find a nearly optimal procedure for evaluating various aspects of covariance structure models. A rule for handling non-convergence problems in bootstrap replications is proposed.

1. Introduction

Covariance structure analysis (CSA) has become one of the most popular methods in multivariate analysis, especially in the social and behavioural sciences. The classical approach to model evaluation uses the likelihood ratio statistic $T_{ML}$ based on the normality assumption (Lawley & Maxwell, 1971). Because data sets in practice are seldom normal, Browne (1984) developed an asymptotically distribution-free statistic, $T_B$. Satorra and Bentler (1988) proposed a rescaled version of the likelihood ratio statistic, $T_{SB}$, which performs quite robustly under violations of the normality assumption (Hu, Bentler, & Kano, 1992). The analytical merits of these statistics are characterized by asymptotics. With practical data, there still exist violations of conditions that will interfere with accurate applications of these statistics. Actually, none of these statistics

\* Requests for reprints should be addressed to Ke-Hai Yuan, Department of Psychology, Laboratory for Social Research, University of Notre Dame, Notre Dame, IN 46556, USA (e-mail: kyuan@nd.edu).
can be well described by a chi-square distribution even for simulated normal data with small to medium sample sizes (Bentler & Yuan, 1999).

Bootstrap methods provide a competitive alternative for statistical inference under violations of standard regularity conditions (Efron & Tibshirani, 1993; Davison & Hinkley, 1997). Bootstrap testing in CSA was developed by Beran and Srivastava (1985). Bollen and Stine (1993) and Yung and Bentler (1994, 1996) further showed how to obtain correct Type I errors with $T_{ML}$ and $T_B$. Zhang and Boos (1992) used the bootstrap to test homogeneity of covariance matrices from multiple samples. Simulation studies by Fouladi (1998) indicate that bootstrap inference with $T_{ML}$ has the best overall performance in controlling Type I errors.

Although the bootstrap has shown promise for CSA, various unknowns related to this procedure still exist when it is applied to the three commonly used statistics $T_{ML}$, $T_{SB}$ and $T_B$. For example, bootstrap theory requires the underlying sampling distribution to have finite fourth-order moments. With practical data possessing large empirical kurtosis, how can the bootstrap be safely applied in CSA? When using a bootstrap method, are $T_{ML}$, $T_B$ and $T_{SB}$ equivalent? If not, which statistic should one choose? When inference is based on asymptotics, the existing literature indicates that $T_B$ has the best power (Fouladi, 2000; Yuan & Bentler, 1997). What are the power properties of the three statistics when they are used with the bootstrap? With a bootstrap method, how should one determine the sample size needed to achieve a certain power in CSA? How should one deal with non-convergence problems in bootstrap replications?

Our purpose in this paper is to address the above issues associated with CSA. In Section 2 we provide details of a power analysis via the bootstrap, including sample-size determination and a test for close fit. In Section 3 we compare the pivotal properties of the three statistics for bootstrap inference. We also discuss limitations of the pivotal properties to be realized with practical data. In Section 4 we propose a downweighting approach that can be applied with the bootstrap to distributions with heavy tails. As will be shown by means of examples in Section 5, for data having heavy tails, using the downweighting procedure with the bootstrap not only makes the bootstrap procedure valid, but also leads to more accurate statistical inference. Examples in Section 5 also illustrate how to find a nearly optimal procedure when using the bootstrap for inference. Through the examples, we also compare the power properties of the three test statistics. In Section 6 we provide a reasonable way of dealing with non-convergence in bootstrap replications. The paper concludes with a brief discussion.

2. Bootstrap approaches to inference and power

Beran (1986) studied the power properties of the bootstrap based on a general statistic whose distribution depends on nuisance parameters. Our development of inference and power analysis can be regarded as an application of Beran’s general theory of resampling to CSA.

2.1. Model test

Let $F_x$ be the cumulative distribution function (cdf) of a $p$-dimensional population from which a sample $x_1, \ldots, x_n$ has been drawn. Let $\hat{F}_x$ be the corresponding empirical distribution function (edf) defined by the sample $x_i$'s. Without loss of generality, we assume $E(x_i) = \mu$ and $\text{Cov}(x_i) = \Sigma$. For a covariance structure $\Sigma(\theta)$, an inference
procedure is to find whether there exists a $\theta_0$ such that the null hypothesis

$$H_0: \Sigma = \Sigma(\theta_0)$$ (1)

holds. A good testing procedure should accept $H_0$ when it is true. When an alternative hypothesis $\Sigma = \Sigma_a$ is true such that

$$H_1: \Sigma_a \neq \Sigma(\theta) \text{ for any admissible } \theta,$$ (2)

it should reject (1).

In order to test (1) with a statistic $T = T(x_1, \ldots, x_n)$ we need a critical value $c_\alpha$ such that

$$P(T > c_\alpha | H_0) = \alpha.$$ (3)

Without knowing $F_x$, it is impossible for us to find $c_\alpha$ when the probability in (3) is measured by $F_x$. However, we can define a $c^*_\alpha$ such that

$$P(T > c^*_\alpha | \hat{F}_0) = \alpha,$$ (4)

where $\hat{F}_0$ is an edf with a covariance matrix satisfying (1). In bootstrap testing for (1), $c^*_\alpha$ is estimated through resampling from $\hat{F}_0$. Let

$$y_i = \hat{\Sigma}^{1/2} S_x^{-1/2} x_i, \quad i = 1, \ldots, n,$$

where $\hat{\Sigma} = \Sigma(\hat{\theta})$ for an admissible estimate $\hat{\theta}$ (Beran & Srivastava, 1985). It is obvious that the edfs $\hat{F}_x$ and $\hat{F}_y$ have the same type of distribution but differ in ‘location and scale’. Actually, the only purpose of (5) is to create an empirical distribution $\hat{F}_0 = \hat{F}_x$ having the same distributional type as that of $\hat{F}_x$ and with a covariance matrix satisfying $H_0$.

Let $\mathcal{Y}^*_b = (y_1^{(b)}, \ldots, y_n^{(b)})$ be a random sample of size $n$ from $\hat{F}_0$ and $T^*_b$ be the corresponding statistic of $T$ evaluated at this sample. With independent samples $\mathcal{Y}^*_b$, $b = 1, \ldots, B_0$, we rearrange the $T^*_b$ in order and denote them by

$$T(1)^* \leq T(2)^* \leq \ldots \leq T(B_0)^*.$$

The bootstrap estimate $\hat{c}^*_\alpha$ of $c^*_\alpha$ is the $B_0(1 - \alpha)$th quantile $T(B_0(1 - \alpha))^*$, and we reject the null hypothesis (1) when $T > \hat{c}^*_\alpha$. Let $\alpha$ be the exact level of a test defined in (3) and $\hat{\alpha}^*$ be the corresponding level when $c_\alpha$ is replaced by $\hat{c}^*_\alpha$. Under quite general conditions, Beran (1986) established the consistency of $\hat{\alpha}^*$ for $\alpha$. These conditions include the fourth-order moments of $F_x$ being finite.

