# Supplemental Material for Multiple Imputation of longitudinal categorical data through Bayesian mixture latent Markov models

## A Setting the prior distribution

As outlined in Section 2.1, independent Dirichlet distributions can be specified for each Multinomial in model (1)-(2). In an MI context, in which the imputation model does not necessarily match the analysis model, it is common to have no previous knowledge about the imputation model parameters. In such a case, symmetric Dirichlet priors can be chosen: $Dirichlet(c_1, c_2, ..., c_D)$ where $c_1 = c_2 = ... = c_D$. This is the approach we used in all the experiments of the paper, and implied in the remaining of the current section.

Rousseau and Mergensen (2011) found out that when a Bayesian mixture model is overfitting the data (as our model selection approach of Section 2.2 implies), units are allocated by the Gibbs sampler to some of the extra LCs if each component of the latent probabilities hyperparameter is at least as large as half times the number of free parameters within each components. For the BMLM model, this means that each pseudo-count of the LSs $\alpha_k \ \forall \ k$ should be set at least equal to $\sum_j (R_j - 1)/2$. Following the guidelines given in Vidotto, Vermunt, and van Deun (2018), who examined the behavior of the prior distribution in standard Bayesian LC models (for the MI of cross-sectional missing data), we suggest increasing $\alpha_k$ and $\gamma_k \ \forall \ k$ in such a way that as many states $s_1, ..., s_T$ as possible are occupied during the imputation stage, which can be assessed with the MCMC output. By manipulating with trial-and-error (before the imputation step) the hyperparameters in the priors of the latent states probabilities, we decided to set $\alpha_k = \gamma_k = 5$ in the study of Section 3, while in the empirical experiment of Section 4 - in which the number of within-state free parameters was equal to 27 - we arbitrarily set

$\alpha_k = \gamma_k = 100$. As reported in Vidotto et al. (2018), full allocation of the latent classes/states helps to capture all relevant associations in the data, preventing the sampler from becoming unstable; in fact, in this way the states are identified by the data, rather than by the prior distribution of the emission probabilities. As a consequence, an unstable Gibbs sampler can yield poor imputations.

In the empirical study we found out by means of pre-imputation inspections that reinforcing the prior persistence probabilities caused the Gibbs sampler to produce higher likelihood values (on average) during its iterations. In turn, this could help the BMLM model to better recover the lagged relationships specified for that study. Persistence probabilities are represented by the diagonal elements of the matrix $\boldsymbol{X}_l$. These probabilities can be reinforced by manipulating the hyperparameter vector of the $q$-th row of $\boldsymbol{X}_l$, by setting it equal to $\boldsymbol{\gamma} = (\gamma_1, ..., \gamma_q^*, ..., \gamma_K)$ with $\gamma_q^* > \gamma_k \; \forall \; k \neq q$. In the empirical study this was achieved by setting $\gamma_q^* > \sum_{k \neq q} \gamma_k$, with $\gamma_k = 100$ and $\gamma_q^* = K\gamma_k = 100K$ (in which $K = 9$). Reinforcing the persistence probabilities in the simulation study of Section 3 was not necessary, since increasing it did not entail any increase in the (averaged) likelihood values produced during the Gibbs sampler iterations.

Concerning the hyperparameters for the weights of the time-constant LCs, we decided to perform the imputations of both the study in Section 3 and the experiment in Section 4 by setting $\eta_l$ equal to the number of free parameters within each time-constant component, i.e., we set $\eta_l = \{(K-1)(K+1) + K(\sum_j R_j - 1) + \sum_p U_p - 1\} \; \forall \; l$.

Lastly, for the time-constant conditional and the time-varying emission probabilities we follow the guidelines of Vidotto et al. (2018) and set $\zeta_{upl} = \delta_{rjkl} = 0.01$ or $0.05 \; \forall \; u, p, r, j, k, l$ (final results are usually similar for these two values). This setting helps to make the prior pseudo-counts of the parameters ruling the conditional distribution of the observed data less influential in the imputation step.

# B BMLM model estimation

In this section, the Gibbs sampler for the BMLM model estimation is described. It is assumed that $L$, $K$, and the model hyperparameters have been established already according to the guidelines of Section 2.2 and Appendix A. Furthermore, also the total number of Gibbs sampler iterations $B$ should be chosen. $I$ of these $B$ iterations will be used as burn-in (such that model

estimation is performed on the last $B - I$ iterations). $I$ should be large enough to make the sampler attain the equilibrium distribution of the model parameter, which can be assessed by typical MCMC output inspection, e.g., by considering the traceplot of the log-likelihood functions generated at each iterations (as suggested by Vidotto et al. (2018)). Additionally, $\boldsymbol{\theta}^{(0)}$ is initialized by sampling all model parameters from uniform Dirichlet distributions, in such a way to increase the likelihood of initializing the sampler in the interior of the parameter space, speeding up convergence.

