Bayesian Analysis of Latent Variable Models using Mplus

Tihomir Asparouhov and Bengt Muthén

Version 2

June 29, 2010

1 Introduction

In this paper we describe some of the modeling possibilities that are now available in Mplus Version 6 with the Bayesian methodology. This new methodology offers many new possibilities but also many challenges. The paper is intended to spur more research rather than to provide complete answers. Recently there have been many papers on Bayesian analysis of latent variable models. In this paper we do not provide a review of the existing literature but rather emphasize the issues that have been overlooked up to now. We use the Bayesian methodology in the frequentist world and compare this methodology with the existing frequentist methods. Here we do not provide details on the algorithms implemented in Mplus, but such details are available in Asparouhov and Muthén (2010). We focus instead on simulation studies that illuminate the advantages and disadvantages of the Bayesian estimation when compared to the classical estimations methods such as the maximum-likelihood and weighted least squares estimation methods.

2 Factor Analysis

2.1 Large vs. Small Number of Continuous Indicators

In this section we evaluate the Bayes estimation of a one factor analysis model with a relatively large number of indicators $P = 30$ and a small number of indicators $P = 5$. As we will see having a large number of indicators is more challenging than one would expect in MCMC because it creates a high correlation between the generated factors and the loadings. We consider three different parameterizations. Denote by parameterization "L" the parameterization where all loadings are estimated and the factor variance is fixed to 1. The parameterization "V" is the parameterization where the first factor loading is fixed to 1 and the variance of the factor is estimated. The parameterization PX is the extended parameter parameterization where both the variance and the first loadings are estimated. This model is formally speaking unidentified however the standardized loadings are still identified. These standardized loadings are obtained by

$$\lambda_s = \lambda\sqrt{\psi}$$

where λ_s is the standardized loading, λ is the unstandardized loading and ψ is the factor variance. The loadings λ_s are essentially equivalent to the

loadings in parameterization "L". The "PX" parameterization idea has been used with the Gibbs sampler for example in van Dyk and Meng (2001) and Gelman et al. (2008b).

We compare the 3 parameterizations in a simulation study with 30 indicators. We generate the data using the following parameters, all loadings are set to 1, all residual variances are set to 1 and the factor variance is set to 1. We generate 100 data sets of various sample sizes. The priors used in the Bayes estimation are the Mplus default priors, see Appendix A. For all loadings and intercepts the prior is uniform on the $(-\infty, \infty)$ interval. The prior for the variance parameter is the inverse-gamma. This prior tends to be important for small sample size models thus we will investigate the various options. We consider three different priors $IG(-1, 0)$, $IG(0, 0)$, and $IG(1, 2)$. The first two can be considered non-informative prior, while the last one can be considered to be a weakly informative prior. We denote the corresponding parameterizations by "V1", "V2", and "V3". The results are presented in Table 1. The parameters are presented in standardized scale. The mean and the residual variance parameters are estimated well in all cases and we do not include those in the table. The loading parameters are symmetric for parameterizations "L" and "PX" and we include only the first loading for these parameterizations. For parameterizations "V" we report the first and the second loading. All loadings after the second loading are equivalent essentially to the second loading.

It it is clear from the results in Table 1 that the "PX" parameterization is superior for small sample sizes such as $N = 50$ and $N = 100$. The "PX" parameterization shows no bias with any sample size. The "L" parameterization shows biased estimates for small sample sizes but for $N = 200$ or more the results are good both in terms of coverage and bias. The parameterization "V" is clearly the worst parameterization for this model. For sample size $N = 50$ and $N = 100$ it showed too many convergence problems in the case of $IG(-1, 0)$ and $IG(0, 0)$ priors (parameterizations "V1" and "V2"). Among the three different prior the $IG(1, 2)$ prior (parameterization "V3") yields the best results. For sample size of 500 or more the prior has no effect on the results. For sample sizes $N = 200$ and even $N = 500$ the estimates show significant bias and coverage problems.

In contrast Table 2 shows the corresponding results with 5 indicators. The results show virtually no problems for all three parameterizations with the exception of the "V" parameterization for "N=50" where there seems to be some bias. The different variance priors has little effect on the estimates.

Table 1: Bias(coverage) for one factor model with 30 indicators.

Parameter-ization	Para-meter	N=50	N=100	N=200	N=500	N=1000	N=5000
L	λ_1	0.59(47)	0.17(76)	0.07(87)	0.02(91)	0.01(97)	0.00(90)
PX	λ_1	0.03(92)	0.00(95)	0.00(96)	-0.01(97)	0.00(98)	0.00(91)
V1	λ_1	-	-	-0.26(41)	-0.07(76)	-0.03(89)	-0.01(89)
V1	λ_2	-	-	0.00(98)	0.00(96)	0.00(96)	0.00(93)
V2	λ_1	-	-	-0.31(34)	-0.08(73)	-0.03(89)	-0.01(88)
V2	λ_2	-	-	0.00(98)	0.00(96)	0.00(96)	0.00(93)
V3	λ_1	-0.59(0)	-0.41(17)	-0.18(47)	-0.07(78)	-0.03(90)	-0.01(89)
V3	λ_2	0.40(64)	0.05(95)	0.01(98)	0.00(97)	0.00(96)	0.00(93)

Thus we conclude that the problems we see in the Bayes estimation are specific to the large number of indicators. The larger the number of indicators the bigger the sample size has to be for some of the above parameterizations to yield good estimates. While the 3 models use different priors settings the difference in the results is not due to the priors but it is due to the different mixing in the MCMC chain. If all chains are run infinitely long the results would be the same or very close, however running the chain infinitely long is only an abstract concept and therefore it is important to know which parameterization is best for which model. The above simulation shows that the more traditional parameterizations "V" and "L" are fine to use unless the number of indicators is large and the sample size is small. In this special case one needs to use the more advanced parameterization "PX". As the number of indicators increases the minimal sample size for the "L" parameterization and the "V" parametrization will grow. The simulations also show that the parameterization "L" is somewhat better than parameterization "V" when the sample size is small.

2.2 Centered Parameterization

Dunson et al. (2005) recommend the use of the so called centered parameterization in structural equation models where not only the factor variance is a free parameter as in the parameterization "V" described in the previous

Table 2: Bias(coverage) for one factor model with 5 indicators.

Parameter-ization	Para-meter	N=50	N=100	N=200	N=500	N=1000	N=5000
L	λ_1	0.04(98)	0.04(95)	0.03(95)	0.02(94)	0.01(93)	0.00(94)
PX	λ_1	0.00(94)	0.02(98)	0.02(95)	0.01(96)	0.01(94)	0.00(93)
V1	λ_1	-0.15(87)	-0.03(97)	0.00(96)	0.00(98)	0.00(96)	0.00(95)
V1	λ_2	0.02(95)	-0.01(96)	0.00(97)	0.00(92)	0.00(91)	0.00(96)
V2	λ_1	-0.07(100)	-0.07(96)	-0.02(97)	0.00(98)	0.00(96)	0.00(95)
V2	λ_2	0.04(97)	0.01(97)	-0.01(97)	0.00(92)	0.00(92)	0.00(96)
V3	λ_1	-0.15(87)	-0.03(97)	0.00(96)	0.00(98)	0.00(96)	0.00(95)
V3	λ_2	0.02(95)	-0.01(96)	0.00(97)	0.00(92)	0.00(91)	0.00(96)

section but also the factor mean is a free parameter as well. In this parameterization for identification purposes the intercept of one of the indicator variables is fixed to 0. We denote this parameterization by "C". The conclusion in Dunson et al. (2005) is that the centered parameterization offers better mixing and faster convergence. The purpose of this section is to evaluate the need for this parameterization in Mplus. The results in Dunson et al. (2005) are obtained from an algorithm that differers from the Mplus algorithm in one important aspect. In Mplus all latent variables are generated as a block, i.e., they are generated simultaneously from a multivariate distribution conditional on the observed data and the parameters. This protects from a potentially poor mixing due to highly correlated draws for the latent variables. In Dunson et al. (2005) as well as the WinBugs software the latent variables are generated one at a time, i.e., from a univariate distribution conditional on the other latent variables, the observed data and the parameters.

We conduct a simulation study to evaluate the performance of parameterizations "C", "V" and "L". Parameterizations "V" and "L" are the most commonly used parameterizations with the ML estimator and would be considered the natural parameterizations for most applications. As an example we use the model described in Dunson et al. (2005) and the parameter estimates reported in Appendix B in Dunson et al. (2005). All variables are generated using a normal distribution. We generate 100 data sets using

samples $N = 75$ and $N = 1000$ and analyze the data sets with the three parameterizations. The small sample size $N = 75$ is chosen to correspond to the actual application reported in Dunson et al. (2005). The model has 11 observed variables and 9 factors, i.e., the example does not appear to be in the realm of the large number of indicators phenomenon reported in the previous section. The following equations describe the model

$$Y = \nu + \Lambda\eta + \Gamma\xi + \varepsilon$$

$$\eta_1 = \alpha_1 + \beta_{13}\eta_3 + \zeta_1$$
$$\eta_2 = \alpha_2 + \beta_{21}\eta_1 + \beta_{23}\eta_3 + \zeta_2$$

where Y is the vector 11 observed dependent variables, η is a vector of 3 latent variables and ξ is a vector with 6 latent variables. The variables ε, ζ_1 and ζ_2 are independent residuals. The structure of the loading matrix is described as follows. In the Λ matrix all the entries are 0 with the exception of λ_{i1} for $i = 1, ..., 4$, λ_{i2} for $i = 5, ..., 8$ and λ_{i3} for $i = 9, ..., 11$, i.e., the first 4 variables are measurements of η_1, the next 4 variables are measurements for η_2 and the last 3 variables are measurements for η_3. The purpose of the latent variables ξ is to pick up unaccounted by η residual correlations among the observed variables. For example, ξ_1 gives a residual correlation between y_1 and y_5, i.e, the first and the fifth loadings of ξ_1 are fixed to 1 and the rest of the ξ_1 loadings are fixed to 0. The remaining 5 residual correlation included in this model corresponding to the remaining five ξ factor variables are the residual correlations between y_2 and y_4; y_2 and y_6; y_3 and y_7; y_4 and y_8; and y_6 and y_8.

Using all 3 parameterizations in Mplus we obtained similar estimates and the coverage of the confidence intervals in all cases were near the 95% nominal rate. Here we do not report the parameter estimates but rather the computational time until convergence which is essentially an indicator for the quality of the mixing as well as the convergence rate. Table 3 contains the average computational time for a single replication as well the convergence rate in the simulation study. The computational time indicates that essentially there is no difference in the mixing and that regardless of which parameterization is used the estimation for this model is straight forward. In addition the convergence rate is 100% for the sample size $N = 1000$ and for the small sample size of $N = 75$ the convergence rate is the best for parameterization "L". Therefore we can make the following conclusions. Parameterization choice

Table 3: Computational time (in seconds) and convergence rate for different SEM parameterizations.

