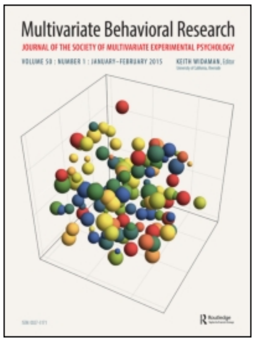

Multivariate Behavioral Research

ISSN: 0027-3171 (Print) 1532-7906 (Online) Journal homepage: https://www.tandfonline.com/loi/hmbr20

Covariance Structure Model Fit Testing Under Missing Data: An Application of the Supplemented EM Algorithm

Li Cai & Taehun Lee

To cite this article: Li Cai & Taehun Lee (2009) Covariance Structure Model Fit Testing Under Missing Data: An Application of the Supplemented EM Algorithm, Multivariate Behavioral Research, 44:2, 281-304, DOI: 10.1080/00273170902794255

To link to this article: https://doi.org/10.1080/00273170902794255

Multivariate Behavioral Research, 44:281–304, 2009

ISSN: 0027-3171 print/1532-7906 online
DOI: 10.1080/00273170902794255

Psychology Press
Taylor & Francis Group

Covariance Structure Model Fit Testing Under Missing Data: An Application of the Supplemented EM Algorithm

Li Cai
Graduate School of Education and Information Studies
University of California, Los Angeles

Taehun Lee
L. L. Thurstone Psychometric Lab
University of North Carolina

We apply the Supplemented EM algorithm (Meng & Rubin, 1991) to address a chronic problem with the "two-stage" fitting of covariance structure models in the presence of ignorable missing data: the lack of an asymptotically chi-square distributed goodness-of-fit statistic. We show that the Supplemented EM algorithm provides a convenient computational procedure that leads to such a chi-square statistic, and we provide a SAS macro implementing this method. Our derivations are corroborated with results from a small simulation study. We also apply the proposed method to 2 empirical data sets: (a) confirmatory factor analysis of Mardia, Kent, & Bibby's 1979 Open-book Closed-book data and (b) conditional latent curve modeling of adolescent aggressive behavior as discussed by Curran (1997).

Fitting covariance structure models (CSM) when some data are missing can be considerably more complicated than when all cases are complete. Under the working assumption that the data are missing at random (MAR) and that the missingness mechanism is ignorable, if we assume multivariate normality of the manifest variables, three methods of estimation are readily available for structural

Correspondence concerning this article should be addressed to Li Cai, Graduate School of Education and Information Studies, University of California, Los Angeles, 2331 Moore Hall, Box 951521, Los Angeles, CA 90095-1521. E-mail: lcai@ucla.edu.

models that can be identified from means and covariances.[1] The first method is known as full-information (or direct) maximum likelihood (FIML; Anderson, 1957; Arbuckle, 1996), in which one maximizes a sum of log-likelihoods for individual cases instead of the log-likelihood based on complete data sufficient statistics. The second method uses multiple imputation (MI; Schafer, 1997) to "fill in" missing data drawn from the posterior predictive distribution of the missing data given the observed data. Generally, $M > 1$ complete data sets are produced and each subsequently analyzed with complete data tools. The M sets of results are then appropriately combined to arrive at a single set of estimates. The third method is the so-called EM "two-stage" procedure (hereafter EM2S). In the first stage one uses the EM algorithm (Dempster, Laird, & Rubin, 1977) to find the maximum likelihood estimate (MLE) of the mean vector and the covariance matrix of the manifest variables based on incomplete data. In the second stage, a structural model is fitted to the EM estimated means and covariances using conventional CSM software. In the research reported here, we focus on EM2S.

The EM2S estimator is consistent and asymptotically normal (Yuan & Bentler, 2000). Though there may be efficiency difference between EM2S and FIML, this is usually not a cause for concern unless the fraction of missing data is extremely large. Empirically, the point estimates produced by EM2S and FIML are often virtually identical (see, e.g., Allison, 2003).

Numerous authors have pointed out desirable features of the two-stage procedure in applied research. First, software packages are readily available for both the EM estimation of means and covariances and estimation of structural parameters based on complete data sufficient statistics. Second, just as what Enders and Peugh (2004) pointed out, EM2S makes it easy to include so called *auxiliary variables* that are not part of the structural model but nonetheless are related to the pattern of missingness into the analysis. The usefulness of this inclusive strategy has been confirmed in a number of studies (e.g., Collins, Schafer, & Kam, 2001) as the MAR assumption seems better satisfied if auxiliary variables are present. Though Graham (2003) showed that one can also make special accommodations for auxiliary variables in FIML estimation, we note that with EM2S, special accommodations are not necessary. One simply estimates a larger covariance matrix with more variables. Third, unlike MI, the EM2S estimator is purely deterministic, that is, no stochastic imputations are generated. This is important because the patterns of missingness in real data sets are generally not monotone,[2] which necessitates the use of Markov chain Monte

[1]We deliberately avoid considering such extended CSM models as nonlinear or mixture models because those models can only be identified from raw data.

[2]Loosely speaking, monotone missing means that the columns of the data set can be arranged such that if a case is missing in one of the columns, then all subsequent columns are missing. The reader should refer to Schafer (1997) for the precise definition of monotone missing data.

Carlo based methods to generate the imputations. Yet the chains converge *weakly* (and asymptotically in time), so there is no finite-time guarantee that the imputations are indeed drawn from the posterior predictive distribution of the missing data. Standard convergence diagnostics such as time-series plots or autocorrelation plots can become cumbersome if the number of variables is large. In sum, proper use of multiple imputation algorithms based on Markov chains requires much greater care and experience than a deterministic method such as EM2S. Finally, we would like to add one more reason for favoring the two-stage procedure—the availability of a single mean vector and covariance matrix for reporting in journal publications. In contrast, because the FIML estimator is based on the raw data, it is difficult to provide enough statistical information in an article so that the reader can replicate the findings without obtaining the raw data set.

Despite the advantages, the lack of an asymptotically chi-square distributed goodness-of-fit (GOF) statistic is a critical flaw of the EM2S estimator. The naive practice of multiplying the minimum fit function value by the number of cases (N) minus one does not lead to a chi-square statistic. In fact, Yuan and Bentler (2000) showed that this naive statistic is distributed as a mixture of one degree-of-freedom chi-square variates. At an intuitive level, the covariance matrix analyzed in the second stage contains "less information" due to the missingness, and a multiplier based on the number of cases tends to inflate the test statistic. This heuristic observation directly motivated Enders and Peugh (2004) to consider alternative choices of the multiplier instead of $N - 1$. Unfortunately, their adjustments are based on ad hoc simulation results without enough analytical justification.

The most thorough treatment of this problem to date is given by Yuan and Bentler (2000). Indeed, the statistic that we propose in this article can be derived along the same line as Yuan & Bentler's T_4 statistic. Yet it is our original contribution to point out the connection between the Supplemented EM algorithm with the GOF problem in EM2S estimation of CSM. Furthermore, despite the generality of Yuan & Bentler's estimating equation approach (as they must because their focus was on nonnormality), we take a rather minimalist, pure likelihood approach that we believe is more accessible. We argue that after using the EM algorithm to estimate the means and covariances, the Supplemented EM algorithm requires only a trivial amount of additional computation, for which we supply a SAS macro that can be obtained free of charge. Tangentially, it is also our intention to introduce the Supplemented EM algorithm in full generality to investigators who use the EM algorithm in contexts other than missing data because the Supplemented EM is applicable whenever the EM is. Other potential applications will become possible once this link is pointed out (see, e.g., Cai, 2008).

Although the original EM algorithm is well known, the Supplemented EM algorithm seems to be a relatively obscure subject among researchers in the behavioral sciences. Therefore, a review of the key technical aspects of the Supplemented EM algorithm is provided in the first section. We also walk the reader through the details of the Supplemented EM with an artificial example. Next, we address the problem of GOF testing for EM2S. It is shown that the key to this problem lies in the computation of an asymptotic covariance matrix that can be readily obtained using the Supplemented EM algorithm. Results from a small simulation are reported to verify the analytical results, and two example data sets are analyzed to demonstrate the use of an SAS macro implementing the proposed method. The article concludes with a discussion about extensions of the proposed procedure and other uses of the Supplemented EM algorithm.

THE SUPPLEMENTED EM ALGORITHM

In order to discuss the Supplemented EM algorithm, we must first review the EM algorithm in the classical incomplete data context. The introductory material is intended to provide a context general enough to accommodate our specific application.

