5

THREE LIKELIHOOD-BASED METHODS FOR MEAN AND COVARIANCE STRUCTURE ANALYSIS WITH NONNORMAL MISSING DATA

*Ke-Hai Yuan**
Peter M. Bentler†

Survey and longitudinal studies in the social and behavioral sciences generally contain missing data. Mean and covariance structure models play an important role in analyzing such data. Two promising methods for dealing with missing data are a direct maximum-likelihood and a two-stage approach based on the unstructured mean and covariance estimates obtained by the EM-algorithm. Typical assumptions under these two methods are ignorable nonresponse and normality of data. However, data sets in social and behavioral sciences are seldom normal, and experience with these procedures indicates that normal theory based methods for nonnormal data very often lead to incorrect model evaluations. By dropping the normal distribution assumption, we develop more accurate procedures for model inference. Based on the theory of generalized estimating equations, a way to obtain consistent standard errors of the two-stage estimates is given. The asymptotic efficiencies of different estimators are compared under various assumptions. We also propose a minimum chi-square approach and show that the estimator obtained by

This work was supported by National Institute on Drug Abuse Grants DA01070 and DA00017. We gratefully acknowledge the constructive feedback of the editors and two referees.

*University of North Texas
†University of California, Los Angeles

this approach is asymptotically at least as efficient as the two likelihood-based estimators for either normal or nonnormal data. The major contribution of this paper is that for each estimator, we give a test statistic whose asymptotic distribution is chi-square as long as the underlying sampling distribution enjoys finite fourth-order moments. We also give a characterization for each of the two likelihood ratio test statistics when the underlying distribution is nonnormal. Modifications to the likelihood ratio statistics are also given. Our working assumption is that the missing data mechanism is missing completely at random. Examples and Monte Carlo studies indicate that, for commonly encountered nonnormal distributions, the procedures developed in this paper are quite reliable even for samples with missing data that are missing at random.

1. INTRODUCTION

Mean and covariance structure models play an important role in understanding relationships among multivariate observations. With popular software—e.g., LISREL (Jöreskog and Sörbom, 1993) and EQS (Bentler 1995)—these kinds of models are widely used in social and behavioral sciences. Special cases include path analysis, confirmatory factor analysis, errors-in-variables, simultaneous equations, and other latent variable structural equation models (e.g., see Kline 1998; Mueller 1996). With a complete sample, various approaches to estimation and testing in mean and covariance structures have been developed (e.g., Bollen 1989; Browne 1984; Browne and Arminger 1995; Satorra and Bentler 1988, 1994; Yuan and Bentler 1997a, b 1998). However, in the social and behavioral sciences, data collection may involve long questionnaires that are tiring to fill out precisely, and a single study may take several years to complete, so that missing data are almost inevitable. Despite this fact, not many statistically sound approaches to mean and covariance structures with missing data are available, especially, for the typical nonnormal data sets that are found in social and behavioral sciences (Micceri 1989). There is no effective way to evaluate a hypothesized mean and covariance structure $\mu = \mu(\theta_0)$ and $\Sigma = \Sigma(\theta_0)$, generally regarded as the most critical part of modeling hypothetical relations among latent variables.

When data are multivariate normal and the missing data mechanism is ignorable, which includes missing at random (MAR) and missing completely at random (MCAR) (Little and Rubin 1987; Rubin 1987; Schafer 1997), estimation and inference can be accomplished by maximum likelihood (ML) facilitated by the EM-algorithm (Dempster et al. 1977; Little and Rubin 1987; Meng and Pedlow 1992; Schafer 1997). One such approach is a two-stage method (e.g., Brown 1983; Finkbeiner 1979; Rovine 1994). In the first stage of this approach, estimates $\overline{X}_n$ and S_n of μ and Σ are obtained through the EM-algorithm based on a multivariate normality assumption. The second stage is to proceed with the analysis as in the complete data case, treating $\overline{X}_n$ and S_n as the sample mean and sample covariance matrix. In this stage, one obtains an estimate $\tilde{\theta}$ of θ_0 by minimizing the likelihood ratio function based on a normality assumption

$$\begin{aligned} F_{ML}(\theta) &= \mathrm{tr}(S_n\Sigma^{-1}(\theta)) - \log|S_n\Sigma^{-1}(\theta)| \\ &\quad + (\overline{X}_n - \mu(\theta))'\Sigma^{-1}(\theta)(\overline{X}_n - \mu(\theta)) - p, \end{aligned} \tag{1}$$

where p is the number of variables, and $T_1 = nF_{ML}(\tilde{\theta})$ is a test statistic to evaluate the model structure (e.g., Browne and Arminger 1995). Another approach to obtain an estimate of θ_0 is a direct ML method (e.g., Allison 1987; Arbuckle 1996; Finkbeiner 1979; Jamshidian and Bentler 1999; Lee 1986; Muthén et al. 1987). For the ith observed case X_i with dimension p_i, $E(X_i) = \mu_i$ and $\mathrm{cov}(X_i) = \Sigma_i$, where the latter are a subvector and a submatrix of μ and Σ respectively. With

$$\begin{aligned} l_i(\theta) = \frac{p_i}{2}\log(2\pi) - \frac{1}{2}\{&\log|\Sigma_i(\theta)| + (X_i - \mu_i(\theta))' \\ &\times \Sigma_i^{-1}(\theta)(X_i - \mu_i(\theta))\}, \end{aligned} \tag{2}$$

the direct maximum-likelihood estimate (MLE) $\hat{\theta}$ can be obtained by maximizing the log-likelihood function

$$l(\theta) = \sum_{i=1}^{n} l_i(\theta).$$

Let vech$(\cdot)$ be an operator that transforms a symmetric matrix into a vector by stacking the columns of the matrix, leaving out the elements above the diagonal. We will use $\sigma = \mathrm{vech}(\Sigma)$ and $\beta = (\sigma', \mu')'$ for convenience. With missing data and a saturated model, the μ_i and Σ_i are functions of β, and the corresponding log-likelihood function is given by

$$l(\beta) = \sum_{i=1}^{n} l_i(\beta). \tag{3}$$

Denote $\hat{\beta}$ the estimator of β_0 obtained by maximizing (3). Based on nesting in the saturated model (μ, Σ), the associated likelihood-ratio statistic T_2 for testing the structure $(\mu(\theta), \Sigma(\theta))$ is (e.g., Arbuckle 1996).

$$T_2 = 2[l(\hat{\beta}) - l(\hat{\theta})].$$

Because no better alternative is available, these two approaches are commonly used in practice even when data are nonnormal. For example, the direct ML approach is implemented in the computer programs LINCS (Schoenberg 1989), AMOS (Arbuckle 1996), Mplus (Muthén and Muthén 1998), and Mx (Neale 1994). Under the assumption of multivariate normality and the null hypothesis, T_2 is asymptotically distributed as $\chi^2_{p+p^*-q}$, where $p^* = p(p+1)/2$ and q is the number of unknown parameters in θ. When data are complete, T_1 equals T_2 and equals the commonly used Wishart likelihood-ratio statistic T_{ML}. For complete data, conditions also exist for T_{ML} to be valid for nonnormal data with some specific models (Amemiya and Anderson 1990; Anderson and Amemiya 1988; Browne 1987; Browne and Shapiro 1988; Mooijaart and Bentler 1991; Satorra and Bentler 1990, 1991; Shapiro 1987; Yuan and Bentler 1999). Unfortunately, there is no effective way of verifying these conditions in practice. When these conditions are not satisfied, normal theory methods generally lead to incorrect model evaluation and misleading substantive conclusion even in the complete data cases (e.g., Hu et al. 1992; Curran et al. 1996; Yuan and Bentler 1998), and it is unlikely that one can avoid the incorrectness with an added missing data problem. Ideally, one would model the data by finding a nonnormal distribution, and inference could proceed with maximum likelihood. Unfortunately, distributions of real data in the social and behavioral sciences tend to be skewed and to have heterogeneous marginal kurtoses. Existing classes of multivariate nonnormal distributions are either too restricted or have unknown distributional forms (e.g., Olkin 1994; Fang et al. 1990; Yuan and Bentler 1999). It is not an exaggeration to say that, in practice, any application of the maximum-likelihood method with multivariate data is at best only an approximation to reality. Normal theory-based methods represent one type of such an approximation.

A breakthrough in mean and covariance structure analysis with nonnormal missing data was made by Arminger and Sobel (1990). Using the pseudo maximum-likelihood (PML) theory developed by Gourieroux et al. (1984), Arminger and Sobel proposed the sandwich-type covariance matrix in describing the distribution of parameter estimates instead of the inverse of the normal theory-based information matrix. Sandwich-type covariance matrices for covariance structure analysis with complete data have been proposed by Dijkstra (1981), Bentler (1983), Shapiro (1983), Browne (1984), Arminger and Schoenberg (1989) and Browne and Arminger (1995). Our experience indicates that standard errors based on the sandwich-type covariance matrix are much more accurate in evaluating parameter significance than those based on normal theory methods. On the other hand, the sandwich-type standard errors behave similarly with those based on normal theory methods when data are multivariate normal (e.g., Yuan and Bentler 1997b). Based on simulation studies with various distribution conditions, Yuan and Bentler recommended the sandwich-type covariance matrix as the default output in structural equation modeling programs.

Almost all the normal theory-based missing data methods assume that the missing data mechanism is MAR. This assumption may not be realistic with practical data (e.g., Allison 1987:77). A more frank attitude should be that some missing variables are MCAR, some are MAR, and some may even be not missing at random (NMAR). Like the normality assumption on distribution, MAR represents only a working assumption for missing data. If data are normal and all the missing variables are either MAR or MCAR, parameter estimates based on maximum likelihood will be consistent (Rubin 1987). However, admitting an incorrect distributional specification, according to Laird (1988) and Rotnitzky and Wypij (1994), the parameter estimates will be inconsistent unless the missing data mechanism is MCAR. With missing data from a distribution having heterogeneous marginal skewness and kurtosis, it is not clear how to specify the likelihood function for obtaining the true MLEs. Normal theory-based maximum likelihood is still a working approach to missing data problems in mean and covariance structures regardless of what the missing data mechanisms are. Actually, whatever the assumption on the missing data mechanism is, once a fitting function as in (1) or (2) is chosen, the MLEs of the model parameters are the same. If there is a bias in a parameter estimate, the bias will not disappear because some untrue assumptions have been accepted. In order to obtain sensible statistical inference, there are two sets of assumptions one can choose: (I) normal data and MAR mechanism; (II) nonnormal data and MCAR mechanism. Besides Arminger and Sobel (1990) and Yuan and Bentler (1996), almost all missing data methods for mean and covariance structure analysis assume normal data. All the likelihood-based methods assume the MAR mechanism. The advantage of picking assumptions (I) is that one can use the likelihood ratio statistic and Fisher information-based standard errors for model inference. Unfortunately, inference based on normal theory can be quite misleading in practice. If assuming (II), one has to develop new statistics and standard errors for model inference.

