Running head: TESTS OF HOMOGENEITY OF MEANS AND COVARIANCE

Tests of Homogeneity of Means and Covariance Matrices for Multivariate Incomplete Data[1]

Kevin H. Kim[2] and Peter M. Bentler

University of California, Los Angeles

[1]Preliminary results on this research were presented at the 1999 Western Psychological Association convention, Irvine, CA.

[2]The assistance of Dr. Ke-Hai Yuan is gratefully acknowledged. Address correspondence to the first author at Department of Psychology, UCLA, Box 951563, Los Angeles, CA 90095-1563, or by email at kevinkim@ucla.edu.

Abstract

Existing test statistics for testing whether incomplete data may represent a missing completely at random sample from a single population with a given mean vector and covariance matrix are based on a likelihood rationale. This rationale cannot be implemented adequately when some patterns of missing data may contain very few subjects. A generalized least squares rationale is used to develop parallel tests that should be more stable in small samples. Several options for weighting the contribution of data from various patterns of missing data are studied. These are asymptotically equivalent, but are shown in a simulation study to perform radically differently in finite samples. Three factors were varied for the simulation: number of variables, percent missing complete at random, and sample size. One thousand data sets were simulated for each condition. Little's test of homogeneity of means and a modified generalized least squares test of homogeneity of covariance matrices performed close to an ideal Type I error rate for most of the conditions. A combined test performed almost as well.

Keyword List: Missing data, Likelihood ratio, Generalized least squares, Multivariate normal distribution

Tests of Homogeneity of Means and Covariance Matrices for Multivariate Incomplete Data

Incomplete data is one of the most pervasive problems in data analyses. The problem occurs for various reasons (e.g., a participant refuses to answer a question, a question is skipped accidentally, equipment fails, a participant drops out). The degree to which it is problematic depends on the pattern of incomplete data and how much is missing (Tabachnick & Fidell, 1996). The pattern of incomplete data is a more important problem than the amount missing. Missing values scattered randomly through a data matrix pose a less serious problem because almost any method can be used to deal with the problem while still obtaining consistent estimates of parameters, tests, and standard errors. Of course, the greater the amount of incomplete data, the greater the loss of information and hence the lower the statistical precision. If only a few data points are missing in a random pattern from a large data set, almost any procedure for handling missing values yields similar results. If, however, a lot of data are missing from a small to moderately sized data set, the problems can be very serious and special procedures have to be used to minimize errors of inference. Unfortunately, there are no firm guidelines for how much incomplete data can be tolerated for a sample of given size.

If the missingness is independent of both the missing values and the observed values of other variables, as in the discussion above, the data is said to be missing completely at random (MCAR). However, if the missingness is only independent of the missing values and not on the observed values of other variables then the data is said to be missing at random (MAR; Little & Rubin, 1987). When data are MCAR or MAR, likelihood based

methods can yield acceptable statistical results. When data are neither MCAR nor MAR, nonrandom patterns of incomplete data, on the other hand, are serious no matter how few of them there are because they affect the generalizability of results. There are three major problems created by incomplete data: (1) if the missingness is neither MCAR or MAR, than any analysis that ignores the nonrandomness may be biased, (2) the existence of incomplete data usually implies a loss of information, so that estimates will be less efficient than planned, and (3) standard statistical methods are designed for complete data; thus any incomplete data analysis is usually more complicated (Little & Schenker, 1995).

An important development in the handling incomplete data is the Expectation - Maximization (EM) algorithm formalized by Dempster, Laird, and Rubin (1977). The EM algorithm is a very general iterative algorithm for obtaining maximum likelihood estimators (MLE) in incomplete data problems. The EM algorithm formalizes a relatively old ad hoc idea for handling incomplete data: (1) replace missing values by regression-imputed estimated values, (2) estimate parameters, (3) reestimate the missing values assuming the new parameter estimates are correct, (4) reestimate parameters, and so forth, iterating until convergence. Dempster et al. (1977) exposed the full generality of the algorithm by proving general results about its behavior, specifically that each iteration increases the likelihood, and by providing a wide range of examples. The range of problems that can be attacked by EM is very broad and includes problems not usually considered to be ones arising from incomplete data (e.g., variance components estimation, iteratively reweighted least squares estimation). In the context of structural equation modeling (SEM), this algorithm can be used to obtain MLEs for unstructured means μ and covariances Σ, or models based on structured means $\mu = \mu(\theta)$ and covariances $\Sigma = \Sigma(\theta)$, where θ is the vector of basic model

parameters. See, for example, Rovine (1994), Arbuckle (1996) and Jamshidian and Bentler (1999). In this paper we deal only with the unstructured means μ and covariances Σ, and the EM algorithm is used to to obtain the MLE $\hat{\theta}$ of θ, which contains the elements of μ and the nonredundant elements of Σ.

The MLE estimates $\hat{\theta}$ obtained from the EM algorithm when applied to unstructured means and covariances are consistent estimates of the population parameters under either MAR or MCAR assumptions. It would be desirable to test these assumptions. In SEM applications of EM, for example, a single vector of population means and covariances are analyzed, but if the data are not MCAR then a single set of mean and covariance parameters will be inadequate to describe the data. Although no tests have been proposed to test whether a data set is MAR, several tests have been proposed to evaluate the MCAR assumption. However, as we shall see, aside from an ad hoc procedure, these tests are based on likelihood formulae that may be unstable in small to medium sized samples. In this paper, we develop a variety of tests for MCAR that are based on a generalized least squares (GLS) rationale that, a priori, should be less likely to break down. Several variants of these tests are developed that are equivalent in large samples, but are shown to perform differently in realistic sized samples.

Notation and Null Hypotheses

A somewhat ad hoc method of testing MCAR when only a single variable contains missing values is to compare the distributions of fully observed variables for missing and observed, using a t-test for the difference in means. The computer program BMDP8D extended the t-test approach to multivariate data with missing values on any of p variables (Dixon, 1988).

For each variable with missing values, the data is divided into two parts; cases with missing and observed data. The means of observed values of the other variables in the two groups are then compared by two-sample independent t-tests. If there are no significant differences between these means, the data is MCAR. However, this method requires $p(p-1)$ t-tests for assessing the MCAR assumption, and since the tests are not independent, the probability levels associated with the tests will be questionable (i.e., inflation of Type I error). Furthermore, means may be MCAR while covariance are not, and so this test is incomplete.

There have been developments that test MCAR in several different areas (e.g., contingency tables, Fuchs, 1982; generalised estimating equations, Chen & Little, 1999). In the area of multivariate normal data, there are currently two proposed test statistics for analyzing whether incomplete data patterns are MCAR (Little, 1988; Tang & Bentler, 1998). Both Little and Tang-Bentler use the EM algorithm to impute incomplete data, obtain MLE of free parameters under the MCAR assumption, and propose a test statistic that can be used to evaluate an MCAR null hypothesis. Little (1988) developed an MCAR test based on evaluating the homogeneity of available means for different patterns of incomplete data. A given variable will be observed under several patterns of missing data, for example variable 1 may be observed for subjects who omit variable 2, or also for subjects who do not omit variable 2 but omit variable 3, and so on. The means on variable 1 from the various sets of subjects should be estimates of the same population mean if the data are MCAR. Similarly, the various means on variable 2 would be homogeneous if the data are MCAR. Little's test evaluates the homogeneity of means for all variables, across data patterns, simultaneously. Little also mentioned, but did not study, a test based on both means and covariances, in which the homogeneity of available covariance matrices is simultaneously studied with homo-

geneity of means. He expected that this test might not perform well due to its typically large degrees of freedom. In quite a different context, Tang and Bentler (1998) studied covariance structures, such as factor analysis models, for incomplete data. As will be shown below, when their test is specialized to that of an unstructured but common covariance matrix for all patterns of incomplete data, it provides a test of the MCAR assumption. In fact, it will be shown that their test, in this case, specializes to a test that can be constructed based on a chi-square difference rationale applied to Little's two proposed tests. These ideas will be made more precise.

