Bayesian Multilevel Latent Class Models for the Multiple Imputation of Nested Categorical Data

Davide Vidotto Jeroen K. Vermunt Katrijn van Deun

Department of Methodology and Statistics, Tilburg University

Abstract

With this paper, we propose using a Bayesian multilevel mixture model for the multiple imputation of nested categorical data. While the literature and standard software show a lack of imputation models in this context, with the current paper we intend to fill this gap by proposing a model that is able to retrieve original associations in the data at both the first and second level of the hierarchy, as well as to respect the original scale of the data. After formally introducing the model and showing how it can be implemented, we carry out a simulation study and a real-data study in order to assess its performance, and compare it with the commonly used listwise deletion and an already available R-routine. Results indicate that the Bayesian Multilevel Latent Class model is able to recover unbiased and efficient parameter estimates of the analysis models considered in our studies, outperforming in this way the competing methods.

Keywords: *Bayesian mixture models, latent class models, missing data, multilevel analysis, multiple imputation.*

1 Introduction

Nested or multilevel data are typical in educational, social, and medical sciences. In this context, level-1 (or lower-level) units, such as students, citizens, patients, are nested within level-2 (or higher-level) units, such as schools, cities, hospitals. When lower-level units within the same group are correlated with each other, the nested structure of the data must be taken into account. While standard single-level analysis assumes independent level-1 observations, multilevel modeling allows to take these dependencies into

account. In addition, variables can be collected and observed at both levels of the dataset, which is another feature not taken into account by single-level analyses.

Akin to single-level analysis, however, the problem of missing data arises and must be properly handled also with multilevel data. While multilevel modeling has in general gained a lot of attention in the last decades, issues related to item nonresponses in this context are still open (Van Buuren, 2011). In this respect, Van Buuren (2011) observed that the most common practice followed by analysts is discarding all the units with nonresponses and performing the analysis with the remaining data, a technique known as *listwise deletion* (LD). While LD can potentially lead to a large waste of data (for instance, with a missing item for a level-2 unit, all the level-1 units belonging to that group are automatically removed), it also introduces bias in the estimates of the analysis model when the missingness is in the predictors. Another missing-data handling technique, *maximum likelihood for incomplete data*, which is considered one of the major methods for missing data in single-level analysis (Allison, 2009; Schafer & Graham, 2002) under the *missing at random* (MAR) assumption[1], has certain drawbacks with multilevel data (Allison, 2009; Van Buuren, 2011). First, the variables that rule the missingness mechanism must be included in the analysis model. As a consequence, specifying and interpreting the joint distribution of such data can become a complex task in this case. Furthermore, departures from the true model can lead to biased estimates, or incorrect standard errors (Van Buuren, 2011). Second, with multilevel models the derivation of the maximum likelihood estimates, for instance through EM algorithm or numerical integration, can be computationally troublesome (Goldstein, Carpenter, Kenward, & Levin, 2009).

A more flexible tool present in the literature is *multiple imputation* (MI; Rubin, 1987). MI can replace the original incomplete dataset with $m > 1$ completed datasets, in which the original missing values have been replaced by means of an imputation model. In this context, the main aim of the imputation model is preserving the original relationships present among the variables (reflected in the imputed data), while the imputation model parameters are not of primary interest. The final goal of the imputation model is drawing imputed values from the posterior distribution of the missing data given the observed data. After this step, standard full-data analysis can be performed on each of the m completed datasets. By doing this, uncertainty coming from the sampling stage can be distinguished from uncertainty due to the imputation step in the pooled estimates and their standard errors. One of the major advantages of MI is that, after the imputation stage, any kind of analysis can be performed on the completed data (Allison, 2009).

Specification of the imputation model is one of the most delicate steps in MI. Two main imputation modeling techniques are present in the literature: full conditional specification (Van Buuren, Brand, Groothuis-Oudshoorn, & Rubin, 2006) and joint modeling (Schafer, 1997). While the former is based on

[1]That is, the distribution of the missing data depends exclusively on other observed data, and not on the missing data itself.

a variable-by-variable imputation, and requires specification of separate conditional models for each item with missing observations, the latter only needs specification of a joint multivariate model of the items in the dataset, from which the imputations are drawn. As a general rule, the imputation model should be at least as complex as the substantive model, in order not to miss important relationships between the variables and the observations that are object of study in the final analysis (Schafer & Graham, 2002).

In a multilevel context, this means that also the sampling design must be taken into account. A number of studies has shown the effect of ignoring the double-level structure of the data when imputing with standard single-level models (Van Buuren, 2011; Carpenter & Kenward, 2013; Drechsler, 2015; Andridge, 2011; Reiter, Raghunathan, & Kinney, 2006). Results indicate that including design effects in the imputation model - when they are not actually needed - can lead in the worst case to a loss of efficiency and conservative inferences, while using single-level imputation models when design effects are present in the data can be detrimental for final inferences. The latter case can result in biased final estimates, as well as in severe under-estimation of the between-groups variation and biased standard errors of the fixed effects (Carpenter & Kenward, 2013). To take the nested structure of the data into account, mixed effects models are better equipped than fixed effects imputation models with dummy variables, since the latter can overestimate the between-groups variance (Andridge, 2011). Furthermore, single-level imputation can yield different values for level-2 variables within the same group, if these are included in the model. Conversely, multilevel modeling automatically incorporates the nested structure of the data, takes into account level-1 units correlations within the same level-2 unit, and imputes the data respecting the exact level of the hierarchy in which the imputations is to be performed.

Survey data often record categorical item responses. While multilevel MI for continuous data has already been discussed in the literature (Schafer & Yucel, 2002; Yucel, 2008; Van Buuren, 2011), to our knowledge no ad-hoc methods have been proposed in the literature for categorical data, and require better coverage (Van Buuren, 2012). Most of the standard software focuses on single-level imputation models (see Andridge, 2011 for a review of software packages wrongly suggested for multilevel studies), or does not allow for the MI of multilevel categorical data, such as the `mice` package (Zhao & Schafer, 2016; Van Buuren & Groothuis-Oudshoorn, 2000), which bases its imputations on full conditional specification modeling. An MI technique based on multilevel joint modeling can be found in the `pan` R-library (Zhao & Schafer, 2016). However, `pan` is also not suited for categorical data, because it does not work with the original scale type and treats all the variables as continuous. The imputed data are then imputed through rounding, which can introduce bias in the MI estimates (Horton, Lipsitz, & Parzen, 2003). An R package that allows for the MI of multilevel mixed type of data (categorical and continuous) is the `jomo` package (Quartagno & Carpenter, 2016), another joint modeling (JOMO) approach. For each categorical variable with missingness, JOMO assumes an underlying latent q-variate normal distribution,

where $q+1$ is the number of categories of each variable, at both levels. The joint distribution of these variables are then estimated, and the imputations are based on the components' scores. For more information about the functioning of JOMO, see Carpenter and Kenward (2013). JOMO works under a Bayesian paradigm and uses the Gibbs sampler (Gelfand & Smith, 1990) to perform imputations. This approach, while representing a further step in the literature, has also limitations. First, similar to the `pan` package, JOMO does not work with the original scale type of the data. Second, by working with multivariate normals, imputations yielded by JOMO can correctly reflect only pairwise relationships in the data, ignoring possible higher-level orders of associations, limiting in this way the flexibility of the method. Lastly, the computational time required by JOMO increases as a function of the number of variables, as well as with the number of their categories, since the number of assumed multivariate normal distributions grows quickly with these two quantities.

In this paper, we propose using the *Bayesian Multilevel Latent Class* (or *mixture*) *model* (BMLC) for the MI of multilevel categorical data. In a single-level context, the standard Latent Class (LC) imputation model was firstly introduced by Vermunt, Van Ginkel, Van der Ark, and Sijtsma (2008) in a frequentist framework and then by Si and Reiter (2013) under a nonparametric Bayesian perspective. The Bayesian setting allows for an easier and more appealing computation in presence of multilevel data (Goldstein et al., 2009; Yucel, 2008) through MCMC algorithms, and is viewed as a natural choice in a MI context (Schafer & Graham, 2002), since the posterior distribution of the missing data given the observed data can be directly specified as a part of the model. The attractive part of the BMLC model is its flexibility, since it can pick up very complex associations in the data at both levels when a large enough number of latent classes (or mixture components) is specified. Furthermore, the model works with the original scale type of the data, preventing rounding bias.

The approach we propose is based on the non-parametric version of the multilevel LC model introduced by Vermunt (2003) in a frequentist setting. Unlike the single-level LC model, the BMLC is able to capture heterogeneity in the data at both levels of the dataset, by clustering the level-2 units into level-2 LCs and, conditioned on these clusters, level-1 units are classified into level-1 LCs. With this setting, units at level-1 within the same level-2 group are assumed independent from each other only when conditioned on the level-2 LC. The BMLC model proposed for the imputation extends the work of Vermunt (2003) to include also level-2 indicators, allowing for correct imputations at both levels of the dataset. In the paper, we will address issues related to the selection of model and priors, as well as the model estimation and imputation steps. A simulation study with different sample size conditions will be presented in which the BMLC imputation is compared with the LD and JOMO methods. Lastly, a real-data application will show the behavior of the BMLC imputation model when executed with a real world dataset.

The outline of the paper is as follows. In Section 2, the BMLC model is introduced, along with model and prior selection and model estimation issues. In Section 3, a simulation study is performed with two different sample size conditions. Section 4 shows an application to a real-data situation. Finally, Section 5 concludes with final remarks by the authors.

2 The Bayesian Multilevel Latent Class Model for Multiple Imputation

In MI, imputations are drawn from the distribution of the missing data conditioned on the observed data. With Bayesian imputations, this is the posterior predictive distribution of the missing data given the observed data and the model parameter π, that is $\Pr(D^{mis}|D^{obs},\pi)$, which can be derived from the posterior of the model parameter given the observed data, $\Pr(\pi|D^{obs})$. This allows for modeling uncertainty about π. Since $\Pr(\pi|D^{obs}) \propto \Pr(\pi)\Pr(D^{obs}|\pi)$, we need to specify a data model - $\Pr(D^{obs}|\pi)$ - and a prior distribution - $\Pr(\pi)$ - in order to obtain the posterior of π. Model estimation, as well as the imputation step, is performed through Gibbs sampling.