Arminger and Sobel (1990) developed the sandwich-type covariance matrix for parameter estimates obtained by maximizing (2). With assumptions (II), the present paper has the following four contributions. First, a new and rather natural approach for estimating θ_0 is proposed. The new estimator $\check{\theta}$ is asymptotically at least as efficient as either $\tilde{\theta}$ or $\hat{\theta}$. Second, we will develop a sandwich-type covariance matrix for $\tilde{\theta}$, and study the asymptotic efficiencies of the estimators $\tilde{\theta}$, $\hat{\theta}$, and $\check{\theta}$. Third, we will characterize the asymptotic distributions of T_1 and T_2. Rescaled versions of these test statistics, whose distributions should be better approximated by the proposed chi-square distributions, will be given. Fourth, for each estimator we give a test statistic whose asymptotic distribution is chi-square regardless of the underlying distribution. In addition to the above four contributions, we also study how much bias in parameter estimates occurs when MCAR is not a realistic assumption. Our statistical development of these procedures is based on the approach of the generalized estimating equation developed in Liang and Zeger (1986) (see also Yuan and Jennrich 1998). When applied to mean and covariance structure analysis, the regularity conditions involved are as follows: the population covariance matrix is positive definite, fourth-order moments exist, and the structural model is identified and twice continuously differentiable. These regularity conditions will be assumed throughout our development, without being specified explicitly. A closely related approach is the PML theory set out by Gourieroux et al. (1984) and applied to mean and covariance structure models by Arminger and Schoenberg (1989) and Arminger and Sobel (1990). Parallel developments for covariance structures with complete data are made by Dijkstra (1981), Shapiro (1983), Browne (1984), Arminger and Schoenberg (1989), and Browne and Arminger (1995), and for arbitrary types of structural models by Bentler and Dijkstra (1985).

The rest of this paper will be arranged as follows: The asymptotic distributions of different estimators and their asymptotic efficiencies will be studied in Section 2. Different test statistics for evaluating model structures will be given in Section 3. Since many applications in practice involve only a structured covariance matrix with a nuisance mean vector, we will turn to this special case in Section 4. Illustrations and comparisons among various procedures are presented in Section 5. Simulation studies on parameter bias will be described in Section 6. Some concluding remarks are given at the end of the paper. An outline of relatively complicated proofs will be given in an appendix.

2. ASYMPTOTIC DISTRIBUTION AND EFFICIENCY

In this section, we will study and compare three estimators: (1) the two-stage estimator $\tilde{\theta}$, (2) the direct MLE $\hat{\theta}$, and (3) a minimum chi-square estimator $\check{\theta}$ that will be introduced in this section.

In order to study the asymptotic distribution of $\tilde{\theta}$, we need to know the distribution of $\overline{X}_n$ and S_n. Since the E-step of the EM-algorithm is based on the normality assumption, the parameter estimate $\hat{\beta} = (\text{vech}'(S_n), \overline{X}_n')'$ of $\beta = (\sigma', \mu')'$ actually maximizes the log-likelihood function in (3). Consequently, $\hat{\beta}$ is a stationary point of the generalized estimating equation

$$G(\hat{\beta}) = 0, \tag{4a}$$

where

$$G(\beta) = \frac{1}{n}\sum_{i=1}^{n} \frac{\partial l_i(\beta)}{\partial \beta}. \tag{4b}$$

Let β_0 denote the population counterpart of β. If $E[G(\beta_0)] = 0$, then under some standard regularity conditions, $\hat{\beta}$ is strongly consistent and asymptotically normally distributed (Yuan and Jennrich 1998), and

$$\sqrt{n}(\hat{\beta} - \beta_0) \xrightarrow{\mathcal{L}} N(0, \Omega_{\hat{\beta}}), \tag{5a}$$

where $\Omega_\beta = A_\beta^{-1} B_\beta A_\beta^{-1}$ with

$$A_\beta = -\lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} \frac{\partial^2 l_i(\beta_0)}{\partial \beta_0 \partial \beta_0'}, \qquad B_\beta = \lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} \frac{\partial l_i(\beta_0)}{\partial \beta_0} \frac{\partial l_i(\beta_0)}{\partial \beta_0'}. \tag{5b}$$

Similar results can be obtained using the PML theory (Gourieroux et al. 1984; Arminger and Sobel 1990). The estimating equation (4) is called unbiased if $E[G(\beta_0)] = 0$ (e.g., Godambe and Kale 1991). In such a case, the result in (5) holds regardless of what the underlying distribution of the data is. When the missing data mechanism is MCAR, (4) is unbiased. When data are normal, (4) is also unbiased for MAR data. When $X_i \sim N(\mu_i(\mu), \Sigma_i(\Sigma))$, then $A_\beta = B_\beta$, which corresponds to the normal theory information matrix and $\Omega_\beta = A_\beta^{-1}$ is the minimum asymptotic covariance matrix any estimator can achieve. For a general nonnormal distribution, $\hat{\beta}$ does not enjoy such an optimum property. However, for the unbiased estimating equation (4), regardless of the distribution of the data, consistent estimates of A_β and B_β can be obtained by replacing the unknown parameter β_0 in (5b) by $\hat{\beta}$ and omitting the limit notation; we will denote such an estimator by $\hat{\Omega}_{\hat{\beta}}$. In addition, note that the form of A_β also does not depend on the underlying distribution. Let

$$\kappa_i = \frac{\partial\,\mathrm{vec}(\Sigma_i)}{\partial\sigma'}, \qquad \text{and} \qquad \tau_i = \frac{\partial\mu_i}{\partial\mu'},$$

which are respectively $p_i^2 \times p^*$ and $p_i \times p$ constant matrices with elements being either 1 or 0, then

$$A_\beta = \begin{pmatrix} \frac{1}{2}\lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}\kappa_i'(\Sigma_i^{-1}\otimes\Sigma_i^{-1})\kappa_i & 0 \\ 0 & \lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}\tau_i'\Sigma_i^{-1}\tau_i \end{pmatrix}.$$

We will denote $H_2 = A_\beta$ for easy comparison of efficiency.

Equation (5) gives a justification for using the normal assumption based EM-algorithm to impute the missing data from a nonnormal distribution. This result is necessary to study the distribution of $\tilde{\theta}$, as well as many other applications in which S_n or $\overline{X}_n$ are used in multivariate analysis.

Now we can study the asymptotic distribution of $\tilde{\theta}$. When $\Sigma_i = \Sigma$, then $\kappa_i = D_p$, which is the duplication matrix as defined in Magnus and Neudecker (1988, p. 49). Let

$$H_1 = \begin{pmatrix} \frac{1}{2}D_p'(\Sigma^{-1}\otimes\Sigma^{-1})D_p & 0 \\ 0 & \Sigma^{-1} \end{pmatrix},$$

then the asymptotic distribution of $\tilde{\theta}$ is given by (e.g., Browne and Arminger 1995)

$$\sqrt{n}(\tilde{\theta} - \theta_0) \xrightarrow{\mathcal{L}} N(0, \Omega_{\tilde{\theta}}), \tag{6a}$$

where

$$\Omega_{\tilde{\theta}} = (\dot{\beta}' H_1 \dot{\beta})^{-1} (\dot{\beta}' H_1 \Omega_{\hat{\beta}} H_1 \dot{\beta}) (\dot{\beta}' H_1 \dot{\beta})^{-1} \tag{6b}$$

with $\dot{\beta} = \partial \beta(\theta_0)/\partial \theta_0'$. A consistent estimate of $\Omega_{\tilde{\theta}}$ can be obtained by replacing Σ by S_n, $\dot{\beta}$ by $\dot{\beta}(\tilde{\theta})$, and $\Omega_{\hat{\beta}}$ by $\hat{\Omega}_{\hat{\beta}}$.Notice that the sandwich-type covariance matrix consists of the normal theory-based covariance matrix $(\dot{\beta}' H_1 \dot{\beta})^{-1}$ on each side, which is due to the normal discrepancy function equation (1). The middle block $(\dot{\beta}' H_1 \Omega_{\hat{\beta}} H_1 \dot{\beta})$ accounts for the missing information and (or) nonnormal data. Actually, for normal data without any missing variables, $\Omega_{\hat{\beta}} = H_1^{-1}$ and $\Omega_{\tilde{\theta}} = (\dot{\beta}' H_1 \dot{\beta})^{-1}$ is just the inverse of the normal theory-based information matrix.

We have obtained the distribution of $\tilde{\theta}$ using the result in (5). Another way of utilizing (5) is to estimate θ_0 by minimizing

$$Q_n(\theta) = (\hat{\beta} - \beta(\theta))' \hat{\Omega}_{\hat{\beta}}^{-1} (\hat{\beta} - \beta(\theta)). \tag{7}$$

This is a minimum chi-square method, as developed in Ferguson (1996, ch. 23), which requires sample size to be large enough, and the number of variables to be not too large, so that $\hat{\Omega}_{\hat{\beta}}$ is invertible. A parallel approach for obtaining parameter estimates with covariance structure analysis for complete data was proposed by Browne (1984). With (5) and the standard regularity conditions, the estimator $\check{\theta}$ is consistent and asymptotically normally distributed with

$$\sqrt{n}(\check{\theta} - \theta_0) \xrightarrow{\mathcal{L}} N(0, \Omega_{\check{\theta}}), \tag{8}$$

where $\Omega_{\check{\theta}} = (\dot{\beta}' \Omega_{\hat{\beta}}^{-1} \dot{\beta})^{-1}$ can be consistently estimated by replacing $\dot{\beta}$ and $\Omega_{\hat{\beta}}$ by $\dot{\beta}(\check{\theta})$ and $\hat{\Omega}_{\hat{\beta}}$ respectively.

Both the estimates $\tilde{\theta}$ and $\check{\theta}$ are two-stage estimates, requiring the initial estimate $\hat{\beta}$. The direct normal theory MLE $\hat{\theta}$ does not need to first obtain $\hat{\beta}$. Using the theory of PML, Arminger and Sobel (1990) gave the asymptotic distribution of $\hat{\theta}$,

$$\sqrt{n}(\hat{\theta} - \theta_0) \xrightarrow{\mathcal{L}} N(0, \Omega_{\hat{\theta}}), \tag{9a}$$

where $\Omega_{\hat{\theta}} = A_\theta^{-1} B_\theta A_\theta^{-1}$ with

$$A_\theta = -\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} \frac{\partial^2 l_i(\theta_0)}{\partial \theta_0 \partial \theta_0'}, \qquad B_\theta = \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} \frac{\partial l_i(\theta_0)}{\partial \theta_0} \frac{\partial l_i(\theta_0)}{\partial \theta_0'}. \tag{9b}$$

When data are normal, $A_\theta = B_\theta$, which is the normal theory information matrix associated with the model, and $\Omega_{\hat\theta} = A_\theta^{-1}$ so that $\hat\theta$ is asymptotically fully efficient.