Consider the data matrix X of N observations (i.e., cases) on p variables with missing data, where $X \sim N_p(\mu, I \otimes \Sigma)$. Let X_{ij} represent the score of the i^{th} case on the j^{th} variable and let Y $(N \times p)$ represent a pattern matrix of X, such that the elements of Y are defined as

$$Y_{ij} = \begin{cases} 0 & : \quad X_{ij} \quad \text{is missing} \\ 1 & : \quad X_{ij} \quad \text{is observed} \end{cases} \qquad \text{for all } i = 1, \ldots, N \text{ and } j = 1, \ldots, p$$

Then the i^{th} row of Y, Y_i, defines a patterns of 1's and 0's indicative of present and absent data, and will be called an incomplete data pattern. Another row of Y, the k^{th} row, Y_k, may contain an identical pattern of 1's and 0's, or a different pattern (e.g., if $p = 5$, $Y_i = [1\,1\,1\,0\,0]$ and $Y_k = [1\,1\,1\,0\,0]$ as opposed to $Y_i = [1\,1\,1\,0\,0]$ and $Y_k = [1\,0\,1\,1\,0]$). Often several cases in the data file will have an identical pattern of 1's and 0's. Any given pattern of incomplete data may occur up to a maximum of N times; this can only occur if there is only one data pattern, either complete (no missing) or incomplete (where all cases are missing certain variable(s)). However, if there are ℓ variables with incomplete data, where $\ell \in \{1, \ldots, p\}$, then there can be up to 2^ℓ incomplete data patterns if $\ell < p$, or $(2^\ell - 1)$ incomplete data patterns if $\ell = p$.

In practice, there may be far fewer. If there is a very small number of patterns of incomplete data, in the context of covariance structure analysis, the multiple group methods of Allison (1987) and Muthén, Kaplan, and Hollis (1987) could be used to develop a test of the MCAR assumption. However, this is very unlikely in practice.

The following notation is used throughout: m is the number of incomplete data patterns; S_i is the $(p_i \times p_i)$ sample covariance matrix for observed variable(s) for pattern i; $\hat{\Sigma}_i$ is the $(p_i \times p_i)$ EM imputed covariance matrix for observed variable(s) for pattern i, and a submatrix of $\hat{\Sigma}$, the $(p \times p)$ EM imputed model covariance matrix; p_i is number of observed variable(s) for pattern i; p is the number of variables; n_i is the number of cases for pattern i; N $(= \sum_{i=1}^{m} n_i)$ is the total number of cases; $\bar{X}_i$ is the $(p_i \times 1)$ vector of means for observed variables for pattern i; and $\hat{\mu}_i$ is the $(p_i \times 1)$ vector of EM imputed means for observed variables for pattern i, a subvector of $\hat{\mu}$, the $(p \times 1)$ vector of EM imputed means. For simplicity, we denote the number of sample means available in an analysis as p_1^* $(= \sum_{i=1}^{m} p_i)$, the number of sample covariances available as p_2^* $(= \sum_{i=1}^{m} \frac{p_i(p_i+1)}{2})$, and the number of sample means and covariances available as p^* $(= \sum_{i=1}^{m} \frac{p_i(p_i+3)}{2})$. We also make use of the sample size proportions: $c_i = \frac{n_i}{N}$ and $k_i = \frac{n_i-(1-c_i)}{N}$.

We may consider three clases of test statistics that test hypotheses that can be generated from the MCAR theory. The first null hypothesis is that the population means μ_i for the various patterns of incomplete data are subsets of a single population mean vector μ ($\mu_i \subseteq \mu$ for all $i = 1, \ldots, m$). A test of this hypothesis will be called the means test. It is based on comparing the p_1^* available sample means to the $q_1 = p$ MLE estimators of the common means, and has $d_1 = (p_1^* - q_1)$ degrees of freedom. The second null hypothesis is that the population covariance matrices Σ_i for the various patterns of incomplete data are subsets

of a single population covariance matrix Σ ($\Sigma_i \subseteq \Sigma$ for all $i = 1, \ldots, m$). A test of this hypothesis will be called the covariance matrices test. It is based on comparing the p_2^* available sample covariances to the $q_2 = \frac{p(p+1)}{2}$ MLE estimators of the common covariances, and has $d_2 = (p_2^* - q_2)$ degrees of freedom. The third null hypothesis is that the population means μ_i and covariance matrices Σ_i for the various patterns of incomplete data are both simultaneously subsets of a single population mean μ and covariance matrix Σ ($\mu_i \subseteq \mu$ and $\Sigma_i \subseteq \Sigma$ for all $i = 1, \ldots, m$). A test of this hypothesis will be called the mean and covariance matrices or combined test. The combined test is based on comparing the p^* available sample means and covariances to the $q = \frac{p(p+3)}{2}$ MLE estimators of the common means and covariances, and has $d_3 = (p^* - q)$ degrees of freedom.

Likelihood Based Tests of MCAR

We are now ready to examine extant tests. Little (1988) developed the means test

$$d^2 = NL_1(\hat{\theta}) = N \sum_{i=1}^{m} c_i[(\bar{X}_i - \mu_i(\hat{\theta}))' \Sigma_i^{-1}(\hat{\theta})(\bar{X}_i - \mu_i(\hat{\theta}))]. \quad (1)$$

The EM algorithm provides $\hat{\theta}$, that is $\hat{\mu}$ and $\hat{\Sigma}$, the common means and covariances, from which the subsets $\hat{\mu}_i$ and $\hat{\Sigma}_i$ are obtained. Little showed that $d^2 \sim \chi^2_{d_1}$ under the null hypothesis. To get some insight into equation (1), if all sample mean vectors $\bar{X}_i$ of incomplete patterns are subsets of a population mean vector and happen to have sample means that are exact subsets of a set of common means (i.e., $\bar{X}_i = \hat{\mu}_i$ and $\mu_i \subseteq \mu$), then the test statistic will equal zero. More realistically, under the null hypothesis, $\bar{X}_i \simeq \mu_i(\hat{\theta}) = \hat{\mu}_i$ and $E(d^2) = d_1$, i.e., the degrees of freedom. Thus all cases would have been sampled from a population with same mean. Rejection of the null hypothesis implies that the population means for

the various patterns are not homogeneous, i.e., that it is not plausible that the data are MCAR. While this is an interesting and valuable test, in most applications of SEM, interest focuses more on covariance matrices than on means. It would be possible for means to be homogeneous while covariances are not. To solve this problem, Little (1988) also proposed, but did not study, a combined test:

$$d^2_{aug} = NL_3(\hat{\theta}) = d^2 + N\sum_{i=1}^{m} c_i[\text{tr}(S_i\Sigma_i^{-1}(\hat{\theta})) - \log|S_i\Sigma_i^{-1}(\hat{\theta})| - p_i]. \quad (2)$$

The asymptotic distribution of d^2_{aug} is $\chi^2_{d_3}$, where d_3 is the degrees of freedom. To understand d^2_{aug}, if all mean and covariance matrices of incomplete patterns are subsets of population mean and covariance matrices and are homogeneous subsets in the various samples, (i.e., with $\bar{X}_i = \hat{\mu}_i$ with $\mu_i \subseteq \mu$ and $S_i = \hat{\Sigma}_i$ with $\Sigma_i \subseteq \Sigma$ for all patterns), then the test statistic will equal zero. More realistically, under the null hypothesis, $\bar{X}_i \simeq \hat{\mu}_i$ and $S_i \simeq \hat{\Sigma}_i$ and $E(d^2_{aug}) = d_3$. In this case the null hypothesis that all cases and patterns of incomplete data are samples from a population with the same mean and covariance matrix would be accepted. This would be an optimal situation for SEM, because it implies that it is meaningful to evaluate a model for the population means and covariances. In contrast, rejection of the null hypothesis would imply that not all population means and covariances are homogeneous across patterns of missing data, and hence to model a single summary mean vector and covariance matrix may not be appropriate.

Next we develop a covariance matrices test following this rationale. Using Little's (1988) means and combined tests, a covariance test can be derived as the difference test

$$d^2_{cov} = d^2_{aug} - d^2 = N\sum_{i=1}^{m} c_i[\text{tr}(S_i\Sigma_i^{-1}(\hat{\theta})) - \log|S_i\Sigma_i^{-1}(\hat{\theta})| - p_i]. \quad (3)$$

This is distributed asymptotically under the null hypothesis as $\chi^2_{d_2}$ with d_2 degrees of freedom.