Algorithm 1 reports the steps for the Gibbs sampler. In order to sample the states of the Markov chain for each subject, multi-move sampling is used. The steps necessary to perform multi-move sampling are shown in Algorithm 2. Multi-move sampling, in turn, requires the calculation of the filtered state probabilities $\Pr(s_t = k|\boldsymbol{\theta}, w = l, \mathbf{y}_{it})$, the computation of which is described in Algorithm 3.

## B.1 The Gibbs sampler

**Algorithm 1**

For b=1,...,B:

1. for $i = 1, ..., n$ sample a LS $w^{(b)}$ from a Multinomial distribution with probabilities

$$\Pr(w^{(b)} = l|\boldsymbol{\theta}^{(b-1)}, \mathbf{z}_i, \mathbf{y}_i) = \frac{\omega_l^{(b-1)} \Lambda_{\mathbf{u}l}^{(b-1)} \pi_{\tilde{\mathbf{r}}l}^{(b-1)}}{\sum_c \omega_c^{(b-1)} \Lambda_{\mathbf{u}c}^{(b-1)} \pi_{\tilde{\mathbf{r}}c}^{(b-1)}}$$

   for each $l = 1, ..., L$, and where $\pi_{\tilde{\mathbf{r}}l}^{(b-1)} = \Pr(\mathbf{y}_i = \tilde{\mathbf{r}}|w = l)^{(b-1)}$ (equation 2);

2. for each $i = 1, ..., n$ and for all time points $t = 1, ..., T$, conditioned on the LC $w^{(b)}$, sample a LS $s_t$ from

$$\Pr(s_t^{(b)}|\boldsymbol{\theta}^{(b-1)}, w^{(b)} = l, \mathbf{y}_{it}).$$

   This can be achieved with multi-move sampling (see Algorithm 2 below);

3. for $l = 1, ..., L$, update the mixture weights $\boldsymbol{\omega}$ with
   $\boldsymbol{\omega}^{(b)}|w^{(b)} = l, \boldsymbol{\eta} \sim$

$$Dirichlet\left(\eta_1 + \sum_{i=1}^{n} \mathcal{I}_i(w^{(b)} = 1), ..., \eta_L + \sum_{i=1}^{n} \mathcal{I}_i(w^{(b)} = L)\right)$$

where $\mathcal{I}_i(w^{(b)} = l) = 1$ if for unit $i$ $w^{(b)} = l$ and 0 otherwise;

4. for $l = 1, ..., L, p = 1, ..., P$ update the conditional probabilities
$\boldsymbol{\lambda}_{pl}^{(b)} | w^{(b)} = l, \mathbf{z}^{obs}, \boldsymbol{\zeta}_{pl} \sim$

$$Dirichlet\left(\zeta_{1pl} + \sum_{i:w^{(b)}=l} \mathcal{I}(z_{ip} = 1), ..., \zeta_{U_p pl} + \sum_{i:w^{(b)}=l} \mathcal{I}(z_{ip} = U_p)\right)$$

where $\mathcal{I}(z_{ip} = u) = 1$ if $z_{ip} = u$ and $z_{ip} \in \mathbf{z}^{obs}$ and 0 otherwise;

5. for $l = 1, ..., L$ compute $\pi_{\tilde{\mathbf{r}}l}^{(b)}$ conditioned on $w^{(b)} = l$ after updating the parameter values of each within-class LM model:

   - for $t = 1$, update the initial state probabilities

   $\boldsymbol{\nu}^{(b)} | s_1^{(b)}, w^{(b)} = l, \boldsymbol{\alpha} \sim$
   $Dirichlet\left(\alpha_1 + \sum_{i:w^{(b)}=l} \mathcal{I}_{i1}(s_1^{(b)} = 1), ..., \alpha_K + \sum_{i:w^{(b)}=l} \mathcal{I}_{i1}(s_1^{(b)} = K, w^{(b)} = l)\right)$

   where $\mathcal{I}_{it}(s_t^{(b)} = k) = 1$ if for unit $i$ $s_t^{(b)} = k$ and 0 otherwise;

   - for $q = 1, ..., K$ and $\forall\ t \geq 2$ update the transition probabilities

   $\boldsymbol{\xi}_q^{(b)} | s_{t-1}^{(b)}, s_t^{(b)}, w^{(b)} = l, \boldsymbol{\gamma} \sim$
   $Dirichlet\left(\gamma_1 + \sum_{i,t:w^{(b)}=l, s_{t-1}^{(b)}=q} \mathcal{I}_{it}(s_t^{(b)} = 1), ..., \gamma_K + \sum_{i,t:w^{(b)}=l, s_{t-1}^{(b)}=q} \mathcal{I}_{it}(s_t^{(b)} = K)\right)$;

   - for $k = 1, ..., K, j = 1, ..., J$ and $\forall\ t$ update the conditional response probabilities

   $\boldsymbol{\phi}_{jk}^{(b)} | s_t^{(b)}, w^{(b)} = l, \mathbf{y}^{obs}, \boldsymbol{\delta}_{jk} \sim$
   $Dirichlet\left(\delta_{1jk} + \sum_{i,t:w^{(b)}=l, s_t^{(b)}=k} \mathcal{I}(y_{itj} = 1), ..., \delta_{R_j jk} + \sum_{i,t:w^{(b)}=l, s_t^{(b)}=k} \mathcal{I}(y_{itj} = R_j)\right)$

   where $\mathcal{I}(y_{itj} = r) = 1$ if $y_{itj} = r$ and $y_{itj} \in \mathbf{y}^{obs}$ and 0 otherwise.

