Analyzing Imputed Data with the Bayesian Estimator in Mplus

Tihomir Asparouhov and Bengt Muthén

March 11, 2021

In this note we describe the methodology used in Mplus to analyze imputed data with the Bayesian estimator. The estimation comprises of estimating each imputed data set separately and then combining the posterior parameter distributions into a total posterior parameter distribution. If p represents the model parameters, X represents the observed data, X^* represents the missing data, and X_i^* represents the imputed data in the $i-$th imputed data set,

$$[p|X, priors] = \int [p|X, X_i^*, priors] \;\; d[X_i^*|X]$$

and therefore

$$[p|X, priors] \approx \frac{1}{M} \sum_{i=1}^{M} [p|X, X_i^*, priors]$$

where M is the number of imputed data sets. The last equation implies that combining the posterior distributions for the imputed data with equal weight produces the correct final posterior distribution. This final posterior is then used to obtain credibility intervals, point estimates as well as standard errors.

To ensure that the posterior distribution $[p|X, X_i^*, priors]$ is equally weighted, it is necessary to estimate each imputed data set with the same number of iterations. As usual in Mplus, the second half of the MCMC chains is used to construct $[p|X, X_i^*, priors]$. If the number of iterations is the same for each imputed data set, the final posterior distribution will be an equally weighted mixture of all posterior distributions. In Mplus this is implemented as follows. The number of iterations is specified as a fixed number of iterations,

using the FBITER option in the ANALYSIS command. Mplus analysis for non-imputed data with this option is generally treated as a converged estimation. The PSR convergence criterion is not checked and it is assumed that the number of iterations FBITER is chosen to be sufficiently large so that convergence is achieved. FBITER can also be used to obtain an approximate posterior distribution of the parameters even if complete convergence is not achieved. This is not the case for Mplus imputed data estimation. For imputed data, Mplus must be provided with the number of iterations FBITER, however, the PSR convergence criterion (controlled via the BCONV option) is evaluated. If convergence is not satisfied at the end of the estimation process, the posterior distribution for that imputed data set is not included. It is therefore important to check the number of completed estimations, either in the summary of the estimation or in the TECH9 output where convergence problems are reported. If the estimation for some of the imputed data sets is not complete, the FBITER option must be increased until all replications have converged. Finding the needed number of iterations FBITER, which ensures that all replications converge, may require running several preliminary runs. One reasonable strategy is to first run one imputed data set without FBITER to determine the number of iterations needed for convergence. In a second step, double the number of these iterations, and analyze all imputed data sets with this FBITER number.

It should be noted here that using the Bayesian estimator to analyze imputed data sets should be limited to a few specific situations only. It should not be used routinely to deal with missing data as in most situations analyzing the incomplete data (with the Bayesian estimator) is simpler and equally efficient. Here we describe some situations where it is reasonable to use this approach.

- In some situations, it may be desirable to impute the missing data from a model that is different from the estimated model. The imputation model may be set to a simpler model if the desired model is difficult to estimate in the presence of missing data. The imputation model can also be set to a more complex model. For example, it is reasonable to impute the data from an unrestricted model and then analyze the imputed data with a structural model. Such an approach could be useful when the structural model is potentially misspecified. The approach would prevent potential problems arising from the fact that the internally imputed data is generated from potentially an incorrect

model.

- In certain situations, it may be necessary to first impute the data due to limitations in the available estimators. For example, covariates in Mplus are not allowed to have missing values. The covariates would have to be converted to dependent endogenous variables to prevent listwise deletion. In some situations, this approach might not be desirable. The missing data can be imputed in one step and analyzed in a separate step where the covariates are treated as exogenous variables.

- In certain situations, it may be necessary to conduct the estimation in two steps. The first step would involve generating plausible values for random effects or latent variables. In the second step a non-linear transformation is applied to those plausible values to construct new predictors or dependent variables. Plausible values are essentially imputed missing values and must be analyzed as such. The second step in the estimation would require analyzing the imputed and transformed data as described in this note. One such example is a multilevel modeling of cycles by sine-cosine where the subject specific random coefficients for the sine and cosine functions are transformed into amplitude and phase for subsequent analysis.

- In some situations, the data set is already imputed and the incomplete/unimputed data is not immediately available, although comparing just two imputed data sets can be used to reveal which values are imputed and which are not, when the data is continuous. When the data is not continuous, all imputed data sets would have to be compared.

The posterior predictive p-value is not computed in Mplus with imputed data as the methodological challenges appear to be currently unresolved. The Wald test, however, can be used via the MODEL TEST command.

When the ML estimator is used with imputed data, the asymptotic variance for the parameter estimates is computed as follows

$$V = \frac{1}{M}\sum_{m=1}^{M} V_m + \frac{M+1}{M(M-1)}\sum_{m=1}^{M}(p_m - \bar{p}_\cdot)^2$$

where p_m is the parameter estimate for the m-th imputed data set, $\bar{p}_\cdot$ is the average estimate, and V_m is its asymptotic variance of p_m. When the

Bayes estimator is used with imputed data, the variance of the posterior distribution is computed as follows

$$V = \frac{1}{M}\sum_{m=1}^{M} V_m + \frac{1}{M}\sum_{m=1}^{M}(p_m - \bar{p}_\cdot)^2$$

where p_m is the sample mean of the posterior distribution for the m-th imputed data set, V_m is the sample variance of the posterior distribution for the m-th imputed data set, and $\bar{p}_\cdot$ is the sample mean of the combined/total posterior distribution. In this case, the variance is also the sample variance of the total/combined posterior distribution. The two summation terms in the above formulas represent the decomposition of the variance as within (variation within imputed data set) and between (variation between the imputed data sets). The two formulas are equivalent under the following asymptotic conditions.

