# Bayesian Analysis Using Mplus: Technical Implementation

*Tihomir Asparouhov and Bengt Muthén*

Version 3

September 29, 2010

# 1 Introduction

In this note we describe the implementation details for estimating latent variable models with the Bayesian estimator in Mplus. The algorithm used in Mplus is Markov Chain Monte Carlo (MCMC) based on the Gibbs sampler, see Gelman et al. (2004).

The Gibbs sampler generates iteratively a sequence of parameters, latent variables, and missing observations, which upon convergence can be used to construct the posterior distribution given the observed data and prior specifications for the parameters. The Gibbs sampler blocks the parameters, latent variables, and missing observations, into groups that are updated sequentially. Denoted by $\theta_t$ a vector of all model parameters, latent variables and missing observations generated during iteration $t$. The Gibbs sampler is based on splitting $\theta_t$ into $d$ groups

$$\theta_t = (\theta_{1t}, ..., \theta_{dt}).$$

and generating the the components of this vector in the following sequence from the following conditional distributions

$$[\theta_{1t}|\theta_{2t-1}, ..., \theta_{d\ t-1}, data, priors]$$

$$[\theta_{2t}|\theta_{1t}, \theta_{3\ t-1}..., \theta_{d\ t-1}, data, priors]$$

$$...$$

$$[\theta_{dt}|\theta_{1t}, ..., \theta_{d-1\ t-1}, data, priors].$$

Upon convergence Mplus uses a segment of the generated sequence (usually the end of the sequence) $\theta_n$,...,$\theta_m$ to construct the posterior distributions of the parameters, latent variables and missing observations, given the observed data and priors, i.e, the draws $\theta_n$,...,$\theta_m$ can be assumed to be independent draws from the posterior distributions. Note that generally they are not independent and $\theta_t$ and $\theta_{t+1}$ are often highly correlated, however when enough iterations have been generated the posterior distribution formed by these generated observations will be the same as the true posterior distribution.

The success of this estimation process depends on correctly diagnosing convergence. Convergence depends very heavily on correct split of the $\theta_t$ vector. Mplus will attempt to choose the most optimal such split for a particular model and perform different updating algorithms in different situations. Generally speaking highly correlated elements in the $\theta_t$ vector have to be in the same updating group, if this is not the case then the sequence will essentially not converge. For example if $\theta_{1t}$ and $\theta_{2t}$ are perfectly correlated, if they are in different updating groups then they will never change in the MCMC sequence and the process will essentially not converge.

Convergence in the MCMC sequence is unfortunately not always easy to diagnose. Mplus provides trace and autocorrelation plots for the parameter estimates as well as the Potential Scale Reduction (PSR) convergence criteria, which compares several independent MCMC sequences. In some applications careful and to a large extent subjective convergence analysis has to be done after the Mplus estimation process to ensure that convergence has really occurred. Models can also be parameterized in various different ways. Certain parameterizations will be better than others, i.e., choosing the optimal parameterizations will affect the rate of convergence as well as the quality of the final results.

# 2 Structural Equation Models with Continuous Variables

## 2.1 Model

Let $y$ be a vector of $p$ observed dependent variables, $\eta$ be a vector of $m$ latent variables, and $x$ be a vector of $q$ independent observed variables. The structural equation model we consider is

$$y = \nu + \Lambda\eta + Kx + \varepsilon \tag{1}$$

$$\eta = \alpha + B\eta + \Gamma x + \zeta \tag{2}$$

where $\varepsilon$ and $\zeta$ are normally distributed zero mean residuals with variance covariance matrix $\Theta$ and $\Psi$. The matrices $\Theta$ and $\Psi$ are assumed to be block diagonal. For example

$$\Psi = \begin{pmatrix} \Psi_{11} & 0 & \cdots & 0 \\ 0 & \Psi_{22} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \Psi_{dd} \end{pmatrix}$$

where each of the matrices $\Psi_{jj}$ is a full variance covariance matrix

$$\Psi_{jj} = \begin{pmatrix} \psi_{j11} & \psi_{j12} & \cdots & \psi_{j1k} \\ \psi_{j21} & \psi_{j22} & \cdots & \psi_{j2k} \\ \vdots & \vdots & \ddots & \vdots \\ \psi_{jk1} & \psi_{jk2} & \cdots & \psi_{jkk} \end{pmatrix}$$

where $k$ is the size of $\Psi_{jj}$ and $k$ is different across the blocks. The same definition applies to $\Theta$.

Models with non-block diagonal $\Theta$ and $\Psi$ can also be included in this framework. Such models can be estimated in Mplus using the ALGORITHM option of the ANALYSIS command. The ALGORITHM option has to be set to GIBBS(RW), which means that a random walk algorithm will be used in the estimation of $\Theta$ and $\Psi$. This Metropolis-Hastings based algorithm can estimate an arbitrary structured variance covariance matrix. However the algorithm is somewhat less efficient in terms of mixing quality than the Mplus default algorithm based on conjugate priors. Details of the random walk algorithm can be found in Chib and Greenberg (1998).

## 2.2 Model Extension

Equation (2) in the above model is more flexible than equation (1) because the dependent variable in (2) can also be a predictor variables. It is possible however to obtain that additional flexibility for the first equation by adding to the model artificial latent variables. For example, suppose that we need to regress $Y_1$ on $Y_2$. Since (1) does not provide this flexibility we can use the following two latent variable setup to incorporate that into the above model. Suppose $\eta_1$ and $\eta_2$ are latent variables and $Y_1 = \eta_1$, $Y_2 = \eta_2$, $\eta_1 = \alpha + \beta\eta_2 + .....$ The first two equations can be part of (1) and the last one is part of (2). In total the above three equations are equivalent to $Y_1 = \alpha + \beta Y_2 + ....$ Mplus will implement the above 3 equation setup automatically.

## 2.3 Priors

Let $\gamma$ be the vector of all of the free parameters among $\nu$, $\alpha$, $\Lambda$, $B$, $\Gamma$, $K$. We assume a normally distributed prior for $\gamma$

$$\gamma \sim N(\gamma_0, \Omega_\gamma).$$

For all blocks in the variances covariances matrices $\Psi$ and $\Theta$ we assume Inverse Wishart prior, for example for a block $\Psi_{jj}$ we assume

$$\Psi_{jj} \sim IW(\Omega_{\Psi_{jj}}, d_{\Psi_{jj}}).$$

There is one exception to this rule. If the size of the block is 1 then we assume an inverse gamma prior

$$\Psi_{jj} \sim IG(\alpha_{\Psi_{jj}}, \beta_{\Psi_{jj}}).$$

All of the above specified priors are conjugate priors, i.e., the conditional distributions in the Gibbs sampler are in the same family of distributions as the prior distribution.

## 2.4 Estimation

The parameters and the latent variables are split in 3 groups: the slope intercept and loading parameters $\gamma$, the variance covariance parameters $\Psi$ and $\Theta$, and the latent variables $\eta$. Thus the Gibbs sampler has 3 steps

- Step 1. Update $\eta$
$$[\eta|\Psi, \Theta, \gamma, Y, priors]$$
- Step 2. Update $\gamma$
$$[\gamma|\Psi, \Theta, \eta, Y, priors]$$
- Step 3. Update $\Psi, \Theta$
$$[\Psi, \Theta|\gamma, \eta, Y, priors]$$

Below we describe the computation of these conditional distributions. All 3 conditional distributions are easy to sample from. These conditional distributions have been derived for example in Lee (2007), Muthen and Arminger (1995), Arminger and Muthen (1998) among others and can be considered well known. We provide these here for completeness.

*Step 1.*

In this step we obtain the conditional distribution of $\eta$ given everything else. Denote this conditional distribution by $[\eta|*]$. Let $B_0 = I - B$ where $I$ is the identity matrix. We can rewrite equation (2) as

$$\eta = B_0^{-1}(\alpha + \Gamma x) + \zeta_0$$

where $\zeta_0 = B_0^{-1}\zeta$ has a variance covariance matrix $\Psi_0 = (B_0^{-1})\Psi(B_0^{-1})^T$. The conditional distribution is then given by

$$[\eta|*] \sim N(Dd, D)$$

where

$$D = \left(\Lambda^T\Theta^{-1}\Lambda + \Psi_0^{-1}\right)^{-1}$$

$$d = \Lambda^T\Theta^{-1}(y - \nu - Kx) + \Psi_0^{-1}B_0^{-1}(\alpha + \Gamma x)$$

*Step 2.*

To obtain the conditional distribution of $\gamma$ given everything else we rewrite (1) and (2) in the following equivalent form

$$z = F\gamma + \epsilon$$

where $\epsilon = (\varepsilon, \zeta)$, $z = (y, \eta)$, and $F$ is a matrix of dimensions $p + m$ by $(p + m)(1 + m + q)$

$$F = I \otimes (1, \eta, x)$$

and $I$ is the identity matrix of size $p + m$. Let $V$ be the variance covariance of $\epsilon$

$$V = \begin{pmatrix} \Theta & 0 \\ 0 & \Psi \end{pmatrix}$$

Let's assume that there are $n$ observations in the data. We will index by $i$ the variables from the $i$-th observation. The conditional distribution of $\gamma$ is given by

$$[\gamma|*] \sim N(Dd, D) \tag{3}$$

where

$$D = \left(\sum_{i=1}^{n} F_i V^{-1} F_i + \Omega_\gamma^{-1}\right)^{-1} \tag{4}$$

$$d = \sum_{i=1}^{n} F_i V^{-1} z_i + \Omega_\gamma^{-1}\gamma_0 \tag{5}$$

*Step 3.*

In this step we need to determine the conditional distribution of the variance covariance matrices given the observed dependent variables $Y$ and the latent variables $\eta$, i.e., in this step of the computation $\eta$ is also observed and is essentially not different from the observed variable $Y$. Therefore we will describe the conditional distribution only for $\Theta$. We consider the two cases separately, blocks of size 1 and blocks of size greater than 1.

