Bayesian latent class models for the multiple imputation of categorical data

Vidotto, Davide; Vermunt, Jeroen K.; Van Deun, Katrijn

Published in:
Methodology: European Journal of Research Methods for the Behavioral and Social Sciences

DOI:
10.1027/1614-2241/a000146

Publication date:
2018

Document Version
Publisher's PDF, also known as Version of record

Original Article

Bayesian Latent Class Models for the Multiple Imputation of Categorical Data

Davide Vidotto, Jeroen K. Vermunt, and Katrijn Van Deun

Department of Methodology and Statistics, Tilburg University, The Netherlands

Abstract: Latent class analysis has been recently proposed for the multiple imputation (MI) of missing categorical data, using either a standard frequentist approach or a nonparametric Bayesian model called Dirichlet process mixture of multinomial distributions (DPMM). The main advantage of using a latent class model for multiple imputation is that it is very flexible in the sense that it can capture complex relationships in the data given that the number of latent classes is large enough. However, the two existing approaches also have certain disadvantages. The frequentist approach is computationally demanding because it requires estimating many LC models: first models with different number of classes should be estimated to determine the required number of classes and subsequently the selected model is reestimated for multiple bootstrap samples to take into account parameter uncertainty during the imputation stage. Whereas the Bayesian Dirichlet process models perform the model selection and the handling of the parameter uncertainty automatically, the disadvantage of this method is that it tends to use a too small number of clusters during the Gibbs sampling, leading to an underfitting model yielding invalid imputations. In this paper, we propose an alternative approach which combined the strengths of the two existing approaches; that is, we use the Bayesian standard latent class model as an imputation model. We show how model selection can be performed prior to the imputation step using a single run of the Gibbs sampler and, moreover, show how underfitting is prevented by using large values for the hyperparameters of the mixture weights. The results of two simulation studies and one real-data study indicate that with a proper setting of the prior distributions, the Bayesian latent class model yields valid imputations and outperforms competing methods.

Keywords: Bayesian mixture models, latent class models, missing data, multiple imputation

Multiple imputation (MI; Rubin, 1987) is a powerful technique to deal with the problem of missing data in a dataset. Unlike other missing data procedures, it allows for separating the missing data handling step and the substantive analysis step under the assumption that data are *missing at random* (MAR). In MI, to account for the uncertainty about the imputations, the original incomplete dataset is replaced by multiple ($m > 1$) complete datasets, in each of which the missing values are replaced by different sets of random values generated from an imputation model. In the substantive analysis, each of the m datasets is analyzed separately and m results are pooled through Rubin's (1987) rules. This yields point estimates of the parameters of interest, such as regression coefficients, along with their standard errors, which also reflect the uncertainty due to the presence of missing data (Allison, 2009; Little & Rubin, 2002; Schafer & Graham, 2002). In order for MI to work well, the imputation model should preserve the important relationships between the variables of interest, which can be simple bivariate associations but also higher-order interactions.

While methods for continuous missing data have been extensively researched in the past, methods to handle nonresponse in categorical variables have not been fully established yet. During the past years, the literature has considered log-linear models (Schafer, 1997) and MI by chained equations (MICE; Van Buuren & Groothuis-Oudshoorn, 2000). The former has the advantage of being able to describe complex associations in the data (through the saturated model), but it can only handle a limited number of variables. MICE can also be used when the number of categorical variables with missing values is large, but since this requires estimating a large number of binary and/or multinomial logistic models, model selection and specification can become a cumbersome task, especially if complex relationships requiring higher-order interactions should be preserved by the imputation model (Si & Reiter, 2013; Vermunt, Van Ginkel, Van der Ark, & Sijtsma, 2008).

