Assessing overall fit is a topic of keen interest to structural equation modelers, yet measuring goodness of fit has been hampered by several factors. First, the assumptions that underlie the chi-square tests of model fit often are violated. Second, many fit measures (e.g., Bentler and Bonett's [1980] normed fit index) have unknown statistical distributions so that hypothesis testing, confidence intervals, or comparisons of significant differences in these fit indices are not possible. Finally, modelers have little knowledge about the distribution and behavior of the fit measures for misspecified models or for nonnested models. Given this situation, bootstrapping techniques would appear to be an ideal means to tackle these problems. Indeed, Bentler's (1989) EQS 3.0 and Jöreskog and Sörbom's (forthcoming) LISREL 8 have bootstrap resampling options to bootstrap fit indices. In this article the authors (a) demonstrate that the usual bootstrapping methods will fail when applied to the original data, (b) explain why this occurs, and, (c) propose a modified bootstrap method for the chi-square test statistic for model fit. They include simulated and empirical examples to illustrate their results.

Bootstrapping Goodness-of-Fit Measures in Structural Equation Models

Kenneth A. Bollen
University of North Carolina–Chapel Hill

Robert A. Stine
University of Pennsylvania

Assessing overall fit is a topic of keen interest to structural equation modelers, yet measuring goodness of fit has been hampered by several factors. First, the assumptions that underlie the chi-square tests of model fit often are violated by the approximate nature of the model, excessive kurtosis of the observed random variables, or a small to moderate sample size. Second, many fit measures (e.g., Bentler and Bonett's [1980] normed fit index) have

AUTHORS' NOTE: *We wish to thank Kwok-fai Ting for RA work done for this project and the anonymous referees for their comments. This research was presented at the Department of Economics, Universitat Pompeu Fabra, Barcelona, Spain, and at the 1992* Social Science Methodology Conference *in Trent, Italy. Partial support for Bollen's research on this project came from NSF, SES-8908361 and SES-9121564.*

SOCIOLOGICAL METHODS & RESEARCH, Vol. 21, No. 2, November 1992 205-229

unknown statistical distributions so that hypothesis testing, confidence intervals, or comparisons of significant differences in these fit indices are not possible.[1] Finally, we have little knowledge about the distribution and behavior of the fit measures for misspecified models or for nonnested models.

Given this situation, bootstrapping techniques would appear to be an ideal means to tackle these problems. Indeed, Bentler's (1989) EQS 3.0 and Jöreskog and Sörbom's (forthcoming) LISREL 8 have bootstrap resampling options to bootstrap fit indices. The purposes of this article are: (a) to demonstrate that the usual bootstrapping methods will fail when applied to the original data, (b) to explain why this occurs, and, (c) to propose a modified bootstrap method for the chi-square test statistic for model fit.

We begin with a section that reviews the bootstrapping method. Two sections on hypothesis testing with bootstrapping procedures follow: the first on tests of means and the second on the likelihood ratio chi-square tests in structural equation models. In these two sections, we demonstrate and explain the failure of the usual bootstrapping scheme and we present modified bootstrap procedures that have superior performance. The brief section that follows explains how to apply the modified bootstrap procedure to chi-square difference tests for nested models. Results of the previous two sections are then illustrated in an empirical example. Next we examine whether such modifications work for other fit indices for structural equation models. The article concludes by reviewing the major results and by describing the remaining problems facing bootstrap applications in this area.

NAIVE BOOTSTRAPPING

Several introductions to the bootstrap methodology are available (e.g., Efron and Tibshirani 1986; Efron 1982; Stine 1989). In brief the usual bootstrap method is as follows. Let $\{X_1, X_2, \ldots, X_N\}$ be a random sample of size N with each X_i independently drawn from the same population that has a cumulative distribution function G that is characterized by a parameter θ. Our estimator of θ is symbolized as $\hat{\theta}$, and we wish to know its sampling distribution. For the random vari-

ables, $\{X_1, X_2, \ldots, X_N\}$, we observe a given sample $\{x_1, x_2, \ldots, x_N\}$. The bootstrap method samples from the population defined by the *empirical* distribution function G_N to estimate the sampling distribution of $\hat{\theta}$. The empirical distribution function, G_N, is that distribution with mass 1/N on $x_1, x_2, \ldots, x_N$. That is, we form a bootstrap sample, $\{X_1^*, X_2^*, \ldots, X_N^*\}$, by taking N independent draws *with replacement* from $\{x_1, x_2, \ldots, x_N\}$. The bootstrap estimate, $\hat{\theta}^*$, is computed using the same formula as that for $\hat{\theta}$, but it is calculated using the bootstrap sample. Repeating this process B times gives $\hat{\theta}^*(1), \hat{\theta}^*(2), \ldots, \hat{\theta}^*(B)$. From these bootstrap replicates of $\hat{\theta}$ we estimate the bootstrap distribution of $\hat{\theta}$ including its mean and variance.

Though this general bootstrap procedure works well in many cases, it can fail. For example, bootstrapping gives a misleading impression of the distribution of the maximum of a sample (Bickel and Freedman 1981). *The success of the bootstrap depends on the sampling behavior of a statistic being the same when the samples are drawn from the empirical distribution and when they are taken from the original population.* In some situations this assumption does not hold and the resulting bootstrap estimates perform poorly.[2] Bootstrapping transformed data can sometimes correct the problem. The sections that follow illustrate these points.

HYPOTHESIS TESTS ON THE MEAN

Consider the distribution of the usual Z statistic when testing the null hypothesis that $\mu = 0$ when $X_1, \ldots, X_N$ are a sample from a normal distribution with mean μ and known variance equal to 1. Then the distribution of the test statistic

$$Z = \frac{\bar{X}}{\frac{\sigma}{\sqrt{N}}} = \sqrt{N}\,\bar{X} \qquad [1]$$

is normal with mean zero and variance 1 under H_o. Thus Z^2 is chi-square with 1 degree of freedom and yields a test equivalent to the usual two-sided Z test. The procedure rejects the null hypothesis when

Z^2 exceeds, say, 3.84 which is the upper 5% point in the chi-square distribution with 1 degree of freedom (df). More generally, the chi-square distribution with 1 df is the basis for finding the *p* value associated with the observed test statistic.

Consider the following bootstrap approach to evaluate the *p* value of this test. As in the "naive" bootstrap scheme, resample the data and generate B copies of the observed statistic. That is, repeat the following procedure B times:

1. resample $\{x_1, x_2, \ldots, x_N\}$ to obtain $\{x_1^*, x_2^*, \ldots, x_N^*\}$
2. compute $\bar{x}^*$, the mean of the bootstrap sample,
3. form $z^{*2} = N\bar{x}^{*2}$.

Though one might be tempted to use the collection $\{z^{*2}(1), z^{*2}(2), \ldots, z^{*2}(B)\}$ to assess the significance of our observed test statistic, to do so would be wrong. To demonstrate this we generated 150 observations from the N(0,1) distribution. Then we applied the preceding bootstrap algorithm. Figure 1 shows the chi-square distribution with 1 degree of freedom (solid line) and the bootstrap distribution of Z^{*2} based on 500 bootstrap samples (long dash, short dash).[3] The difference is dramatic. The bootstrap density is shifted to the right with a peak near 2. In contrast the chi-square density is monotonically decreasing over the same range. The 0.05 critical value based on the bootstrap is about 13 compared to 3.84 based on a chi-square with 1 df. The bootstrap critical value of 13 is determined so that only 5% of the bootstrap replications Z^{*2} exceed this value.