Suppose that $\theta_{11}$ is a block of size 1, with prior distribution

$$\theta_{11} \sim IG(\alpha_0, \beta_0).$$

The conditional distribution of $\theta_{11}$ is

$$[\theta_{11}|*] \sim IG(\alpha_0 + \frac{n}{2}, \beta_0 + \beta_1) \tag{6}$$

where

$$\beta_1 = \frac{1}{2}\sum_{i=1}^{n}\left(y_{i1} - \nu_1 - \sum_{j=1}^{m}\lambda_{1j}\eta_{ij} - \sum_{j=1}^{q}K_{1j}x_{ij}\right)^2.$$

Now suppose that $\Theta_{11}$ is a block of size $d$

$$\Theta_{11} = \begin{pmatrix} \theta_{11} & \theta_{12} & \cdots & \theta_{1d} \\ \theta_{21} & \theta_{22} & \cdots & \theta_{2d} \\ \vdots & \vdots & \ddots & \vdots \\ \theta_{d1} & \theta_{d2} & \cdots & \theta_{dd} \end{pmatrix}$$

with prior distribution

$$\Theta_{11} \sim IW(\Omega, f). \tag{7}$$

The conditional distribution of $\Theta_{11}$ is

$$[\Theta_{11}|*] \sim IW(E + \Omega, n + f) \tag{8}$$

where $E$ is the first diagonal block of size $d$ in the matrix

$$\sum_{i=1}^{n}(y_i - \nu - \Lambda\eta_i - Kx_i)(y_i - \nu - \Lambda\eta_i - Kx_i)^T \tag{9}$$

## 2.5 Convergence

The main criterion used in Mplus for determining convergence of the MCMC sequence is based on the potential scale reduction (PSR). Mplus will run several different MCMC chains (2 by default). From each chain the first half of the iterations is considered preliminary. Only the second half is used for forming the posterior distribution and for evaluating convergence. There is one exception to this rule. If Mplus is run with only one chain the first half of the iterations are removed but the second half itself is split in half and those two halves are treated as if they were two different chains for the purpose of computing PSR. Suppose that there are $m$ chains and $n$ iterations (after the preliminary iterations are removed). Let $\theta$ be a parameter in the model and denote by $\theta_{ij}$ the value of $\theta$ in iteration $i$ in chain $j$. The PSR for this parameter is computed as follows.

$$\bar{\theta}_{.j} = \frac{1}{n}\sum_{i=1}^{n}\theta_{ij}$$

$$\bar{\theta}_{..} = \frac{1}{m}\sum_{j=1}^{m}\bar{\theta}_{.j}$$

$$B = \frac{1}{m-1}\sum_{j=1}^{m}(\bar{\theta}_{.j} - \bar{\theta}_{..})^2$$

$$W = \frac{1}{m}\sum_{j=1}^{m}\frac{1}{n}\sum_{i=1}^{n}(\theta_{ij} - \bar{\theta}_{.j})^2$$

$$PSR = \sqrt{\frac{W+B}{W}}$$

If PSR is less than $1 + \epsilon$ for all the parameters in the model Mplus will conclude that convergence has occurred. The convergence criterion is checked every 100-th iteration. Here $\epsilon = fc$ where $c$ is controlled by the user with the *bconvergence* command in Mplus. The factor $f$ is a multiplicity factor that makes the convergence criteria more lenient when there are more parameters in the model. For most models $1 + \epsilon$ is between 1.05 and 1.1, using the default value of $c = 0.05$.

Essentially the PSR convergence criteria is equivalent to monitoring the ICC for the parameters (the chains being the clusters) and concluding convergence if ICC is below a certain value. For example if the convergence criterion is PSR$< 1.05$, that is equivalent to ICC$< 0.09$ because

$$ICC = \frac{B}{W+B} = 1 - \frac{1}{PSR^2} < 1 - \frac{1}{1.05^2} \approx 0.09.$$

Small ICC values imply that the chains are so similar that given a particular parameter value it is not possible to determine which chain it comes from.

# 3 Structural Equation Models with Continuous and Categorical Variables

## 3.1 Model

For each categorical variable $Y_j$ in the model, taking the values from 1 to $k$, we assume that there is a latent variable $Y_j^*$ and threshold parameters $\tau_{1j}, ..., \tau_{k-1j}$ such that

$$Y_j = t \Leftrightarrow \tau_{t-1j} \leq Y_j^* < \tau_{tj} \tag{10}$$

where we assume $\tau_{0j} = -\infty$ and $\tau_{kj} = \infty$. The above definition essentially converts a categorical variable $Y_j$ into an unobserved continuous variable $Y_j^*$. The rest of the model is then defined as in (1) and (2) where for each categorical variable we use $Y_j^*$ instead of $Y_j$. For identification purposes the intercept of $Y_j^*$ is assumed to be $\nu_j = 0$ and the residual variance $\theta_{jj} = 1$.

## 3.2 Priors

The new parameters in the model are the thresholds parameters $\tau_{tj}$. The prior for these parameters can be of any kind, i.e., there are no known conjugate prior distributions. There is one exception to this rule. When a binary variable $Y_j$ is used in the model there is only one threshold parameter $\tau_{1j}$. In certain models an alternative parameterization is used because it allows conjugate priors. The threshold parameter $\tau_{1j}$ and the intercept parameter $\nu_j$ are essentially perfectly correlated parameters with correlation -1. Instead of using a parameterization where $\tau_{1j} = a$ and $\nu_j = 0$, we can use a parameterization where $\tau_{1j} = 0$ and $\nu_j = -a$. This way there is no threshold parameter to be estimated and the intercept parameter $\nu_j$ can be estimated with the conjugate normal prior. The above parameterization is however not possible when the variable $Y_j^*$ is not only a dependent variable but also a predictor variable in the general model.

In addition to the new $\tau_{tj}$ parameters we now have a new type of variance covariance matrices in the model. A variance covariance block can now include categorical variables for which the residual variance is fixed to 1, i.e., a $\Theta$ block can be an unrestricted correlation matrix

$$\begin{pmatrix} 1 & \theta_{12} & \cdots & \theta_{1d} \\ \theta_{21} & 1 & \cdots & \theta_{2d} \\ \vdots & \vdots & \ddots & \vdots \\ \theta_{d1} & \theta_{d2} & \cdots & 1 \end{pmatrix}$$

or a matrix that is partially a correlation matrix and partially a covariance matrix with some elements on the diagonal fixed to 1 and some free parameters

$$\Theta = \begin{pmatrix} 1 & \cdots & \theta_{1d} & \theta_{1d+1} & \cdots & \theta_{1k} \\ \vdots & \ddots & \vdots & \vdots & \vdots & \vdots \\ \theta_{d1} & \cdots & 1 & \theta_{dd+1} & \cdots & \theta_{dk} \\ \theta_{d+1\ 1} & \cdots & \theta_{d+1\ d} & \theta_{d+1\ d+1} & \cdots & \theta_{d+1\ k} \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ \theta_{k1} & \cdots & \theta_{kd} & \theta_{k\ d+1} & \cdots & \theta_{kk} \end{pmatrix}. \quad (11)$$

We will call the above matrices *partial correlation matrices.* The parameters in a partial correlation matrix are essentially new parameters as well. We need to define priors for such partial correlation matrices and derive their conditional distribution to be used in the Gibbs sampler. Consider the parameter matrix expanded to include the non-identified variance parameters $v_1$,...,$v_d$

$$\Theta^* = \begin{pmatrix} v_1 & \cdots & \theta_{1d}\sqrt{v_1 v_d} & \theta_{1d+1}\sqrt{v_1} & \cdots & \theta_{1k}\sqrt{v_1} \\ \vdots & \ddots & \vdots & \vdots & \vdots & \vdots \\ \theta_{d1}\sqrt{v_1 v_d} & \cdots & v_d & \theta_{dd+1}\sqrt{v_d} & \cdots & \theta_{dk}\sqrt{v_d} \\ \theta_{d+1\ 1}\sqrt{v_1} & \cdots & \theta_{d+1\ d}\sqrt{v_d} & \theta_{d+1\ d+1} & \cdots & \theta_{d+1\ k} \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ \theta_{k1}\sqrt{v_1} & \cdots & \theta_{kd}\sqrt{v_d} & \theta_{k\ d+1} & \cdots & \theta_{kk} \end{pmatrix}. \quad (12)$$

To obtain a prior for $\Theta$ we specify an Inverse Wishart prior $IW(\Omega, f)$ for the expanded matrix $\Theta^* = (\Theta, v_1, v_2, ..., v_d)$ and we use the marginal distribution of $\Theta$ as the prior specification for $\Theta$. Mplus will not allow a complete specification of the matrix $\Omega$, it will allow specification for all of the entries except the diagonal entries corresponding to the fixed entries of 1 on the main diagonal. This actually does not limit in any way the shapes of the marginal distributions of $\Theta$ that can be obtained. In fact it only affects the distribution of the expanded parameters $v_1, ..., v_d$. Internally Mplus will complete the $\Omega$ matrix with the following entries corresponding to the 1's on the main diagonal: $f+k+1$. This has to be taken into account when off diagonal elements are given for the prior matrix $\Omega$.

