Multiple Imputation of longitudinal categorical data through Bayesian mixture latent Markov models

Davide Vidotto Jeroen K. Vermunt Katrijn van Deun

Department of Methodology and Statistics, Tilburg University

April 11, 2018

Abstract

Standard latent class modeling has recently been shown to provide a flexible tool for the multiple imputation (MI) of missing categorical data in cross-sectional studies. This article introduces an analogous tool for longitudinal studies: MI using Bayesian mixture Latent Markov (BMLM) models. Besides retaining the benefits of latent class models, i.e., respecting the (categorical) measurement scale of the variables and preserving possibly complex relationships between variables within a measurement occasion, the Markov dependence structure of the proposed BMLM model allows capturing lagged dependencies between adjacent time points, while the time-constant mixture structure allows capturing dependencies across all time points, as well as retrieving associations between time-varying and time-constant variables. The performance of the BMLM model for MI is evaluated by means of a simulation study and an empirical experiment, in which it is compared with complete case analysis and MICE. Results show good performance of the proposed method in retrieving the parameters of the analysis model. In contrast, competing methods could provide correct estimates only for some aspects of the data.

Keywords: *Bayesian mixture latent Markov models, missing data, longitudinal analysis, multiple imputation.*

1 Introduction

Sociological, psychological and medical research studies are often performed by means of longitudinal designs, and with variables measured on a categorical scale. An example is the LISS (Longitudinal Internet Studies for the Social Sciences) panel study consisting of periodically administered internet surveys by CentERData (Tilburg University, The Netherlands) to a representative sample of the Dutch population, and covering a broad range of topics such as health, religion, work, and the like.

Different from cross-sectional studies, missing data in longitudinal studies may not only concern partial missingness within a single measurement occasion, but may also take the form of complete missing information for certain occasions as a result of *missing visits* (or *complete missingness*) or subjects dropping out from the study.[1] It is well known that the presence of missing data can cause biased or inaccurate inferences, as well as loss of power, if it is not cautiously handled either before or during the actual statistical analysis. Multiple Imputation (MI) is a method developed by Rubin (1987) which allows separating the missing data handling from the substantive analyses of interest, and moreover takes the additional uncertainty resulting from the missing values into account. Assuming that data are *missing at random* (MAR)[2], in MI the missing values in a dataset are replaced with $M > 1$ sets of values sampled from the distribution of the missing data given the observed data, $\Pr(\mathbf{y}^{mis}|\mathbf{y}^{obs})$. In order to be able to do this, we have to build an imputation model. The substantive model of interest is then estimated on each of the M completed datasets, where the M sets of estimates can be pooled through the rules provided by Rubin (1987).

When imputing missing longitudinal data, the imputation model must fulfill several requirements in order to produce valid imputations. In particular, an imputation model for longitudinal analysis should:

1. capture dependencies among variables within measurement occasions;
2. capture overall dependencies between time points resulting from the fact that individuals differ from one another in a systematic way;
3. capture potential stronger relationships between adjacent time points;
4. automatically (i.e., without explicit specification) capture complex relationships in the data, such as higher-order interactions and non-linear associations;
5. respect the measurement scale of the variables (continuous/categorical).

[1] In the first case (missing visits), subjects fail or refuse to provide information for all variables at one or more time occasions. In the second case (drop-out), a subject stops providing information for all variables from a specific time point until the end of the study. Even though this paper generally deals with partial missingness, we will also test the performance of the BMLM model for MI in presence of missing visits by means of a simulation study and an empirical experiment. In the latter, few cases of drop-out are also present in the dataset.

[2] That is, the probability of missingness depends exclusively on the observed data.

In particular, requirement 4 is motivated by the fact that the imputed datasets could be re-used for several types of analyses, in which different aspects of the data need to be taken into account. An imputation model that can automatically describe all the relevant associations of the data provides datasets that can be re-used in different contexts. Conversely, if an imputation model requires explicit specification of interaction terms and other complex relationships, the imputed datasets are likely to be tailored only for some specific analyses, and the imputation step should be re-performed according to the particular problem under investigation. Furthermore, specifying all the complex interactions that might arise in a dataset can be a difficult and tedious task (Vermunt, Van Ginkel, Van der Ark, & Sijtsma, 2008).

While for longitudinal continuous data the joint-modeling approach with the multivariate normal model (Schafer, 1997) and the full conditional specification with the MICE technique (Van Buuren & Oudshoorn, 1999; Van Buuren & Groothuis-Oudshoorn, 2000) have been proposed and evaluated in the literature (Romaniuk, Patton, & Carling, 2014), for categorical data the problem has not yet been settled.

One possible approach is implementing MICE with generalized linear models using a logistic link function after converting the data from long to wide format.[3] In such a way, relationships among the variables at different time points can correctly be captured by MICE and reproduced in the imputations (Allison, 2009; White, Royston, & Wood, 2011). Despite the advantages and the ease of implementation of the method, MICE is not always guaranteed to work. In the first place, notwithstanding its good performances in simulation studies, convergence to the true distribution of the missing data is not ensured, since the method lacks of theoretical and statistical foundation (Vermunt et al., 2008). Second, conversion from long to wide format causes the number of variables to be imputed (and to be used as predictors) to grow linearly with the number of time points T, slowing down computations and requiring regularization techniques if the sample size is small. Lastly, by default MICE only includes linear main effects into the imputation model, necessitating explicit specification of more complex relationships when those are needed in the analysis model, and thus failing to meet requirement 4 above.

An alternative solution for categorical data is represented by mixture or latent class (LC) models (Lazarsfeld, 1950), proposed and shown to provide good results as imputation models by Vermunt et al. (2008). Mixture modeling allows for flexible joint-density estimation of the categorical variables in the dataset, and requires only the specification of the number of LCs K. When K is set large enough, the model can automatically capture the relevant associations of the joint distribution of the variables (McLachlan & Peel, 2000; Vermunt et al., 2008), achieving requirement 4. However, standard LC models

[3] That is, converting the dataset in such a way that the different time points (the single rows of the dataset in the long format) become columns in the wide format. In this way, each row in the wide format corresponds to a single unit of analysis.

are better suited for cross-sectional datasets, because they do not account for the longitudinal architecture of the data, and, accordingly, do not satisfy requirement 3 above.

A natural extension of the LC model to longitudinal categorical data, which in addition accounts for unobserved heterogeneity between units, is represented by the *mixture Latent Markov* (MLM) model (Vermunt, 2010). With the MLM model subjects are clustered at two levels. At the higher level, a time-constant LC variable groups the units with similar time-varying patterns with each other, meeting in this way requirement 2. At the within-subject level, dynamic latent states (LSs; i.e., LCs that can vary over time) are specified for each time point, and -with the first-order Markov assumption- the LS distribution at time t depends only on the LS occupied at time $t-1$. From a MI point of view, the dynamic LSs help accounting for stronger dependencies across adjacent time points, satisfying requirement 3 above. Furthermore, the distribution of the observed variables at a specific time point depends not only on the time-constant LCs but also on the dynamic LSs, allowing to take dependencies within time points into account, thus meeting requirements 1 and 4. Lastly, the model respects the data scale (requirement 5) by assuming Multinomial distributions for all variables in the measurement model. As a further advantage, the MLM model can produce imputations also for time-constant variables with missing values, when present in the dataset at hand.