Now we have three methods to estimate θ_0. Even though all the estimators are consistent and asymptotically normally distributed, there exist differences in their efficiencies. When data are normally distributed, we know that $\hat\theta$ is asymptotically fully efficient with $\Omega_{\hat\theta} = A_\theta^{-1}$ and $\Omega_{\tilde\theta} \geq \Omega_{\hat\theta}$ in such a case. Since $\hat\beta$ is also asymptotically efficient when data are normal with $\Omega_{\hat\beta} = A_\beta^{-1}$, we hope this efficiency could be inherited by $\check\theta$. This is actually true. Notice that the $l_i(\theta)$ in equation (2) depends on θ through $\beta = \beta(\theta)$. It follows from a chain rule of differentiation

$$\frac{\partial^2 l_i(\theta)}{\partial\theta\partial\theta'} = \frac{\partial\beta'(\theta)}{\partial\theta}\,\frac{\partial^2 l_i(\beta)}{\partial\beta\partial\beta'}\,\frac{\partial\beta(\theta)}{\partial\theta'} \tag{10}$$

that $A_\theta = \dot\beta' H_2 \dot\beta$. It follows from (8) and (9) that $\Omega_{\check\theta} = \Omega_{\hat\theta}$ for normal distributions. When data are not normal, we cannot compare the efficiency of $\hat\theta$ and $\tilde\theta$, but $\check\theta$ is asymptotically at least as efficient as either $\hat\theta$ or $\tilde\theta$. This is shown in the following. Let $Y \sim N(\xi, \Omega_{\hat\beta})$, where ξ is a mean vector of proper dimension, then $\dot\beta'\Omega_{\hat\beta}^{-1} Y$ regressed on $\dot\beta' H_1 Y$ gives

$$\begin{aligned} E(\dot\beta'\Omega_{\hat\beta}^{-1} Y | \dot\beta' H_1 Y) = \dot\beta'\Omega_{\hat\beta}^{-1}\xi + (\dot\beta' H_1 \dot\beta)(\dot\beta' H_1 \Omega_{\hat\beta} H_1 \dot\beta)^{-1} \\ \times \dot\beta' H_1 (Y - \xi). \end{aligned} \tag{11}$$

It follows from (11) that the covariance matrix of the residual $\dot\beta'\Omega_{\hat\beta}^{-1} Y - E(\dot\beta'\Omega_{\hat\beta}^{-1} Y | \dot\beta' H_1 Y)$ is given by

$$\dot\beta'\Omega_{\hat\beta}^{-1}\dot\beta - (\dot\beta' H_1 \dot\beta)(\dot\beta' H_1 \Omega_{\hat\beta} H_1 \dot\beta)^{-1}(\dot\beta' H_1 \dot\beta), \tag{12}$$

which is at least a nonnegative definite matrix. Notice that the first term in (12) is the inverse of $\Omega_{\check\theta}$ and the second term is the inverse of $\Omega_{\tilde\theta}$, it immediately follows that $\Omega_{\check\theta} \leq \Omega_{\tilde\theta}$. Similarly, when sample size is large, $\check\theta$ also has a theoretical advantage over $\hat\theta$ for general nonnormal data. Using (5) and

$$\frac{\partial l_i(\theta)}{\partial\theta'} = \frac{\partial l_i(\beta)}{\partial\beta'}\,\frac{\partial\beta(\theta)}{\partial\theta'}, \tag{13}$$

we have $B_\theta = \dot\beta' B_\beta \dot\beta = \dot\beta' H_2 \Omega_{\hat\beta} H_2 \dot\beta$ and it follows that

$$\Omega_{\hat\theta} = (\dot\beta' H_2 \dot\beta)^{-1}(\dot\beta' H_2 \Omega_{\hat\beta} H_2 \dot\beta)(\dot\beta' H_2 \dot\beta)^{-1}. \tag{14}$$

Replacing H_1 in (12) by H_2, one obtains the residual covariance matrix of $\dot{\beta}'\Omega_{\tilde{\beta}}^{-1}Y$ regressing on $\dot{\beta}'H_2Y$ as

$$\dot{\beta}'\Omega_{\tilde{\beta}}^{-1}\dot{\beta} - (\dot{\beta}'H_2\dot{\beta})(\dot{\beta}'H_2\Omega_{\hat{\beta}}H_2\dot{\beta})^{-1}(\dot{\beta}'H_2\dot{\beta}), \tag{15}$$

which implies that $\Omega_{\check{\theta}} \leq \Omega_{\hat{\theta}}$. So for both normal and nonnormal data, $\check{\theta}$ is asymptotically at least as efficient as either $\hat{\theta}$ or $\tilde{\theta}$. The inequality in $\Omega_{\check{\theta}} \leq \Omega_{\tilde{\theta}}$ is strict unless data are normal and complete, and the inequality in $\Omega_{\check{\theta}} \leq \Omega_{\hat{\theta}}$ is strict unless data are normal. Notice that the asymptotic efficiency possessed by $\check{\theta}$ may not hold for finite sample sizes, as was demonstrated in Yuan and Bentler (1997b) for complete data. This implies that unless sample size is large or data are extremely nonnormal, $\hat{\theta}$ or $\tilde{\theta}$ may be preferable in practice.

For nonnormal data, the comparison between $\hat{\theta}$ and $\tilde{\theta}$ is not so clear, and their asymptotic efficiencies depend on the underlying sampling distribution. Notice that $H_2 \approx H_1$ if the number of missing cases $n_m \ll n$. From (6b) and (14), it follows that $\Omega_{\tilde{\theta}} \approx \Omega_{\hat{\theta}}$ in such a case. Even though $\hat{\theta}$ is asymptotically more efficient when data are normal, the advantage of using the direct ML method may not be so obvious when the number of missing cases is small compared to the sample size. Also, the direct ML approach may need special programming (Jamshidian and Bentler 1999) while $(\overline{X}_n, S_n)$ can be obtained through the EM-algorithm, which is straightforward. The implication is that the two-stage method may generally be preferred in practice.

3. TEST STATISTICS

In the last section, we have shown the large sample advantage of the minimum chi-square estimator over both the direct ML and the two-stage estimators. The minimum chi-square method also automatically produces a test statistic for evaluating the structural model $\beta = \beta(\theta)$. This is because

$$T_3 = nQ(\check{\theta}) \xrightarrow{\mathcal{L}} \chi_{p+p^*-q}. \tag{16}$$

The statistic T_3 is a generalization of the asymptotically distribution free statistic proposed by Browne (1984) for the complete data case, which has been available in several major statistical software packages (e.g., LISREL, EQS).

Before we construct new test statistics associated with $\tilde{\theta}$ and $\hat{\theta}$, we would like to show that the statistics T_1 and T_2 generally do not asymptotically follow chi-square distributions with nonnormal data. As characterized in the following Lemmas 3.1 and 3.2 whose proofs are given in the appendix to this chapter, these statistics approach some mixtures of chi-square distributions instead. Based on these lemmas, we propose a rescaled version for each of the statistics. The new rescaled statistics are better approximated by the reference chi-square distribution $\chi^2_{p+p^*-q}$.

We will work on the two-stage likelihood ratio test statistic T_1 first. Let

$$V_1 = H_1 - H_1\dot{\beta}(\dot{\beta}'H_1\dot{\beta})^{-1}\dot{\beta}'H_1 \tag{17}$$

and $\lambda_j^{(1)}, j = 1, \cdots, p + p^* - q$ be the nonzero eigenvalues of $\Omega_{\hat{\beta}} V_1$. The following lemma characterizes the large sample property of T_1.

Lemma 3.1. Under some standard regularity conditions,

$$T_1 \xrightarrow{\mathcal{L}} \sum_{j=1}^{p+p^*-q} \lambda_j^{(1)} \chi^2_{j1},$$

where χ^2_{j1} are independent chi-square random variables each with degree of freedom 1.

So the two-stage likelihood ratio test statistic generally does not approach $\chi^2_{p+p^*-q}$ in distribution. When $\Omega_{\hat{\beta}}^{-1} = H_1$, $\Omega_{\hat{\beta}}^{1/2} V_1 \Omega_{\hat{\beta}}^{1/2}$ is a projection matrix, and all the $\lambda_j^{(1)}$s are equal to 1, $T_1 \xrightarrow{\mathcal{L}} \chi_{p+p^*-q}$. This is the case with complete data sampled from a normal distribution. Since T_1 approaches a distribution of linear combination of chi-squares, and no commonly used distribution is available to describe the behavior of the linear combination of chi-squares, we may rescale T_1 so that it can be better approximated by its proposed distribution. A simple choice is to rescale T_1 to have the asymptotic mean $p + p^* - q$, giving the statistic

$$T_1^* = (p + p^* - q)T_1/\mathrm{tr}(\hat{\Omega}_{\hat{\beta}} \hat{V}_1), \tag{18}$$

where $\hat{V}_1$ is a consistent estimate of V_1. The statistic T_1^* is parallel to one proposed by Satorra and Bentler (1988, 1994) for the complete data case. Existing empirical experience with nonnormal complete data indicates that the distribution of T_1^* can be approximated by $\chi^2_{p+p^*-q}$ very well (Hu et al. 1992; Curran et al. 1996). More experience needs to be gained to evaluate the performance of this statistic with missing data sets.

In a similar way, the direct likelihood ratio statistic T_2 also has a distribution that approaches a linear combination of chi-square variates. Let

$$V_2 = H_2 - H_2\dot{\beta}(\dot{\beta}'H_2\dot{\beta})^{-1}\dot{\beta}'H_2 \tag{19}$$

and $\lambda_j^{(2)}, j = 1, \cdots, p + p^* - q$ be the nonzero eigenvalues of $\Omega_{\hat{\beta}}V_2$. Then we have the following lemma.

Lemma 3.2. Under regularity conditions stated earlier,

$$T_2 \xrightarrow{\mathcal{L}} \sum_{j=1}^{p+p^*-q} \lambda_j^{(2)} \chi_{j1}^2,$$

where χ_{j1}^2 are independent chi-square random variables each with degree of freedom 1.

Since $\Omega_{\hat{\beta}} = H_2^{-1}$ when data are normal, T_2 will approach $\chi_{p+p^*-q}^2$ in this case. Generally, we may use a rescaled version of this statistic

$$T_2^* = (p + p^* - q)T_2/\mathrm{tr}(\hat{\Omega}_{\hat{\beta}}\hat{V}_2) \tag{20}$$

in applications. As with T_1^*, more experience with T_2^* for different nonnormal data would be useful in guiding the application of such a statistic in data modeling practice.

Since neither T_1 nor T_2 or their rescaled versions asymptotically follow chi-square distributions, we need to have other statistics to better evaluate a structural model associated with the two types of ML-based procedures. This is given in the following lemma and its derivation is given in the appendix to this chapter.

Lemma 3.3. Let $\hat{e}(\theta) = \hat{\beta} - \beta(\theta)$ and

$$T(\theta) = n\hat{e}'(\theta)\{\hat{\Omega}_{\hat{\beta}}^{-1} - \hat{\Omega}_{\hat{\beta}}^{-1}\dot{\beta}(\theta)[\dot{\beta}'(\theta)\hat{\Omega}_{\hat{\beta}}^{-1}\dot{\beta}(\theta)]^{-1}\dot{\beta}'(\theta)\hat{\Omega}_{\hat{\beta}}^{-1}\}\hat{e}(\theta). \tag{21}$$

Then under the regularity conditions stated earlier, the distributions of both $T_4 = T(\tilde{\theta})$ and $T_5 = T(\hat{\theta})$ asymptotically follow $\chi_{p+p^*-q}^2$.

When data are not normal, the improper behavior of likelihood ratio type test statistics is well known. Within the class of elliptical distributions, Muirhead and Waternaux (1980) and Shapiro and Browne (1987) studied rescaled versions of likelihood ratio statistics. When data are not elliptical, the rescaled statistics may not behave properly even when sample size gets larger. However, the statistics T_4 and T_5 always approach

$\chi^2_{p+p^*-q}$ regardless of what the sampling distribution is. Actually, the result in Lemma 3.3 can also be applied to the estimator $\check{\theta}$, which will result in the statistic T_3 defined previously. So statistics T_4 and T_5 can be regarded as a generalization of the minimum chi-square test statistic.

We need to emphasize that the statistics T_3, T_4, and T_5 may not be computable unless $\hat{\Omega}_{\hat{\beta}}$ is invertable. A large sample size is usually needed in order to get a nonsingular $\hat{\Omega}_{\hat{\beta}}$, and even larger sample sizes may be needed for T_3, T_4, and T_5 to behave as nominal chi-squares. With complete data, Yuan and Bentler (1997a, 1998) proposed some asymptotically equivalent versions for these statistics that behave more like the nominal chi-squares for small to medium sample sizes. It would be interesting to see a generalization of these statistics to missing data cases in future studies.

4. STRUCTURAL EQUATION MODELING WITH A NUISANCE MEAN

In many aspects of multivariate analysis—e.g., factor analysis and principal components—the only interest is in covariance matrices, and the population means are best considered to be free nuisance parameters. In such a case, we denote the parameter in the covariance matrix as γ, so the unknown parameter vector now is $\theta' = (\gamma', \mu')'$. In principle, the theory developed in the last two sections also works for this case, but some of the results can be simplified. Our purpose is to give a simplified form of the different test statistics. We will use q to denote the number of unknown parameters in γ.