Although this test seems not previously to have been described, it can be obtained from Tang and Bentler (1998). They proposed minimizing

$$L_2(\theta) = \sum_{i=1}^{m} c_i\{\log|\Sigma_i| + \mathrm{tr}[\Sigma_i^{-1}(S_i^* - \mu_i\bar{X}_i' - \bar{X}_i\mu_i' + \mu_i\mu_i')] - \log|S_i| - p_i\},$$

where $S_i^* = S_i + \bar{X}_i\bar{X}_i'$ and the argument θ has been omitted for simplicity. However, this can be rewritten as

$$L_2(\theta) = \sum_{i=1}^{m} c_i\{\log|\Sigma_i| + \mathrm{tr}(S_i\Sigma_i^{-1}) + (\bar{X}_i - \mu_i)'\Sigma_i^{-1}(\bar{X}_i - \mu_i) - \log|S_i| - p_i\}.$$

If we allow $\bar{X}_i = \hat{\mu}_i$ with $\hat{\mu}_i \not\subseteq \hat{\mu}$, i.e., permit the means not to be structured or held equal by patterns, it follows that $NL_2(\hat{\theta}) = d_{cov}^2$ as defined in (3).

The asymptotic distribution of the parameters of the MLE for means and covariances is given in the first part of the appendix. It will be clear from that development, as well as the forms of equations (1)-(3), that the test statistics described above are all based on a log-likelihood function and rationale. We use $L(\theta)$ mnemonically to provide a reminder of the likelihood origin of these tests. Asymptotically, likelihood-based tests are optimal when its assumed conditions, such as multivariate normality and asymptotic sample size, are met. However, such tests may not be optimal in practice in small to intermediate sized samples, even if normality is not an issue. Little (1988) cited a restriction for his combined test (2) which evidently must also apply to the covariance matrices test (3). Since S_i is singular if the number of cases for that pattern is smaller than the number of observed variables in the pattern, $|S_i|$ will be zero and the $\log|S_i|$ will go to negative infinity. Therefore, these tests will only be applicable if the number of cases is greater than or equal to the number of observed variables in a given pattern. In practice this implies that it will not be possible to use all cases in an analysis, since patterns with very small n_i will often exist. In fact, S_i may

be near singular when n_i is only somewhat larger than p_i, and hence the log likelihood and the covariance and combined tests derived from it can theoretically be quite unstable in small to intermediate sized samples. Surprisingly, there is essentially no published information on how problematic these likelihood tests might be in practice. To overcome these problems, new covariance and combined tests of MCAR are developed herein that would permit using all cases. These new statistics are developed based on the GLS approach.

Generalized Least Squares Tests of MCAR

Browne (1974) showed that the normal theory GLS estimator and test statistic for covariance structures perform equivalently, in large samples, to their MLE counterparts. In this paper, an iteratively updated GLS weight matrix based on the inverse model covariance matrix, $\Sigma_i^{-1}(\hat{\theta})$ is used rather than the inverse of the sample covariance matrix S_i^{-1}which is typically used in GLS estimation but may not be invertible in small samples. In complete data applications, Lee and Jennrich (1979) showed that minimizing a GLS function by iteratively updating the GLS weight matrix Σ^{-1}, based on the current values of the parameters $\hat{\theta}$, produces the MLE. Our tests are of the form studied by Lee and Tsui (1982), though their application was in the context of multiple population analyses and general covariance structures. Bentler (1989) also extended this method to mean structures, in parallel to that described for multiple populations. We develop means, covariances, and combined tests based on the GLS rationale provided by these prior writers.

Using the weighting coefficients c_i for the various patterns in the GLS function as has traditionally been proposed in multisample GLS estimation, and as was implicitly utilized by Little (1988) in his means test, the GLS means test is the same as Little's means test,

equation (1). The covariance test is obtained by calculating:

$$G_2(\hat{\theta}) = \sum_{i=1}^{m} \frac{c_i}{2} \mathrm{tr}[(S_i - \Sigma_i(\hat{\theta}))\Sigma_i^{-1}(\hat{\theta})]^2. \quad (4)$$

Asymptotically, $NG_2(\hat{\theta}) \sim \chi^2_{d_2}$, the same as the likelihood based covariance test. However, the possible singularity of S_i does not disturb this test. In (4), as in Little's means test, only the EM-based MLE model covariance matrix $\hat{\Sigma}$ has to be invertible. The individual pattern model covariance matrices Σ_i are all subsets of this one matrix, and hence will be invertible if the MLE matrix is positive definite. In general, this should occur. As before, under the null hypothesis, the expected value of $NG_2(\hat{\theta})$ is given by d_2, the degrees of freedom. If $NG_2(\hat{\theta})$ is small compared to degrees of freedom, the null hypothesis that all cases represent samples from a population with same covariance matrix would be accepted. This would permit covariance structure analysis of the single covariance matrix $\hat{\Sigma}$ rather than potentially requiring separate models for the various S_i. On the other hand, if $NG_2(\hat{\theta})$ is large compared to degrees of freedom, the null hypothesis would be rejected, and structural modeling of the common estimate $\hat{\Sigma}$ would not adequately mirror the heterogeneity in S_i.

As in the case of likelihood analyses, the means and covariances functions can be combined to yield

$$G_3(\hat{\theta}) = \sum_{i=1}^{m} \{c_i[(\bar{X}_i - \mu_i(\hat{\theta}))' \Sigma_i^{-1}(\hat{\theta})(\bar{X}_i - \mu_i(\hat{\theta}))] + \frac{c_i}{2} \mathrm{tr}[(S_i - \Sigma_i(\hat{\theta}))\Sigma_i^{-1}(\hat{\theta})]^2\}. \quad (5)$$

The technical basis for concluding that, under the null hypothesis asymptotically $NG_3(\hat{\theta}) \sim \chi^2_{d_3}$ is given in the second part of the Appendix. Unlike its likelihood counterpart, this test will not break down with singular sample covariance matrices for any pattern of missing data.

Variants of Generalized Least Squares Tests

The results of a pilot study showed that Little's combined test (2) and our version of the Tang-Bentler covariance tests (4) produced a very large inflation of Type I error, even when considering only data patterns meeting the restriction that sample size is larger than the number of variables for that pattern. While the GLS combined and covariances test results were a significant improvement, they did not produce the desired Type I error rate. Therefore, we developed new weighting constants that hopefully would adjust or reflect appropriately the contribution of small sample size to χ^2. There are two variants of our new tests, several modified GLS (MGLS) tests, and a mixed-weight GLS (MWGLS) test. In the MGLS tests, the weighting variable from GLS, $c_i = n_i/N$ is replaced by $k_i = [n_i-(1-c_i)]/N$, a weighting variable derived for this study. In the MWGLS combined test, the means weights are the traditional GLS weights c_i, while the covariance weights are the new weights k_i. A study of the weighting variables c_i and k_i will reveal that they are similar at two extremes, where $n_i << N$ and $n_i = N$. For instance, if $n_i << N$, then both $c_i \approx 0$ and $k_i \approx 0$. At the other extreme, $n_i = N$, both the weighting variables c_i and k_i will equal 1, and thus be identical to the weight with complete data. Where the two weighting variables differ is between these two extremes. For example, if $n_i = 2$ and $N = 100$, then $c_i = 0.02$ whereas $k_i = 0.0102$. The weighting variable has been reduced 49 percent, a large reduction of its original weight. However, if $n_i = 50$, then $c_i = 0.5$ and $k_i = 0.495$, a reduction of only 1 percent, a negligible amount. Since asymptotically c_i and k_i are equal, it was anticipated that these relatively small differences in weights might lead to very different small sample behavior of tests defined on them.