The above approach of using parameters that are full or partial correlation matrices is used in Mplus not just for categorical variables but also for other variables, for example, in some cases for identification purposes the variance covariance matrix for latent factors is specified as correlation matrices.

## 3.3 Estimation

The estimation of the model for the combination of categorical and continuous variables described above builds naturally on the estimation of the model with all continuous variables. Basically we need to derive Step 4 in the Gibbs sampler which generates the underlying continuous variables $[Y_j^*|*]$ for all categorical variables $Y_j$. Steps 1-3 will then remain the same with the exception that $Y_j^*$ will be used in the place of $Y_j$. Note however that Step 5 has to be derived in which we generate the thresholds parameters. In addition, Step 3 needs to be expanded to handle the cases where the variance covariance matrices include partial correlation blocks.

In the Mplus implementation however Steps 4 and 5 are intertwined and are done in 3 different ways depending on the model. This is done so that the most optimal and efficient algorithm is used for each model. But even before we start considering which algorithm is used in which case we have to again mention the fact that for binary variables, in most cases, the thresholds are simply substituted by means for the underlying continuous variables, see Section 3.2, i.e., the discussion below will not apply for most models with binary variables. More precisely, for each binary variable that is not a predictor of any other variable the threshold parameter is substituted with a mean parameter for the underlying continuous variable. In that case there is no need to generate the threshold. However, $Y^*$ still has to be generated.

The three algorithms that Mplus uses can be described as follows.

*Method 1.* $\tau$ and $Y^*$ are one block of parameters and are generated together in two steps using the following decomposition

$$[\tau, Y^*|*] \sim [\tau|*][Y^*|\tau, *]$$

*Method 2.* If there are $d$ categorical variables, say $Y_1$,...,$Y_d$ and $\tau$ and $Y^*$ are separated into $d$ groups $\{Y_j^*, \tau_{.j}\}$, for $j=1,...,d$, and essentially the

Gibbs sampler gets $d$ new separate steps

$$[\tau_{.1}, Y_1^*|*, \tau_{.j}, Y_j^*, j \neq 1]$$

$$[\tau_{.2}, Y_2^*|*, \tau_{.j}, Y_j^*, j \neq 2]$$

$$.$$

$$[\tau_{.d}, Y_d^*|*, \tau_{.j}, Y_j^*, j \neq d]$$

Each of the above steps is again separated into two substeps just like in Method 1. For example for the $i$-th categorical variable we have

$$[\tau_{.i}, Y_i^*|*, \tau_{.j}, Y_j^*, j \neq i] \sim$$

$$[\tau_{.i}|*, \tau_{.j}, Y_j^*, j \neq i][Y_i^*|\tau_{.i}, *, \tau_{.j}, Y_j^*, j \neq i]$$

*Method 3.* If there are $d$ categorical variables, say $Y_1$,...,$Y_d$ and $\tau$ and $Y^*$ are separated into $d+1$ groups $\{Y_j^*\}$, for $j = 1, ..., d$ form $d$ groups and all threshold parameters form another group. Gibbs sampler gets $d+1$ new steps

$$[Y_1^*|*, \tau, Y_j^*, j \neq 1]$$

$$[Y_2^*|*, \tau, Y_j^*, j \neq 2]$$

$$.$$

$$[Y_d^*|*, \tau, Y_j^*, j \neq d]$$

as well as one additional step that generates all thresholds

$$[\tau|*, Y^*]$$

Mplus will automatically determine which method is best for each model and will apply that method in the estimation. It is not possible for the user to select the estimation method. Generally Method 1 is the most efficient, Method 2 is less efficient than Method 1 but more efficient than Method 3 and Method 3 is the least efficient. The reason we need however the additional methods is because Method 1 is not available for all models, nor is Method 2. Method 3 is available for all models.

To determine which method can be used we need to consider the conditional distribution of the vector $Y^* = (Y_1^*, ..., Y_d^*)$ given $Y_o = (Y_{d+1}, ..., Y_p)$, $X$

and $\eta$ as well as all model parameters. It is easy to show that this conditional distribution is normal

$$Y^* \sim N(\mu + \beta_1 Y_o + \beta_2 X + \beta_3 \eta, V) \tag{13}$$

where the values of $\mu$, $\beta_1$, $\beta_2$, $\beta_3$ and $V$ can easily be obtained with matrix algebra from the model parameters. If $V$ is a diagonal matrix then Method 1 can be applied. If $V$ is not a diagonal matrix then Method 2 is applied with the exception to the case when the model also contains parameter equalities between the $\tau$ groups, which basically implies that Method 2 can not be defined since $\tau$ groups with model constraints can not be separated into different blocks. Therefore Method 3 is applied only in the case when there are threshold equalities as well as non-zero off diagonal elements in $V$.

Now we describe in detail the three methods. Method 1 and Method 2 can be considered multivariate extensions of the algorithm proposed by Cowles (1996) for estimating univariate probit regression model, see also Johnson and Albert (1999). This method amounts to a simple Metropolis-Hastings sampling for the thresholds using a normal distribution as the proposal (jumping) distribution. For completeness we will describe this algorithm. Method 3 is a multivariate extension of the Albert and Chib (1993) method.

### 3.3.1 Method 1

Let's first focus on Method 1. Notice that when $V$ is diagonal (and the prior distributions for the $\tau$ groups are independent), then the $\tau$ groups are conditionally independent, that is

$$[\tau|*] \sim [\tau_{.1}|*][\tau_{.2}|*]...[\tau_{.d}|*]$$

therefore we can simply focus on a single group $[\tau_{.j}|*]$, i.e., the threshold set for the $j-$th categorical variable. We use Cowles (1996) approach to sample from this conditional distribution. This consists of two steps

Step 1. Generate sequentially a proposed new threshold set $g = (g_1, ..., g_k)$ where $k$ is the number of thresholds for the $j-$th categorical variable. For $t = 1, ..., k$ generate $g_t$ from the normal distribution $N(\tau_{tj}, \sigma^2)$ truncated to the interval $(g_{t-1}, \tau_{t+1\ j})$, assuming as usual the notation that $g_0 = -\infty$ and $\tau_{k+1\ j} = \infty$. The variance of the proposal/jumping distribution $\sigma^2$ is chosen to be a small value such as 0.1 and is adjusted during a preliminary stage of the estimation to obtain optimal mixing, i.e., optimal acceptance rate in the

Metropolis-Hastings algorithm. The optimal acceptance rate is considered to be between .25 and 0.50.

Step2. Compute the acceptance ratio $R$

$$R = \prod_{i=1}^{n} \frac{\Phi(v_{jj}^{-0.5}(g_{y_{ij}} - m_{ij})) - \Phi(v_{jj}^{-0.5}(g_{y_{ij}-1} - m_{ij}))}{\Phi(v_{jj}^{-0.5}(\tau_{y_{ij}j} - m_{ij})) - \Phi(v_{jj}^{-0.5}(\tau_{y_{ij}-1\ j} - m_{ij}))}$$

$$\times \prod_{i=1}^{k} \frac{\Phi(\sigma^{-1}(\tau_{i+1\ j} - \tau_{i\ j})) - \Phi(\sigma^{-1}(g_{i-1} - \tau_{i\ j}))}{\Phi(\sigma^{-1}(g_{i+1} - g_i)) - \Phi(\sigma^{-1}(\tau_{i-1\ j} - g_i))}$$

$$\times \frac{P(g)}{P(\tau_{.j})}$$

where $P()$ is the prior distribution for $\tau_{.j}$, $m_{ij}$ is the conditional mean given in (13) and $v_{jj}$ is the diagonal entry in the matrix $V$ from the conditional distribution in (13). The Metropolis-Hastings algorithm accepts the proposed new threshold set $g$ with probability $min(1, R)$. If the new thresholds set is rejected the old threshold set is retained. Thus using the Metropolis-Hastings algorithm we can sample from $[\tau_{.j}|*]$ without having an explicit derivation for this distribution.

Method 1 is completed by specifying the conditional distribution $[Y^*|\tau, *]$. Under the assumption of diagonal $V$, the $Y_j^*$ are independent

$$[Y^*|\tau, *] \sim [Y_1^*|\tau, *]...[Y_d^*|\tau, *]$$

and we can simply describe the univariate conditional distribution $[Y_j^*|\tau, *]$. The conditional distribution of $Y_{ij}^*$ is (13) truncated to the interval $(\tau_{y_{ij}j}, \tau_{y_{ij}+1\ j})$.