The EM Algorithm

The essence of the EM algorithm is to transform the intractable incomplete data estimation problem into iteratively solving a sequence of simple complete data problems. Suppose the observed data $\mathbf{Y}$ can be augmented by the missing data $\mathbf{X}$ to permit the representation of the complete data as $\mathbf{Z} = (\mathbf{Y}, \mathbf{X})$. Suppose further that the task is to find the MLE $\hat{\boldsymbol{\sigma}}$ of a d-dimensional parameter vector $\boldsymbol{\sigma}$ based on a parametric model for the observed data, whose density is $f(\mathbf{Y}|\boldsymbol{\sigma})$.

Let $\pi(\mathbf{X}|\mathbf{Y}, \boldsymbol{\sigma})$ be the conditional density of the missing data given the observed data, so that the complete data density is given by

$$f(\mathbf{Z}|\boldsymbol{\sigma}) = f(\mathbf{Y}|\boldsymbol{\sigma})\pi(\mathbf{X}|\mathbf{Y}, \boldsymbol{\sigma}).$$

It then follows that the observed data log-likelihood can be written as

$$l(\boldsymbol{\sigma}|\mathbf{Y}) = l(\boldsymbol{\sigma}|\mathbf{Z}) - \log \pi(\mathbf{X}|\mathbf{Y}, \boldsymbol{\sigma}), \tag{1}$$

where $l(\boldsymbol{\sigma}|\mathbf{Y})$ is the observed data log-likelihood, and $l(\boldsymbol{\sigma}|\mathbf{Z})$ the complete data log-likelihood.

Dempster et al. (1977) noticed that had the missing data been observed, the complete data log-likelihood would often become easy to maximize. Suppose a provisional estimate of $\boldsymbol{\sigma}$ is $\boldsymbol{\sigma}^*$, one iteration of the EM algorithm consists of

(a) the E(xpectation) step, in which the expected complete data log-likelihood is computed as

$$Q(\boldsymbol{\sigma}|\boldsymbol{\sigma}^*) = \int l(\boldsymbol{\sigma}|\mathbf{Z})\pi(\mathbf{X}|\mathbf{Y}, \boldsymbol{\sigma}^*)d\mathbf{X}, \tag{2}$$

and (b) the M(aximization) step, in which $Q(\boldsymbol{\sigma}|\boldsymbol{\sigma}^*)$ is maximized to yield an updated estimate of $\boldsymbol{\sigma}$. The two steps are iterated until convergence.

Once the MLE is obtained, a natural next step is to assess its variability. It is well known that the negative of the second derivative matrix of $l(\boldsymbol{\sigma}|\mathbf{Y})$ evaluated at the MLE is the (observed) information matrix, whose inverse is the large-sample covariance matrix of the parameter estimates. However, having incomplete data makes the analytical evaluation of $l(\boldsymbol{\sigma}|\mathbf{Y})$ and its derivatives "difficult or at least tedious" (McLachlan & Krishnan, 1996, p. 111). The EM algorithm circumvents the difficulty by not computing the derivatives, but this implies that the EM does not provide a covariance matrix at convergence. Since its inception, much criticism has been leveled at the EM algorithm regarding this apparent deficiency. A number of authors have proposed methods for computing the information matrix (e.g., Louis, 1982; Meilijson, 1989). The most easily accessible procedure for such a purpose is the Supplemented EM algorithm (Meng & Rubin, 1991), which is introduced below in three parts. Schafer (1997) also contains a description of the procedure.

Part One: The Missing Information Principle

First, let

$$\mathscr{I}(\boldsymbol{\sigma}|\mathbf{Y}) = -\frac{\partial^2 l(\boldsymbol{\sigma}|\mathbf{Y})}{\partial\boldsymbol{\sigma}\,\partial\boldsymbol{\sigma}^T}, \qquad \mathscr{I}(\boldsymbol{\sigma}|\mathbf{Z}) = -\frac{\partial^2 l(\boldsymbol{\sigma}|\mathbf{Z})}{\partial\boldsymbol{\sigma}\,\partial\boldsymbol{\sigma}^T} \tag{3}$$

be the observed data and the complete data information matrix, respectively. Next, let

$$\mathscr{I}_c(\hat{\boldsymbol{\sigma}}|\mathbf{Y}) = \int \mathscr{I}(\hat{\boldsymbol{\sigma}}|\mathbf{Z})\pi(\mathbf{X}|\mathbf{Y}, \hat{\boldsymbol{\sigma}})d\mathbf{X} \tag{4}$$

be the conditional expectation of the complete data information matrix, and let

$$\mathscr{I}_m(\hat{\boldsymbol{\sigma}}|\mathbf{Y}) = \int -\frac{\partial^2 \log \pi(\mathbf{X}|\mathbf{Y}, \hat{\boldsymbol{\sigma}})}{\partial\boldsymbol{\sigma}\,\partial\boldsymbol{\sigma}^T}\pi(\mathbf{X}|\mathbf{Y}, \hat{\boldsymbol{\sigma}})d\mathbf{X} \tag{5}$$

be the missing information matrix, both evaluated at the MLE. Then after taking second derivatives on both sides of Equation (1), and integrating over

$\pi(\mathbf{X}|\mathbf{Y}, \hat{\boldsymbol{\sigma}})$, one arrives at the *missing information principle* of Orchard and Woodbury (1972):

$$\begin{aligned} \mathscr{I}(\hat{\boldsymbol{\sigma}}|\mathbf{Y}) &= \mathscr{I}_c(\hat{\boldsymbol{\sigma}}|\mathbf{Y}) - \mathscr{I}_m(\hat{\boldsymbol{\sigma}}|\mathbf{Y}) \\ &= \left\{\mathbf{I}_d - \mathscr{I}_m(\hat{\boldsymbol{\sigma}}|\mathbf{Y})\mathscr{I}_c^{-1}(\hat{\boldsymbol{\sigma}}|\mathbf{Y})\right\} \mathscr{I}_c(\hat{\boldsymbol{\sigma}}|\mathbf{Y}), \end{aligned} \tag{6}$$

where $\mathscr{I}_m(\hat{\boldsymbol{\sigma}}|\mathbf{Y})\mathscr{I}_c^{-1}(\hat{\boldsymbol{\sigma}}|\mathbf{Y})$ is a matrix representation of the *fraction of missing information*, and $\mathbf{I}_d$ is a $d \times d$ identity matrix. In other words, the first line of Equation (6) says that the observed data information is exactly equal to the complete data information minus the missing information, whereas the second line says that the observed data information matrix can be obtained by adjusting the complete data information matrix according to the (matrix) fraction of missing information.

Part Two: The EM Map

Suppose the parameter space of $\boldsymbol{\sigma}$ is $\mathscr{S}$, a subset of $\mathbb{R}^d$. The EM algorithm defines a vector-valued mapping, $\boldsymbol{\sigma} \to M(\boldsymbol{\sigma})$, from $\mathscr{S}$ to itself. Let the estimate of $\boldsymbol{\sigma}$ in the kth iteration be $\boldsymbol{\sigma}^{(k)}$. Then the EM map can be written as

$$\boldsymbol{\sigma}^{(k+1)} = M(\boldsymbol{\sigma}^{(k)}). \tag{7}$$

It is clear that $\hat{\boldsymbol{\sigma}} = M(\hat{\boldsymbol{\sigma}})$, so the MLE is referred to as a fixed point of the EM map. Suppose further that the mapping is continuous, then in the vicinity of $\hat{\boldsymbol{\sigma}}$, a Taylor series expansion of $M(\boldsymbol{\sigma})$ yields, to a first approximation,

$$\boldsymbol{\sigma}^{(k+1)} = M(\boldsymbol{\sigma}^{(k)}) \approx M(\hat{\boldsymbol{\sigma}}) + \boldsymbol{\Delta}(\hat{\boldsymbol{\sigma}})(\boldsymbol{\sigma}^{(k)} - \hat{\boldsymbol{\sigma}}) = \hat{\boldsymbol{\sigma}} + \boldsymbol{\Delta}(\hat{\boldsymbol{\sigma}})(\boldsymbol{\sigma}^{(k)} - \hat{\boldsymbol{\sigma}}), \tag{8}$$

where

$$\boldsymbol{\Delta}(\hat{\boldsymbol{\sigma}}) = \frac{\partial M(\hat{\boldsymbol{\sigma}})}{\partial \boldsymbol{\sigma}} \tag{9}$$

is the $(d \times d)$ Jacobian matrix of $M(\cdot)$ evaluated at the MLE. If we let $M_j(\boldsymbol{\sigma})$ be the jth element of $M(\boldsymbol{\sigma})$, and σ_i be the ith element of $\boldsymbol{\sigma}$, then the (i, j)th element of $\boldsymbol{\Delta}(\hat{\boldsymbol{\sigma}})$ is

$$\hat{\delta}_{ij} = \frac{\partial M_j(\hat{\boldsymbol{\sigma}})}{\partial \sigma_i}.$$

Rearrangement of Equation (8) leads to

$$\boldsymbol{\sigma}^{(k+1)} - \hat{\boldsymbol{\sigma}} \approx \boldsymbol{\Delta}(\hat{\boldsymbol{\sigma}})(\boldsymbol{\sigma}^{(k)} - \hat{\boldsymbol{\sigma}}),$$

which shows that the EM algorithm is linearly convergent and the rate of convergence is precisely $\boldsymbol{\Delta}(\hat{\boldsymbol{\sigma}})$.