As an obvious modification of Little's means test, the MGLS means test is given by $NM_1(\hat{\theta})$, where

$$M_1(\hat{\theta}) = \sum_{i=1}^{m} k_i[(\bar{X}_i - \mu_i(\hat{\theta}))'\Sigma_i^{-1}(\hat{\theta})(\bar{X}_i - \mu_i(\hat{\theta}))]. \tag{6}$$

The MGLS covariance matrices test is given as $NM_2(\hat{\theta})$, where

$$M_2(\hat{\theta}) = \sum_{i=1}^{m} \frac{k_i}{2}\mathrm{tr}[(S_i - \Sigma_i(\hat{\theta}))\Sigma_i^{-1}(\hat{\theta})]^2. \tag{7}$$

And the combined test is given by $NM_3(\hat{\theta})$, where

$$M_3(\hat{\theta}) = \sum_{i=1}^{m}\{k_i[(\bar{X}_i - \mu_i(\hat{\theta}))'\Sigma_i^{-1}(\hat{\theta})(\bar{X}_i - \mu_i(\hat{\theta}))] + \frac{k_i}{2}\mathrm{tr}[(S_i - \Sigma_i(\hat{\theta}))\Sigma_i^{-1}(\hat{\theta})]^2\}. \tag{8}$$

The asymptotic distributions and degrees of freedom of these tests are identical to their likelihood counterparts. Preliminary simulation runs had established that Little's (1988) means test generally performs quite well. Thus there is some question whether the use of k_i in the means test based on (6) and also in the combined test based on (8) above might not actually be harmful rather than helpful. In contrast, we expected the new covariance matrices test based on (7) to perform well. As a result a new combined test, the MWGLS test was developed. It is given as $NF_3(\hat{\theta})$, where

$$F_3(\hat{\theta}) = \sum_{i=1}^{m}\{c_i(\bar{X}_i - \mu_i(\hat{\theta}))'\Sigma_i^{-1}(\hat{\theta})(\bar{X}_i - \mu_i(\hat{\theta})) + \frac{k_i}{2}\mathrm{tr}[(S_i - \Sigma_i(\hat{\theta}))\Sigma_i^{-1}(\hat{\theta})]^2\}. \tag{9}$$

Since c_i and k_i are consistent though different estimates of a common population proportion γ_i, it is anticipated that the MWGLS test based on (9), being based on component functions (1) and (7) that we anticipate to work well in practice, would improve error rates under the null hypothesis in small to intermediate samples. Because the combined test $NF_3(\hat{\theta})$ involves the unusual feature of being based on two different weighting constants, in the Appendix we provide a proof that the asymptotic distribution of the test is $\chi^2_{d_3}$.

In summary, there are two means tests, three covariance tests, and three combined tests that have been proposed to evaluate MCAR assumption. Asymptotically, that is, with very large samples and under the null hypothesis and the multinormality assumption, the various means tests should perform equivalently. Similarly, the covariance matrix tests should perform equivalently, as should the combined tests. However, realistic data sets often contain many patterns of missing data with very small samples, and asymptotically equivalent tests may perform very differently. Aside from a small simulation done by Little (1988) on his means test only, essentially nothing is known about how well these tests work in practice. Thus, a Monte Carlo study was carried out to study the properties of the test statistics under various conditions.

Monte Carlo Study

In order to evaluate the relative performance of the various test statistics under null conditions, a series of simulated experiments were conducted. In each of those experiments, 1000 samples were generated from a specific population under known MCAR conditions. The various proposed MCAR tests were then computed in each sample. The performance of these tests across all samples is then evaluated empirically and compared to the proposed reference distributions with the given degrees of freedom. This process was then repeated 140 times under 140 different conditions, so that 140,000 data sets were used. The data sets were simulated by manipulating three factors: 1) number of variables, 2) sample size, and 3) percent MCAR.

Materials

A C++ computer program was written that used parallel processing (i.e., Message-

Passing Interface) to simulate data, complete the data analyses, and record the results. The computer program ran on the Unix operating system.

Design and Procedure

A $4 \times 5 \times 7$ between-subjects Monte Carlo study was conducted to study the properties of tests of homogeneity of means and covariance matrices for incomplete data. The test properties were examined by manipulating three factors. First, there were four different numbers of variables; 5, 10, 20, and 30. Second, there were five levels of MCAR; 10, 20, 30, 40, and 50 percent. And lastly, there were seven different sample sizes, 100, 200, 300, 400, 500, 1000, and 1500. One thousand data sets were simulated for each condition.

Univariate data were randomly sampled from a normal distribution with a mean equal to zero and standard deviation of one using an algorithm adopted from Odeh and Evans (1974). In order to control for possible univariate outliers, any datum with a higher absolute value of more than 3.3 was replaced with zero. The process described above was repeated for all cases and variables. The variables in the data sets were uncorrelated. After data was created, missing values were generated by associating a uniform random number between 0 and 1 (inclusive) to each number in the data set; i.e., created a uniform random probability associated with each number. If a probability of a number was less than or equal to the percent MCAR, a number was coded missing. Missing values were generated among the first five variables of the data set according to percent MCAR. The first five variables contain the same amount of percent MCAR. Up to 32 distinct incomplete patterns were simulated. The probability of observing any particular incomplete pattern could be computed by multiplying the probabilities of the observed pattern of variables. For example, the probability of observing an incomplete pattern, 00111 (i.e., first two variables are missing), for five variables

with 20 percent MCAR would be (.20)(.20)(.80)(.80)(.80) = .020, or 2 percent of a data set would have this particular incomplete pattern.

The EM algorithm was used to estimate the missing data and obtain maximum likelihood estimates of the common mean $\hat{\mu}$, and common covariance matrix $\hat{\Sigma}$. These estimates are also the iteratively reweighted GLS estimates. The EM algorithm was set at maximum of 500 iterations, and a difference of .00001 in the root mean square of the determinant of the covariance matrices across iterations was used as the convergence criterion. The MCAR test statistics (equations (1)-(9)) were computed using the maximum likelihood and sample mean and covariance matrices. For Little's covariance and combined test as well as the Tang-Bentler test, i.e., for all ML tests, only those incomplete patterns that met Little's restriction were used to compute the test statistics. The results of the analyses were written to an external file.

Scoring

Type I error rates were computed by tabulating the percent of cases within each condition where the test statistic was significant at $\alpha = .05$. The Type I error rates for 30 variables with 100 cases were dropped from computing the mean Type I error rates due to their inconsistency with rest of the results (i.e., outliers) and due to a small ratio between sample size and number of variables (100/30 = 3.33). Furthermore, the Type I error rates of 30 variables with 50 percent MCAR at 200 sample size for all percent MCAR were also not used to compute the mean Type I error rates of covariance or combined test statistics.

There were a limited number of data sets which met the restriction placed by Little's test of homogeneity of covariance matrices and where the sample covariance matrices were positive definite for 30 variables with high percent MCAR. Since only 14 percent of available

cases were available for computation of ML covariance and combined test statistics for 30 variables with 30 to 50 percent MCAR, a Type I error rates of the ML covariance and combined test statistics were not computed for these conditions. Furthermore, Type I error rates of the ML covariance and combined test statistics could not be computed for 20 variables with 100 cases for all percent MCAR, as well as 200 sample sizes for 40 and 50 percent MCAR.

All data were used to compute Type I error rates for other test statistics except for the few conditions where the EM algorithm failed to converge. For 20 variables, only about 87 percent of data sets were available for tabulation for 50 percent MCAR at 100 cases. Available sample sizes increased to 99 percent for 200 cases. And only one data set failed to converge for 30 percent MCAR at 100 cases. For 30 variables, 99.8 and 95 percent of data sets were used for 40 and 50 percent MCAR at 200 sample size. The only other restriction was with five variables, where about 99 percent of data sets were used for 30 percent MCAR at 200 cases and 50 percent MCAR at 100 cases.

Results

The overall mean of the variables in the 140,000 data sets simulated was 0.00 (SD = 0.02). Also the mean standard deviation and covariance of the variables in the data sets were 0.99 (SD = 0.01) and 0.00 (SD = 0.01) respectively. These numbers indicate that the data generation worked correctly.

The number of incomplete patterns generated in the data sets increased with increase in sample size and percent MCAR (see Table 1). The maximum number of incomplete patterns, 32, was reached quickly for 40 and 50 percent MCAR, even for small sample sizes. The number of incomplete patterns generated were not different between number of variables,

since all data sets had same number missing.

The mean Type I error rates for 4 (number of variables) × 5 (percent MCAR) × 7 (sample size) Monte Carlo study were computed for each test statistic in every condition. There were 140 mean Type I error rates for each test statistic.

Means test: the mean Type I error rate of Little's ML-based test of homogeneity of means for incomplete data was 3.67 percent with standard deviation of 0.97 percent across number of variables, percent MCAR, and sample sizes (see Table 2). The mean Type I error rate did not deviate far from the ideal Type I error rate of five percent (see Table 3 & Figure 1). The minimum Type I error rate was 1.0 percent for five variables with 50 percent MCAR and 100 cases. The maximum Type I error rate was 6.2 percent for five variables with 10 percent MCAR and 1500 cases. The ML and GLS tests are identical, and hence GLS is not tabulated separately.