### 3.3.2 Method 2

Here we need to describe the following two conditional distributions

$$[\tau_{.i}|*, \tau_{.j}, Y_j^*, j \neq i]$$

and

$$[Y_i^*|\tau_{.i}, *, \tau_{.j}, Y_j^*, j \neq i].$$

First consider the conditional distribution of $Y_i^*$ conditional on $X$, $\eta$, $Y_o$ and $Y_j^*$ for all $j \neq i$. Let's denote the group of variables $Y_j^*$ for all $j \neq i$ by $Y_{-i}^*$. This conditional distribution is normal

$$Y_i^* \sim N(\mu_i + \beta_{1i} Y_o + \beta_{2i} X + \beta_{3i} \eta + \beta_{4i} Y_{-i}^*, v_i) \tag{14}$$

where all the coefficients in the above distribution can be obtained algebraically from the coefficients in (13). To generate the distribution

$$[\tau_{.i}|*, \tau_{.j}, Y_j^*, j \neq i]$$

we can use the Cowles (1996) algorithm just as we did in Method 1 but using the conditional distribution (14) instead of the conditional distribution (13). The conditional distribution

$$[Y_i^*|\tau_{.i}, *, \tau_{.j}, Y_j^*, j \neq i].$$

can also be described just as it was described for Method 1 but using the conditional distribution (14) instead of the conditional distribution (13).

### 3.3.3 Method 3

The conditional distribution

$$[Y_j^*|*, \tau, Y_{-j}^*]$$

here is given the same way as it is for Method 2. Therefore we only need to describe the conditional distribution

$$[\tau|*, Y^*].$$

Assuming that the $\tau$ blocks have independent priors the conditional distributions are also independent

$$[\tau|*, Y^*] \sim [\tau_{.1}|*, Y^*]...[\tau_{.d}|*, Y^*]$$

and therefore we can simply describe $[\tau_{.j}|*, Y^*]$, or more precisely how to sample from this distribution. We use the Albert and Chib (1993) algorithm. The conditional distribution for $[\tau_{tj}|*, Y^*, \tau_{-tj}]$ is simply the prior distribution of $\tau_{tj}$ truncated to the interval

$$(max(\tau_{t-1\ j}, max(Y_{ij}^* : Y_{ij} = t)), min(\tau_{t+1\ j}, min(Y_{ij}^* : Y_{ij} = t + 1))).$$

### 3.3.4 Generating partial correlation matrices

In this section we discuss the algorithms implemented in Mplus for generating the partial correlation blocks in the variance covariance matrices. There are four different algorithms implemented in Mplus.

The first algorithm is a simple parameter extended estimation where the variances for $Y^*$ are estimated as free parameters. These parameters are formally unidentified if estimated with classical frequentist estimators. However within the Bayesian estimation framework they are identified as long as their prior distribution is a proper prior. The resulting posterior distribution is the same as the prior distribution for these parameters. During the estimation a posterior distribution is built for the full variance covariance block. Subsequently that posterior distribution is used to obtain the posterior distribution for the actual parameters of interest, i.e., the identified correlation parameters. This algorithm is the default algorithm in Mplus and it can also be specified using the ALGORITHM option of the ANALYSIS command, using the setting ALGORITHM=GIBBS(PX1).

The second algorithm implemented in Mplus for the estimation of correlation matrices is the algorithm described in Boscardin et al. (2008). This algorithm can be specified using the command ALGORITHM=GIBBS(PX2).

The third algorithm implemented in Mplus for the estimation of correlation matrices is the algorithm described in Liu and Daniels (2006). This algorithm can be specified using the command ALGORITHM=GIBBS(PX3).

The forth algorithm implemented in Mplus for the estimation of correlation matrices is the algorithm described in Chib and Greenberg (1998). This algorithm can be specified using the command ALGORITHM=GIBBS(RW).

Extensive simulations studies, not reported here, have shown that the simple parameter extended algorithm used in Mplus as a default method is the best algorithm in terms of optimal mixing and fast convergence.

# 4 Estimating Structural Equation Models with Missing Data

If a categorical variable is missing the change that is needed in the above algorithm is simply that during the generation of $Y^*$ the normal distribution that produces $Y^*$ is not truncated to the corresponding threshold interval, instead it is not truncated at all, or simply put it is truncated to the interval

$(-\infty, \infty)$. If a continuous variable, say for example $Y_j$, has missing values, then we include a new step in the Gibbs sampler where the missing values of $Y_j$ are generated from the univariate $[Y_j|*]$ which is a univariate conditionally normal distribution

$$N(\mu_j + \beta_{j1}\eta + \beta_{j2}X + \beta_{j3}Y_{-j}, v_j) \tag{15}$$

where the parameters $m_j$, $\beta_{j1}$, $\beta_{j2}$, $\beta_{j3}$, and $v_j$ can be obtained from the model parameters using matrix algebra. Note that if multiple continuous variables have missing values they are updated one at a time. For example, if each of the continuous variable $Y_1, ..., Y_k$ have missing values then the Gibbs sampler gets $k$ new steps

$$[Y_1|*, Y_{-1}]$$
$$[Y_2|*, Y_{-2}]$$
$$\dots$$
$$[Y_k|*, Y_{-k}]$$

All of the above Gibbs steps are done using univariate distributions as described in (15). Just like the maximum-likelihood estimation method this algorithm will correctly estimate the model under the assumption that the missing data is missing at random (MAR), see Little and Rubin (1987).

# 5 Fixed Parameters and Equalities Between the Parameters

Fixed parameters in the model are simply not updated in the Gibbs sampler, i.e., fixed parameters do not present any estimation challenge. However, ceratin parameters in the current Mplus implementation are updated as blocks of parameters, for example all the threshold parameters for a particular categorical variable or all the parameters in a variance covariance block. Therefore it is not possible with the current implementation in Mplus to fix only one parameter in such a block, either all parameters have to be free or all the parameters have to be fixed within each of these blocks. One exception to this rule is the variance parameters in a variance covariance block which can be fixed to 1 and thereby converting the variance covariance matrix to a partial correlation matrix.

Equalities between parameters are slightly more challenging than fixed parameters. First note that the above model has three different types of parameters: thresholds, variance/covariance parameters and the third group is the parameters $\gamma$ which consists of all loadings, intercepts and slopes. Equalities between parameters from different groups are not possible. For example a threshold parameter can not be held equal to a variance parameter. In addition groups such as variance covariance blocks and threshold groups (the thresholds of a single categorical variable) must simultaneously be held equal to another group. Partial equality constraints is not possible.

For thresholds and variance covariance parameters equalities simply amounts to combining the "data" part of the conditional distribution, i.e., if two parameters are held equal then the data driven portion of the conditional likelihood will essentially change from $n$ observations to $2n$ observations. The first $n$ will be those that are used for the conditional distribution of the first parameters and the second $n$ are those that are used for the conditional distribution of the second likelihood. For example, let $\theta_{11}$ and $\theta_{22}$ be the variances of $Y_1$ and $Y_2$. If the two parameters are estimated as unequal then their conditional distributions are obtained as follows. If the prior distribution of $\theta_{jj}$ is

$$\theta_{jj} \sim IG(\alpha_{0j}, \beta_{0j})$$

then the conditional distribution of $\theta_{jj}$ is

$$[\theta_{jj}|*] \sim IG(\alpha_{0j} + n/2, \beta_{0j} + \beta_{1j})$$

where

$$\beta_{1j} = \frac{1}{2}\sum_{i=1}^{n}\left(y_{ij} - \nu_j - \sum_{l=1}^{m}\lambda_{1l}\eta_{il} - \sum_{l=1}^{q}K_{1l}x_{il}\right)^2.$$

If the parameters are held equal the conditional distribution of $\theta_{11}$ is

$$[\theta_{11}|*] \sim IG(\alpha_{01} + n/2 + n/2, \beta_{01} + \beta_{11} + \beta_{12}).$$

For the $\gamma$ parameters the idea is slightly different. The conditional distribution of $\gamma$ is given by

$$[\gamma|*] \sim N(Dd, D) \tag{16}$$

where $D$ and $d$ are given in (4) and (5). Both $d$ and $D^{-1}$ are sums of a term due to the prior and a term due to the data. Let's ignore for a second the term due to the prior (this term will not change under parameter equalities). Assuming that there is no prior, $d$ is simply the first derivative of the log-likelihood evaluated at 0 and $D^{-1}$ is the second derivative of the log-likelihood. Now using simple derivatives rule we obtain the method for deriving the $d$ and $D^{-1}$ under equality constraints. If parameter $j_1$ and $j_2$ are held equal then in the vector $d$ the values $d_{j_1}$ and $d_{j_2}$ are combined into one and also in the matrix $D^{-1}$ the four values with coordinates $(j_1, j_1)$, $(j_1, j_2)$, $(j_2, j_1)$ and $(j_2, j_2)$ are combined into one.

# 6 Non-conjugate Priors

Non-conjugate priors are implemented in Mplus for the threshold parameters and also for the variance covariance parameters. The derivations given above for the threshold parameters were done using arbitrary priors. In fact all priors for the thresholds are non-conjugate priors. Therefore in this section we focus only on the estimation of the variance covariance parameters with non-conjugate priors. It is well known that variance parameters priors for the variances of random effects and factors have relatively high influence on the posterior estimates. Therefore it is important in practice to be able to specify various different priors for these parameters.