### 2.2. Power evaluation

The bootstrap can also be used to evaluate the power $\beta_\alpha$ of a statistic $T$ defined by

$$\beta_\alpha = P(T > c_\alpha | H_1).$$ (6)

Let $\Sigma_a$ be the covariance matrix in (2) and

$$z_i = \Sigma_a^{1/2} S_x^{-1/2} x_i.$$ (7)

Then $\hat{F}_x = \hat{F}_x$ represents the bootstrap population under $H_1$. For $b = 1, 2, \ldots, B_1$, let $Z_b^* = (z_1^{(b)}, \ldots, z_n^{(b)})$ be random samples of size $n$ from $\hat{F}_x$ and $T_b^*$ be the corresponding statistic of $T$ evaluated at $Z_b^*$. Then the bootstrap estimate of $\beta_\alpha$ in (6) is

$$\hat{\beta}_\alpha^* = \frac{\# \{ T_b^* > \hat{c}^*_\alpha \}}{B_1},$$ (8)

where $\hat{c}^*_\alpha$ is the bootstrap estimate of the $c^*_\alpha$ in (4). Under fairly general conditions, Beran (1986) established the consistency of $\hat{\beta}_\alpha^*$ for $\beta_\alpha$ as $n$ approaches infinity. This consistency
should be understood in the following sense: when $\Sigma_\delta$ is the population covariance matrix of $F_\delta$ that generates the observed sample $x_1, \ldots, x_n$, then $\beta_\alpha^* \text{ in (7) is consistent}$ for the probability $\beta_\alpha$ in (6), and the latter is the exact power measured by the cdf $F_\delta$.

In contrast to (5), the $\Sigma_\delta$ in (7) cannot be replaced by a consistent estimate. For example, we will not obtain a consistent estimate of power for $T$ to reject the null hypothesis $H_0$ under the true covariance matrix $\Sigma_0 = \text{Cov}(x_i)$ when replacing $\Sigma_\delta$ by the sample covariance matrix $S_\delta$ of the $x_i$'s. The inconsistency here can be well understood through $T_\text{ML}$. Actually, with $\Sigma_\delta$ being $O(1/\sqrt{n})$ apart from $\Sigma_0$, the bootstrap approach to power is asymptotically equivalent to the approach developed in Satorra and Saris (1985) for normally distributed data (see also Steiger, Shapiro, & Browne, 1985). These authors used a non-central chi-square distribution to describe the behaviour of $T_\text{ML}$ under $H_1$ with a non-centrality parameter (NCP)

$$\delta = n \min_{\theta} D(\Sigma_\delta, \Sigma(\theta)), \quad (9)$$

where $D = D_\text{ML}$ is the Wishart likelihood discrepancy function. When replacing $\Sigma_\delta$ by a consistent estimate $\hat{\Sigma}_\delta$, even though $\hat{\Sigma}_\delta = \Sigma_\delta + O_p(1/\sqrt{n})$, the NCP estimate based on $\hat{\Sigma}_\delta$ can be $O_p(1)$ apart from that based on $\Sigma_\delta$. This limitation of power evaluation also exists in other classical procedures for power analysis (e.g., Cohen, 1988). The fact is that, without knowledge of the true effect size, one cannot consistently estimate the power of a statistic in rejecting the null.

Although the bootstrap approach to power analysis is equivalent to the approach proposed by Satorra and Saris (1985) for normally distributed data when $\Sigma_\delta$ is $O(1/\sqrt{n})$ apart from $\Sigma(\theta)$, there are fundamental differences between the two approaches. With the bootstrap approach, there is neither a central chi-square distribution when $H_0$ holds nor a non-central chi-square distribution when $H_1$ holds. Instead, the distribution of $T$ is judged by its empirical behaviour based on resampling from observed data. We may regard $\lambda = \min_{\theta} D(\Sigma_\delta, \Sigma(\theta))$ as a measure of effect size. For a given sample size, the greater $\lambda$ is, the higher the power for a bootstrap test to distinguish between $H_0$ and $H_1$. Because there is no chi-square distribution assumption with the bootstrap approach, power analysis cannot resort to a chi-square table, and has to be evaluated separately for different samples. This tailor-made type of analysis can be regarded as an advantage of the bootstrap methodology.

### 2.3. Determination of sample size

Let $x_1, \ldots, x_n$ be a pilot sample. Under $H_1$ in (2), we want to find the smallest sample size for a statistic $T$ to reject $H_0$ with probability $\beta_0$. For this purpose, we first draw $B_0$ independent samples $y^{(b)}_r = (y_1^{(b)}, \ldots, y_m^{(b)})$ of size $m$ from $\hat{F}_0$ to estimate the critical value $c^*_\alpha(m)$ as in Section 2.1. We then draw $B_1$ independent samples $z^{(b)}_r = \hat{F}_z$ of size $m$, where the covariance matrix of $z \sim F_z$ is $\Sigma_\delta$. Evaluating $T^*_\delta(m)$ at each sample $z^{(b)}_r$, as in (8), the estimated power for sample size $m$ is

$$\beta_\alpha^*_\delta(m) = \# \{ T^*_\delta(m) > c^*_\alpha(m) \} / B_1.$$ 

A smaller sample size $m_1$ is needed if $\beta_\alpha^*_\delta(m) > \beta_0$, otherwise a greater sample size $m_2$ is needed. Finding the minimum $m$ such that $\beta_\alpha^*_\delta(m) \geq \beta_0$ may require a series of trials. The interval-halving procedure in the Appendix of MacCallum, Browne, and Sugawara (1996), who studied sample-size determination using non-central chi-squares, can be equally applied here. More discussion on the bootstrap approach to sample size and
Since where Although it is unclear which of the statistics 3. Pivotal property of $i$ $h$ $D$ $etal$ A covariance structure model is at best only an approximation to the real world. Any interesting model will be rejected when the sample size is large enough. Based on the root-mean-square error of approximation (RMSEA) (Steiger & Lind, 1980), MacCallum et al. (1996) proposed testing for close fit rather than exact fit in (1). Using a non-central chi-square to describe the behaviour of $T_{ML}$, their approach is equivalent to testing if the NCP is less than a prespecified value. The test for close fit can also be easily implemented with the bootstrap. Let $\Sigma_c$ be a covariance matrix with close fit such that

$$\delta_c = \min_\theta D(\Sigma_c, \Sigma(\theta)), \quad (10)$$

where $\delta_c$ is the ‘non-centrality parameter’ in equation (8) of MacCallum et al. (1996). Such a covariance matrix can be found in the form of $\Sigma^b = b \Sigma_x + (1 - b) \Sigma(\theta)$. When $D = D_{ML}$ is the Wishart likelihood discrepancy function, Yuan and Hayashi (2001) showed that $\min_\theta D(\Sigma^b, \Sigma(\theta)) = D(\Sigma^b, \Sigma(\theta))$ is a strictly increasing function of $b \in [0, 1]$, and it is straightforward to find a $\delta_c$ once a $\delta_c$ is given. Let $y_i = \Sigma_1^{1/2} S_x^{-1/2} x_i$, $i = 1, \ldots, n$; then the covariance matrix of $y \sim \tilde{F}_y$ is $\Sigma_c$. Let $T^b_\theta$, $b = 1, \ldots, B_0$ be the corresponding statistics of $T$ evaluated respectively at $B_0$ independent samples from $\tilde{F}_y$. Then $\hat{c}_a^* = T_{(1 - a)\alpha}^b$ is the critical value estimate of the test for close fit, and the model will be rejected when $T = T(x_1, \ldots, x_n) > \hat{c}_a^*$. It is also straightforward to perform power analysis with the bootstrap when a less-close-fit covariance matrix describes the population.

