Bootstrapping and the Identification of Exogenous Latent Variables Within Structural Equation Models

Gregory R. Hancock and Jonathan Nevitt
Department of Measurement, Statistics, and Evaluation
University of Maryland, College Park

In traditional applications of latent variable models, each exogenous latent variable must either have its variance parameter fixed or a loading path to a measured indicator variable fixed (either customarily to 1). Without doing so the measurement model will suffer from underidentification, thereby yielding no unique solution when estimating the parameters of interest. The choice of whether to fix the variance or the loading is somewhat arbitrary, guided primarily by the researcher’s need for inference regarding particular parameters within the model.

Under conditions of multivariate nonnormal data, the method by which one makes identified the measurement of exogenous latent variables may not be as arbitrary. Specifically, as addressed briefly by Arbuckle (1997), when one is utilizing a bootstrapping approach for generating empirical standard errors for parameters of interest, the researcher must choose to fix an indicator path rather than the latent variable variance in order for the empirical standard errors to be generated properly. This article offers an illustrative explanation of why such an approach is necessary. Given the increased attention toward bootstrapping techniques within structural equation modeling, our hope is that a greater awareness and understanding of this unique situation will be facilitated.

Requests for reprints should be sent to Gregory R. Hancock, 1230 Benjamin Building, University of Maryland, College Park, MD 20742-1115. E-mail: ghancock@wam.umd.edu
THEORETICAL EXPLANATION

Consider the simple one-factor model displayed twice in Figure 1. With only three indicators, this model is just-identified. (Such a model was chosen for the theoretical portion of this article purely for pedagogical purposes and in no way detracts from the generalizability of the explanations contained herein.) In both models within the figure, population path values and exogenous variable (i.e., factor and error) variances are indicated by adjacent numbers when fixed and adjacent letters when free to vary. The left side of the figure shows Model 1 in which the factor variance is free to vary and the first measured variable serves as the factor's scale indicator; the right side of the figure shows Model 2 in which the factor variance is constrained to unity and all indicator paths are free to vary.

In order to fit the models represented in Figure 1, a sample covariance matrix is required. Imagine that a sample of \( N = 200 \) cases has been randomly drawn from a trivariate population. The sample covariance matrix shown on the next page (diagonal and lower triangle) results (note that, although all positive covariances were chosen here for simplicity, the principles conveyed in the example extend to situations where such is not the case, as will be seen later).
Returning to Figure 1, the model-implied covariance matrix for Model 1 is as follows:

\[
\begin{bmatrix}
14.000 \\
7.560 & 15.208 \\
30.660 & 55.188 & 228.518
\end{bmatrix}
\]

Setting these model-implied expressions equal to their corresponding sample variances and covariances yields six equations with six unknowns. If one were to approach this system of equations manually, deriving the value of \(a\) would proceed by using the covariance expressions below the diagonal:

\[
\frac{(ab)(ac)}{(abc)} = a
\]

\[
\frac{(7.560)(30.660)}{(55.188)} = 4.2 = a
\]

Systematically plugging this value back into the set of model-implied equations yields the remaining parameter estimates for this sample: \(b = 1.8, c = 7.3, d = 9.8, e = 1.6,\) and \(f = 4.7.\)

With respect to bootstrapping from the parent sample that yielded the sample covariance matrix, each bootstrap sample will yield a just-identified solution for all parameter estimates in Model 1 as the multivariate fit function of choice is minimized to exactly zero. Estimated empirical standard errors for these parameters may then be derived from the sampling distributions of parameter estimates from the bootstrap samples.

Now consider Model 2 in Figure 1, in which the factor variance has been fixed to unity and all loadings are free to vary. The following model-implied covariance matrix arises:

\[
\begin{bmatrix}
g^2 + j \\
gh & h^2 + k \\
gi & hi & i^2 + m
\end{bmatrix}
\]

As before, one can start to solve the system of equations arising when setting these expressions equal to the corresponding elements of the sample covariance matrix. Be-
gaining with \( g \), the expressions below the diagonal can be used as shown in the following. This time, however, a solution can only be derived for \( g^2 \), rather than for \( g \) itself:

\[
(gh)(gi)/(hi) = g^2
\]

\[
(7.560)(30.660)/(55.188) = 4.2 = g^2
\]

From this expression alone, \( g \) could be either positive or negative. Unfortunately, no other equations inform our choice of sign. If one chooses the positive square root of 4.2, then \( g = 2.05 \). Systematically plugging this value back into the set of equations yields the remaining parameter approximations for this sample: \( h = 3.69 \), \( i = 14.96 \), \( j = 9.80 \), \( k = 1.59 \), and \( m = 4.72 \). On the other hand, if one chooses the negative square root of 4.2, then \( g = -2.05 \) and the remaining parameter estimates are \( h = -3.69 \), \( i = -14.96 \), \( j = 9.80 \), \( k = 1.59 \), and \( m = 4.72 \). Thus, whereas the error variance estimates \( (j, k, \text{ and } m) \) are identical in both solutions, the loading values are all opposite in sign.