Two separate algorithms are implemented in Mplus for generating variance covariance parameters with non-conjugate priors. The first algorithm is used for variance parameters only, i.e., for variance covariance blocks of size 1. A separate algorithm is used for variance covariance blocks of size bigger than 1. First we describe the algorithm for variance covariance blocks of size 1. In this case Mplus simply implements a Metropolis-Hastings step to generate the variance parameter using normally distributed random walk jumping rule. Let's illustrate this algorithm using $\theta_{11}$. At each MCMC iteration we generate a new value $\theta_{11}^*$ value from the proposal distribution $N(\theta_{11}, \sigma^2)$. The value is accepted with probability $min(1, R)$ where $R$ is the acceptance ratio

$$R = \frac{P(\theta_{11}^*|*)}{P(\theta_{11}|*)} \frac{P_0(\theta_{11}^*)}{P_0(\theta_{11})}$$

where $P_0(\theta_{11})$ is the prior non-conjugate density of $\theta_{11}$ and $P(\theta_{11}|*)$ is the density given in (6) with $\alpha_0 = \beta_0 = 0$ (to eliminate the conjugate prior). If the proposed value is rejected the old value of $\theta_{11}$ is retained. The variance

$\sigma^2$ of the jumping distribution can be adapted within a preliminary stage of the estimation so that an acceptance rate between 0.25 and 0.50 is obtained.

Now we describe the Metropolis-Hastings algorithm for generating variance covariance matrices of size bigger than 1 with non-conjugate priors. Suppose that the prior for a $\Theta_{11}$ block is $P_0(\Theta_{11})$. Let's call this prior the target prior, using the terminology of Liu and Daniels (2006). Let's call the prior given in (7) the proposal prior. That proposal prior will be used as if it is the true target prior but only to construct a jumping distribution. If the prior is given by (7) then the posterior of $\Theta_{11}$ is given by (8). We use that posterior as the jumping distribution $J(\Theta_{11}) \sim IW(E + \Omega, n + f)$, i.e., we draw a new $\Theta^*_{11}$ matrix from that distribution. Let's now compute the acceptance probability

$$R = \frac{J(\Theta_{11})}{J(\Theta^*_{11})} \frac{P(\Theta^*_{11}|*)}{P(\Theta_{11}|*)} =$$

$$\frac{|\Theta_{11}|^{-0.5(n+f+p+1)} Exp(-0.5Tr((E+\Omega)\Theta_{11}^{-1}))}{|\Theta^*_{11}|^{-0.5(n+f+p+1)} Exp(-0.5Tr((E+\Omega)\Theta_{11}^{*-1}))}$$

$$\frac{|\Theta^*_{11}|^{-0.5(n)} Exp(-0.5Tr(E\Theta_{11}^{*-1}))}{|\Theta_{11}|^{-0.5(n)} Exp(-0.5Tr(E\Theta_{11}^{-1}))} \frac{P_0(\Theta^*_{11})}{P_0(\Theta_{11})} =$$

$$(|\Theta^*_{11}|/|\Theta_{11}|)^{0.5(f+p+1)} Exp(-0.5Tr(\Omega(\Theta_{11}^{-1} - \Theta_{11}^{*-1}))) \frac{P_0(\Theta^*_{11})}{P_0(\Theta_{11})}.$$

If we now choose the parameters in the proposal prior to be $f = -p - 1$ and $\Omega = 0$ the above formula simplifies to

$$R = \frac{P_0(\Theta^*_{11})}{P_0(\Theta_{11})}.$$

This approach generally works quite well unless the desired prior variance is very small. In such a case however the estimated parameter can be generally converted to a fixed parameter.

The above algorithm is used for non-conjugate priors for variance covariance matrices as well as partial correlation matrices when the default algorithm is used in Mplus for variance covariance generation. In Mplus this algorithm is specified as ALGORITHM=GIBBS(PX1). Several other algorithms are available for the generation of variance covariance matrices and partial correlation matrices, see Section 3.3.4. The availability of non-conjugate priors is specific for each of these algorithms. For ALGORITHM=GIBBS(PX2) non-conjugate priors are not available in Mplus, i.e., the priors are limited to the Inverse Wishart prior. For ALGORITHM=GIBBS(PX3) only the constant density prior is available. For ALGORITHM=GIBBS(RW) all priors are non-conjugate as this algorithm is already based on the Metropolis-Hastings algorithm.

# 7 Observed Mediator

There are two different ways to use a categorical variable as a predictor. One way is to use the underlying $Y^*$ variable as the predictor, which we call a latent mediator. Another way is to use the observed categorized value $Y$ as the predictor, which we call the observed mediator. Both options are implemented in Mplus. So far in this paper only $Y^*$ has been used in equations (1) and (2), since we simply converted categorical $Y$ into continuous $Y^*$ and used the estimation algorithm for continuous variables. Thus the algorithm described previously can be used to estimate the latent mediator models. In this section we briefly describe the implementation that is used in Mplus for the observed mediator model. For each categorical mediator variable we essentially create a new predictor $X$ variable that is equal to the categorized value. If the categorical variable has no missing values then nothing else is needed. However if the mediator has missing values those values need to be generated within the Gibbs sampler. So far we had not discussed generation of missing categorized value because it is not used anywhere, although that is a straight forward process because $Y^*$ and $\tau$ determines the categorized value completely. When the mediator is observed however we use the following approach to generate the missing categorical value $Y$. Let $Y_j$ is the categorical variable which has missing values and let $w_1$ represent the dependent variables that can be predicted by $Y_j$ and $w_2$ represents all other variables with the exception of $Y_j^*$, which is not a part of either $w_1$ and $w_2$. Under the assumption of model consistency one can split the variables into variables that can be predicted by $Y_j$, these are the variables in the $w_1$ vector, and variables that can predict $Y_j$, these are the variables in the $w_2$ vector. Then

$$P(Y_j = k|w_1, w_2) = \frac{P(w_1|Y_j = k, w_2)P(Y_j = k|w_2)}{\sum_k P(w_1|Y_j = k, w_2)P(Y_j = k|w_2)}$$

where $P(Y_j = k|w_2)$ is computed from the probit regression implied from the model for $Y_j$ and $P(w_1|Y_j = k, w_2)$ represents a conditionally normal distribution where $Y_j$ and $w_2$ are the predictors.

Using the above computation we can sample from the distribution $[Y_j|*]$ when $Y_j$ is missing. This step is then followed by the generation of $[Y_j^*|*, Y_j]$ which is done the way as in the case when $Y_j$ is not missing.

# 8 Mixture Models

Let $C$ be a categorical latent variable, which takes the values from $1, ..., k$ with probabilities $p_1$,...,$p_k$. The mixture model is described by the following two equations

$$[y|C = j] = \nu_j + \Lambda_j \eta + K_j x + \varepsilon \tag{17}$$

$$[\eta|C = j] = \alpha_j + B_j \eta + \Gamma_j x + \zeta \tag{18}$$

where $\varepsilon$ and $\zeta$ in class $C = j$ have variance covariance matrices $\Theta_j$ and $\Psi_j$. In addition if $Y_i$ is a categorical variables we have

$$Y_i = t|C = j \Leftrightarrow \tau_{t-1\ ij} \leq Y_i^* < \tau_{tij}$$

## 8.1 Estimation

If the latent variable $C$ is actually observed then the model is essentially a multiple group structural equation model. The parameters $p_1$,...,$p_k$ are essentially fixed parameters and the estimation of such a model essentially repeats the estimation for the non-mixture model $k$ times at every step in the Gibbs sampler. When $C$ is truly a latent variable then we have to include in the Gibbs sampler the following two new steps

Steps 1. Generate $C$ from the conditional distribution $[C|*]$.

Steps 2. Generate the parameters $p_1$,...,$p_k$ from the conditional distribution $[p_j|*]$.

The second step is quite simple. Assuming a conjugate Dirichlet prior $D(a_1, ..., a_k)$ the conditional distribution $[p_j|*]$ is simply $D(a_1+n_1, ..., a_k+n_k)$, where $n_j$ is the number of observations in class $j$, i.e., $n_j = \sum_{i=1}^n \delta_{ij}$ where $\delta_{ij} = 1$ if $C_i = j$ and is 0 otherwise.

The first step however is more complicated. Three different methods are implemented in Mplus for generating the $C$ variable. The methods differ in the way the $C$ variables are grouped for use in the Gibbs sampler. Mplus will automatically attempt to determine the best method for each model and will use that method, however the Methods can also be pre-specified by the user as well.

### 8.1.1 Method 1

Method 1 uses the $C$ variables as a separate group and generates $C$ from

$$[C|Y, \eta, Y^*, X]. \tag{19}$$

Let $w$ be the vector of all $\eta$, all $Y^*$ for all categorical dependent variables and all $Y$ for all continuous dependent variables. The conditional distribution of $w$ is

$$[w|X, C = j] \sim N(a_j + b_j X, v_j) \tag{20}$$

where $a_j$, $b_j$ and $v_j$ can be obtained from (17) and (18) using simple matrix algebra. The conditional distribution (19) is then given by

$$P(C = j|Y, \eta, Y^*, X) = \frac{p_j P(w|X, C = j)}{\sum_j p_j P(w|X, C = j)}$$

where $P(w|X, C = j)$ is the density implied by (20).