In this article, we investigate the performance of MLM modeling as a MI tool for missing categorical longitudinal data. The model is implemented under a Bayesian paradigm. The choice of Bayesian modeling in MI is mainly motivated by two arguments: (a) it naturally yields the posterior distribution of the missing data given the observed data; and (b) it automatically takes into account the variability of the imputation model parameter, yielding proper imputations (Schafer & Graham, 2002).

The outline of the paper is as follows. In Section 2, the model is formally introduced, and the model selection issue is addressed. Sections 3 and 4 describe a simulation and an empirical study evaluating the performance of the Bayesian MLM (BMLM) imputation model. The authors provide final remarks in Section 5.

2 The Bayesian mixture Latent Markov Model for Multiple Imputation

Bayesian estimation of the MLM model requires defining the exact data generating model, such as the number of classes for the mixture part and the number of states for the latent Markov chain, as well as the prior distribution of the model parameters. This allows obtaining $\Pr(\boldsymbol{\theta}|\mathbf{y}^{obs})$, the posterior distribution of the unknown model parameters given the observed data $\mathbf{y}^{obs}$. In MI, the M sets of imputations are obtained from the posterior predictive distribution of the missing data, i.e.

$\Pr(\mathbf{y}^{mis}|\mathbf{y}^{obs}) = \int \Pr(\mathbf{y}^{mis}|\boldsymbol{\theta}) \Pr(\boldsymbol{\theta}|\mathbf{y}^{obs}) d\boldsymbol{\theta}$. To achieve this, M parameter values $\boldsymbol{\theta}^{(m)}$ $(m = 1, ..., M)$ are first sampled from $Pr(\boldsymbol{\theta}|\mathbf{y}^{obs})$, and subsequently the imputations are drawn from $Pr(\mathbf{y}^{mis}|\boldsymbol{\theta}^{(m)})$.

2.1 Data generating model and prior distribution

We will assume fixed measurement occasions t $(t = 1, ..., T)$ over all subjects and variables. For the i-th unit $(i = 1, ..., n)$, y_{itj} indicates the value observed for the j-th time-varying categorical variable $(j = 1, ..., J)$ at time t, with $y_{itj} \in \{1, ..., r, ..., R_j\}$ (therefore R_j represents the number of categories for the j-th variable). The J-dimensional vector of observed values for unit i at time t is denoted by $\mathbf{y}_{it} = \mathbf{r}_t$, where $\mathbf{r}$ represents a generic pattern, and $\mathbf{y}_i = \mathbf{r}^*$ is the vector of responses at all time points for unit i.

Often, also time-constant variables (such as the subject's gender) are present in the dataset. When this is the case, z_{ip} is used to denote the value on the p-th $(p = 1, ..., P)$ time-constant variable observed for unit i. Here $z_{ip} \in \{1, ..., u, ..., U_p\}$ and the P-dimensional time-constant pattern observed for i is given by $\mathbf{z}_i = \mathbf{u}$.

The MLM describes the joint distribution of the data $\Pr(\mathbf{z}_i, \mathbf{y}_i)$ by introducing two types of categorical latent variables: a time-constant LC variable w $(w \in \{1, ..., l, ..., L\})$ and a sequence of dynamic LSs $s_1, s_2, ..., s_t, ..., s_T|w = l$ $(s_t \in \{1, ..., k, ..., K\}\ \forall\ t)$. For the first-order Markov assumption, the distribution of the LSs at time t is dependent on the past only through state at time $t-1$, that is $\Pr(s_t|s_{t-1}, ..., s_1, w = l) = \Pr(s_t|s_{t-1}, w = l)$. Furthermore, the model assumes local independence for the distribution of both time-constant and time-varying variables conditioned on the latent variables: $\Pr(\mathbf{y}_{it} = \mathbf{r}_t|s_t = k, w = l) = \prod_j \Pr(y_{itj} = r|s_t = k, w = l)$ and $\Pr(\mathbf{z}_i = \mathbf{u}|w = l) = \prod_p \Pr(z_{ip} = u|w = l)$.

The MLM model is composed of four parts:

- the *latent class probabilities* for the time-constant latent clusters, expressed by $\Pr(w = l) = \omega_l\ \forall\ l$;
- the *latent states probabilities*, which represent the distribution of the LSs at each time point; these are given by:
  - the *initial state probabilities*, which describe the distribution of the latent states at time $t = 1$, and denoted by $\Pr(s_1 = \kappa|w = l) = \nu_{\kappa l}\ \forall\ \kappa, l$;
  - the *transition probabilities*, the probabilities for a unit to switch from state $s_{t-1}|w = l$ to state $s_t|w = l$ $(t = 2, ..., T)$, and indicated with $\Pr(s_t = k|s_{t-1} = q, w = l) = \xi_{q,k(t)l}$;
- the *conditional response probabilities* of the time-constant variables given the LC w, denoted with $\Pr(z_{ip} = u|w = l) = \lambda_{upl}$ for the p-th variable and $\Pr(\mathbf{z}_i = \mathbf{u}|w = l) = \Lambda_{\mathbf{u}l}$ for the whole pattern: under local independence, $\Lambda_{\mathbf{u}l} = \prod_p \lambda_{upl}$;

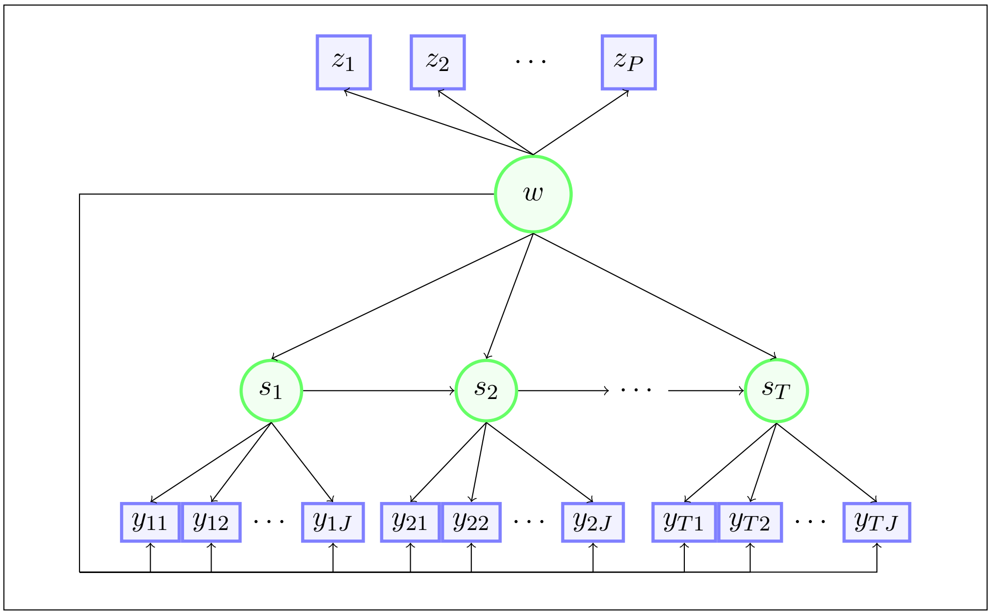

Figure 1: MLM model, graphical representation. w: time-constant latent class variable; z: time-constant variables; s: dynamic latent variable; y: time-varying variables.