Parameter-ization	Comp. Time $N = 75$	Comp. Time $N = 1000$	Conv. Rate $N = 75$	Conv. Rate $N = 1000$
L	1	4	100%	100%
V	1	3	92%	100%
C	1	5	93%	100%

is tied to the estimation algorithm and parameterization recommendations obtained with different algorithms do not necessarily apply to Mplus. In particular the centered parameterization recommended in Dunson et al. (2005) does not seem to have any advantages over the standard parameterizations when used with Mplus. In addition, the choice of the parameterization has a fairly limited impact when the sample size is large, although "large" is a relative term and it is different for different models. When the sample size is small it appears from the simulation study in this and the previous section that the "L" parameterization is more robust.

2.3 Factor Analysis with Binary Indicators

In this section we consider some of the issues that arise in Bayesian factor analysis with binary variables. Binary indicators provide more limited information than continuous variables. If also the sample size is small and the number of indicators is small there will be more limited information in the data about the quantities that we are estimating. In this situation the estimates will depend on the priors. Consider a one factor model with five binary indicators where all the thresholds are 0, all the loadings are 1, and the factor variance is 1. We estimate this model with the parameterization "L". The default prior on each loading is the normal distribution with zero mean and infinite variance, which we denote by $N(0, \infty)$. We also consider three other priors: $N(0, 1)$, $N(0, 4)$ and $N(0, 20)$. All of these are to some extent non-informative and diffuse. Such priors are common in the IRT literature. For example, in Fox and Glas (2001) the prior for the loadings are $N(0, \infty)$ constrained to all positive values. In Segawa et al. (2008) the prior for the loadings is $N(0, 10^5)$ constrained to (-25,5). In Patz and Junker (1999) the

prior for the loading is log-normal with mean 0 and variance 0.5. In Song et al. (2009) and Lee et al. (2010) similar but more complicated priors were used. In all of these papers however the sample size were large and the effect of the priors on the estimation is very small. Small sample size situations were not investigated.

In Gelman et al. (2008a) weakly informative priors, such as $N(0, 1)$, $N(0, 4)$ and $N(0, 20)$, are recommenced for logistic and probit regression although preference is given there to priors based on the T-distribution and the Cauchy distribution.

In this simulation we generate 100 data sets of different sample sizes and analyze the data sets with each of these 4 prior assumptions for the loadings. The results are presented in Table 4. The table contains the bias and coverage only for the first loading parameter. The remaining loading parameters are similar to the first loading. The threshold parameters are estimated well in all cases and we do not include these results. It is clear from these results that the prior choice affects the results quite substantially for sample size $N = 50$ and $N = 100$ while for sample size $N = 200$ and bigger the effect of the prior is small or none at all. As the sample size increases the effect of the prior essentially disappear and the parameter estimates become the same as the ML estimates. For small sample sizes the point parameter estimates are affected dramatically by the prior. The best results are obtained with the $N(0, 1)$ prior. Using $N(0, \infty)$ for parameters on the logit scale may actually be a very poor choice. Consider the implications of such a prior on R^2 of the regression equation for each of the Y. The probability that $R^2 < 99\%$ is smaller than $R^2 > 99\%$. It is obvious that this prior will be inappropriate for most situations and given that the priors have an effect on the results when the sample size is known a more thoughtful choice is needed. If no informative prior is available, a prior related to the ML or WLSMV estimates for the same model may be a viable choice.

In Lee (2010) a similar conclusion has been reached. The authors state that large sample sizes are required to achieve accurate results. Let's again iterate and clarify this point. Bayesian estimation of structural models with categorical variables show **prior assumption dependence** when the sample size is small, for example $N = 100$. The results of the Bayesian estimation are accurate, however they depend on the prior assumptions. In these models all priors are informative since they affect the final result substantially. Thus setting up meaningful priors is very important. Setting up priors in these models is also challenging because of the fact that the parameter are on

probit scale. It would be much easier for setup priors on probability scales but that is not always possible.

The **prior assumption dependence** may be a substantial hurdle in practice, simply because good and meaningful priors are not easy to specify. This is particularly problematic when the Bayesian estimation is simply used to obtain good point estimates and standard errors, because it is not clear how to select priors that provide estimates with small or no bias. Our example shows however that simply avoiding the so called generic non-informative priors is a good first step. For all parameters that are on probit scale it seems that specifying priors that have unlimited uniform range is not a good idea because such priors induce skewed priors on probability scale. Instead, selecting priors with a reasonable and finite range is likely to yield better results.

The **prior assumption dependence** occurs for small sample sizes. Every model however will have a model specific sample size range where this dependence occurs and we can not provide a general sample size range. However the **prior assumption dependence** is easy to check. Simply estimating the model with various different priors will show the degree to which the estimates depend on the priors.

Finally we are going to provide a frequentist interpretation on the posteriors and explain why priors can affect the results. Suppose that we draw parameters form the priors and then from those parameters and the model we draw data sets similar to the observed data. We then remove all such draws that produce data different from the observed data and we retain all draws that produced data sets that are the same as the observed. The retained parameters form the posterior distribution. Thus if two parameter values θ_1 and θ_2 have been equally likely to produce the data the frequency ratio in the posterior will be the same as that in the prior, i.e., the prior will have a tremendous effect on the posterior when the data can not discriminate very well between the parameters.

3 Estimating Structural Equation Models With Categorical Variables and Missing Data

The most popular method for estimating structural equation models with categorical variables is the weighted least squares method (estimator=WLSMV

Table 4: Bias (percent coverage) for $\lambda_{1,1} = 1$ in factor analysis with binary indicators.

Prior	N=50	N=100	N=200	N=500
$N(0, \infty)$	3.41(84)	0.56(89)	0.11(92)	0.04(91)
$N(0, 20)$	0.55(87)	0.31(89)	0.13(92)	0.04(92)
$N(0, 4)$	0.27(93)	0.18(91)	0.09(93)	0.04(94)
$N(0, 1)$	0.04(95)	0.07(97)	0.04(94)	0.01(93)

in Mplus). This method however has certain limitations when dealing with missing data. The method is based on sequentially estimating the univariate likelihood and then conditional on the univariate estimates the bivariate model is estimated. The problem with this approach is that when the missing data is MAR and one dependent variable Y_1 affects the missing data mechanism for another variable Y_2, the two variables have to be estimated simultaneously in all stages of the estimation otherwise the estimates will be biased.

Consider the following simple example. Let Y_1 and Y_2 be binary variables taking values 0 and 1 and let Y_1^* and Y_2^* be the underlying normal variables. The relationship between Y_i and Y_i^* is given by

$$Y_i = 0 \Leftrightarrow Y_i^* < \tau_i$$

for $i = 1, 2$ and parameters τ_i. Let $\tau_1 = \tau_2 = 0$. In that case $P(Y_1 = 0) = P(Y_2 = 0) = 50\%$. Suppose also that the tetrachoric correlation between the two variables is $\rho = 0.5$. Suppose that the variable Y_2 has missing values and that the missing data mechanism is

$$P(Y_2 \text{ is missing}|Y_1 = 0) = Exp(-2)/(1 + Exp(-2)) \approx 12\% \qquad (1)$$

$$P(Y_2 \text{ is missing}|Y_1 = 1) = Exp(1)/(1 + Exp(1)) \approx 73\%. \qquad (2)$$

This missing data mechanism is MAR (missing at random). We simulate 100 data sets according to this bivariate probit model of size 1000 and generate the missing data according to (1) and (2). We then estimate the unrestricted bivariate probit model with both the WLSMV and Bayes estimators in Mplus.

Table 5: Comparing the WLSMV and Bayes estimators on bivariate MAR dichotomous data.

Parameter	True Value	WLSMV Estimates	Bayes Estimates	WLSMV Coverage	Bayes Coverage
ρ	0.50	0.35	0.51	0.14	0.91
τ_1	0.00	0.01	0.01	0.96	0.96
τ_2	0.00	0.23	-0.01	0.00	0.95

The results of the simulation study are given in Table 5. It is clear from these results that the WLSMV estimates are biased while the Bayes estimates are unbiased. The bias in the WLSMV estimates results in poor coverage for that estimator, while the coverage for the Bayes estimator is near the nominal 95% level.

The same problem occurs for two-level models. Suppose that we have 500 clusters of size 10 and two observed binary variables. The corresponding basic bivariate two-level probit model for the two binary variables is given by

$$Y_i = 0 \Leftrightarrow Y_i^* < \tau_i$$

$$Y_i^* = Y_{iw} + Y_{ib}$$

for i=1,2. Here Y_{iw} is a standard normal variable, i.e., Y_{iw} has mean zero and variance 1. The variable Y_{ib} has zero mean and variance v_i. Both Y_{iw} and Y_{ib} are unobserved. There are two types of tetrachoric parameters in this model. On the within level we have the within level tetrachoric correlation ρ_w between Y_{1w} and Y_{2w}. On the between level we have the between level tetrachoric covariance ρ_b between Y_{1b} and Y_{2b}. We generate 100 data sets and we generate missing data using the missing data mechanism (1) and (2). We then analyze the data with the Bayes and WLSMV estimators in Mplus. The results of the simulation study are given in Table 6. We see that for two-level models the WLSMV estimates are again biased while the Bayes estimates are unbiased. The tetrachoric WLSMV estimates on both levels are biased as well as the threshold WLSMV estimates.

The weighted least squares estimator relies on unbiased estimates of tetrachoric, polychoric and polyserial correlations to build estimates for any structural model. If these correlation estimates are biased the structural parame-

Table 6: Comparing the WLSMV and Bayes estimators on bivariate two-level MAR dichotomous data.

Parameter	True Value	WLSMV Estimates	Bayes Estimates	WLSMV Coverage	Bayes Coverage
ρ_w	0.50	0.40	0.50	0.08	0.88
ρ_b	0.20	0.17	0.20	0.70	0.96
v_1	0.30	0.30	0.30	0.94	0.97
v_2	0.30	0.28	0.31	0.93	0.93
τ_1	0.00	0.00	0.00	0.95	0.93
τ_2	0.00	0.24	0.00	0.00	0.96

ters estimates will also be biased. Consider for example the growth model of 5 binary variables observed at times $t = 0, 1, 2, 3, 4$. The model is described by the following equation

$$P(Y_{it} = 1) = \Phi(\eta_{1i} + t\eta_{2i}).$$

where Φ is the standard normal distribution function. The model has 5 parameters: the mean μ_1 of the random intercept η_{1i} and the mean μ_2 of the random slope η_{2i} as well as the variance covariance Ψ of these two random effects which has 3 more parameters. We generate 100 data sets of size 1000 and we generate missing data for y_2, y_3, y_4 and y_5 via the missing data mechanism described in (1) and (2), i.e., y_1 affects the missing data mechanism for $y_2, ..., y_4$.