3. Pivotal property of $T_{ML}$, $T_{SB}$ and $T_B$

Although it is unclear which of the statistics $T_{ML}$, $T_{SB}$ and $T_B$ is best for bootstrapping with a specific data set, there is a general result for choosing a statistic (Beran, 1988; Hall, 1992). In the following, we will discuss this result and its relevance to CSA.

A statistic is pivotal if its distribution does not depend on any parameters in the underlying sampling distribution, and it is asymptotically pivotal if its asymptotic distribution does not depend on unknown parameters. Let a statistic $T = f(S_x)$ be a smooth function of the sample covariance matrix $S_x$ of $x_1, \ldots, x_n$. Suppose $T$ is asymptotically pivotal and an Edgeworth expansion applies to its distribution function under $H_0$ (Barndorff-Nielsen & Cox, 1984; Wakaki, Eguchi, & Fujikoshi, 1990),

$$P(T \leq t \mid F_0) = G(t) + n^{-1} g(t) + O(n^{-3/2}), \quad (11a)$$

where $G(t)$ is the asymptotic distribution function of $T$ and $g(t)$ is a smooth function that depends on some unknown population parameters. As discussed in Hall (1992) and Davison and Hinkley (1997, Section 2.6.1), (11a) generally holds for smooth functions of sample moments. Let $T^*$ be the corresponding statistic based on resampling from the corresponding $\tilde{F}_0$. Its Edgeworth expansion is

$$P(T^* < t \mid \tilde{F}_0) = G(t) + n^{-1} \hat{g}(t) + O_p(n^{-3/2}). \quad (11b)$$

Since $\hat{g}(t) = g(t) + O_p(n^{-1/2})$ in general,

$$P(T^* \leq t \mid \tilde{F}_0) - P(T \leq t \mid F_0) = O_p(n^{-3/2}). \quad (12)$$

power can be found in Beran (1986), Efron and Tibshirani (1993, Chapter 25) and Davison and Hinkley (1997, Section 4.6).
Suppose we know $G$, then asymptotic inference is based on comparing the observed statistic $T = T(x_1, \ldots, x_n)$ to the critical value $G^{-1}(1 - \alpha)$. Equation (11a) implies that asymptotic inference can achieve a designated level $\alpha$ to the order of accuracy of $O_p(1/n)$. According to (12), bootstrap inference based on an asymptotically pivotal statistic can achieve the level $\alpha$ to a higher order of accuracy. When $T$ is not asymptotically pivotal, $G(t)$ in (11) depends on some unknown parameters. One has to estimate $G(t)$ by $\hat{G}(t)$ for asymptotic inference. Similarly, with the bootstrap one has to implicitly estimate the unknown parameters in $G(t)$, and consequently it achieves the same order of accuracy to the level $\alpha$ as that based on $\hat{G}(t)$. So for more accurate inference we should choose a statistic that is as nearly pivotal as possible. Beran (1988) gave a nice discussion on using a pivotal statistic for bootstrap and the order of accuracy for Type I errors.

When data are normally distributed, $(T_{ML} | H_0)$ is asymptotically distributed as $\chi^2_{p^* - q}$, where $p^* = p(p + 1)/2$ and $q$ is the number of unknown parameters in $\Sigma(\theta)$. Consequently, $T_{ML}$ is asymptotically pivotal (Davison & Hinkley, 1997, p. 139). More general conditions also exist for $T_{ML}$ to be asymptotically pivotal with some specific models (Amemiya & Anderson, 1990; Browne & Shapiro, 1988; Mooijaart & Bentler, 1991; Satorra & Bentler, 1990; Shapiro, 1987; Yuan & Bentler, 1999). Unfortunately, there is no effective way of verifying these conditions in practice. The statistic $T_{SB}$ is obtained by rescaling $T_{ML}$ using fourth-order sample moments (Satorra & Bentler, 1988). Within the class of elliptical or pseudo-elliptical distributions with finite fourth-order moments, $(T_{SB} | H_0)$ approaches $\chi^2_{p^* - q}$ (Yuan & Bentler, 1999). So $T_{SB}$ is asymptotically pivotal when the sampling distribution is elliptical or pseudo-elliptical. It is not asymptotically pivotal for other types of non-normal distribution. In contrast to $T_{ML}$ and $T_{SB}$, $T_B$ is asymptotically pivotal for any sampling distribution with finite fourth-order moments.

The above discussion may imply that $T_B$ is the preferred statistic for bootstrap inference in CSA. However, there exist conditions that may hide the pivotal property of $T_B$ with finite samples. In addition to the asymptotically pivotal property, Davison and Hinkley (1997, Section 2.5.1) discussed nearly pivotal properties of a statistic $T$, which requires the distribution of $T$ to approximately follow the same distribution when sampling from $F_x$ or $\hat{F}_x$. Since the target function $G(t)$ for the three test statistics is the cdf of $\chi^2_{p^* - q}$, the better a statistic is approximated by a chi-square distribution, the more accurate its bootstrap inference is in controlling Type I errors. Because sample size is a serious issue with CSA, a statistic that is asymptotically pivotal may not be nearly pivotal, as will be shown in Section 5.

4. Heavy tails with practical data

Bootstrap inference based on $T_{ML}$, $T_B$ or $T_{SB}$ requires $F_x$ to have finite fourth-order moments. Of course, the sample fourth-order moments of any sample will always be finite. However, if some of the fourth-order moments (e.g., kurtoses) are quite large, the corresponding population fourth-order moments may actually be infinite. Even when all the population fourth-order moments are finite, inference based on the bootstrap will not be accurate when $F_x$ possesses heavy tails. Here, we propose applying the bootstrap procedure to a sample in which cases contributing to heavy tails are properly down-weighted. Following the study in Yuan, Chan, and Bentler (2000), we only use the Huber-type weights (Huber, 1981; Tyler, 1983) in controlling those cases.
For a $p$-variate sample $\mathbf{x}_1, \ldots, \mathbf{x}_n$, let

$$d_i = d(\mathbf{x}_i, \mu, \Sigma) = \left(|(\mathbf{x}_i - \mu)\Sigma^{-1}(\mathbf{x}_i - \mu)|\right)^{1/2}$$

be the Mahalanobis distance and $\mu_1(t)$ and $\mu_2(t)$ be non-negative scalar functions. A robust $M$-estimator $(\hat{\mu}, \hat{\Sigma})$ can be obtained by solving (Maronna, 1976)

$$\mu = \frac{\sum_{i=1}^n u_1(d_i)\mathbf{x}_i}{\sum_{i=1}^n u_1(d_i)} \quad (13a)$$

$$\Sigma = \frac{\sum_{i=1}^n u_2(d_i^2)(\mathbf{x}_i - \mu)(\mathbf{x}_i - \mu)/n.} \quad (13b)$$

With decreasing functions $u_1(t)$ and $u_2(t)$, cases having greater $d_i$s receive smaller weights and thus their effects are controlled. Let $\rho$ represent the proportion of outlying cases one wants to control, and $r$ be a constant defined by $P(\chi^2_p > r^2) = \rho$. The Huber-type weights corresponding to this $\rho$ are

$$u_1(d) = \begin{cases} 1, & \text{if } d \leq r, \\ \frac{r}{d}, & \text{if } d > r, \end{cases} \quad (14)$$

and $u_2(d^2) = \left\{u_1(d^2)\right\}/\varphi$, where $\varphi$ is a constant such that $E\{\chi^2_p, u_2(\chi^2_p)\} = p$ which makes the estimate $\hat{\Sigma}$ unbiased if $F_\Sigma = N_p(\mu, \Sigma)$. We may regard $\rho$ as a tuning parameter which can be adjusted according to the tails of a specific data set. Applying different types of weights to several data sets, Yuan, Bentler, and Chan (2001) found that the most efficient parameter estimates often go with Huber-type weights.

