A Simple Approach to Inference in Covariance Structure Modeling With Missing Data: Bayesian Analysis

CSE Technical Report 411

Bengt Muthén
CRESST/University of California, Los Angeles

May 1996

National Center for Research on Evaluation,
Standards, and Student Testing (CRESST)
Graduate School of Education & Information Studies
University of California, Los Angeles
Los Angeles, CA 90095-1522
(310) 206-1532

The work reported herein was supported under the Educational Research and Development Center Program, cooperative agreement number R117G10027 and CFDA catalog number 84.117G, as administered by the Office of Educational Research and Improvement, U.S. Department of Education.

A SIMPLE APPROACH TO INFERENCE IN COVARIANCE STRUCTURE MODELING WITH MISSING DATA: BAYESIAN ANALYSIS[1]

Bengt Muthén
CRESST/University of California, Los Angeles

Introduction

Well-known educational achievement studies such as the National Assessment of Educational Progress (NAEP), the National Education Longitudinal Study (NELS), and the Longitudinal Study of American Youth (LSAY) often exhibit complex patterns of missing data, both due to design and involuntarily, for example, due to attrition of students. Muthén, Kaplan, and Hollis (1987) showed that a wide variety of analyses with missing data can be performed using existing covariance structure software such as LISREL and LISCOMP. Muthén et al. (1987) used standard missing data theory (Little & Rubin, 1987) to solve the problem as a multiple-group analysis with one group per missing data pattern. This methodology has recently been applied in the context of multidimensionality of achievement by Muthén, Khoo, and Nelson Goff (1994) and achievement growth modeling by Muthén (1994).

Due to the method of data collection used in typical educational achievement studies, however, the mechanics of carrying out these analyses is rather awkward. For example, in NAEP no student takes all the achievement test items that are used (BIB spiraling of test forms) resulting in 26 different subgroups of students. In LSAY, there are ten patterns due to an adaptive testing design, where the test form a student obtains in one grade depends on his/her performance the previous grade, but many more patterns arise due to students missing from one or more of the testing occasions. To analyze data using existing covariance structure software, each pattern needs to be considered as a separate group. The more groups there are, the more complicated the analysis is. These practical difficulties may hinder widespread adoption of the new techniques because of the extensive control language that needs to be used.

Achievement item analyses typically use IRT techniques to estimate abilities from the item responses. These techniques deal with missing data well

[1] I thank Ginger Nelson Goff for expert assistance.

using standard missing data theory. However, IRT software is limited in its handling of complex modeling such as multifactorial analysis and growth analysis. In contrast, current covariance structure software for continuous variables, such as test scores, is flexible in handling the modeling of multifactorial analysis and growth but is limited in its handling of missing data. Following are examples of the limitations of IRT software that can be avoided by using covariance structure software for continuous variables.

When modeling growth, typical IRT software produces a score for a single dimension for each time point. These scores are then analyzed using growth modeling. For example, in LSAY, NAEP math items are used to form a unidimensional scale across Grades 7 to 12 which can then be modeled. Growth studies therefore become limited to the study of that single general dimension even though interesting growth may take place with respect to other, more specific dimensions. In contrast, using covariance structure software for continuous variables, several dimensions can be modeled over time simultaneously.

Even when IRT software produces multiple dimensions, these dimensions cannot be specified in a flexible way. For example, one cannot specify a confirmatory factor model as can be done in covariance structure software. In addition, specialized IRT software used for NAEP produces scores based on conditioning variables such as gender, ethnicity, etc., limiting the possibility of using these and related variables as covariates in later analyses. Public access tapes for NAEP provide plausible values which are IRT scores deduced from test results as well as from student-related background information (conditioning variables). While scores on several dimensions are produced, these dimensions are not the only ones of interest, and the validity of the scores for studying relationships to variables related to the conditioning variables is not clear. In contrast, covariance structure software allows for flexible modeling of the various dimensions, and because individual scores are not estimated, the problem of using the conditioning variables in later analyses is avoided.

The use of covariance structure software for continuous variables requires the creation of item parcels that can be seen as fallible indicators of latent variables. Using item parcels, complex modeling such as the analysis of growth is well defined for continuous variables by using latent variable covariance structure modeling. The use of such approaches provides a way to validate findings based on IRT scores and enables the investigator to go beyond the analyses that are

possible with IRT scores. But this is where the problem comes in: Covariance structure modeling is presently not well suited to handle missing data in an easy fashion.