The results of this simulation study can be found in Table 7. We analyze the data using the true model with several different estimators. We analyze the data again with the WLSMV estimator directly and with the Bayes estimators directly. The Bayes estimator we use use two different priors for the Ψ, the default prior of uniform improper prior for all positive definite matrices $IW(0, -3)$ and the proper prior $IW(I, 3)$ which implies a more reasonable range for the variance parameters as well as a uniform prior on $(-1, 1)$ for the correlation parameter, see Appendix A. In addition to these estimators we also analyze the data with the following estimators. Using the Mplus imputation method we analyze the data with the WLSMV estimator with 5 imputed data sets as well as 50 imputed data sets. The multiple imputation method is based on a Bayesian estimation of an unrestricted model

which is then used to impute the missing values. Multiple and independent imputations are created which are then analyzed using Rubin (1987) method. The unrestricted model used for imputation is the the sequential regression with observed mediators model which is the default method in Mplus for this kind of data. This approach was pioneered by Raghunathan et al. (2001). In addition, this model can be analyzed with the ML estimator. The ML estimator is guaranteed to provide consistent estimates since the missing data is MAR. Note again however that the ML estimator has some shortcomings that can not be overcome in general but for this particular model do not apply. The shortcomings of the ML estimators is that it can only be used with no more than 3 or 4 latent variables, otherwise the computational burden is so large that it becomes impractical. Another shortcoming of the ML estimator is that it can not be used with residual correlations between the categorical variables because that would lead to the use of the multivariate probit function that requires computationally demanding numerical integration. Finally the ML estimator does not provide a model fit based on an unrestricted multivariate probit model. The WLSMV estimator and the Bayes estimator both avoid the above mentioned shortcomings. The parameter values used in this simulation study are as follows $\mu_1 = 0.00$, $\mu_2 = 0.20$, $\psi_{11} = 0.50$, $\psi_{22} = 0.50$, and $\psi_{12} = 0.30$.

As expected again we see that the WLSMV estimates are biased while the Bayes estimates are close to the true values. In particular the mean of the random slope is underestimated dramatically by the WLSMV estimator while the Bayes estimator is consistent. Also the coverage for the WLSMV estimator is unacceptable. In addition we see that the results between the Bayes estimators with the two different priors are different and thus we have prior assumption dependence for this model. Even though we have a sample size of 1000, the growth model is somewhat difficult to identify because it uses only 5 binary variables. In this example again we see a clear advantage of using proper prior with bounded range rather than uniform improper prior. The proper prior leads to a decrease in the bias and improved coverage. Overall the four estimators, the Bayes estimator with $IW(I, 3)$ prior, the ML estimator, the WLSMV estimator with 5 imputed data sets, and the WLSMV estimator with 50 imputed data sets performed very well and there doesn't appear to be a substantial difference between these estimators. Increasing the number of imputed data sets from 5 to 50 does not seem to improve the results, i.e., 5 imputed data sets are sufficient. All of these 4 estimators are computationally fast and while they are more involved than the

Table 7: Bias(Coverage) for MAR dichotomous growth model.

Estimator	μ_1	μ_2	ψ_{11}	ψ_{22}	ψ_{12}
WLSMV	-0.03(.92)	-0.16(.02)	-0.23(.62)	0.09(.96)	-0.08(.68)
Bayes $IW(0,-3)$	-0.01(.94)	0.00(.93)	0.12(.88)	0.06(.85)	-0.02(.93)
Bayes $IW(I,3)$	-0.01(.92)	0.00(.93)	0.06(.97)	0.03(.89)	-0.02(.92)
ML	0.00(.95)	-0.01(.93)	0.05(.89)	0.01(.96)	-0.01(.97)
WLSMV (5 Imput.)	-0.01(.96)	0.00(.92)	0.06(.94)	0.04(.91)	0.00(.93)
WLSMV (50 Imput.)	-0.01(.93)	0.00(.93)	0.05(.95)	0.04(.91)	0.00(.94)

traditional WLSMV estimator, the improvement in the results is substantial. We conclude that in the presence of missing data the Bayes estimator offers a valuable alternative to the WLSMV estimator, both as a direct estimator or as an imputation method followed by the WLSMV estimator. The imputation method followed by the WLSMV estimator however has the advantage that is does not depend on priors selection and therefore can be considered as the most straight forward approach.

4 Small Sample Size

In this section we will show how to use the Bayesian estimator to overcome small sample size shortcomings of the maximum likelihood estimator. The ML estimates are consistent and when the sample size is sufficiently large the point estimates will be essentially unbiased and the asymptotic estimates for the standard errors can be used to provide a 95% confidence intervals. For small sample size however there is no guarantee that the point estimates will be unbiased nor that the confidence intervals have the desired coverage. In this section we provide a simple example that shows these shortcomings and also shows how the Bayesian estimator can be used to resolve this problems. We consider a two-level regression with a random intercept. In two-level models the parameters on the between level are essentially estimated by as many observations as there are clusters in the data, i.e., for the asymptotic to hold the number of clusters has to be sufficiently large. In practice however the number of clusters is often small, i.e., the ML asymptotic formulas are

often unreliable. The model that we investigate in this section is given by

$$Y_{ij} = \alpha + \beta X_{ij} + \eta_j + \varepsilon_{ij}$$

where Y_{ij} is the $i-$th observation in cluster j, X_{ij} is a standard normal covariate, η_j is zero mean normally distributed random effect, and ε_{ij} is a zero mean normal residual. The model has 4 parameters α, β, $\psi = Var(\eta_j)$, and $\theta = Var(\varepsilon_{ij})$. We generate data according to this model using the following parameter values $\alpha = 0$, $\beta = 1$, $\psi = 1$, and $\theta = 2$. We generate 100 data sets with M clusters each with 50 observations, i.e., the total sample size in each data set is $50M$. We analyze the generated data with the ML estimator as well as 3 different Bayes estimators. The Bayes estimators differ in the choice of prior distribution for the parameter ψ. The Mplus default prior is the improper prior $IG(-1, 0)$ which is equivalent to a uniform prior on $[0, \infty)$. We also estimate the model with the priors $IG(0, 0)$ and $IG(0.001, 0.001)$ which have been considered in two-level models, see Browne and Draper (2006) and Gelman (2006).

The results for the parameter ψ are presented in Table 8. The best results in terms of bias and coverage are obtained with the Bayes estimator and priors $IG(0, 0)$ and $IG(0.001, 0.001)$. The difference between the two priors are essentially non-existent. The ML estimator shows low coverage even for $M = 20$ and bigger bias even for $M = 60$. The Bayes estimator with default prior also performs poorly in terms of bias. Similar advantage of the Bayes estimator also occurs for the α parameter. This simulation shows that when the number of clusters is smaller than 50 the Bayes estimator can be used to obtain better estimates and more accurate confidence intervals in two-level models particularly when the between level variance components use priors such as $IG(0, 0)$.

5 Bayesian Estimation as the Computationally Most Efficient Method

In this section we describe models that are computationally challenging for traditional estimation methods such as maximum-likelihood but are now doable with the Bayesian estimation.

Table 8: Bias(Coverage) for ψ in two-level regression.

Estimator	M=5	M=10	M=20	M=40	M=60
ML	-.22(.64)	-.13(.82)	-.09(.86)	-.04(.91)	-.03(.93)
Bayes IG(0.001,0.001)	.13(.96)	.01(.94)	.03(.91)	.01(.93)	.01(.95)
Bayes IG(0,0)	.13(.96)	.01(.94)	.03(.93)	.01(.93)	.01(.95)
Bayes IG(-1,0)	1.88(.89)	.35(.95)	.16(.88)	.07(.91)	.04(.93)

5.1 Multilevel Random Slopes Models With Categorical Variables

Random effect models with categorical variables are usually estimated with the ML estimator however each random effect accounts for one dimension of numerical integration. The ML estimator is not feasible when the dimension of numerical integration is more than 3. Thus the maximum number of random effect the ML estimator can estimate in a two-level probit regression is 3 (one intercept and 2 random slopes). It is possible to use Montecarlo integration method with more than 3 random effects, however, such an approach usually require careful monitoring of the estimation. In particular the convergence in a maximum-likelihood estimation with Montecarlo integration is somewhat tricky because the usual methods that are based on monitoring the log-likelihood value or the log-likelihood derivatives will be difficult to use due to larger numerical integration error.

In this section we will evaluate the performance of the Bayes estimator for q random effects for $q = 1, ...6$. The estimation time for the ML estimator grows exponentially as a function of the number of random effects. This however is not the case for the Bayesian estimation where the computational time grows linearly as a function of the number of random effects, i.e., increasing the number of random effects is not an obstacle for the Bayes estimation.

We conduct simulations with different number of random effects starting from 1 effect to 6 random effects. Let the number of random effects be q. To generate data for our simulation studies we first generate q covariates X_k for $k = 1, ..., q$. We set $X_1 = 1$ so that we include the random intercept in this model. For $k = 2, ..., q$ we generate X_k as independent normally

distributed covariate with mean zero and variance 1. We also generate a normally distributed between level covariate W with mean 0 and variance 1. Let the binary dependent variable be U and denote by U_{ij} the observation i in cluster j. The two-level binary probit regression is given by

$$P(U_{ij} = 1) = \Phi\Big(\sum_{k=1}^{q} s_{kj} X_{kij}\Big)$$

$$s_{kj} = \alpha_k + \beta_k W_j + \varepsilon_{kj}$$

where ε_{kj} is a zero mean residual with variance covariance matrix Ψ of size q, i.e., the random effects are assumed to be correlated. We generate data using these parameter $\alpha_k = 0$, $\Psi = diag(0.5)$, $\beta_k = 0.7$.

Using the Bayes estimator we analyze the generated data using each of these 3 different prior for Ψ: $IW(0, -q-1)$, $IW(I, q+1)$, and $IW(2I, q+1)$, where I is the identity matrix, see Appendix A. The prior $IW(0, -q-1)$ is basically the default prior in Mplus and it has a constant density function over the definition domain, i.e., it is improper uniform prior. The $IW(I, q+1)$ is usually selected as a proper prior that has non-informative marginal distributions for the correlation parameters, i.e., the marginal prior distribution for the correlations between the random effects is uniform on $[-1, 1]$. The marginal distribution for the diagonal elements is $IG(1, 0.5)$ which has mode at 0.25. The only difference between the prior $IW(2I, q+1)$ and $IW(I, q+1)$ is that the marginal prior distribution for the diagonal element of Ψ is $IG(1, 1)$ which has a mode of 0.5. If we have a good reason to believe that the variances of the random effects are near 0.5 (which is the true value here) then we can choose $IW(2I, q+1)$ instead of $IW(I, q+1)$ as a proper prior. In these simulations studies we generate 200 clusters with 20 observations each for a total sample size of 4000. Note however that for the purpose of estimating the between level random effects distribution the sample size is 200, i.e., this sample size is within a range where the the prior assumptions can potentially affect the posterior.

The results of the simulations are presented in Table 9. We provide the results only for the parameter ψ_{11}. The quality of the estimation of the rest of the parameters is similar to that of $\psi_{1,1}$. Since we are holding the sample size constant here the "prior dependence issue" is arising as the model becomes more and more complicated. As we increase the number of random effects in the model there is less and less information about these random effects in the data which in turn results in bigger dependence on the prior assumptions.