Let $(\hat{\mu}, \hat{\Sigma})$ be the converged solution to (13), $u_2 = u_2(d^2(\mathbf{x}_i, \hat{\mu}, \hat{\Sigma}))$ and

$$\mathbf{x}_i^{(\rho)} = \{\sqrt{u_2}(\mathbf{x}_i - \hat{\mu})\}. \quad (15)$$

Then we can rewrite (13b) as

$$\hat{\Sigma} = \sum_{i=1}^n \mathbf{x}_i^{(\rho)}\mathbf{x}_i^{(\rho)}/n,$$

which is just the sample covariance matrix of the $\mathbf{x}_i^{(\rho)}$. Yuan et al. (2000) proposed using (15) as a downweighting transformation formula. Working with several practical data sets, they found that the transformed sample $\mathbf{x}_i^{(\rho)}$ is much better approximated by a normal distribution than the original sample $\mathbf{x}_i$. As we shall see, when applying the bootstrap to a sample $\mathbf{x}_i^{(\rho)}$, the test statistic $T_{\text{ML}}$ is approximately pivotal. For $\mathbf{x}_i^{(\rho)}$ we have

$$E(\mathbf{x}_{i1}^{(\rho)} \mathbf{x}_{i2}^{(\rho)} \mathbf{x}_{i3}^{(\rho)} \mathbf{x}_{i4}^{(\rho)}) = E[u_2^2(d_i^2)(x_{i1} - \mu_1)(x_{i2} - \mu_2)(x_{i3} - \mu_3)(x_{i4} - \mu_4)] + o(1), \quad (16)$$

where $\mathbf{x}_{ij}^{(\rho)}$ is the $j$th element of $\mathbf{x}_i^{(\rho)}$. Because the denominator of $u_2(d_i^2)$ in (14) contains $d_i^2$ when $d_i^2 > r^2$, (16) is bounded. Even when the fourth-order moments of $F_\Sigma$ do not exist, the corresponding fourth-order moments of the $\mathbf{x}_i^{(\rho)}$ are still finite.

When interest lies in the structure of the population covariance matrix $\Sigma = \text{Cov}(\mathbf{x}_i)$, one needs to know whether transformation (15) changes the structure of $\Sigma$. Let $\Sigma^{(\rho)}$ be the population covariance matrix of $\mathbf{x}_i^{(\rho)}$; then $\Sigma^{(\rho)} = \kappa \Sigma$ when the sampling distribution is elliptically symmetric. For almost all the commonly used models in the social sciences, modelling $\Sigma$ is equivalent to modelling $\Sigma^{(\rho)}$ (Browne, 1984). Yuan et al. (2000) further illustrated that outlying cases create significant multivariate skewness, and that the sample covariance matrix $\hat{\mathbf{S}}_\mathbf{x}$ is biased in such a situation. Analysis based on $\mathbf{x}_i^{(\rho)}$ can
successfully remove the bias in $S_x$ and recover the correct model structure. More details on the merits of robust approaches to CSA can be found in Yuan and Bentler (1998) and Yuan et al. (2000).

5. Applications

Our purpose in this section is to compare the performance of $T_{ML}$, $T_B$ and $T_{SB}$ using real data. We will show that, by combining the bootstrap and a downweighting transformation, a nearly optimal procedure may be found for analysing a given data set. If a test statistic $T$ is nearly pivotal, then its evaluations $T_b^*$ at the $B_0$ bootstrap replications should approximately follow a chi-square distribution. There are a variety of tools to evaluate this property. We favour the quantile–quantile (QQ) plot because of its visualization value. If $T$ follows $\chi^2_d$, the plot of ordered $T_b$’s against the quantiles of $\chi^2_d$ should form an approximately straight line with slope $\sqrt{2/2}$. A significant departure from this line indicates violations of the pivotal property. To save space, the QQ plots are presented in a document available on the Internet (http://www.nd.edu/~kyuan/papers/bsinfpowfig.pdf). We choose $B_0 = B_1 = B = 1000$ in estimating $\hat{\beta}^*_\alpha$ and $\hat{\beta}^*_\alpha$, although a smaller $B$ may be enough in obtaining these estimates (Efron & Tibshirani, 1993).

Example 1. Holzinger and Swineford (1939) provide a data set consisting of 24 cognitive variables on 145 subjects. We use nine of the variables in our study: visual perception, cubes, and lozenges, measuring a spatial ability factor; paragraph comprehension, sentence completion, and word meaning, measuring a verbal ability factor; addition, counting dots, and straight-curved capitals, measuring (via the imposition of a time limit) a speed factor. Let $x$ represent the nine variables; then the confirmatory factor model

$$\mathbf{x} = \Lambda \mathbf{f} + \mathbf{e}, \quad \text{Cov}(\mathbf{x}) = \Lambda \Phi \Lambda' + \Psi$$

(17a)

with

$$\Lambda = \begin{pmatrix} 1.0 & \lambda_{21} & \lambda_{31} & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1.0 & \lambda_{52} & \lambda_{62} & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1.0 & \lambda_{83} & \lambda_{93} \end{pmatrix},$$

$$\Phi = \begin{pmatrix} \phi_{11} & \phi_{12} & \phi_{13} \\ \phi_{21} & \phi_{22} & \phi_{23} \\ \phi_{31} & \phi_{32} & \phi_{33} \end{pmatrix},$$

(17b)

represents Holzinger and Swineford’s hypothesis. We assume that the measurement errors are uncorrelated, with $\Psi = \text{Cov}(\mathbf{e})$ being a diagonal matrix. There are $q = 21$ unknown parameters, and the model degrees of freedom are 24.

Mardia’s (1970) multivariate kurtosis for the nine-variable model is 3.04, implying that the data may come from a distribution with heavier tails than those of a normal distribution. We therefore apply the downweighting transformation (15) in order for $T_{ML}$ to be approximately pivotal. Based on previous research (Yuan et al., 2000, 2001) we apply the three statistics to three samples: the raw sample $\mathbf{x}_i$, and the transformed samples $\mathbf{x}_i^{(0.25)}$ and $\mathbf{x}_i^{(0.25)}$. The following bootstrap procedures include further transforming these samples to satisfy $H_0$ or $H_1$ as given in Section 2.
QQ plots for the three statistics on the raw data set suggest that no statistic is approximately pivotal. All have heavier right tails than that of $\chi^2_{24}$. This is expected for $T_{ML}$ because the sample $x_i$ exhibits heavier tails than those of a normal distribution. Previous studies suggest that a large sample size is needed for $T_B$ to behave like a chi-square random variable. A sample size of $n = 145$ may not be large enough or there may be other unknown factors that prevent $T_B$ from behaving like a chi-square variable. Even though many simulation studies recommend $T_{SB}$, this statistic certainly has a heavier right tail than that of $\chi^2_{24}$ for this data set.