2.1 The Data Model

Let $D = (\mathbf{Z}, \mathbf{Y})$ denote a nested dataset with J level-2 units and n_j level-1 units within level-2 unit j ($j = 1, ..., J$), with a total sample size of $n = \sum_j n_j$. Suppose, furthermore, that the dataset contains T level-2 categorical items $Z_1, ..., Z_t, ..., Z_T$, each with R_t observed categories ($t = 1, ..., T$) and S level-1 categorical items $Y_1, ..., Y_S$, each with U_s ($s = 1, ..., S$) observed categories.

We denote with $\mathbf{z}_j = (z_{j1}, ..., z_{jT})$ the vector of the T level-2 item scores for level-2 unit j, and with $\mathbf{y}_j = (\mathbf{y}_{j1}, ..., \mathbf{y}_{ji}, ...\mathbf{y}_{jn_j})$ the full vector of the level-1 observations within the level-2 unit j, in which $\mathbf{y}_{ji} = (y_{ji1}, ..., y_{jiS})$ is the vector of the S level-1 item scores for level-1 unit i within the level-2 unit j. The data model consists of two parts, a part for the level-2 (or higher-level) units and a part for the level-1 (or lower-level) units. Let us introduce the level-2 LCs variable W_j with L classes (W_j can take on values $1, ..., l, ..., L$), and the level-1 LCs variables $X_{ji}|W_j$ - with K classes - within the l-th level-2 LC (with X_{ji} ranging in $1, ..., k, ..., K$).

The higher-level data model for unit j can then be expressed by

$$\Pr(\mathbf{Z}_j = \mathbf{z}_j, \mathbf{Y}_j = \mathbf{y}_j) = \sum_{l=1}^{L} \Pr(W_j = l) \prod_{t=1}^{T} \Pr(Z_{jt} = z_{jt}|W_j = l) \prod_{i=1}^{n_j} \Pr(\mathbf{Y}_{ji} = \mathbf{y}_{ji}|W_j = l).$$

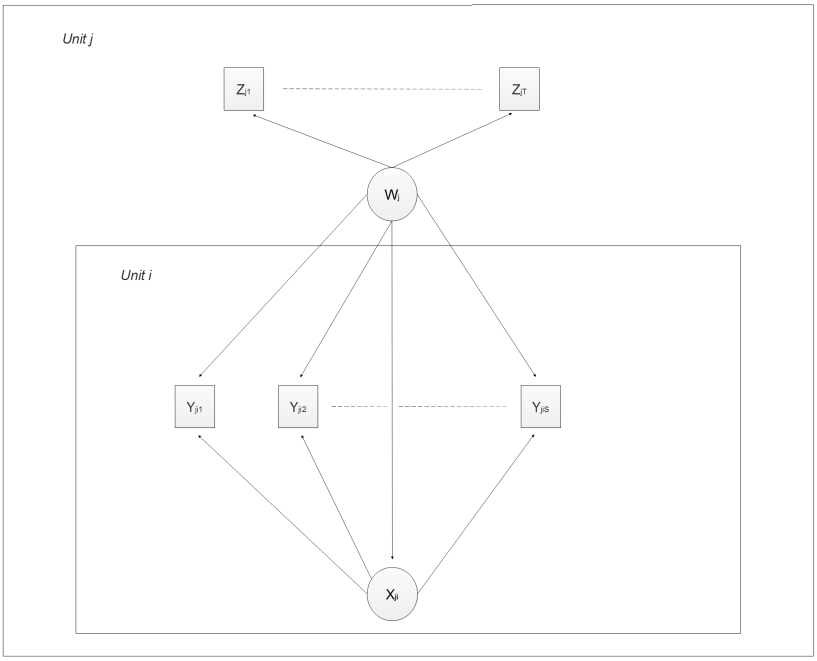


Figure 1: Graphical representation of the multilevel LC model with observed items at both levels of the hierarchy.

This model is linked to the lower-level data model for the level-1 unit i within the level-2 unit j through

$$\Pr(\mathbf{Y}_{ji} = \mathbf{y}_{ji}|W_j = l) = \sum_{k=1}^{K} \Pr(X_{ji} = k|W_j = l) \prod_{s=1}^{S} \Pr(Y_{jis} = y_{jis}|W_j = l, X_{ji} = k).$$

Figure 1 represents the underlying graphical model. From the figure, it is possible to notice both how the number of level-1 latent variables is allowed to vary with j (because within each level-2 unit we have n_j level-1 units and, accordingly, n_j latent variables X_{ji}) and how W_j affects Z_j, X_{ji} and Y_{ji} simultaneously.

As in a standard LC analysis, we will assume Multinomial distributions for the level-1 LCs variable $X|W$ and the conditional response distributions $\Pr(Y_s|W, X)$. Additionally, we will assume Multinomial distributions also for the conditional responses at the higher level $\Pr(Z_t|W)$ and, as we are considering the non-parametric[2] version of the multilevel LC model, also the level-2 mixture variable W is assumed to follow a Multinomial distribution. Formally,

$W \sim \text{Multinom}(\pi_W)$

$X|W = l \sim \text{Multinom}(\pi_{lX})$ for $l = 1, ..., L$

$Z_t|W = l \sim \text{Multinom}(\pi_{lt})$ for $t = 1, ..., T,\ l = 1, ..., L$

$Y_s|W = l, X = k \sim \text{Multinom}(\pi_{lks})$ for $s = 1, ..., S,\ l = 1, ..., L,\ k = 1, ..., K.$

[2]Vermunt (2003) observed that, since we are assuming a Multinomial distribution for the random effects, the term 'non-parametric' does not mean distribution-free in this context. Rather, Vermunt (2003) opposed the non-parametric version to the parametric one, which assumes normality of the random effect.

The parameters denote a vector containing the probabilities of each category of the corresponding Multinomial distribution. That is, $\pi_W = (\pi_1, ..., \pi_l, ..., \pi_L), \pi_{lX} = (\pi_{l1}, ..., \pi_{lk}, ..., \pi_{lK}), \pi_{lt} = (\pi_{lt1}, ..., \pi_{ltr}, ...\pi_{ltR_t})$, $\pi_{lks} = (\pi_{lks1}, ..., \pi_{lksu}, ..., \pi_{lksU_s})$. The whole parameter vector is $\pi = (\pi_W, \pi_{lX}, \pi_{lt}, \pi_{lks})$.

Assuming Multinomiality for all the (latent and observed) items of the model, we can rewrite the model for $\Pr(\mathbf{z}_j, \mathbf{y}_j)$ as

$$\Pr(\mathbf{Z}_j = \mathbf{z}_j, \mathbf{Y}_j = \mathbf{y}_j; \pi) = \sum_{l=1}^{L} \pi_l \prod_{t=1}^{T} \prod_{r=1}^{R_t} (\pi_{ltr})^{\mathcal{I}_{jt}^r} \prod_{i=1}^{n_j} \pi_{jil}, \tag{1}$$

in which $\mathcal{I}_{jt}^r = 1$ if $z_{jt} = r$ and 0 otherwise, and $\pi_{jil} = \Pr(\mathbf{Y}_{ji} = \mathbf{y}_{ji} | W_j = l)$. The latter quantity is derived from the lower-level data model, given by

$$\pi_{jil} = \sum_{k=1}^{K} \pi_{lk} \prod_{s=1}^{S} \prod_{u=1}^{U_s} (\pi_{lksu})^{\mathcal{I}_{jis}^u} \tag{2}$$

where $\mathcal{I}_{jis}^u = 1$ if $y_{jis} = u$ and 0 otherwise.

The model is capable of capturing between- and within-level-2 units variability, by first classifying the J groups in one of the L clusters of the mixture variable W and subsequently, given a latent level of W, classifying the level-1 units within j in one of the K clusters of the mixture variable $X|W$. In order to capture heterogeneity at both levels, the model makes two important assumptions:

- the *local independence* assumption, according to which items at level-2 are independent from each other within each LC W_j and items at level-1 are independent from each other given the level-2 LC W_j and the level-1 LC $X_{ji}|W_j$;
- the *conditional independence* assumption, where level-1 observations within the level-2 unit j are independent from each other once conditioned on the level-2 LC W_j.

According to these assumptions, the mixture variable W is able to pick up both dependencies between the level-2 variables and dependencies among the level-1 units belonging to level-2 unit j, while the mixture variable X is able to capture dependencies among the level-1 items. Both equations (1) and (2) incorporate these assumptions through their product terms.

It is also noteworthy that, by excluding the last product (over i) in equation (1) we obtain the standard LC model for the level-2 units, while, by excluding the product over t in equation 1 and setting $L = 1$, we obtain the standard LC model for the level-1 units.

In Bayesian MI, the quantity $\Pr(\mathbf{Z}_j, \mathbf{Y}_j; \pi)$ tends to dominate the (usually non-informative) prior distribution of the parameter, because the primary interest of an imputation model is the estimation of distribution of the observed data, which determines the imputations. Thus, as remarked by Vermunt et al. (2008), we do not need to interpret π, but rather obtain a good description of the joint distribution of

the items. Moreover, since an imputation model should be as general as possible (that is, it should make as few assumptions as possible) in order to be able to describe all the possible relationships between the items needed in the post-imputation analysis (Schafer & Graham, 2002), we will work with the unrestricted version of the multilevel LC model proposed by Vermunt (2003). In such a version, both the level-1 latent proportions and the level-1 conditional response probabilities are free to vary across the L level-2 LCs.

For a deeper insight into the (frequentist) multilevel LC model we refer to Vermunt (2003, 2008).