## B.2 Multi-move sampling

**Algorithm 2**:

1. For i=1,...,n calculate and store the filtered state probabilities $\Pr(s_t^{(b)}|\boldsymbol{\theta}^{(b-1)}, w^{(b)} = l, \mathbf{y}_{it})$ for $t = 1, ..., T$ (see Algorithm 3);

2. for $i = 1, ..., n$ sample $s_T^{(b)}$ from $\Pr(s_T^{(b)}|\boldsymbol{\theta}^{(b-1)}, w^{(b)} = l, \mathbf{y}_{iT})$;

3. for $t = T-1, ..., 1$ and $i = 1, ..., n$, given the known state $s_{t+1}^{(b)} = k$ sample $s_t^{(b)}$ from $\Pr(s_t^{(b)} = q|s_{t+1}^{(b)} = k, \boldsymbol{\theta}^{(b-1)}, w^{(b)} = l, \mathbf{y}_{it}) =$

$$\frac{\xi_{q,kl}^{(b-1)} \Pr(s_t^{(b)} = q|\boldsymbol{\theta}^{(b-1)}, w^{(b)} = l, \mathbf{y}_{it})}{\sum_q \xi_{q,kl}^{(b-1)} \Pr(s_t^{(b)} = q|\boldsymbol{\theta}^{(b-1)}, w^{(b)} = l, \mathbf{y}_{it})}.$$

## B.3 Filtered State Probabilities

**Algorithm 3**:

1. At t=1, for $i = 1, ..., n, \kappa = 1, ..., K$ compute

$$\Pr(s_1^{(b)} = \kappa|\boldsymbol{\theta}^{(b-1)}, w^{(b)} = l, \mathbf{y}_{i1} = \mathbf{r}) = \frac{\nu_{\kappa l}^{(b-1)} \Phi_{\mathbf{r}\kappa l}^{*(b-1)}}{\sum_c \nu_{cl}^{(b-1)} \Phi_{\mathbf{r}cl}^{*(b-1)}}.$$

Since we are estimating the model only on $\mathbf{y}^{obs}$, we define $\Phi_{\mathbf{r}kl}^{*(b-1)} = \prod_j \phi_{rjkl}^{*(b-1)}$ where

$$\phi_{rjkl}^{*(b-1)} = \begin{cases} \phi_{rjkl}^{(b-1)} & \text{if } y_{itj} = r \text{ and } y_{itj} \in \mathbf{y}^{obs} \\ 1 & \text{otherwise} \end{cases}$$

$\forall\ t, i, j, r.$

2. for $t = 2, ..., T$:

- for $i = 1, ..., n, k = 1, ..., K$ compute
  $\Pr(s_t^{(b)} = k|\boldsymbol{\theta}^{(b-1)}, \mathbf{y}_{i(t-1)}) = \sum_q \xi_{q,kl}^{(b-1)} \Pr(s_{t-1}^{(b)} = q|\boldsymbol{\theta}^{(b-1)}, w^{(b)} = l, \mathbf{y}_{i(t-1)})$;

- for $i = 1, ..., n, k = 1, ..., K$ compute the filtered state probabilities through
  $\Pr(s_t^{(b)} = k|\boldsymbol{\theta}^{(b-1)}, w^{(b)} = l, \mathbf{y}_{it} = \mathbf{r}_t) =$

$$\frac{\Phi_{\mathbf{r}kl}^{*(b-1)} \Pr(s_t^{(b)} = k|\boldsymbol{\theta}^{(b-1)}, w^{(b)} = l, \mathbf{y}_{i(t-1)})}{\Pr(\mathbf{y}_{it} = \mathbf{r}_t|\boldsymbol{\theta}, w^{(b)} = l, \mathbf{y}_{i(t-1)})}$$

  where

  $\Pr(\mathbf{y}_{it} = \mathbf{r}_t|\boldsymbol{\theta}^{(b-1)}, w^{(b)} = l, \mathbf{y}_{i(t-1)}) =$

$$\sum_c \Phi_{\mathbf{r}cl}^{*(b-1)} \Pr(s_t^{(b)} = c|\boldsymbol{\theta}^{(b-1)}, w^{(b)} = l, \mathbf{y}_{i(t-1)}).$$

# References


Rousseau, J., & Mergensen, K. (2011). Asymptotic behaviour of the posterior distribution in overfitted mixture models. *Journal of the Royal Statistical Society Series B (Statistical Metodology), 73(5)*, 689-710.

Vidotto, D., Vermunt, J. K., & van Deun, K. (2018). Bayesian latent class Models for the multiple imputation of categorical data. *Methodology, 14*, 56-68.