When applied to the transformed sample $x_i^{(25)}$, all the three statistics still have heavier right tails than that of $\chi^2_{24}$. When applied to $x_i^{(25)}$, the QQ plots indicate that $T_{ML}$ is approximately described by $\chi^2_{24}$, but $T^*_B$ or $T^*_B$ are not. Consequently, the suggested procedure for this data set is to apply $T_{ML}$ to the transformed sample $x_i^{(25)}$. Bootstrapping with $T_{ML}$ on this sample not only leads to more efficient parameter estimates (Yuan et al., 2001) but also provides a more accurate Type I error estimate.

Table 1 gives the significance levels of model (17) evaluated by various procedures. Several differences exist among these procedures. First, all the bootstrap $p$-values ($p_B$) are greater than those ($p_{\chi^2}$) obtained by referring the test statistics to $\chi^2_{24}$. The only comparable pair of $p_B$ and $p_{\chi^2}$ is when $T_{ML}$ is applied to the sample $x_i^{(25)}$. Second, the $T_B$ statistic for each sample is the largest among the three statistics, but the $p_B$ for $T_B$ is also the largest for any of the three samples. This is in contrast to conclusions regarding the performance of $T_B$ when referring to a chi-square distribution (Hu et al., 1992; Fouladi, 2000; Yuan & Bentler, 1997). Third, all the $p_B$s based on any statistic with any of the samples are quite comparable, implying the robustness of bootstrap inference. The above phenomena can be further explained by the critical values $\hat{c}^{*}_{25}$ in Table 2. All the $\hat{c}^{*}_{25}$s are greater than $\chi^2_{24}(95)^{-1} = 36.415$, the 95th percentile of $\chi^2_{24}$. When the heavier tails are downweighted in the sample, the $T^*$s corresponding to each of the three statistics behave more like $\chi^2_{24}$ and the corresponding $\hat{c}^{*}_{25}$ is nearer 36.415. Although $T_B$ is the largest among the three statistics on any of the samples, the corresponding $\hat{c}^{*}_{25}$ for $T_B$ is also the largest, explaining why the associated $p_B$s for $T_B$ can also be the largest. On the other hand, the traditional inference procedure is to refer $T_B$ to $\chi^2_{24}$ for significance. With such a fixed reference distribution that does not take the increased variability of $T_B$ into account, a larger $T_B$ generally corresponds to a smaller $p_{\chi^2}$.

Table 1. Statistics, bootstrap $p$-values ($p_B$) and $p$-values ($p_{\chi^2}$) referring to $\chi^2_{24}$

<table>
<thead>
<tr>
<th>Statistic</th>
<th>Sample $x_i$</th>
<th>$p_B$</th>
<th>$p_{\chi^2}$</th>
<th>Sample $x_i^{(25)}$</th>
<th>$p_B$</th>
<th>$p_{\chi^2}$</th>
<th>Sample $x_i^{(25)}$</th>
<th>$p_B$</th>
<th>$p_{\chi^2}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$T_{ML}$</td>
<td>51.187</td>
<td>.017</td>
<td>.001</td>
<td>49.408</td>
<td>.010</td>
<td>.002</td>
<td>48.883</td>
<td>.004</td>
<td>.002</td>
</tr>
<tr>
<td>$T_{SB}$</td>
<td>48.696</td>
<td>.020</td>
<td>.002</td>
<td>50.318</td>
<td>.013</td>
<td>.001</td>
<td>53.674</td>
<td>.003</td>
<td>.000</td>
</tr>
<tr>
<td>$T_B$</td>
<td>56.726</td>
<td>.055</td>
<td>.000</td>
<td>62.095</td>
<td>.024</td>
<td>.000</td>
<td>66.910</td>
<td>.014</td>
<td>.000</td>
</tr>
</tbody>
</table>

The power properties of $T_{ML}$, $T_B$ and $T_{SB}$ are evaluated at two alternatives. Starting with model (17), the parameter $\lambda_{a1}$ is the only significant path identified by the Lagrange multiplier test in EQS (Bentler, 1995). Adding this extra parameter to (17b), the first alternative $\Sigma_{a1}$ is the estimated covariance matrix obtained by fitting this model
to the raw sample \( x_i \) by minimizing \( D_{ML} \). The second alternative \( \Sigma_{a2} \) is the sample covariance matrix of the raw sample. With these two alternatives the NCPs in (9) for the three statistics are given in Table 3. Notice that, once a \( \Sigma \) is given, the \( d \) in (9) corresponding to \( T_{ML} \) is fixed and does not depend on the sample. Because both \( T_{SB} \) and \( T_B \) involve fourth-order sample moments, their corresponding \( d \)'s change as the tails of the data set change. QQ plots for the ordered \( T^* \) s against quantiles of the \( \chi^2_{24}(\delta) \) indicate that the distributions of \( T_{ML}^* \) and \( T_{SB}^* \) applied to \( x \) are well approximated by the corresponding non-central chi-squares. Because \( T_{ML} \) under \( H_0 \) applied to \( x_i \) is well approximated by \( \chi^2_{24}(22.60) \) and \( \chi^2_{24}(51.19) \) respectively. This expectation is not fulfilled as judged from the corresponding QQ plots. The reason for this is that, even for perfectly normal data simulated from \( N_p(\mu, \Sigma) \), the statistic \( T_{ML} \) will not behave like a non-central chi-square unless \( \delta \) is quite small and \( n \) is large. Because the NCP tends to be overestimated when it is large (Satorra, Saris, & de Pijper, 1991), the corresponding QQ plots are below the \( x = y \) line.

Can we trust the results of power analysis in the traditional approach when \( T_{ML} \) and \( T_{SB} \) are well approximated by non-central chi-square distributions in this example? The answer is no. This is because the critical value \( \chi^2_{24}(0.95)^{-1} = 36.415 \), which does not take the actual variability of \( T_{ML} \) or \( T_{SB} \) into account, is not a good estimate of \( c_\alpha \) in (3). As can be seen from Table 2, there exist quite substantial differences between \( \chi^2_{24}(0.95)^{-1} \) and the \( \hat{c}_{05} \)'s. Actually, power analysis based on a non-central chi-square table is not reliable for even perfectly normal data when \( \Sigma(\theta) \) is not near \( \Sigma_a \). Table 4 contrasts the power estimates based on the bootstrap (\( \hat{\beta}_\alpha \)) with those based on the traditional approach (\( \hat{\beta}_\alpha \)) using NCPs in Table 3. The \( \hat{\beta}_\alpha \)'s are uniformly smaller than the \( \hat{\beta}_\alpha \)'s, especially when the NCPs are not huge. When an NCP is large enough, there is almost no overlap between the distribution of \( (T|H_a) \) and that of \( (T|H_1) \) Any sensible inference procedure can tell the difference between the two, and consequently the power estimates under \( H_{a2} \) in Table 4 are approximately the same.