In contrast, the mean Type I error rate of the MGLS test of homogeneity of means for incomplete data was 0.36 percent with standard deviation of 0.48 percent (see Table 2). The minimum Type I error rate for MGLS test of homogeneity of means was 0.0 percent in many conditions where the sample size was low. The maximum rate of Type I error rate for MGLS test of homogeneity of means was 3.7 percent for five variables with 50 percent MCAR and 1500 cases. Clearly, Little's ML/GLS means test performed more adequately than the MGLS means test.

Covariance matrices test: the mean Type I error rate of the ML-based test of homogeneity of covariance matrices for incomplete data was 85.15 percent with standard deviation of 12.37 percent across number of variables, percent MCAR, and sample sizes (see Table 2). The minimum Type I error rate for the ML test of homogeneity of covariance matrices was

8.0 percent for five variables with 40 percent MCAR and 1500 cases. Only two mean Type I error rates were under 10 percent, which occurred for five variables with 1500 cases at two extreme levels of MCAR (i.e., 40 and 50 percents). The maximum Type I error rate for ML was 100.0 percent for many of conditions where number of variables were high. Most data sets with 10 variables and all with 20 variables produced a Type I error rate of 100.0 percent.

Similarly, the mean Type I error rate of the GLS test of homogeneity of covariance matrices for incomplete data was 54.43 percent with standard deviation of 15.92 percent, far above the desired five percent rate (see Table 2). The mean Type I error rates increased as number of variables and percent MCAR increased. The minimum Type I error rate was 3.8 percent for five variables with 50 percent MCAR and 1500 cases. The maximum Type I error rate was 100.0 percent for 30 variables with moderate sample sizes.

Clearly, both ML and GLS tests have unacceptably high error rates. In contrast, the mean Type I error rate of the MGLS test of homogeneity of covariance matrices for incomplete data was 5.17 percent with standard deviation of 2.45 percent (see Table 2). The mean Type I error rates for five and ten variables across percent MCAR are very close to five percent with very little variation. However for 20 and 30 variables at an extreme percent MCAR, there are more deviations from the ideal five percent rate (see Table 4 & Figure 2). The minimum Type I error rate was 0.2 percent for 30 variables with 10 percent MCAR and 1000 cases. The maximum Type I error rate was 15.4 percent for 30 variables with 40 percent MCAR and 300 cases.

Combined test: the mean Type I error rate of Little's ML-based test of homogeneity of means and covariance matrices for incomplete data was 82.78 percent with a standard deviation of 12.99 percent across number of variables, percent MCAR and sample sizes (see

Table 2). The minimum Type I error rate was 7.6 percent for five variables with 40 percent MCAR and 1500 cases. The maximum Type I error rate was 100.0 percent for many of the conditions. Most of the 10 and all of 20 variable data sets produced Type I error rates of 100 percent.

Similarly, the mean Type I error rate of the GLS test of homogeneity of means and covariance matrices for incomplete data was 52.53 percent with a standard deviation of 15.76 percent (see Table 2). The minimum Type I error rate for the GLS test was 3.4 percent for five variables with 40 percent MCAR and 1000 cases. The maximum Type I error rate for the GLS was 100.0 percent for 30 variables with 50 percent MCAR and 300 and 400 cases.

Clearly, both ML and GLS-based tests have unacceptably high error rates. In contrast, the mean Type I error rate of the MGLS test of homogeneity of means and covariance matrices for incomplete data was 2.42 percent with a standard deviation of 1.44 percent (see Table 1). The minimum Type I error rate for the MGLS test was 0.1 percent for several conditions where the number of variables was high (i.e., 20 and 30 variables) with moderate percent MCAR (i.e., 20 and 30 percent) and small sample sizes (i.e., 100 and 200 cases). The maximum Type I error rate for the MGLS test was 15.6 percent for 30 variables with 50 percent MCAR and 200 cases. While the MGLS test performed in a far superior manner as compared to ML and GLS-based tests, the MGLS test had a slightly too low error rate.

The mean Type I error rate of the MWGLS test of homogeneity of means and covariance matrices was 4.34 percent with a standard deviation of 2.09 percent (see Table 2). The minimum Type I error rate for this test was 0.2 percent for a high number of variables with low percent MCAR and small sample sizes (see Table 5 & Figure 3). The maximum Type

I error rate for this test was 14.3 percent for 30 variables and 40 percent MCAR and 400 cases.

Discussion

In addition to reviewing existing tests, we have developed a number of new tests for the homogeneity of means and covariance matrices when data are missing completely at random. These tests are useful to determine whether a structural modeling analysis of a single overall mean vector or covariance matrix makes sense. Although all of the tests within a given class are asymptotically equivalent, their finite sample behavior was found to be remarkably different in a series of simulation studies. Essentially no prior empirical work of any kind, except that by Little (1988) on his means test, could be found for any of the extant tests. Of course none existed for the new tests.

The best test for the test of homogeneity of means for incomplete data was Little's (1988) ML-based test. This is also the GLS test. Little's test of homogeneity of means did not differ far from the ideal Type I error rate of five percent (see Figure 1). The test was very stable throughout sample sizes as well as percent MCAR and number of variables. The test behaved very well not only with large samples, but also with small samples.

For evaluating the homogeneity of covariance matrices for incomplete data, the best test of the null hypothesis was the MGLS test. The overall mean Type I error rate was very close to five percent (see Figure 2). However, this test produced inflated Type I error rates at 20 and 30 variables with an extreme percent MCAR (i.e., 40 and 50 percents). Evidently, covariance matrices are much harder to estimate with precision when the amount of missing data is quite large, as compared to the estimation of means. This test statistic performed very

well with 5 and 10 variables throughout all conditions of sample sizes and percent MCAR. For modest percents of MCAR (i.e., 10 and 20 percents) with 20 and 30 variables, the test produced slight underestimates of Type I error rates. However, since the number of variables with missing scores was held at a constant five variables, it is understandable that missing values would have less impact on data sets with more variables. The maximum Type I error rate of the MGLS test of homogeneity of covariance matrices for a modest percent MCAR was only 7.2 percent, slightly above the ideal five percent, while the standard deviation of the Type I error rates dropped from 2.45 percent to 1.48 percent.

The best performing test for evaluating the combined null hypothesis of homogeneity of means and covariance matrices was the new MWGLS test. This test is a combination of Little's test of homogeneity of means and the MGLS test of homogeneity of covariance matrices. The MWGLS test performed somewhere between the performances of its two components (see Figure 3). Although the error rates did vary somewhat (see Table 5), on average performance was remarkably better than that of the ML test proposed by Little (1988) and our GLS test (see Table 2). While our MGLS combined test already improved over these ML and GLS tests, the MWGLS test achieved still further improvements.

The empirical part of this research has some limitations. The number of variables with missing scores was held at a constant of five variables. So, the percent of variables with missing scores varied from a low of 17 percent to a high of 100 percent in the data sets. It is not known whether the test statistics will behave similarly for five variables containing five variables with missing scores (the condition that was studied) as compared to ten variables containing ten variables with missing scores (a condition not studied) at the same overall rate of missingness. A further limitation is that the amount of missingness in the variables was

held constant across variables within each data set. The behavior of these test statistics is not known for conditions in which variables may have different percent MCAR, for instance, one variable might have ten percent missing while an other variable might have 20 percent missing. It is expected that the test statistics would behave similarly to conditions with the same average missing amount of MCAR, and perhaps would not exceed the behavior observed here under highest MCAR. But only further research can evaluate this expectation. Finally, all the variables in this study were uncorrelated. It would be interesting to determine whether a similar pattern of results would be obtained where variables with missing scores are highly correlated with other variables.

The proposed test statistics developed in this study and Little's test of homogeneity of means provides researchers with greater latitude in exploring data with missing scores. While the MGLS test of homogeneity of covariance matrices provides researchers with a valuable tool to check for similarity of relationships among variables, the test remains a combined test of variances as well as covariances. Further work can separate this test into a test on homogeneity of correlations and homogeneity of variances. It is entirely possible, for example, that correlations are homogeneous while variances are not. Such differences cannot be determined by the omnibus test we developed here.

