Using Mplus Monte Carlo Simulations In Practice: A Note On Non-Normal Missing Data In Latent Variable Models

Bengt Muthén
University of California, Los Angeles

Tihomir Asparouhov
Muthén & Muthén *

Mplus Web Notes: No. 2
Version 2, March 22, 2002

*The research of the first author was supported under grant K02 AA 00230-01 from NIAAA. The email address is bmuthen@ucla.edu.

1 Introduction

This note discusses the use of Mplus Monte Carlo simulations to study parameter estimates, standard errors, and coverage in latent variable modeling in the common situation of having both non-normal data and missing data. The results indicate that satisfactory results can be obtained using normality-based maximum-likelihood estimation with robust standard errors in Mplus. The presentation builds on an earlier note, Muthén (2002a), which introduces Mplus Monte Carlo simulations. Readers not familiar with Mplus mixture Monte Carlo studies may consult the earlier note. The current note introduces more advanced features. As with the first note, all examples can be run using the free Mplus Demo version available at www.statmodel.com/mplus/demo.html. A summary of the Mplus language can be obtained at www.statmodel.com/mplus/language.html.

2 Brief Technical Background

It is well-known in the literature that maximum-likelihood estimation under the assumption of multivariate normality produces good parameter estimates even when data are strongly non-normal, but may give underestimated standard errors and inflated chi-square leading to too frequent rejections. For a good recent overview, see Enders (2001). Model tests of fit will not be discussed in this note. To protect against non-normality, standard error computations may be carried out using robust versions. Such robust standard errors are available in Mplus using the MLM estimator (see Muthén & Muthén, 1998-2001; Appendix 4, p. 357), and also using the MLR estimator in the mixture track (see Muthén & Muthén, 1998-2001; Appendix 8, p. 370). The MLM estimator does not allow missing data in the current Mplus version. The MLR estimator does allow missing data and can be used also in non-mixture situations using a single-class mixture analysis. This approach will be studied here.

Maximum-likelihood (ML) methods to deal with missing data (Little & Rubin, 1987) are typically based on assumptions of normality, although procedures based on t distributions have also been developed to protect against outliers. With normal data, the normality-based ML approach retains consistency of estimates with missing data under both the MCAR and MAR assumptions. Less is known about the properties of the normality-based ML approach in the more realistic case of MAR missingness with non-normal data. A recent discussion of related issues is given in Yuan and Bentler (2000) referring to the normality-based ML approach for non-normal data discussed in Arminger and Sobel (1990) under the term pseudo ML. The pseudo ML standard error computation is also in line with standard results on extremum estimators as given in Amemiya (1985, chapter 4) and is used by the Mplus MLR estimator. Writing the observed-data log likelihood as

$$logL = \sum_{i=1}^{n} logL_i, \tag{1}$$

and defining

$$\mathbf{B} = \sum_{i=1}^{n} \frac{\partial log L_i}{\partial \boldsymbol{\pi}} \times \frac{\partial log L_i}{\partial \boldsymbol{\pi}'}, \tag{2}$$

and

$$\mathbf{A} = -\sum_{i=1}^{n} \frac{\partial^2 log L_i}{\partial \boldsymbol{\pi}\, \partial \boldsymbol{\pi}'}, \tag{3}$$

the MLR approach approximates the Fisher information matrix using

$$I_{MLR} = \mathbf{A}^{-1}\, \mathbf{B}\, \mathbf{A}^{-1}. \tag{4}$$

This approach is discussed in Yuan and Bentler (2000, p. 173) and recommended for small and medium-sized samples.

3 Non-Normal Data Generation

As an illustrative example a linear growth model with selective attrition is considered. In Study A, there are five time points and no covariates. Here, non-normal data are introduced with no missing data. In Study B, a covariate is added, where the covariate is predictive of missingness in line with MAR. Here, the problems of non-normal and missing data are considered together.

For both studies, non-normal data are generated in the Mplus Monte Carlo track for mixture analysis. Non-normality is obtained by generating data from two classes of individuals with different growth model parameter values. This may be a realistic cause of non-normality in many applications. Although this may call for a mixture analysis with two classes, in the present setting only the resulting non-normality is of interest. The majority class contains 88% of the subjects and has a low intercept growth factor mean, whereas a minority class has a high intercept growth factor mean. The minority class also has a considerably larger intercept factor variance and a larger slope factor mean. For each class, normal data are generated. It follows that a single-class analysis of data generated in this fashion considers a mixture that features non-normal data with large positive skewness. The parameter values chosen in the 2-class data generation correspond to univariate skewness values of about 2 and univariate kurtosis values of about 3 - 4. These values are commonly seen in practice and represent settings studied both in Enders (2001) and in earlier literature on non-normality (Muthén & Kaplan, 1985). Because the classes only vary in the growth factor parameter values, the growth model for the single-class analysis is correct in the sense that it holds perfectly in terms of the mean and covariance structure, resulting in zero residuals and a perfect conventional chi-square fit in the population. Similar mixture issues for factor analysis models are covered e.g. in the discussion of heterogeneity in latent variable models presented in Muthén (1989, p. 559) and in the mixture modeling in Muthén (2002b, p. 102). True values for the single-class analysis may be conveniently obtained as estimates from a single replication with a very large sample size, say $n = 100,000$.