**Table 2. Bootstrap critical values \( \hat{c}_{05} \)**

<table>
<thead>
<tr>
<th>Statistic</th>
<th>Sample ( x_i )</th>
<th>Sample ( x_{i}^{(05)} )</th>
<th>Sample ( x_{i}^{(25)} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( T_{ML} )</td>
<td>42.348</td>
<td>39.991</td>
<td>37.022</td>
</tr>
<tr>
<td>( T_{SB} )</td>
<td>41.694</td>
<td>41.067</td>
<td>40.587</td>
</tr>
<tr>
<td>( T_B )</td>
<td>57.930</td>
<td>55.120</td>
<td>53.925</td>
</tr>
</tbody>
</table>

**Table 3. Non-centrality parameter estimates associated with each statistic and sample**

<table>
<thead>
<tr>
<th>Statistic</th>
<th>( H_{a1} )</th>
<th>( H_{a2} )</th>
<th>( H_{a1} )</th>
<th>( H_{a2} )</th>
<th>( H_{a1} )</th>
<th>( H_{a2} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( T_{ML} )</td>
<td>22.603</td>
<td>51.187</td>
<td>22.603</td>
<td>51.187</td>
<td>22.603</td>
<td>51.187</td>
</tr>
<tr>
<td>( T_{SB} )</td>
<td>21.763</td>
<td>48.695</td>
<td>23.107</td>
<td>51.901</td>
<td>24.875</td>
<td>55.948</td>
</tr>
<tr>
<td>( T_B )</td>
<td>20.289</td>
<td>56.726</td>
<td>20.130</td>
<td>63.187</td>
<td>20.683</td>
<td>69.516</td>
</tr>
</tbody>
</table>
Table 4 illustrates the power properties of each statistic when the data change. When the outlying cases are downweighted, not only do the parameter estimates become more efficient (Yuan et al., 2001), but also the powers for identifying an incorrect model are higher for all the statistics. The powers of $T_{ML}$ and $T_{SB}$ are quite comparable, while the power of $T_B$ is much lower. This is in contrast to the conclusion for power analysis based on the non-central chi-square, where $T_B$ has the highest power in detecting a wrong model (Fouladi, 2000; Yuan & Bentler, 1997).

Next, we focus on testing model (17) for close fit. We will only present results of the bootstrap using $T_{ML}$. The use of $T_{SB}$ or $T_B$ is essentially the same. MacCallum et al. (1996) recommended testing close fit (RMSEA $\leq .05$), fair fit ($.05 <$ RMSEA $\leq .08$), mediocre fit ($.08 \leq$ RMSEA $\leq .10$), and poor fit (.10 < RMSEA). We will conduct a bootstrap procedure for these tests. With (10), it is easy to see that RMSEA $= [\min_{\theta} D(\Sigma_c, \Sigma(\theta))/(p^* - q)]^{1/2} = \epsilon$ corresponds to $\delta_c = n(p^* - q)\epsilon^2$. The $\delta_c$'s corresponding to $\epsilon = .05, .08, \text{ and } .10$ are given in the second row of Table 5. The 95th percentiles of the corresponding $\chi^2_{24}(\delta_c)$ are in the third row of the table. These are, respectively, fixed critical values for testing close fit, fair fit, and mediocre fit in the approach of MacCallum et al. (1996).

Table 5. Test for close fit

<table>
<thead>
<tr>
<th>Statistic</th>
<th>Sample $x_i$</th>
<th>Sample $x_i^{(05)}$</th>
<th>Sample $x_i^{(25)}$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$H_{a1}$</td>
<td>$H_{a2}$</td>
<td>$H_{a1}$</td>
</tr>
<tr>
<td>$T_{ML}$</td>
<td>$\hat{\beta}_a$</td>
<td>.627</td>
<td>.993</td>
</tr>
<tr>
<td></td>
<td>$\hat{\beta}_a$</td>
<td>.803</td>
<td>.998</td>
</tr>
<tr>
<td>$T_{SB}$</td>
<td>$\hat{\beta}_a$</td>
<td>.613</td>
<td>.989</td>
</tr>
<tr>
<td></td>
<td>$\hat{\beta}_a$</td>
<td>.782</td>
<td>.997</td>
</tr>
<tr>
<td>$T_B$</td>
<td>$\hat{\beta}_a$</td>
<td>.374</td>
<td>.983</td>
</tr>
<tr>
<td></td>
<td>$\hat{\beta}_a$</td>
<td>.743</td>
<td>.999</td>
</tr>
</tbody>
</table>

Table 4. Bootstrap power estimates ($\hat{\beta}_a^*$) and power estimates ($\hat{\beta}_a$) referring to $\chi^2_{24}(\delta_c)$
Let $S_x$ be the sample covariance matrix of $x_i$ and $\hat{\theta}$ be the corresponding maximum likelihood estimate of $\theta$. The values of $b$ in the solution for $\Sigma_c$ in the form of $\Sigma_c = (1-b)\Sigma(\hat{\theta}) + bS_x$, corresponding to $\epsilon = .05$, .08, and .10, are respectively 0.42161, 0.66981 and 0.83083. Applying the bootstrap procedure, outlined in Section 2.4, to the samples $x_i$, $x_i^{(.05)}$ and $x_i^{(.25)}$, the estimates $\hat{c}_a^*$ and $p$-values for each of the samples are reported in Table 5. The results indicate that there are substantial differences between the traditional $p$-values $p_{\chi^2}$ and the bootstrap $p$-values $p_B$, especially when RMSEA = .10 for the downweighted samples. It is obvious that $T_{ML}$ when RMSEA = .10 has a much shorter right tail than that of $\chi^2_{24}(\delta_c)$. This fact is also reflected in the corresponding $\hat{c}_a^*$'s. Table 5 once again illustrates the fact that, given a close fit, the behaviour of $T_{ML}$ cannot be described by a non-central chi-square unless the corresponding $T_{ML}$ under $H_0$ in (1) has a heavy right tail. As when testing for exact fit, using a chi-square table to judge the significance of a statistic in testing model (17) for close fit is misleading.