This paper investigates methods that avoid using multiple groups to represent the missing data patterns in covariance structure modeling. The aim is instead to do a single-group analysis where the only action the analyst has to take is to indicate missingness in his/her data. The simplification would be a big step forward for data analysts and may make the difference between an investigator actually choosing to attempt the analysis or not, particularly in the context of a complex analysis such as growth modeling.

One possibility for providing such an approach is to base the analysis on means and covariances that are created while properly taking missing data into account. Missing data theory exists for getting such sample statistics, and one can develop appropriate model testing and standard error procedures. Such a method is not fully efficient, however, in that the resulting estimates are not maximum-likelihood (ML). Instead, a new covariance structure approach developed by Muthén and Arminger (1994) will be utilized. This approach draws on Bayesian theory and is a full-information estimator as is ML estimation. The approach promises to be especially useful in small-sample situations. It is of interest to compare the Muthén-Arminger approach to that of the maximum-likelihood techniques based on multiple-group analysis.

In the second section of this paper, the proposed methodology is briefly described in a nontechnical way. The third section presents tests of the performance of this approach on simulated data under various forms of missing data. The last section concludes.

A Bayesian Approach to Missing Data

With continuous observed variables, the covariance structure model describes the latent variables as predictors of the observed variables, the measurements or indicators, using linear regressions. The parameters consist of the measurement parameters describing these linear regressions and the structural parameters describing the distribution of the latent variables. This results in three sets of unknowns for the Bayesian analysis: the latent variable values, the measurement parameter values, and the structural parameter

values. The chosen analysis approach is Bayesian in that the posterior distribution for the unknowns given the observed data will be considered. In this framework, parameters are viewed as random variables and the posterior distribution describes the likely location and uncertainty of the parameter values.

Recently, Muthén and Arminger (1994) studied Bayesian estimation in latent variable regression models. The computations in such analyses are often very cumbersome. To reduce the computational burden, Muthén and Arminger (1994) used the Gibbs sampler Markov Chain Monte Carlo algorithm (see, e.g., Tanner, 1993). This allows the posterior distribution to be described by a series of random draws from a set of simple conditional distributions. This paper focuses on the special case of Muthén and Arminger (1994) where all observed variables are continuous and are seen as indicators of latent variables as in conventional covariance structure modeling. The simulation study expands on the simulations of Muthén and Arminger (1994), which mostly concerned binary response variables and simpler forms of missing data.

There are three conditional distributions involved in the Gibbs calculations with continuous indicators. One is the conditional distribution of the latent variables given the observed variable values (the data), the measurement parameters, and the structural parameters. The second is the conditional distribution of the measurement parameters given the latent variable values, the data, and the structural parameters. The third is the conditional distribution of the structural parameters given the latent variable values, the data, and the measurement parameters.

The Gibbs sampler algorithm simply goes through these three steps repeatedly, drawing random values from each of the conditional distributions in turn. At the end of a sufficient number of cycles, the marginal posterior distribution for each of the three types of unknowns is obtained. In particular, the distribution for each of the model parameters is obtained. The mean, or the mode, of the distribution may be taken as the counterpart of a conventional, frequentist, parameter estimate. The standard error of the distribution may be taken as the standard error of the parameter estimate.

The first conditional distribution is the one commonly used to estimate "factor scores." The model assumes that this distribution is multivariate normal with a certain known mean vector, depending on the data, and covariance matrix

and is therefore easy to sample from. Given values for the latent variables obtained from this first step, the values for the measurement parameters of the second step are easy to describe given that they correspond to regression intercepts and slopes in regressions of observed dependent variables (the data) on observed predictors (the latent variables). It can be shown that they have a multivariate normal distribution. The third step can often be avoided entirely by setting the metric of the latent variables by choosing unit factor variances, having orthogonal factors, and standardizing the factor means to zero. The structural parameters of the factor covariance can otherwise be obtained by random draws from an inverse chi-square distribution or an inverse Wishart distribution (see Muthén & Arminger, 1994).

When data are missing for a certain observed variable, these missing values are simply viewed as yet another set of unknowns in the Bayesian framework. The Gibbs algorithm is extended so that data are generated for these variables by random draws from another normal conditional distribution. With the conventional assumption of uncorrelated measurement errors, the conditional distribution drawn from is a function of the latent variable values and the measurement parameters.