The bootstrap fails in this simple example because the fundamental bootstrap assumption does not hold. That is, bootstrap resampling from the observations does not resemble sampling from a population in which the null hypothesis holds. This failure manifests itself in a variety of ways, but it is perhaps most easily seen by comparing the expected value of Z^2 to that of the bootstrap replications. Notice that we do not need simulations; all of the calculations are exact, as though we had performed an infinite number of replications ($B = \infty$).

First, consider the expected value of Z^2. Suppose that $X_1, X_2, \ldots, X_N$ are a sample from a population with mean μ and variance σ^2. It follows that

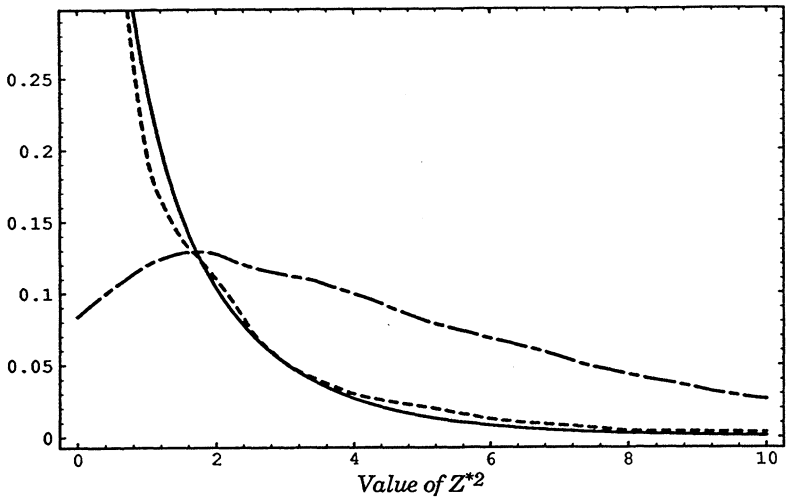


Figure 1: Comparison of the Chi-Square Density With 1 Degree of Freedom (Solid Line) to the Naive Bootstrap Density (— - —) and Modified Bootstrap Density (- - -) of Z^{*2}

$$\begin{aligned} E(Z^2) &= N\,E(\bar{X} - \mu + \mu)^2 \\ &= N\,E[(\bar{X} - \mu)^2] + N\mu^2 \\ &= \sigma^2 + N\mu^2, \end{aligned} \qquad [2]$$

where the second equality follows from noting that $E(\bar{X} - \mu) = 0$ and the third is a restatement of the result that $VAR(\bar{X}) = \sigma^2/N$. In the prior example, $X_1, X_2, \ldots, X_N$ are N(0,1) and $E(Z^2)$ reduces to 1, which is the expected value of a chi-square random variable having 1 degree of freedom.

To find the expected value of Z^{*2}, we basically repeat the familiar steps leading to equation 2. We must, however, compute the expected value with respect to the population associated with the bootstrap observations X_i^*. Since the X_i^* are drawn from the empirical distribution G_N, we need to compute the expected values with respect to this population rather than to G. Let E* denote the expected value that puts weight 1/N on each of $x_1, x_2, \ldots, x_N$. The expected value with respect to E* gives averages associated with infinite bootstrap resampling

from G_N without needing simulation. The two essential results we require are

$$E^*(X_i^*) = \sum_{i=1}^{N} x_i/N = \bar{x} \quad [3]$$

and

$$VAR^*(X_i^*) = E^*(X_i^* - \bar{x})^2 = v^2 \quad [4]$$

where

$$v^2 = \sum_{i=1}^{N} (x_i - \bar{x})^2/N \quad [5]$$

As a consequence, when computing expected values with E*, one replaces the usual parameters μ and σ^2 by the sample values $\bar{x}$ and v^2. Note that v^2 is the Maximum Likelihood Estimator (MLE) of σ^2 under normality and it is almost the usual unbiased estimator of σ^2, differing by having the divisor of N rather than N – 1. It follows from equations 3 and 4 that the expected value of bootstrap replications of the sample mean is

$$E^*(\bar{X}^*) = \bar{x} \quad [6]$$

and the variance is

$$VAR^*(\bar{X}^*) = v^2/N. \quad [7]$$

Note that the original observations $x_1, x_2, \ldots, x_N$ are treated as fixed constants when applying E*, so all of these expectations are in fact conditional on the observed sample.

With these results in hand, we can compute the expected value of the bootstrap replications of Z^{*2} and discover the source of the problem in this naive resampling scheme. Following the derivations of equations 2, 6, and 7, they imply that

$$\begin{aligned} E^*(Z^{*2}) &= N E^*(\bar{X}^*)^2 \\ &= N E^*(\bar{X}^* - \bar{x} + \bar{x})^2 \\ &= N E^*[(\bar{X}^* - \bar{x})]^2 + N\bar{x}^2 \\ &= v^2 + N\bar{x}^2 \\ &= v^2 + z^2 \end{aligned} \quad [8]$$

Thus, the expected value of the bootstrap replicates Z^{*2} is the observed value of the test statistic plus an additional term that approaches σ^2 as N increases. Considering the specific illustration with $X_1, X_2, \ldots, X_N$ distributed N(0,1), equation 2 implies that the $E(Z^2) = 1$, whereas equation 8 shows that the expected value of Z^{*2} is

$$E(V^2 + Z^2) = \frac{N-1}{N} + 1 = 2 - 1/N \qquad [9]$$

As a result, Z^2 averages near 1 but the average of Z^{*2} is about 2.

We should point out that other features of the distribution of Z^{*2} are also misleading. For example, Z^{*2} has greater sampling variance than Z^2 and its distribution is fundamentally different. One cannot correct this problem by a simple change of location with rescaling. Babu (1984) gives a thorough, technical description of these issues, and Freedman (1981) describes a related problem occurring in regression analysis without an intercept.

So where does this bootstrapping scheme go wrong? Recall that the significance of a test statistic like Z^2 is the probability of observing a statistic of that size or larger *when the null hypothesis is true*. Consider the test of H_o: $\mu = 0$ using Z^2. Under the naive resampling as done in the example, the mean of the bootstrap population, namely the average of the observed sample, is almost surely not zero. Thus these bootstrap samples are drawn from a population for which the null hypothesis does not hold, regardless of whether H_o holds for the unknown population from which $X_1, X_2, \ldots, X_N$ are drawn. Hence the bootstrap values of the test statistic reject H_o too often. It is only natural that Z^2 appears in the expression in equation 8 for the expected value of Z^{*2} since it represents how poorly the null hypothesis holds in the observed sample.

