Plausible Values for Latent Variables Using Mplus

Tihomir Asparouhov and Bengt Muthén

August 21, 2010

1 Introduction

Plausible values are imputed values for latent variables. All latent variables can be thought of as observed variables that have missing data for all observations. Using Mplus imputation utilities based on the MCMC Bayesian estimation, see Asparouhov and Muthén (2010), we can produce imputed values for each latent variable. If sufficient number of imputed values are drawn are we essentially obtain the entire posterior distribution of the latent variables.

There are two types of applications that we describe below. The first applications is when the individual level latent variable is of interest. One such example is the case when we want to construct a factor scores estimate for a latent variable and a standard error for that factor score estimate. To be able to do this it is necessary to use many imputed values. For example 100 or 500 such values can yield a precise posterior distribution for a latent variable which can be used to compute the posterior mean factor score estimate. In this note we will use 500 plausible values for the purpose of computing factor scores and their standard errors.

The second type of application is the case when a population level statistic is of interest. One such example is when the plausible values are used into a secondary model. In this case only 5 imputed data sets can be used. These plausible value data sets are analyzed just like missing data imputed data sets, i.e, by combining the results across the imputations using Rubin's method (1987).

For particular applications of plausible values see also von Davier et al. (2009), Mislevy et al. (1992) and Carlin (1992).

2 Factor Score Estimates For Small Sample Sizes

In many SEM applications the ML or WLS estimators yield negative residual variances. This is commonly referred to as Heywood Case and it often occurs for no other reason but a small sample size. In that case the frequentist estimation methods will not be able to produce factor score estimates at all. The Bayes estimator however can be used to construct factor score estimates by using the plausible values utility to compute the posterior mean for each latent variable. In this section however we focus on a more intricate situation when a Heywood case does not occur but one of the residual variances is near zero. In such a case the the frequentist methods can produce factor score estimates, however, these estimates are likely to be of poor quality because the factor score estimates will be almost perfectly correlated with the measurement variable with small residual variance. Again these problems can also occur for no other reason but a small sample size.

Consider the following factor analysis example

$$Y_j = \mu_j + \lambda_j \eta + \varepsilon_j$$

where Y_j are the observed variables for $j = 1, ..., 3$, η is the latent factor and ε_j are the residual variables. We generate a single data set with $N = 45$ observations using the following parameter values $\mu_j = 0$, $\lambda_j = 1$ and the factor variance as well as the residual variances are set to 1. All variables are generated from a normal distribution. Using this data set we estimate the factor model with the ML and the Bayes estimators. The model parameter estimates are presented in Table 1. The two estimators yield similar results

Table 1: ML and Bayes Estimates for Factor Analysis Model

Parameter	ML	Bayes
μ_1	0.307	0.270
μ_2	0.008	-0.053
μ_3	0.296	0.212
λ_2	0.873	0.976
λ_3	1.665	1.543
θ_1	1.298	1.407
θ_2	1.587	1.603
θ_3	0.034	0.589
ψ	0.966	0.946

with the exception of the estimates for the residual variance parameter θ_3. The ML estimator in this example produced a near Heywood case, i.e., the ML estimate for this residual variance is a small but positive number.

Next we estimate the factor scores for η and the factor score standard errors using both the ML and the Bayes estimators. To evaluate the performance of the factor score estimates we compute

$$MSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(\hat{\eta}_i - \eta_i)^2}$$

where $\hat{\eta}_i$ is the factor score estimate and η_i is the true factor score value. For the ML estimator the MSE is 0.636 while for the Bayes estimator it is 0.563. This indicates that the Bayes estimator produces more accurate factor score estimates. In addition MSE should be near the average SE values for

the factor scores. In this example the average SE for the ML estimator is 0.109 while for the Bayes estimator it is 0.484, i.e., the Bayes estimator yields more accurate SE estimates. This can also be seen by computing the coverage probability for both estimators, that is, the percentage of observations for which the 95% symmetric confidence interval around the factor score estimate contains the true factor score value. Using the ML estimator we obtain a coverage of 20%, while using the Bayes estimator we obtain a coverage of 89%. The underestimation of the factor score standard errors is rooted in the small sample size, which leads to a small residual variance estimate, which in turn leads to underestimation of the factor score standard errors.