As in Muthén and Arminger (1994), the Gibbs sampler is here applied by first going through 500 "burn-in" cycles and then recording the values from the next 2,000 cycles.

A Monte Carlo Study

A Monte Carlo study was carried out for a covariance structure model with missing data. A single-factor model with ten continuous indicators was chosen. The sample size was set at 500. Two cases of missing data were studied, randomly missing data and selectively missing data. Data were generated and analyzed over ten replications.

The first missing data situation corresponds to "missing completely at random," or MCAR using terminology from the missing data literature (see, e.g., Little & Rubin, 1987). Here, data are missing randomly for the last five observed variables. The second situation corresponds to "missing at random," or MAR. The MAR term implies that estimation which ignores the missing data mechanism is correct even if data are missing selectively. This holds as long as the missingness

depends on observed variables in the model for which no data are missing. In this case, data are missing on the last five observed variables if the first observed variable has a value less than zero. This type of selective missingness is common in, for example, longitudinal studies where attrition at later time points is a function of the values of variables observed at earlier time points. It is interesting to note that this selective missingness does not invalidate the estimation of the parameters. In both the MCAR and MAR case, data are missing for half of the sample.

The measurement parameter values chosen for the intercepts (nu) and slopes (lambda) are given in Table 1 (the MCAR case) and Table 2 (the MAR case). The factor variance is two and the factor mean is zero. The measurement error variances are all one. The reliability for each observed variable is therefore 0.5. The "Probn" column gives tests of univariate normality for each marginal posterior distribution. The "Lw" and "Up" columns give lower and upper limits for the counterpart to conventional 90% confidence intervals for each parameter, while the column "Cover" gives the proportion of the ten replications for which the intervals cover the true value (these should be 0.9).

The MCAR case of Table 1 shows that the Bayesian approach works very well despite the fact that data are missing for half of the variables for half of the sample. The posterior parameter distributions appear close to normal and the parameter estimates (e.g., taken at the mean) show no bias. The posterior variation (measured by "Std") is only slightly increased for variables 6-10 as compared to variables 1-5.

The MAR case of Table 2 shows that the Bayesian approach also works very well with data that are missing selectively. There is still no parameter bias. The parameter variation is now slightly higher for the parameters associated with variables with missing data (6-10), but the increase is not dramatic. The coverage is about as good in the MAR case as in the MCAR case.

Table 1

Averages of 10 Monte Carlo Replications From Bayesian Analysis Using the Gibbs Sampler (Sample size = 500; Model: p=10 cont. vars., m=1, q=0; Gibbs: N=2000, burn-in=500; Missing data for first 250 obs. for Y6 to Y10)

	Method	Gibbs Sampler						
PARA	True	Mean	Median	Std	Probn	Lw	Up	Cover
NU1	0.000	0.037	0.037	0.064	0.362	-0.068	0.142	0.700
NU2	0.000	0.035	0.035	0.063	0.559	-0.069	0.140	0.800
NU3	0.000	0.038	0.037	0.064	0.317	-0.067	0.143	0.900
NU4	0.000	0.018	0.018	0.063	0.632	-0.085	0.122	0.900
NU5	0.000	0.043	0.043	0.062	0.564	-0.060	0.146	0.800
NU6	0.000	0.001	0.000	0.081	0.278	-0.130	0.135	0.800
NU7	0.000	0.013	0.012	0.079	0.452	-0.117	0.143	1.000
NU8	0.000	0.051	0.051	0.081	0.480	-0.082	0.184	0.900
NU9	0.000	0.018	0.020	0.082	0.445	-0.117	0.151	0.800
NU10	0.000	-0.005	-0.005	0.080	0.268	-0.138	0.126	0.900
MEAN	0.000	0.025	0.025	0.072	0.436	-0.093	0.143	0.850
LA1	0.700	0.714	0.713	0.043	0.475	0.645	0.784	1.000
LA2	0.700	0.719	0.718	0.042	0.385	0.651	0.790	1.000
LA3	0.700	0.711	0.711	0.042	0.418	0.643	0.781	1.000
LA4	0.700	0.706	0.705	0.042	0.275	0.638	0.777	0.900
LA5	0.700	0.712	0.712	0.042	0.462	0.645	0.781	1.000
LA6	0.700	0.746	0.745	0.056	0.236	0.656	0.839	0.900
LA7	0.700	0.726	0.725	0.054	0.393	0.638	0.817	0.700
LA8	0.700	0.729	0.729	0.055	0.380	0.640	0.821	0.800
LA9	0.700	0.712	0.712	0.055	0.456	0.621	0.803	0.900
LA10	0.700	0.735	0.734	0.055	0.293	0.646	0.826	0.700
MEAN	0.700	0.721	0.720	0.049	0.377	0.642	0.802	0.890