In this simple case, it is relatively easy to repair the bootstrap scheme. To make the null hypothesis true under bootstrapping resampling, center the data around the sample mean $\bar{x}$ and draw bootstrap samples from $\{(x_1 - \bar{x}), (x_2 - \bar{x}), \ldots, (x_N - \bar{x})\}$. Resampling from these centered values forces the mean of the bootstrap population to be zero so that H_o holds. The resulting bootstrap replicates provide an estimated p value of the test statistic. To illustrate this we took the same simulated data as before except that we formed deviation scores before

applying the bootstrapping resampling. The third curve (short dashes) in Figure 1 is the modified bootstrap distribution. The centered resampling scheme leads to a more accurate estimate of the distribution of Z^2 and is closer to the chi-square distribution (solid line). For example, the χ^2 with 1 df critical value is 3.84 for a Type I error of 0.05. Centered resampling gives a mean of 1.2 and a critical value of 4.9, that is, 4.9 is greater than all but 5% of the bootstrap replicates. Recall that for the naive bootstrap distribution, the corresponding critical value is about 13.[4]

LIKELIHOOD RATIO TEST STATISTIC

The subject of this section is the application of the bootstrap method to the distribution of the likelihood-ratio-based test statistic used in structural equation models. The Maximum Likelihood (ML) fitting function is

$$F_{ML}(S, \Sigma(\theta)) = \ln|\Sigma(\theta)| + tr(S\Sigma^{-1}(\theta)) - \ln|S| - p \qquad [10]$$

where $\Sigma(\theta)$ is the population covariance matrix implied by a model, θ is the $t \times 1$ vector that contains the free parameters of the model, **S** is the unbiased sample covariance matrix of the observed variables, and p is the number of observed variables.[5] A vector $\hat{\theta}$ is chosen to minimize equation 10. Then the fitting function is used to test the null hypothesis, H_o: $\Sigma = \Sigma(\theta)$. The test statistic $T = (N - 1)\, F_{ML}(\mathbf{S}, \Sigma(\hat{\theta}))$ is asymptotically distributed as a noncentral chi-square variate with degrees of freedom equal to $\frac{1}{2}(p + 1)p - t$ and noncentrality parameter κ, provided that the observed variables have no excessive multivariate kurtosis. The noncentrality parameter κ equals $(N - 1)F_{ML}(\Sigma, \Sigma(\theta_o))$ with θ_o chosen to minimize $F_{ML}(\Sigma, \Sigma(\theta))$ over choices of θ (Browne 1984, p. 70, corollary 4.1).

Naive bootstrapping of the test statistic involves repeating the following procedure B times:

1. resample the original data and form **S***, the covariance matrix for the bootstrap sample,
2. fit the hypothesized model to **S*** by minimizing $F_{ML}(\mathbf{S}^*, \Sigma(\theta^*))$ over choices of θ^*, and,

3. compute $T^* = (N-1)\,F_{ML}(S^*, \Sigma(\hat{\theta}^*))$ where $\hat{\theta}^*$ is the value of θ^* that minimizes $F_{ML}(S^*, \Sigma(\theta^*))$.

The test statistic T for the original sample would then be compared to the bootstrap distribution of T*(1), T*(2), . . . , T*(B) for tests of statistical significance.

Naive bootstrapping of the chi-square statistic for structural equation models is inaccurate. To understand why consider the expected values and variances of T and T*. Since T [$= (N-1)F_{ML}(S, \Sigma(\hat{\theta}))$] is distributed as a noncentral chi-square variable in large samples, we have

$$AE(T) = df + \kappa \qquad [11]$$

where AE(•) is the asymptotic expectation, df stands for the degrees of freedom, and κ is the noncentrality parameter. If the model is valid and the other assumptions for the test hold, then κ is zero and T follows a central chi-square distribution. The variance of T is

$$AVAR(T) = 2df + 4\kappa \qquad [12]$$

where AVAR(•) denotes the asymptotic variance. For a valid model where κ is zero, equation 12 reduces to 2 df.

The distribution of T* approximates a noncentral chi-square distribution when resampling from the bootstrap population. In the original population, the noncentrality parameter κ equals $(N-1)$ $F_{ML}(\Sigma, \Sigma(\theta_o))$. In the bootstrap population, the noncentrality parameter is $(N-1)F_{ML}(S, \Sigma(\hat{\theta}))$ which equals T.[6] Consequently, we are led to the approximation

$$E^*(T^*) \approx df + T \qquad [13]$$

Taking expectations with respect to the original population and substituting from equation 13 gives

$$\begin{aligned} E[E^*(T^*)] &\approx df + df + \kappa \\ &\approx 2df + \kappa \end{aligned} \qquad [14]$$

Similarly, the approximate bootstrap variance of T* is

$$VAR^*(T^*) \approx 2df + 4T \qquad [15]$$

where VAR*(•) represents the variance over the bootstrap samples. The expectation of equation 15 with respect to the original population is

$$E[VAR^* (T^*)] \approx 6df + 4\kappa \quad [16]$$

Hence, the mean of the bootstrap distribution of T* is larger than the mean of T by approximately df, and the variance of T* exceeds that of T. The observed value of the test statistic T is the noncentrality value in the bootstrap population. H_0 is violated when bootstrapping regardless of whether it holds in the actual population. As a result, the bootstrap approximation fails to match the sampling distribution associated with the original population.

These results are *approximations* for two reasons. The noncentral chi-square distribution for T is an asymptotic result, and the expression for the noncentrality parameter κ assumes that the observed variables originate from a distribution with no excess kurtosis.[7] Even if the original data are a sample from a multinormal distribution, a given sample can appear very nonnormal leading to excessive kurtosis in the bootstrap population. Also, the bootstrap distribution is discrete whereas the original population distribution is continuous. Despite these limitations we have found these approximations helpful in understanding and predicting the results of naive bootstrapping of the test statistic.

As with the chi-square test of H_0: $\mu = 0$, it is possible to develop a bootstrap procedure that applies to this more complex model. The idea again is to resample a set of observations for which the null hypothesis is true. For a test of H_0: $\Sigma = \Sigma(\theta)$, the bootstrap population from which we resample must have the covariance structure specified by the null hypothesis.[8]

To generate such a population requires a little matrix algebra, including the idea of the square root of a matrix. Let $\mathbf{Y}$ denote the $N \times p$ data matrix of the centered observed variables. Also, $\mathbf{S} = \mathbf{Y}'\mathbf{Y}/(N - 1)$ denotes the sample covariance matrix of $\mathbf{Y}$ and $\hat{\Sigma}$ $(= \Sigma(\hat{\theta}))$ is the estimated implied covariance matrix. Let $\mathbf{M} = \mathbf{M}^{1/2\prime}\mathbf{M}^{1/2}$ represent a square root factorization of the positive definite matrix $\mathbf{M}$, such as that given by a Cholesky factorization.[9] Then the covariance matrix of

$$\mathbf{Z} = \mathbf{Y}\mathbf{S}^{-1/2}\hat{\boldsymbol{\Sigma}}^{1/2} \quad [17]$$

is indeed $\hat{\boldsymbol{\Sigma}}$ since

$$\begin{aligned}\mathbf{Z}'\mathbf{Z}/(N-1) &= \hat{\boldsymbol{\Sigma}}^{1/2}\mathbf{S}^{-1/2}\mathbf{Y}'\mathbf{Y}\mathbf{S}^{-1/2}\hat{\boldsymbol{\Sigma}}^{1/2}/(N-1)\\ &= \hat{\boldsymbol{\Sigma}}^{1/2}\mathbf{S}^{-1/2}\mathbf{S}\mathbf{S}^{-1/2}\boldsymbol{\Sigma}^{1/2}\\ &= \hat{\boldsymbol{\Sigma}}^{1/2}\mathbf{S}^{-1/2}\mathbf{S}^{1/2}\mathbf{S}^{1/2}\mathbf{S}^{-1/2}\hat{\boldsymbol{\Sigma}}^{1/2}\\ &= \hat{\boldsymbol{\Sigma}}^{1/2}\hat{\boldsymbol{\Sigma}}^{1/2}\\ &= \hat{\boldsymbol{\Sigma}}. \end{aligned} \quad [18]$$

Now perform the bootstrap by resampling the rows of **Z** rather than the original observations in the rows of **Y**. The test statistic T calculated for the covariance matrix of **Z** will be zero since the sample covariance matrix of **Z** equals $\hat{\boldsymbol{\Sigma}}$.