The current study only examined the test statistics for testing the null hypothesis that data are missing completely at random. There are currently no test statistics to evaluate the less restrictive hypothesis that data are missing at random. An interesting clash arises when considering these two types of hypotheses. It is possible that our tests would verify that a data set is not MCAR, that is, all means and covariances cannot be considered to be subset samples of a single population mean vector and covariance matrix. In that case, it

would seem prudent not to undertake SEM under a single modeling hypothesis. However, the MLE estimates that are obtained from the EM algorithm are claimed to be consistent estimates of a single set of population parameters when data are MAR. However, what these population parameters might be when the various patterns of data are not from the same population (assuming for simplicity no type I errors in the judgment) is not entirely clear. Certainly, further development is necessary to incorporate these contradictory conclusions and ensure confidence in any results based on missing data.

While not central to the proposes of studying missing data methodology, it is possible that some of the new statistics developed here also may apply to complete data contexts with multiple groups. For example, Lee and Tsui (1982) use the weight c_i, to generate a multiple sample GLS methodology for covariance structures. It is possible that weights based on k_i will perform better in small samples, though both have the same asymptotic limit. Similarly, it is possible that the MWGLS χ^2, test, based on equation (9), would do better in multisample mean and covariance structure analysis as compared to current methodology that is based on equation (5), which was not found to work well here. Of course, these are topics for further research.

References

Allison, P. D. (1987). Estimation of linear models with incomplete data. In C. Clogg (Ed.), *Sociological methodology 1987* (pp. 71-103). San Francisco: Jossey Bass.

Arbuckle, J. L. (1996). Full information estimation in the presence of incomplete data. In G. A. Marcoulides and R. E. Schumacker (eds.) *Advanced structural equation modeling: Issues and techniques* (pp. 243–277). New Jersey: Lawrence Erlbaum Associates.

Bentler, P. M. (1989). *EQS structural equations program manual.* Los Angeles: BMDP Statistical Software.

Browne, M. W. (1974). Generalized least squares estimators in the analysis of covariance structures. *South African Statistical Journal*, *8*, 1-24.

Chen, H. Y., & Little, R. (1999). A test of missing completely at random for generalised estimating equation with missing data. *Biometrika*, *86*, 1-13.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Serial B*, *39*, 1-38.

Dixon, W. J., ed. (1988). *BMDP statistical software.* Los Angeles: University of California Press.

Fuchs, C. (1982). Maximum likelihood estimation and model selection in contingency tables with missing data. *Journal of the American Statistical Association*, *77*, 270-278.

Jamshidian, M., & Bentler, P. M. (1999). ML estimation of mean and covariance structures with missing data using complete data routines. *Journal of Educational and Behavioral Statistics*, *24*, 21-41.

Lee, S. -Y., & Jennrich, R. I. (1979). A study of algorithms for covariance structure analysis with specific comparisons using factor analysis. *Psychometrika*, *44*, 99–114.

Lee, S. -Y., & Tsui, K. -L. (1982). Covariance structure analysis in several populations. *Psychometrika*, *47*, 297-308.

Little, R. J. A. (1988). A test of missing completely at random for multivariate data with missing values. *Journal of the American Statistical Association*, *83*, 1198-1202.

Little, R. J. A., & Rubin, D. B. (1987). *Statistical analysis with missing data.* New York: John Wiley & Sons.

Little, R. J. A., & Schenker, N. (1995). Missing Data. In G. Arminger, C. C. Clogg, & M. E. Sobel (eds.), *Handbook of statistical modeling for the social and behavioral sciences* (pp. 39-75). New York: Plenum Press.

Magnus, J. R., & Neudecker, H. (1988). *Matrix Differential Calculus with Application in Statistics and Econometrics.* Wiley, New York.

Muthén, B., Kaplan, D., & Hollis, M. (1987). On structural equation modeling with data that are not missing completely at random. *Psychometrika*, *52*(3), 431-462.

Odeh, R. E., & Evans, J. O. (1974). Algorithm AS 70. The percentage points of the normal distribution. *Applied Statistics*, *23*, 96 - 97.

Rovine, M. J. (1994). Latent variables models and missing data analysis. In A. von Eye and C. C. Clogg (eds.) *Latent variables analysis: Applications for developmental research* (pp. 181-225). Thousand Oaks, CA: Sage.

Schott, J. (1997). *Matrix analysis for statistics.* Wiley, New York.

Tabachnick, B. G., & Fidell, L. S. (1996). *Using multivariate statistics, 3rd ed.* New York: Harper Collins Publishers Inc.

Tang, M., & Bentler, P. M. (1998). Theory and method for constrained estimation in structural equation models with incomplete data. *Computational Statistics & Data Analysis*, *27*, 257-270.

Appendix

Asymptotic Distribution

There are m missing patterns with sample mean vector $\bar{X}_i$ and sample covariance matrix S_i based on sample size n_i. The population counterparts are μ_i and Σ_i. Let vec$(\cdot)$ be an operator which transforms a matrix into a vector by stacking the columns of the matrix and let vech$(\cdot)$ be an operator which transforms a symmetric matrix into a vector by picking the nonduplicated elements of the matrix, $s_i = \text{vech}(S_i)$ and $\sigma_i = \text{vech}(\Sigma_i)$, then

$$\sqrt{n_i}(s_i - \sigma_i) \xrightarrow{\mathcal{L}} N(0, \Gamma_i), \tag{A1}$$

where $\Gamma_i = 2D_{p_i}^{+}(\Sigma_i \otimes \Sigma_i)D_{p_i}^{+\prime}$; duplication matrix D_{p_i} is a $(p_i^2 \times \frac{p_i(p_i+1)}{2})$ such that $\text{vec}(\Sigma_i) = D_{p_i}\sigma_i$ (Magnus and Neudecker (1988, p. 49); Schott (1997, p. 283)) and $D_{p_i}^{+}$ is the Moore-Penrose inverse of D_{p_i}. Let $z_i = (\bar{X}_i', s_i')'$ and $\xi_i = (\mu_i', \sigma_i')'$, then

$$\sqrt{n_i}(z_i - \xi_i) \xrightarrow{\mathcal{L}} N(0, \Pi_i), \tag{A2}$$

where $\Pi_i = \text{diag}(\Sigma_i, \Gamma_i)$. In missing data SEM applications, of interest is the mean and covariance structure $(\mu(\theta), \Sigma(\theta))$. So each $(\mu_i(\theta), \sigma_i(\theta))$ is a subvector of $(\mu(\theta), \Sigma(\theta))$.

Let $N = n_1 + \cdots + n_m$, $c_i = \frac{n_i}{N} \to \gamma_i$, S_i be of dimension p_i with $p_i^* = p_i + p_i(p_i+1)/2$, $p^* = p_1^* + \cdots + p_m^*$, and q be the number of unknown parameters in θ. Let

$$L(\theta) = \frac{1}{N}\sum_{i=1}^{m} n_i L_i(\theta),$$

where

$$L_i(\theta) = (\bar{X}_i - \mu_i(\theta))'\Sigma_i^{-1}(\theta)(\bar{X}_i - \mu_i(\theta)) + [\text{tr}(S_i\Sigma_i^{-1}(\theta)) - \log|S_i\Sigma_i^{-1}(\theta)| - p_i].$$

First we will consider the distribution of estimator $\hat{\theta}$ by minimizing $L(\theta)$. We will use dot on top of a function to imply derivative, e.g,

$$\dot{L}(\theta) = \frac{\partial L(\theta)}{\partial \theta}, \quad \ddot{L}(\theta) = \frac{\partial^2 L(\theta)}{\partial \theta \partial \theta'}.$$

We may omit the argument of a function if evaluated at the population value, e.g., $\sigma = \sigma(\theta_0)$.