Since one still needs to estimate μ in the direct ML approach, this approach is the same even though there is no interest in μ. But in the second stage of the two-stage approach, $\tilde{\gamma}$ will be estimated by minimizing the Wishart likelihood function

$$F_{WL}(\gamma) = \mathrm{tr}(S_n\Sigma^{-1}(\gamma)) - \log|S_n\Sigma^{-1}(\gamma)| - p. \tag{22}$$

So the statistic T_1 is now $T_1 = nF_{WL}(\tilde{\gamma})$. Let $\hat{\Omega}_{\hat{\sigma}} = \hat{\Omega}^{(11)}_{\hat{\beta}}$ be the corresponding consistent estimate of the covariance of $\hat{\sigma}$, the statistic T_3 now becomes $T_3 = nQ(\check{\gamma})$, where $\check{\gamma}$ is obtained by minimizing

$$Q(\gamma) = (\hat{\sigma} - \sigma(\gamma))'\hat{\Omega}_{\hat{\sigma}}^{-1}(\hat{\sigma} - \sigma(\gamma)). \tag{23}$$

Similarly, the statistic T_4 and T_5 now become $T_4 = T(\tilde{\gamma})$ and $T_5 = T(\hat{\gamma})$, respectively, where

$$T(\gamma) = n\hat{e}'(\gamma)\{\hat{\Omega}_{\hat{\sigma}}^{-1} - \hat{\Omega}_{\hat{\sigma}}^{-1}\dot{\sigma}(\gamma)[\dot{\sigma}'(\gamma)\hat{\Omega}_{\hat{\sigma}}^{-1}\dot{\sigma}(\gamma)]^{-1}\dot{\sigma}'(\gamma)\Omega_{\hat{\sigma}}^{-1}\}\hat{e}(\gamma)$$

with $\hat{e}(\gamma) = \hat{\sigma} - \sigma(\gamma)$. Each of the statistics T_3, T_4, and T_5 has an asymptotic distribution $\chi^2_{p^*-q}$, regardless of what the underlying distribution of the data is. In order to get the counterparts of T_1^* and T_2^*, let

$$W_1 = \frac{1}{2} D_p'(\Sigma^{-1} \otimes \Sigma^{-1}) D_p, \qquad W_2 = \frac{1}{2n} \sum_{i=1}^{n} \kappa_i'(\Sigma_i^{-1} \otimes \Sigma_i^{-1}) \kappa_i,$$

$$U_1 = W_1 - W_1 \dot{\sigma} (\dot{\sigma}' W_1 \dot{\sigma})^{-1} \dot{\sigma}' W_1, \qquad \text{and}$$

$$U_2 = W_2 - W_2 \dot{\sigma} (\dot{\sigma}' W_2 \dot{\sigma})^{-1} \dot{\sigma}' W_2.$$

Now, both $T_1^* = (p^* - q) T_1 / \mathrm{tr}(\hat{\Omega}_{\hat{\sigma}} \hat{U}_1)$ and $T_2^* = (p^* - q) T_2 / \mathrm{tr}(\hat{\Omega}_{\hat{\sigma}} \hat{U}_2)$ asymptotically approach mixtures of chi-square distributions, with mean $p^* - q$.

When the sampling distribution is elliptical, all the estimators of γ are asymptotically independent with the estimators of μ. It is known that both T_1^* and T_2^* asymptotically follow $\chi^2_{p^*-q}$ with a complete sample from an elliptical distribution, but we are unable to generalize this property to the missing data case. Notice that the mean vector μ is explicitly estimated in the direct ML approach, and it is implicitly estimated through ML as defined in (4), even when (22) and (23) do not involve mean structures. Although μ is a vector of nuisance parameters, using methods other than ML to obtain its estimate will result in bias in the parameter estimates of γ. More detailed discussion of this aspect was given in Allison (1987).

5. IMPLEMENTATION AND NUMERICAL COMPARISON

We will use a data set from Mardia et al. (1979) to illustrate the steps in implementing the various procedures developed in this paper and to compare the effect of different missing data mechanisms and distribution conditions on these procedures. Mardia et al. (1979, table 1.2.1) give test scores of $n = 88$ students on five subjects. The five subjects are (1) Mechanics, (2) Vectors, (3) Algebra, (4) Analysis, and (5) Statistics. The first two subjects were tested with closed book exams and the last three were tested with open book exams. The original design of this data set is to study if different examination methods measure different abilities of the students. This hypothesis was confirmed by Tanaka et al. (1991) with a two-factor model

$$X = \mu + \Lambda f + e, \qquad \text{and} \qquad \Sigma(\theta) = \Lambda \Phi \Lambda' + \Psi, \tag{24}$$

where

$$\Lambda = \begin{pmatrix} \lambda_{11} & \lambda_{21} & 0 & 0 & 0 \\ 0 & 0 & \lambda_{32} & \lambda_{42} & \lambda_{52} \end{pmatrix}',$$

Φ is a correlation matrix for identification purposes and Ψ is a diagonal matrix. Since the interest is in the factor structure (24), μ is a nuisance parameter. So there are $q = 11$ covariance parameters with $\theta = (\lambda_{11}, \lambda_{21}, \lambda_{32}, \lambda_{42}, \lambda_{52}, \phi_{12}, \psi_{11}, \psi_{22}, \psi_{33}, \psi_{44}, \psi_{55})$. Since $p^* = 15$, the model degrees of freedom are 4. From this complete data set, we created the following missing data samples: (I) The last variable is removed 33 percent, realized by deleting the last variable of the first case in every three cases. So the missing data mechanism in this sample can be regarded as MCAR. (II) To create nonresponses that are missing only at random, we remove the last two variables if the summation of the first three scores is less than 113, which results in 17 (about 20 percent) cases with missing variables. The sample in (II) is a simulation of the situation in which a student may quit the later ones if he or she did not perform well on the first three tests. The advantage here is that we have the parameter estimates for the complete data, which will facilitate the comparison on biases and test statistics of different methods. For all the missing data sets in this section, we use divisor $n - 1$ instead of n when computing the covariance estimates in the M-step of the EM-algorithm, as recommended by Beale and Little (1975).

Considering that none of the current version of the standard software contains the procedures developed here,[1] we would like to outline the necessary steps for implementing these procedures. This will be given next before numerical comparison on these procedures with several missing data sets.

5.1. *Implementation*

For any of the procedures, the MLE S_n of Σ is needed, while $\bar{X}_n$ is also needed if a mean structure is also of interest. They can be obtained by setting the structured model as the saturated model in a program with the direct ML missing data procedure (e.g., LINCS, Mplus, Mx, or AMOS), or one can use the EM-algorithm based on normal assumptions (e.g., Little

[1] We would like to note that EQS 6 (Bentler, in press) will have the new missing data procedures in its future versions.

and Rubin 1987, sec. 8.2). A consistent $\hat{\Omega}_{\hat{\beta}}$ based on (5b) also has to be computed. For the two-stage approach, the S_n and $\overline{X}_n$ can be used as input data in any of the standard programs to get parameter estimates $\tilde{\theta}$ and $T_1 = nF_{ML}(\tilde{\theta})$. For example, for the missing data sample (I) the factor loadings in $\tilde{\theta}$ are given later in Table 2(a) and $T_1 = 2.69$. Standard errors based on the sandwich-type covariance matrix should be evaluated according to (6b). If the rescaled statistic is used for inference, $\hat{V}_1$ based on (17) should be evaluated, which, together with $\hat{\Omega}_{\hat{\beta}}$, will lead to the statistic T_1^* in (18). For example, $T_1^* = 2.04$ for the missing data sample (I). If the asymptotic chi-square statistic T_4 is wanted, one does not need to compute $\hat{V}_1$, instead, $\dot{\beta}(\tilde{\theta})$ should be evaluated and T_4 be obtained according to (21). For the missing data sample (I), $T_4 = 1.74$.

For the direct ML method $\hat{\theta}$ and T_2 are available through standard software (e.g., AMOS, LINCS, Mplus, or Mx). The sandwich-type covariance matrix of $\hat{\theta}$ can be obtained by (9). If the rescaled statistic in (19) is sought, one has to compute $\hat{V}_2$, which together with $\hat{\Omega}_{\hat{\beta}}$ will facilitate the computation of T_2^*. As in the two-stage ML method, one has to compute $\dot{\beta}(\hat{\theta})$ instead of $\hat{V}_2$ if the asymptotic chi-square statistic T_5 is wanted. For the missing data sample (I), $T_2 = 1.96$, $T_2^* = 1.92$, and $T_5 = 1.72$, and the corresponding factor loadings in $\hat{\theta}$ are given in Table 2(a).

The minimum chi-square approach is the easiest one to implement with a programing language such as SAS IML. With $\hat{\beta}$ and $\hat{\Omega}_{\hat{\beta}}$, $\check{\theta}$ is just the generalized least squares estimate with objective function (7). For the missing data sample (I), $T_3 = 1.72$ and the factor loadings corresponding to $\check{\theta}$ are given later in Table 2(a).

5.2. *Comparison*

Previous analysis for the complete data set by Tanaka et al. (1991) indicates that model (24) fits the data very well by either the likelihood method or the minimum chi-square method. Actually, this data set basically follows a multivariate normal distribution with standardized multivariate skewness and kurtosis (Mardia et al. 1979, p. 148) being 3.24 and .057 respectively. Referring these two numbers to distributions χ^2_{35} and $N(0,1)$ respectively, both are far from significant. So we would expect that normal theory methods also work well on missing data samples (I) and (II). Each missing data method was used on each of the two missing data samples. Test statistics corresponding to different methods are given in the upper panel of Table 1. For comparison purpose, the test

TABLE 1
Various Test Statistics with Normal and Nonnormal Data Under Different Missing Mechanisms

Samples	T_1	T_2	T_1^*	T_2^*	T_3	T_4	T_5
Complete Data	2.07	2.07	2.13	2.13	2.01	1.98	1.98
(I)	2.69	1.96	2.04	1.92	1.72	1.74	1.72
(II)	1.10	.81	.12	.11	.85	.85	.85
Complete Data	42.23	42.23	7.80	7.80	4.14	4.11	4.11
(III)	40.68	35.66	5.64	6.42	3.79	3.82	3.97
(IV)	50.34	37.71	7.34	7.04	2.88	3.04	3.18

statistics for complete data are also included. None of the statistics is statistically significant when referred to χ_4^2, indicating that the proposed model structure cannot be rejected. This suggests that all the statistics give reliable inference when data are approximately normal and the mechanism of nonresponse is ignorable.