2.2 The Prior Distribution

In order to obtain a Bayesian estimation of the model defined by equations (1) and (2), a prior distribution for π is needed. For the Multinomial distribution, a class of conjugate priors widely used in the literature is the Dirichlet distribution. The Dirichlet distribution gives a probability measure in the simplex $\{(q_1, ..., q_D)|q_d > 0 \; \forall d \text{ and } \sum_d q_d = 1\}$ (where D represents the number of categories of the Multinomial distribution) and its parameter can be seen as a *pseudo-count* artificially added by the analyst in the model. Thus, for the BMLC model we assume as priors:

(a) $\pi_W \sim \text{Dir}(\alpha_W)$,

(b) $\pi_{lX} \sim \text{Dir}(\alpha_{lX})$,

(c) $\pi_{lt} \sim \text{Dir}(\alpha_{lt})$,

(d)$\pi_{lks} \sim \text{Dir}(\alpha_{lks})$.

Under this notation, the hyperparameters of the Dirichlet distribution denote vectors, in which each single value is the pseudo-count placed on the corresponding probability value. Thus, α_W corresponds to the vector $(\alpha_1, ..., \alpha_l, ..., \alpha_L)$, and similarly $\alpha_{lX} = (\alpha_{l1}, ..., \alpha_{lk}, ..., \alpha_{lK}) \; \forall \; l$, $\alpha_{lt} = (\alpha_{lt1}, ..., \alpha_{ltr}, ..., \alpha_{ltR_t}) \; \forall \; l, t$ and $\alpha_{lks} = (\alpha_{lks1}, ..., \alpha_{lksu}, ..., \alpha_{lksU_s}) \; \forall \; l, k, s$. The vector containing all the hyperparameter values will be indicated by $\alpha = (\alpha_W, ..., \alpha_{lks}) \; \forall \; l, k, s, t$.

Because in our MI application we will work with symmetric Dirichlet priors[3], in the remainder of the paper we will use the value of a single pseudo-count to denote the value of the whole corresponding vector. For instance, the notation $\alpha_l = 1$ will indicate that the whole vector α_W will be a vector of 1's.

Vidotto, Vermunt, and Van Deun (submitted) studied the effect of different Dirichlet prior specification for the MI through LC models in single-level datasets. Their results showed that unbiased parameter estimates and confidence intervals with coverage close to their nominal level are achievable by setting informative symmetric Dirichlet priors for the latent mixture proportions and non informative Dirichlet priors for the conditional response probabilities. In fact, setting a large value[4] for the pseudo-counts

[3]That is, Dirichlet distributions whose all the pseudo-counts are equal to each other.

[4]Precisely, the value of the pseudo-count should be at least equal to half times the number of free parameters to be estimated within each LC. See Rousseau and Mergensen (2011) for technical details.

of the LC probabilities will cause the Gibbs sampler (see Section 2.4) to distribute the units across all the specified latent clusters. When estimating an overfitting Bayesian LC model[5], which is less of a problem in MI as remarked by Vermunt et al. (2008), there will always be a non-zero probability that at least one of the classes will be empty if smaller values of the latent probabilities hyperparameters are set. Thus, making this prior distribution more informative will help the Gibbs sampler to draw from the equilibrium distribution $\pi|\mathbf{Z}_j, \mathbf{Y}_j$ and, accordingly, will produce imputations that will lead to correct inferences, because the model will exploit all the selected classes. Furthermore, since - as already mentioned - the imputation model parameter values are of no concern in MI, making the prior distribution for the mixture components more informative will not hamper correct post-imputation inferences.

Conversely, Vidotto et al. (submitted) found out that the prior distribution for the conditional response probabilities should have as little influence on the posterior distribution as possible, say, for instance, with a hyperparameter value equal to 0.01 or 0.05 for all the categories of every item within each LC. This will ensure as little influence as possible of the pseudo-counts in the imputation stage: as it will be clear from Algorithm 1 of Section 2.4, the imputations will be performed through the conditional distribution of the variable with missingness, given the membership of a unit to a specific LC. Making the prior distribution of the conditional response probabilities as non-informative as possible will help creating imputations that are almost exclusively based on the observed data.

Concerning the BMLC model, little is known about the effect of the choice of prior distributions for Model (1) because the model has not been extensively explored in the literature. Nonetheless, we suspect that behaviors observed for single-level LC imputation models will also hold at the higher level of the hierarchy. In order to assess the effect of different prior specifications for the level-2 model parameters, we will manipulate α_l and α_{ltr} in the study of Section 3. For the lower-level model (Model (2) in the previous section), we will assume that the findings of Vidotto et al. (submitted) hold[6]. Therefore, we will set informative values for $\alpha_{lk}\ \forall\ l, k$ and non-informative values for $\alpha_{lksu}\ \forall\ l, k, s$.

2.3 Model Selection

In MI, mis-specifying a model in the direction of over-fitting is less problematic than mis-specifying towards under-fitting (Carpenter & Kenward, 2013; Vermunt et al., 2008). While the former case, in fact, might lead to slightly over-conservative inferences in the worst scenario, the latter case is likely to introduce bias (and too liberal inferences) since important features of the data are omitted. For this reason, with single-level LC imputation models, Vidotto et al. (submitted) suggested to define an arbitrarily large number of classes, expressed as a function of the number of patterns in the dataset.

[5]That is, a model with a number of classes larger than what is actually supported by the data; see next section.

[6]This conjecture is justified by noticing that, given a level-2 LC W_j, the lower level model corresponds to a standard LC model.

That is, if Q denotes the number of occurring unique patterns in the data, their method implies fixing the number of LCs equal to Q/C, with $C > 0$ a tunable constant. In this way, the model can take into account sample size and number of observed categories at the same time. Furthermore, C should be selected such that the model can pick up a significant level of detail from the observed patterns, while maintaining reasonable computational speed.

For the BMLC, when variables at both levels are available we follow the above guidelines and select L and K depending on the number of observed patterns. Let us denote with $Q^{(2)}$ the number of observed patterns at level-2 of the dataset (that is, the number of observed patterns among the variables $Z_1, ..., Z_T$) and with $Q^{(1)}$ the number of observed patterns at level-1 (that is, the number of observed patterns among the items $Y_1, ..., Y_S$ for all the J level-2 units). For the level-2 number of components, we propose

$$L = \frac{Q^{(2)}}{C_2} \tag{3}$$

and for the level-1 number of components we propose

$$K = \frac{1}{L}\frac{Q^{(1)}}{C_1}, \tag{4}$$

where $C_2 > 0$ and $C_1 > 0$ are suitably chosen constants. C_2 and C_1 can be viewed as the expected number of patterns to be allocated within each LC. This criterion offers different advantages. Besides taking into account the number of observed categories and the sample sizes at both levels, it automatically regulates K according to L: a larger L will need a smaller K in order to distribute all the level-1 units across the lower-level classes of the models, while with a smaller L a larger K will be needed. As a rule of thumb, we propose a compromise between a good level of detail picked by the BMLC model and acceptable computational speed, achievable by setting $C_2 = 10$ and $C_1 = 4$. Increasing these values will decrease the computing time, but also the level of detail captured by the model, while decreasing them will have the opposite effect. These values should be further investigated in future research.

Selection of L and K through the presented formulas should enable the level-2 mixture variable W to capture both level-2 items variability and level-1 subjects heterogeneity. However, in some cases, the researcher might doubt that the selected L is large enought to pick up both these sources of variability. In other cases, items at level-2 $\{Z_1, ..., Z_T\}$ may not be available, because they are not necessary for the study design or for the research question. In other circumstances, furthermore, variables are available at level-2 but the dataset at hand might be so large that the number of classes selected by the above criterion would not allow for practical computing time. Under all these circumstances, model selection can be performed as described in Gelman, Carlin, Stern, and Rubin (2013), chapter 22. The procedure requires running the Gibbs sampler described in Algorithm 1 of the next section (without Step 7) with

arbitrarily large L^* and K^*, and setting hyperparameters for the LC probabilities that can favor empty superfluous components. As suggested in Gelman et al. (2013)[7], this values could be $\alpha_l = 1/L^* \; \forall \; l$ and $\alpha_K = 1/K^* \; \forall \; k$. At the end of every iteration of the Gibbs sampler, we keep track of how many LCs have been actually allocated, in order to obtain a distribution for L and K when the algorithm is over. If the posterior modes L_{mod} and K_{mod} of such distributions are smaller than L^* e K^*, in the next step we perform the imputations with L_{mod} and K^* if the value of K^* does not involve intensive computations, or a value for K included between K_{mod} and K^* otherwise. However, if at least one between L_{mod} and K_{mod} equals $L*$ and/or K^*, we re-run the preliminary Gibbs sampler by increasing the corresponding value(s), and repeat the procedure until optimal L and K are found.

2.4 Estimation and Imputation

Since we are dealing with unobserved variables (W and X), model estimation is performed through a Gibbs sampler with Data Augmentation scheme (Tanner & Wong, 1987). Following the estimation and imputation scheme proposed for single-level LC imputation models by Vermunt et al. (2008) , we will perform the estimation only on the observed part of the dataset (denoted by $\{\mathbf{Y}^{obs}, \mathbf{Z}^{obs}\}$). In particular, in the first part of Algorithm 1 (see below) the BMLC model is estimated by first assigning the units to the LCs (steps 1-2) through the *posterior membership probabilities* -the probability for a unit to belong to a certain LC conditioned on the observed data, $\Pr(W_j|\mathbf{Y}_j^{obs}, \mathbf{Z}_j^{obs})$ and $\Pr(X_{ji}|W_j, \mathbf{Y}_j^{obs}, \mathbf{Z}_j^{obs}) \; \forall \; i, j$- and subsequently by updating the model parameter (steps 3,4,5,6). At the end of the Gibbs sampler (step 7), after the model has been estimated, we impute the missing data through m draws from $\Pr(\pi|\mathbf{Y}^{obs}, \mathbf{Z}^{obs})$.

After fixing K, L and α, we must establish I, the number of total iterations for the Gibbs sampler. If we denote with b the number of the iterations necessary for the burn-in, we will set I such that $I = b+(I-b)$, where $I-b$ is the number of iterations used for the estimation of the equilibrium distribution $\Pr(\pi|\mathbf{Y}^{obs}, \mathbf{Z}^{obs})$, from which we will draw the parameter values necessary for the imputations. Of course, b must be large enough to ensure convergence of the chain to its equilibrium (which can be assessed from the output of the Gibbs sampler).