- The sample size is sufficiently large, which ensures that the posterior distribution is asymptotically normal/symmetric. In this case the ML mode estimate would be the same as the median estimate used with the Bayes estimator, i.e., p_m and V_m are asymptotically the same for the two estimators
- The number of iterations in the MCMC estimation is sufficiently large so that the sample mean and variance p_m and V_m are asymptotically the same as the true mean and variance of the posterior distribution
- The number of imputed data sets M is sufficiently large so the term $(M+1)/(M-1)$ converges to 1

The last requirement can be quantified as follows. Using 50 imputed data sets, the standard errors obtained with the Bayes estimator would be within 2% of the standard error obtained with the ML estimator. Using 100 imputed data sets, the standard errors obtained with the Bayes estimator would be within 1% of the standard error obtained with the ML estimator. For smaller M, the Bayes estimator standard errors would be smaller than those obtained with the ML estimator. At least $M = 50$ imputed data sets should be used with the Bayes estimator. However, the standard error estimates would be very close even with a smaller M. If the amount of missing data information (ICC) is large and M is very small, such as $M = 5$, a more noticeable difference between the two estimators can occur.

Table 1: Comparing the ML and Bayes Estimator results on multiply imputed data: Estimate(SE)

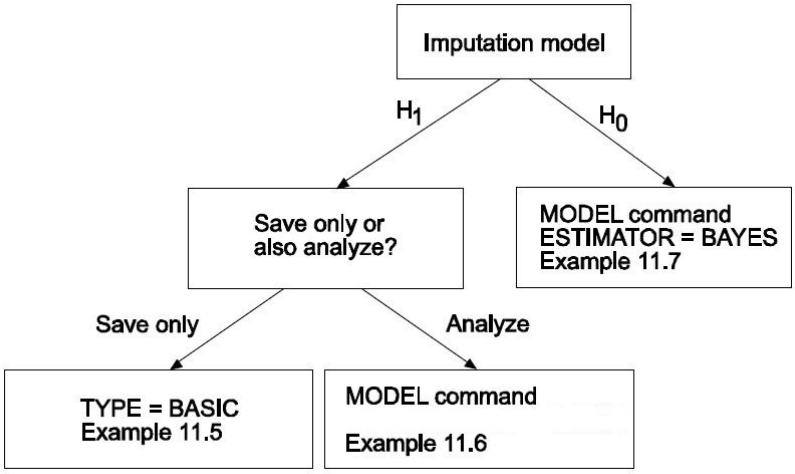
Estimator	M	μ_1	v_1	ρ
Bayes	5	-0.06(.066)	1.05(.088)	0.55(.068)
ML	5	-0.06(.073)	1.04(.092)	0.54(.073)
Bayes	10	-0.05(.060)	1.02(.085)	0.55(.068)
ML	10	-0.06(.063)	1.01(.086)	0.53(.068)
Bayes	20	-0.03(.068)	1.04(.103)	0.55(.071)
ML	20	-0.03(.069)	1.04(.104)	0.55(.071)
Bayes	50	-0.03(.068)	1.02(.099)	0.54(.071)
ML	50	-0.02(.070)	1.02(.098)	0.53(.070)
Bayes	100	-0.03(.066)	1.02(.097)	0.54(.071)
ML	100	-0.03(.067)	1.01(.096)	0.53(.070)
Bayes	Incomplete	-0.02(.063)	1.02(.092)	0.54(.068)
ML	Incomplete	-0.03(.063)	1.00(.092)	0.53(.068)

Next we illustrate the performance of the Bayes estimator with a simulated example. We generate two variables Y_1 and Y_2 with a standard normal distribution and correlation of 0.5. Missing data is generated for Y_1 as follows

$$P(Y_1 \text{ is missing}) = \frac{1}{1 + Exp(Y_2)}.$$

This missing data mechanism is MAR (not MCAR) and yields approximately 50% missing data for Y_1. A single data set is generated with 500 observations. The data set is used to obtain M imputed data sets using the unrestricted bivariate model. The data are then analyzed as incomplete and as imputed with both the ML and the Bayes estimator. The results for the 3 parameters affected by the missing data treatment: $\mu_1 = E(Y_1)$, $v_1 = Var(Y_1)$ and $\rho = Cov(Y_1, Y_2)$ are reported in Table 1. The results indicate that the ML and the Bayes estimators perform similarly and that the incomplete data estimation results are similar to the imputed data estimation results. The results also indicate that some improvement in the precision of the estimates and the standard errors is obtained by increasing M to 100 but the gains are rather small.

Figure 1: Mplus options for generating and analyzing imputed data



Analyzing imputed data with the Bayesian estimator is implemented in Mplus version 8.6, and it is not available in earlier versions. On page 576 of Mplus User's Guide Version 8, a diagram presents the possible options in generating and analyzing imputed data. An updated version of this diagram is presented in Figure 1 which reflects the fact that imputed data can now be analyzed also with the Bayes estimator.