In order to further compare the different test statistics, two partially artificial data sets were created. Let r_i, $i = 1, \cdots, n$ be a sample from the population $r = (\chi_3^2 - 3)/\sqrt{6}$, and

$$Y_i = r_i(X_i - E(X)). \tag{25}$$

As $E(r^2) = 1$, and r_i is independent with X_i, the covariance structure of Y_i is

$$\begin{aligned}\operatorname{cov}(Y) &= E(YY') - E(Y)E(Y') = E(r^2)E[(X_i - E(X))(X_i - E(X))'] \\ &= \operatorname{cov}(X)\end{aligned}$$

which is the same as that of the original sample X_i. Applying the transformation (25) to each of the missing data samples (I) and (II) respectively, we obtain two more missing data samples (III) and (IV).[2] Notice that the nonresponses in sample (III) are still completely at random; and the nonresponses in (IV) are still at random. However, since some of the r_i will be

[2] To permit anyone to further study samples (III) and (IV) in this example, we would like to note that the sampling from χ_3^2 was created by 2 × rangam(seed, 1.5) in SAS IML with initial seed = 1234567, where the first 50 random variables were discarded in order to minimize the effects of the arbitrariness of the initial seed.

negative, the nonresponses in sample (IV) will not represent the scores whose first three variables are the lowest. Also, since $E(r^4) = 7$, we can not expect that the sample Y_i in either (III) or (IV) is normally distributed any more. Actually, the standardized Mardia's multivariate kurtosis for the transformed complete sample is 29.60, which is highly significant when referred to $N(0,1)$. So, theoretically, the normal assumption based methods cannot give reliable inference. Each of the methods was applied to samples (III) and (IV). Results are presented in the lower panel of Table 1. When referring to χ_4^2, both T_1 and T_2 are significant in either sample (III) or (IV), indicating incorrect rejection of a reasonable model. However, as expected, all the other statistics behave very stably and none is significant, yielding the correct conclusion that the proposed model is acceptable. Actually, all the statistics for the two missing data samples behave similarly to their counterparts for the complete data.

Even though we do not know the true population parameters in this example, we know the estimates for the complete sample and also the missing data mechanism for each missing data set. The difference between estimates based on the complete sample and the corresponding population value is because of sampling error. The differences between parameter estimates based on the complete sample and those based on missing data samples are because of loss of efficiency and possible bias due to missing data. Comparing the parameter estimates by different missing data methods with those by complete data will give us valuable information on possible biases. For this, we list only the factor loading estimates in Table 2 to save space, where θ^0 and θ^1 are respectively the MLE and minimum chi-square estimates based on the complete data. Notice that when applying the two-stage and the direct ML methods to the complete data we have $\theta^0 = \tilde{\theta} = \hat{\theta}$, which are generally different from $\theta^1 = \check{\theta}$, because they are obtained from fitting different objective functions. Based on distributions and missing data mechanisms of samples (I) to (IV), we know that theoretically there are no biases for estimates in Table 2 (a) and estimates for sample (III) in Table 2 (b). Discrepancies among different estimates are due to sampling errors or finite sample effects. For easy comparison, the largest discrepancy D^* between parameter estimates for each missing data sample and those for the complete sample are given; an asterisk is used to indicate the specific parameter estimate. Contrasting the D^*s for sample (II) with those for sample (I) in Table 2 (a), we may notice a little bit larger D^* for MAR data than those for MCAR data due to a finite sample effect. This phenomenon will also be observed in the next section. Comparing the

TABLE 2
Estimates of Parameters by Different Methods

(a) Normal Data

	Complete Data		(I)			(II)		
θ	θ^0	θ^1	$\tilde{\theta}$	$\hat{\theta}$	$\check{\theta}$	$\tilde{\theta}$	$\hat{\theta}$	$\check{\theta}$
λ_{11}	12.18	12.53	12.15	12.21	12.43	12.25	12.26	12.31
λ_{21}	10.32	10.25	10.47	10.42	10.34	10.39	10.37	10.32
λ_{32}	9.78	9.71	9.69	9.80	9.74	10.14	10.13	10.01
λ_{42}	11.42	11.47	11.64	11.48	11.40	*13.22	*13.42	*13.88
λ_{52}	12.45	12.64	*11.79	*11.60	*11.38	13.50	13.49	13.90
D^*			$-.66$	$-.85$	-1.26	1.80	2.00	2.14

(b) Nonnormal Data

	Complete Data		(III)			(IV)		
θ	θ^0	θ^1	$\tilde{\theta}$	$\hat{\theta}$	$\check{\theta}$	$\tilde{\theta}$	$\hat{\theta}$	$\check{\theta}$
λ_{11}	12.28	13.55	12.58	12.91	13.50	12.79	12.98	14.42
λ_{21}	10.78	9.61	10.64	10.37	10.00	10.47	10.32	9.30
λ_{32}	9.15	8.35	9.52	10.08	8.54	9.90	10.15	9.18
λ_{42}	14.51	12.15	14.61	14.36	12.12	*16.55	*18.09	*13.52
λ_{52}	12.41	12.78	*10.10	*8.74	*10.76	12.53	11.49	13.18
D^*			-2.13	-3.67	-2.02	2.04	3.58	1.37

Note: $\theta^0 = \tilde{\theta} = \hat{\theta}$ is the MLE based on the complete data; $\theta^1 = \check{\theta}$ is the minimum chi-square estimate based on the complete data; D^* is the largest discrepancy between parameter estimates for each missing data sample and those for the complete sample.

D^* for sample (IV) with those for sample (III), one cannot observe greater discrepancies for MAR data than those for MCAR data. From the above comparison we conclude that there is no noticeable bias in MLE based on the wrongly specified distribution for the missing data sample (IV) whose missing data mechanism is MAR.

6. BIAS IN NORMAL THEORY MLE FOR NONNORMAL DATA

Through Monte Carlo, we will continue to study the bias associated with mean and covariance parameters in normal theory-based likelihood estimates when data are not normal. For simplicity and without loss of generality, we choose $p = 2$, $\mu_1 = \mu_2 = 0$, $\sigma_{11} = \sigma_{22} = 1$, and $\sigma_{12} = \sigma_{21} = \rho$

with $\rho = .5$ and .9, respectively. Let x_1 be always observed with sample size N. Only N_1 cases are observed for x_2 according to a missing data mechanism. The biases in parameter estimates of μ and Σ can be obtained by comparing the estimates with the population values. Based on Anderson (1957), there exist analytical solutions for the parameter estimates if a normal distribution assumption is assumed, thus no iteration is necessary for this simple design. Notice that if there is no bias in the parameter estimates $\hat{\beta}$ defined in (4), then there will be no bias in parameter estimates of θ_0. So all possible biases should be reflected in $\hat{\beta}$. Without loss of generality, our interest is in the possible bias in $\hat{\beta}$. Due to a saturated model, parameter estimates for the three estimation methods are identical.

Let e_1 and e_2 be independent standardized random variables, the joint distribution of (x_1, x_2) is generated through

$$x_1 = e_1, \qquad x_2 = \sqrt{1-\rho^2}e_2 + \rho e_1.$$

For different distribution conditions, we choose e_1 as $N(0,1)$ and $e_1 = (\chi_5^2 - 5)/\sqrt{10}$ respectively for symmetric and skewed distributions. Since estimates of μ_1 and σ_{11} are based on complete samples, we would like to create more distributional conditions for x_2 to study the possible biases in estimates of μ_2, σ_{22}, and ρ. For each of the e_1, e_2 is generated respectively from five distributions:

$N(0,1)$;
the standardized t-distribution $t_5/\sqrt{5/3}$;
the standardized uniform distribution $(U(0,1) - 1/2)/\sqrt{1/12}$;
the standardized chi-square distribution $(\chi_3^3 - 3)/\sqrt{6}$;
and the standardized lognormal distribution $(\exp(z) - \exp(1/2))/\sqrt{e(e-1)}$, where $z \sim N(0,1)$.

This design creates a variety of skewnesses and kurtoses in the variable x_2 as given in Table 3. A very severe departure from normality occurs in the case where e_1 follows $N(0,1)$ and e_2 follows lognormal$(0,1)$; the skewness and kurtosis of x_2 are respectively 4.02 and 62.40 with $\rho = .5$. Similarly, when e_1 is χ_5^2 with lognormal e_2, skewness and kurtosis are equally large.

All three missing data mechanisms—MCAR, MAR and NMAR—are included. For comparison purposes, we also include a complete sample for each of the distribution conditions. Since we are interested in possible large sample biases of the MLEs with an incorrect distributional assump-

TABLE 3
Skewness and Kurtosis of x_2

	$\rho = .5$		$\rho = .9$	
Data e_1 & e_2	Skewness	Kurtosis	Skewness	Kurtosis
N(0,1) & N(0,1)	0.00	0.00	0.00	0.00
N(0,1) & t_5	0.00	3.37	0.00	0.22
N(0,1) & U(0,1)	0.00	-0.67	0.00	-0.04
N(0,1) & χ_3^2	1.06	2.25	0.14	0.14
N(0,1) & LN(0,1)	4.02	62.40	0.51	4.00
χ_5^2 & N(0,1)	0.16	0.15	0.92	1.57
χ_5^2 & t_5	0.16	3.52	0.92	1.79
χ_5^2 & U(0,1)	0.16	-0.52	0.92	1.53
χ_5^2 & χ_3^2	1.22	2.40	1.06	1.72
χ_5^2 & LN(0,1)	4.18	62.55	1.43	5.58

tion, we choose $N = 1000$ for the complete data case and $N = 1000$ and 2000 for each of the missing data cases. The MCAR mechanism is created by removing x_2 in every even numbered case. When $x_1 \sim N(0,1)$ the MAR mechanism is realized by removing the corresponding x_2 when $x_1 < 0$; and for x_1 following the standardized χ_5^2, the MAR mechanism is realized by removing x_2 when x_1 is greater than its population median. So for both MCAR and MAR, the missing percentage is about 50 percent. The NMAR mechanism depends on the actual observation of x_2 whose median is not straightforward to obtain. This mechanism was created by removing x_2 if it is less than 0 when $x_1 \sim N(0,1)$ or if it is greater than the median of x_1 when x_1 follows the standardized χ_5^2. For each distribution condition, the average number of observations on x_2 for each of the NMAR samples is reported in Tables 4 and 5.

For each of the designed conditions, 500 replications are used. Since estimates of μ_1 and σ_{11} are just the sample mean and sample variance of x_1, which is completely observed, the asymptotic bias can only be observed on estimates of $\theta = (\mu_2, \sigma_{22}, \rho)'$. Let $\theta^{(i)}$ be the estimate of θ in the ith replication and $\bar{\theta} = \sum_{i=1}^{500} \theta^{(i)}/500$, then typical bias is calculated as $\bar{\theta} - \theta_0$. Since our interest is in contrasting the biases of MLEs from a misspecified distribution for MAR data with MLEs that are known to have zero systematic bias, the biases in Tables 4 and 5 are calculated according to

$$\text{Bias} = (\bar{\theta} - \theta_0)'(\bar{\theta} - \theta_0). \tag{26}$$

TABLE 4
Variance and Bias (ρ = .5)

(a) ($N = 1000$)

	Complete		MCAR		MAR		NMAR		
Data e_1 & e_2	Var $\times 10^3$	Bias $\times 10^6$	Var $\times 10^3$	Bias $\times 10^6$	Var $\times 10^3$	Bias $\times 10^6$	N_1	Var $\times 10^3$	Bias $\times 10^6$
N(0,1) & N(0,1)	4.22	1.59	7.34	1.55	17.28	26.16	500	2.36	9.84
N(0,1) & t_5	6.87	4.01	12.35	2.83	21.67	15.41	500	5.77	8.67
N(0,1) & U(0,1)	3.73	3.86	6.51	8.21	16.32	42.85	500	1.69	11.01
N(0,1) & χ_3^2	6.36	5.79	12.17	1.72	19.07	6.42	434	10.49	9.97
N(0,1) & LN(0,1)	83.17	3.76	88.86	101.40	218.79	85.97	414	420.73	9.55
χ_5^2 & N(0,1)	4.78	1.57	7.71	21.07	41.29	75.51	427	2.83	13.01
χ_5^2 & t_5	7.78	2.58	13.68	76.15	46.11	52.00	421	10.76	12.15
χ_5^2 & U(0,1)	4.54	10.41	7.36	11.77	40.62	82.61	435	1.66	13.49
χ_5^2 & χ_3^2	7.18	38.22	12.12	56.38	42.45	110.89	492	1.20	11.09
χ_5^2 & LN(0,1)	70.06	6.47	76.99	436.57	241.07	409.17	507	1.20	9.06

(b) ($N = 2000$)

	MCAR		MAR		NMAR		
Data e_1 & e_2	Var $\times 10^3$	Bias $\times 10^6$	Var $\times 10^3$	Bias $\times 10^6$	N_1	Var $\times 10^3$	Bias $\times 10^6$
N(0,1) & N(0,1)	3.78	15.03	7.49	.90	1000	1.27	9.83
N(0,1) & t_5	6.50	9.47	9.68	4.34	1001	2.99	8.65
N(0,1) & U(0,1)	3.22	2.28	7.34	5.32	999	.87	10.99
N(0,1) & χ_3^2	6.08	4.76	9.14	2.45	868	5.27	9.97
N(0,1) & LN(0,1)	56.92	68.23	87.00	101.23	828	159.10	9.47
χ_5^2 & N(0,1)	3.82	9.87	21.99	69.66	854	1.42	13.02
χ_5^2 & t_5	6.85	17.56	23.07	20.76	842	4.84	12.12
χ_5^2 & U(0,1)	3.70	9.17	22.11	27.34	868	.87	13.49
χ_5^2 & χ_3^2	6.71	10.31	20.94	71.95	985	.63	11.10
χ_5^2 & LN(0,1)	57.51	484.27	88.32	9.00	1015	.60	9.05

Note: $N_1 = N$ for complete data; $N_1 = N/2$ for MCAR data; $N_1 \approx N/2$ for MAR data.