The "prior dependence" is visible for $q = 4, 5, 6$ and it increases as q increases. The results obtained with the uniform improper prior $IW(0, -q-1)$ appear to be the worst, both in terms of bias and coverage. Similar to the results we obtained in Section 2.3 we have the problem that the uniform improper prior puts too much weight on out of reasonable range values. In contrast the two proper priors seem to yield a small bias and good coverage. In particular the bias obtained with the prior $IW(2I, q+1)$ is close to 0 in all cases.

How do we in practice choose priors that are appropriate and are likely to give minimal bias in the point estimates? The first step in this process is to choose several different priors that are quite vague such as the three priors used above and to compare the results. If the results are similar then we can conclude that the sample size is big enough and that choosing among different vague priors will not change the results. If however the results are not similar then we can conclude that we have "prior dependence" in the Bayes estimation and that we need to carefully consider the appropriateness of the different priors. In most cases improper uniform priors for categorical data should be considered unreliable. We can also choose among the different priors the prior that appears to be most in agreement with the various posteriors that we have obtained. In our example, both $IW(I, q+1)$ and $IW(2I, q+1)$ show that the variance point estimates are near 0.5 and thus the prior $IW(2I, q+1)$ can be considered more appropriate as it has a marginal distribution for $\psi_{1,1}$ with a mode of 0.5. When we are unable to provide a meaningful prior for parameters but we are forced to do so by the small sample size there is nothing wrong in choosing a prior that agrees to some extent with the posterior and this is exactly the strategy we followed in this example. We used a prior dependent posterior to make a more informative choice on the prior. The differences between the various posterior distributions are much smaller then those in the prior, and those selecting a prior that does not contradict any of the posterior distributions is a valid choice.

5.2 Small Random Effect Variance

When the variance of a random effect is very small the EM algorithm has a very slow rate of convergence even when aided by acceleration methods. On the other hand the Bayes estimator can take advantage of a prior specification that avoids the variances collapsing to zero problem. The inverse gamma prior will generally work well for this purpose, see Appendix A. The

Table 9: Bias (percent coverage) for $\psi_{1,1} = 0.5$ in two-level probit regression with q random effects.

Prior	$q = 1$	$q = 2$	$q = 3$	$q = 4$	$q = 5$	$q = 6$
$IW(0, -q - 1)$	0.03(90)	0.04(92)	0.04(96)	0.08(81)	0.10(84)	0.19(60)
$IW(I, q + 1)$	0.03(89)	0.02(93)	-0.01(97)	-0.01(95)	-0.04(97)	-0.05(92)
$IW(2I, q + 1)$	0.03(90)	0.03(93)	0.01(96)	0.02(97)	-0.01(97)	-0.01(97)

near zero variance for the random effect is a common problem in two-level regression models. Once it is known that the variance is near 0 the correct modeling approach is to replace the random effect coefficient with a standard regression coefficient. With such a specification the model will be much easier to estimate, however, typically we do not know that the variance is near zero and thus estimate the effect as a random effect and frequently the ML estimation method will produce either very slow convergence or non-convergence. To illustrate this we conduct the following simulation study. Consider again the two-level binary regression model as in the previous section

$$P(U_{ij} = 1) = \Phi\Big(\sum_{k=1}^{q} s_{kj} X_{kij}\Big)$$

$$s_{kj} = \alpha_k + \beta_k W_j + \varepsilon_{kj}$$

where $q = 3$ and again the first covariate is set to the constant 1 to include the random intercept in the model. Here we have one random intercept and two random slopes, i.e., the ML estimation will use only 3 dimensional integration which usually is not computationally heavy. We generate the data using the following parameters $\alpha_1 = 0$, $\alpha_2 = 0.2$ and $\alpha_3 = 0.2$, $\beta_k = 0.7$ for $k = 1, 2, 3$. The two covariates on the within level X_2 and X_3 and the between level covariate W are generated as standard normal variables. The ε_k variables are generated as independent with variance $\psi_{11} = \psi_{22} = 0.5$ and $\psi_{33} = 0$. The last parameter is the key parameter, i.e., the second random slope has a zero variance. We generate 100 data sets each with 200 clusters of size 20 and analyze them with both the ML and the Bayes estimator. For the Bayes estimation we use the Inverse-Wishart prior $IW(I, 4)$ for the random effects variance covariance parameters, see Appendix A. For the α_k and β_k

parameters we use uniform prior on $(-\infty, \infty)$ interval, i.e., a non-informative improper prior.

The ML estimator took on average 19 minutes to complete each replication while the Bayes estimator used only 5 seconds, i.e., the Bayes estimator is about 200 times faster than the ML estimator. The results are presented in Table 10. The bias for both methods is near 0 and the coverage near the 95% nominal rate with one exception. The Bayes estimator does not ever include the 0 value in the confidence interval for the variance parameter ψ_{33}, i.e., the coverage here is 0. This is always going to be the case. The posterior distribution for a variance parameter consists only of positive values and since the inverse-gamma prior (which is the marginal distribution obtained from the inverse-wishart prior, see Appendix A) has low prior for near zero values than even small positive values are not included in the posterior. Thus we see that the Bayes estimator will necessarily estimate zero variances to small but positive values, which essentially leads to the bias of 0.06 seen in Table 10. The Bayes estimator can not be used to test significance of the random effect variance. The same is true for the standard LRT test because of borderline issues that distort the LRT distribution. Instead DIC and BIC can be used to evaluate the need for a random effect. Finally we conclude that the Bayes estimator avoids the collapse at zero of the variance parameter which in turn results in mush faster estimation. However when the Bayes estimator gives a small positive value for a variance parameter we should always suspect that the true value is actually zero.

6 Posterior Predictive P-value

Several discrepancy functions have been implemented in Mplus to obtain posterior predictive P-values (PPP). The main one is the chi-square test of fit discrepancy function, see Asparouhov and Muthén (2010). This discrepancy function can be used to detect structural misspecifications in the model, i.e., the PPP method based on the classical test of fit discrepancy function can be used to test the structural model for misspecifications. Note also that the PPP method uses the estimated posterior distribution and evaluates how that posterior distribution and the model fits the data. Therefore the PPP method can be used also as a check for the posterior distribution of the parameter estimates. Since the posterior distribution depends on the prior distribution, the PPP method is also a test for the prior specifications for

Table 10: Bias (percent coverage) for small random effect variance estimation.

Parameter	ML	Bayes
α_1	0.01(90)	0.01(93)
α_2	0.01(95)	0.01(89)
α_3	0.00(96)	0.01(98)
β_1	0.01(96)	0.00(97)
β_2	0.00(98)	0.01(95)
β_3	0.00(95)	0.03(95)
ψ_{11}	0.01(96)	0.02(94)
ψ_{22}	0.01(93)	0.03(94)
ψ_{33}	0.01(99)	0.06(0)
ψ_{12}	0.00(97)	-0.01(95)
ψ_{13}	0.00(97)	0.00(98)
ψ_{23}	0.00(97)	0.01(97)

the parameters estimates.

In this section we illustrate the advantages and disadvantages of the PPP method. Four models are considered below: SEM with continuous variables, SEM with categorical variables, mixture with continuous variables and mixture with categorical variables.

6.1 Posterior Predictive P-value in SEM with Continuous Variables

In this section we study the power and the type I error for the PPP method and compare it to the classical likelihood ratio test (LRT) of fit, based on the ML estimator, and also the weighted least squares (WLSMV) test of fit. For the ML estimator we include the various robust LRT statistics given with the MLR, MLM and MLMV estimators. All of these tests are available in Mplus.

We begin by conducting a power analysis for a structural equation model with 15 indicator variables $y_1, ..., y_{15}$ measuring 3 latent factors η_1, η_2, and

η_3. The following equation describes the model

$$y = \mu + \Lambda\eta + \varepsilon$$

where y is the vector of 15 variables, η is the vector of the 3 latent factors with variance Ψ, and ε is the vector of 15 independent residuals with a variance covariance Θ, which is assumed to be a diagonal matrix. In this simulation we will study the ability of the tests of fit to reject the model when some small loadings in Λ are misspecified. We vary the sample size in the simulation to obtain approximate power curves for the three tests. Data is generated using the following parameters $\theta_i = 0.5$, $\mu_i = 0$, Ψ is the identity matrix of size 3 and

$$\Lambda = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0.2 \\ 1 & 0 & 0.2 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0.3 & 1 & 0 \\ 0.3 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{pmatrix}.$$

We estimate a misspecified model where the two loadings of size 0.2 are omitted from the model while all other non-zero loadings are estimated with one exception. For identification purposes we fix $\lambda_{1,1} = \lambda_{6,1} = \lambda_{11,1} = 1$. We also estimate the matrix Ψ. Since the model is misspecified we expect the tests of fit to reject the model for sufficiently large sample size. Table 11 contain the results from this simulation for various sample sizes. It is clear that the LRT test is the most powerful and it would in general reject the incorrect model more often than both the PPP and the WLSMV. All tests however will reject the incorrect model with sufficient sample size.

We now conduct a different simulation study designed to check the type I error of the tests, i.e., to determine how often the correct model is incorrectly

rejected. We generate the data using the same Λ matrix but with $\lambda_{4,3} = \lambda_{5,3} = 0$ and analyze the model using the correct specification. The rejection rates are presented in Table 12. The LRT rejects incorrectly the correct model much more often than both PPP and WLSMV. The rejection rates for LRT are much higher than the nominal 5% level for sample size 50 and 100. Among the four versions of the LRT test the MLMV performs best however the type I error for sample size 50 is still too high.

The conclusion of the above simulation study is that the LRT appears to be more powerful than the PPP and WLSMV but this is at the cost of incorrect type I error for small sample size cases, i.e., the use of LRT is only reliable when the sample size is sufficiently large. On the other hand the PPP is always reliable and for sufficiently large sample size has the same performance as the LRT, i.e., the PPP test is just as capable of rejecting incorrect models as LRT. Overall however the WLSMV test seems to perform better than both PPP and LRT. The WLSMV type I error is near the nominal level even when the sample size is small and for certain sample size values it appears to be more powerful than PPP. However the WLSMV would not be a possibility when there is missing MAR data. Thus the PPP seems to be the only universal test that works well in all cases.

Using simulation studies Savalei (2010) found that the MLMV estimator performs best in small sample sizes, however in our simulation we found a different result. The WLSMV estimator performed better than MLMV and in fact in the absence of missing data the WLSMV test statistic outperforms all other statistics. This exposes the problems with frequentist inference. Both MLMV and WLSMV methods are based on and designed for large sample size and have no guarantee to work well in small sample size. Simulation studies can favor one method over another however there is no guarantee that such a simulation result would be replicated in different settings. On the other hand the PPP method is designed so that it works independently of the sample size.