The modified bootstrapping procedure involves repeating the following procedure B times:

1. resample the modified data, **Z**, and form $\mathbf{S}_m^*$, the covariance matrix for the bootstrap sample of the modified data,
2. fit the hypothesized model to $\mathbf{S}_m^*$ by minimizing $F_{ML}(\mathbf{S}_m^*, \Sigma(\theta_m^*))$ over choices of θ_m^*, and,
3. compute $T_m^* = (N-1)\,F_{ML}(\mathbf{S}_m^*, \Sigma(\hat{\theta}_m^*))$ where $\hat{\theta}_m^*$ is the value of θ_m^* that minimizes $F_{ML}(\mathbf{S}_m^*, \Sigma(\theta_m^*))$.

The means and variances of the modified test statistic, T_m^*, are

$$\begin{aligned} E^*(T_m^*) &\approx df \\ E[E^*(T_m^*)] &\approx df \end{aligned} \quad [19]$$

and

$$\begin{aligned} VAR^*(T_m^*) &\approx 2df \\ E[VAR^*(T_m^*)] &\approx 2df \end{aligned} \quad [20]$$

The bootstrap sampling distribution of T_m^*, the modified test statistic, behaves as the sampling distribution of T from the original population when the null hypothesis is true. As before, we note that these results are approximations for the reasons stated previously.

We illustrate these ideas with a simulation of a single latent variable (ξ_1) measured with eight indicators (x_1 to x_8):

$$x_i = \lambda_{i1}\xi_1 + \delta_i \quad [21]$$

where i = 1 to 8, λ_{i1} is a coefficient giving the effect of ξ_1 on x_i, δ_i represents random measurement error with $E(\delta_i) = 0$ and $COV(\delta_i, \xi_1) = 0$. The eight values of the λ_{i1}s are 1, 1.2, 1.4, 1.6, 1, 1.2, 1.4, and 1.6, respectively. The variances of the δ_is are 2, 4, 6, 8, 2, 4, 6, and 8, respectively, and the variance of the factor is 1. We generated data from normal distributions that conform to the model for an N of 150. The test statistic T for this sample is 23.1. The critical value for a chi-square variate with 20 df is 31.4 for a Type I error of 0.05. Thus, the data are consistent with H_o: $\Sigma = \Sigma(\theta)$.

We then applied the naive bootstrap method to this sample, that is, we resampled with replacement of the original sample, estimated the covariance matrix for the bootstrap sample, and calculated T* for each bootstrap sample. Next, we used the previously described modified bootstrap procedure where we transformed the original **Y** to **Z** and then bootstrapped the data of **Z**. Figure 2a overlays the bootstrap distribution of T_m^* from the modified procedure (short dash) on top of the Monte Carlo simulated distribution of T (long dash) and a chi-square distribution with 20 degrees of freedom (solid line). In addition the naive bootstrap distribution of T* is plotted (long dash, short dash).

The bootstrap approximation to the density based on T* is shifted to the right and has greater variance than the chi-square density. These properties are consistent with the approximations of equations 14 and 16. Clearly the naive bootstrap distribution is quite inaccurate. The density for the modified bootstrap density and the simulated density are both less peaked and more long-tailed than the chi-square density. Figure 2b magnifies the right tails of these three densities. The modified bootstrap approximation is the most long-tailed, with the simulated estimate between the bootstrap and the chi-square. The longer tail for the Monte Carlo simulation than for the chi-square is consistent with other Monte Carlo simulation studies and is consistent with the finding that the usual test statistic rejects the null hypothesis too frequently in small samples (e.g., Boomsma 1982).

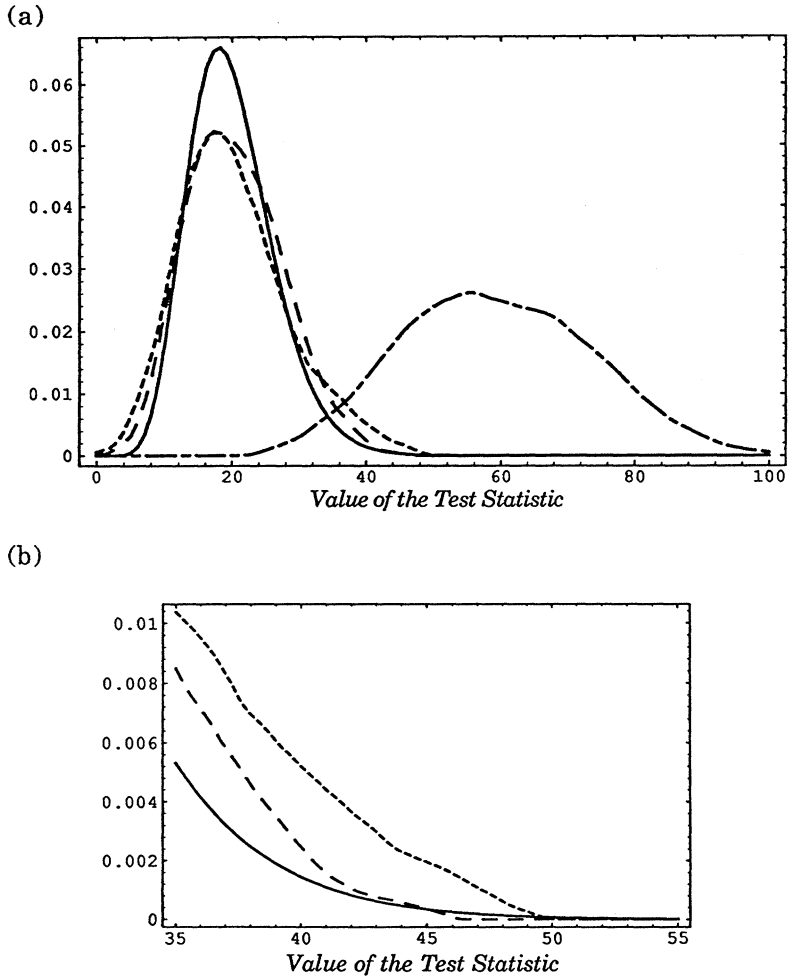


Figure 2: **Comparison of the Chi-Square Approximation With 20 Degrees of Freedom (Solid Line) and the Simulated Density of T (— — —) to the Naive Bootstrap Density of T* (— - —) and Modified Bootstrap Density of T^*_m(- - - -): (a) Complete Densities, (b) Right Tails, Omitting the Naive Bootstrap Density**

TESTING NESTED MODELS

Chi-square difference tests are commonly used for comparing the fit of nested models. It would be useful to have a method to bootstrap

a test statistic that enables us to compare nested models. The preceding modified bootstrapping procedure applies to this test statistic as well. The null hypothesis for the chi-square test is that the more restrictive model is valid and the less restrictive model is not required. The key to applying the modified method is to transform the data so that the null hypothesis is true in the bootstrap population. That is, apply the transformation of $\mathbf{Y}$ to $\mathbf{Z}$ (see equation 17) using the $\hat{\Sigma}$ fitted under the most restrictive model. Fit both models to the same bootstrapped covariance matrix and form the difference in the test statistics. The empirical example that follows illustrates this procedure.