An important result in Dempster et al. (1977) connects the rate of convergence to the fraction of missing information. It can be shown that

$$\boldsymbol{\Delta}(\hat{\boldsymbol{\sigma}}) = \mathscr{I}_m(\hat{\boldsymbol{\sigma}}|\mathbf{Y})\mathscr{I}_c^{-1}(\hat{\boldsymbol{\sigma}}|\mathbf{Y}) \tag{10}$$

holds in the vicinity of the MLE as long as the expected complete-data log-likelihood is maximized in each M-step. On substituting (10)) into (6) and inverting, the large-sample covariance matrix of the MLE can be written as

$$V(\hat{\boldsymbol{\sigma}}|\mathbf{Y}) = \mathscr{I}^{-1}(\hat{\boldsymbol{\sigma}}|\mathbf{Y}) = \mathscr{I}_c^{-1}(\hat{\boldsymbol{\sigma}}|\mathbf{Y})\{\mathbf{I}_d - \boldsymbol{\Delta}(\hat{\boldsymbol{\sigma}})\}^{-1}. \tag{11}$$

Part Three: The Supplemented EM Algorithm

Equation (11) forms the basis of the Supplemented EM algorithm. By construction, $\mathscr{I}_c(\hat{\boldsymbol{\sigma}}|\mathbf{Y})$ depends only on the the complete data sufficient statistics so it should be easy to compute. The rate matrix $\boldsymbol{\Delta}(\hat{\boldsymbol{\sigma}})$ can be computed using the code and iteration history of the original EM, as shown later.

Assuming the original EM converged in K iterations, and the iteration history has been saved as $\boldsymbol{\sigma}^{(0)}, \ldots, \boldsymbol{\sigma}^{(K)}$. Let

$$\boldsymbol{\sigma}_{(i)}^{(k)} = (\hat{\sigma}_1, \ldots, \hat{\sigma}_{i-1}, \sigma_i^{(k)}, \hat{\sigma}_{i+1}, \ldots, \hat{\sigma}_d), \tag{12}$$

where $\sigma_i^{(k)}$ is the ith element of $\boldsymbol{\sigma}^{(k)}$. In other words, $\boldsymbol{\sigma}_{(i)}^{(k)}$ is equal to $\hat{\boldsymbol{\sigma}}$ except that its ith element is replaced by its value at the kth iteration. From the basic definition of the derivative, we have

$$\begin{aligned}
\hat{\delta}_{ij} &= \frac{\partial M_j(\hat{\boldsymbol{\sigma}})}{\partial \sigma_i} \\
&= \lim_{\sigma_i \to \hat{\sigma}_i} \frac{M_j(\hat{\sigma}_1, \ldots, \hat{\sigma}_{i-1}, \sigma_i, \hat{\sigma}_{i+1}, \ldots, \hat{\sigma}_d) - M_j(\hat{\boldsymbol{\sigma}})}{\sigma_i - \hat{\sigma}_i} \\
&= \lim_{k \to \infty} \frac{M_j(\boldsymbol{\sigma}_{(i)}^{(k)}) - \hat{\sigma}_j}{\sigma_i^{(k)} - \hat{\sigma}_i} = \lim_{k \to \infty} \delta_{ij}^{(k)},
\end{aligned}$$

where the forward difference

$$\delta_{ij}^{(k)} = \frac{M_j(\boldsymbol{\sigma}_{(i)}^{(k)}) - \hat{\sigma}_j}{\sigma_i^{(k)} - \hat{\sigma}_i} \tag{13}$$

is the (i, j)th element of a $d \times d$ matrix $\boldsymbol{\Delta}^{(k)}$ such that $\lim_{k\to\infty} \boldsymbol{\Delta}^{(k)} = \boldsymbol{\Delta}(\hat{\boldsymbol{\sigma}})$, element by element.

Equation (13) suggests that the sequence of $\boldsymbol{\Delta}^{(k)}$'s can be obtained by the following supplemental cycles after the convergence of the original EM. Because these additional cycles are run after the main EM, Meng and Rubin (1991) called the procedure Supplemented EM.

Set $i = 1$, and retrieve the output of the original EM at the kth iteration, that is, $\boldsymbol{\sigma}^{(k)}$:

1. Compute $\boldsymbol{\sigma}_{(i)}^{(k)}$ as in Equation (12) and set it as a new starting value;
2. Run *one* iteration of the original EM started from $\boldsymbol{\sigma}_{(i)}^{(k)}$ to obtain $M(\boldsymbol{\sigma}_{(i)}^{(k)})$;
3. Compute $\delta_{ij}^{(k)}$ for $j = 1, \ldots, d$ and this completes the ith row of $\boldsymbol{\Delta}^{(k)}$;
4. Increase i by 1 and go back to step 1 or stop if $i > d$.

This completes the kth cycle of the Supplemented EM, and the result is $\boldsymbol{\Delta}^{(k)}$. Next, we move on to the output of the original EM at the $(k + 1)$th iteration, that is, $\boldsymbol{\sigma}^{(k+1)}$, and restart the process to compute $\boldsymbol{\Delta}^{(k+1)}$, and so on and so forth until the $\boldsymbol{\Delta}$'s stabilize. Note that with a matrix-oriented programming language such as GAUSS or SAS/IML, the ith row of $\boldsymbol{\Delta}^{(k)}$ can be computed as an element-by-element division between $[M(\boldsymbol{\sigma}_{(i)}^{(k)}) - \hat{\boldsymbol{\sigma}}]^T$ and $(\sigma_i^{(k)} - \hat{\sigma}_i)$.

Our strategy for monitoring the convergence of the Supplemented EM algorithm is to start it at $\boldsymbol{\sigma}^{(1)}$ and run it for $K^* \leq K$ cycles, where K^* is determined if the maximum element-wise difference between $\boldsymbol{\Delta}^{(K^*-1)}$ and $\boldsymbol{\Delta}^{(K^*)}$ is sufficiently small. Meng and Rubin (1991) noted that the Supplemented EM algorithm amounts to numerically differentiating the EM Map, so the stopping criterion should be less stringent than the original EM. They proposed that one use the square root of the original EM's stopping criterion. In our experience, that suggestion has performed well.

An Illustrative Example of Supplemented EM at Work

We use an artificial bivariate data set with partial missing observations to better explicate the details of the Supplemented EM. The data can be found in Table 1. The two variables are denoted as Z_1 and Z_2. The task is to compute the MLE

TABLE 1
Artificial Bivariate Data Set

Z_1	4.5	−2.2	9.7	28.0	—	−19.8	6.1	—	−3.3	−17.8
Z_2	—	2.7	5.5	13.5	−7.2	−10.0	−6.2	−0.8	−13.2	—

of the mean vector, along with its standard error. To do this, we first use the EM algorithm for multivariate normal missing data problems to compute the estimates of the means $(\overline{Z}_1, \overline{Z}_2)^T$, the variances S_{11} and S_{22}, and the covariance S_{21}, based on the observed incomplete data. Let the MLEs be ordered as $\hat{\boldsymbol{\sigma}} = (\overline{Z}_1, \overline{Z}_2, S_{11}, S_{21}, S_{22})^T$. Details on implementing this particular EM algorithm are discussed in the next section, but for now let us assume that such an algorithm is available.