We initialize $\pi^{(0)}$ through draws from uniform Dirichlet distributions (that is, Dirichlet distributions with all their parameter values set equal to 1), in order to obtain $\pi_W^{(0)}$, $\pi_{lX}^{(0)}$, $\pi_{lt}^{(0)}$ and $\pi_{lks}^{(0)} \; \forall \; l, k, t, s$. After all these preliminary steps are performed, the Gibbs sampler is run as shown in Algorithm 1.

[7]Importantly, while Gelman et al. (2013)'s goal was to find a minimum number of interpretable clusters for inference purposes, here our goal is to find a value for L and K for imputation purposes. Moreover, Gelman et al. (2013)'s method was designed for single-level mixture models. We extend here the mechanism to the level-2 mixture variable.

Algorithm 1:

(A) Part 1. For $h = 1, ..., I$:

1. for $j = 1, ..., J$ sample $W_j^{(h)} \in \{1, ..., L\}$ from a Multinomial distribution with the posterior membership probabilities at level-two as parameters (and sample size 1), calculated through $\Pr(W_j^{(h)} = l | \mathbf{Y}_j^{obs}, \mathbf{Z}_j^{obs}, \pi^{(h-1)}) =$

$$\frac{\pi_l^{(h-1)} \left\{ \prod_{t=1}^{T} \prod_{r=1}^{R_t} \left(\pi_{ltr}^{(h-1)} \right)^{\mathcal{I}_{jt}^{r*}} \right\} \left\{ \prod_{i=1}^{n_j} \sum_{k=1}^{K} \pi_{lk}^{(h-1)} \prod_{s=1}^{S} \prod_{u=1}^{U_s} \left(\pi_{lksu}^{(h-1)} \right)^{\mathcal{I}_{jis}^{u*}} \right\}}{\sum_{p=1}^{L} \pi_p^{(h-1)} \left\{ \prod_{t=1}^{T} \prod_{r=1}^{R_t} \left(\pi_{ptr}^{(h-1)} \right)^{\mathcal{I}_{jt}^{r*}} \right\} \left\{ \prod_{i=1}^{n_j} \sum_{k=1}^{K} \pi_{pk}^{(h-1)} \prod_{s=1}^{S} \prod_{u=1}^{U_s} \left(\pi_{pksu}^{(h-1)} \right)^{\mathcal{I}_{jis}^{u*}} \right\}},$$

in which $\mathcal{I}_{jt}^{r*} = 1$ if $Z_{jt} = r$ and $Z_{jt} \in \mathbf{Z}^{obs}$ or $\mathcal{I}_{jt}^{r*} = 0$ otherwise, and similarly $\mathcal{I}_{jis}^{u*} = 1$ if $Y_{jis} = u$ and $Y_{jis} \in \mathbf{Y}^{obs}$ or $\mathcal{I}_{jis}^{u*} = 0$ otherwise;

2. for $i = 1, ..., n_j \ \forall \ j$, and given $W_j^{(h)}$, sample $X_{ji}^{(h)} \in \{1, ..., K\}$ from a Multinomial distribution with the posterior membership probabilities at level-one as parameters (and sample size 1), calculated through

$$\Pr(X_{ji}^{(h)} = k | W_j^{(h)} = l, \mathbf{Y}_j^{obs}, \mathbf{Z}_j^{obs}, \pi^{(h-1)}) = \frac{\pi_{lk}^{(h-1)} \left\{ \prod_{s=1}^{S} \prod_{u=1}^{U_s} \left(\pi_{lksu}^{(h-1)} \right)^{\mathcal{I}_{jis}^{u*}} \right\}}{\sum_{v=1}^{V} \pi_{lv}^{(h-1)} \left\{ \prod_{s=1}^{S} \prod_{u=1}^{U_s} \left(\pi_{lvsu}^{(h-1)} \right)^{\mathcal{I}_{jis}^{u*}} \right\}};$$

3. draw

$$\left(\pi_W^{(h)} | W^{(h)}, \alpha_W \right) \sim Dir \left(\alpha_1 + \sum_{j=1}^{J} \mathcal{I} \left(W_j^{(h)} = 1 \right), \cdots, \alpha_L + \sum_{j=1}^{J} \mathcal{I} \left(W_j^{(h)} = L \right) \right)$$

where $\mathcal{I}(w_j^{(h)} = l) = 1$ if $w_j^{(h)} = l$ and 0 otherwise;

4. for $l = 1, ..., L$ draw
$\left(\pi_{lX}^{(h)} | W^{(h)} = l, X^{(h)}, \alpha_{lX} \right) \sim$

$$Dir \left(\alpha_{l1} + \sum_{j,i:W_j^{(h)}=l} \mathcal{I} \left(X_{ji}^{(h)} = 1 \right), \cdots, \alpha_{lK} + \sum_{j,i:W_j^{(h)}=l} \mathcal{I} \left(X_{ji}^{(h)} = K \right) \right)$$

where $\mathcal{I}(X_{ji}^{(h)} = k) = 1$ if $X_{ji}^{(h)} = k$ and 0 otherwise;

5. for $l = 1, ..., L$, $t = 1, ..., T$ draw

$$\left(\pi_{lt} | W^{(h)} = l, \mathbf{Z}_t^{obs}, \alpha_{lt} \right) \sim Dir \left(\alpha_{lt1} + \sum_{j:W_j^{(h)}=l} \mathcal{I}_{jt}^{1*}, \cdots, \alpha_{ltR_t} + \sum_{j:W_j^{(h)}=l} \mathcal{I}_{jt}^{R_t*} \right);$$

6. for $l = 1, ..., L$, $k = 1, ..., K$, $s = 1, ..., S$ draw
$\left(\pi_{lks}|W^{(h)} = l, X^{(h)} = k, \mathbf{Y}_s^{obs}, \alpha_{lks}\right) \sim$

$$Dir\left(\alpha_{lks1} + \sum_{j,i:W_j^{(h)}=l \cap X_{ji}^{(h)}=k} \mathcal{I}_{jis}^{1*}, \cdots, \alpha_{lksU_s} + \sum_{j,i:W_j^{(h)}=l \cap X_{ji}^{(h)}=k} \mathcal{I}_{jis}^{U_s*}\right).$$

(B) Part 2. After I iterations:

7. (*imputation step*) perform m draws from the distribution $\Pr(\pi|\mathbf{Y}^{obs}, \mathbf{Z}^{obs})$ estimated in Steps 1-6; in particular, the f-th draw ($f = 1, ..., m$) must include $w_j^{(f)}$, $x_{ji}^{(f)}$, $\pi_{lt}^{(f)}$ and $\pi_{lks}^{(f)}$ $\forall j, i, t, s \in \{\mathbf{Y}^{mis}, \mathbf{Z}^{mis}\}$, the missing part of the dataset. Perform the f-th imputation for the items at level-2 by drawing

$$\left(Z_{jt}|W_j^{(f)} = l, \pi^{(f)}\right) \sim Multinom\left(\pi_{lt}^{(f)}\right),$$

and the f-th imputation for the items at level-1 by drawing

$$\left(Y_{jis}|W_j^{(f)} = l, X_{ji}^{(f)} = k, \pi^{(f)}\right) \sim Multinom\left(\pi_{lks}^{(f)}\right),$$

$\forall\ Z_{jt}, Y_{jis} \in \{\mathbf{Y}^{mis}, \mathbf{Z}^{mis}\}$.

Clearly, the m parameter values obtained at Step 7 should be independent, such that no autocorrelations are present among them. This can be achieved by selecting I large enough and performing m equally spaced draws between iteration $b + 1$ and iteration I. The Gibbs sampler output can help to assess the convergence of the chain.

3 Study 1: Simulation Study

3.1 Study Set-up

In Study 1, we evaluated the performance of the BMLC model and compared it with the performance of the LD and the JOMO methods.

We generated 500 datasets from a population model, created missing data through a MAR mechanism, and then applied the JOMO and BMLC imputation methods, as well as the LD technique, to the incomplete datasets. To assess the performance of the missing data methods the bias, stability and coverage rates of the 95% confidence intervals were compared, where the results of the complete-data case (that is, the results obtained if there were no missingness in the dataset) were taken as benchmark.

Parameter	β_{00}	β_1	β_2	β_3	β_4	β_5	β_{24}	γ_1	γ_2	γ_3	γ_4	γ_5	γ_{35}	τ^2
Value	-0.5	1.35	-1	-0.4	0.8	-0.75	0.25	0.5	0.85	0.45	-0.6	0.3	0.15	1

Table 1: Parameter values for Model 5-6.

Population Model. For each of the 500 datasets, we generated $T = 5$ binary level-2 predictors $\mathbf{Z}_j = (Z_{j1}, ..., Z_{j5})$ for each higher-level unit $j = 1, ..., J$ from the log-linear model

$$log\Pr(\mathbf{Z}_j) = -.1\sum_{t=1}^{5} Z_{jt} + .1\sum_{t=1}^{4}\sum_{u=(t+1)}^{5} Z_{jt}Z_{ju} + .8Z_{j1}Z_{j2}Z_{j4}.$$

Within each level-2 unit j, $S = 5$ binary level-1 predictors $\mathbf{Y}_{ji} = (Y_{ji1}, ..., Y_{ji5})$ were generated for each level-1 unit $i = i, ..., n_j$ from the (conditional) log-linear model

$$log\Pr(\mathbf{Y}_{ji}|\mathbf{Z}_j) = 1.5\sum_{s=1}^{5} Y_{jis} - .5\sum_{s=1}^{4}\sum_{v=(s+1)}^{5} Y_{jis}Y_{jiv} - 1.5Y_{ji1}Y_{ji2}Y_{ji3} + Y_{ji3}Y_{ji4}Y_{ji5}$$
$$+ 2.25Y_{ji4}Z_{j1} + 1.5Y_{j2}Z_{j2} - 2.3Y_{j3}Z_{j4},$$

where cross-level interactions were inserted to introduce some intra-class correlation between the level-1 units. Finally, we generated the binary outcome Y_6 from a random intercept logistic model, where

$$logit\Pr(Y_{ji6}|\mathbf{Y}_{ji}, \mathbf{Z}_j) = \beta_{j0} + \beta_1 Y_{ji1} + \beta_2 Y_{ji2} + \beta_3 Y_{ji3} + \beta_4 Y_{ji4} + (\beta_5 + \gamma_{35}Z_{j3})Y_{ji5} + \beta_{24}Y_{ji2}Y_{ji4} \quad (5)$$

was the level-1 response model and

$$\beta_{j0} = \beta_{00} + \gamma_1 Z_{j1} + \gamma_2 Z_{j2} + \gamma_3 Z_{j3} + \gamma_4 Z_{j4} + \gamma_5 Z_{j5} + u_j, \text{ with } u_j \sim N(0, \tau^2) \quad (6)$$

was the level-2 model. Table 1 shows the numerical values of the level-1 parameter $\beta = (\beta_{00}, ..., \beta_{24}, \gamma_{35})$ and the level-2 parameter $\gamma = (\gamma_1, ..., \gamma_5)$, which include the cross-level interaction γ_{35}. Table 1 also reports the value of the variance of the random effects, τ^2. Model 5-6 was the analysis model of our study, in which the main goal was recovering its parameter estimates after generating missingness.