- the *emission probabilities*, which define the probability of the time-varying variables conditioned on the LC w and the LS at time t: $\Pr(y_{itj} = r|s_t = k, w = l) = \phi_{rtjkl}$, and -for the local independence- $\Pr(\mathbf{y}_{it} = \mathbf{r}_t|s_t = k, w = l) = \Phi_{\mathbf{r}tkl} = \prod_j \phi_{rtjk}$.

Given the model components above, the MLM model describes the probability of the observed variables as

$$\Pr(\mathbf{z}_i = \mathbf{u}, \mathbf{y}_i = \mathbf{r}^*) = \sum_l \omega_l \Lambda_{\mathbf{u}l} \pi_{\mathbf{r}^*l} \tag{1}$$

where, at the within-subject level,

$$\pi_{\mathbf{r}^*l} = \Pr(y_i = \mathbf{r}^*|w = l) = \sum_{s_1,\ldots,s_T} \nu_{\kappa l} \Phi_{\mathbf{r}1kl} \prod_{t>1} \xi_{q,k(t)l} \Phi_{\mathbf{r}tkl}. \tag{2}$$

Figure 1 represents the path diagram of the data generating model. The picture stresses the double task executed by the subject-level mixture component w: capturing dependencies among the time constant variables and overall dependencies between all time points. Figure 1 also shows how the LS s_t at time t affects the distribution of both s_{t+1} and $\mathbf{y}_{it}$, capturing dependencies between variables within time point t (by means of the emission probabilities) as well as relationships between adjacent time points (by means of the transition probabilities). With such a model configuration, requirement 2 of Section 1 is satisfied with the time-constant latent variable w, while requirements 1 and 3 are met by means of the latent Markov structure assumed upon the time-varying variables.

Importantly, the model can also be implemented in absence of the time-constant variables, which involves dropping the term $\Lambda_{\mathbf{u}l}$ from equation (1) and the nodes representing the time-constant variables $z_{i1}, ..., z_{iP}$ from Figure 1.

The transition probabilities $\xi_{q,k(t)l}$ are stored in T $K \times K$ squared matrices $\boldsymbol{X}_l^t \ \forall \ t \geq 2$. $\boldsymbol{X}_l^t$ is a stochastic matrix, the rows of which must sum to 1: an entry in row q and column k of the matrix represents the probability for a unit to switch from state q at time $t-1$ to state k at time t. The q-th row of $\boldsymbol{X}_l^t$ will be denoted by $\boldsymbol{\xi}_{ql}^t$.

In order to improve class identification, and to reduce the computational burden during the estimation step, we will assume homogeneous transition and emission probabilities across time points: $\xi_{q,k(t)l} = \xi_{q,k(h)l} \ \forall \ t \neq h$ and $t, h \geq 2$ and $\phi_{rtjkl} = \phi_{rhjkl}$, which entails $\Phi_{\mathbf{r}tk} = \Phi_{\mathbf{r}hk} \ \forall \ t \neq h$ and $t, h \geq 1$. Thus, the time-identifier subscript will be dropped from the transition and emission probabilities in the remainder of this article, i.e., $\xi_{q,k(t)l} = \xi_{q,kl}, \boldsymbol{X}_l^t = \boldsymbol{X}_l$ and $\boldsymbol{\xi}_{ql}^t = \boldsymbol{\xi}_{ql} \forall \ t \geq 2$, and $\phi_{rtjk} = \phi_{rjk}, \Phi_{\mathbf{r}tk} = \Phi_{\mathbf{r}k} \ \forall \ t \geq 1$.

For the Bayesian specification of the model, distributional assumptions must be made for all variables and parameters in model (1)-(2). Since all (latent and observed) variables in the model are categorical, a Multinomial distribution will be adopted for each of them. Formally:

- $w \sim Multinomial(\boldsymbol{\omega})$, with $\boldsymbol{\omega}$ the latent weights vector $(\omega_1, ..., \omega_L)$;
- $z_{ip}|w = l \sim Multinomial(\boldsymbol{\lambda}_{pl})$, with $\boldsymbol{\lambda}_{pl} = (\lambda_{1pl}, ..., \lambda_{U_p pl}) \ \forall \ p, l$;
- $s_1|w = l \sim Multinomial(\boldsymbol{\nu}_l)$, where $\boldsymbol{\nu}_l$ is the initial state probabilities vector $(\nu_{1l}, ..., \nu_{Kl}) \ \forall \ l$;
- $s_t|s_{t-1} = q, w = l \sim Multinomial(\boldsymbol{\xi}_{ql}) \ \forall \ t > 1, l$;
- $y_{itj}|s_t = k, w = l \sim Multinomial(\boldsymbol{\phi}_{jkl})$, with $\boldsymbol{\phi}_{jkl}$ the probability vector $(\phi_{1jkl}, ..., \phi_{rjkl}, ..., \phi_{R_j jkl})$ $\forall \ j, k, l$.

We denote by $\boldsymbol{\theta}$ the whole parameter vector, i.e. $\boldsymbol{\theta} = (\boldsymbol{\omega}, \boldsymbol{\lambda}_{11}, ..., \boldsymbol{\lambda}_{PL}, \boldsymbol{\nu}_1, ..., \boldsymbol{\nu}_L, \boldsymbol{X}_1, ..., \boldsymbol{X}_L, \boldsymbol{\phi}_{111}, ..., \boldsymbol{\phi}_{JKL})$. The conjugate of the Multinomial is the Dirichlet distribution. Hence we will set:

- $\boldsymbol{\omega} \sim Dirichlet(\boldsymbol{\eta})$, with $\eta = (\eta_1, ..., \eta_L), \ \eta_l > 0 \ \forall \ l$;
- $\boldsymbol{\lambda}_{pl} \sim Dirichlet(\boldsymbol{\zeta}_{pl})$, with $\boldsymbol{\zeta}_{pl} = (\zeta_{1pl}, ..., \zeta_{U_p pl})$ and $\zeta_{upl} > 0 \ \forall \ u, p, l$.
- $\boldsymbol{\nu}_l \sim Dirichlet(\boldsymbol{\alpha})$, with $\boldsymbol{\alpha} = (\alpha_1, ..., \alpha_K), \alpha_\kappa > 0 \ \forall \ \kappa, l$;
- $\boldsymbol{\xi}_{ql} \sim Dirichlet(\boldsymbol{\gamma})$, with $\boldsymbol{\gamma} = (\gamma_1, ..., \gamma_K), \gamma_k > 0 \ \forall \ k, l$;
- $\boldsymbol{\phi}_{jkl} \sim Dirichlet(\boldsymbol{\delta}_{jk})$, with $\boldsymbol{\delta}_{jk} = (\delta_{1jk}, ..., \delta_{R_j jk}), \delta_{rjk} > 0 \ \forall r, j, k, l$.