Table 2 shows a sequence of estimates from the iteration history of the EM algorithm, up to Cycle 15. One can see that the means are started from zeros, and the starting covariance matrix is an identity matrix. The vector of MLEs is found to be $\hat{\boldsymbol{\sigma}} = (0.266, -2.778, 184.552, 86.993, 69.243)^T$. At this point, one would reason that the standard errors of the means could simply be obtained by dividing the estimated variances by sample size (10 in this case) and taking square root. However, as it is empirically demonstrated, this is incorrect and will lead to standard errors that are too small. The problem with this line of reasoning is that it ignores the fact that the amount of information the data set actually contains is less than what the sample size would imply due to the missing observations. The standard errors need to be adjusted, and we use the Supplemented EM for such a purpose.

We start the Supplemented EM from the iteration history, beginning with estimates from Cycle 1. First, we take the MLEs $\hat{\boldsymbol{\sigma}}$ and replace the first element by its value from Cycle 1, resulting in $\boldsymbol{\sigma}_{(1)}^{(1)} = (0.520, -2.778, 184.552, 86.993, 69.243)^T$. Next, we run one cycle of EM, with $\boldsymbol{\sigma}_{(1)}^{(1)}$ as its starting value. The result is $M(\boldsymbol{\sigma}_{(1)}^{(1)}) = (0.317, -2.801, 184.406, 87.098, 69.401)^T$. Using Equation (13), we can compute the $(1, 1)$th element of $\boldsymbol{\Delta}^{(1)}$ as

$$\delta_{11}^{(1)} = \frac{M_1(\boldsymbol{\sigma}_{(1)}^{(1)}) - \hat{\sigma}_1}{\sigma_1^{(1)} - \hat{\sigma}_1} = \frac{0.317 - 0.266}{0.520 - 0.266} = .200.$$

TABLE 2
A Possible Sequence of EM Iteration History

		Cycle						
	Start	*1*	*2*	*3*	$\cdots$	*13*	*14*	*15*
$\overline{Z}_1$	0	0.520	0.073	0.145		0.266	0.266	0.266
$\overline{Z}_2$	0	−1.570	−2.435	−2.646		−2.777	−2.777	−2.778
S_{11}	1	165.946	188.382	187.018	$\cdots$	184.553	184.552	184.552
S_{21}	0	63.731	82.501	85.389		86.993	86.993	86.993
S_{22}	1	56.230	66.324	68.089		69.242	69.243	69.243

By the same token, the $(1,2)$th element is

$$\delta_{12}^{(1)} = \frac{M_2(\boldsymbol{\sigma}_{(1)}^{(1)}) - \hat{\sigma}_2}{\sigma_1^{(1)} - \hat{\sigma}_1} = \frac{-2.801 + 2.778}{0.520 - 0.266} = -0.094.$$

We continue until all five elements in the first row of $\boldsymbol{\Delta}^{(1)}$ are computed.

For the second row of $\boldsymbol{\Delta}^{(1)}$, we follow the aforementioned routine. We go back to the MLE and now replace its *second* element by its value from Cycle 1, resulting in a new starting value $\boldsymbol{\sigma}_{(2)}^{(1)} = (0.266, -1.570, 184.552, 86.993, 69.243)^T$. Again, we run one cycle of the original EM to get $M(\boldsymbol{\sigma}_{(2)}^{(1)}) = (-0.037, -2.536, 185.852, 85.767, 67.902)^T$. The $(2,1)$th element of $\boldsymbol{\Delta}^{(1)}$ is therefore

$$\delta_{21}^{(1)} = \frac{M_1(\boldsymbol{\sigma}_{(2)}^{(1)}) - \hat{\sigma}_1}{\sigma_2^{(1)} - \hat{\sigma}_2} = \frac{-0.037 - 0.266}{-1.570 + 2.778} = -0.251,$$

and other elements in the second row of $\boldsymbol{\Delta}^{(1)}$ are similarly computed. When the entire $\boldsymbol{\Delta}^{(1)}$ is assembled, we have

$$\boldsymbol{\Delta}^{(1)} = \begin{pmatrix} 0.200 & -0.094 & -0.574 & 0.412 & 0.624 \\ -0.251 & 0.200 & 1.077 & -1.015 & -1.111 \\ 0.000 & 0.004 & 0.200 & -0.098 & -0.048 \\ -0.004 & -0.007 & -0.361 & 0.221 & -0.010 \\ 0.005 & 0.000 & 0.242 & -0.052 & 0.200 \end{pmatrix}.$$

This completes the first cycle of the Supplemented EM, and we move on to the next cycle in the EM iteration history to compute $\boldsymbol{\Delta}^{(2)}$, and so on and so forth.

The sequence of $\boldsymbol{\Delta}^{(k)}$'s eventually converges element by element to the Jacobian matrix of the EM Map $M(\cdot)$. For our data set, the converged $\boldsymbol{\Delta}$ is

$$\boldsymbol{\Delta}(\hat{\boldsymbol{\sigma}}) = \begin{pmatrix} 0.200 & -0.094 & -0.614 & 0.408 & 0.615 \\ -0.251 & 0.200 & 0.772 & -1.076 & -1.304 \\ 0.000 & 0.004 & 0.200 & -0.088 & -0.038 \\ -0.004 & -0.007 & -0.417 & 0.220 & -0.013 \\ 0.004 & 0.000 & 0.209 & -0.043 & 0.200 \end{pmatrix}.$$

Given the MLEs of the means and covariance matrix, standard normal theory results (e.g., Rao, 1973) can be used to establish the complete data covariance matrix of $\hat{\boldsymbol{\sigma}}$, which is equal to the inverse of the conditional expectation of the complete data information matrix

$$\mathscr{I}_c^{-1}(\hat{\boldsymbol{\sigma}}\,|\,\mathbf{Y}) = \begin{pmatrix} 18.455 & 8.699 & 0.000 & 0.000 & 0.000 \\ 8.699 & 6.924 & 0.000 & 0.000 & 0.000 \\ 0.000 & 0.000 & 6811.882 & 3210.954 & 1513.565 \\ 0.000 & 0.000 & 3210.954 & 2034.674 & 1204.733 \\ 0.000 & 0.000 & 1513.565 & 1204.733 & 958.916 \end{pmatrix}.$$

Plugging $\boldsymbol{\Delta}(\hat{\boldsymbol{\sigma}})$ into Equation (11), the observed data covariance matrix of $\hat{\boldsymbol{\sigma}}$ is

$$V(\hat{\boldsymbol{\sigma}}\,|\,\mathbf{Y}) = \begin{pmatrix} 20.432 & 8.439 & -6.785 & -0.327 & 2.274 \\ 8.439 & 7.731 & 2.619 & -6.219 & -6.144 \\ -6.784 & 2.619 & 7232.400 & 3214.923 & 1483.815 \\ -0.327 & -6.219 & 3214.935 & 2183.355 & 1326.649 \\ 2.274 & -6.144 & 1483.817 & 1326.642 & 1118.056 \end{pmatrix}.$$

The diagonal elements of $V(\hat{\boldsymbol{\sigma}}\,|\,\mathbf{Y})$ are, without exception, larger than the corresponding elements in $\mathscr{I}_c^{-1}(\hat{\boldsymbol{\sigma}}\,|\,\mathbf{Y})$. The adjustment effect of $\boldsymbol{\Delta}(\hat{\boldsymbol{\sigma}})$ is evident.

Finally, we are ready to compute the standard errors of the means. Table 3 shows three methods for computing these standard errors. The first one, under the heading "complete data covariance," is simply the square root of the first two diagonal elements of $\mathscr{I}_c^{-1}(\hat{\boldsymbol{\sigma}}\,|\,\mathbf{Y})$. The second one uses the Supplemented EM adjustment. Purely for the sake of comparison, we have explicitly calculated the second derivative matrices of the log-likelihoods for each individual case. Note that in general latent variable modeling problems, the attempt at calculating such a second derivative matrix would not only be difficult but also would defeat the purpose of using the EM algorithm in the first place. Nevertheless, the inverse of this Hessian times -1 is the observed data covariance matrix that the Supplemented EM is trying to approximate. The square root of its diagonal elements are shown in the last column of the table. One can see that (a) the complete data standard errors are appreciably smaller than the other two sets of standard errors, and (b) the quality of the Supplemented EM approximation is excellent in this case.

TABLE 3
Estimated Standard Errors of the Means

		SE		
	Estimate	*Complete Data Covariance*	*Supplemented EM Approx.*	*Inverse of Hessian*
$\overline{Z}_1$	0.266	4.30	4.52	4.52
$\overline{Z}_2$	−2.778	2.63	2.78	2.77

APPLICATION OF SUPPLEMENTED EM TO EM2S TESTING OF CSM

Equipped with the theory developed in the previous section, we are now ready to tackle the problem of GOF testing for the EM2S estimator. We first formalize the EM2S estimator. A chi-square GOF statistic is then derived, followed by some simulation results.