Now, if one were to draw bootstrap samples from the parent sample, Model 2 parameter estimates could yield both types of solutions (i.e., differing in sign of loadings). In general, the relative frequency of these two solutions would depend on such conditions as sample size, shape of score distributions in the parent sample, strength and sign of covariances, and start values used by the structural equation modeling software. Thus, because two global minima exist for the multivariate fit function, the bootstrap sampling distribution for the loading parameter estimates should be somewhat bimodal. This will result in inflated bootstrap standard errors, which will yield downwardly biased test statistics (e.g., \( z \)) of the parent sample’s loading parameter estimates. This phenomenon is illustrated in a brief example presented in the following.

**EXAMPLE**

Consider for demonstration purposes a one-factor population model, this time with four indicator variables to illustrate an overidentified model scenario. Let the factor variance and the four error variances equal 1, and the loadings for \( V1 \), \( V2 \), \( V3 \), and \( V4 \) equal 1, 1, -1, and -1, respectively. The corresponding model-implied population covariance matrix \( \Sigma \) and correlation matrix \( P \) are as follows:

\[
\begin{bmatrix}
2 & 1 & -1 & -1 \\
1 & 2 & -1 & -1 \\
-1 & -1 & 2 & 1 \\
-1 & -1 & 1 & 2
\end{bmatrix}
\text{ and } \begin{bmatrix}
1 & & & \\
.5 & 1 & & \\
-1 & .5 & 1 & \\
-1 & .5 & .5 & 1
\end{bmatrix}
\]
Simulated data were generated using GAUSS (Aptech Systems, 1996). To emulate a bootstrapping process in applied modeling, a random sample of size $N = 200$ was drawn from a multivariate normal population with covariance matrix $\Sigma$ as given previously. The resulting $200 \times 4$ data matrix was used as a parent sample from which to draw bootstrap samples. (Although bootstrapping is commonly applied to nonnormal data, normal data were generated in this example for simplicity and in no way detracts from the example's illustrative value.) Bootstrap samples of size $N = 200$ were randomly drawn with replacement from the parent sample and modeled using EQS (Bentler, 1996). One thousand bootstrap samples were drawn, with each sample modeled in EQS in two separate model fittings. In one fitting the model was identified by fixing the V1 loading to 1, and in the other fitting the model was identified by fixing the factor variance to unity. For each bootstrap sample the loading for V2 was treated as the parameter of interest and was thereby monitored under both modeling strategies.

Figure 2 presents histograms of the distributions of bootstrap parameter estimates for the V2 loading under the two identification approaches. The top panel shows the distribution of estimates when the loading for V1 was fixed to 1. Notice the parameter estimates approximate a unimodal distribution, in this case with mean of 0.96 and standard deviation of 0.10. The distribution’s mean differs from 1 because of sampling variability both in bootstrap samples from the parent sample and in the parent sample from the population, whereas the standard deviation represents the desired bootstrap standard error for the V2 loading. The lower panel in Figure 2 shows the distribution of parameter estimates when the factor variance was fixed to unity. Under this scenario the distribution of estimates is bimodal, with estimated loadings randomly fluctuating near -1 and near 1; this yields a bootstrap standard error of 0.87 (almost nine times larger than the previous .10) and an overall distribution mean of 0.57. In this example one might expect more equal frequencies of parameter estimates near the high and low modes; however, the asymmetry present in this distribution is simply a function of sampling error locked into the parent sample relative to the population.

This small simulation has demonstrated the phenomenon that was algebraically illustrated earlier in this article. The bootstrap standard error for the model in which the factor variance was fixed was greatly inflated as compared to the bootstrap standard error for the model in which the factor loading was fixed. To be fair, this is an extreme example, chosen to illustrate potentially dramatic problems when incorrectly identifying an exogenous factor. In practice, incorrect identification may not be as troublesome if, for example, the majority of loadings are of the same sign and if the modeling software chooses start values of that sign. In such a case, little difference may exist between bootstrap standard errors under both identification configurations. Notwithstanding, improperly bootstrapped standard errors may greatly reduce statistical power to detect true relations within the model and may be avoided quite simply as discussed and illustrated in this article.
FIGURE 2 Histograms of bootstrap parameter estimates.

ACKNOWLEDGMENT

We wish to thank Werner Wothke for helpful comments during the preparation of this article.

REFERENCES