Vermunt et al. (2008) proposed using a *frequentist latent class* (FLC), or finite mixture, model for the MI of categorical data. LC models overcome the difficulties encountered

https://doi.org/10.1027/1614-2241/a000146

with log-linear models and chained equations. Firstly, the model specification only requires specifying the number of latent classes (or mixture components) K. When K is set large enough, LC models can estimate the joint distribution of the data and automatically capture important associations among the variables at hand (Vermunt et al., 2008). Secondly, the particular form of the model and the *local independence* assumption offer easy computation even with a large number of variables. Furthermore, Vermunt et al. (2008) showed by means of a simulation study that MI via FLC modeling yields correct parameter estimates of the substantive model. With the FLC model, the uncertainty about the imputation model parameters is accounted for by bootstrapping. Using a similar model but with a Bayesian nonparametric approach, Si and Reiter (2013) introduced imputation of categorical data with the *Dirichlet Process Mixture of Multinomial Distributions* (DPMM). While the DPMM assumes a (theoretically) infinite number of mixture components, in practice an arbitrarily large number of clusters is selected during the Gibbs sampling iterations (Gelfand & Smith, 1990) to perform the actual imputations.

Albeit appealing, both the FLC and the DPMM models have certain disadvantages. The former requires multiple, sequential runs of the *expectation-maximization* (EM) algorithm, first for determining the number of classes using a model selection criterion like the *Akaike information criteria* (AIC), and subsequently for obtaining the m imputations, which involves reestimating the selected FLC imputation model using m bootstrap samples. Hence, imputing with the frequentist model can be time-consuming, especially for large datasets when various models with large numbers of classes have to be compared and/or when a large number of imputations has to be performed. The DPMM overcomes these problems by performing the selection of the number of classes and the actual imputations as part of a single run of the Gibbs sampling procedure. However, this method is prone to data underfitting; that is, relevant associations in the data may not be picked up because not all the necessary LCs get filled during the Gibbs sampling. This can be deleterious for the resulting imputations: Vermunt et al. (2008) observed that underfitting in MI is undesirable, because it causes the imputation model to disregard important relationships in the data, leading to biased and inaccurate final inferences. On the other hand, overfitting is of small concern, since picking up particular features which are sample specific does not introduce bias in the final imputations.

In the current paper, we propose performing MI using a Bayesian LC (BLC) model, which overcomes the disadvantages of the FLC and the DPMM approaches. One of the new features of our approach is that the number of classes needed for the imputation model is determined using a 0single, preliminary run of the Gibbs sampler in which a model is used with a large number of classes and with prior distributions that favor the emptying of extra components. The m imputations can subsequently be obtained in a second run, in which the number of LCs is fixed at the value determined in the first stage. A second special feature of our approach is that the prior distribution of the mixture weights are set in such a way that the units are allocated across all the LCs during the Gibbs sampler, helping the BLC model to prevent underfitting, and leading to more accurate imputations than the DPMM.

The outline of the remainder of this paper is as follows. In the Bayesian Latent Class Imputation section, the BLC model for the MI of categorical data is introduced, along with its estimation and set-up. The Simulation Studies section describes two simulation studies which compare the BLC model with different prior specifications, as well as with the DPMM, FLC, and MICE approaches. The Real-Data Study section reports the results of a real-data experiment. The Discussion section concludes with final remarks by the authors.

Bayesian Latent Class Imputation

Bayesian imputations are derived from the posterior predictive distribution of the missing data given the observed data, that is, $\Pr(Y_{\text{mis}}|Y_{\text{obs}}) = \int \Pr(Y_{\text{mis}}|\pi)\Pr(\pi|Y_{\text{obs}})\text{d}\pi$, in which π is the model parameter vector. Thus, imputations are performed by first drawing m values from the posterior distribution of the model parameter $\Pr(\pi|Y_{\text{obs}})$, and then by sampling from the predictive distribution $\Pr(Y_{\text{mis}}|\pi^{*(l)})$, $l = 1, \ldots, m$. The posterior $\Pr(\pi|Y_{\text{obs}})$ is estimated via Gibbs sampling and derived from two quantities: a probabilistic model for the data (the likelihood) and a prior distribution for π.