### 8.1.2 Method 2

Method 2 groups $C$ and $\eta$ in one Gibbs sampler group and generates $C$ and $\eta$ in the following sequence

$$[\eta, C|Y, Y^*, X] = [C|Y, Y^*, X][\eta|C, Y, Y^*, X].$$

Note that sampling from $[\eta|C, Y, Y^*, X]$ is the same as for non-mixture models so here only the first conditional distribution

$$[C|Y, Y^*, X] \tag{21}$$

is described. Let $w$ be the vector of all $Y^*$ for all categorical dependent variables and all $Y$ for all continuous dependent variables. The conditional distribution of $w$ is

$$[w|X, C = j] \; N(a_j + b_j X, v_j) \tag{22}$$

where $a_j$, $b_j$ and $v_j$ can be obtained from (17) and (18) using simple matrix algebra. The conditional distribution (21) is then given by

$$P(C=j|Y,Y^*,X)=\frac{p_jP(w|X,C=j)}{\sum_j p_jP(w|X,C=j)}$$

where $P(w|X,C=j)$ is the density implied by (22). Method 2 is available for all models and it is usually more efficient (i.e. provides better mixing and faster convergence) than Method 1.

### 8.1.3 Method 3

Method 3 is available only for models without continuous latent variables $\eta$. This method groups $C$ and $Y^*$ in one Gibbs sampler group and generates $C$ and $Y^*$ in the following sequence

$$[C,Y^*|Y,X]=[C|Y,X][Y^*|C,Y,X].$$

Note that the last conditional distribution $[Y^*|C,Y,X]$ is the same as for non-mixture models so here only the first conditional distribution $[C|Y,X]$ is described. Let $w_1$ be the vector of all $Y$ for all continuous variables and let $w_2$ be the vector of all $Y^*$ for all categorical variables and let $w_3$ be the vector of all categorical $Y$ variables.

Using (17) and (18) we obtain the following conditional distributions

$$[w_1|X,C=j]\sim N(a_j+b_jX,v_{1j}) \tag{23}$$

$$[w_2|X,w_1,C=j]\sim N(c_j+d_jXe_jw_1,v_{2j}) \tag{24}$$

using simple matrix algebra. The conditional distribution $[C|Y,X]$ is now given by

$$P(C=j|Y,X)=\frac{p_jP(w_3|X,w_1,C=j)P(w_1|X,C=j)}{\sum_j p_jP(w_3|X,w_1,C=j)P(w_1|X,C=j)}$$

where $P(w_1|X,C=j)$ is the density implied by (23) and $P(w_3|X,w_1,C=j)$ is essentially given by (24) with the probit link function. Note however that if $v_{2j}$ is not diagonal $P(w_3|X,w_1,C=j)$ will involve the multivariate probit function which is very computationally demanding and thus this Method 3 is only available when $v_{2j}$ is diagonal. One example where $v_{2j}$ is not diagonal is the case when the residual terms corresponding to categorical variables in (17) are correlated.

There are certain models that can be done only with Method 3. We will describe two such examples. The first example is an LCA model with a perfect binary indicator $Y_1$ in class 1. Suppose that $P(Y_1 = 2|C = 1) = 1$. If $C$ and $Y_1^*$ are not generated in the same step then in the MCMC simulation $C$ and $Y_1^*$ will never change. That is because $C = 1$ will imply a large value for $Y_1^*$ which in turn will imply $C = 1$, i.e., if $Y_1 = 2$ and the starting value for $C$ is 1 - the MCMC will not generate another value regardless of the rest of the data for that observation. The second example is an LCA model with categorical variables that have more than 2 categories. In this case if $C$ is not generated directly from the observed categorical $Y$ values, but from $Y^*$, we again will have convergence problems because $Y^*$ is simply a zero mean standard normal variable that carries no information about $C$. The convergence problems described above may actually be difficult to detect in practice because they are based to a large extent on problems with some of the latent variables rather than with the parameter estimates which are closely monitored by Mplus. Method 3 would be the preferred method by Mplus when the method is available.

## 8.2 Label Switching

Label switching is a well documented estimation problem for Bayesian estimation of mixture models, see for example Celeux et al. (2000). Most Mixture models do not uniquely determine the class labelling, i.e., the first class in a model can actually be set as the second class by simply switching the parameters in the first two classes. This means that the posterior distribution will contain essentially $k!$ symmetric peaks since there are $k!$ possible ordering of the classes. If these peaks are not disjoint an MCMC sequence that is run long enough will provide this posterior distribution. The problem however with this posterior distribution is that we are not interested in the entire posterior with all $k!$ peaks but only in one of these peaks. If we look for the mean or the median of the entire posterior we can obtain values that may be between the peaks and are so unlikely that as point estimates are entirely useless. The mode point estimates however do not have that problem. The estimates for the multivariate mode will always be in just one of these symmetric peaks, whichever peak has had the most values occur during the finite number of iterations.

If we need more than just point estimates however, using the mode point estimates will not resolve the problems. Note that if the peaks are connected it is not clear how to define $1/k!$ portion of the posterior distribution that can be used for inference. So the problem is that the one peak search is not even well defined. For large samples the peaks are not connected and then there is no such problem. For moderately large samples the peaks might be connected only in their tails which means that we could probably ignore the ”undefined” problem because, regardless of the way it is defined it will only affects the tails of the posterior distribution. For small samples however the peaks may be connected more heavily and the precise definition of what is $1/k!$ of the posterior will affect all results quite heavily.

Label switching can be prevented by identifying the labelling uniquely and that can be done by introducing parameter constraints (inequalities) among the model parameters. Depending on which inequalities are used however a different definition of $1/k!$ arises.

There are two types of label switching that are possible. Both of these are described in the following two sections.

### 8.2.1 Label switching between chains

This label switching means that one MCMC chain has converged to one peak in the posterior, while another chain has converged to another peak. Using PSR convergence criteria will basically conclude non-convergence even though both chains have converged well. One possible solution to this problem is to simply run one chain only to estimate the model. Alternative solution is that Mplus runs a number of identical iterations in all of the chains within a preliminary stage and then runs non-identical iterations among the chains to estimate the posterior and determine convergence. Because of the initial identical iterations phase it is likely that all the chains will near the same peak and in the second stage when the iterations are different among the chains the chains will stay in that same peak. The longer the preliminary phase is the greater the chance for success will be in this alternative solution. This method is implemented in Mplus and by default 50 identical iterations are done prior to generating the independent chains.

Note also that this kind of label switching can occur even for large samples simply because of starting values pointing towards different peaks.

### 8.2.2 Label switching within chains

When the sample size is not sufficiently large and the symmetric peaks of the posterior are connected the MCMC chain will run through all the peaks in the posterior distribution if it is run long enough. Because the peaks are connected via their tails the MCMC chains will not usually reach these tails very easily and the actual switches between the peaks may be so unfrequent that obtaining a satisfactory estimate for the posterior distribution will require basically a vast number of iterations. Using a convergence criteria such as PSR will basically lead to non-convergence conclusion due to the fact that it will likely compare posteriors from different peaks in some way. The solution to this problem is to either use parameter inequalities as model constraints to identify the class ordering uniquely or just monitor the estimation process very closely and use a section of the posterior that is visibly clear from a single peak. Note however that when using parameter constraints to identify the class ordering we need to choose parameters that are truly different among the classes. If we chose parameters that are not very different among the classes we could define the $1/k!$ component very inappropriately. This means that often some kind of pre-processing will be needed to identify which parameters could be used for class labelling identification.

Another solution for preventing label switching that is available in Mplus is to have several observations preassigned with preassigned class membership. If these observations are chosen properly this may prevent label switching. In Mplus this can be implemented by introducing a perfect latent class indicator which has missing values for all observations that are not preassigned and is observed only for those that are preassigned. A perfect latent class indicator is such that $P(Y = i|C = i) = 1$. It is easy to include such indicator in the model.

Note also that if only point estimates are desired it is safe to use the multivariate mode as point estimator and simply ignore label switching as well as the convergence monitoring via PSR.

## 9 Multilevel Models

Suppose that the observations are not independent but are clustered into higher level units such as classrooms or cities. The multilevel model we consider is based on the following decomposition

$$Y = Y_w + Y_b$$

here $Y_w$ is the within part of the dependent variable and $Y_b$ is the cluster level dependent variable. Both $Y_w$ and $Y_b$ are unobserved. Each of these two variables are assumed to satisfy a structural equation model as described in (1) and (2).

Estimating such a model with the Bayes estimator amounts to adding an additional step into the Gibbs sampler that will be able to sample $Y_b$ from the conditional distribution $[Y_b|*]$. After $Y_b$ is sampled, we compute $Y_w = Y - Y_b$ and then the within and the between models are estimated separately using the same method as for single level models.

The conditional distribution $[Y_b|*]$ is fairly complex, however, it is well known because it is used in the maximum-likelihood estimation of this model with the EM algorithm. This can be found among other places in Asparouhov and Muthén (2003) as well as Bentler and Liang (2003). In particular, Asparouhov and Muthén (2003) includes this conditional distribution for the two-level model even when there are random slopes, i.e., any of the coefficients $K$ and $\Gamma$ can vary across clusters and can be between level variables that are included in the structural model on the between level. Thus the Bayesian estimation of the multilevel model can also easily include random slopes.