$\boldsymbol{\eta}, \boldsymbol{\zeta}_{pl}, \boldsymbol{\alpha}, \boldsymbol{\gamma}$ and $\boldsymbol{\delta}_{jk}$ are called *hyperparameters* of the model. Appendix A gives some guidelines about how to set the priors for MI purposes.

2.2 Model Selection

In MI the imputation model parameters need not be interpreted, and performing imputations with a model that takes into account sample-specific aspects (i.e., a model that overfit the data) is of little concern here (Vermunt et al., 2008). Much more problematic is performing imputations with models that disregard important associations in the data (i.e., models that underfit the data).

Overfitting the data with the BMLM model, and with mixture models in general, means that a number of LCs and LSs (L and K) has been selected for the imputations that is larger than what is needed for the data. When this happens, the BMLM model can carefully capture all relevant associations among the variables as well as sample-specific fluctuations, similar to log-linear imputation models that include non-significant terms (Vermunt et al., 2008). Therefore, to perform imputations a large L and a large K can be chosen. However, it is not always clear whether the selected number of LCs/LSs is large enough; at the same time, too large values might unnecessarily slow down computations, specially with large datasets.

Bayesian modeling offers a simple solution to detect the number of LSs to be used in the imputation model. The method is described by Gelman, Carlin, Stern, and Rubin (2013), chapter 22 for standard mixture models (i.e., for $T = 1$). Their method consists of preliminarily processing the data by estimating a LC model (by means of the Gibbs sampler) with an arbitrarily large number of classes ($K = K^*$) and prior distributions for the latent variable parameter that favor the occurrence of empty components (e.g., with $\alpha_k = 1/K^* \ \forall \ k$) during the iterations of the Gibbs sampler. Counting the number of latent clusters (at each time point) occupied by the units during every iteration leads to a probability distribution for K once the Gibbs sampler is terminated. Gelman et al. (2013), who developed the method for substantive analysis, suggested to use the posterior mode of such distributions to perform inference and obtain interpretable classes. For MI purposes, Vidotto, Vermunt, and van Deun (2018) proposed using the posterior maximum of the resulting posterior distribution.[4] Once K has been chosen, the mixture model can be re-run (with prior distributions set as described in Appendix A) and the imputations can then be performed.

For the BMLM model (case $T > 1$), Gelman et al. (2013)'s method (modified for this kind of model) can be used to determine both L and K (as shown in the simulation study of Section 3 and in the application of Section 4), by setting arbitrarily large values for the number of latent classes and states ($L = L^*$ and $K = K^*$) when running the preliminary Gibbs sampler, and hyperparameters for the latent classes proportions and transition probabilities equal to $\eta_l = 1/L^* \ \forall \ l$ and $\alpha_k = \gamma_k = 1/K^* \ \forall \ k$. The number of clusters to be used for the mixture components can then be chosen to be equal to the posterior maximum of the resulting distribution for L. The number of latent states can be chosen to be the largest

[4] That is, the largest $\bar{K}$ such that $\Pr(K = \bar{K}) > 0$.

among the L posterior maxima observed across time points (chosen, for each l, to be the smallest across t) . That is, we would first consider for each latent cluster $l = 1, ..., L$ the smallest posterior maxima of the number of latent states occupied across the various time points, and subsequently we would choose K as the maximum of the resulting L-dimensional vector. We opt for the smallest posterior maxima across time points, rather than for the largest ones, in order not to incur into the risk of leaving some of the latent states empty during the imputation stage, which could make the Gibbs sampler unstable, as explained in Appendix A.

2.3 Model Estimation and Imputation Step

In presence of the latent variable w and the dynamic states $s_1, ..., s_T$, model estimation occurs through Gibbs sampling with Data Augmentation scheme[5] (Geman & Geman, 1984; Tanner & Wong, 1987).

Appendix B reports the Gibbs sampler (Algorithm 1) used to estimate model (1)-(2). For MI, model estimation is performed only on $\mathbf{z}^{obs}, \mathbf{y}^{obs}$, as in Vermunt et al. (2008). During one iteration, units are first allocated to the time-constant classes according to the *posterior membership probabilities* $\Pr(w|\boldsymbol{\theta}, \mathbf{z}_i, \mathbf{y}_i)$ and then, conditioned on the sampled w, units are assigned to the states of the LM chain at each time point. For each subject, the sequence $s_1, ..., s_T$ is drawn via *multi-move sampling* (Chib, 1996; Fruhwirth-Schnatter, 2006) through their posterior distribution $\Pr(s_1, ..., s_T|w = l, \boldsymbol{\theta}, \mathbf{y}^{obs})$. Multi-move sampling requires to store the *filtered state probabilities* $\Pr(s_t|\mathbf{y}_{it}, \boldsymbol{\theta})$ for each time point. How to perform multi-move sampling and compute the filtered-state probabilities is reported in Algorithms 2 and 3 of Appendix B. After units have been allocated to the LSs, the model parameters are updated using subsequent steps of Algorithm 1.

For each subject with missing values, M values of the LCs w and the LSs s_t (for any t in which the subject provided one or more missing values) should be drawn, along with the conditional distribution probabilities and emission probabilities corresponding to the variables with missing information. These draws must be performed during M of the (post-burn in) Gibbs sampler iterations and should be as spaced from each other as to resemble i.i.d. samples. The sampled values can then be used to perform the imputations: $\forall\ z_{ip} \in \mathbf{z}^{mis}$ and $y_{itj} \in \mathbf{y}^{mis}$,

$$\Pr(z_{ip}^{mis}|w^{(m)} = l) \sim Multinomial(\boldsymbol{\lambda}_{pl}^{(m)})$$

and

$$\Pr(y_{itj}^{mis}|s_t^{(m)} = l, w^{(m)} = l) \sim Multinomial(\boldsymbol{\phi}_{jkl}^{(m)})$$

for $m = 1, ..., M$.

[5] In Data Augmentation units are assigned to the LCs in a first step, and -accordingly- model parameters are updated in the subsequent step. These two main steps are then iterated.

3 Simulation Study

Performance of the BMLM imputation model was assessed by means of a simulation study, and compared with the *complete case* (CC) analysis and MICE techniques. In the study we used four time-varying and four time-constant variables, and we included missing visits (typical of multilevel analysis) to make the parameter retrieval more challenging for the missing data routines. In both studies, analyses were carried out with R version 3.3.0.