Sample size conditions. We fixed the total level-1 sample size to $n = \sum_j n_j = 1000$, and generated 500 datasets for two different level-2 and level-1 sample size conditions. In the first condition, $J = 50$ and $n_j = 20 \; \forall \; j$, while in the second condition $J = 200$ and $n_j = 5 \; \forall \; j$.

Generating missing data. From each dataset, we generated missingness according to the following MAR mechanism. For each combination of the variables (Y_3, Y_4) observations were made missing in Y_1 with probabilities $(0.05, 0.55, 0.4, 0.14)$; for each combination of the variables (Y_3, Y_6) observations were made missing in Y_2 with probabilities $(0.15, 0.25, 0.65, 0.35)$; for each combination of (Y_4, Z_4) observations were made missing in Y_5 with probabilities $(0.01, 0.1, 0.55, 0.2)$; for each possible value of the variable Z_2 missingness was generated on Z_1 with probabilities $(0.15, 0.4)$; finally, for each of the values taken on by Z_5 missingness was generated on Z_2 with probabilities $(0.1, 0.5)$. Through such a mechanism, the rate of nonresponses across the 500 datasets was on average 30% for each item with missingness.

Missing-data methods. We applied three missing data techniques to the incomplete datasets: LD, JOMO and BMLC imputation, with the latter set up as follows. Since we are dealing with small datasets, we selected the number of classes L and K according to the method exposed in Section 2.3, with $C_2 = 10$ and $C_1 = 4$. This led to an average number of classes equal to $L = 3.77$ at level-2 and $K = 10.57$ at level-1 when $J = 50,\ n_j = 20$ and to $L = 6.35$ at level-2 and $K = 6.58$ at level-1 when $J = 200,\ n_j = 5$. Hyperparameters of the level-1 LCs and conditional responses (namely α_{lx} and $\alpha_{lks}\ \forall\, l, k, s$) were set following the guidelines of Section 2.2, that is, with informative prior distributions[8] for the parameters π_{lX} and with a non-informative prior distribution for the parameters π_{lks}. In order to asses the performance of the BMLC model under different level-2 prior specifications, we manipulated the level-2 hyperparameters α_l and α_{ltr}. Each possible variant of the BMLC model will be denoted by $BMLC(\alpha_l, \alpha_{ltr})$. In particular, we tested the BMLC model with uniform priors for both the level-2 LC variable parameters and the level-2 conditional response parameters - the BMLC(1,1) model - or with non-informative prior for the conditional responses - the BMLC(1,.01) model. We alternated the same values for the conditional response pseudo-counts with a more informative value for the level-2 mixture variable parameter, the BMLC(*,1) and the BMLC(*,.01) model. Here, the '*' denotes the hyperparameter choice based on the number of free parameters[9] within each class $l = 1, ..., L$; since this number could change with K, different values for this hyperparameter were used across the 500 datasets. For each dataset, $m = 5$ imputations were used and a total of $I = 5000$ Gibbs sampler iterations were run, of which $b = 2000$ were used for the burn-in and $I - b = 3000$ for the imputations.

For the JOMO imputation method, which also performs imputation through Gibbs sampling, we specified a joint model for the categorical variables with missingness, and used the variables with completely observed data as predictors. We set the number of burn-in iterations equal to $b = 10000$, and performed the 5 imputations for each dataset across $I - b = 3000$ iterations, in order to have a number of iterations for the imputations equal to the Gibbs sampler of the BMLC method. We ran the algorithm with its default settings: non-informative priors and common covariance matrix across clusters.

[8] Calculated through $\alpha_{lk} = \left(\sum_s (U_s - 1)\right)/2\ \forall\, l, k$.
[9] Calculated through $\alpha_l = \left(\sum_t (R_t - 1) + (K - 1) + K(\sum_s U_s - 1)\right)/2\ \forall\, l$.

In order to have a benchmark for results comparison, we also estimated Model 5-6 to the complete data, before generating the missingness.

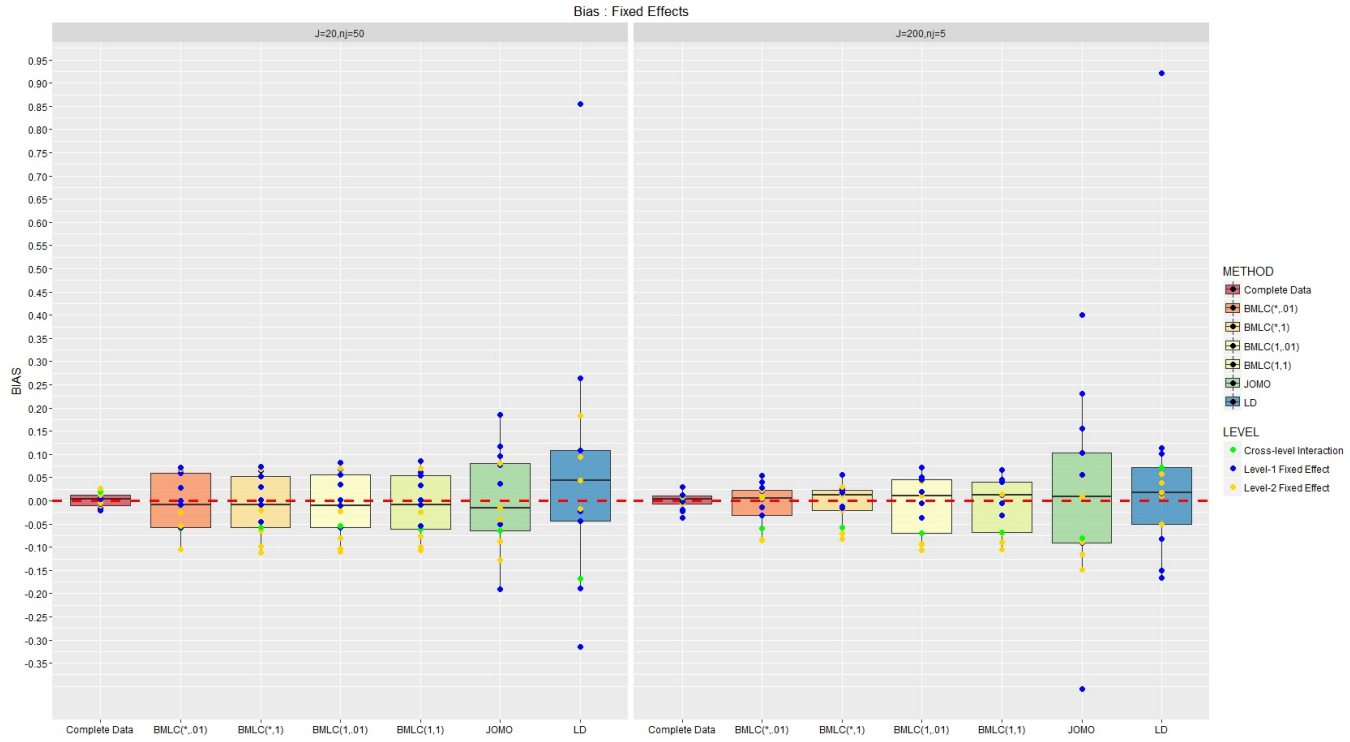
Study outcomes. For each parameter of Model 5-6, we compared the bias of the estimates, along with their standard deviation (to assess stability) and coverage rate of the 95% confidence intervals. Analyses were performed with R version 3.3.0. JOMO was run from the `jomo` R-library.