Since $\hat{\theta}$ satisfies $\dot{L}(\hat{\theta}) = 0$, using a Taylor expansion on $\dot{L}(\hat{\theta}) = 0$ we obtain

$$\sqrt{N}(\hat{\theta} - \theta_0) = -\ddot{L}^{-1}(\theta_0)\sqrt{N}\dot{L}(\theta_0) + o_p(1), \tag{A3}$$

where $o_p(1)$ will approach 0 in probability when all the n_i approach infinity. Now let

$$W_i = \begin{pmatrix} \Sigma_i^{-1} & 0 \\ 0 & W_{2i} \end{pmatrix}$$

with

$$W_{2i} = \frac{1}{2} D'_{p_i} (\Sigma_i^{-1} \otimes \Sigma_i^{-1}) D_{p_i}.$$

Then

$$\ddot{L} = \frac{1}{N} \sum_{i=1}^{m} n_i \ddot{L}_i \xrightarrow{P} 2[\sum_{i=1}^{m} \gamma_i (\dot{\xi}_i' W_i \dot{\xi}_i)]. \tag{A4}$$

Let $e_i = \sqrt{n_i}(z_i - \xi_i)$, then

$$\sqrt{N}\dot{L}(\theta_0) = -2 \sum_{i=1}^{m} \sqrt{\gamma_i} \dot{\xi}_i' W_i e_i + o_p(1). \tag{A5}$$

Notice that since Γ_i is the inverse of W_{2i}, we have $\Pi_i^{-1} = W_i$. It follows from (A3) to (A5) that

$$\sqrt{N}(\hat{\theta} - \theta_0) \xrightarrow{\mathcal{L}} N(0, \Omega),$$

where

$$\Omega = [\sum_{i=1}^{m} \gamma_i (\dot{\xi}_i' W_i \dot{\xi}_i)]^{-1}.$$

Mixed Weight GLS Test Statistic

Remember from equation (9) that

$$F_3(\hat{\theta}) = \sum_{i=1}^{m}\{c_i(\bar{X}_i - \mu_i(\hat{\theta}))'\Sigma_i^{-1}(\hat{\theta})(\bar{X}_i - \mu_i(\hat{\theta})) + \frac{k_i}{2}\mathrm{tr}[(S_i - \Sigma_i(\hat{\theta}))\Sigma_i^{-1}(\hat{\theta})]^2\}.$$

Both constants $c_i = \frac{n_i}{N} \rightarrow \gamma_i$ and $k_i = \frac{n_i-(1-c_i)}{N} \rightarrow \gamma_i$ so the differences in weights are unimportant in the limit. Hence, the following also applies to $G_3(\hat{\theta})$ in equation (5) and $M_3(\hat{\theta})$ in equation (8). Let

$$\begin{aligned} W &= \mathrm{diag}(W_1, \cdots, W_m), \\ W_\gamma &= \mathrm{diag}(\gamma_1 W_1, \cdots, \gamma_m W_m), \\ \xi &= (\xi_1', \cdots, \xi_m')' \\ e &= (e_1', \cdots, e_m')'. \end{aligned}$$

Then the asymptotic distribution of $W^{\frac{1}{2}}e$ is given by

$$W^{\frac{1}{2}}e \xrightarrow{\mathcal{L}} N(0, I). \qquad (A6)$$

After simplification we can write $T = NF_3(\hat{\theta})$ as

$$T = [W^{\frac{1}{2}}e - W_\gamma^{\frac{1}{2}}\sqrt{N}(\hat{\xi} - \xi)]'[W^{\frac{1}{2}}e - W_\gamma^{\frac{1}{2}}\sqrt{N}(\hat{\xi} - \xi)] + o_p(1) \qquad (A7)$$

Based on (A4) and (A5) we can rewrite (A3) as

$$\sqrt{N}(\hat{\theta} - \theta_0) = (\dot{\xi}'W_\gamma\dot{\xi})^{-1}\dot{\xi}'W_\gamma^{\frac{1}{2}}W^{\frac{1}{2}}e + o_p(1),$$

which leads to

$$\sqrt{N}(\hat{\xi} - \xi) = \dot{\xi}(\dot{\xi}'W_\gamma\dot{\xi})^{-1}\dot{\xi}'W_\gamma^{\frac{1}{2}}W^{\frac{1}{2}}e + o_p(1).$$

So

$$W^{\frac{1}{2}}e - W_{\gamma}^{\frac{1}{2}}\sqrt{N}(\hat{\xi} - \xi) = Q(W^{\frac{1}{2}}e) + o_p(1), \tag{A8}$$

where

$$Q = I - W_{\gamma}^{\frac{1}{2}}\dot{\xi}(\dot{\xi}'W_{\gamma}\dot{\xi})^{-1}\dot{\xi}'W_{\gamma}^{\frac{1}{2}}.$$

Notice that Q is a projection matrix. Thus(A7) and (A8) lead to

$$T = (W^{\frac{1}{2}}e)'Q(W^{\frac{1}{2}}e) + o_p(1). \tag{A9}$$

Since the rank of Q is $d_3 = p^* - q$, it follows from (A6) and (A9) that

$$T \xrightarrow{\mathcal{L}} \chi^2_{d_3}.$$

Note that this result holds for GLS, MGLS, and MWGLS methods equally. Of course, this is an asymptotic result, and the finite sample behavior of the various tests may be different.

TABLE 1

Mean Number of Incomplete Patterns Generated Averaged across Number of Variables.

	Percent MCAR				
Sample Size	10%	20%	30%	40%	50%
100	12	19	25	29	30
200	15	23	29	31	32
300	17	26	30	32	32
400	18	27	31	32	32
500	19	28	31	32	32
1000	22	30	32	32	32
1500	24	31	32	32	32

TABLE 2

Overall Type I Error Rates of Test of Assumptions for MCAR Averaged across Number of Variables, Sample Size, and Percent MCAR.

	MCAR Test	Mean	Standard Deviation	Minimum	Maximum
Mean	ML	3.67	0.97	1.0	6.2
	MGLS	0.36	0.48	0.0	3.7
Covariance	ML	85.15	12.37	8.0	100.0
	GLS	54.43	15.92	3.8	100.0
	MGLS	5.17	2.45	0.2	15.4
Combined	ML	82.78	12.99	7.6	100.0
	GLS	52.53	15.76	3.4	100.0
	MGLS	2.42	1.44	0.1	15.6
	MWGLS	4.34	2.09	0.2	14.3

TABLE 3

Mean Type I Error Rates of ML Test of Homogeneity of Means for Incomplete Data Averaged across Sample Sizes.

Number of variables	Percent MCAR	Mean	Standard Deviation	Minimum	Maximum
5	10	4.4	0.9	3.3	6.2
	20	4.3	1.0	2.8	5.7
	30	4.3	0.9	3.1	5.5
	40	3.7	0.9	2.1	4.8
	50	3.5	1.3	1.0	4.8
10	10	4.2	0.8	3.0	5.4
	20	4.3	1.2	2.8	5.9
	30	3.8	0.9	2.1	4.9
	40	3.7	1.1	1.4	4.5
	50	3.5	1.3	1.5	5.1
20	10	4.0	1.2	1.6	5.4
	20	3.3	0.6	2.2	4.0
	30	3.9	1.4	1.5	5.5
	40	3.4	1.2	1.5	4.9
	50	3.6	0.9	2.8	5.3
30	10	4.0	0.9	2.8	5.2
	20	3.7	0.7	2.6	4.4
	30	4.0	0.8	3.1	5.0
	40	3.9	1.3	2.4	5.2
	50	3.5	1.0	2.3	4.8

TABLE 4

Mean Type I Error Rates of the MGLS Test of Homogeneity of Covariance Matrices for Incomplete Data Averaged across Sample Sizes.

Number of variables	Percent MCAR	Mean	Standard Deviation	Minimum	Maximum
5	10	4.7	1.1	3.6	6.8
	20	5.3	0.9	3.5	6.3
	30	5.2	0.4	4.4	5.7
	40	4.3	1.2	2.5	6.1
	50	3.8	1.3	2.6	6.0
10	10	2.8	1.3	1.7	5.1
	20	3.7	1.7	1.7	6.1
	30	5.3	2.0	1.6	7.3
	40	5.4	2.9	2.2	10.0
	50	5.1	2.4	2.6	9.2
20	10	1.6	1.0	0.3	3.4
	20	2.7	2.6	0.3	7.2
	30	6.8	4.0	0.4	11.1
	40	7.8	4.3	1.4	12.7
	50	7.9	4.7	2.9	15.1
30	10	1.2	0.8	0.2	2.5
	20	2.5	2.4	0.3	6.2
	30	7.8	5.0	1.4	13.7
	40	11.2	4.2	5.9	15.4
	50	8.2	4.7	2.6	14.4

TABLE 5

Mean Type I Error Rates of the MWGLS Test of Homogeneity of Means and Covariance Matrices for Incomplete Data Averaged across Sample Sizes.