3.1 Set-up

Population Model. Four time-constant binary predictors $Z_1, ..., Z_4$ were generated from

$$log\,\Pr(Z_1, Z_2, Z_3, Z_4) \propto 0.5 \sum_p Z_p - \sum_{p=1}^{3} \sum_{p'=p+1}^{4} Z_p Z_{p'} + 2.8 Z_1 Z_2 Z_3 \tag{3}$$

For the time-varying variables, we started by defining the predictors of a potential substantive model at time point $t = 1$. Therefore, we generated $J = 3$ binary variables Y_{11}, Y_{12}, Y_{13} with the log-linear model:

$$log\,\Pr(Y_{11}, Y_{12}, Y_{13}) \propto -0.5 \sum_j Y_{1j} + \sum_{j=1}^{2} \sum_{j'=j+1}^{3} Y_{1j} Y_{1j'} - 0.5 Y_{11} Y_{12} Y_{13}. \tag{4}$$

For $t > 1$, the binary predictors Y_{t1}, Y_{t2} and Y_{t3} were generated through auto-regressive (AR) logistic models

$$\text{logit}\,\Pr(Y_{tj}) = 0.5 Y_{(t-1)j} - 0.15 \sum_{j' \neq j} Y_{(t-1)j'}, \tag{5}$$

for $j = 1, ..., 3$ and $\forall\ t > 1$. In this way we created predictors that are auto-correlated with each other in time. After generating the 3 predictors, we created at each time point the outcome variable Y_{t4} through the AR logistic model

$$logit\,\Pr(Y_{t4}) = \begin{cases} \beta_0 + \beta_1 Y_{t1} + \beta_2 Y_{t2} + \beta_3 Y_{t3} + \beta_{12} Y_{t1} Y_{t2} + \mu_1 Z_1 + \mu_2 Z_2 \\ \qquad + \mu_3 Z_3 + \mu_4 Z_4 & \text{if } t = 1 \\ \beta_0 + \beta_1 Y_{t1} + \beta_2 Y_{t2} + \beta_3 Y_{t3} + \beta_{12} Y_{t1} Y_{t2} + \mu_1 Z_1 + \mu_2 Z_2 \\ \qquad + \mu_3 Z_3 + \mu_4 Z_4 + \rho Y_{(t-1)4} + \tau Y_{(t-1)3} & \text{if } t > 1. \end{cases} \tag{6}$$

Table 1 shows the parameter values chosen for β_0,...,β_{12}, ρ, τ, and $\mu_1, ..., \mu_4$. These parameters were chosen in order to assess how the missing data techniques could capture different aspects of the data:

- $\beta_0, \beta_1, \beta_2, \beta_3, \beta_{12}$ were used to assess how the techniques recovered relationships among variables at the same time point;

Table 1: Values of the parameters in model (6).

Parameter	β_0	β_1	β_2	β_3	β_{12}	μ_1	μ_2	μ_3	μ_4	ρ	τ
Value	-0.8	0.6	-0.9	0.8	-1	0.3	-0.2	0.75	0.6	0.75	0.2

- ρ was used to assess how the models could recover auto-correlations in Y_4 at lag-1;
- τ served to determine whether the models could recover crossed-lagged associations (between Y_3 and Y_4) at lag-1;
- $\mu_1, ..., \mu_4$ served to monitor how the missing data models could retrieve the relationships between the time-varying outcome and the time-constant variables.

From the population model (3)-(4)-(5)-(6), we generated $N = 200$ datasets with $n = 200$ units and $T = 10$ time points.

Generating missingness. MAR missingness was generated in Z_1, Z_2, Y_1 and Y_3. Defining R_p equal to 1 when Z_p was missing and 0 otherwise for $p \in \{1, 2\}$, and R_{tj} equal to 1 when Y_{tj} was missing ($j \in \{1, 3\}$) and 0 when Y_{tj} was observed, missingness was created as follows. For the subject-level variable Z_1,

$$\Pr(R_1 = 1) = \begin{cases} 0.1 & \text{if } Z_3 = 0 \\ 0.3 & \text{if } Z_3 = 1, \end{cases}$$

while for Z_2

$$\Pr(R_2 = 1) = \begin{cases} 0.15 & \text{if } Z_4 = 0 \\ 0.35 & \text{if } Z_4 = 1. \end{cases}$$

As far as the time-varying variables are concerned, the mechanisms were specified as follows. For Y_{t1},

$$\Pr(R_{t1} = 1) = \begin{cases} 0.30 & \text{if } t = 1 \\ 0.35 & \text{if } Y_{(t-1)4} = 0 \text{ and } t > 1 \\ 0.25 & \text{if } Y_{(t-1)4} = 1 \text{ and } t > 1, \end{cases}$$

and for Y_{t3}

$$\Pr(R_{t3} = 1) = \begin{cases} 0.45 & \text{if } Y_{t2} = 0 \\ 0.20 & \text{if } Y_{t2} = 1. \end{cases}$$

While for Y_{t3} missingness was fully MAR and dependent on present values of Y_{t2}, for Y_{t1} the missingness mechanism depended on the time indicator t. In particular, at $t = 1$ missing values were entered according to a MCAR mechanism. For $t > 1$, missingness in Y_{t1} was MAR with a probability depending on the value of $Y_{(t-1)4}$. In such a way, we allowed the missingness mechanism to depend also on past values.

Furthermore, we entered missing visits at each time point by removing for some units simultaneous values of Y_{t1}, Y_{t2}, Y_{t3} and Y_{t4} with probability equal to 0.05 $\forall$ t. These mechanisms yielded about 35% missing observations in Y_1 and Y_3 (across the whole dataset and for each time point), about 20% in Z_1 and Z_2, and about 5% in Y_2 and Y_4.

Missing data methods. After missingness was generated, we implemented three missing data techniques on the dataset. The first one was CC analysis. The second was the BMLM imputation technique presented in this article. For the selection of L and K, we used Gelman et al. (2013)'s method described in Section 2.2. Running a preliminary Gibbs sampler for each dataset led to select an average number of LCs equal to $L = 7.76$ and average number number of LSs equal to $K = 10.54$ (starting with $L^* = 10$ and $K^* = 15$, with 3000 iterations for the Gibbs sampler, of which 1000 used for the burn-in). Appendix A reports how the prior distributions for the BMLM model were set. $B = 3000$ iterations were run for the imputation step, including $I = 1000$ of burn-in. For each dataset, $M = 20$ imputations were performed.

The third missing data technique was the MICE imputation method via logistic regression. For MICE, the datasets were transformed from long to wide format. Notice that, in this case, MICE used an imputation model with $JT = 40$ time-varying variables (plus the 4 time-constant ones). MICE was implemented with its default settings and run for 20 iterations per imputation, with which $M = 20$ imputations were obtained.

Outcomes. Bias, stability (in terms of standard deviation of the produced estimates) and coverage rates of the 95% confidence intervals of the parameters in model (6) were used in order to evaluate the performance of each method.

3.2 Results

Results of the simulation study are shown in Table 2. The BMLM imputation method could, overall, retrieve approximately unbiased parameter estimates not only for the predictors of the time-varying variables, but also for the parameters of the time-constant variables, $\mu_1, ..., \mu_4$. CC analysis retrieved unbiased parameter estimates for the main effects parameters of the time-varying variables (as well as the main effects of the subject-specific variables), but retrieved biased intercept and lagged-relationships. The MICE imputation technique could not pick up the estimates of the main and interaction effects of time-varying variables (specially β_1 and β_{12}), but could recover unbiased lagged relationships (ρ and τ) and parameters of the time-constant effects.

CC analysis produced the most unstable estimates among the three methods. Estimates yielded by

Table 2: Simulation Study: results observed for the estimates of the AR logistic regression coefficients in model (6) for three missing data methods: CC (complete case analysis), BMLM (Bayesian Mixture Latent Markov model) imputation, MICE imputation. Large bias (in absolute value) and too low coverage rates are marked in boldface.