Note however that the PPP test appears to show a bias of some sort. Typical tests of fit will reach the nominal 5% rejection rate when the model is correct. Here we see however that the PPP is below the 5% rejection rate even for large sample size cases. This discrepancy is due to the fact that the PPP value is not uniformly distributed as the P-value in classical likelihood ratio tests, see Hjort et al. (2006).

Table 11: Rejection rates of LRT, PPP and WLSMV for a misspecified CFA.

Sample Size	50	100	200	300	500	1000	5000
LRT-ML	0.54	0.42	0.46	0.58	0.94	1.00	1.00
LRT-MLM	0.64	0.43	0.47	0.61	0.94	1.00	1.00
LRT-MLMV	0.25	0.20	0.36	0.49	0.91	1.00	1.00
LRT-MLR	0.67	0.44	0.48	0.61	0.94	1.00	1.00
PPP	0.04	0.16	0.29	0.47	0.77	0.99	1.00
WLSMV	0.09	0.18	0.44	0.69	0.95	1.00	1.00

Table 12: Rejection rates of LRT, PPP and WLSMV for a correctly specified CFA model.

Sample Size	50	100	200	300	500	1000	5000
LRT-ML	0.37	0.21	0.11	0.08	0.07	0.02	0.04
LRT-MLM	0.46	0.21	0.11	0.08	0.07	0.02	0.04
LRT-MLMV	0.20	0.08	0.07	0.08	0.05	0.02	0.04
LRT-MLR	0.47	0.24	0.11	0.08	0.07	0.02	0.04
PPP	0.01	0.05	0.02	0.00	0.02	0.01	0.01
WLSMV	0.04	0.05	0.05	0.04	0.03	0.03	0.06

Table 13: Rejection rates of LRT and PPP of 0.1 misspecified loadings.

Sample Size	300	500	1000
LRT	0.19	0.21	0.44
PPP	0.06	0.12	0.29

Table 14: Rejection rates of LRT and PPP of 0.3 misspecified loadings.

Sample Size	300	500	1000
LRT	0.96	1.00	1.00
PPP	0.87	0.99	1.00

6.2 Posterior Predictive P-value as an Approximate Fit

From the previous section we see that the PPP rejects less often than the LRT test. In practice this can be viewed as a positive contribution rather than as a lack of power. It is often the case that the LRT chi-square test of fit rejects a model because of misspecifications that are too small from practical point of view. In this section we explore the possibility to use PPP instead of the LRT as a test that is less sensitive to misspecifications. Using the same example as in the previous section we consider the rejection rates when omitting the cross loadings $\lambda_{4,1}$ and $\lambda_{5,1}$. When the true values of these loadings are less than 0.1 on standardized scale we would want the test not to reject the model and if the true values are above 0.3 on standardized scale we would want the test to reject the model. We construct two simulation studies. In the first we generate the data using crossloadings $\lambda_{4,1} = \lambda_{5,1} = 0.1$ and analyze it without these loadings, i.e., assuming that they are zero. In the second simulation study we generate the data using crossloadings $\lambda_{4,1} = \lambda_{5,1} = 0.3$ and analyze it without these loadings. The rejection rates are presented in Tables 13 and 14. From these results it seems that to some extent the PPP fulfills this role. For a small loss of power to reject the big misspecifications we reduce the "type I error" of rejecting the model because of small misspecifications.

Table 15: Rejection rates of PPP and WLSMV for a misspecified CFA with categorical variables.

Sample Size	200	300	500	1000	2000	5000
PPP	0.00	0.00	0.00	0.03	0.36	1.00
WLSMV	0.56	0.86	0.99	1.00	1.00	1.00

6.3 Posterior Predictive P-value in SEM with Categorical Variables

When the variables are categorical we generate the underlying continuous variables Y^*. Using Y^* we can compute the LRT test of fit function just as we do for the models with continuous variables. Using this function as a discrepancy function we compute the PPP value to evaluate the model.

In this section we will conduct a simulation study to evaluate the performance of the PPP test and to compare it to the WLSMV test of fit. Both the type I error and the power of the tests are considered.

We evaluate the power of the two tests on a factor model with 15 binary variables and 3 factors. We use the same parameter setup as in the previous section, with the exception of the Θ matrix which for identification purposes is fixed to the identity matrix. All the thresholds are zero. We generate the data according to this model but analyze the data according to the model which does not include the small cross loadings $\lambda_{9,1}$, $\lambda_{10,1}$, $\lambda_{4,3}$, $\lambda_{5,3}$. The rejection rates for the two tests are presented in Table 15. The results suggest that the WLSMV chi-square is much more powerful than PPP. Unlike in the continuous case the difference between the power is quite dramatic. One reason for why this may be the case is because the Y^* are generated from the estimated model, i.e., it will be difficult to detect misspecification that way. In contrast the WLSMV chi-square test is directly related to the data via the polychoric matrix. Nevertheless we see from the results in Table 15 that given sufficient sample size the PPP will reject the incorrect model with certainty. Smaller sample size cases were not included in the above comparison because the Bayes estimator had convergence problems with the default non-informative priors.

Next we consider the type I error for the two tests. We generate the data using a loading matrix as above but without the small cross loadings, i.e.,

Table 16: Rejection rates of PPP and WLSMV for a correctly specified CFA with categorical variables.

Sample Size	200	300	500	1000	2000	5000
PPP	0.00	0.00	0.00	0.00	0.00	0.00
WLSMV	0.03	0.03	0.05	0.03	0.06	0.05

$\lambda_{9,1} = \lambda_{10,1} = \lambda_{4,3} = \lambda_{5,3} = 0$ and we analyze the data according to the correct model. The rejection rates are presented in Table 16. Both tests do not exceed the nominal 5% rate and therefore have an acceptable type I error. The fact that all rejection rates for PPP are zero also suggests that there is a conservative bias in the PPP procedure.

With the exception of binary variables the above PPP method does not address the model fit when it comes to thresholds and mean structures for categorical variables. The Mplus technical output 10 will provide PPP values that address this part of the model using discrepancy functions such as univariate likelihoods and the observed proportion for each category.

We conclude that the WLSMV test of fit is more powerful than the PPP test. The PPP test however is useful in situations where the WLSMV test can not be used. For example situations where there is missing MAR data or when there are informative priors in the Bayes estimation.

6.4 Using Posterior Predictive P-value to Determine the Number of Factors

In this section we consider the power of PPP to determine the number of factors in a factor analysis model with binary indicators. Data is generated according to a two factor analysis model

$$Y^* = \Lambda\eta + \varepsilon$$

where Y^* is a vector of 20 normally distributed variables, η is a vector of 2 normally distributed factors and ε is a vector of 20 normally distributed residuals with mean zero and variance one. The observed binary variables are obtained by

$$Y_i = 0 \Leftrightarrow Y_i^* < \tau_i$$

Table 17: Rejection rates of PPP and WLSMV for a two factor model specified as one factor model.

Sample Size	50	100	200	300	500
PPP	0.00	0.02	0.40	0.93	1.00
WLSMV	0.66	0.96	1.00	1.00	1.00

for $i = 1, ..., 20$. The loading matrix is such that the first 10 observed variables load on the first factor and the second 10 load on the second factor, i.e., $\lambda_{i,1} = 1$ for $i = 1, ..., 10$ and $\lambda_{i,2} = 1$ for $i = 11, ..., 20$. All other entries are 0 in the loading matrix. The τ parameters are $\tau_i = 0$ for $i \leq 10$ and $\tau_i = -0.5$ for $i > 10$. The variance covariance matrix for η is

$$\Psi = \begin{pmatrix} 1 & 0.5 \\ 0.5 & 0.8 \end{pmatrix}.$$

We generate 100 data sets of various sample sizes and analyze it as a one factor model with the Bayes and the WLSMV estimators. Table 17 shows the rejection rates for the PPP and the WLSMV tests. The results here again confirm that WLSMV is more powerful than PPP however we can also see here that the power of PPP is quite good as well. For sample size of 300 and more the rejection rate is near 100%. The results from a separate simulation study, not reported here, using data generated and analyzed according to a 20 binary indicators and one factor model showed that both tests have acceptable type I error and the correct model was rejected below or near the 5% nominal rate.

6.5 Posterior Predictive P-value in Mixture Analysis

In mixture analysis there is no natural unrestricted model that can be used to test against and thus the LRT is not available as a test of fit. On the other hand the PPP based on the chi-square test of fit is available because at each MCMC iteration the C variable is generated and thus the chi-square test of fit is simply computed as it would be computed in a multiple group analysis. In this section we evaluate the performance of the PPP test in latent class analysis (LCA) and latent profile analysis (LPA).

6.5.1 PPP in LCA

Consider a two class LCA model with 10 binary indicators. Using the probit link in class 1 all thresholds are -1 and in class 2 all thresholds are 1. The two classes are of equal size. The most common assumption in Mixture models is that the class indicators are conditionally independent, i.e., conditional on the latent class variable the indicators are independent. In this section we study the ability of the PPP method to detect such misspecifications. We consider three simulation studies. In simulation A we generate data so that all indicators are conditionally independent and analyze it that way. In simulation B we generate data so that all indicators are conditionally independent with the exception of $[Corr(Y_1^*, Y_2^*)|C = 1] = 0.7$, but analyze it as if the indicators are independent, i.e., this model is misspecified. In simulation C we generate data as in simulation B but analyze it using the correct model specification, i.e., this model is correctly specified. Note that simulation C can be analyzed in Mplus only with the Bayes estimator, but not with the ML estimator because it requires a multivariate probit function. Simulations A and B can be analyzed with the Bayes and the ML estimators, however the ML estimator does not provide a test of fit.

The results of the simulation study are presented in Table 18. The rejection rates for the correctly specified models are all 0. In addition, the incorrectly specified model is rejected with certainty as the sample size increases.

It is also seen in this table as well as in the previous simulation study that the power of the PPP is low and large sample sizes are needed for the test to reject misspecified models.

It is interesting to note here that there are no other well-established reliable tests of fit for this LCA model. The Pearson and the log-likelihood ration chi-square tests have more than 1000 degrees of freedom and when the degrees of freedom are so large it is well known that the tests are unreliable.

Note also that in this estimation process the mixture latent variable can absorb some of the model misspecifications when the number of indicators is smaller. With 10 indicators however the classes are quite well separated and the misspecifications remain in the model.

Table 18: Rejection rates of PPP in LCA models. In simulation A and C the mode is correctly specified while in simulation B it is misspecified.

Sample Size	500	1000	2000	5000
A	0.00	0.00	0.00	0.00
B	0.01	0.03	0.28	1.00
C	0.00	0.00	0.00	0.00

Table 19: Rejection rates of PPP in LCA models. In simulation A and C the mode is correctly specified while in simulation B it is misspecified.