The EM2S Estimator

Consider a multivariate normal data matrix $\mathbf{Z}$ with N independent rows and p variables. If the ith row of $\mathbf{Z}$ is $\mathbf{z}_i^T$, then $\mathbf{z}_i$ follows a p dimensional multivariate normal distribution with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Omega}$. Let $\boldsymbol{\sigma} = [\boldsymbol{\mu}^T, \text{vech}(\boldsymbol{\Omega})^T]^T$, where the operator vech$(\cdot)$ stacks the unique elements of a symmetric matrix. Clearly, the dimension of $\boldsymbol{\sigma}$ is $d = p + p(p+1)/2$. Let the MLEs of the mean vector and the covariance matrix be $\hat{\boldsymbol{\mu}}$ and $\hat{\boldsymbol{\Omega}}$, respectively, then $\hat{\boldsymbol{\sigma}} = [\hat{\boldsymbol{\mu}}^T, \text{vech}(\hat{\boldsymbol{\Omega}})^T]^T$. For these moments, consider a structural model:

$$\boldsymbol{\sigma}(\boldsymbol{\theta}) = \begin{pmatrix} \boldsymbol{\mu}(\boldsymbol{\theta}) \\ \text{vech}[\boldsymbol{\Omega}(\boldsymbol{\theta})] \end{pmatrix}, \tag{14}$$

where $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ is a q-dimensional parameter vector in a subset of $\mathbb{R}^q$. If there are no missing data in $\mathbf{Z}$, we can use standard CSM software such as SAS PROC CALIS to estimate $\boldsymbol{\theta}$ from sample moments. If missing values are present in $\mathbf{Z}$, estimation of $\boldsymbol{\theta}$ can be carried out in two stages.

In the first stage, the EM algorithm is used to obtain the MLEs of $\boldsymbol{\mu}$ and $\boldsymbol{\Omega}$. With no loss of generality, suppose $\mathbf{z}_i$ can be partitioned as $\mathbf{z}_i = (\mathbf{y}_i^T, \mathbf{x}_i^T)^T$, where $\mathbf{y}_i$ and $\mathbf{x}_i$ are the observed and missing part of $\mathbf{z}_i$, respectively. Let $\boldsymbol{\mu}$ and $\boldsymbol{\Omega}$ be partitioned accordingly into missing and observed parts, that is, $\boldsymbol{\mu} = (\boldsymbol{\mu}_{i,y}^T, \boldsymbol{\mu}_{i,x}^T)^T$, and

$$\boldsymbol{\Omega} = \begin{pmatrix} \boldsymbol{\Omega}_{i,yy} & \boldsymbol{\Omega}_{i,xy}^T \\ \boldsymbol{\Omega}_{i,xy} & \boldsymbol{\Omega}_{i,xx} \end{pmatrix}.$$

Given a provisional estimate of $\boldsymbol{\sigma}$, say, $\boldsymbol{\sigma}^{(k)} = [(\boldsymbol{\mu}^{(k)})^T, \text{vech}(\boldsymbol{\Omega}^{(k)})^T]^T$, the conditional distribution of $\mathbf{x}_i$ is multivariate normal with mean vector

$$\boldsymbol{\mu}_{i,x|y}^{(k)} = \boldsymbol{\mu}_{i,x}^{(k)} + \boldsymbol{\Omega}_{i,xy}^{(k)}(\boldsymbol{\Omega}_{i,yy}^{(k)})^{-1}(\mathbf{y}_i - \boldsymbol{\mu}_{i,y}^{(k)}),$$

and covariance matrix

$$\boldsymbol{\Omega}_{i,x|y}^{(k)} = \boldsymbol{\Omega}_{i,xx}^{(k)} - \boldsymbol{\Omega}_{i,xy}^{(k)}(\boldsymbol{\Omega}_{i,yy}^{(k)})^{-1}(\boldsymbol{\Omega}_{i,xy}^{(k)})^T.$$

In the E-step, the missing values are replaced by their conditional means, that is,

$$\mathbf{z}_i^{(k)} = \begin{pmatrix} \mathbf{y}_i \\ \boldsymbol{\mu}_{i,x|y}^{(k)} \end{pmatrix},$$

and the matrix of sums-of-squares and cross products is formed as

$$\mathbf{C}^{(k)} = \sum_{i=1}^{N} \left\{ \mathbf{z}_i^{(k)} (\mathbf{z}_i^{(k)})^T + \begin{pmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Omega}_{i,x|y}^{(k)} \end{pmatrix} \right\}.$$

In the M-step, the updated mean vector is computed as $\boldsymbol{\mu}^{(k+1)} = N^{-1} \sum_{i=1}^{N} \mathbf{z}_i^{(k)}$, and the updated estimate of covariance matrix is

$$\boldsymbol{\Omega}^{(k+1)} = N^{-1}\mathbf{C}^{(k)} - \boldsymbol{\mu}^{(k+1)}(\boldsymbol{\mu}^{(k+1)})^T.$$

Suppose convergence is achieved at iteration K, the aforementioned EM algorithm produces the MLE $\hat{\boldsymbol{\sigma}}$ as well as the iteration history $\{\boldsymbol{\sigma}^{(0)}, \ldots, \boldsymbol{\sigma}^{(K)}\}$.

In the second stage, the structural model (14) is fit to the MLE $\hat{\boldsymbol{\sigma}} = [\hat{\boldsymbol{\mu}}^T, \text{vech}(\hat{\boldsymbol{\Omega}})^T]^T$ by minimizing the following discrepancy function:

$$\begin{aligned} F(\boldsymbol{\theta}; \hat{\boldsymbol{\sigma}}) &= \text{tr}\{\hat{\boldsymbol{\Omega}}[\boldsymbol{\Omega}(\boldsymbol{\theta})]^{-1}\} - \log|\hat{\boldsymbol{\Omega}}[\boldsymbol{\Omega}(\boldsymbol{\theta})]^{-1}| - p \\ &\quad + [\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}(\boldsymbol{\theta})]^T[\boldsymbol{\Omega}(\boldsymbol{\theta})]^{-1}[\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}(\boldsymbol{\theta})]. \end{aligned} \tag{15}$$

We denote the EM2S estimator that minimizes $F(\boldsymbol{\theta}; \hat{\boldsymbol{\sigma}})$ over $\boldsymbol{\Theta}$ as $\hat{\boldsymbol{\theta}}$, and at the minimum, the following GOF statistic is printed by all conventional CSM software:

$$T_F = (N-1)F(\hat{\boldsymbol{\theta}}; \hat{\boldsymbol{\sigma}}). \tag{16}$$

GOF Testing

The naive GOF statistic T_F is not asymptotically chi-square distributed. It is distributed as a mixture of one degree-of-freedom chi-square variates (Yuan & Bentler, 2000). It turns out that the key to finding a chi-square GOF statistic lies in the asymptotic covariance matrix of the means and covariances. Once this asymptotic covariance matrix is obtained, Browne's (1984) residual-based statistic may be used, which is asymptotically chi-square distributed for a consistent and asymptotically normal estimator of $\boldsymbol{\theta}$, such as EM2S. Let

$$\mathbf{J}(\boldsymbol{\theta}) = \frac{\partial \boldsymbol{\sigma}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}$$

be the $d \times q$ Jacobian of the structural model. Assuming Browne's (1984) regularity conditions, $\mathbf{J}(\boldsymbol{\theta})$ is of full column rank, so there exists a $d \times (d-q)$ matrix $\mathbf{J}^c(\boldsymbol{\theta})$ that is an orthogonal complement of $\mathbf{J}(\boldsymbol{\theta})$, such that $[\mathbf{J}^c(\boldsymbol{\theta})]^T\mathbf{J}(\boldsymbol{\theta}) = \mathbf{0}$.