Table 2

Averages of 10 Monte Carlo Replications From Bayesian Analysis Using the Gibbs Sampler (Sample size = 500; Model: p=10 cont. vars., m=1, q=0; Gibbs: N=2000, burn-in=500; Missing data for Y6 to Y10 if Y1 less than or equal to zero)

	Method	Gibbs Sampler						
PARA	True	Mean	Median	Std	Probn	Lw	Up	Cover
NU1	0.000	-0.013	-0.013	0.062	0.484	-0.114	0.089	1.000
NU2	0.000	0.012	0.013	0.062	0.331	-0.089	0.114	1.000
NU3	0.000	0.026	0.026	0.063	0.374	-0.078	0.130	1.000
NU4	0.000	0.010	0.010	0.062	0.657	-0.093	0.111	0.900
NU5	0.000	0.027	0.027	0.062	0.428	-0.075	0.130	0.800
NU6	0.000	-0.035	-0.033	0.095	0.348	-0.196	0.120	1.000
NU7	0.000	0.022	0.024	0.094	0.195	-0.137	0.174	0.700
NU8	0.000	0.005	0.006	0.092	0.182	-0.149	0.152	1.000
NU9	0.000	-0.015	-0.015	0.094	0.428	-0.169	0.139	0.800
NU10	0.000	-0.006	-0.005	0.095	0.127	-0.162	0.147	0.900
MEAN	0.000	0.003	0.004	0.078	0.355	-0.126	0.131	0.910
LA1	0.700	0.694	0.693	0.042	0.524	0.625	0.763	1.000
LA2	0.700	0.684	0.684	0.041	0.350	0.618	0.752	1.000
LA3	0.700	0.691	0.690	0.042	0.579	0.622	0.761	0.900
LA4	0.700	0.701	0.700	0.041	0.388	0.633	0.770	0.900
LA5	0.700	0.710	0.709	0.041	0.428	0.643	0.778	1.000
LA6	0.700	0.744	0.743	0.067	0.247	0.634	0.856	0.800
LA7	0.700	0.697	0.698	0.067	0.188	0.588	0.807	0.800
LA8	0.700	0.693	0.691	0.066	0.301	0.586	0.802	1.000
LA9	0.700	0.691	0.691	0.067	0.403	0.582	0.804	0.900
LA10	0.700	0.687	0.686	0.067	0.221	0.578	0.799	0.900
MEAN	0.700	0.699	0.699	0.054	0.363	0.611	0.789	0.920

Conclusions

The Bayesian approach appears very promising for missing data covariance structure modeling. First of all, it results in good properties for the parameter estimates. Second, it provides an analysis method that is very easy to use. The ease comes from the fact that a multiple-group approach is not necessary, but all the user has to specify is what is customary, namely a missing value code for observations that are missing. The technique is, however, not yet available in covariance structure software. Developments are taking place for remedying this.

References

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

Muthén, B. (1994, August). *Latent variable modeling of longitudinal and multilevel data*. Presented at the annual meeting of the American Sociological Association, Section on Methodology, Showcase Session, Los Angeles.

Muthén, B., & Arminger, G. (1994). *Bayesian latent variable regression for binary and continuous response variables using the Gibbs sampler* (Technical Report). Los Angeles: University of California, Graduate School of Education & Information Studies.

Muthén, B., Kaplan, D., & Hollis, M. (1987). On structural equation modeling with data that are not missing completely at random. *Psychometrika*, *42*, 431-462.

Muthén, B., Khoo, S. T., & Nelson Goff, G. (1994). *Multidimensional description of subgroup differences in mathematics achievement data from the 1992 National Assessment of Educational Progress* (Technical Report). Los Angeles: University of California, Graduate School of Education & Information Studies.

Tanner, M. A. (1993). Tools for statistical inference. *Methods for exploration of posterior distributions and likelihood functions* (2nd ed.). New York: Springer-Verlag.