Sample Size	50	100	200	300	500	1000	5000
A	0.05	0.05	0.03	0.01	0.02	0.02	0.07
B	0.17	0.55	0.90	0.99	1.00	1.00	1.00
C	0.07	0.02	0.06	0.03	0.03	0.02	0.02

6.5.2 PPP in LPA

Latent profile analysis (LPA) is similar to LCA except that here all class indicator variables are continuous. In the simulations we use 10 continuous indicators with means -1 in class one and 1 in class two. The variance of the variables is the same across the classes and it is set to 1. As in the previous section we study the ability of PPP to detect within class dependence. The simulation A, B and C are constructed as in the previous section. The results are presented in Table 19. As expected in simulation studies A and C the models are rejected near the 5% nominal rate while in simulation B the model is rejected almost with certainty even when the sample is small. This indicates that the PPP test is fairly powerful for Mixture models with continuous variables. In this case there is no alternative test of fit also.

6.5.3 Using PPP to Determine the Number of Classes in Mixture Models

In certain mixture models, such as growth mixture models, the within class model is usually not modified, but rather, the number of classes is increased

until a satisfactory fit to the data is obtained. Thus the main modelling question boils down to determining the number of classes needed in the model to fit the data well. Several techniques have been proposed and are widely used in practice, see Nylund et al. (2007). None of these however has been universally accepted because of various shortcomings which will not be discussed here. In this section we illustrate how to use the PPP method as a class enumeration technique. Consider a quadratic growth mixture model

$$Y_{it} = \eta_{0i} + \eta_{1i}t + \eta_{2i}t^2 + \varepsilon_{it}$$

where Y_{it} is a normally distributed outcome for observation i at time t and the distribution of the random effects η_{ji} is given by

$$[\eta_{ji}|C_i = k] = \alpha_{jk} + \zeta_{ji}$$

where C_i is the latent class variable. The residual variable

$$\zeta_i = (\zeta_{0i}, \zeta_{1i}, \zeta_{2i})$$

has a variance covariance Ψ that is the same across the classes, i.e., the three random effects have means varying across the classes but their variance covariance is the same across the classes. The residuals ε_{it} have a diagonal variance covariance Θ which is also independent of the class variable.

We generate data according to a 4 class model and estimate it according to a 4 class model and a 3 class model. We expect the PPP method to reject the 3 class model and accept the 4 class model. We use the parameter estimates obtained for the 4-class model estimated for the STAR*D antidepressant data, see Muthén et. al. (2010). Table 20 contains the rejection rates for the estimated models. It is clear from these results that the PPP works correctly and it can be used to determine the number of classes. For sample size of 1000 or more the test has substantial power to reject the model with insufficient number of classes.

Note that there are certain similarity between PPP and the Bootstrap LRT test for testing between k class model and a $k-1$ class model (implemented in Mplus technical 14 output). Both are based on simulating the LRT values. The difference is however that the Bootstrap LRT is based on the LRT between the k class model and a $k-1$ class model, while the PPP is based on the k class model and the unrestricted k class model. A full comparisons between these methods is beyond the scope if this article.

Table 20: Using PPP as a class enumeration method. Rejection rates for the 3-class and 4-class models.

Sample Size	500	1000	2000	4000
3-class model	0.13	0.66	0.99	1.00
4-class model	0.00	0.00	0.00	0.01

7 Two-Part Growth Modeling

Two-part modeling is used to model distributions that have a large percentage of zero values. Consider as an example the outcome of heavy drinking, measured by the question: How often have you had 6 or more drinks on one occasion during the last 30 days? The answer to this question essentially has two separate pieces that can be modeled separately. The first piece is if the value is positive or not, i.e., if the subject engages in heavy drinking. The second piece is if the subject engages in the heavy drinking activity, how often it happens. If for example 75% of the observations have a 0 value it would not be appropriate to fit a normal distribution to this data. Instead modeling the two separate pieces of information would lead to a better model fit for this kind of data. To formalize this suppose Z is the observed variable. We define two new variables a binary variable U to indicate activity engagement and the second variable Y is the actual value when such an activity exist.

$$U = \begin{cases} 0 & \text{if } Z = 0 \\ 1 & \text{if } Z \geq 0 \end{cases} \tag{3}$$

$$Y = \begin{cases} \text{missing} & \text{if } Z = 0 \\ Z & \text{if } Z \geq 0 \end{cases} \tag{4}$$

At this point one can use standard models such as logistic or probit regression for U and linear regression for Y to construct elaborate models. For more information on two-part modeling see Olsen and Schafer (2001). In longitudinal setting we have multiple observations for the same individual and thus we can estimate a growth model for the U variables as well as the Y variables. Typically the random effects in the two growth models are correlated and therefore we estimate a joint model for U and Y. Because the missing values

for Y are directly predicted by the U variables we have MAR situation which is not MCAR. This means that the WLSMV estimator is not appropriate. The WLSMV estimates for this model are usually very biased and thus we will not include the WLSMV estimator in this discussion. The ML estimates are unbiased however numerical integration is required. In the example that we describe below we use a quadratic growth model for the U variables which means that the numerical integration is 3 dimensional, which is feasible but can be quite slow. Thus the Bayes estimator is of interest as a potentially more computationally efficient estimator. Another reason to use the Bayes estimator for this model is the fact that it provides a model fit through the PPP value. Such model fit is not available for the ML estimator. The PPP value essentially will compare the two-part model against an unrestricted variance covariance matrix for all Y and U^* variables thus it will be useful in evaluating the fit of the growth curves. It is not possible to estimate the same unrestricted model with the ML estimator because it typically leads to high dimensional numerical integration which is not feasible.

In this section we conduct a simulation study based on the two-part growth example presented in Muthén (2010). There are 5 repeated measures for each individual and there are 1192 individuals in the sample. Both the U part of the model and the Y part of the model are fitted to a quadratic growth curve. The model can be described as follows.

$$P(U_{ij} = 1) = \Phi(\eta_{1i} + \eta_{2i}t_j + \eta_{3i}t_j^2)$$

$$Y_{ij} = \eta_{4i} + \eta_{5i}t_j + \eta_{6i}t_j^2 + \varepsilon_{ij}$$

The random effects $\eta_{1i}, ..., \eta_{6i}$ are estimated as correlated random effects with mean α and variance covariance Ψ. These parameters amount to 27 parameters, in addition we have the 5 residual variance parameters θ_j for ε_{ij} so in total the model has 32 parameters. To generate data we use the parameter values reported in Muthén (2010) with one exception. The variance parameter ψ_{66} was estimated as 0.02. As it was explained in Section 5.2 such a value should be interpreted as 0 in Bayes estimation. Thus we generate the data using $\psi_{6k} = 0$ for $k = 1, .., 6$ and analyze the data with the same model, where these parameters are fixed to 0, i.e., the quadratic effect is a fixed effect for the Y part of the model. Eliminating the 6 Ψ parameters we get a model with 26 parameters. Here we describe the simulation study based on this model. We also conducted a simulation study with the small ψ_{6k} values however the Bayes estimator will always push the values away from

zero and that creates bias not just for the variance parameter but also for other parameters in the model. The most appropriate approach is to eliminate small variance random effects and convert them into fixed effects. If the small variance random effect is considered important the time variables t_j can be rescaled (for example they can be divided by a factor of 2) so that variance parameter is estimated to a larger value.

In the simulation study we use an Inverse-Wishart prior IW(I,6) for the variance covariance matrix Ψ, see Appendix A. Using this prior is important. As we have seen previously for models that are somewhat difficult to identify the priors can have an effect on the results even when the sample size is large. For this example using the default Mplus prior of IW(0,-6) leads to worse estimates and coverage. The prior for the α parameters is uniform on $(-\infty, \infty)$ and the prior for the θ parameters is $IG(-1, 0)$. Both of these are the Mplus defaults and these priors as usual have little influence on the estimates. The results of the simulation study for the Bayes and the ML estimator are presented in Table 21. The computational time for the ML estimator is twice as much as that for the Bayes estimator (2 minutes per replication v.s. 1 minute per replication). The convergence rate for the Bayes estimator is 100% while for ML it is 95%. Both the Bayes estimator and the ML estimator produce small bias and good coverage. The bias of the Bayes estimates is slightly bigger and for some of the parameter the coverage drops down to 83% however. Since the model is somewhat difficult to identify (for example here 3 random effects are identified from 5 binary variables) we can assume that if the MCMC sequence is run longer the results will improve. In the simulation study we use 20000 MCMC iterations for each replication. The PPP value did not reject the model, i.e., the model was accepted in every replication in the simulation study. On the other hand when we analyze the same data but using only a linear growth model rather than a quadratic in both the U and the Y parts of the model then the model is rejected 100% of the time. Therefore the PPP can indeed be used as a test of fit for the two-part model. This feature is very valuable alone in the absence of any other alternatives.

Table 21: Absolute bias(percent coverage) for two-part model

Parameter	True Value	ML	Bayes
α_1	-0.81	0.01(93)	0.00(97)
α_2	-0.68	0.00(95)	0.12(88)
α_3	-0.31	0.00(96)	0.06(83)
α_4	0.42	0.00(95)	0.00(98)
α_5	-0.16	0.00(92)	0.02(91)
α_6	-0.09	0.00(94)	0.00(90)
ψ_{11}	4.11	0.06(98)	0.03(95)
ψ_{22}	3.03	0.09(96)	0.28(90)
ψ_{33}	0.33	0.01(96)	0.08(83)
ψ_{44}	0.22	0.00(98)	0.00(98)
ψ_{55}	0.18	0.00(93)	0.00(96)
ψ_{21}	1.96	0.03(97)	0.15(98)
ψ_{31}	0.54	0.01(97)	0.10(95)
ψ_{32}	0.94	0.03(97)	0.14(89)
ψ_{41}	0.72	0.02(97)	0.01(98)
ψ_{42}	0.31	0.00(97)	0.04(96)
ψ_{43}	0.09	0.00(98)	0.02(90)
ψ_{51}	-0.01	0.01(95)	0.01(93)
ψ_{52}	0.33	0.00(97)	0.00(92)
ψ_{53}	0.10	0.00(99)	0.00(93)
ψ_{54}	0.06	0.00(94)	0.00(96)
θ_1	0.10	0.00(92)	0.00(94)
θ_2	0.20	0.00(94)	0.00(95)
θ_3	0.21	0.00(95)	0.00(95)
θ_4	0.19	0.00(98)	0.00(96)
θ_5	0.18	0.00(91)	0.00(93)

8 Multiple Indicator Growth Modeling for Categorical Variables. Comparing the Efficiency of the Bayes and the WSLMV Estimators.

In this section we simply demonstrate the quality of the Bayes estimation when there is a large number of factors and a large number of categorical indicators. If there is missing data the WLSMV estimator could be inappropriate because the missing data could be MAR and only full information estimation methods such as the Bayes and the ML estimators are appropriate. Another reason for preferring a full information method is the fact that full information methods are asymptotically the most efficient, i.e., yield the minimal mean squared error MSE. When the number of indicators is large however the ML estimator is not computationally feasible if the factors are measured by categorical variables because that leads to numerical integration that is extremely heavy computationally with 4 or more factors. It is possible to use Montecarlo integration methods for such models to obtain the ML estimates however that method is somewhat more difficult to conduct. It suffers from frequent non-convergence problems due to the fact that estimates for the log-likelihood can contain substantial error which will leads to difficulty in maximizing the log-likelihood. Typically the EM algorithm used for the ML estimation produces monotonically increasing likelihood which enables us to easily monitor convergence. With the Montecarlo integration however that happens only for certain parameterizations. In general the integration error will results in a non-monotonic log-likelihood which becomes a problem in evaluating the convergence.