Example 2. Neumann (1994) presented an alcohol and psychological symptom data set consisting of 10 variables and 335 cases. The two variables in $x = (x_1, x_2)'$ are respectively family history of psychopathology and family history of alcoholism, which are indicators for a latent construct of family history. The eight variables in $y = (y_1, \ldots, y_8)'$ are respectively: the age of first problem with alcohol; age of first detoxification from alcohol; alcohol severity score; alcohol use inventory; SCL-90 psychological inventory; the sum of the Minnesota Multiphasic Personality Inventory scores; the lowest level of psychosocial functioning during the past year; and the highest level of psychosocial functioning during the past year. With two indicators for each latent construct, these eight variables respectively measure: age of onset; alcohol symptoms; psychopathology symptoms; and global functioning. Neumann's (1994) theoretical model for this data set is

\begin{align}
\mathbf{x} &= \Lambda_x \xi + \delta, \quad \mathbf{y} = \Lambda_y \eta + \epsilon, \\
\eta &= \mathbf{B} \eta + \Gamma \xi + \zeta,
\end{align}

where

\begin{align}
\Lambda_x &= \begin{pmatrix} 1.0 \\ \lambda_1 \end{pmatrix}, \quad \Lambda_y = \begin{pmatrix} 1 & \lambda_2 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & \lambda_3 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & \lambda_4 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & \lambda_5 \end{pmatrix}, \\
\mathbf{B} &= \begin{pmatrix} \beta_{21} & 0 & 0 & 0 \\ \beta_{31} & \beta_{32} & 0 & 0 \\ 0 & \beta_{42} & \beta_{43} & 0 \end{pmatrix}, \quad \Gamma = \begin{pmatrix} \gamma_{11} & 0 \\ 0 & 0 \end{pmatrix}, \quad \Phi = \text{Var}(\xi),
\end{align}

and $\epsilon$, $\delta$, and $\zeta$ are vectors of errors whose elements are all uncorrelated. The model degrees of freedom are 29.

With Mardia's multivariate kurtosis equal to 14.76, the data set may come from a distribution with heavy tails. A bootstrap procedure is more appropriate after the heavy tails are properly downweighted. Actually, Yuan et al. (2001) found that the sample $\mathbf{x}_i^{(.10)}$ leads to the most efficient parameter estimates in (18) among various procedures. Here we apply the three statistics to the two samples $\mathbf{x}_i$ and $\mathbf{x}_i^{(.10)}$. Our purpose is to
explore the pivotal property of $T_{ML}$, $T_{SB}$ and $T_B$ on each sample. After noticing that a large portion of the QQ plot for $T_{ML}$ applying to $x_i^{(10)}$ is below the $x = y$ line, our analysis also includes the sample $x_i^{(05)}$.

The QQ plots of the three statistics applied to $x_i$ indicate that none of them is nearly pivotal. $T_{ML}^*$ and $T_{SB}^*$ are for the most part above the $x = y$ line; however, both their left tails are slightly below the $x = y$ line. This implies that some bootstrap samples fit model (18) extremely well, a phenomenon which can be caused by too many data points near the centre of the distribution. A downweighting procedure only affects data points that cause a test statistic to possess a heavy right tail. It is not clear to us how to deal with a sample when a test statistic has a light left tail. The QQ plots of $T_{ML}^*$ and $T_{SB}^*$ on $x_i^{(05)}$ and $x_i^{(10)}$ suggest that their heavier right tails on $x_i$ are under control. However, the right tail of $T_B$ is still quite heavy when compared to $\chi^2_{29}$. The downweighting transformation $H_i(\cdot, 0.30)$ not only just controls the right tail of $T_{ML}^*$ but also makes most of the QQ plot fall below the $x = y$ line. Visually inspecting the QQ plots of $T_{ML}^*$ on $x_i$, $x_i^{(05)}$, and $x_i^{(10)}$ suggests that a downweighting transformation by $H_i(\cdot, 0.05)$ with $0 < \rho < 0.05$, may lead to a better procedure for analysing the alcohol and psychological symptom data set based on $T_{ML}$. Since any procedure is only an approximation to the real world and $T_{ML}$ is nearly pivotal when applying to $x_i^{(05)}$, we recommend the analysis using $T_{ML}$ on $x_i^{(05)}$ for this data set. This leads to $T_{ML} = 42.43$ with a bootstrap $p$-value of 0.040, implying that the model (18) marginally fits the alcohol and psychological symptom data set.

Both the raw data sets in Examples 1 and 2 have significant multivariate kurtoses and the downweighting transformation (15) achieves approximately pivotal behaviour of the statistics. We may wonder how these test statistics behave when applied to a data set that does not have a significant multivariate kurtosis. This is demonstrated in the following example.

Example 3. Table 1.2.1 of Mardia, Kent, and Bibby (1979) contains test scores of $n = 88$ students on $p = 5$ topics: mechanics, vectors, algebra, analysis, and statistics. The first two topics were tested with closed-book exams and the last three with open-book exams. Since these two examination methods may tap different abilities, a two-factor model as in (17a) with

$$
\Lambda = \begin{pmatrix}
1.0 & \lambda_{21} & 0 & 0 & 0 \\
0 & 0 & 1.0 & \lambda_{42} & \lambda_{52}
\end{pmatrix}, \quad \Phi = \begin{pmatrix}
\phi_{11} & \phi_{12} \\
\phi_{21} & \phi_{22}
\end{pmatrix}
$$

was proposed and confirmed by Tanaka, Watadani, and Moon (1991). Mardia’s multivariate kurtosis for this data set, equal to 0.057, is not significant. It would be interesting to see how $T_{ML}$, $T_B$ and $T_{SB}$ perform on this data set.

QQ plots for the three statistics applied to $x_i$ indicate that all of them have heavier right tails than that of $\chi^2_{0.2}$. The QQ plot for $T_{ML}^*$ applied to $x_i^{(\rho)}$ continues to exhibit a heavier right tail until $\rho$ reaches 0.30. $T_{SB}^*$ and $T_B$ still possess quite heavy right tails even when $\rho = 0.30$.

This data set has been used as an example for influence analysis. Previous studies indicate that case number 81 is the most influential point (Lee & Wang, 1996; Fung & Kwan, 1995). So it is enticing to apply the three statistics to the $x_i$’s without the 81st case. The QQ plots of $T_{ML}^*$, $T_{SB}^*$ and $T_B^*$ applied to the remaining 87 cases indicate that all three statistics still have heavier right tails than that of $\chi^2_{0.2}$. Actually, compared to those based on all 88 cases, the right tails of the three statistics on the 87 cases are even heavier! For this data set, our recommended analysis is to use $T_{ML}$ applied to $x_i^{(0.30)}$. With
6. Non-convergence with Bootstrap Replications

Non-convergence issues exist with resampling and caution is needed for bootstrap inference (Ichikawa & Konishi, 1995). This generally happens when the sample size is not large enough, and especially when a model structure is wrong. This issue was discussed in Yuan and Bentler (1997) with Monte Carlo studies on a covariance structure model. Here, we propose a reasonable way of handling non-convergence with bootstrap replications.

With an iterative algorithm like Newton’s, convergence is generally defined as \(|\Delta \theta^{(j)}| < \epsilon\), where \(\epsilon\) is a small number and \(\Delta \theta^{(j)} = \theta^{(j)} - \theta^{(j-1)}\), with \(\theta^{(j)}\) being the \(j\)th step solution. Let \(\sigma = \text{vech}(\Sigma)\) be the factor formed by stacking the non-duplicated elements of \(\Sigma\), and let its sample counterpart be \(s = \text{vech}(\hat{S})\). Then \(\Delta \theta^{(j+1)} = (\sigma_j W_j \sigma)^{-1} \sigma_j W_j (s - \sigma)\), where \(\sigma_j = [d\sigma_j / d\theta_j \theta^{(j)}]\), \(\sigma_j = \sigma \theta^{(j)}\) and \(W_j\) is the corresponding weight matrix evaluated at \(\theta^{(j)}\). So \(\Delta \theta^{(j)}\) is proportional to \(\hat{S} - \Sigma(\theta^{(j-1)})\), and it is impossible for \(|\Delta \theta^{(j)}|\) to be smaller than \(\epsilon\) if \(\Sigma(\theta)\) is far from \(\hat{E}(S)\). Although a model is correct in a bootstrap population, with bootstrap replications result in convergent solutions, although a model is correct in a bootstrap population, and it is impossible for \(|\Delta \theta^{(j)}|\) to be smaller than \(\epsilon\) if \(\Sigma(\theta)\) is far from \(\hat{E}(S)\). Although a model is correct in a bootstrap population, and it is impossible for \(|\Delta \theta^{(j)}|\) to be smaller than \(\epsilon\) if \(\Sigma(\theta)\) is far from \(\hat{E}(S)\). Although a model is correct in a bootstrap population.