We conclude that for small sample size the Bayes factor score estimates and standard errors are more reliable than those obtained by the ML estimator.

Also note here that the ML factor score standard errors are the same across all observations. Only the factor score estimate changes across the observations but the factor score standard error remains constant. This is not the case for the Bayes estimator because with the Bayes estimator the parameters are not assumed to be constants as in the ML estimator. For the Bayes estimator the parameters are assumed to vary according to the estimated posterior distribution. This is another advantage of the Bayes estimator because the uncertainty in the model parameter estimates is taken into account. The correlation between the absolute factor score estimates and the factor score standard error estimates is 0.76, i.e., a strong correlation exist between these estimates. The larger the absolute factor score value is the larger the standard error is. This result has a natural explanation.

Large absolute factor score values are generally in the tail of the factor score distribution, i.e., in a region with fewer observations. Thus having larger standard errors in that region is natural. When the sample size is large however the uncertainty in the parameter estimates becomes so small that the Bayes standard errors also become nearly constant, i.e., the varying standard error phenomenon is only relevant for small sample sizes.

3 Factor Score Estimates For Large Sample Sizes and Non-normally Distributed Factors

In this section we explore the effect of factor non-normality on the factor score estimates. We use the same model as in the previous section however to avoid any small sample size related problems we generate a very large data set of size $N = 10000$. The parameters used in the data generation are as in the previous section. The factor variable η is generated either from a standard normal distribution or an exponential distribution with mean and variance 1. The residual variables are always generated from a normal distribution. Asymptotically the ML and the Bayes estimator yield the same results not just for the parameters but also for the posterior distribution of the factors. This occurs regardless of whether the data is normally distributed or not. We analyze the data using the true one factor model with the parameterization where all the loadings are estimated and the factor variance is assumed to be fixed to 1.

In this example we first analyze the data with normally distributed factors and we obtain the following results. The coverage for the factor scores for both the ML and the Bayes estimator is 94%. MSE for both estimators is 0.507 which is close to the average factor score SE of 0.496 for ML and 0.495 for Bayes. When we analyze the data with non-normally distributed factors we obtain the following results. The MSE for the ML estimators is 0.502 and for the Bayes estimator it is 0.496 while the average factor score standard error for both estimators is 0.496. The coverage for both estimators is 95%.

From these results we make the following conclusions. The Bayes and the ML estimators asymptotically yield the same factor score estimates and factor score standard errors. In addition non-normality of the factor does not affect the quality of the factor score estimates or the quality of the factor score standard error.

4 Using Plausible Values in Secondary Analysis

4.1 Plausible Values for Categorical Latent Variables

In this section we demonstrate how plausible values for categorical latent variables can be used for secondary modeling. The imputed plausible values can be used as observed values to build models after the latent variable has been imputed. As an example we will use one of the latent class models described in Clark and Muthén (2009). In this model there are 10 binary latent class indicators U_j and a single latent class predictor variable X. Denote by

C the latent class variables. The variables U_j as well as the variable C take values 1 and 2. The model is described as follows.

$$P(U_j = 1|C = 1) = p_{1j} \tag{1}$$

$$P(U_j = 1|C = 2) = p_{2j} \tag{2}$$

for $j = 1, ..., 10$. In addition

$$P(C = 1|X) = \frac{1}{1 + Exp(\alpha + \beta X)}. \tag{3}$$

We generate 100 data sets according to the above model all of size $N = 1000$. We use the following parameter values to generate the data: $\alpha = 0$, $\beta = 0.5$, $p_{1j} = 0.73$, and $p_{2j} = 0.27$. The entropy for this mixture model is 0.8.

The goal of the following simulation study is to demonstrate how the latent variable C can be imputed from a latent class model and then used separately to estimate the logistic regression equation (3). We consider two imputation models for the latent class variable. The first model is the latent class model given by equations (1) and (2), i.e., this imputation model does not include X. The second model we use as an imputation model is the model where the X variable is included as an indicator variable, i.e., in addition to equations (1) and (2) we estimate the following equation

$$X|C = \alpha_C + \varepsilon. \tag{4}$$

In general imputation models should include as many variables as possible because each observed variable can carry some unique information about the latent variables. The exact specification for the imputation model is not very important as long as it reasonably flexible.