In this section we use as an example the multiple indicator growth model described in Muthén (2010). This example combines IRT and growth modeling using nine binary indicators of a factor measured at eight time points. The ML estimation requires eight dimensions of integration and thus it is not feasible. We use the parameter values reported in Muthén (2010) to construct a simulation study. We generate 100 data sets and analyze them with the Bayes and the WLSMV estimators. The sample size in the original example is 1174 and we generate all data sets to be of that size. The model can be described as follows. For $t = 1, ..., 8$; $j = 1, ..., 9$; $k = 0, 1, 2$ and $i = 1, ..., 1174$

$$P(U_{ijt} = 1) = \Phi(\nu_j + \lambda_j \eta_{it})$$

$$\eta_{it} = \zeta_{0i} + \zeta_{1i}t + \zeta_{2i}t^2 + \varepsilon_{it}.$$
$$\zeta_{ki} = \alpha_k + \beta_k W_i + \xi_{ik}.$$

where all of the latent variables η_{it}, ζ_{ki}, ε_{it}, and ξ_{ik} are normally distributed and W_i is a binary covariate. For identification $\lambda_1 = 1$ and $\alpha_0 = 0$. The free parameters in the model are the 9 parameters ν_j, the 8 parameters λ_j , the 8 residual variances θ_t of ε_{it}, the 2 parameters α_k, the 3 parameters β_k, and the 6 parameters in the multivariate $N(0, \Psi)$ distribution of ξ_{ik} for a total of 36 parameters. The variance ψ_{22} is estimated to a value near 0 in the original example so we are going to again fix the quadratic random effect to 0, essentially eliminating the 3 parameters ψ_{k2}, i.e., the model has 33 parameters and this is the model we use for generating the data and the model we estimate. For the Bayes estimation we use the following priors. For the Ψ matrix we use the inverse-wishart prior $IW(I, 3)$, for the parameters θ_t we use the inverse-gamma prior $IG(-1, 0)$ and for all other parameters we use the uniform prior on $(-\infty, \infty)$. The results of the simulation study are presented in Table 22. Both estimators yield unbiased estimates and confidence interval coverage near the 95% nominal level. The Bayes estimates yield smaller MSE than the WLSMV estimates for most parameters, i.e., the Bayes estimates can be considered more accurate than the WLSMV estimates. This confirms the well known theoretical results that full information methods are asymptotically most efficient. Note that the Bayes and the ML estimator are asymptotically equivalent. In addition the PPP value accepted the model 100% of the time while the WLSMV estimator accepted the model 93% of the time. The computational time for the Bayes estimator is about 1.5 times the computational time for the WLSMV estimator (2.5 v.s. 3.5 minutes per replication). The convergence criteria used with the Bayes estimator is the automated PSR convergence criterion implemented in Mplus. We conclude that the Bayes estimator provides a valid full-information alternative to the WLSMV estimator and can be used for example to ensure that missing data is properly accounted for or to ensure that the most efficient estimates are obtained.

Table 22: Absolute bias(percent coverage) for multiple indicator growth model

Parameter	True Value	WLSMV	Bayes	WLSMV-MSE / Bayes-MSE
ν_1	0.39	0.00(94)	0.00(90)	1.21
ν_2	0.2	0.00(94)	0.00(88)	1.34
ν_3	-1.66	0.02(92)	0.00(87)	1.35
ν_4	-1.81	0.02(96)	0.00(94)	1.51
ν_5	-0.39	0.00(92)	0.00(90)	1.46
ν_6	-1.35	0.01(93)	0.01(93)	1.83
ν_7	-0.83	0.00(92)	0.00(93)	1.34
ν_8	-0.67	0.01(98)	0.00(90)	1.33
ν_9	-0.05	0.00(92)	0.00(90)	1.19
λ_2	1.40	0.00(97)	0.01(89)	1.22
λ_3	1.89	0.01(95)	0.02(87)	1.73
λ_4	1.34	0.01(96)	0.02(91)	1.74
λ_5	1.18	0.00(94)	0.02(87)	1.29
λ_6	1.42	0.01(92)	0.01(88)	1.76
λ_7	1.42	0.00(96)	0.01(89)	1.30
λ_8	1.32	0.01(93)	0.01(90)	1.60
λ_9	1.29	0.01(96)	0.01(88)	1.19
θ_1	0.15	0.01(96)	0.02(92)	1.36
θ_2	0.49	0.02(94)	0.02(91)	1.82
θ_3	1.72	0.04(97)	0.04(92)	2.05
θ_4	0.87	0.02(97)	0.00(92)	1.61
θ_5	0.69	0.00(94)	0.01(95)	2.17
θ_6	0.75	0.01(92)	0.03(93)	1.84
θ_7	0.96	0.01(98)	0.08(87)	0.85
θ_8	0.53	0.02(94)	0.06(84)	1.16
α_1	0.00	0.00(97)	0.00(90)	0.91
α_2	0.00	0.00(92)	0.00(94)	1.00
β_0	0.62	0.01(97)	0.00(90)	0.89
β_1	0.11	0.00(100)	0.00(92)	0.78
β_2	-0.02	0.00(96)	0.00(93)	1.00
ψ_{11}	1.77	0.05(93)	0.04(91)	1.81
ψ_{22}	0.25	0.01(93)	0.01(91)	1.25
ψ_{21}	-0.41	0.01(92)	0.00(92)	1.83

9 Appendix A. Priors

Choosing meaningful priors is important when the estimates and the posterior distribution show "prior dependence". In addition when informative priors are available it is important to specify these priors correctly. In this section we provide a brief tutorial for how to setup the Inverse Gamma and Inverse Wishart priors.

9.1 Inverse Gamma Prior

The inverse gamma prior $IG(\alpha, \beta)$ has a positive density on $(0, \infty)$. Figures 1-6 show some examples of inverse gamma density function. The density function is given by

$$\frac{\beta^{\alpha}}{\Gamma(\alpha)} x^{-\alpha-1} Exp(-\beta/X)$$

where Γ is the gamma function. The mean of the distribution is

$$\frac{\beta}{\alpha - 1}$$

when $\alpha > 1$, otherwise it is infinity. The variance of the distribution is

$$\frac{\beta^2}{(\alpha-1)^2(\alpha-2)}$$

when $\alpha > 2$, otherwise it infinity. To setup an informative conjugate inverse gamma prior for a variance parameter one can simply solve the system of equations for α and β

$$m = \frac{\beta}{\alpha - 1}$$

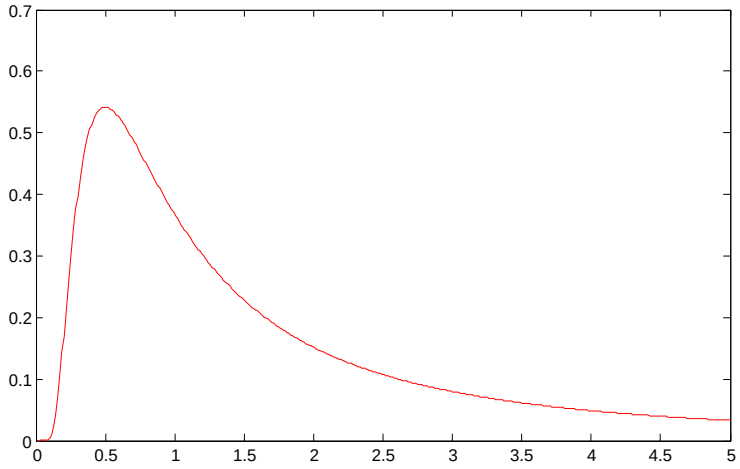
$$v = \frac{\beta^2}{(\alpha-1)^2(\alpha-2)}$$

where m and v are the mean and the variance of the informative prior. Fortunately this system is very easy to solve.

$$\alpha = 2 + \frac{m^2}{v}$$

$$\beta = m + \frac{m^3}{v}$$

Figure 1: $IG(1,1)$ density



When α is small and the variance or the mean is infinity it is useful to consider the mode of the distribution

$$\frac{\beta}{\alpha+1}.$$

The parameter β is a scale parameter for the inverse gamma distribution. Thus if $X \sim IG(\alpha,1)$ then $\beta X \sim IG(\alpha,\beta)$. For this distribution to be proper both parameters α and β should be positive. The Mplus default of $IG(-1,0)$ is basically the uniform prior on $(0,\infty)$.

9.2 Inverse Wishart Distribution

The domain of the Inverse Wishart distribution $IW(\Psi,m)$ is all positive definite matrices of size p. The density is given by

$$\frac{|\Psi|^{m/2}|X|^{-(m+p+1)/2}Exp(-Tr(\Psi X^{-1})/2)}{2^{mp/2}\Gamma_p(m/2)}$$

where Γ_p is the multivariate gamma function and the argument X of the density is a positive density function. To set informative prior with certain expected value we can use the fact that the mean of the distribution is