TABLE 5
Variance and Bias ($\rho = .9$)

(a) ($N = 1000$)

	Complete		MCAR		MAR		NMAR		
Data e_1 & e_2	Var $\times 10^3$	Bias $\times 10^6$	Var $\times 10^3$	Bias $\times 10^6$	Var $\times 10^3$	Bias $\times 10^6$	N_1	Var $\times 10^3$	Bias $\times 10^6$
N(0,1) & N(0,1)	4.70	10.26	5.87	7.91	10.18	8.86	500	5.25	3.10
N(0,1) & t_5	4.83	9.16	5.99	7.71	10.38	7.96	500	6.03	2.84
N(0,1) & U(0,1)	4.86	7.64	5.85	5.00	10.46	7.92	500	4.79	3.24
N(0,1) & χ_3^2	4.65	9.46	5.72	3.81	8.83	.92	493	6.50	3.67
N(0,1) & LN(0,1)	9.59	21.75	10.63	54.42	21.83	17.36	488	25.04	2.70
χ_5^2 & N(0,1)	7.76	5.78	8.95	13.23	24.61	8.15	473	5.28	7.33
χ_5^2 & t_5	8.62	2.13	9.80	4.71	24.54	4.56	475	8.30	7.07
χ_5^2 & U(0,1)	8.83	31.16	10.04	28.29	25.06	90.26	470	4.39	7.70
χ_5^2 & χ_3^2	8.53	63.50	9.62	64.47	23.57	44.04	485	5.47	4.03
χ_5^2 & LN(0,1)	11.84	13.27	13.31	81.72	37.32	19.04	495	6.32	2.06

(b) ($N = 2000$)

	MCAR		MAR		NMAR		
Data e_1 & e_2	Var $\times 10^3$	Bias $\times 10^6$	Var $\times 10^3$	Bias $\times 10^6$	N_1	Var $\times 10^3$	Bias $\times 10^6$
N(0,1) & N(0,1)	3.10	9.90	4.62	3.63	999	2.61	3.10
N(0,1) & t_5	3.21	6.47	4.74	3.09	999	3.00	2.84
N(0,1) & U(0,1)	2.93	2.04	4.75	3.20	999	2.31	3.24
N(0,1) & χ_3^2	2.92	.53	4.29	3.31	985	3.08	3.71
N(0,1) & LN(0,1)	6.25	22.00	9.55	4.84	976	11.34	2.66
χ_5^2 & N(0,1)	4.53	2.37	13.31	9.21	945	2.61	7.32
χ_5^2 & t_5	4.01	.53	11.45	.92	950	3.49	7.09
χ_5^2 & U(0,1)	4.78	24.13	12.96	45.28	940	2.09	7.69
χ_5^2 & χ_3^2	4.98	23.63	11.55	6.39	970	2.84	4.00
χ_5^2 & LN(0,1)	7.94	59.11	17.92	2.94	991	3.21	2.04

Note: $N_1 = N$ for complete data; $N_1 = N/2$ for MCAR data; $N_1 \approx N/2$ for MAR data.

It is obvious that any large discrepancy between $\bar{\theta}$ and θ_0 will be reflected in (26). As sampling variation influences the accuracy of an estimate, we also calculated the sample variance of the estimates among the 500 replications according to

$$\text{Var} = \frac{1}{500}\sum_{i=1}^{500}(\theta^{(i)} - \bar{\theta})'(\theta^{(i)} - \bar{\theta}). \tag{27}$$

Table 4 gives the bias and variance corresponding to each condition for $\rho = .5$. Since theoretically there is no asymptotic bias with complete data and MCAR data, the biases reflected by the second and fourth columns of Table 4 (a) and the second column of Table 4 (b) reflect only a finite sample effect. Because of a correct distribution assumption, the corresponding biases for the condition e_1 & e_2 being $N(0,1)$ and the missing data mechanism being MAR in Table 4 also reflect the finite sample effect. We may notice that this effect can be quite large even for complete samples. We will mainly compare biases under MAR and MCAR since both are based on approximately equal sample sizes. From Table 4 (a) we may notice that with the normal sample, the bias under MAR is about 17 times that under MCAR, even though all this is due to a finite sample effect. Similarly, for the normal sample, the variance under MAR is also several times that under MCAR. This indicates that estimates based on MAR data may not be as accurate as estimates based on complete data and MCAR data even though the distributional assumption is correct. A similar proportion of inaccuracy can be observed when the distributional assumption is incorrect. For example, for $N(0,1)$ & t_5 and $N(0,1)$ & $U(0,1)$, biases under MAR are about five times those of MCAR; for conditions $N(0,1)$ & χ_3^2 and χ_5^2 & $N(0,1)$, the biases under MAR are about four times those corresponding to MCAR. For nonnormal data, the largest proportion occurs with χ_5^2 & $U(0,1)$ where the bias corresponding to MAR is about seven times that of MCAR. It is important to note, however, that this is still smaller than that for the normal data. Actually, for conditions $N(0,1)$ & $LN(0,1)$, χ_5^2 & t_5, and χ_5^2 & $LN(0,1)$, the biases under MAR are smaller than those under MCAR, even though these distributions are quite different from normal.

In the last two columns of Table 4 (a) are the variances and biases for data that are NMAR. Regardless of the actual observed sample sizes and the underlying distribution of the data, the bias in the estimates are about 10^4 to 10^5 times of those when data are MCAR or MAR. Even though the maximum-likelihood procedure may perform better than an ad hoc procedure such as listwise deletion (e.g., Schafer 1997, sec. 2.5.2), with an average of about ten times the standard error, the magnitude of biases may render inference on parameters meaningless.

Turning to Table 4 (b), except for a particular phenomenon that for normal data the bias under MAR is much smaller than that under MCAR, which may be due to a finite sample effect, the comparison of biases corresponding to MCAR, MAR, and NMAR is similar to those in Table 4 (a). The results for $\rho = .9$ corresponding to different conditions are given in Table 5, where similar comparisons can be found as with $\rho = .5$ in Table 4. We may also notice from both Tables 4 and 5 with data being MAR and MCAR that larger biases generally go with larger variances. This may indicate that differences between different finite sample estimators may only reflect different efficiencies and not especially biases.

We may conclude from the above comparison that, if there is any large sample bias with normal theory MLE for some commonly encountered nonnormal distributions in practice, the bias is not large enough to worry about. This is rather fortunate since in practice almost any distributional assumption with high dimensional data is likely to be incorrect.

7. DISCUSSION

Since not many multivariate distributions are available to describe the nonnormality of practical data, the normal distribution assumption is commonly used with the analysis of nonnormal missing data in mean and covariance structure analysis. Unfortunately, this generally leads to inaccurate inferences about model structure. We propose several new procedures that do not need any specific distribution assumptions. Statistical development and numerical examples illustrate the merit of procedures that are newly developed over those that are based on a normality assumption. By dropping this assumption, one has to assume the missing data mechanism is MCAR according to Laird (1988) and Rotnitzky and Wypij (1994). We may need to reemphasize that using a normal distribution and a MAR missing data mechanism leads to the same parameter estimates as using an unknown distribution and a MCAR missing data mechanism. Fortunately, our simulation results and examples do not indicate noticeable biases for nonnormal data that are MAR. Taking into consideration the analytical and empirical results, we make the following recommendations: Use the minimum chi-square method for inference when sample size is large; use the direct or the two-stage methods with the rescaled statistics for model inference and sandwich-type covariance matrices for standard errors when sample size is medium. The small sample problem is still open even for complete normal data (e.g., Bentler and Yuan 1999).

When facing a missing data set with nonnormal distributions, we can consider another possible approach: model the data with a multivariate *t*-distribution, as developed in Little (1988). However, according to Gourieroux et al. (1984), even when data are MCAR, imputation based on such a distribution may not yield consistent estimates of the population covariances unless the data truly follow the multivariate *t*-distribution. The practical aspect of this inconsistency may not be so serious, as was observed in Lange et al. (1989).

In our development of methods for nonnormal missing data, the focus has been on extending and correcting maximum-likelihood–based methods. Of course, ad hoc methods have been used in data analysis for decades and provide another option in handling incomplete data. These include mean imputation, listwise deletion, pairwise computations, hot deck imputation as well as more recently developed methods such as similar response pattern imputation. In these approaches, a modified data set or a covariance matrix is created that subsequently can be analyzed by any existing standard method designed for complete data. An advantage of these approaches is that they are relatively practical to implement; indeed such methods for dealing with incomplete data can be found in most well-known statistical program packages. Furthermore, nonnormality can be routinely handled when an imputed data matrix is analyzed with a distribution-free method. These methods are all appropriate when the amount of missing data is extremely small. However, there exist several drawbacks of these nonprincipled methods. For example, listwise deletion can render a longitudinal study with few cases left, resulting in grossly inefficient estimates (e.g., Brown 1994). When the missing data mechanism is MAR, existing simulation results indicate that listwise deletion causes parameter estimates to be biased even for normal data (Little and Rubin 1987; Schafer 1997). Similarly, a recent study with a confirmatory factor analysis model by Marsh (1998) indicates that the pairwise computation method leads to substantially biased test statistics, depending on the percent of missing data and its interaction with sample size. On the other hand, the simulation results in the last section imply that there is no noticeable bias even for MLE based on a wrongly specified distribution when the missing data are MAR.

In addition to likelihood-based methods, the multiple imputation technique developed by Rubin (1987) has showed its potential in handling incomplete data problems. In this method each missing value in a data set is replaced by a vector of m simulated values, thus creating m complete data sets that agree with the original incomplete data set on the observed values. Then each of the m imputed data sets is analyzed using a standard complete data routine and the result of the complete data analyses are combined to make inference. Because multiple imputation can remove the difficulty of modeling missing data mechanisms and the computational complications of incomplete data, a variety of multiple imputation techniques have been developed recently (e.g., see Meng 1994; Rubin 1996; Schafer 1997). Extending these techniques to mean and covariance structure analysis would be highly valuable. Due to the typical nonnormality of social science data, however, any such an extension still remains a challenge. For example, when complete data exhibit heterogeneous marginal skewness and kurtosis, it is nearly impossible to find a correct model to generate multiple imputations that conform with the randomness of the missing values. Suppose one uses a model based on the normal distribution to generate the imputed values, then the consequences of replacing the missing values by normal variables on the combined results is not clear (e.g., Rubin 1996; Schafer 1997). As discussed in the introduction, nonnormality is a problem not just with missing data. Even with a complete nonnormal data set, rescaled and generalized least squares type of statistics or recently developed bootstrap techniques (e.g., Bollen and Stine 1993; Yung and Bentler 1996) may have to be used in order to obtain reliable model evaluation.