Table 2: Using plausible values to estimate a logistic regression for a latent class variable.

Imputation Model	Bias	Coverage	MSE
Model without X	-0.11	75%	0.123
Model with X	-0.02	97%	0.076

For each of the 100 generated data sets we use the two imputation models to impute the latent class variable C. We generate 5 imputed data sets. The imputed data sets are then used to estimate a logistic regression of the imputed values for C on the predictor variable X. This is done as in the usual imputation analysis using the Mplus implementation of Rubin (1987) method. The results of this simulation study are presented in Table 2. In this table we only report the results for the logistic regression coefficient β. The true value for this regression coefficient is 0.5. The latent class imputation where X is included in the model clearly outperforms the latent class imputation without X. When X is included in the imputation the estimate of β is unbiased, the coverage is near the nominal 95% and the MSE is better than for the imputation model without X. In fact this imputation method shows that plausible values can indeed be used for secondary analysis. Using the imputation model without the X variable we get slightly biased estimates, which leads to low coverage and increase in MSE. These results also match the results obtained by the pseudo draw method reported in Clark and Muthén (2009) in the case when the entropy is 0.8. The pseudo draw method is very similar to the plausible values method. We conclude that plausible

values method can be used to perform secondary analysis using latent variables measured and imputed from a previously estimated model, however, the imputation models should be as broad as possible and should include all relevant variables, otherwise the secondary analysis could be slightly biased. We also demonstrated in the above simulation study that 5 imputed data sets are sufficient for the purpose of secondary model estimation.

4.2 Plausible Values for Continuous Latent Variables

In this section we show the advantages of plausible values over traditional factor score estimates for the use in secondary analysis. Consider the following factor analysis model with two factors each measured by three dependent variables. The model is given by the following two equations. For $j = 1, ..., 3$

$$Y_j = \mu_j + \lambda_j \eta_1 + \varepsilon_j$$

and for $j = 4, ..., 6$

$$Y_j = \mu_j + \lambda_j \eta_2 + \varepsilon_j.$$

We generate a single data set of size 10000 using the following parameter values $\mu_j = 0$, $\lambda_j = 1$, the variance θ_j of ε_j is 1, the variance ψ_{ii} of η_i is 1 and the covariance ψ_{12} between η_1 and η_2 is 0.6. We estimate the true model using the ML estimator and we estimate the factor scores for both factors. In addition we estimate the true model with the Bayes estimator and generate 5 sets of plausible values. We use the factor scores and the plausible values to simply estimate the factor means, the factor variance and the factor correlation in a secondary run. The plausible values are used again as multiple imputation values. The results of this simulation are presented

Table 3: Using plausible values and factor scores to estimate factor means, factor variances and covariances.

Parameter	True Value	Factor Scores	Plausible Values
α_1	0	0.00	0.00
α_1	0	0.00	0.00
ψ_{11}	1	0.76	1.03
ψ_{22}	1	0.80	1.05
ψ_{12}	0.6	0.57	0.63
ρ	0.6	0.73	0.61

in Table 3. The factor means are denoted by α_i. The factor correlation is denoted by ρ. It is clear from these results that the plausible values method yields more accurate estimates for the factor variances and the factor correlation. The factor score method overestimates the factor correlation and underestimates the factor variances.

References

[1] Asparouhov T. & Muthén B. (2010). Multiple Imputation with Mplus. Technical Report. www.statmodel.com

[2] Carlin JB (1992) Meta analysis for 2 x 2 tables: a Bayesian approach. Stats Med. 11,141-158.

[3] Clark S. & Muthén B. (2009). Relating latent class analysis results to variables not included in the analysis. Submitted for publication.

[4] Mislevy R., Johnson E., & Muraki E. (1992) Scaling Procedures in NAEP. Journal of Educational Statistics, Vol. 17, No. 2, Special Issue: National Assessment of Educational Progress, pp. 131-154.

[5] Rubin, D.B. (1987) Multiple Imputation for Nonresponse in Surveys. J. Wiley & Sons, New York.

[6] von Davier M., Gonzalez E. & Mislevy R. (2009) What are plausible values and why are they useful? IERI Monograph Series Issues and Methodologies in Large-Scale Assessments. IER Institute. Educational Testing Service.