		Missing data method		
	Parameter	CC	BMLM	MICE
Bias	$\beta_0 = -0.80$	**0.36**	0.10	**0.18**
	$\beta_1 = 0.60$	0.01	0.00	**-0.19**
	$\beta_2 = -0.90$	-0.02	0.00	-0.14
	$\beta_3 = 0.80$	0.01	-0.02	-0.10
	$\beta_{12} = -1$	-0.03	0.00	**0.33**
	$\mu_1 = 0.30$	0.03	-0.04	-0.03
	$\mu_2 = -0.20$	-0.05	0.00	0.01
	$\mu_3 = 0.75$	0.09	-0.01	-0.01
	$\mu_4 = 0.60$	0.08	-0.02	-0.01
	$\rho = 0.75$	**-0.22**	-0.05	-0.04
	$\tau = 0.20$	**-0.24**	-0.05	-0.01
Stability	$\beta_0 = -0.80$	0.30	0.18	0.18
	$\beta_1 = 0.60$	0.32	0.19	0.18
	$\beta_2 = -0.90$	0.28	0.16	0.15
	$\beta_3 = 0.80$	0.19	0.13	0.12
	$\beta_{12} = -1$	0.40	0.25	0.23
	$\mu_1 = 0.30$	0.20	0.12	0.12
	$\mu_2 = -0.20$	0.20	0.12	0.12
	$\mu_3 = 0.75$	0.20	0.11	0.11
	$\mu_4 = 0.60$	0.23	0.13	0.13
	$\rho = 0.75$	0.27	0.11	0.11
	$\tau = 0.20$	0.27	0.12	0.12
Coverage Rate	$\beta_0 = -0.80$	**0.76**	0.92	**0.84**
	$\beta_1 = 0.60$	0.96	0.94	**0.84**
	$\beta_2 = -0.90$	0.95	0.96	0.91
	$\beta_3 = 0.80$	0.94	0.94	0.90
	$\beta_{12} = -1$	0.98	0.97	**0.72**
	$\mu_1 = 0.30$	0.93	0.97	0.96
	$\mu_2 = -0.20$	0.98	0.97	0.95
	$\mu_3 = 0.75$	0.94	0.95	0.97
	$\mu_4 = 0.60$	0.92	0.94	0.96
	$\rho = 0.75$	**0.88**	0.94	0.92
	$\tau = 0.20$	**0.82**	0.96	0.94

the BMLM technique and MICE had, overall, similar stability for all types of regression coefficients, although the main and interaction effects of time-varying predictors produced by the BMLM model tended to vary more. The BMLM method yielded confidence intervals that were mostly close to their nominal level. MICE produced confidence intervals for the time-constant and lagged effects with coverage rates rather close to their nominal level, but intervals with too low coverage for main and interaction effects of the time-varying items. The confidence intervals computed after CC analysis were close to their nominal coverage level, excluding the intervals of β_0, ρ and τ, which resulted in a too low coverage.

4 Empirical Study

While in the previous section the parameters of the BMLM MI method was evaluated using simulated datasets from constructed populations, in this section we focus on a real dataset. More specifically, we make use of the associations as present in a real longitudinal dataset rather than specifying these ourselves, and investigate whether these associations are retained when introducing missing values (including missing visits) and imputing these using the BMLM model. For this application we create the missing values in the dataset ourselves, in such a way to have a benchmark (the results obtained with the complete data) for the estimates retrieved by the missing-data methods.

We used data collected by CentERData through their LISS panel, which consists of a (representative) sample of Dutch individuals, who participate in monthly Internet surveys. Key topics surveyed once per year include work, education, income, housing, time use, political views, values, and personality.[6] For our experiment, we selected the first 4 yearly waves ($T = 4$, from June 2008 until June 2011) of the Housing questionnaire.

4.1 Study set-up

The data and the analysis model. The original datasets consisted of about a hundred variables (which included survey-specific and background variables) and sample sizes that varied from wave to wave, ranging from 4411 (Wave 3) to 5018 (Wave 4) cases. We merged the datasets coming from the four surveys, retained only those units with complete information for all four waves, and selected only those cases who were owners of the dwellings where they had residence (this was functional to the analysis model we decided to estimate). This resulted in a dataset with sample size of $n = 257$ (and 1028 rows in total for the four time points).

Next, using this dataset, we estimated a panel regression model with random intercept and autoregressive errors for the outcome variable 'House Satisfaction'[7]; this variable is denoted by Y_{t0} in Table

[6] More information about the LISS panel can be found at `www.lissdata.nl`.

[7] The name of the variable was `cd08a001` in the original dataset.

Table 3: Real-data experiment: variables used in the panel regression model (7) (top part) and to generate missingness (bottom part). Type of variables: TV = time-varying; TC = time-constant. R = respondent.

Variables for the analysis model		
Variable ID	Description	Values (range)
Y_{t0} (TV)	R.'s house satisfaction	1 Very unsatisfied; 4 Very satisfied
Y_{t1} (TV)	R.'s vicinity satisfaction	1 Very unsatisfied; 4 Very satisfied
Y_{t2} (TV)	R.'s opinion about the value of the dwelling	1 Low; 5 High
Y_{t3} (TV)	Type of R.'s dwelling	1 Single family; 7 With shop or workplace
Y_{t4} (TV)	The dwelling has damp walls or floors	0 No; 1 Yes
Y_{t5} (TV)	Number of living-at-home children	$0 = 0$; $3 \geq 3$
Y_{t6} (TV)	Personal net income	0 No income; $7 \geq 3000$ euros
Y_{t7} (TV)	Paid service costs to associations of owners	1 Yes; 2 No
t (TV)	Wave indicator	1 = 1st wave; 4 = 4th wave
Extra variables used to generate missingness		
Variable ID	Description	Values (range)
Z_1 (TC)	R.'s gender	0 Female; 1 Male

3. Among the remaining variables, we detected 7 (time-varying) predictors ($Y_{t1}, ..., Y_{t7}$ in Table 3) that were significant at the 5% level, yielding a total of $J = 8$ variables in the analysis model. Descriptions of these variables, including the time indicator t, are given in Table 3 (top part). Some of these were re-coded (transformed from continuous to categorical) and for others we collapsed some categories (so that their frequencies were not too small).

The panel regression model we estimated was

$$Y_{it0} = \beta_0 + \sum_{j=1}^{6} \beta_j Y_{itj} + \beta_{16} Y_{it1} Y_{it6} + \tau_1 Y_{i(t-1)1} + \tau_7 Y_{i(t-1)7} + u_{i0} + \epsilon_{it} \tag{7}$$

where the random effects u_{i0} were assumed to be normally distributed:

$$u_{i0} \sim N(0, \sigma_1^2).$$

The errors ϵ_{it} were assumed to be the components of a Multivariate Normal, with auto-regressive (AR(1)) covariance structure:

$$\boldsymbol{\epsilon}_i \sim MVN\left(\begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}, \sigma_2^2 \begin{bmatrix} 1 & \rho & \rho^2 & \rho^2 \\ \rho & 1 & \rho & \rho^2 \\ \rho^2 & \rho & 1 & \rho \\ \rho^3 & \rho^2 & \rho & 1 \end{bmatrix}\right).$$

The values of the model parameters $\beta_0, ...\beta_6, \beta_{16}, \tau_1, \tau_7, \sigma_1^2, \sigma_2^2, \rho$ estimated on the complete data are reported in the first columns of Table 4 below, along with their standard errors. All predictor effects were significant at 5% level as highlighted, except for Y_{t6}, one of the variables yielding the significant interaction term β_{16}.