3.2 Study Results

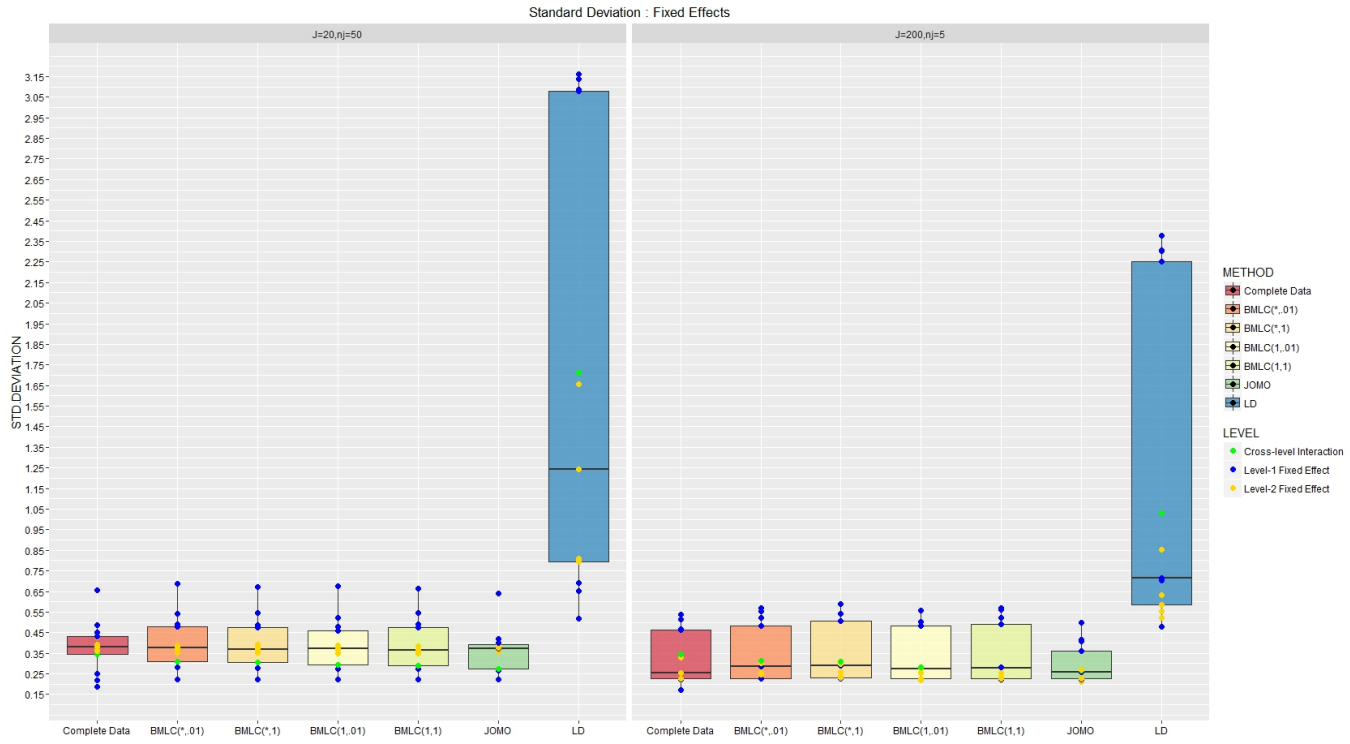
Figures 2a, 2b and 3 show the averaged bias, standard deviation and coverage rates of the 95% confidence intervals for the thirteen fixed effect coefficients of Model 5-6, averaged over the 500 datasets. The figures also show point estimates of each coefficient, distinguishing between level-1, level-2 and cross-level interaction fixed effects.

Figure 2a reports the bias of the fixed-effects estimates. Under both scenarios, BMLC imputation appeared as the missing-data method which produced the least biased estimates (the boxplots are fairly centered around 0). When $J = 50$ and $n_j = 20$, the choice of the prior distribution for the BMLC model did not seem to affect the final results in term of bias. JOMO imputation produced also on average unbiased estimates, but some parameters resulted in larger bias than the BMLC imputation technique. This is probably due to a big limitation of JOMO: by working with multivariate normal distributions, it can correctly capture (and manifest in the imputations) two-way associations between the items, but higher-order relationships (present in the datasets of this simulation study) are most likely ignored by the method, causing bias in the final estimates. The LD method, which was negatively affected by a smaller sample size, yielded the most biased coefficients. In particular, some of the level-1 fixed effects appeared heavily biased both down- and up-ward. In the $J = 200$, $n_j = 5$ condition, the bias for the BMLC models was reduced with respect to the previous scenario. Under this condition, the specification of the prior distribution seemed to have an effect in the final estimates produced by the BMLC model. In particular, priors which favored a full allocation of the level-2 units across all the L classes, namely models BMLC(*,.01) and BMLC(*,1), resulted with a slightly smaller bias than priors which not favored full allocation, namely BMLC(1,.01) and BMLC(1,1). With $J = 200$, the bias of the level-2 fixed effects resulting from BMLC imputation was lower than in the condition with $J = 50$. LD method also yielded estimates with smaller bias in the second condition (with the exception of one level-1 fixed effect, β_3), although still more biased, in general, than the ones produced by the BMLC models. As far as the JOMO imputation was concerned, no particular improvements were observed in the bias of the estimates from the scenario with $J = 50$ to the scenario with $J = 200$. On the contrary, some of the level-1 fixed effects ($\beta_{00}, \beta_2, \beta_4$) resulted in a larger bias than in the the previous case.

Figure 2b shows the stability of the estimates produced by all models, represented by their standard



(a)



(b)

Figure 2: Bias (a) and standard deviation (b) observed for the thirteen fixed multilevel logistic regression level-1, level-2, and cross-level coefficients obtained with complete data and the missing data methods BMLC(*,.01), BMLC(*,1), BMLC(1,.01), BMLC(1,1), JOMO and LD. Left: $J = 50,\ n_j = 20$. Right: $J = 200,\ n_j = 5$.

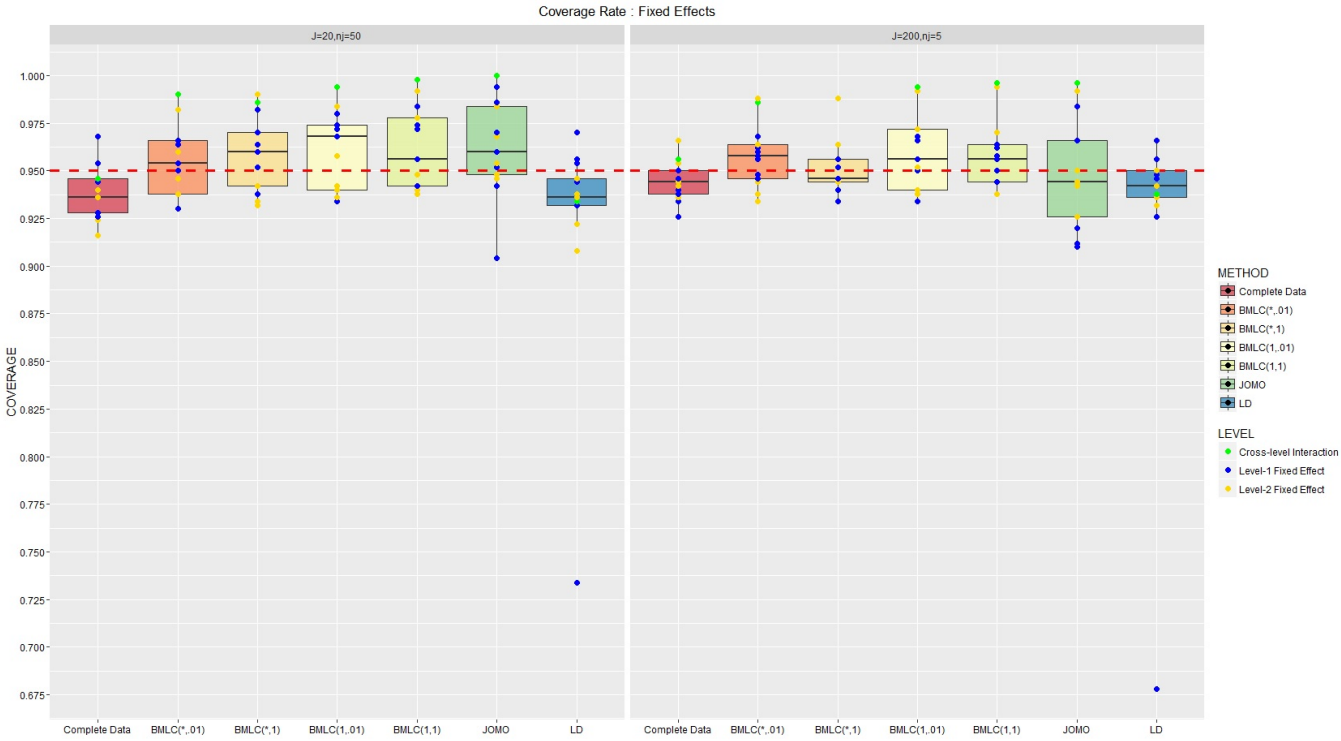


Figure 3: Coverage rates observed for the confidence intervals of the thirteen fixed multilevel logistic regression level-1, level-2, and cross-level coefficients obtained with complete data and the missing data methods BMLC(*,.01), BMLC(*,1), BMLC(1,.01), BMLC(1,1), JOMO and LD. Left: J = 50; nj = 20. Right: J = 200; nj = 5.

deviations across replications. The BMLC methods were the most similar - in terms of magnitude - to the Complete Data case, with both $J = 50$ and $J = 200$. For such models, the prior distribution did not seem to have an influence on the stability of the estimates. LD technique estimates were the most unstable, as a result of a smaller sample size. The JOMO imputation technique, on the other hand, resulted with the most stable estimates, even more than the Complete Data case. As already observed, this was probably due to the fact that the JOMO method, ignoring higher orders relationships, was an imputation model simpler than what was required by the data, and produced estimates that did not vary as they should.

Figure 3 displays the coverage rates of the 95% confidence intervals obtained with each method: some of the coverages resulting from the JOMO imputation, in particular the coverage rate of one level-1 fixed effect (β_1) under the $J = 50$ condition and three level-1 fixed effects ($\beta_0, \beta_1, \beta_2$) in the $J = 200$ condition, were too small (between 0.9 and 0.92), as a result of a too simple imputation model. LD produced, overall, coverage rates closed to the ones obtained under the Complete Data case. However, the coverages of the confidence intervals yielded by the LD method were the result of a large bias and large standard errors of the parameter estimates, which led to too wide intervals. Furthermore, the

$\tau^2 = 1$: Bias		
Method	$J = 50, n_j = 20$	$J = 200, n_j = 5$
Complete Data	-0.14	-0.05
BMLC(*,.01)	-0.17	-0.08
BMLC(*,1)	-0.16	-0.05
BMLC(1,.01)	-0.16	-0.08
BMLC(1,1)	-0.16	-0.06
JOMO	-0.12	-0.05
LD	**-0.31**	0.05

Table 2: Bias of the variance of the random effect for the complete data and the missing data methods BMLC(*,.01), BMLC(*,1), BMLC(1,.01), BMLC(1,1), JOMO and LD. Significant bias (w.r.t. the complete data estimator) is marked in boldface.