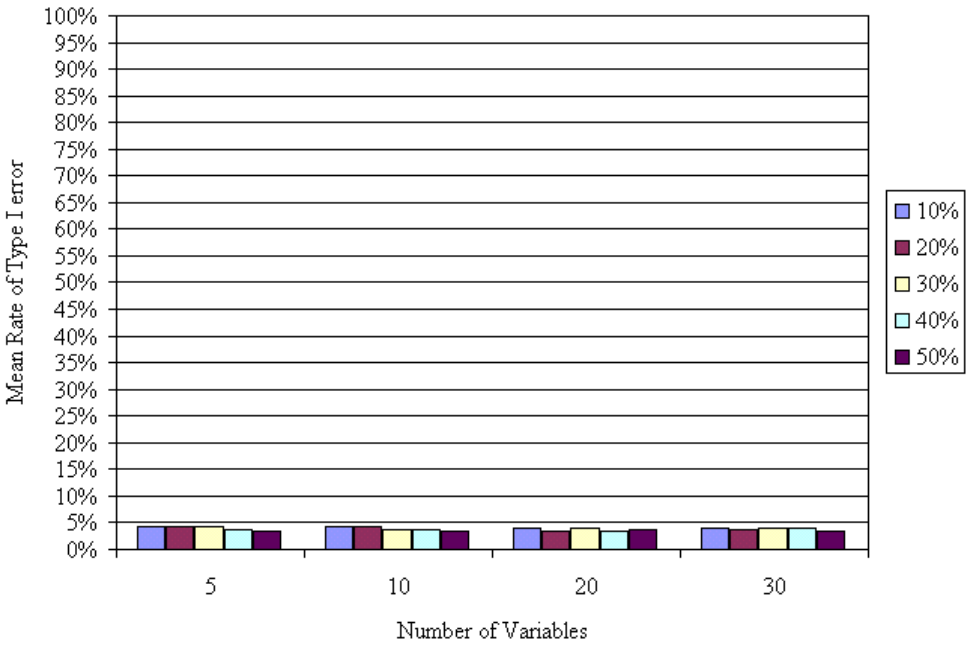
Number of variables	Percent MCAR	Mean	Standard Deviation	Minimum	Maximum
5	10	4.1	0.8	3.3	5.6
	20	4.2	0.8	2.6	5.1
	30	3.8	0.4	2.9	4.2
	40	3.0	0.4	2.5	3.5
	50	2.8	0.5	2.1	3.6
10	10	2.6	1.1	1.8	4.5
	20	3.4	1.6	1.5	5.9
	30	4.0	1.6	1.1	5.7
	40	4.1	2.1	1.6	7.6
	50	3.6	1.7	1.6	6.1
20	10	1.4	1.0	0.2	3.1
	20	2.5	2.4	0.2	6.5
	30	6.3	3.7	0.3	9.9
	40	6.6	3.6	1.4	10.6
	50	6.4	4.2	2.0	13.0
30	10	1.1	0.8	0.2	2.4
	20	2.4	2.3	0.2	5.6
	30	7.4	4.8	1.4	13.0
	40	10.3	3.7	5.3	14.3
	50	6.7	4.2	2.3	12.9

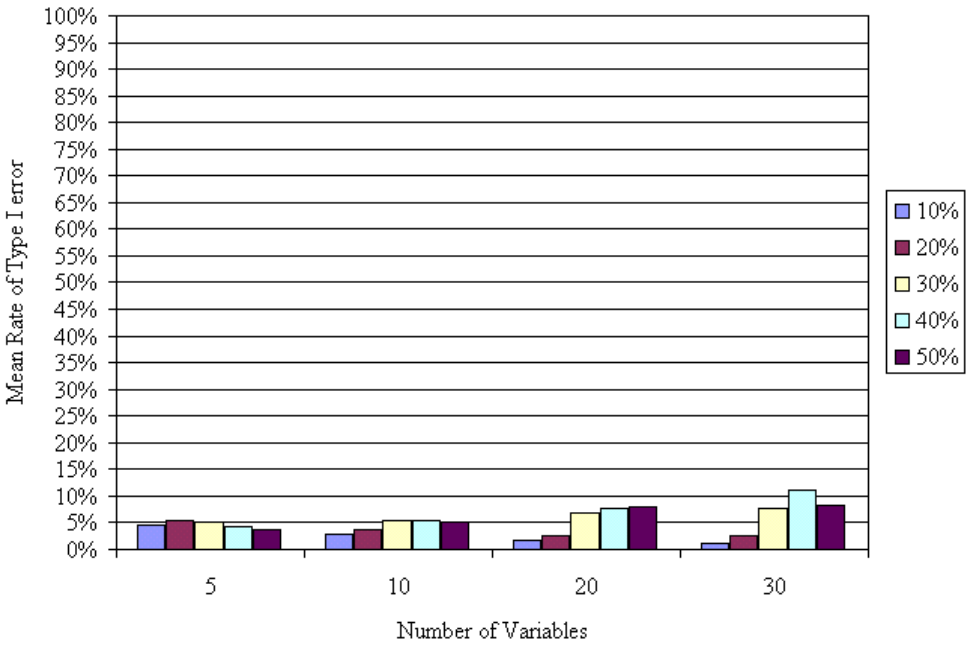
Figure Caption

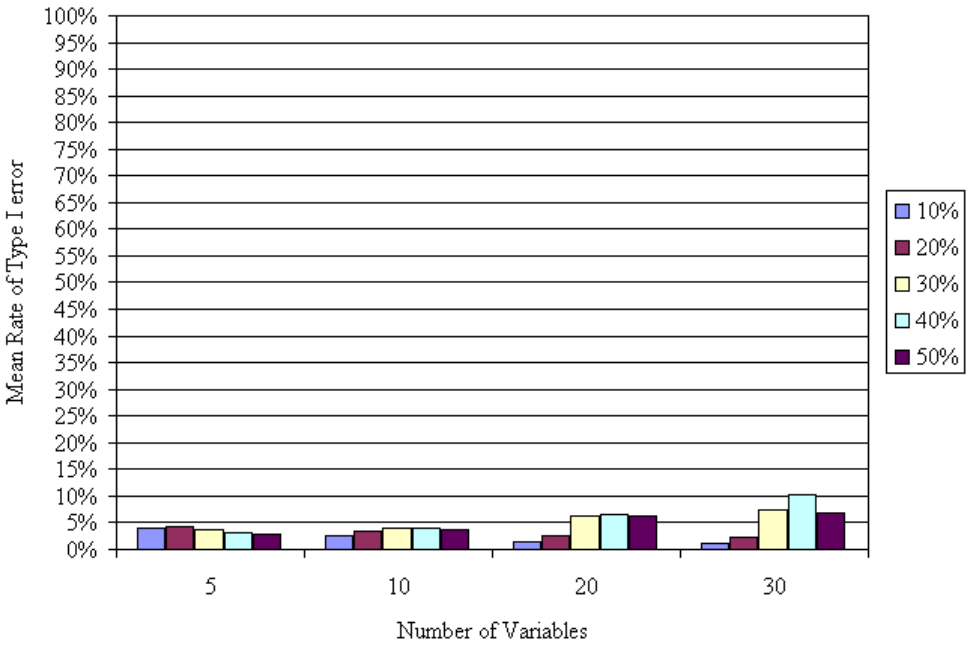
Figure 1. Mean Type I error rates for ML test of homogeneity of means as a function of number of variables and percent MCAR.

Figure 2. Mean Type I error rates for the MGLS test of homogeneity of covariance matrices as a function of number of variables and percent MCAR.

Figure 3. Mean Type I error rates for the MWGLS test of homogeneity of means and covariance matrices as a function of number of variables and percent MCAR.


100%
95%
90%
85%
80%
75%
70%
65%
60%
55%
50%
45%
40%
35%
30%
25%
20%
15%
10%
5%
0%
Mean Rate of Type I error
5
10
20
30
Number of Variables
10%
20%
30%
40%
50%


100%
95%
90%
85%
80%
75%
70%
65%
60%
55%
50%
45%
40%
35%
30%
25%
20%
15%
10%
5%
0%
Mean Rate of Type I error
5
10
20
30
Number of Variables
10%
20%
30%
40%
50%


100%
95%
90%
85%
80%
75%
70%
65%
60%
55%
50%
45%
40%
35%
30%
25%
20%
15%
10%
5%
0%
Mean Rate of Type I error
5
10
20
30
Number of Variables
10%
20%
30%
40%
50%