$$\frac{\Psi}{m-p-1}.$$

Figure 2: $IG(2,1)$ density

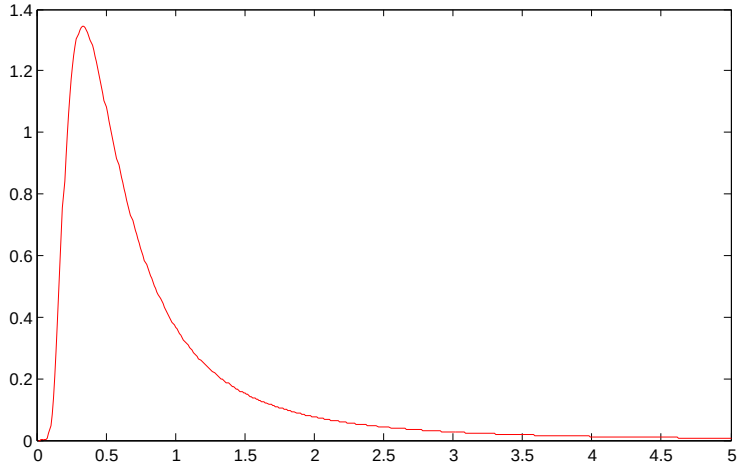


Figure 3: $IG(3,1)$ density

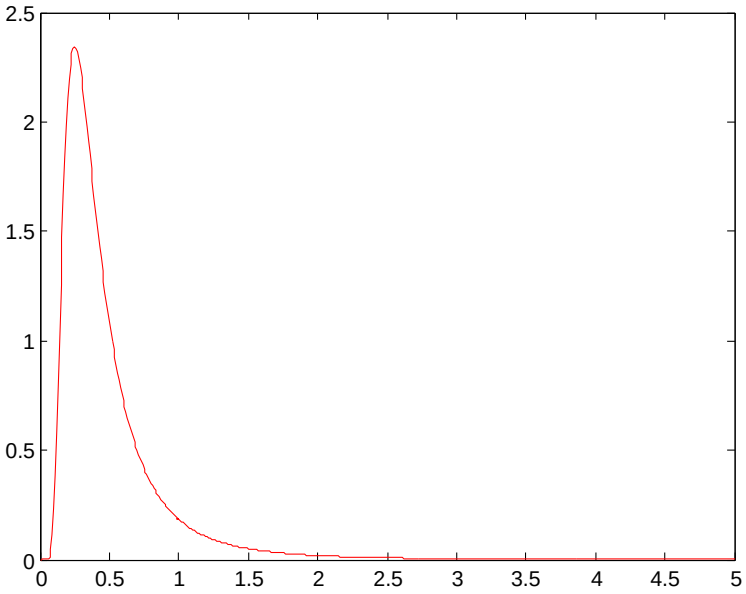


Figure 4: $IG(1, 2)$ density

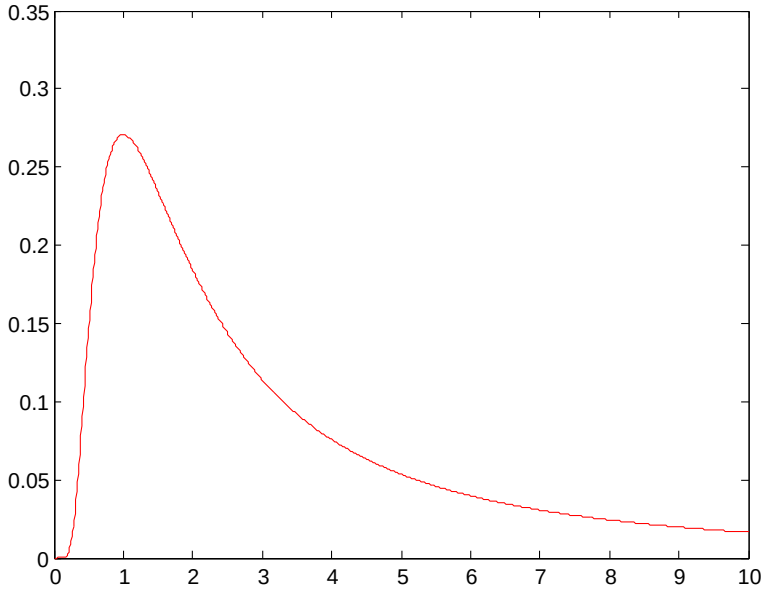


Figure 5: $IG(0.001, 0.001)$ density

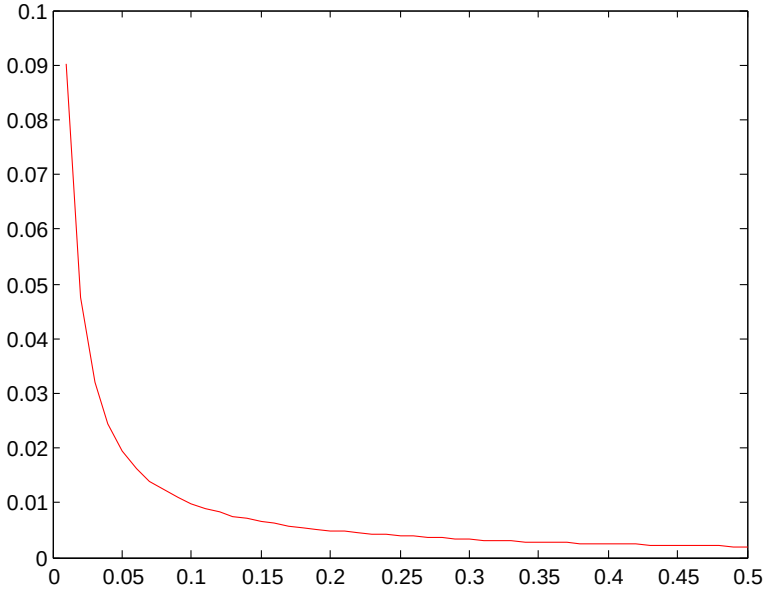
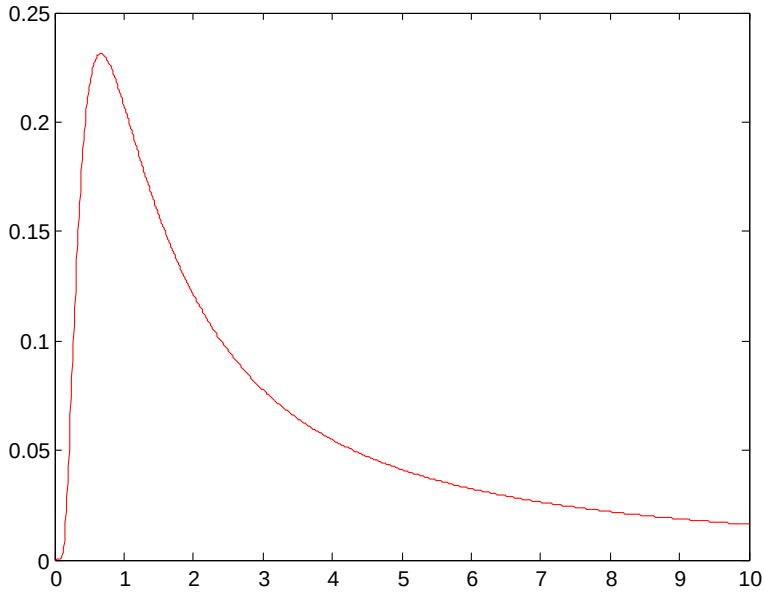


Figure 6: $IG(0.5, 1)$ density



The mean exist and is finite only if $m > p+1$. If $m \leq p+1$ then we can use the fact that the mode of the distribution is

$$\frac{\Psi}{m+p+1}.$$

The variance, i.e., the level of informativeness is controlled exclusively by the parameter m. The larger the value of m the more informative the prior is. To evaluate the informativeness of the prior one should consider the marginal distribution of the diagonal elements. The marginal distribution of the $i-$th diagonal entry is

$$IG((m-p+1)/2, \psi_{ii}/2).$$

To set informative prior with certain variance we can use the approach described in the previous section for this marginal prior which will determine the value of m. Then we multiply the desired expected value by $(m-p-1)$ to get Ψ. It is clear that in this process the level of informativeness of Inverse Wishart priors is rigid and the informativeness of one parameter in the matrix determines the informativeness of all other parameters. This shows that the Inverse Wishart prior may be insufficiently flexible for some applications.

Let's describe also a special case of a prior that is particularly useful. If you set the prior to $IW(D, p+1)$ where D is a diagonal matrix then marginal distribution for all correlations is uniform on the interval $(-1, 1)$ while the marginal distributions of the variance is $IG(1, d_{ii}/2)$. The values of the diagonal elements d_{ii} can be set to match the mode of the desired prior with the mode of $IG(1, d_{ii}/2)$ which is $d_{ii}/4$.

More information on the Inverse Wishart distribution and the marginal distributions of all the entries in the matrix can be found in Barnard et al. (2000).

References

[1] Asparouhov, T. and Muthén, B. (2010) Bayesian Analysis Using Mplus. Mplus Technical Report. http:\\www.statmodel.com

[2] Barnard, J., McCulloch, R. E., and Meng, X. L. (2000). Modeling covariance matrices in terms of standard deviations and correlations, with application to shrinkage. Statistica Sinica 10, 1281-1311.

[3] Browne, W. J., and Draper, D. (2006). A comparison of Bayesian and likelihood-based methods for fitting multilevel models. Bayesian Analysis, 1, 473-514.

[4] Dunson D., Palomo J., and Bollen K. (2005) Bayesian struc- tural equation modeling. SAMSI TR2005-5. http:\\www.samsi.info\TR\tr2005-05.pdf

[5] Fox, J.P., and Glas, C. A.W. (2001). Bayesian estimation of a multilevel IRT model using Gibbs sampling. Psychometrika, 66, 271-288.

[6] Gelman A. (2006) Prior distributions for variance parameters in hierarchical models. Bayesian Analysis 1, 515-533.

[7] Gelman A., Jakulin A., Pittau M. G., Su Y. S. (2008a). A weakly informative default prior distribution for logistic and other regression models. The Annals of Applied Statistics, 2, 1360-1383.

[8] Gelman A., van Dyk, D. A., Huang Z., Boscardin J. (2008b) Using Redundant Parameterizations to Fit Hierarchical Models. Journal Of Computational And Graphical Statistics, 17, 95-122.

[9] Hjort, N. L., Dahl, F. A. and Steinbakk, G. H. (2006). Post-processing posterior predictive p-values. J. Amer. Statist. Assoc. 101, 1157-1174.

[10] Muthén, B. (2010) Bayesian Analysis In Mplus: A Brief Introduction. Mplus Technical Report. http:\\www.statmodel.com

[11] Muthén, B., Asparouhov, T., Hunter, A. & Leuchter, A. (2010). Growth modeling with non-ignorable dropout: Alternative analyses of the STAR*D antidepressant trial. Submitted for publication.

[12] Nylund, K. L., Asparouhov, T., & Muthn, B. O. (2007). Deciding on the number of classes in latent class analysis and growth mixture modeling: A Monte Carlo simulation. Structural Equation Modeling, 14, 535-569.

[13] Lee S. Y., Song X. Y., Cai J. H. (2010) A Bayesian Approach for Nonlinear Structural Equation Models With Dichotomous Variables Using Logit and Probit Links. Structural Equation Modeling, 17, 280-302.

[14] Song X.Y., Xia Y. M., Lee S. Y. (2009) Bayesian semiparametric analysis of structural equation models with mixed continuous and unordered categorical variables. Statistics in Medicine, 28, 2253-2276.

[15] Olsen, M.K. & Schafer, J.L. (2001). A two-part random effects model for semicontinuous longitudinal data. Journal of the American Statistical Association, 96, 730-745.

[16] Patz R. and Junker B. (1999) A Straightforward Approach to Markov Chain Monte Carlo Methods for Item Response Models Journal of Educational and Behavioral Statistics Summer, 24, 146-178.

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

[18] Raghunathan, T. E., Lepkowski, J. M., Van Hoewyk, J., and Solenberger, P. (2001) A Multivariate Technique for Multiply Imputing Missing Values Using a Sequence of Regression Models. Survey Methodology, Vol. 27, No 1. 85-95.

[19] Savalei, V. (2010) Small Sample Statistics for Incomplete Nonnormal Data: Extensions of Complete Data Formulae and a Monte Carlo Comparison. Structural Equation Modeling, 17, 241-264.

[20] Segawa E., Emery S., and Curry S. (2008) Extended Generalized Linear Latent and Mixed Model, Journal of Educational and Behavioral Statistics, 33, 464-484.

[21] van Dyk, D. A., and Meng, X. L. (2001), The Art of Data Augmentation. Journal of Computational and Graphical Statistics, 10, 1-111.