LD method generated coefficients for one of the parameters (β_3) with a too low coverage (about 0.7). The BMLC imputation method produced more conservative confidence intervals when $J = 50$ than the $J = 200$ condition. In this latter case, the intervals appeared closer to their nominal level. Behavior of the confidence intervals for the BMLC models also depended on the prior distribution used by the model. In fact, priors which favored full allocation of the level-2 LCs led to confidence intervals slightly closer to the nominal level. Interestingly, both imputation methods (BMLC and JOMO) produced confidence intervals that, overall, had a coverage rate larger than their nominal level. This can be the consequence of the large amount of missingness (about 30% for each item) entered in the data. Moreover, for the BMLC model, this can also be attributed to the over-fitting strategy pursued in this paper.

Table 2 reports the results obtained for the variance of the random effects, in term of bias. All the BMLC models, as well as the JOMO imputation model, yielded a random effect variance very close to the Complete Data case one under both scenarios. In the $J = 50$ condition, the variance estimated by JOMO resulted less biased than the Complete Data estimator. This might be the consequence of the biased fixed effect parameters produced by this imputation technique, shown in Figure 2. Finally, the LD method produced the most biased variance of the random effects, in particular when the number of level-2 units was equal to $J = 50$.

4 Study 2: Real-data case

The European Social Survey (NSD: Norwegian Centre for Research Data, 2012), or ESS, collects sociological, economical and behavioral data from European citizens. The survey is performed by the NSD (Norwegian Centre for Research Data) every two years, and consists of items both at the individual (level-1) and at the country (level-2) level. The data are freely available at the website http://www.europeansocialsurvey.org/. In order to assess the performance of the BMLC model with

real data, we carried out an analysis using the ESS data of Round 6, which consists of multilevel data collected in 2012.

After cleaning the dataset, we estimated a possible analysis model using one of the items as outcome variable. Subsequently, we introduced missingess according to a MAR mechanism. Finally, the results (bias of the estimates, standard errors and p-values) obtained after BMLC imputation were compared with the results obtained under the Complete Data case and the LD method. We also made an attempt to perform imputations with the JOMO technique, but the dataset was too large for this routine. After 4 days of computation on a normal calculator (Intel Core i7), JOMO had no completed the burn-in iterations yet, and we decided to stop the process. This highlights one of the main issues of the JOMO procedure: when dealing with large datasets, this package must deal with too many multivariate normal variables and random effects, and becomes extremely slow. As a comparison, computations with the BMLC model required less than two days on the same machine for both the model selection and the imputation stages (see below for details).

4.1 Study Set-up

Data preparation. The original datasets consisted of $n = 54673$ level-1 respondents within $J = 29$ countries and 36 variables, of which $T = 15$ were observed at the country level, $S = 20$ at the person level and one item was the country indicator. At level-1, items consisted either of social, political, economical and behavioral questions, which the respondent were asked to rate (e.g., from 0 to 10) according to their opinion, or of background variables, such as age and education. At level-2, some economical and political (continuous) indicator related to the countries were reported. Some of the units (at both levels) contained missing or meaningless values (such as "Not Applicable"), and those units were removed from the dataset, in order to work with "clean" data. Furthermore, we recoded the qualitative levels of the rating scales and converted them to numbered categories, and transformed some continuous variables (such as Age or all the level-2 items) into integer valued categories[10]. This enabled us to run the BMLC model on this dataset.

After removing level-1 items related with the study design and least "recent" versions of the items (i.e., all the replicated items across the survey waves, observed before 2010), and discarding units younger than 18 years old and/or not eligible for voting (in the next sub-paragraph we will explain the reason of this choice), $T = 11$ level-2 and $S = 17$ level-1 items were left, observed across $n = 28704$ level-1 units within $J = 21$ countries. These countries were Belgium ($n_j = 1497$), Switzerland ($n_j = 1002$), Czech Republic ($n_j = 1308$), Germany ($n_j = 2285$), Denmark ($n_j = 1321$), Estonia ($n_j = 1485$), Spain ($n_j = 1429$), Finland ($n_j = 1772$), France ($n_j = 1581$), UK ($n_j = 1575$), Hungary ($n_j = 1327$), Ireland

[10]In particular, percentiles were used to create break-points and allocate units into the new categories. The choice of the percentiles depended on the number of categories used for each item.

($n_j = 1948$), Iceland ($n_j = 519$), Italy ($n_j = 623$), Netherlands ($n_j = 1591$), Norway ($n_j = 1312$), Poland ($n_j = 1281$), Portugal ($n_j = 1263$), Sweden ($n_j = 1473$), Slovenia ($n_j = 706$) and Slovakia ($n_j = 1406$).

Analysis model. We further restricted the dataset (in terms of number of items) by looking for a possible model of interest that can be estimated with the data at hand. First, we selected the binary variable "Voted in the last elections" (Y_0) as outcome. This is why we deleted the level-1 units "Not eligible for voting" from the dataset in the previous step. Second, we looked for possible items that could significantly explain the variability of this item through a multilevel logistic model. Selection of the predictors (and of the random effects) was performed through stepwise forward selection, including in the model only the significant predictors (i.e., with p-values lower than 0.05) which led to a drop of the AIC index of the model. The final model for "Voted in the last elections" was a multilevel logistic model with random intercept and random slope, and was specified as

$$\begin{aligned} logit\,\Pr(Y_{ji0}|\mathbf{Y}_{ji}, \mathbf{Z}_j) =& \beta_{j0} + (\beta_1 + \gamma_{11} Z_{j1}) Y_{ji1} + \beta_2 Y_{ji2} + \beta_3 Y_{ji3} + \beta_4 Y_{ji4} + \beta_5 Y_{ji5} \\ &+ \beta_6 Y_{ji6} + \beta_{j7} Y_{ji7} + \beta_8 Y_{ji8} + \beta_9 Y_{ji9} \end{aligned} \tag{7}$$

at level-1 and

$$\begin{aligned} \beta_{j0} =& \beta_{00} + \gamma_1 Z_{j1} + u_{j0}, \text{ with } u_{j0} \sim N(0, \tau_0^2 = 0.29), \\ \beta_{j7} =& \beta_{70} + u_{j1}, \text{ with } u_{j1} \sim N(0, \tau_1^2 = 0.02) \end{aligned} \tag{8}$$

at level-2. A description of the 11 variables used in the model can be found at the top of Table 3, while the values of the coefficients (both fixed and random) are reported in the second column of Table 5 below. Furthermore, columns 5 and 8 of Table 5 show standard errors and p-values (for the hypothesis of null coefficients) of the fixed effect parameters, obtained with the original data.

Entering missingness. Subsequently, we entered MAR missingness in the dataset. Missingness was generated on Y_2, Y_4, Y_7, Y_6 and Z_1 through logistic models for the nonresponse indicator. We did not only use the variables in Model 7-8 in order to generate the missingness, but also other items still present in the dataset. The latter are listed in the bottom part of Table 3. Table 4 shows the logistic models used to create missingness. The coefficients of these models were chosen in such a way to ensure between (about) 14% and 25% of missingness for each of the selected items. At the end of the process, only 18 countries and 9871 level-1 units (about one third of the dataset) were left with fully observed data.

Missing data methods. We applied LD and BMLC to the sample with nonresponses. The BMLC was run with all the 23 variables listed in Table 3, and was set as follows. Since the dataset was too large to use the criterion given by (3)-(4) in Section 2.3 (with $C_2 = 10$ and $C_1 = 4$, we should set $L = 2$

Item Name	**Description**	**Coding**
Y_0	Voted in the last elections	0 No, 1 Yes
Y_1	TV watching: news and politics	0 No time, 7 >3 hours
Y_2	Trust in politicians	0 No trust, 10 Complete trust
Y_3	Placement in the right/left scale	1 Left, 5 Right
Y_4	Life satisfaction	0 Dissatisfied, 10 Satisfied
Y_5	Immigration is bad/good for economy	0 Bad, 10 Good
Y_6	National elections are free and fair	0 Not important, 10 Extremely important
Y_7	Age (Range)	1 (18/34), 5 (68/103)
Y_8	Marital status	0 Not married, 1 Married
Y_9	Highest level of education	1 <Secondary, 7 >Tertiary
Z_1	Social Expenditure (Country level)	1 Low, 2 High
$Y_1 Z_1$	Cross-level interaction between Y_1 and Z_1	-
	Other Items used to generate missingess:	
Item Name	**Description**	
Y_{10}	Subjective general health	
Y_{11}	Political parties offer alternatives	
Y_{12}	Media provide reliable information	
Z_2	Area (Country level)	
Z_3	Median age (Country level)	
Z_4	Population size (Country level)	
Z_5	Unemployment level (Country level)	
Z_6	Number of students (primary - secondary education) (Country level)	
Z_7	Number of students (tertiary education) (Country level)	
Z_8	Governmental capabilities (Country level)	
Z_9	Transparency (Country level)	
Z_{10}	Health Expenditure (Country level)	

Table 3: ESS data items used in the Study 2.

Missingness in...	Missingness generating model
Y_2	$1.3 + 0.1Y_{11} - 0.4Y_{12} - 0.15Z_7$
Y_4	$0.5 - 0.5Y_{10} - 0.5Y_9 + Z_5$
Y_6	$-1 - 1.7Y_0 + 0.3Z_{10} + 0.15Z_8$
Y_7	$-0.5 + 0.2Y_3 + 0.25Z_3 - 1.5Z_4$
Z_1	$-1 - Z_9 - 0.5Z_6 + Z_2$

Table 4: Missingness generating mechanism for the items of the ESS dataset.