INDUSTRIALIZATION AND POLITICAL DEMOCRACY EXAMPLE

To further illustrate the modified bootstrap procedure, we use a model of the relationship between industrialization and political democracy in 75 developing countries (see the path diagram in Figure 3). It is a panel model with two latent endogenous variables, political democracy in 1960 (η_1) and 1965 (η_2), and one latent exogenous variable, industrialization in 1960 (ξ_1). Industrialization is measured with three indicators, and political democracy for 1960 and 1965 is measured with the same four indicators. For model A, shown in Figure 3, the measurement errors for the same indicator at two points in time are allowed to correlate, as are the errors for measures that are ratings from the same judge. For model B, we restrict all measurement error covariances to zero. Detailed descriptions of the variables, models, and the ML estimates are in Bollen (1989).[10]

Bootstrap methods are helpful in this situation since we do not know whether the sample is sufficiently large to rely on the asymptotic chi-square distribution for the test. Using the ML fitting function and associated test statistics we find T_A is 39.6 with 38 df for model A and T_B is 73.6 with 44 df for model B. Comparing these to the corresponding chi-square distributions, we find that model A has quite a good fit ($p = 0.40$) whereas model B has a worse fit ($p < 0.01$). Performing a nested chi-square difference test we have ($T_B - T_A$) equal to 34 with 6 df which is highly significant ($p < 0.001$). Thus the usual methods

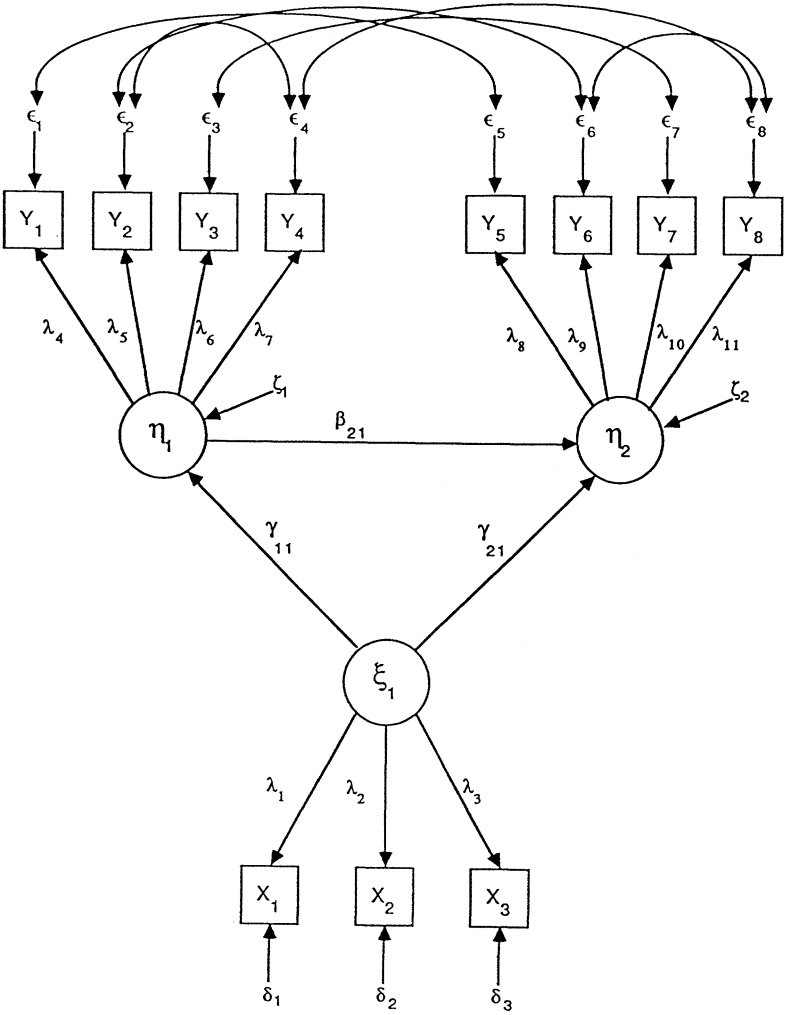


Figure 3: Path Diagram, Model A

would lead us to conclude that model A fits well but model B without correlated measurement errors has a very poor fit.

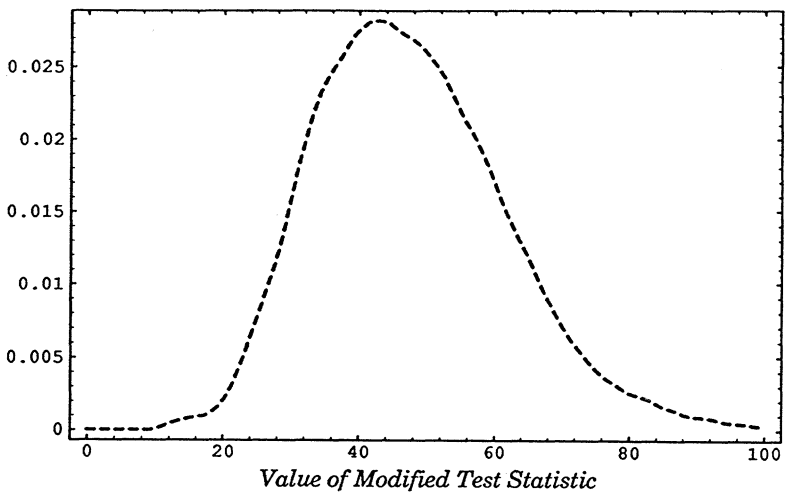


Figure 4: Modified Bootstrap Density of the Test Statistic T_m^* in Industrialization and Political Democracy Model B

We then formed the modified bootstrap distribution assuming that the more restrictive, simpler model B is correct. Figure 4 shows this distribution based on 250 replications. The T_B of 73.6 has a bootstrap *p* value of 0.055, that is, T_B is larger than all but 5.5% of the bootstrap replications under H_o. The resulting *p* value is considerably higher than that found with the usual methods. For this example, model B has a much better fit based on the bootstrap distribution than was found using the usual chi-square-based test. We also used the modified bootstrap procedure to analyze the change in fit of model B versus model A. Here again we assume model B is correct for the modified bootstrap. Figure 5 plots the bootstrap density for differences in the bootstrap test statistics for models B and A (dash) and a chi-square distribution with 6 df (solid). Note also the difference in the chi-square and bootstrap distributions in this case. The bootstrap *p* value for a difference of 34 is near zero since none of the bootstrapped differences was this large. So for this example the modified bootstrapped test statistic suggests a different *p* value for model B, but still indicates that model A is much superior, giving a highly significant improvement in the test statistic.