Let $\boldsymbol{\sigma}(\hat{\boldsymbol{\theta}})$ be the model-implied moments. The residual moments are simply $\mathbf{e} = \hat{\boldsymbol{\sigma}} - \boldsymbol{\sigma}(\hat{\boldsymbol{\theta}})$. Furthermore, let $\boldsymbol{\Xi} = V(\hat{\boldsymbol{\sigma}}|\mathbf{Y})$ be the $d \times d$ asymptotic covariance matrix of $\sqrt{N}\hat{\boldsymbol{\sigma}}$. Then for $H_0 : \boldsymbol{\sigma} - \boldsymbol{\sigma}(\boldsymbol{\theta}) = \mathbf{0}$ for some $\boldsymbol{\theta}_0 \in \boldsymbol{\Theta}$ vs. $H_1 : \boldsymbol{\sigma} - \boldsymbol{\sigma}(\boldsymbol{\theta}) \neq \mathbf{0}$ for any $\boldsymbol{\theta}$, the following GOF statistic

$$T_B = N\mathbf{e}^T\boldsymbol{\Gamma}\mathbf{e} \tag{17}$$

is asymptotically distributed as a central chi-square variable with $d-q$ degrees of freedom under the null hypothesis, where

$$\boldsymbol{\Gamma} = [\mathbf{J}^c(\hat{\boldsymbol{\theta}})]\{[\mathbf{J}^c(\hat{\boldsymbol{\theta}})]^T\boldsymbol{\Xi}[\mathbf{J}^c(\hat{\boldsymbol{\theta}})]\}^{-1}[\mathbf{J}^c(\hat{\boldsymbol{\theta}})]^T.$$

The EM2S estimator provides consistent estimates of every matrix in Equation (17) except $\boldsymbol{\Xi}$, but it should be clear by now that the easiest method for computing $\boldsymbol{\Xi}$ is the Supplemented EM algorithm, as long as the iteration history $\{\boldsymbol{\sigma}^{(0)}, \ldots, \boldsymbol{\sigma}^{(K)}\}$ is saved. The covariance matrix $\boldsymbol{\Xi}$ is a rate adjusted complete data covariance matrix $\mathscr{I}_c^{-1}(\hat{\boldsymbol{\sigma}}|\mathbf{Y})\{\mathbf{I}_d - \boldsymbol{\Delta}(\hat{\boldsymbol{\sigma}})\}^{-1}$, as in Equation (11), where $\mathscr{I}_c(\hat{\boldsymbol{\sigma}}|\mathbf{Y})$ takes an extremely simple form in this case due to multivariate normality:

$$\mathscr{I}_c(\hat{\boldsymbol{\sigma}}|\mathbf{Y}) = \begin{pmatrix} \hat{\boldsymbol{\Omega}}^{-1} & \mathbf{0} \\ \mathbf{0} & \frac{1}{2}\mathbf{D}_p^T(\hat{\boldsymbol{\Omega}}^{-1} \otimes \hat{\boldsymbol{\Omega}}^{-1})\mathbf{D}_p \end{pmatrix},$$

where $\mathbf{D}_p$ is the duplication matrix as defined in Schott (1997).

Standard Errors for Structural Parameter Estimates

It follows from results in Browne and Arminger (1995), particularly a Lemma in Section 2.2, that the limiting covariance matrix of $\sqrt{N}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)$ is

$$\boldsymbol{\Upsilon} = \left\{\mathbf{J}(\boldsymbol{\theta}_0)^T\boldsymbol{\Xi}^{-1}\mathbf{J}(\boldsymbol{\theta}_0)\right\}^{-1}. \tag{18}$$

This can be consistently estimated by evaluating $\mathbf{J}$ at $\hat{\boldsymbol{\theta}}$. The diagonal elements of $\boldsymbol{\Upsilon}$ are the squared asymptotic standard errors for the structural parameters.

Simulation One: Type I Error

To support our claim that T_B is indeed asymptotically chi-square distributed under the null hypothesis, a small simulation was conducted. A confirmatory factor analysis model with six manifest variables and two correlated factors was used in the simulation. The covariance structure may be written as

$$\boldsymbol{\Omega}(\boldsymbol{\theta}) = \boldsymbol{\Lambda}\boldsymbol{\Phi}\boldsymbol{\Lambda}^T + \boldsymbol{\Psi}. \tag{19}$$

The generating loading matrix is

$$\boldsymbol{\Lambda}^T = \begin{pmatrix} 0.8 & 0.8 & 0.8 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0.7 & 0.7 & 0.7 \end{pmatrix},$$

and $\boldsymbol{\Phi}$ is set to a correlation matrix with $\phi_{21} = 0.2$. The unique variances are given by $\boldsymbol{\Psi} = \text{diag}(\psi_{11}, \ldots, \psi_{66})$, where all the diagonal elements are equal to 0.5. The number of free parameters q is equal to 13. The generating mean vector is a null vector, and because the confirmatory factor model does not have an explicit mean structure, the means are treated as nuisance parameters fixed to their MLEs during estimation.

Multivariate normal data were simulated from the covariance structure in Equation (19) under two kinds of missing data mechanisms: missing completely at random (MCAR) and MAR. For the MCAR condition, the following procedure was used. First, for each observation, a fair coin was flipped to decide whether there should be any missing values. Next, for the cases with missing values, two uniform integers between 1 and 6, inclusive, were generated and the values of the manifest variables indexed by the two integers were set to missing. This produces about 16% missing observations. For MAR, there are two conditions for the proportion of missing observations. The first three manifest variables were set to missing if the mean of the last three manifest variables were less than either -0.67 (MAR1) or -1 (MAR2), leaving either about 25 % or 15% of all cases with missing observations, respectively. In addition, a condition with no missing data (NOMIS) is also included as a benchmark.

The four missing data conditions (NOMIS, MCAR, MAR1, MAR2) were crossed with three sample sizes: $N = 100, 300, 500$, resulting in 12 simulation conditions. In each condition, 500 replications were attempted. For each replication, the EM2S estimator was used to fit the confirmatory factor analysis model. The stopping criterion for the EM was 1×10^{-10}. For the Supplemented EM algorithm, the stopping criterion was 1×10^{-5}. A BFGS solver was used to optimize the log-likelihood in the second-stage structural model fitting. Nonconverged replications were discarded. The code for data generation, first-stage EM estimation, and second-stage model fitting were all programmed in GAUSS (Version 6.08, Aptech Systems, Inc., 2003).

Two GOF statistics were calculated in each converged replication, T_F from Equation (16) and T_B from Equation (17). The reference distribution for both statistics is a central chi-square distribution with eight degrees of freedom.

The results are summarized in Table 4. We tabulate observed Type I error rates at three significance levels: .01, .05, and .10. It is evident that T_F is not distributed as a chi-square variable when there are missing data. The rejection rates are much too high. On the other hand, T_B maintains good Type I error rate control under missing or no missing data conditions. The distribution of T_B clearly approaches that of a chi-square variable as N increases.

TABLE 4
Simulation One: Type I Error

							Significance Level		
N	*Converged*	*Statistic*	*M*	*SD*	*Min*	*Max*	*0.01*	*0.05*	*0.10*
				NOMIS					
100	500	T_F	8.26	3.98	0.63	23.91	.010	.056	.136
		T_B	7.91	3.77	0.61	20.84	.006	.040	.090
300	500	T_F	8.01	4.15	0.98	26.32	.016	.052	.114
		T_B	8.01	3.92	1.20	22.84	.008	.058	.104
500	500	T_F	8.19	4.09	1.13	24.22	.014	.056	.112
		T_B	8.07	4.00	1.12	25.49	.010	.052	.106
				MCAR					
100	480	T_F	12.94	6.73	1.46	41.56	.135	.312	.427
		T_B	8.09	4.02	0.80	23.15	.008	.051	.106
300	488	T_F	12.02	6.00	1.72	36.67	.105	.240	.352
		T_B	8.06	3.97	1.20	25.15	.014	.043	.105
500	490	T_F	11.81	5.93	1.46	38.27	.098	.233	.331
		T_B	7.98	3.95	1.00	26.03	.010	.050	.104
				MAR1					
100	495	T_F	11.97	6.46	0.92	38.90	.103	.259	.386
		T_B	7.79	3.77	0.64	20.85	.006	.040	.075
300	500	T_F	11.38	5.76	1.32	39.85	.068	.210	.310
		T_B	7.92	3.98	1.03	28.33	.010	.052	.084
500	500	T_F	11.38	5.84	2.09	36.93	.082	.194	.308
		T_B	7.99	4.00	1.55	25.66	.012	.054	.090
				MAR2					
100	498	T_F	10.28	5.28	1.10	31.36	.060	.165	.261
		T_B	7.88	3.72	0.92	20.29	.002	.040	.092
300	500	T_F	9.75	4.93	1.39	32.20	.030	.124	.206
		T_B	7.92	3.95	1.21	27.60	.008	.044	.094
500	500	T_F	9.81	4.93	1.71	28.94	.040	.128	.198
		T_B	8.02	3.98	1.33	26.18	.012	.054	.104

Note. The entries in the Converged column refer to the number of converged replications in each condition.