	Estimates			Standard errors			p-values		
Parameter	Complete Data	LD	BMLC	Complete Data	LD	BMLC	Complete Data	LD	BMLC
β_{00}	-3.45	-2.72	-3.26	0.33	0.44	0.37	0.00	0.00	0.00
β_1	0.15	0.07	0.16	0.04	0.08	0.04	0.00	**0.42**	0.00
β_2	0.07	0.07	0.07	0.01	0.02	0.01	0.00	0.00	0.00
β_3	0.05	0.01	0.05	0.02	0.03	0.02	0.00	**0.82**	0.00
β_4	0.06	0.07	0.07	0.01	0.02	0.01	0.00	0.00	0.00
β_5	0.02	0.04	0.02	0.01	0.01	0.01	0.02	0.01	0.02
β_6	0.12	0.10	0.11	0.01	0.02	0.01	0.00	0.00	0.00
β_{70}	0.34	0.35	0.33	0.03	0.05	0.03	0.00	0.00	0.00
β_8	0.39	0.33	0.39	0.03	0.07	0.03	0.00	0.00	0.00
β_9	0.23	0.23	0.22	0.01	0.02	0.01	0.00	0.00	0.00
γ_1	0.71	0.57	0.59	0.20	0.24	0.23	0.00	0.03	0.02
γ_{11}	-0.06	-0.01	-0.07	0.03	0.06	0.03	0.02	**0.87**	0.04
τ_0^2	0.29	0.42	0.32						
τ_1^2	0.02	0.03	0.01						

Table 5: Study 2: Estimates, standard errors and p-values obtained with Complete Data, LD and BMLC methods for the fixed and random effects parameters of Model 7-8, attained after applying each method to the (fully or partially) observed data. Not-significant 5% p-values are marked in boldface.

and $K = 1250$, or $K = 500$ if we increase C_1 to be equal to 10, which is clearly infeasible in practical time), we decided to perform model selection using the second technique report in the same Section 2.3. A preliminary run of the Gibbs sampler with $L^* = 8$ and $K^* = 30$ indicated that running Algorithm 1 with $L = 2$ (the posterior mode of L) and $K = 30$ (the largest posterior mode for K was equal to 21) was sufficient to perform imputations. We set the hyperparameter priors $\alpha_{ltr} = \alpha_{lksu} = 0.05$ for each l, k, t, s, r, u, and the prior hyperparameters for the mixture weights $\alpha_l = 1500$ for each l at level-2 and $\alpha_{lk} = 50$ for each l, k at level-1. [11] $m = 100$ imputations were performed across 25000 iterations after a burn-in period of $b = 5000$ iterations, for a total of $I = 30000$ iterations.

Outcomes. We applied the considered methods (LD and BMLC), and evaluated bias, standard errors and p-values of the final estimates, and compared them with the Complete Data case.

4.2 Study Results

Table 5 show the results of the experiment. From the table, it is possible to observe how the BMLC method led to final parameter estimates very close to the Complete Data case. Only two coefficients (β_{00} and γ_1) were slightly off the Complete Data case value. The LD method tended to retrieve slightly more biased estimates (in particular β_{00},β_1 and γ_1), but overall the retrieved values with such technique were acceptable. In columns 5-7 of the table standard errors of the estimates are reported. The standard errors obtained with the LD method were larger than the ones yielded by the BMLC imputation model, as a consequence of a smaller sample size. On the other hand, the BMLC imputation model could exploit the full sample size, and retrieved standard errors very close to the Complete Data Case. The effect of the smaller standard errors obtained with the BMLC imputation model can be observed in the last three

[11] Given the large number of level-1 units within each country, multiple starts were needed in order to obtain the Gibbs sampler with the desired number of fully allocated LCs.

columns of Table 5, reporting the p-values of the fixed effects: the fixed effects estimated through the BMLC imputation resulted all significant ($p < 0.05$), as they were supposed to be. The LD technique, on the other hand, produced some non-significant coefficients (β_1, β_3 and γ_{11}), showing how this method, unlike MI, could lead to loss of power in statistical tests.

With respect to the variance of the random components (reported in the bottom of Table 5), the Complete Data case and the BMLC imputation method yielded roughly similar values of τ_0^2 and τ_1^2. Conversely, the LD method led to an overly large estimate of the random intercept τ_0^2.

5 Discussion

In this paper we proposed the use of Bayesian Multilevel Latent Class (BMLC) models for the MI of multilevel categorical data. After presenting the model and its configurations in Section 2, we performed two studies in order to assess its performance under different conditions.

In Study 1, a simulation study with two sample size conditions was carried out in which the BMLC imputation method was compared to the LD method (still one of the most applied techniques in the presence of multilevel missing data according to Van Buuren, 2011) and the JOMO technique, one of the few available routines which allow for the MI of multilevel categorical data. The analysis model used was a random intercept logistic model. In Study 2, data coming from the ESS survey were used to investigate the behavior of the BMLC model with real-case data, and to compare it with the LD method results. In this second study, the analysis model was a multilevel logistic model with random intercept and slope.

Overall, the BMLC model showed a good performance in terms of bias, stability of the estimates and coverage rates of the coefficient intervals of the final estimates. Unlike the LD and the JOMO methods, which had limitations either because of a too small sample size used (LD) or because of a too simple imputation model (JOMO), the BMLC model offers a flexible imputation technique, able to pick up complex orders of associations among the items of the dataset at both levels, returning unbiased and stable parameter estimates of the analysis model. This imputation model can be a useful tool for applied researchers that need to deal with multilevel categorical data (e.g., coming from surveys) that are only partially observed, since it can help to recover potentially valuable information that could be lost if the subjects with missingness were simply discarded, as the results coming from the LD method have shown in both Study 1 and Study 2 of this paper.

Despite the proven utility of the BMLC imputation model, some issues still need to be better crystallized by further studies. First, in the current article we gave a rough rule of thumb for both model and prior specification. Future research should focus on how these criteria can be refined. Second, the behavior of the Gibbs sampler seemed to be unstable when dealing with the BMLC model with a small number of level-2 units and a large number of level-1 units, and very sensitive to the choice of the pseudo-counts

entered as hyperparameters. A better understanding of the functioning of the Gibbs sampler when run with the BMLC model can help improving the allocation of units across the LCs, the quality of the imputations and the computing time needed to run the algorithm. Nevertheless, the behavior of the sampler can easily be checked via MCMC output (e.g., by checking the number of non-empty classes at the end of Gibbs sampling), and multiple starts can be performed until the desired number of classes is fully occupied across its iterations.

Finally, the proposed approach can be extended in various meaningful ways. First, the BMLC model can be also applied to longitudinal data, in which multiple observations in time (level-1 units) are nested within individuals (level-2 units). If the level-1 observations within the same subject are independent with each other, but depend on a (discrete) time indicator, it suffices to include the latter in the BMLC model as level-1 indicator and perform the imputations. Second, while we dealt with multilevel categorical data, the BMLC model can also be applied to continuous or mixed type of data. In addition, the model can be easily extended to deal with three or more levels of hierarchy. This can be the case, for instance, when a sample of students (level-1) is drawn from a sample of schools (level-2) which, in turn, is drawn from a sample of countries (level-3).

References

Allison, P. D. (2009). Missing data. *The SAGE Handbook of Quantitative Methods in Psychology, 4*, 72-89.

Andridge, R. (2011). Quantifying the impact of fixed effects modeling of clusters in multiple imputation for cluster randomized trials. *Biometrical Journal, 53*, 57-74.

Carpenter, R., & Kenward, M. (2013). *Multiple imputation and its application*. New York: John Wiley & Sons.: Wiley.

Drechsler, J. (2015). Multiple imputation of multilevel missing data - Rigor versus simplicity. *Journal of Educational and Behavioral Statistics, 40*, 69-95.

Gelfand, A. E., & Smith, A. F. M. (1990). Sampling based approaches to calculating marginal densities. *Journal of the American Statistical Association, 85*, 398-409.

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

Goldstein, H., Carpenter, J. R., Kenward, M., & Levin, K. (2009). Multilevel models with multivariate mixed response types. *Statistical modelling, 9*, 173-197.

Horton, N. J., Lipsitz, S., & Parzen, M. (2003). A potential for bias when rounding in Multiple Imputation. *American Statistician, 57*, 229-232.

NSD: Norwegian Centre for Research Data. (2012). *ESS Round 6: European Social Survey Round 6 Data.* Data file edition 2.2. Norway: Data Archive and distributor of ESS data for ESS ERIC.

Quartagno, M., & Carpenter, J. (2016). `jomo`: A package for multilevel joint modelling multiple imputation [Computer software manual]. Retrieved from `https://CRAN.R-project.org/package=jomo`

Reiter, P., Raghunathan, T. E., & Kinney, S. (2006). The importance of modeling the survey design in multiple imputation for missing data. *Survey Methodology,32*, 143-150.

Rousseau, J., & Mergensen, K. (2011). Asymptotic behaviour of the posterior distribution in overfitted mixture models. *Journal of the Royal Statistical Society: Series B, 73*, 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*, 147-177.

Schafer, J. L., & Yucel, R. M. (2002). Computational strategies for multivariate linear mixed-effects models with missing values. *Journal of computational and graphical statistics, 11*, 437-457.

Si, Y., & Reiter, J. P. (2013). Nonparametric Bayesian Multiple Imputation for Incomplete Categorical Variables in Large-Scale Assessment Surveys. *Journal of Educational and Behavioral Statistics, 38*, 499-521.

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

Van Buuren, S. (2011). Multiple imputation of multilevel data. In Eds, Hox J, J. & Roberts J, K. *The Handbook of Advanced Multilevel Analysis(10)*, pp. 173-196. Routledge, Milton Park, UK.

Van Buuren, S. (2012). *Flexible imputation of missing data.* Boca Raton, FL.: Chapman & Hall/CRC.

Van Buuren, S., Brand, J. P. L., Groothuis-Oudshoorn, C. G. M., & Rubin, D. B. (2006). Fully conditional specification in multivariate imputation. *Journal of Statistical Computation and Simulation, 76*, 1049-1064.

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.

Vermunt, J. K. (2003). Multilevel latent class models. *Social methodology, 33*, 213-239.

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*, 369-397.

Vidotto, D., Vermunt, J. K., & Van Deun, K. (submitted). Bayesian latent class Models for the multiple imputation of categorical data.

Yucel, R. (2008). Multiple imputation for multilevel continuous data. *Philosophical transactions of the*

Royal Society, A, 2, 2389-2403.

Zhao, J. H., & Schafer, J. L. (2016). `pan`: Multiple imputation for multivariate panel or clustered data [Computer software manual]. (R package version 1.4)