# 10 Posterior Predictive Checking

Mplus implements posterior predictive model checking for all single level models using the classical chi-square fit function as the discrepancy function as well as some other discrepancy functions. The classical chi-square fit function is simply the likelihood ratio test (LRT) between the structural equation model and an unrestricted mean and variance covariance model.

First let's define the posterior predictive P-value (PPP) for an arbitrary discrepancy function f. Suppose that $\theta$ is the vector of all model parameters. The discrepancy function uses as arguments the observed data $Y$ and $X$ as well as model parameters $\theta$, i.e., $f = f(Y, X, \theta)$. First assume that small values of $f$ indicate better fit between the data and the model. This is precisely the case when the discrepancy function is the classical chi-square fit function.

At each MCMC iteration $t$ the discrepancy function $f(Y, X, \theta_t)$ is computed with the current parameter estimates and the data. In addition, at each MCMC iteration $t$ we generate a new data set $\tilde{Y}_t$ of the same size as the original data set from the estimated model (1) and (2) using the current parameter estimates $\theta_t$. The discrepancy function $f(\tilde{Y}_t, X, \theta_t)$ is then computed using the new data set $\tilde{Y}_t$. The posterior predictive P-value (PPP) is then defined as

$$PPP = P(f(Y, X, \theta) < f(\tilde{Y}, X, \theta)) \approx \frac{1}{m} \sum_{t=1}^{m} \delta_t$$

where $\delta_t = 1$ if $f(Y, X, \theta_t) < f(\tilde{Y}, X, \theta_t)$ and 0 otherwise. It is important to reach convergence in the MCMC process before using the iterations for computing the PPP value. Mplus uses every 10-th iteration for computing the PPP value to reduce the computational burden. Mplus will use the same iterations for computing PPP as those used for computing the point estimates and standard errors etc., with the modification that it will use every 10-th iteration.

If the discrepancy function is not a measure of fit then much smaller $f(Y, X, \theta)$ values then those obtained from $f(\tilde{Y}, X, \theta))$ also can indicate model misfit of some kind. Thus the PPP value for a non-fit function is defined as the smaller of the PPP values for $f$ and $-f$.

Low PPP value indicates clearly that the model is not appropriate for this data and that there is some kind of misspecification. The PPP value tests the model against an unspecified alternative. It can reject the model even in cases when the structural model is actually an unrestricted variance covariance matrix. That is because the discrepancy function is not maximized over the parameter, but rather is computed with the current parameter estimates. This implies that even for the unrestricted model the discrepancy function is not 0 and that allows us to detect misspecifications even in the unrestricted model.

The misspecification in that case can simply be due to non-normality of the data. In some respects the PPP value is more attractive than a classical P-value because it takes the variability of the parameters into account. For example the classical P-value tests the model with the maximum-likelihood estimates assumed as the actual estimates, ignoring the fact that the estimates are measured with error. The PPP value takes that error into account. In addition the PPP value does not depend on asymptotic theory.

The discrepancy functions used in Mplus are as follows. The classical LRT chi-square test of fit function

$$f = 0.5n(log|\Sigma| + Tr(\Sigma^{-1}(S + (\mu - m)(\mu - m))) - log|S| - p - q) \quad (25)$$

where $S$ is the sample variance covariance, $m$ is the sample mean, $\Sigma$ is the model implied variance covariance and $\mu$ is the model implied means. This discrepancy function has been used for example in Scheines et al. (1999). Covariates $X$ in the above formula are treated the same way as dependent variable, except that the model is complimented with an unrestricted mean and variance model for $X$. This is just an artificial treatment that simplifies the expressions. An actual conditioning on the covariates leads to exactly the same chi-square values. When there are categorical variables in the model we use $Y^*$ to evaluate $f$. For mixture models the fit functions across the classes are added up. When there are missing values in the original data the discrepancy function $f$ is computed using the current sampled values. Mplus will not compute the classical LRT chi-square test of fit function using only the observed data, because that test is much more computationally intensive and it requires pattern by pattern computations. Instead Mplus will compute (25) using the observed $Y$ values as well as the generated $Y$ values for those values that are missing. Thus in the case of missing data $f(Y, X, \theta)$ will vary not only due to changes in $\theta$ but also in $Y$. This discrepancy function can be used among other things for the same purpose as the classical chi-square test, i.e., to evaluate whether or not a factor model accounts for all the correlations among the factor indicators.

Mplus computes a number of other PPP values using other fit functions designed specifically for categorical variables. For each categorical variable Mplus computes the PPP value using as discrepancy function the univariate log-likelihood for that variable. Multivariate log-likelihood is not possible because it involves computationally intensive multivariate probit function. The sum of all univariate likelihoods for all categorical variables is also used as a discrepancy function. In addition, for each categorical variable and each category for that variable the observed percentage/frequency of that category is also used as a double sided discrepancy function.

# 11 Multiple Imputations

In Mplus Version 6 the MCMC simulation can also be used for multiple imputation (MI) of missing data. This method was pioneered in Rubin (1987) and Schafer (1997). The imputed data can be analyzed in Mplus using any classical estimation methods such a maximum-likelihood and weighted least squares (WLS). This is particularly of interest for the WLS estimator which is biased when the missing data is MAR and the dependent variables affect the missing data mechanism. Using the MI method in combination with WLS resolves that problem.

The missing data is imputed after the MCMC sequence has converged. Mplus runs 100 MCMC iterations and then stores the generated missing data values. The process is repeated until the desired number of imputations have been stored. These imputed missing data sets are essentially independent draws from the missing data posterior. The missing data can be imputed in Mplus from a single-level or from a two-level model. The data can be imputed from an unrestricted model (H1 model), which we call H1 imputation, or it can be imputed from any other model that can be estimated in Mplus with the Bayesian estimator, which we call H0 imputation. Unrestricted models are general enough so that model misspecification can not occur. However, these models have a large number of parameters and convergence is sometimes difficult to achieve, particularly for large multivariate sets with many variables that include combinations of categorical and continuous. Unrestricted two-level models can also have convergence problems because of the large number of parameters estimated on the between level sometimes using only a limited number of two-level units/clusters. In case of convergence problems with the H1 imputations, the H0 imputation offers a viable alternative as long as the estimated model used for the imputation fits the data well. With H0 imputation some ground breaking opportunities arise, such as, imputation from LCA models and factor analysis models.

## 11.1 Unrestricted Imputation Models

Three different unrestricted H1 models have been implemented in Mplus for the H1 imputation. All three models are defined for the combination of categorical and continuous variables. Prior to estimating the H1 model all continuous data is standardized so that the mean is zero and the variance is 1 for each continuous variable. After estimation the continuous variable are transformed back to their original scale. In the following sections we describe the three H1 imputation models.

### 11.1.1 Variance Covariance Model

In this model all variables in the data set are assumed to be dependent variables. If $Y$ is the vector of all of these dependent variables the model is given by

$$y = \nu + \varepsilon \tag{26}$$

where $\varepsilon$ is a zero mean vector with variance covariance matrix $\Theta$ which is one full block of unrestricted variance covariance matrix with 1s on the diagonal for each categorical variable in the model, see (11). In addition the vector $\nu$ has means fixed to 0 for all categorical variables and free for all continuous variables. For all categorical variables we estimate also all thresholds as defined in (10).

The two-level version of this model is as follows

$$y = \nu + \varepsilon_w + \varepsilon_b \tag{27}$$

where $\varepsilon_w$ and $\varepsilon_b$ are zero mean vectors defined on the within and the between level respectively with variance covariance matrices $\Theta_w$ and $\Theta_b$. Both of these matrices are one full block of unrestricted variance covariance. Again the vector $\nu$ has means fixed to 0 for all categorical variables and free for all continuous variables. For all categorical variables we estimate also all thresholds again as defined in (10). If a variable is specified as within-only variable the corresponding component in the $\varepsilon_b$ vector is simply assumed to be 0, which implies that also the corresponding parameters in the variance covariance matrix $\Theta_b$ are 0. Similarly if a variable is specified as between-only variable the corresponding component in the $\varepsilon_w$ vector is simply assumed to be 0, which implies that also the corresponding parameters in the variance covariance matrix $\Theta_w$ are 0. For categorical variables for identification purposes again the variance of the variable in $\Theta_w$ is fixed to 1, with the exception of the case when the categorical variable is between-only. In that case the variance on the between level in $\Theta_b$ is fixed to 1.

This model is the default imputation model in all cases.

### 11.1.2 Sequential Regression Model

In this model all variables in the data set are assumed to be dependent variables as well. The model is defined by the following equations

$$y_1 = \nu_1 + \beta_{12} y_2 + \beta_{13} y_3 + ... + \beta_{1p} y_p + \varepsilon_1 \tag{28}$$

$$y_2 = \nu_2 + \beta_{23} y_3 + \beta_{24} y_4 + ... + \beta_{2p} y_p + \varepsilon_2 \tag{29}$$

$$...$$

$$y_p = \nu_p + \varepsilon_p \tag{30}$$

where $\varepsilon_1$,...,$\varepsilon_p$ are independent residuals with variances $\theta_{11}$,...,$\theta_{pp}$. Essentially in this model we have replaced the parameters $\theta_{ij}$, $i < j$ in the variance covariance model described in the previous section with the regression parameters $\beta_{ij}$, $i < j$. For two-level models this $\theta_{ij}$ to $\beta_{ij}$ conversion is basically applied to both levels. The identification restrictions needed for categorical variables are as for the variance covariance model.