Generating missingness. Apart from the variables $Y_{t0}, ..., Y_{t7}$, we used the time-constant variable gender denoted with Z_1 in Table 3, to generate MAR missingness in the variable Y_{t1} (Z_1 was thus also included in the imputation models as a time-constant variable). In particular, by denoting the missingness of Y_{t1} with R_{t1}, we created missing values for Y_{t1} with the logistic model

$$\operatorname{logit} \Pr(R_{t1} = 1) = -3 + 1.9Z_1.$$

Furthermore, we entered MAR missingness in Y_{t2} - conditioned on Y_{t3} - with the logistic model

$$\operatorname{logit} \Pr(R_{t2} = 1) = 2.5 - 1.6Y_{t3},$$

where R_{t2} is defined in a way similar to R_{t1}. The parameters of both logistic models were chosen in such a way to obtain marginal missingness rates of about 20% for each of these two variabes.

Furthermore, we generated missing visits in the dataset; thus, for some units, we removed the observations for all the time-varying variables $Y_{t0}, ..., Y_{t7}$ with increasing probability at each time point. If $R_{MV(t)}$ is the indicator equal to 1 for those units with missing visits at time t and equal to 0 otherwise, the mechanism we used was

$$\operatorname{logit} \Pr(R_{MV(t)} = 1) = -4.5 + 0.55t,$$

which generated missing visits for about 1% of the cases at the first wave, and for about 9% of the cases at the fourth wave.

Overall, all the time-varying variables had a marginal (i.e., across all time points) rate of missingness equal to about 5%, except for Y_{t1} and Y_{t2}, which had a marginal rate of missingness roughly equal to 25%.

Missing data methods. As done for Section 3, we compared the performance of three missing data methods to retrieve the parameters of model 7: CC analysis, BMLM MI and MICE.

With CC analysis we estimated model 7 on the dataset with only complete observations, i.e., excluding all cases with missing data. This left a dataset with 591 rows, with sample sizes ranging from $n = 129$ at wave four to $n = 171$ at wave one.

For the BMLM model, we performed model selection with Gelman et al. (2013) 's method reported in Section 2.2. We ran the preliminary Gibbs sampler with $L^* = 20$ and $K^* = 20$, and the same number of iterations as the previous case. This led us to choose $L = 18$ and $K = 9$. In the subsequent step, $M = 50$ imputations were performed during 50000 iterations (plus 10000 iterations for the burn-in).

Lastly, MICE was implemented with its default settings, and its algorithm was run for 50 iterations for each of the $M = 50$ produced imputations.

Table 4: Real-data experiment: results for the parameters in model 7. Est. = point estimate. S.E. = standard error. 5% significant predictors are denoted with a '*' next to the point estimates obtained with each method.

			Missing Data Method					
Parameter	Complete Data		CC analysis		BMLM		MICE	
	Est.	S.E.	Est.	S.E.	Est.	S.E.	Est.	S.E.
β_0	0.86*	0.23	1.03*	0.30	0.99*	0.29	1.04*	0.27
β_1	0.73*	0.08	0.67*	0.11	0.66*	0.10	0.65*	0.11
β_2	0.12*	0.02	0.12*	0.03	0.09*	0.03	0.10*	0.03
β_3	-0.05*	0.02	-0.06	0.03	-0.06*	0.03	-0.06*	0.03
β_4	-0.52*	0.16	-0.49*	0.22	-0.48*	0.20	-0.40*	0.19
β_5	-0.09*	0.03	-0.12*	0.04	-0.08*	0.04	-0.08	0.04
β_6	0.07	0.04	0.03	0.06	0.11	0.05	0.07	0.05
β_{16}	-0.05*	0.02	-0.03	0.02	-0.05*	0.02	-0.05*	0.02
τ_1	0.11*	0.02	0.12*	0.03	0.11*	0.03	0.12*	0.03
τ_7	-0.10*	0.03	-0.11*	0.05	-0.09*	0.04	-0.12*	0.04
σ_1^2	0.19	-	0.20	-	0.21	-	0.21	-
σ_2^2	0.25	-	0.26	-	0.28	-	0.30	-
ρ	0.13	-	0.07	-	0.11	-	0.10	-

Outcomes. We compared the results provided by each missing data method with the results observed for the complete-data case. In particular, we focused on the point estimates of all parameters in model 7 as well as the standard errors for the fixed effects ($\beta_0, ..., \tau_7$). We also examined which fixed effect estimates were significant at a 5% level.

4.2 Results

The results are reported in Table 4. Both CC analysis and the two versions of the BMLM imputation model retrieved point estimates of the fixed effects rather close to those of the complete-data analysis. Exceptions for the CC analysis were the main effects β_1 and β_6 and the interaction term β_{16}, which were slightly different from the corresponding values obtained with the complete data. Some of the standard errors yielded by CC analysis were inflated because of the limited sample size exploited by this method, which made some parameter estimates no longer significant at the 5% level (in Table 4, some fixed effects are no longer marked with a '*'). Conversely, despite a couple of values being slightly off (the intercept β_0 and the main effect β_1), BMLM could exploit the original sample size, causing the standard errors to be only slightly larger than those of complete-data analysis (reflecting in this way the imputation step uncertainty). As a result, all parameters that were significant with the full data were also significant after imputing the missing values with the BMLM model. The MICE method did not manage to recover well all parameter estimates; for instance, the intercept β_0 and the main effects β_1 and β_4 were (in a more or less pronounced manner) far from the estimates of the complete-data condition, while the standard errors observed after imputing the data with MICE were close to the BMLM MI estimates. Nevertheless, the parameter β_5 which was significant with the complete data and the BMLM imputation method, was no longer significant with the MICE.

Concerning the parameters of the random part of the models, all missing data techniques could retrieve good estimates for the variances of the random effect σ_1^2, as well as the variance for residuals σ_2^2, although the latter was slightly overestimated by all imputation methods. The auto-regressive coefficient ρ, on the other hand, was well retrieved by all MI techniques, and considerably underestimated by CC analysis.

5 Discussion

We introduced the use of the BMLM model for the MI of missing categorical longitudinal data. With a limited amount of model specification (only the number of time-constant clusters L and the number of dynamic states K), the model is flexible enough to automatically recover relationships arising between time-varying and time-constant variables, as well as lagged relationships and auto-correlations. Lastly, the model reflects the correct (categorical) scale with which the variables are measured.