Based on this fact, we should distinguish two kinds of non-convergence in bootstrap or general Monte Carlo studies. The first is where a sample contains enough distinct points and still cannot reach convergence with a model, which should be treated as a significant replication or a ‘bad model’. The second is where a sample does not contain enough distinct observations to fit a model. For obtaining a \(T_{\text{ML}}\), this number is \(p + 1\); the number is \(p^* + 1\) for obtaining a \(T_{\text{SB}}^*\). Although a \(T_{\text{SB}}^*\) can be obtained once a \(T_{\text{ML}}\) is available, we need to have a positive definite sample covariance matrix \(S_y\) of \(y_i = \text{vech}(x_i - \mathbf{X})(x_i - \mathbf{X})'\) in order for \(T_{\text{SB}}^*\) to make sense (Bentler & Yuan, 1999), and the minimum number of distinct data points for \(S_y\) to be positive definite is \(p^* + 1\). We should treat the second case as a bad sample and ignore it in bootstrap replications.

In practice, instead of estimating \(c_{a}^*\) one generally reports the \(p\)-value of a statistic \(T\). If all \(B\) bootstrap replications result in convergent solutions,

\[
p_B = \frac{B_0 + 1 - M}{B_0 + 1},
\]

where \(M = \#(T_{\text{SB}}^* > T)\). When non-convergences occur, \(B_0\) should be defined as the number of converged samples plus the number of significant samples due to a bad model; \(M\) should be defined as the number of non-converged samples due to the bad model plus the number of converged samples that result in \(T_{\text{SB}}^* > T\). With this modification, there is no problem with hypothesis testing. A similar modification applies to formula (8) for power evaluation. With \(c_{a}^* = T_{(1-\alpha)}^*\) in determining the numerator in (8), one needs at least \(B_0(1 - \alpha)\) converged samples when sampling from \(\hat{F}_0\).

The proposed procedure fails to generate a power estimate when the number of convergences under \(H_0\) is below \(B_0(1 - \alpha)\).

Let \(\lambda_1(S_x) \geq \lambda_2(S_x) \geq \ldots \geq \lambda_p(S_x)\) be the eigenvalues of the sample covariance matrix \(S_x\), and \(\gamma_1(S_y) \geq \gamma_2(S_y) \geq \ldots \geq \gamma_{p^*}(S_y)\) be the eigenvalues of the sample

\(T_{\text{ML}} = 1.89\) and a bootstrap \(p\)-value of .71, model (19) is more than good enough in explaining the relationship of the five variables.
covariance matrix \( S_y \) of \( y_i = \text{vech}[(x - \overline{x})(x - \overline{x})'] \). In computing the three examples, our criteria for bad samples are \( \lambda_p(S_x)/\lambda_1(S_x) \leq 10^{-20} \) for obtaining \( T_{ML} \) and \( \gamma_p(S_y)/\gamma_1(S_y) \leq 10^{-20} \) for obtaining \( T_{SB} \) and \( T_B \). If these criteria are satisfied and the model still cannot reach convergence (\( \|\Delta \theta\| < 10^{-5} \)) in 100 iterations, we treat the corresponding statistic as infinity. All non-convergences with the three examples were due to significant replications; these are reported in Table 6. Table 6(a) implies that, for a given sample, the further a model is from \( H_0 \), the more often non-convergences occur. The table also suggests that, with a given model, the heavier the tails of a data set, the more often one obtains non-convergences.

**Table 6. Non-convergences due to significant samples**

(a) Example 1

<table>
<thead>
<tr>
<th>Statistic</th>
<th>Sample ( x_i )</th>
<th>Sample ( x_i^{(0.05)} )</th>
<th>Sample ( x_i^{(0.25)} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( T_{ML} )</td>
<td>0</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>( T_{SB} )</td>
<td>0</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>( T_B )</td>
<td>0</td>
<td>10</td>
<td>21</td>
</tr>
</tbody>
</table>

(b) Example 2

<table>
<thead>
<tr>
<th>Statistic</th>
<th>Sample ( x_i )</th>
<th>Sample ( x_i^{(0.05)} )</th>
<th>Sample ( x_i^{(0.10)} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( T_{ML} )</td>
<td>9</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>( T_{SB} )</td>
<td>9</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>( T_B )</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

(c) Example 3

<table>
<thead>
<tr>
<th>Statistic</th>
<th>Sample ( x_i )</th>
<th>Sample ( x_i^{(0.30)} )</th>
<th>Sample ( x_i ), (81st case removed)</th>
</tr>
</thead>
<tbody>
<tr>
<td>( T_{ML} )</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>( T_{SB} )</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>( T_B )</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

7. Discussion and conclusions

Existing conclusions regarding the three commonly used statistics \( T_{ML} \), \( T_{SB} \) and \( T_B \) are based on asymptotics and Monte Carlo studies. The properties inherent in either of these approaches may not be enjoyed by these statistics in practice because of either a finite sample or an unknown sampling distribution. By applying them to a specific data
set through resampling, properties of the three statistics can be visually examined by means of QQplots. When a data set possesses heavy tails, \( T_{ML} \) will inherit these tails by having a heavy right tail. Actually, all three statistics need the sampling distribution to have finite fourth-order moments. With possible violation of this assumption by practical data, we propose applying a bootstrap procedure to a transformed sample \( x_i^{(\rho)} \) through downweighting. The combination of bootstrapping and downweighting not only offers a theoretical justification for applying the bootstrap to a data set with heavy tails, but also provides a quite flexible tool for exploring the properties of each of the three statistics. Even if inference is based on referring a statistic to a chi-square distribution, one will obtain a more accurate model evaluation by applying \( T_{ML} \) to a transformed sample \( x_i^{(\rho)} \).

Asymptotics justify the three statistics from different perspectives, and \( T_B \) is asymptotically distribution-free. Previous Monte Carlo studies mainly support \( T_{SB} \). However, with proper down weighting, \( T_{ML} \) is generally the one that is best described by a chi-square distribution. However, the conclusion that \( T_{ML} \) is the best statistic for bootstrap inference has to be preliminary. Future studies may find \( T_B \) or \( T_{SB} \) more appropriate for other data sets. We recommend exploring the different procedures for a given data set, as illustrated in Section 5.

**Acknowledgements**

The authors would like to thank Professor Peter M. Bentler and two anonymous referees whose constructive comments lead to an improved version of this paper. Correspondence concerning this article should be addressed to Ke-Hai Yuan (kyuan@nd.edu).

**References**


Received 9 November 2000; revised version received 22 March 2002