The above model was pioneered in Raghunathan et al. (2001). It is particularly powerful and useful in the case of combination of categorical and continuous variables when used also in the framework of observed mediators as defined in Section 7. Note that depending on how the mediator is treated we actually have two different models for H1 imputation defined here, i.e., sequential regression with observed mediators and sequential regression with latent mediators. The default is the observed mediator model. This model is the easier to estimate among the two models.

### 11.1.3 Regression Model

In this model all variables in the data set that have missing data are assumed to be dependent variables $Y$ and all variables that do not have missing data are assumed to be covarites $X$. The model is defined by

$$y = \nu + Kx + \varepsilon \tag{31}$$

where $\nu$ and $\varepsilon$ are as in the variance covariance model. The two level generalization for this model is also simply a generalization of the two-level variance covariance model with the addition to the covariates. For two-level models each covariate is classified as either within-only or between-only, i.e., each covariate is used on just one of the two levels.

One advantage of this model is that if only a few variables have missing values the unrestricted model will have much fewer number of parameters then the previous two models and will likely reach convergence faster.

## 11.2 Plausible Values

Plausible values are essentially imputed values for latent variables. All latent variables can essentially be thought of as observed variables that have missing data for all observations. When imputation is done with an H0 model that includes latent variables Mplus can stores the current values for all latent variables. If the imputations are taken 100 iterations apart we can assume that these are independent draws from the posterior distribution of the latent variables. The plausible values can be used in secondary analysis the same way missing data imputations are used, i.e, by combining the results across the imputations, see Rubin (1987).

Using a sufficient number of plausible values the posterior distribution for every latent variable can be constructed. Empirical Bayes analysis, see for example Carlin (1992), can be conducted using plausible values. For other applications of plausible values see also Mislevy et al. (1992).

# 12 Informative and Non-informative Priors

In this section we describe all priors that can be used in Mplus and will discuss methods for selecting informative an non-informative priors.

The normal distribution prior is specified as $N(\mu, v)$ where $\mu$ is the mean parameter and $v$ is the variance parameter. To specify a non-informative prior you can select $N(0, 10^{10})$. This prior is numerically equivalent to the improper prior which has constant density of 1 on the interval $(-\infty, \infty)$. In Mplus, the default prior for all intercepts, loadings and slopes for normally distributed variables is $N(0, 10^{10})$. The default prior for all intercepts, loadings and slopes for categorical variables is $N(0, 5)$. The default prior for all thresholds is $N(0, 10^{10})$.

Informative priors can be specified easily with the normal prior and can be used to accumulate information across different studies, see for example Yuan and MacKinnon (2009).

The Inverse Gamma distribution is specified as $IG(\alpha, \beta)$. The density function defined for $x > 0$ is

$$f(x) \sim x^{-\alpha-1} Exp(-\beta/x).$$

The mode of the density is $\beta/(\alpha+1)$. The default non-informative prior in Mplus is $IG(-1,0)$ which has constant density of 1 on the interval $(-\infty,\infty)$. Other popular choices for non-informative priors are $IG(0,0)$ (which produces a prior with density proportional to $1/x$) or $IG(0.001,0.001)$, see Browne and Draper (2006). In some small sample size cases we have found that $IG(1,2)$ works quite well in preventing random effect variance collapsing or exploding.

The uniform distribution can be specified in Mplus as $U(a,b)$ where this is the uniform distribution with equal/constant density over the interval $(a,b)$. To choose a non-informative prior one can choose $(-10^{10},10^{10})$ or $(0,10^{10})$ if the prior is for a variance parameter, see for example Browne and Draper (2006).

The Gamma distribution is specified as $G(\alpha,\beta)$. The density function defined for $x > 0$ is

$$f(x) \sim x^{\alpha-1} Exp(-\beta x).$$

The log-normal distribution is specified as $LN(\mu,v)$. The density function is

$$f(x) \sim x^{-1} Exp(-(log(x)-\mu)^2/(2v)).$$

The Inverse Wishart prior is specified as $IW(\Omega,d)$, where $\Omega$ is a positive definite matrix of size $p$ and $d$ is an integer. The density function is

$$f(\Sigma) \sim |\Sigma|^{-(d+p+1)/2} Exp(-Tr(\Omega\Sigma^{-1})/2)$$

The mode of this distribution is $\Omega/(d+p+1)$. To specify a non-informative prior one can specify $IW(0,-p-1)$, i.e., $\Omega$ has all entries 0, and $d=-p-1$. This prior is essentially the uniform prior on $(-\infty,\infty)$ for all $\Sigma$ parameters. Other popular non-informative priors are $IW(0,0)$ and $IW(I,p+1)$. The specification $IW(I,p+1)$ produces marginal priors for the correlation parameters that is uniform on the interval $(-1,1)$ and $IG(1,0.5)$ marginal prior for the variance parameters. To specify in Mplus an informative prior $IW(\Omega,d)$ for $\Sigma$, one has to specify $\sigma_{ij} \sim IW(\Omega_{ij},d)$ for $i=1,..,p$ and $j=1,..,i$, where $\sigma_{ij}$ is the corresponding parameter label and $\Omega_{ij}$ is the $(i,j)$ entry in the $\Omega$ matrix.

The default prior in Mplus for variance covariance matrices for continuous variables is $IW(0, -p-1)$. The default for categorical variables or for the combination of categorical and continuous variables is $IW(I, p+1)$.

The Dirichlet distribution with parameters $\alpha_1$, $\alpha_2$, ...,$\alpha_K$ has a density

$$f(x_1, x_2, ..., x_K) = x_1^{\alpha_1-1} x_2^{\alpha_2-1} ... x_K^{\alpha_K-1}$$

under the constraint $x_1+x_2+...+x_K = 1$. In Mplus this distribution is used as the prior for the class proportions in Mixture models. The common non-informative prior is given by $\alpha_i = 1$. The Mplus default is however $\alpha_i = 10$ to prevent the formation of small class solutions. Typically classes that are formed by only a few observations are not of substantive interest. For small sample size problems however the Mplus default may be inappropriate. To specify an informative prior in Mplus one has to specify $p_i \sim D(\alpha_i, \alpha_K)$ where $i = 1, ....K-1$ and $p_i$ is the label of the parameter $[C\#i]$.

# References

[1] Albert, J. H. & Chib, S. (1993) Bayesian Analysis of Binary and Polychotomous Response Data. Journal of the American Statistical Association, Vol. 88, No. 422., pp. 669-679.

[2] Arminger, G. and Muthen, B. (1998) A Bayesian approach to nonlinear latent variable models using the Gibbs sampler and the Metropolis-Hastings algorithm. Psychometrika, 63, 271-300.

[3] Asparouhov, T., & Muthén, B. (2003) Full-information maximum-likelihood estimation of general two-level latent variable models with missing data: A technical report. Los Angeles: Muthén & Muthén.

[4] Bentler, P. M., & Liang, J. (2003) Two-level mean and covariance structures: Maximum likelihood via an EM algorithm. In S. P. Reise & N. Duan (Eds.)3 Multilevel modeling: Methodological advances, issues, and applications (pp. 53-70). Mahwah, NJ: Erlbaum.

[5] Boscardin, J; Zhang, X. & Belin, T. (2008) Modeling a mixture of ordinal and continuous repeated measures. Journal of Statistical Computation and Simulation Vol. 78, No. 10, 873-886.

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

[7] Chib, S. and Greenberg, E. (1998). Bayesian analysis of multivariate probit models. Biometrika 85, 347-361.

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

[9] Celeux G., Hurn M., and Robert C.P. (2000) Computational and inferential difficulties with mixture posterior distributions. Journal of the American Statistical Association 95: 957-970.

[10] Cowles, M.K. (1996) Accelerating Monte Carlo Markov Chain Convergence for Cumulative-Link Generalized Linear Models. Statistics and Computing 6, 101-111.

[11] Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B.(2004) Bayesian Data Analysis. London, Chapman & Hall.

[12] Johnson, V.E., and Albert, J.H. (1999) Ordinal Data Modeling, New York: Springer.

[13] Lee S.Y. (2007) Structural Equation Modelling: A Bayesian Approach, London: John Wiley & Sons.

[14] Little, R.J.A. and Rubin, D.B. (1987) Statistical Analysis with Missing Data. J. Wiley & Sons, New York.

[15] Liu, X. and Daniels, M.J. (2006). A new algorithm for simulating a correlation matrix based on parameter expansion and re-parameterization. Journal of Computational and Graphical Statistics, 15, 897-914.

[16] 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.

[17] Muthen, B. and Arminger, G. (1995) Bayesian Latent Variable Regression for Binary and Continuous Response Variables Using the Gibbs Sampler. Draft.

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

[20] Scheines, R.; Hoijtink, H. and Boomsma, A. (1999) Bayesian estimation and testing of structural equation models. Psychometrika, 64,37-52.

[21] Schafer, J.L. (1997) Analysis of Incomplete Multivariate Data. Chapman & Hall, London.

[22] Yuan, Y. & MacKinnon, D. (2009) Bayesian Mediation Analysis, Psychological Methods, Vol. 14, No. 4, 301-322.