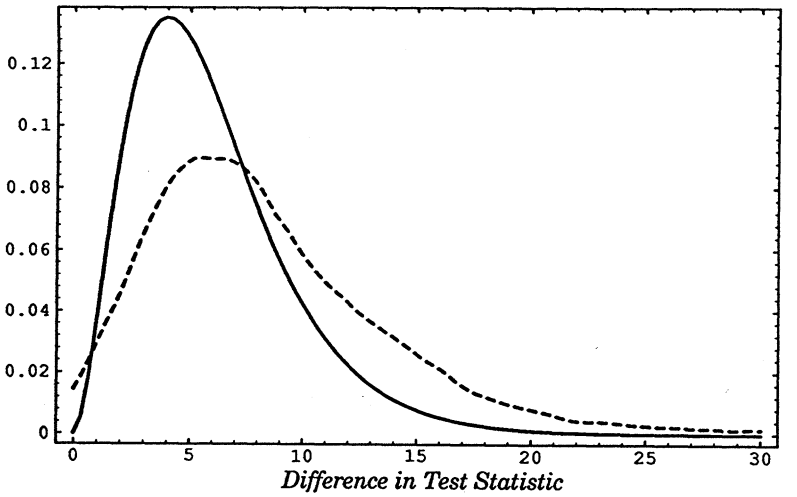


Figure 5: Comparison of the Chi-Square Density With 6 Degrees of Freedom (Solid Line) to the Modified Bootstrap Density (- - -) of the Difference in the Test Statistics for Models A and B

MODIFIED BOOTSTRAP METHOD AND OTHER FIT INDICES

Naive bootstrapping also leads to inaccurate representations of the distribution of other overall fit indices. Table 1 lists some overall fit indices used in structural equation models. For instance, Bentler and Bonett's (1980) normed fit index ($\hat{\Delta}_1$) is

$$\hat{\Delta}_1 = \frac{T_b - T_h}{T_b} \qquad [22]$$

where the subscript of b stands for the baseline model, usually taken to be a model of uncorrelated variables, and h denotes the hypothesized model. The asymptotic expectation of its numerator is

$$AE(T_b - T_h) = df_b + \kappa_b - (df_h + \kappa_h) \qquad [23]$$

The asymptotic expectation of the denominator is

$$AE(T_b) = df_b + \kappa_b \qquad [24]$$

TABLE 1: Overall Fit Indices in Structural Equation Models[a]

$\hat{\Delta}_1 = \dfrac{T_b - T_h}{T_b}$	$\hat{\Delta}_2 = \dfrac{T_b - T_h}{T_b - df_h}$
$\hat{\rho}_2 = \dfrac{\dfrac{T_b}{df_b} - \dfrac{T_h}{df_h}}{\dfrac{T_b}{df_b} - 1}$	$\widehat{AIC} = T_h - 2df_h$
$\widehat{GFI} = 1 - \dfrac{tr[(\hat{\Sigma}^{-1}S - I)^2]}{tr[(\hat{\Sigma}^{-1}S)^2]}$	$\widehat{RMR} = [2 \sum_{i=1}^{q} \sum_{j=1}^{i} \dfrac{(s_{ij} - \hat{\sigma}_{ij})^2}{q(q+1)}]^{1/2}$

a. q = number of observed variables in the analysis, s_{ij} = sample covariance between the i and j observed variables, and $\hat{\sigma}_{ij}$ = model implied estimated covariance between the i and j observed variables. See Bollen (1989, pp. 269-81) and references therein for further discussion of these measures and their properties. The Akaike Information measure (AIC) is scaled slightly differently from that appearing in Bollen (1989).

In contrast, approximations for the corresponding quantities for the bootstrapped normed fit index are

$$E[E^*(T_b^* - T_h^*)] \approx 2df_b + \kappa_b - (2df_h + \kappa_h) \quad [25]$$

and

$$E[E^*(T_b^*)] \approx 2df_b + \kappa_b \quad [26]$$

A comparison of the asymptotic expectations of the numerators and denominators from the original population and the corresponding approximations for the bootstrapped samples reveals that they do not match. Though it is not true that AE(numerator/denominator) equals AE(numerator)/AE(denominator), the bootstrap distribution of the normed fit index will differ from the distribution obtained by repeated sampling of the original population. A similar series of steps would reveal the differences in the approximations to the expected values of the numerators and denominators for the other fit indices in Table 1.

In the preceding section we showed how to modify the bootstrap procedure to allow hypothesis testing with the usual test statistic. Does this modification work with the other fit indices? The answer depends on whether the distributions of the fit indices from the modified

bootstrap procedure mirror the corresponding distributions obtained if we resample from the original population. Figure 6 is the plot of the fit indices for the 8-indicator 1-factor simulation model. We plot the bootstrap densities of the fit indices from the modified (short dash) and naive (long dash, short dash) procedures and densities from a Monte Carlo simulation from a multinormal population (solid) having a covariance matrix given by the single-factor model. In Figure 6 the modified bootstrap densities of the fit indices are much closer to the corresponding Monte Carlo densities than are the naive bootstrap densities.

All the indices in Table 1 are at their maxima when the usual null hypothesis [H_o: $\Sigma = \Sigma(\theta)$] is true. Thus, the chi-square test statistic implicitly tests whether the fit indices are at their perfect fit values. In this sense we already have a significance test for the fit indices.[11] However, as with the usual chi-square test statistic, we may find that in a sufficiently large sample we routinely reject the null hypothesis of a perfect fit due to the approximate nature of the models.

An alternative is to apply the modified bootstrap procedure to data where a restrictive baseline model rather than the hypothesized model is taken to be the true model. For a fit index this would correspond to an implicit null hypothesis that its population value takes a value that indicates a poor fit. With the typical baseline model of uncorrelated variables, this hypothesis would be rejected in nearly all cases. With less restrictive baseline models, this need not be true. However, the usual chi-square test statistic based on the modified bootstrap could be used in this situation as well so that direct hypothesis testing of perfect fit or no fit with the fit indices may not be needed.

CONCLUSIONS

In many areas of statistics the bootstrap has proved to be a very useful tool to approximate the distribution of parameter estimates and other statistics. The structural equation area is beginning to see more applications of the bootstrap and the usage is likely to increase with the availability of bootstrapping resampling in EQS 3.0 and LISREL 8. In this article we sounded a warning that the naive bootstrap generally does not work with the test statistics and fit indices commonly