In spite of our development, additional technical problems for structural equation modeling with missing data remain to be studied in future research. For example, the asymptotic efficiency characterized in Section 3 may not hold for all finite sample sizes. Also, even though the statistics T_3, T_4, and T_5 are asymptotically distribution free, their small sample behavior may not be well described by a chi-square distribution. Furthermore, $\hat{\Omega}_{\hat{\beta}}$ may not be of full rank for smaller sample sizes with a large p. We may have to turn to the statistics T_1^* or T_2^* in such a case, though these are generally not distributed as chi-square even for large sample sizes. More research is necessary for these small sample inference issues to be fully addressed. Another problem is related to the missing data mechanism. Even though the ML-based procedure has little bias when the missing data mechanism is ignorable, MAR is still a strong assumption in practice. It is necessary to develop procedures for dealing with missing data that are NMAR for safer inferences with mean and covariance structure analysis.

Finally, our experience with missing data is limited. The procedures developed in this paper are subject to more empirical verification and modification.

APPENDIX

Proof of Lemma 3.1: Using a Taylor expansion on $F(\tilde{\theta})$, we have

$$F(\tilde{\theta}) = F(\theta_0) + \frac{\partial F(\theta_0)}{\partial \theta_0'}(\tilde{\theta} - \theta_0) + \frac{1}{2}(\tilde{\theta} - \theta_0)' \frac{\partial^2 F(\bar{\theta}_n)}{\partial \bar{\theta}_n \partial \bar{\theta}_n'}(\tilde{\theta} - \theta_0), \tag{A1}$$

where $\bar{\theta}_n$ lies between θ_0 and $\tilde{\theta}$. Using an equation from Muirhead (1982, eq. 15, p. 363),

$$-\log|S_n\Sigma^{-1}| = \text{tr}(I - S_n\Sigma^{-1}) + \frac{1}{2}\text{tr}(I - S_n\Sigma^{-1})^2 + O_p\left(\frac{1}{n^{3/2}}\right).$$

So

$$\begin{aligned} F(\theta_0) &= \frac{1}{2}\text{tr}(I - S_n\Sigma^{-1})^2 + (\overline{X}_n - \mu)'\Sigma^{-1}(\overline{X}_n - \mu) + O_p\left(\frac{1}{n^{3/2}}\right) \\ &= (\hat{\beta} - \beta_0)'H_1(\hat{\beta} - \beta_0) + O_p\left(\frac{1}{n^{3/2}}\right). \end{aligned} \tag{A2}$$

It follows from direct calculations that

$$\sqrt{n}\frac{\partial F(\theta_0)}{\partial \theta_0} = -2\dot{\beta}'H_1\sqrt{n}(\hat{\beta} - \beta_0) + o_p(1), \tag{A3}$$

$$\frac{\partial^2 F(\bar{\theta}_n)}{\partial \bar{\theta}_n \partial \bar{\theta}_n'} \xrightarrow{p} 2\dot{\beta}'H_1\dot{\beta}, \tag{A4}$$

and

$$\sqrt{n}(\tilde{\theta} - \theta_0) = (\dot{\beta}'H_1\dot{\beta})^{-1}\dot{\beta}'H_1\sqrt{n}(\hat{\beta} - \beta_0) + o_p(1). \tag{A5}$$

By putting (A2) to (A5) into (A1) we obtain

$$T_1 = \sqrt{n}(\hat{\beta} - \beta_0)'\{H_1 - H_1\dot{\beta}(\dot{\beta}'H_1\dot{\beta})^{-1}\dot{\beta}'H_1\}\sqrt{n}(\hat{\beta} - \beta_0) + o_p(1). \tag{A6}$$

Lemma 3.1 follows from (A6).

Proof of Lemma 3.2: Using Taylor expansions on $l(\hat{\beta})$ and $l(\hat{\theta})$ at β_0 and θ_0 respectively, we obtain

$$l(\hat{\beta}) = l(\beta_0) + \frac{\partial l(\beta_0)}{\partial \beta_0'}(\hat{\beta} - \beta_0) - \frac{n}{2}(\hat{\beta} - \beta_0)'A_\beta(\hat{\beta} - \beta_0) + o_p(1/n) \tag{A7}$$

and

$$l(\hat{\theta}) = l(\theta_0) + \frac{\partial l(\theta_0)}{\partial \theta_0'}(\hat{\theta} - \theta_0) - \frac{n}{2}(\hat{\theta} - \theta_0)'A_\theta(\hat{\theta} - \theta_0) + o_p(1/n), \tag{A8}$$

where we have used (5) and (10), respectively. Similarly, using a Taylor expansion on $\partial l(\hat{\beta})/\partial\hat{\beta} = 0$ and $\partial l(\hat{\theta})/\partial\hat{\theta} = 0$, we have

$$\sqrt{n}(\hat{\beta} - \beta_0) = A_\beta^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{\partial l_i(\beta_0)}{\partial \beta_0} + o_p(1) \tag{A9}$$

and

$$\sqrt{n}(\hat{\theta} - \theta_0) = A_\theta^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{\partial l_i(\theta_0)}{\partial \theta_0} + o_p(1). \tag{A10}$$

Using (12) on the right-hand side of (A10), it follows from (A9) and (A10) that

$$\begin{aligned}\sqrt{n}(\hat{\theta} - \theta_0) &= A_\theta^{-1}\dot{\beta}'\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{\partial l_i(\beta_0)}{\partial \beta_0} + o_p(1)\\ &= A_\theta^{-1}\dot{\beta}'A_\beta\sqrt{n}(\hat{\beta} - \beta_0) + o_p(1).\end{aligned} \tag{A11}$$

From (A9) to (A11), we also get the following relations

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{\partial l_i(\beta_0)}{\partial\beta_0} = A_\beta\sqrt{n}(\hat\beta-\beta_0)+o_p(1); \tag{A12}$$

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{\partial l_i(\theta_0)}{\partial\theta_0} = \dot\beta' A_\beta\sqrt{n}(\hat\beta-\beta_0)+o_p(1).$$

Since $A_\theta = \dot\beta' A_\beta\dot\beta$, $l(\beta_0) = l(\theta_0)$ and $T_2 = 2(l(\hat\beta) - l(\hat\theta))$, it follows from (A7) to (A12) that

$$T_2 = n(\hat\beta-\beta_0)'\{A_\beta - A_\beta\dot\beta(\dot\beta' A_\beta\dot\beta)^{-1}\dot\beta' A_\beta\}(\hat\beta-\beta_0)+o_p(1)$$

and the lemma follows by recalling that $H_2 = A_\beta$.

Proof of Lemma 3.3: Since the proofs for T_4 and T_5 are the same, we outline only the proof for T_5. First notice that $\dot\beta$ is a $(p+p^*)\times q$ matrix for which $\hat{\dot\beta} = \dot\beta(\hat\theta)$ is a consistent estimate. Let $\hat{\dot\beta}_c$ be a full column rank $(p+p^*)\times(p+p^*-q)$ matrix whose columns are orthogonal to those of $\hat{\dot\beta}$, then $\hat{\dot\beta}_c \xrightarrow{p} \dot\beta_c$. It follows from (A11) that

$$\begin{aligned}\sqrt{n}(\hat\beta-\beta(\hat\theta)) &= \sqrt{n}\{[\hat\beta-\beta_0]-[\beta(\hat\theta)-\beta(\theta_0)]\}\\ &= \sqrt{n}\{[\hat\beta-\beta_0]-\dot\beta(\hat\theta-\theta_0)\}+o_p(1)\\ &= \{I-\dot\beta(\dot\beta' A_\beta\dot\beta)^{-1}\dot\beta' A_\beta\}\sqrt{n}(\hat\beta-\beta_0)+o_p(1),\end{aligned}$$

and

$$\begin{aligned}T_5 &= n\hat e'(\hat\theta)\hat{\dot\beta}_c\{\hat{\dot\beta}_c'\hat\Omega_{\hat\beta}\hat{\dot\beta}_c\}^{-1}\hat{\dot\beta}_c'\hat e'(\hat\theta)\\ &= \sqrt{n}[\dot\beta_c'(\hat\beta-\beta_0)]'(\dot\beta_c'\Omega_{\hat\beta}\dot\beta_c)^{-1}\sqrt{n}[\dot\beta_c'(\hat\beta-\beta_0)]+o_p(1)\\ &\xrightarrow{\mathcal{L}} \chi^2_{p+p^*-q}\end{aligned}$$

The lemma follows from the equality

$$\hat{\dot\beta}_c(\hat{\dot\beta}_c'\hat\Omega_{\hat\beta}\hat{\dot\beta}_c)^{-1}\hat{\dot\beta}_c' = \hat\Omega_{\hat\beta}^{-1} - \hat\Omega_{\hat\beta}^{-1}\hat{\dot\beta}(\hat{\dot\beta}'\hat\Omega_{\hat\beta}^{-1}\hat{\dot\beta})^{-1}\hat{\dot\beta}'\hat\Omega_{\hat\beta}^{-1}$$

which is from lemma 1 of Khatri (1966).

REFERENCES

Allison, Paul D. 1987. "Estimation of Linear Models with Incomplete Data." Pp. 71–103 in *Sociological Methodology 1987*, edited by C.C. Clogg. San Francisco: Jossey-Bass.

Amemiya, Yasuo, and Theodore W. Anderson. 1990. “Asymptotic Chi-Square Tests for a Large Class of Factor Analysis Models.” *Annals of Statistics* 18:1453–63.

Anderson, Theodore W. 1957. “Maximum Likelihood Estimates for the Multivariate Normal Distribution When Some Observations are Missing.” *Journal of the American Statistical Association* 52:200–203.

Anderson, Theodore W., and Yasuo Amemiya. 1988. “The Asymptotic Normal Distribution of Estimators in Factor Analysis Under General Conditions.” *Annals of Statistics* 16:759–71.

Arbuckle, James L. 1996. “Full Information Estimation in the Presence of Incomplete Data.” Pp. 243–77 in *Advanced Structural Equation Modeling: Issues and Techniques*, edited by G.A. Marcoulides and R.E. Schumacker. Mahwah, NJ: Lawrence Erlbaum.

Arminger, Gerhard, and Ronald Schoenberg. 1989. “Pseudo Maximum Likelihood Estimation and a Test for Misspecification in Mean and Covariance Structure Models.” *Psychometrika*, 54:409–26.

Arminger, Gerhard, and Michael E. Sobel. 1990. “Pseudo-Maximum Likelihood Estimation of Mean and Covariance Structures with Missing Data.” *Journal of the American Statistical Association* 85:195–203.

Beale, Evelyn M.L., and Roderick J.A. Little. 1975. “Missing Data in Multivariate Analysis.” *Journal of the Royal Statistical Society*, ser. B, 37:129–45.

Bentler, Peter M. 1983. “Some Contributions to Efficient Statistics in Structural Models: Specification and Estimation of Moment Structures.” *Psychometrika* 48:493–517.

———. 1995. *EQS Structural Equations Program Manual*. Encino, CA: Multivariate Software.

———. In press. *EQS 6 Structural Equations Program Manual*. Encino, CA: Multivariate Software.