The Data Model

Let y_i be a vector of length J, denoting the observed response pattern for unit i ($i = 1, \ldots, n$) on J categorical variables, so that $y_{ij} = s$ is unit i's value on the jth variable ($j = 1, \ldots, J$; $s = 1, \ldots, s_j$). Furthermore, let $x_i = k$ be a realization of the latent categorical variable X for person i, taking on one of the possible values $k \in \{1, \ldots, K\}$. The latent class (LC) model (Goodman, 1974; Lazarsfeld, 1950) describes the joint distribution of the observed variables $(Y_1, \ldots, Y_J)$ through the well known form

$$\Pr(y_i) = \sum_{k=1}^{K} \Pr(x_i = k) \prod_{j=1}^{J} \Pr\left(y_{ij} = s|x_i = k\right),$$

in which the $\Pr(x_i = k)$ are the latent class weights and the $\Pr(y_{ij} = s|x_i = k)$ are the conditional response probabilities. By assuming a Multinomial distribution for both X and $Y_j|X$, with parameters denoted by $\Pr(x_i = k) = \pi_k$ and $\Pr(y_{ij} = s|x_i = k) = \pi_{kjs}$, respectively, the model can be rewritten in terms of the Multinomial parameters as

$$\Pr(y_i; \pi) = \sum_{k=1}^{K} \pi_k \prod_{j=1}^{J} \prod_{s=1}^{s_j} (\pi_{kjs})^{I_{ijs}}, \quad (1)$$

where I_{ijs} is an indicator variable equal to 1 when $y_{ij} = s$ and zero otherwise. Below, we will use the symbols π_x and π_{kj} to refer to the two sets of model parameters, that is, $\pi_x = (\pi_1, \ldots, \pi_K)$ and $\pi_{kj} = (\pi_{kj1}, \ldots, \pi_{kjs_j})$, while $\pi = (\pi_x, \pi_{11}, \ldots, \pi_{KJ})$.

With a sufficiently large number of classes, the LC model can capture the first- and higher-order moments of the joint distribution of the J categorical variables (McLachlan & Peel, 2000). The resulting density is a weighted average (i.e., a mixture) of class-specific Multinomial densities, where the probabilities π_k act as weights. Furthermore, the *local independence* assumption makes the conditional density $\Pr(Y_j|X = k)$ independent of the other response variables given the kth latent class. As a result, the estimation of a LC model involves processing J two-way K-by-s_j tables, instead of the full multi-way table involving all J variables (as done by, e.g., the log-linear model). For this reason, especially when the number of variables is large, the LC model is computationally appealing for MI. Details about MI through FLC models can be found in Vermunt et al. (2008).

The Prior Distributions

Model (1) can be turned into a Bayesian LC (BLC) model by placing prior distributions upon the latent class proportions π_x and the conditional response probabilities π_{kj}. A common choice conjugate to the Multinomial distribution is the Dirichlet prior. Therefore, we will assume that

$$\pi_x \sim \mathrm{Dir}(\alpha_x)$$

and

$$\pi_{kj} \sim \mathrm{Dir}(\alpha_{kj})$$

$\forall\, k, j$. Here the vectors α_x (from here on referred to as the *latent hyperparameter*) and α_{kj} (from here on referred to as *conditional hyperparameter*) are defined as

$$\alpha_x = (\alpha_1, \ldots, \alpha_k, \ldots, \alpha_K)$$

and

$$\alpha_{kj} = \left(\alpha_{kj1}, \ldots, \alpha_{kjs}, \ldots, \alpha_{kjs_j}\right),$$

with $\alpha_k > 0$ and $\alpha_{kjs} > 0\ \forall\ k, j, s$.

The most common setting is to use a single value for the hyperparameters α, yielding symmetric Dirichlet distributions with constant α values; that is, $\alpha_x = (c_1, \ldots, c_1)$ and $\alpha_{kj} = (c_2, \ldots, c_2)$. Below, we will use the fact that the magnitude of c_1 parameters affects the shape of the posterior class distribution: the larger c_1 the more the observations will tend to be evenly distributed across all latent classes, while with c_1 close to zero only some of the classes will have a nonnegligible posterior probability mass.

BLC Model Estimation and Imputation

Model estimation is performed via a Gibbs sampling algorithm. In our implementation, we separate the Gibbs sampling of the LC model parameters from the imputation of the missing values. That is, we first run the Gibbs sampler for a certain number of iterations and store m sets of parameters from iterations which are spaced enough to prevent autocorrelations among the draws. Subsequently, m imputed datasets are created using these m sets of stored parameters. An alternative would be to impute the missing values as a part of the Gibbs sampling iterations, and base the *posterior class membership probabilities* used in the Gibbs sampler on both the observed and the imputed values rather than on the observed part of the data only. Our implementation is computationally more efficient, because there is no need to update the missing data at each iteration, nor to take imputed values into account when the *posterior membership probabilities* of Step 1 are calculated (e.g., Si & Reiter, 2013).