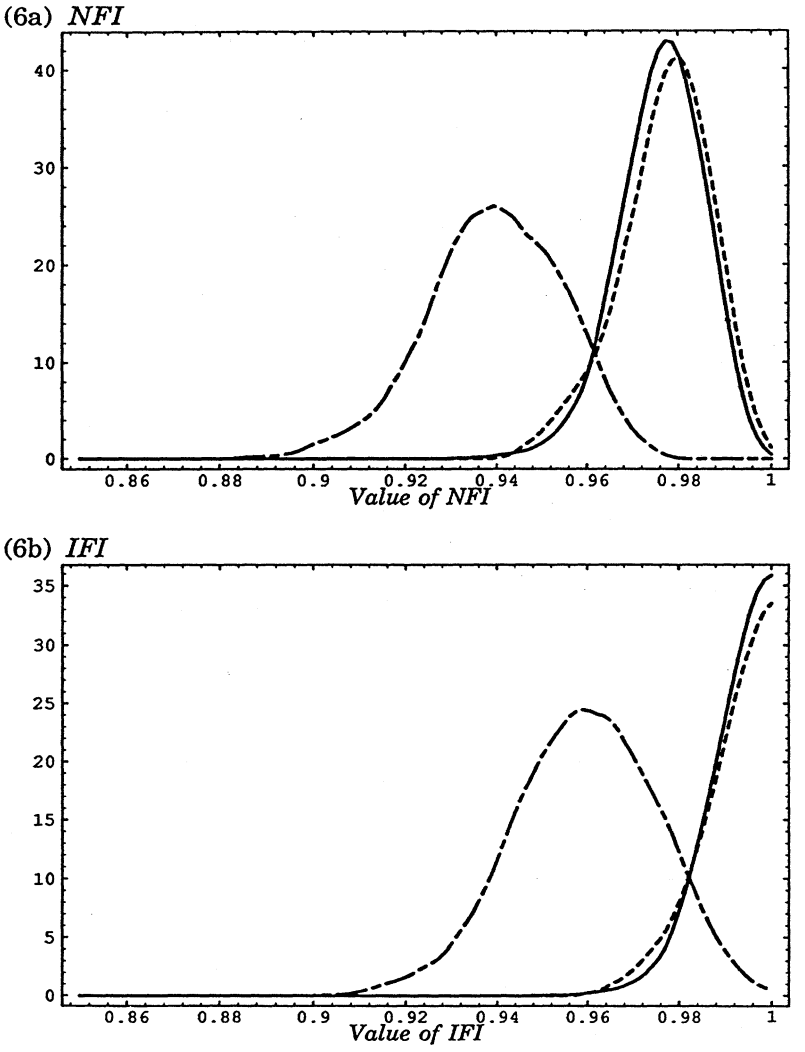
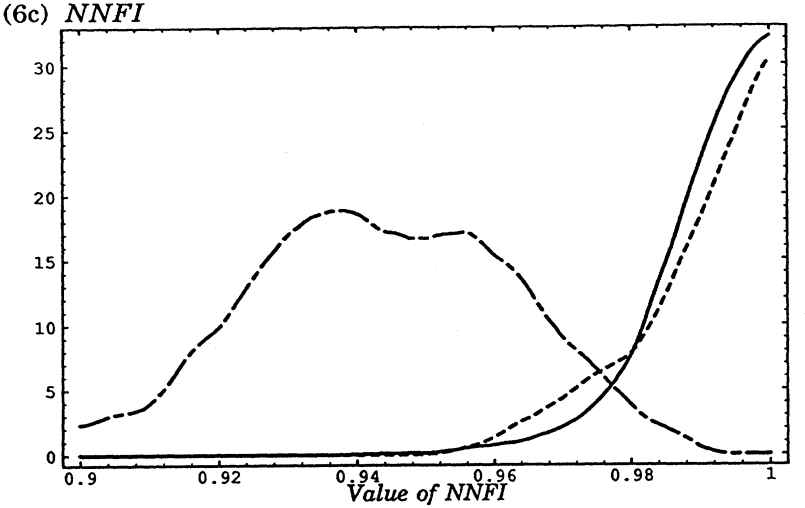


Figure 6: Comparison of Simulated Densities (Solid Line) of Goodness-of-Fit Measures From Table 1 to Naive (— - —) and Modified (- - -) Bootstrap Densities

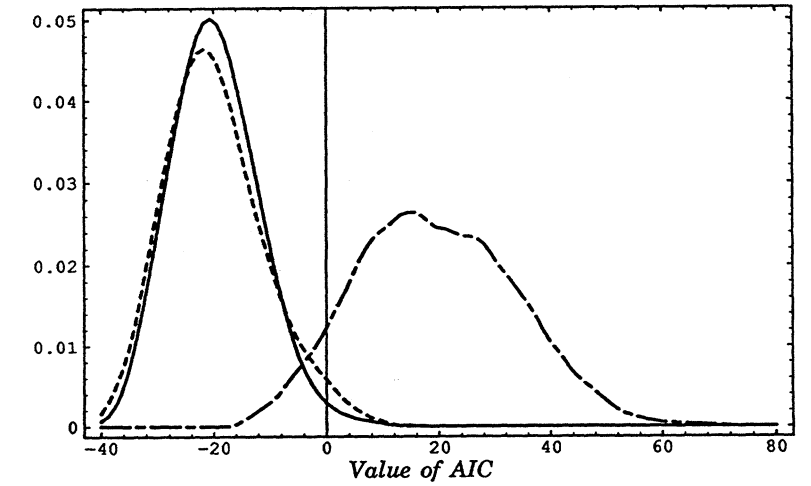
NOTE: NFI, IFI, and NNFI in the figures represent $\hat{\Delta}_1$, $\hat{\Delta}_2$, and $\hat{\rho}_2$, respectively, from Table 1.

(Figure 6 continued)

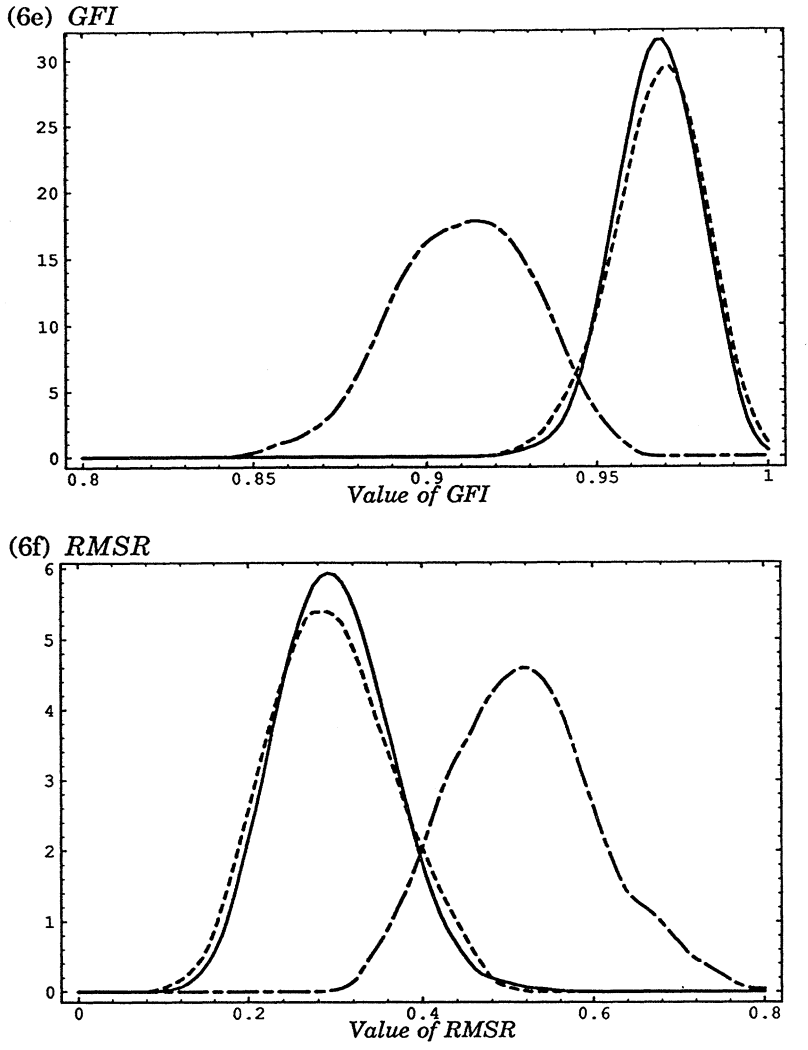
(6c) *NNFI*



(6d) *AIC*



(Figure 6 continued)



employed to gauge overall model fit.[12] Our analytical approximations explained why this was true and our simulation examples illustrated the issue.

In response to this problem we present a modified bootstrap procedure that applies to hypothesis testing with the usual test statistic. With the modification, researchers can gain insight into the behavior of the test statistic with nonnormal data or in moderate sample sizes. We also explained how to test nested hypotheses with the modified bootstrap method.

To implement the modification requires access to matrix programming software such as APL, GAUSS, or SAS's Proc IML that have the ability to find a square root of a positive-definite matrix. With this, the original data can be transformed as described in the article and the bootstrap applied to the transformed data.