Bentler, Peter M., and Theo K. Dijkstra. 1985. “Efficient Estimation via Linearization in Structural Models.” Pp. 9–42 in *Multivariate Analysis VI*, edited by P.R. Krishnaiah. Amsterdam: North-Holland.

Bentler, Peter M., and Ke-Hai Yuan. 1999. “Structural Equation Modeling with Small Samples: Test Statistics.” *Multivariate Behavioral Research* 34:181-97.

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

Bollen, Kenneth A. and Robert Stine. 1993. “Bootstrapping Goodness-of-fit Measures in Structural Equation Models.” Pp. 111–135 in *Testing Structural Equation Models*, edited by K.A. Bollen and J.S. Long. Newbury Park, CA: Sage.

Brown, C. Hendricks. 1983. “Asymptotic Comparison of Missing Data Procedures for Estimating Factor Loadings.” *Psychometrika* 48:269–91.

Brown, Roger L. 1994. “Efficacy of the Indirect Approach for Estimating Structural Equation Models with Missing Data: A Comparison of Five Methods.” *Structural Equation Modeling* 1:287–316.

Browne, Michael W. 1984. “Asymptotic Distribution-free Methods for the Analysis of Covariance Structures.” *British Journal of Mathematical and Statistical Psychology* 37:62–83.

———. 1987. “Robustness of Statistical Inference in Factor Analysis and Related Models.” *Biometrika* 74:375–84.

Browne, Michael W., and Gerhard Arminger. 1995. “Specification and Estimation of Mean and Covariance Structure Models.” Pp. 185–249 in *Handbook of Statistical Modeling for the Social and Behavioral Sciences*, edited by G. Arminger, C.C. Clogg, and M.E. Sobel. New York: Plenum.

Browne, Michael W., and Alexander Shapiro. 1988. “Robustness of Normal Theory Methods in the Analysis of Linear Latent Variate Models.” *British Journal of Mathematical and Statistical Psychology* 41:193–208.

Curran, Patrick S., Stephen G. West, and John F. Finch. 1996. “The Robustness of Test Statistics to Nonnormality and Specification Error in Confirmatory Factor Analysis.” *Psychological Methods* 1:16–29.

Dempster, Arthur P., Nan M. Laird, and Donald B. Rubin. 1977. “Maximum Likelihood Estimation from Incomplete Data via the EM Algorithm” (with discussion). *Journal of the Royal Statistical Society*, ser. B, 39:1–38.

Dijkstra, Theo K. 1981. *Latent Variables in Linear Stochastic Models: Reflections on “Maximum Likelihood” and “Partial Least Squares” Methods*. Ph.D. dissertation, University of Groningen.

Fang, Kai-Tai, Samuel Kotz, and Kaiwang Ng. 1990. *Symmetric Multivariate and Related Distributions*. London: Chapman and Hall.

Ferguson, Thomas S. 1996. *A Course in Large Sample Theory*. London: Chapman and Hall.

Finkbeiner, Carl. 1979. “Estimation for the Multiple Factor Model When Data Are Missing.” *Psychometrika* 44:409–20.

Godambe, Vidyadhar P., and Belvant K. Kale. 1991. “Estimating Function: An Overview.” Pp. 3–20 in *Estimating Functions*, edited by V.P. Godambe. New York: Oxford University Press.

Gourieroux, Christian, Alain Monfort, and Alain Trognon. 1984. “Pseudo Maximum Likelihood Methods: Theory.” *Econometrica* 52:681–700.

Hu, Litze, Peter M. Bentler, and Yutaka Kano. 1992. “Can Test Statistics in Covariance Structure Analysis Be Trusted?” *Psychological Bulletin* 112:351–62.

Jamshidian, Mortaza, and Peter M. Bentler. 1999. “Using Complete Data Routines for ML Estimation of Mean and Covariance Structures with Missing Data.” *Journal of Educational and Behavioral Statistics* 23:21–41.

Jöreskog, Karl G., and Dag Sörbom. 1993. *LISREL 8 User’s Reference Guide,* Chicago: Scientific Software International.

Khatri, C. G. 1966. “A Note on a MANOVA Model Applied to Problems in Growth Curves.” *Annals of the Institute of Statistical Mathematics* 18:75–86.

Kline, Rex B. 1998. *Principles and Practice of Structural Equation Modeling*. New York: Guilford.

Laird, Nan M. 1988. “Missing Data in Longitudinal Studies.” *Statistics in Medicine* 7: 305–15.

Lange, Kenneth L., Roderick J.A. Little, and Jeremy M.G. Taylor. 1989. “Robust Statistical Modeling Using the t Distribution.” *Journal of the American Statistical Association* 84:881–96.

Lee, Sik-Yum. 1986. “Estimation for Structural Equation Models with Missing Data.” *Psychometrika* 51:93–99.

Liang, Kung-Yee, and Scott L. Zeger. 1986. “Longitudinal Data Analysis Using Generalized Linear Models.” *Biometrika* 73:13–22.

Little, Roderick J.A. 1988. “Robust Estimation of the Mean and Covariance Matrix from Data with Missing Values.” *Applied Statistics* 37:23–38.

Little, Roderick J.A., and Donald E. Rubin. 1987. *Statistical Analysis with Missing Data*. New York: Wiley.

Magnus, Jan R., and Heinz Neudecker. 1988. *Matrix Differential Calculus with Applications in Statistics and Econometrics*. New York: Wiley.

Mardia, Kanti V., John T. Kent, and John M. Bibby. 1979. *Multivariate Analysis*. New York: Academic Press.

Marsh, Herbert W. 1998. “Pairwise Deletion for Missing Data in Structural Equation Models: Nonpositive Definite Matrices, Parameter Estimates, Goodness of Fit, and Adjusted Sample Sizes.” *Structural Equation Modeling* 5:22–36.

Meng, Xiao-Li. 1994. “Multiple Imputation Inferences with Uncongenial Sources of Input” (with discussion). *Statistical Science* 9:538–73.

Meng, Xiao-Li, and Steven Pedlow. 1992. “EM: A Bibliographic Review with Missing Articles.” Pp. 24–27 in *Statistical Computing Section, Proceedings of the American Statistical Association*.

Micceri, Theodore. 1989. “The Unicorn, the Normal Curve, and Other Improbable Creatures.” *Psychological Bulletin* 105:156–66.

Mooijaart, Ab, and Peter M. Bentler. 1991. “Robustness of Normal Theory Statistics in Structural Equation Models.” *Statistica Neerlandica* 45:159–71.

Mueller, Ralph O. 1996. *Basic Principles of Structural Equation Modeling*. New York: Springer Verlag.

Muirhead, R. J. 1982. *Aspects of Multivariate Statistical Theory*. New York: Wiley.

Muirhead, Robb J., and Christine M. Waternaux. 1980. “Asymptotic Distributions in Canonical Correlation Analysis and Other Multivariate Procedures for Nonnormal Populations.” *Biometrika* 67:31–43.

Muthén, Bengt, David Kaplan, and Michael Hollis. 1987. “On Structural Equation Modeling with Data that Are Not Missing Completely at Random.” *Psychometrika* 52:431–62.

Muthén, Linda, and Bengt Muthén. 1998. *Mplus User’s Guide*. Los Angeles: Muthén and Muthén.

Neale, Michael C. 1994. “Mx: Statistical Modeling,” 2nd ed. Box 710 MCV, Richmond, VA 23298: Department of Psychiatry, Medical College of Virginia.

Olkin, Ingram. 1994. “Multivariate Nonnormal Distributions and Models of Dependency.” Pp. 37–53 in *Multivariate Analysis and Its Applications*, edited by T.W. Anderson, K.T. Fang, and I. Olkin. Hayward, CA: IMS.

Rotnitzky, Andrea, and David Wypij. 1994. “A Note on the Bias of Estimators with Missing Data.” *Biometrics* 50:1163–70.

Rovine, Michael J. 1994. “Latent Variables Models and Missing Data Analysis.” Pp. 181–225 in *Latent Variables Analysis: Applications for Developmental Research*, edited by A. von Eye and C.C. Clogg. Thousand Oaks, CA: Sage.

Rubin, Donald B. 1987. *Multiple Imputation for Nonresponse in Surveys*. New York: Wiley.

———. 1996. “Multiple Imputation after 18 Years.” *Journal of the American Statistical Association* 91:473–89.

Satorra, Albert, and Peter M. Bentler. 1988. “Scaling Corrections for Chi-Square Statistics in Covariance Structure Analysis.” Pp. 308–13 in *American Statistical Association 1988 Proceedings of Business and Economics Sections*. Alexandria, VA: American Statistical Association.

———. 1990. “Model Conditions for Asymptotic Robustness in the Analysis of Linear Relations.” *Computational Statistics and Data Analysis* 10:235–49.

———. 1991. “Goodness-of-fit Test under IV Estimation: Asymptotic Robustness of a NT Test Statistic.” Pp. 555–67 in *Applied Stochastic Models and Data Analysis*, edited by R. Gutiérrez and M.J. Valderrama. Singapore: World Scientific.

———. 1994. “Corrections to Test Statistics and Standard Errors in Covariance Structure Analysis.” Pp. 399–419 in *Latent Variables Analysis: Applications for Developmental Research*, edited by A. von Eye and C.C. Clogg. Newbury Park, CA: Sage.

Schafer, Joseph L. 1997. *Analysis of Incomplete Multivariate Data*. London: Chapman and Hall.

Schoenberg, Ronald. 1989. *LINCS: Linear Covariance Structure Analysis. User’s Guide*. Kent, WA: RJS Software.

Shapiro, Alexander. 1983. “Asymptotic Distribution Theory in the Analysis of Covariance Structures (a Unified Approach).” *South African Statistical Journal* 17:33–81.

———. 1987. “Robustness Properties of the MDF Analysis of Moment Structures.” *South African Statistical Journal* 21:39–62.

Shapiro, Alexander, and Michael Browne. 1987. “Analysis of Covariance Structures under Elliptical Distributions.” *Journal of the American Statistical Association* 82:1092–97.

Tanaka, Yutaka, Shingo Watadani, and Sung Ho Moon. 1991. “Influence in Covariance Structure Analysis: With an Application to Confirmatory Factor Analysis.” *Communication in Statistics-Theory and Method* 20:3805–21.

Yuan, Ke-Hai, and Peter M. Bentler. 1996. “Mean and Covariance Structure Analysis with Missing Data.” Pp. 307–26 in *Multidimensional Statistical Analysis and Theory of Random Matrices: Proceedings of Sixth Eugene Lukacs Symposium*, edited by A. Gupta and V. Girko. Utrecht, Netherlands: VSP.

———. 1997a. “Mean and Covariance Structure Analysis: Theoretical and Practical Improvements.” *Journal of the American Statistical Association* 92:767–74.

———. 1997b. “Improving Parameter Tests in Covariance Structure Analysis.” *Computational Statistics and Data Analysis* 26:177–98.

———. 1998. “Normal Theory Based Test Statistics in Structural Equation Modelling.” *British Journal of Mathematical and Statistical Psychology* 51:289–309.

———. 1999. “On Normal Theory and Associated Test Statistics in Covariance Structure Analysis Under Two Classes of Nonnormal Distributions.” *Statistica Sinica* 9:831–53.

Yuan, Ke-Hai, and Robert I. Jennrich. 1998. “Asymptotics of Estimating Equations Under Natural Conditions.” *Journal of Multivariate Analysis* 65:245–60.

Yung, Yiu-Fai, and Peter M. Bentler. 1996. “Bootstrapping Techniques in Analysis of Mean and Covariance Structures.” Pp. 195–226 in *Advanced Structural Equation Modeling Techniques*, edited by G.A. Marcoulides & R.E. Schumacker. Mahwah, NJ: Erlbaum.