Simulation Two: Power

Although having properly maintained Type I error rates is obligatory, the proposed T_B statistic would not be useful if it had no power to detect model misspecification. To investigate this issue, we conducted another simulation, using exactly the same generating confirmatory factor model, sample size, and missing data conditions as in the first simulation. The difference is that the two factors in the fitted model are constrained to be orthogonal. With the factor correlation set to .2 in the generating model, the fitted model contains a relatively small amount of misspecification.

Results are summarized in in Table 5. The reference distribution for both T_F and T_B is a central chi-square with nine degrees of freedom. The power estimates are in the last three columns of the table. Generally speaking, T_B is markedly less powerful than T_F when some data are missing, but this advantage of T_F is mostly due to the elevated Type I error rates. In the NOMIS condition, the power of T_B is also less than T_F but not appreciably so, especially when N is large. Though T_F may dominate T_B's power curve, the lack of a well-defined null distribution makes its usefulness for applied research less clear. On the other hand, for about 5% less power in the NOMIS condition, one can now have a GOF statistic (T_B) with a known distribution. We believe that the balance clearly favors T_B.

EXAMPLES

We use two examples to illustrate the foregoing theoretical and simulation results. The computations were carried out using an SAS macro that we wrote. The interested reader may download a free copy from Li Cai's Web site at http://lcai.bol.ucla.edu/programs.html. The usage of the macro is documented in the Appendix.

Confirmatory Factor Model

Following Yuan and Bentler (2000), we first apply T_B to Mardia, Kent, and Bibby's (1979) Open-book Closed-book data set (pp. 3–4). This data set contains test scores on $p = 5$ subject areas from $N = 88$ examinees. For the original complete data set, a correlated two-factor solution fits well. The factor pattern is

$$\boldsymbol{\Lambda}^T = \begin{pmatrix} \lambda_{11} & \lambda_{21} & 0 & 0 & 0 \\ 0 & 0 & \lambda_{32} & \lambda_{42} & \lambda_{52} \end{pmatrix}.$$

This model has 11 free parameters, so there are four degrees of freedom. For the complete data, $T_F = 2.07$ and $T_B = 1.84$. To create data that are MAR, we

TABLE 5
Simulation Two: Power

N	*Converged*	*Statistic*	*M*	*SD*	*Min*	*Max*	*Significance Level* 0.01	0.05	0.10
				NOMIS					
100	500	T_F	11.77	5.58	0.97	34.70	.052	.182	.260
		T_B	10.37	4.34	0.96	26.50	.018	.080	.168
300	500	T_F	16.70	7.52	2.87	47.89	.208	.416	.540
		T_B	15.62	6.55	2.85	39.37	.164	.366	.482
500	500	T_F	20.66	8.12	3.75	48.05	.410	.636	.746
		T_B	19.65	7.37	3.72	44.56	.360	.592	.724
				MCAR					
100	480	T_F	16.49	7.76	1.76	50.18	.206	.435	.529
		T_B	10.47	4.34	1.26	25.55	.019	.081	.152
300	484	T_F	20.67	8.59	3.63	60.53	.409	.626	.725
		T_B	15.05	5.92	2.90	41.83	.120	.347	.492
500	481	T_F	25.34	9.74	4.69	67.98	.599	.796	.879
		T_B	19.28	7.21	3.38	50.00	.343	.590	.690
				MAR1					
100	483	T_F	17.70	8.98	2.81	63.87	.286	.441	.561
		T_B	9.72	4.17	2.00	25.34	.006	.060	.126
300	496	T_F	21.43	10.97	2.03	67.95	.407	.619	.690
		T_B	12.11	5.60	1.56	38.36	.067	.165	.260
500	495	T_F	27.07	12.43	5.63	77.64	.600	.784	.846
		T_B	14.72	6.24	3.23	40.17	.127	.317	.446
				MAR2					
100	486	T_F	14.60	7.12	3.07	40.89	.148	.298	.430
		T_B	9.87	4.12	2.20	24.38	.006	.064	.138
300	500	T_F	18.76	8.57	3.18	61.14	.322	.510	.638
		T_B	12.91	5.29	2.21	32.45	.072	.200	.328
500	500	T_F	23.39	10.10	3.83	65.94	.508	.710	.808
		T_B	15.76	6.17	2.96	39.80	.166	.372	.540

Note. The entries in the "Converged" column refer to the number of converged replications in each condition.

used a method similar to Yuan & Bentler's (2000) method II (p. 180), wherein the scores on the last three subjects were set to missing if the sum of the first two was less than 80. As a result, 28 cases have missing values.

For this data set with missing observations, $T_F = 11.92$, $p = 0.02$, and $T_B = 5.82$, $p = 0.21$. This example reflects the difference between T_F and T_B seen in the simulations. As a comparison, we also used the FIML procedure implemented in LISREL (Jöreskog & Sörbom, 2001) to fit the same model. The

TABLE 6
Standard Errors of Factor Loadings

		SE	
	Estimate	*Complete Data Covariance*	*Supplemented EM Approx.*
λ_{11}	11.79	1.76	1.87
λ_{21}	10.67	1.29	1.40
λ_{32}	10.47	0.88	1.33
λ_{42}	13.67	1.43	2.05
λ_{52}	14.44	1.65	2.49

FIML chi-square came out to be 7.00, $p = 0.12$. Due to the fortunate availability of complete data, we know that the two-factor model fits well. However, using the naive GOF statistic T_F may lead to the erroneous rejection of the two-factor model, whereas according to either T_B or the FIML chi-square, the conclusion may be quite the opposite.

For this model we also computed the standard errors. As an illustration we only look at the five factor loadings. Table 6 shows the EM2S estimates, along with two sets of standard errors: unadjusted complete data standard errors and Supplemented EM standard errors. One can see that although the Supplemented EM standard errors are generally larger, the standard errors for the three loadings on the second factor are noticeably larger than their complete data counterparts. This is to be expected as only the last three variables have missing observations. The inflated standard errors show that Supplemented EM has properly accounted for the missing information.

Conditional Latent Curve Model

The data analyzed here come from a symposium at the 1997 meeting of the Society for Research on Child Development (Curran, 1997). The data set contains four repeated measures of $N = 405$ participants from the National Longitudinal Survey of Youth on their aggressive behavior. A number of time-invariant covariates, including gender and mother's age, are also available. This particular data set can be retrieved from http://www.unc.edu/~curran/example.htm. The data set exhibits substantial attrition. Roughly half of the cases have missing observations on one or more measurement occasions beyond baseline.

A linear latent curve model (Bollen & Curran, 2006) for the repeated measurements of aggressive behavior was fitted using the EM2S estimator, with gender and mother's age serving as time-invariant covariates. We freely estimate the covariance matrix of the intercept and slope. The time-specific residual variances are allowed to be heteroskedastic.

This model has nine degrees of freedom. The naive GOF statistic T_F is equal to 21.07, leading to a p value of .012. On the other hand, T_B is 16.82, and the corresponding p value is .052. Once again, the difference between T_F and T_B may be large enough to lead to qualitatively different conclusions about model fit. It is of interest to note that in this case the FIML chi-square is equal to 16.56, $p = .056$, a number much closer to T_B than T_F.

DISCUSSION

In this article, the Supplemented EM algorithm is applied in a novel way to solve a long-standing problem in the goodness-of-fit testing of covariance structure models fitted to incomplete data using the EM two-stage estimator. We have shown that the central idea is to use the Supplemented EM algorithm to compute an asymptotic covariance matrix of the EM estimated moments based on incomplete data. Although the Supplemented EM algorithm is not new and the EM2S estimator well known, we believe that it is important to point out their relationship because there exists a class of problems, all involving the EM algorithm, that would be easy to solve as applications of the Supplemented EM algorithm (see, e.g., Cai, 2008). Furthermore, given the popularity of the EM algorithm, the Supplemented EM algorithm should be a useful addition to the toolbox of data analysts in the behavioral sciences.

This article serves two purposes. First, the statistic T_B is new and potentially useful. Second, we show that it takes only minimal additional programming to add Supplemented EM computations to the existing EM code. We hope that this article provides enough motivation for authors of popular CSM software packages to start investigating the possibility of including the Supplemented EM algorithm.