The benefits of applying the modified bootstrap procedure to fit indices such as the normed or incremental fit index were less clear-cut. The modification led to bootstrap distributions of the fit indices that were more similar to the distributions that would occur in resampling a population where the hypothesized model is valid. But it may not be necessary to bootstrap the fit indices under the hypothesis of a perfect fit. We can use the chi-square test statistic from the modified bootstrap procedure to test whether the hypothesized model holds and this hypothesis is typically equivalent to the null hypothesis that a fit index in the population equals its maximum (or sometimes minimum) value. Or for nested models, we can test the null hypothesis that the fit indices are equal in the population by using the modified bootstrap procedure for the usual test statistic for nested hypotheses that we proposed. Another alternative for fit indices that contain baseline models is to use the modified bootstrap procedure and generate the transformation so that the baseline model is valid, but this typically creates an overly restrictive model that we will nearly always reject.

A remaining problem is to devise a test of the null hypothesis so that a fit index takes some specific value other than its maximum or minimum (e.g., H_0: incremental fit index = .95). Cudeck and Browne (forthcoming) recently proposed a method to generate data that leads to both a specific minimum value for a fitting function and a corresponding implied covariance matrix. With their method it may be possible to generate an implied covariance matrix that corresponds to a fit between the maximum and minimum. Substituting such an implied covariance matrix into the modified bootstrap procedure may enable tests in which a fit index takes a specific value in the population.

NOTES

1. Recent work by Maiti and Mukherjee (1990) has identified the distribution of Jöreskog and Sörbom's (1986) GFI and AGFI, but their results are based on asymptotic theory so even in this case we do not know when a sample is sufficiently large for such theory to apply.

2. Efron and Tibshirani (1986) give sufficient conditions that permit the use of bootstrap approximations.

3. See Silberman (1986) for a description of the procedures that we use for density estimation in this and in the other figures.

4. Fisher and Hall (1990) provide further discussions about the problems that might be encountered when using bootstrapping methods for hypothesis testing.

5. $F_{ML}(S,\Sigma(\theta))$ derives from either assuming a multinormal distribution for the observed variables or a Wishart distribution for S (see Bollen 1989, pp. 131-35). Browne (1984) justifies the same fitting function by Generalized Least Squares principles using the less restrictive assumption that the observed variables come from a distribution with kurtosis the same as a multinormal distribution, a condition of "no excess kurtosis." No excess kurtosis is less restrictive than multinormality, so we use it throughout the discussion as a distributional assumption for the fitting function in equation 10. Strictly speaking, equation 10 no longer gives the Maximum Likelihood fitting function for nonnormal variables even though it is justified as long as the kurtosis is the same as that for normal variables. However, to avoid new names for the same fitting function and to proceed with the less restrictive assumption of no excess kurtosis, we continue to refer to equation 10 as $F_{ML}(S,\Sigma(\theta))$.

6. This is analogous to the $E^*(Z^{*2})$ result where the observed test statistic is part of the expectation in equation 8.

7. In some problems excessive kurtosis does not matter. See Satorra (1990) and references therein. Also see footnote 3 for a discussion of the distributional assumptions underlying the fitting function.

8. We considered the possibility of modifying T* as a way to correct the bootstrap distribution. For example, the bootstrap distribution of T* – df has the same mean as T but its variance and higher order moments are incorrect.

9. See Beran and Srivastava (1985).

10. See Bollen and Arminger (1991) for an outlier analysis for this model and data.

11. It is possible that the statistical tests based on the modified bootstrap distributions of the fit indices will have different statistical power than the usual test statistic. And this could lead to the desire to use tests based on the fit indices rather than the usual test statistic.

12. Our results do not contradict those in Bollen and Stine (1990) where we showed the usefulness of bootstrap methods in estimating the variability of direct, indirect, and total effects.

REFERENCES

Babu, Gutti Jogesh. 1984. "Bootstrapping Statistics With Linear Combinations of Chi-Squares as Weak Limits." *Sankhya Series A* 46:85-93.

Beran, R. and M. S. Srivastava. 1985. "Bootstrap and Confidence Regions for Functions of a Covariance Matrix." *Annals of Statistics* 13:95-115.

Bentler, Peter M. 1989. *EQS Program Manual.* Los Angeles: BMDP.

Bentler, Peter M. and D. G. Bonett. 1980. "Significance Tests and Goodness-of-Fit in the Analysis of Covariance Structures." *Psychological Bulletin* 88:588-600.

Bickel, P. and D. Freedman. 1981. "Some Asymptotic Theory for the Bootstrap." *Annals of Statistics* 9:1196-1217.

Bollen, Kenneth A. 1989. *Structural Equations With Latent Variables*. New York: Wiley.

Bollen, Kenneth A. and Gerhard Arminger. 1991. "Observational Residuals in Factor Analysis and Structural Equation Models." In *Sociological Methodology 1991*, edited by Peter Marsden. Oxford: Basil-Blackwell.

Bollen, Kenneth A. and Robert Stine. 1990. "Direct and Indirect Effects: Classical and Bootstrap Estimates of Variability." Pp. 115-40 in *Sociological Methodology 1990*, edited by Clifford C. Clogg. Oxford: Basil-Blackwell.

Boomsma, Anne. 1982. "The Robustness of LISREL Against Small Sample Sizes in Factor Analysis Models." Pp. 149-73 in *Systems Under Indirect Observation, Part I*, edited by K. G. Jöreskog and H. Wold. Amsterdam: North-Holland.

Browne, Michael W. 1984. "Asymptotic Distribution Free Methods in the Analysis of Covariance Structures." *British Journal of Mathematical and Statistical Psychology* 37:62-83.

Cudeck, Robert and Michael W. Browne. Forthcoming. "Constructing a Covariance Matrix That Yields a Specified Minimizer and a Specified Minimum Discrepancy Function Value." *Psychometrika*.

Efron, Bradley. 1982. "The Jackknife, the Bootstrap and Other Resampling Plans." CBMS 38, Philadelphia: FIAM. Monograph.

Efron, Bradley and R. Tibshirani. 1986. "Bootstrap Methods for Standard Errors, Confidence Intervals and Other Measures of Statistical Accuracy." *Statistical Science* 1:54-74.

Fisher, Nicholas and Peter Hall. 1990. "On Bootstrap Hypothesis Testing." *Australian Journal of Statistics* 32:177-90.

Freedman, David A. 1981. "Bootstrapping Regression Models." *Annals of Statistics* 9:1218-28.

Jöreskog, Karl and Dag Sörbom. 1986. *LISREL VI*. Mooresville, IN: Scientific Software, Inc.

———. Forthcoming. *LISREL 8*.

Maiti, Sadhan Samar and Bishwa Nath Mukherjee. 1990. "A Note on Distributional Properties of the Jöreskog-Sörbom Fit Indices." *Psychometrika* 55:721-26.

Satorra, Albert. 1990. "Robustness Issues in Structural Equation Modeling: A Review of Recent Developments." *Quality and Quantity* 24:367-86.

Silberman, B. W. 1986. *Density Estimation for Statistics and Data Analysis*. London: Chapman and Hall.

Stine, Robert. 1989. "An Introduction to Bootstrap Methods: Examples and Ideas." *Sociological Methods & Research* 8:243-91.

Kenneth A. Bollen is a professor of sociology at the University of North Carolina–Chapel Hill. His major research interests are international development and statistics. He is the author of Structural Equations With Latent Variables *(1989), published in Wiley's Series in Probability and Mathematical Statistics.*

Robert A. Stine is an associate professor of statistics in the Wharton School of the University of Pennsylvania. His current research areas include resampling methods, time-series analysis, and statistical computing.