4 Study A. Linear Growth Model With Non-Normal Data: No Covariate, No Missing Data

The input and output for Study A are found in the files labelled mc2a and mc2b under Mplus Web Notes at www.statmodel.com. Mc2a considers the normality-based ML estimator using the regular standard error computations assuming normality (estimator = ml in the mixture track), whereas mc2b considers the MLR approach (estimator = mlr in the mixture track) for the same data using (4). For a general discussion of the Mplus Monte Carlo input, see Muthén (2002a). The following are new features used here.

The MONTECARLO command statements

nclasses = 1;

gclasses = 2;

indicate that 2 classes are used for data generation while 1 class is used for analysis.

The MODEL MONTECARLO command

[c#1@-2];

indicates that class 1 contains 88% of the individuals, obtained by translating the logit of -2 to a probability using $P = 1/(1 + e^{-L})$, where P is the probability and L is the logit.

The Monte Carlo summary for mc2a shows that the results for most parameters are surprisingly good in terms of estimates, standard errors, and coverage. Due to the non-normality of the data, however, the standard error for the growth factor variance is underestimated by 38% and the standard error for the growth factor covariance is underestimated by 30%. The coverage for these two estimates is also too low, 78% and 84%, respectively.

The Monte Carlo summary for mc2b shows the corresponding results using the standard error computations of the MLR estimator of (4). Here, the results for the intercept variance and covariance standard errors are coverage are very good, reflecting the robustness to non-normality.

In this study, the non-normality of the outcomes is influenced by the across-class variation in the growth factor means and variances. For the minority class (class 1) the intercept factor mean and variance are $15, 5$, while for the majority class (class 2) they are $0, 1$. Changing the minority class intercept mean and variance values to $2.5, 1$ reduces the univariate skewness values from about 2 to about $0.2-0.6$ and the univariate kurtosis values from about $3-4$ to about $0.2-0.6$. At this modest level of non-normality, all results are satisfactory for the conventional ML estimator. Problems are not severe until the level of non-normality is increased to that used in Study A.

5 Study B. Linear Growth Model With Non-Normal Data: Covariate, MAR Missing Data

The input and output for Study B are found in the files labelled mc2c and mc2d under Mplus Web Notes at www.statmodel.com. Mc2c considers the normality-based ML estimator using the regular standard error computations assuming normality (estimator = ml in the mixture track), whereas mc2d considers the MLR approach (estimator = mlr in the mixture track) for the same data using (4).

The addition of the covariate x and the missingness on $y_1 - y_5$ add the following input statements.

The MONTECARLO command includes the option missing = y1 - y5.

The ANALYSIS command specifies the option type = mixture missing.

The MODEL MISSING command specifies the probability of missing data on the outcomes given in logit scale. For y_1, the bracket logit value of -1 translates into a probability of 0.27 using the formula $P = 1/(1 + e^{-L})$. For $y_2 - y_5$ four different logistic regressions describe the probability of missingness as a function of x. The bracket value gives the intercept and the ON statement gives the slope of the logistic regression (for further details on missing data specifications, see Muthén, 2002a, Study C).

The output from mc2c contains information from the first replication about missing data patterns as well as sample coverage for each variable and pairs of variables. The first replication shows 73% missingness at time 1 and 67% missingness at time 5. The Monte Carlo summary shows a similar degree of misestimation for the same parameters as in Study A, namely misestimated standard errors and coverage for the variance of the intercept growth factor and growth factor covariance, although here referring to the residuals given x.

The output from mc2d shows that the MLR approach of (4) gives satisfactory results for all parameters. To study the generalizability of these findindg, it may be of interest to study variations on the Monte Carlo setup, varying the sample size and the degree of missingness. Readers are also referred to Yuan and Bentler (2000) and Enders (2001) for further studies.

6 Discussion

This note provides a description of possibilities to study estimation quality with non-normal and missing data using Monte Carlo simulations in Mplus. Many variations of the two studies are possible. For example, non-ignorable missing data can be generated to study biases when using the standard MAR assumption. This can be accomplished in Mplus by letting the logistic regression coefficients for the missingness vary as a function

of latent classes.

References

Amemiya, T. (1985). *Advanced Econometrics*. Cambridge, Mass.: Harvard University Press.

Arminger, G. & Sobel, M. (1990). Pseudo-maximum likelihood estimation of mean and covariance structures with missing data. Journal of the American Statistical Association, 85, 195-203.

Enders, C.K. (2001). The impact of nonnormality on full information maximum-likelihood estimation for structural equation models with missing data. Psychological Methods, 6, 352-370.

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

Muthén, B. (1989). Latent variable modeling in heterogeneous populations. Psychometrika, 54, 557-585.

Muthén, B. (2002a). Using Mplus Monte Carlo simulations in practice: A note on assessing estimation quality and power in latent variable models. Mplus Web Notes: No. 1.

Muthén, B. (2002b). Beyond SEM: General latent variable modeling. Behaviormetrika, 29, 81-117.

Muthén, B., & Kaplan D. (1985). A comparison of some methodologies for the factor analysis of non-normal Likert variables. British Journal of Mathematical and Statistical Psychology, 38, 171-189.

Muthén, L. & Muthén, B. (1998-2001). Mplus User's Guide. Los Angeles, CA: Muthén & Muthén.

Yuan, K.H. & Bentler, P.M. (2000). Three likelihood-based methods for mean and covariance structure analysis with nonnormal missing data. In Sobel, M.E. & Becker, M.P. (eds.), Sociological Methodology 2000 (pp. 165-200). Washington, D.C.: ASA.