There are of course limitations to our results. First, as a reviewer pointed out, we have not systematically compared the performance of T_B with the FIML chi-square statistic. The limited empirical evidence we have suggests that they behave similarly. However, we cannot be certain unless a comprehensive simulation study is conducted on that specific topic. Second, a reviewer remarked that the overall chi-square is only one aspect of evaluating model fit, with well-known weaknesses. We wholeheartedly agree. Indeed, we see the research presented here as merely a first step toward a better understanding of CSM fit evaluation under missing data, and we will look into other model fit indices in future work. Third, the macro is currently restricted to the SAS software. However, the Supplemented EM covariance matrix is saved as a SAS data set by the macro and can, in principle, be brought into any other CSM software. The

Appendix describes this output data set. Finally, we have made no provisions for nonnormality in the observed variables. This is partly intended because we do not wish to be bogged down by the technical details. On the other hand, the Supplemented EM algorithm turns out to be a convenient method for handling nonnormal incomplete data estimation problems. As shown by Yuan and Bentler (2000), one can use a two-stage weighted least squares approach with the first stage being the same as in EM2S, but in the second stage the structural parameters are estimated by minimizing a weighted least squares discrepancy function with the weight matrix being the inverse of a "sandwich" covariance matrix of the EM estimated moments. We note that the two pieces of "toast" in Yuan & Bentler's "sandwich" covariance matrix have already been computed with the Supplemented EM algorithm.

Before concluding, we point out that the Supplemented EM algorithm has still other uses. For instance, it can help diagnose the convergence of the EM algorithm and potentially also detect programming errors in the EM code. Meng and Rubin (1991) discuss the alternative uses in detail.

ACKNOWLEDGMENTS

We thank the editor and the reviewers for helpful comments and suggestions that led to an improved article. Li Cai also gratefully acknowledges financial support from the following funding agencies: National Science Foundation (# SES-0717941), the National Center for Research on Evaluation, Standards and Student Testing (CRESST) through an award from the U.S. Department of Education's Institute of Education Sciences (IES) (# R305A050004), and a pre-doctoral advanced quantitative methods training grant awarded to the UCLA Departments of Education and Psychology by IES. The views expressed in this paper are of the authors alone and do not reflect the views or policies of the funding agencies.

REFERENCES

Allison, P. D. (2003). Missing data techniques for structural equation modeling. *Journal of Abnormal Psychology*, *112*, 545–557.

Anderson, T. W. (1957). Maximum likelihood estimates for the multivariate normal distribution when some observations are missing. *Journal of the American Statistical Association*, *52*, 200–203.

Aptech Systems, Inc. (2003). GAUSS (Version 6.08) [Computer software]. Maple Valley, WA: Author.

Arbuckle, J. L. (1996). Full information estimation in the presence of incomplete data. In G. A. Marcoulides & R. E. Schumacker (Eds.), *Advanced structural equation modeling: Issues and techniques* (pp. 697–715). Mahwah, NJ: Erlbaum.

Bollen, K. A., & Curran, P. J. (2006). *Latent curve models: A structural equation perspective*. New York: Wiley.
Browne, M. W. (1984). Asymptotically distribution-free methods for the analysis of covariance structures. *British Journal of Mathematical and Statistical Psychology*, *37*, 62–83.
Browne, M. W., & Arminger, G. (1995). Specification and estimation of mean-and covariance-structure models. In G. Arminger, C. C. Clogg, & M. E. Sobel (Eds.), *Handbook of statistical modeling for social an behavioral sciences* (pp. 185–249). New York: Plenum.
Cai, L. (2008). SEM of another flavour: Two new applications of the supplemented EM algorithm. *British Journal of Mathematical and Statistical Psychology*, *61*, 309–329.
Collins, L. M., Schafer, J. L., & Kam, C.-M. (2001). A comparison of inclusive and restrictive strategies in modern missing-data procedures. *Psychological Methods*, *6*, 330–351.
Curran, P. J. (1997). *Comparing three modern approaches to longitudinal data analysis: An examination of a single developmental sample.* Symposium presented at the 1997 Biennial Meeting of the Society for Research in Child Development, Washington, DC.
Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood estimation from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society—Series B*, *39*, 1–38.
Enders, C. K., & Peugh, J. L. (2004). Using an EM covariance matrix to estimate structural equation models with missing data: Choosing an adjusted sample size to improve the accuracy of inferences. *Structural Equation Modeling*, *11*, 1–19.
Graham, J. W. (2003). Adding missing-data relevant variables to FIML-based structural equation models. *Structural Equation Modeling*, *10*, 80–100.
Jöreskog, K., & Sörbom, D. (2001). *LISREL user's guide*. Chicago: SSI International.
Louis, T. A. (1982). Finding the observed information matrix when using the EM algorithm. *Journal of the Royal Statistical Society—Series B*, *44*, 226–233.
Mardia, K. V., Kent, J. T., & Bibby, J. M. (1979). *Multivariate analysis*. London: Academic.
McLachlan, G. J., & Krishnan, T. (1996). *The EM algorithm and extensions*. New York: Wiley.
Meilijson, I. (1989). A fast improvement to the EM algorithm on its own terms. *Journal of the Royal Statistical Society—Series B*, *51*, 127–138.
Meng, X.-L., & Rubin, D. B. (1991). Using EM to obtain asymptotic variance-covariance matrices: The SEM algorithm. *Journal of the American Statistical Association*, *86*, 899–909.
Orchard, T., & Woodbury, M. A. (1972). A missing information principle: Theory and application. In L. M. Lecam, J. Neyman, & E. L. Scott (Eds.), *Proceedings of the sixth Berkeley symposium on mathematical statistics and probability* (pp. 697–715). Berkeley: University of California Press.
Rao, C. R. (1973). *Linear statistical inference and its applications* (2nd ed.). New York: Wiley.
Schafer, J. L. (1997). *Analysis of incomplete multivariate data*. New York: Chapman & Hall/CRC.
Schott, J. R. (1997). *Matrix analysis for statistics*. New York: Wiley.
Yuan, K.-H., & Bentler, P. M. (2000). Three likelihood-based methods for mean and covariance structure analysis with nonnormal missing data. *Sociological Methodology*, *30*, 165–200.

APPENDIX

Usage of the SAS Macro *SEM*

We use the conditional latent curve model as an example to describe the usage of the SAS macro. Suppose an SAS data set called *srcd* is present in the *work* library, and it contains the following six variables: *anti1–anti4*, *gen*, *momage*. The first four *anti* variables are the repeated measures of aggressive behavior, and *gen* and *momage* are the gender and mother's age variables at baseline. First, the SAS macro definitions should be included using the *%include* statement.

Next, one must define the structural model to be fitted to the EM estimated means and covariances using one of the programming statements acceptable to PROC CALIS. The model definition *must* be a quoted SAS macro string and in this example we use the LINEQS-style statements:

```
%let mymodel=%str(
  lineqs
    anti1 = 1f1 + 0f2 + e1,
    anti2 = 1f1 + 1f2 + e2,
    anti3 = 1f1 + 2f2 + e3,
    anti4 = 1f1 + 3f2 + e4,
    f1 = al1 intercept + gamma1 gen + gamma2 momage + d1,
    f2 = al2 intercept + gamma3 gen + gamma4 momage + d2,
    gen = al3 intercept + d3,
    momage = al4 intercept + d4;
  std
    e1-e4 = th1 th2 th3 th4,
    d1-d4 = ph11 ph22 ph33 ph44;
  cov
    d1 d2 = ph12,
    d3 d4 = ph34;
);
```

Finally, the macro *SEM* is invoked.

```
%SEM(indata=srcd,var=anti1 anti2 anti3 anti4 gen momage,
     nvar=6,nobs=405,calismodel=&mymodel);
```

The *indata* and *var* options tell the macro the data set name and the names of the variables. The *nvar* and *nobs* options give the dimensions of the analysis in terms of the number of manifest variables and the number of observations. Finally, the quoted macro string *&mymodel* is passed on to the macro to define the structural model.

The macro produces the following goodness-of-fit output:

```
            EM Iteration Number
                              42
```

```
          SEM Iteration Number
                            15

 Goodness-of-fit Test for EM2S Estimator
       T_B             df        p-value
   16.8207         9.0000         0.0516
```

The EM iteration number and the SEM iteration number are, respectively, the number of cycles required for the EM and the SEM algorithms to reach convergence. The value of T_B, the degrees of freedom, and the p value are printed at the end. The macro also produces adjusted standard errors for the structural parameters using Equation (18).

The macro will place an SAS data set named *Vmatrix* under the *work* library that contains the Supplemented EM covariance matrix for the estimated means and covariances. This data set can be imported into other CSM software packages to serve as a weight matrix.