The performance of BMLM-based MI approach was evaluated and compared with other two missing data methods, CC analysis and MICE, by means of a simulations studies and an empirical experiment. In the simulation study, the analysis model used was a logistic model including an auto-regression term and a crossed-lagged relationship coefficient, as well as main effects of time-constant predictors. Results showed a good (overall) performance of the BMLM imputation model compared with the competing methods, since it could retrieve (approximately) unbiased estimates for all types of parameters specified in the substantive models, with coverage rates of the confidence intervals that were never too small compared to their 95% nominal level. The good performance of the BMLM model showed that the model can also cope with missing visits when these are present at any time point. Conversely, CC analysis could not recover well the lagged relationships in terms of both bias and confidence intervals, with coverage rates that were too low for their nominal level, while MICE provided biased time-varying main and interaction effects, with corresponding confidence intervals that tended to be too narrow.

In the empirical experiment we estimated a panel regression model using data from the LISS panel. The model included main and interaction fixed effects, along with crossed-lagged relationships and random intercept. Furthermore, the distribution of the residuals was described by a variance and an auto-correlation coefficient. We also included cases with missing visits in the LISS dataset as a further challenge for the missing data methods. Results demonstrated the superiority of the BMLM model when compared to competing methods; in particular, the same conclusions (i.e., the same terms were statistically significant) were drawn for the complete-data case and the BMLM imputation method. This did not happen with the CC and the MICE techniques, for which some terms were not significant anymore. In addition, the BMLM method retrieved variance and error components close to the complete-data analysis.

In the light of the results of the studies carried out in this article, we recommend the applied researcher that needs to deal with missing longitudinal categorical data to consider the BMLM model as a possible MI tool. However, some issues still need to be better analyzed in future research. For instance, whereas in this article we aimed to introduce the use of the BMLM model for MI purposes, some more extensive simulation experiments (in which the model is tested with different sample size and missingness conditions, such as systematic drop-out) should be performed in future studies. In addition, while we showed that our model can deal with MAR missing data, a version of the BMLM model for *missing not at random* data (MNAR; i.e., the distribution of the missingness depends on the unobserved data), which are likely to occur in longitudinal analysis, should be developed in future research.

Furthermore, the proposed imputation model itself can be extended in various useful ways. Firstly, while we dealt with categorical (both ordinal and nominal) variables, the BMLM model can be extended to accommodate mixed types of data, i.e., it can be implemented on datasets containing both categorical and continuous variables. This can be achieved, for instance, by specifying mixtures of univariate Normal and Multinomial distributions. Second, although we assumed the BMLM model to have a Markov chain of order 1, it is possible to consider lags of higher orders by conditioning the distribution of the dynamic LSs at time t on the configuration of the states at earlier time points, e.g. $t-2$, $t-3$, etc., if these kinds of lags are needed in the substantive analysis. Third, when the measurement may occur at different continuous time points rather than at fixed discrete occasions, imputations of the missing data can be provided by assuming a continuous-time latent Markov chain for the distribution of the LSs. Last, for applications in which the subjects observed across time are coming from different groups (e.g., patients coming from different hospitals), the model can be moved towards a multilevel framework, for instance, by adding a further LC variable at the group-level.

Declaration of Conflicting Interests. The authors declare no potential conflicts of interest about the publication of this paper.

Acknowledgements. The research was supported by the Netherlands Organisation for Scientific Research (NWO), grant project number 406-13-048.

References

Allison, P. D. (2009). Missing data. In R. E. Millsap & A. Maydeu-Olivares (Eds.), *The SAGE Handbook of Quantitative Methods in Psychology*, pp. 72-89. Thousand Oaks, CA: Sage.

Chib, S. (1996). Calculating posterior distributions and modal estimates in Markov mixture models. *Journal of Econometrics, 75(1)*, 79-97.

Fruhwirth-Schnatter, S. (2006). *Finite Mixture and Markov Switching Models* (First ed.). New York: Springer-Verlag.

Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (2013). *Bayesian Data Analysis* (Third ed.). London: Chapman and Hall.

Geman, S., & Geman, D. (1984). Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transacations on Pattern Analysis and Machine Intelligence, 6(6)*, 721-741.

Lazarsfeld, P. F. (1950). The logical and mathematical foundation of latent structure analysis. In S. A. Stouffer, L. Guttman, E. A. Suchman, P. F. Lazarsfeld, S. A. Star & J.A. Clausen (Eds.), *Measurement and prediction*, pp. 361-412. Princeton: Princeton University Press.

McLachlan, G. J., & Peel, D. (2000). *Finite mixture models.* New York: Wiley.

Romaniuk, H., Patton, G., & Carling, J. (2014). Multiple Imputation in a Longitudinal Cohort Study: A Case Study of Sensitivity to Imputation Methods. *Am Journal of Epidemiology, 180(9)*, 920-932.

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.

Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys.* New York: Wiley.

Schafer, J. L. (1997). *Analysis of incomplete multivariate data.* London: Chapman & Hall.

Schafer, J. L., & Graham, J. W. (2002). Missing data: Our view of the state of the art. *Psychological methods, 7(2)*, 147-177.

Tanner, A. M., & Wong, W. H. (1987). The calculation of posterior distributions by Data Augmentation. *Journal of the American Statistical Association, 82(398)*, 528-540.

Van Buuren, S., & Groothuis-Oudshoorn, K. (2000). *Multivariate imputation by Chained equations: MICE V.1.0 User's manual.* Leiden, The Netherlands: Toegepast Natuurwetenschappelijk Onderzoek (TNO) Report PG/VGZ/00.038.

Van Buuren, S., & Oudshoorn, C. (1999). *Flexible multivariate imputation by MICE* (Tech. rep. TNO/VGZ/PG 99.054). Leiden: TNO Preventie en Gezondheid.

Vermunt, J. K. (2010). Longitudinal research using mixture models. In *Longitudinal research with latent variables*, Montfort, V.K., Oud, J. and Satorra, A., Eds., Springer, Verlag, Berlin and Heidelberg, 2010, pp. 119-152.

Vermunt, J. K., Van Ginkel, J. R., Van der Ark, L. A., & Sijtsma, K. (2008). Multiple imputation of incomplete categorical data using latent class analysis. *Sociological Methodology, 38(1)*, 369-397.

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

White, I. R., Royston, P., & Wood, A. M. (2011). Multiple imputation using chained equations: issues and guidance for practice. *Statistics in Medicine, 30(4)*, 377-399.

Appendix 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 a 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 α_k $\forall$ k should be set at least equal to $\sum_j (R_j - 1)/2$. Following the guidelines given in Vidotto et al. (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 α_k and γ_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.

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 η_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.

Appendix 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

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

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

$$\boldsymbol{\phi}_{jk}^{(b)} | s_t^{(b)}, w^{(b)} = l, \mathbf{y}^{obs}, \boldsymbol{\delta}_{jk} \sim$$

$$\textit{Dirichlet}\left(\delta_{1jk} + \sum_{i,t:w^{(b)}=l,s_t^{(b)}=k} \mathcal{I}(y_{itj} = 1), ..., \delta_{R_j jk} + \sum_{i,t:w^{(b)}=l,s_t^{(b)}=k} \mathcal{I}(y_{itj} = R_j)\right)$$

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 $\Phi^{*(b-1)}_{\mathbf{r}kl} = \prod_j \phi^{*(b-1)}_{rjkl}$ where

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

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

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

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

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

where

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

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