Here, we assume that both the number of classes K and the hyperparameter values have been previously chosen. The next section discusses how to perform these choices. The parameters of both the latent variable X and the conditional distributions of the jth item given the kth latent class, $Y_j|X = k$, can be initialized through random draws from uniform Dirichlet distributions: $\pi_x^0 \sim \mathrm{Dir}(1, \ldots, 1)$ and $\pi_{kj}^0 \sim \mathrm{Dir}(1, \ldots, 1)\ \forall\ k, j$, in order to increase the likelihood of initializing the sampler from the interior of the parameter space. The total number of iterations (T) depends on the number of burn-in iterations (b), the number draws used for the imputations (m), and the spacing between these m draws (d); that is, $T = b + d \cdot m$. The value of b should be large enough to ensure convergence of the chain to its equilibrium distribution $\Pr(\pi|Y_{\mathrm{obs}})$. Since a BLC imputation model may consist of a large number of parameters and since the quantity of interest in MI is the likelihood $\Pr(Y_{\mathrm{obs}}|\pi)$, convergence is assessed by inspecting the trace plot of the log-likelihood function calculated at each iteration, as suggested by Schafer (1997).

The Gibbs sampler proceeds as follows, for $t = 1, \ldots, T$:

Algorithm 1:

1. Sample $x_i^{(t)} \in \{1, \ldots, K\} \, \forall \, \mathrm{i} = 1, \ldots, n$ from the Multinomial distribution with the *posterior membership probabilities* as parameters, defined as

$$\Pr(x_i^{(t)} = k | Y_{\text{obs}}, \pi^{(t-1)}) = \frac{\pi_k^{(t-1)} \prod_{j=1}^{J} \left(\prod_{s=1}^{s_j} \left(\pi_{kjs}^{(t-1)} \right)^{I_{ijs}^*} \right)}{\sum_{h=1}^{K} \pi_h^{(t-1)} \prod_{j=1}^{J} \left(\prod_{s=1}^{s_j} \left(\pi_{kjs}^{(t-1)} \right)^{I_{ijs}^*} \right)},$$

   in which I_{ijs}^* equals 1 when $y_{ij} = s$ and $y_{ij} \in Y_{\text{obs}}$, and zero otherwise;
2. Sample

$$\left(\pi_x^{(t)} Y_{\text{obs}}, x^{(t)}, \alpha_x\right) \sim$$

$$\text{Dir}\left(\alpha_1 + \sum_{i=1}^{n} I\left(x_i^{(t)} = 1\right), \cdots, \alpha_K + \sum_{i=1}^{n} I\left(x_i^{(t)} = K\right) \right)$$

   where $I(x_i^{(t)} = k)$ is equal to 1 if $x_i^{(t)} = k$ and zero elsewhere;
3. Draw

$$\left(\pi_{kj}^{(t)} Y_{\text{obs}}, x^{(t)}, \alpha_{kj}\right) \sim$$

$$\text{Dir}\left(\alpha_{kj1} + \sum_{i \sim x_i^{(t)} = \mathrm{k}} I_{ij1}^*, \cdots, \alpha_{kjs_j} + \sum_{i \sim x_i^{(t)} = k} I_{ijs_j}^* \right), \quad \forall \, k, j.$$

After ruling out the first b iterations for the burn-in, the BLC model is estimated with the remaining $d \cdot m$ iterations, which are draws from the conditional distribution $\Pr(\pi | Y_{\text{obs}})$. For the imputations, at each dth iteration we store the sampled parameters and class memberships, yielding $\pi^{*(1)}, \ldots, \pi^{*(m)}$ from $\Pr(\pi | Y_{\text{obs}})$ and $x_i^{(1)}, \ldots, x_i^{(m)}$. The imputed values are subsequently drawn from the posterior predictive distribution of the missing data, denoted by $\Pr(Y_{\text{mis}}^{*(l)} | Y_{\text{obs}}, \pi^{*(l)})$, $l = 1, \ldots, m$. These simulated values will be then entered in the blank part of the original incomplete dataset, replicated m times. Formally:

4. *Imputation step*: with each of m parameter sets selected for the imputations, $l = 1, \ldots, m$, given the sampled value $x_i^{(l)} = k$ of each unit, and for each $\{i, j\} \in Y_{\text{mis}}$, sample from

$$\left(Y_{ij} | Y_{\text{obs}}, \pi^{(l)}, x_i^{(l)} = k\right) \sim \text{Multinom}\left(\pi_{kj}^{*(l)}\right)$$

and store the imputed values.

In the experiments described in sections Simulation Studies and Real-Data Study, Algorithm (1) is run with a routine we implemented in R, which is available upon request from the first author.

Setting Up the Model

Model Selection: Number of Classes

For Bayesian finite mixture models, Gelman, Carlin, Stern, and Rubin (2013; chapter 22) proposed performing model selection by resorting to a computational expedient. In particular, they noticed that by starting with arbitrarily large K and latent hyperparameters supporting the occurrence of empty components while the Gibbs sampler is running, it is possible to obtain a posterior distribution for the number of clusters by counting the number of classes filled at each iteration of Algorithm 1 (without Step 4). A possible value for the latent hyperparameter that encourages the realization of empty components is given by $\alpha_k = 1/K \; \forall \; k$, which as indicated by, Gelman et al. (2013) is insensitive to the choice of the starting K. Hence, their approach consists of two main steps: (1) preliminarily run the Gibbs sampler (Steps 1–3 of Algorithm 1) and obtain the posterior distribution of $K|Y_{\text{obs}}$; (2) set K equal to the posterior mode of this distribution, and re-run the Gibbs sampler with this value of K to perform inference. Whereas setting the number of classes equal to the posterior mode is a logical choice in a substantive LC analysis (i.e., for model interpretation), in MI a number of components larger than the one used for substantive analysis are usually required (Vermunt et al., 2008). Therefore, we suggest using the posterior maximum of the distribution of $K|Y_{\text{obs}}$, that is, the largest K^* such that $\Pr(K = K^*|Y_{\text{obs}}) > 0$. Afterwards, it is possible to perform the imputations (Algorithm 1 including Step 4) with a second run of the Gibbs sampler, with K selected at the previous stage and a latent hyperparameter that supports the allocation of the units across all the mixture components (see below). In the experiments of Simulation Studies and Real-Data Study sections, this model selection method was tested for the BLC model, as well as for the FLC imputation model to assess whether this is a good and fast alternative for the model selection step of the FLC model.

Hyperparameter Selection

Latent Hyperparameter

Hoijtink and Notenboom (2004) noticed that when standard priors (e.g., the uniform prior) for the latent weights are used, the probability of obtaining empty classes increases with K. In these situations, sampling from the true posterior becomes difficult for the Gibbs sampler, since the (conditional distribution) parameters of the empty components are fully determined by their prior distributions, making the Gibbs sampler unstable.

As mentioned in the previous section, the assumed prior distribution for the mixture weights strongly affects the shape of the posterior when the Gibbs sampler is run with a large number of classes. In particular, α_x can be set in such a way that all the specified LCs are filled during the Gibbs sampler iterations. Rousseau and Mergensen (2011) showed that, when an overfitting mixture model is estimated with $\max(\alpha_1, \ldots, \alpha_K) < p/2$, where p is the number of free parameters to be estimated within each mixture component,[1] the latent proportions of the extra classes will approach zero, while with $\min(\alpha_1, \ldots, \alpha_K) > p/2$, the possibly redundant classes will be given a nonnegligible weight. The larger the value of α_k is, the larger the number of filled LCs will be. Obtaining full allocation of the components is desirable, because in this way the Gibbs sampler avoids to sample from the prior distribution of the empty components parameters, making the composition of the clusters fully determined by the data. The Markov Chain Monte Carlo (MCMC) output can be used to assess whether all the LCs have been filled during the Gibbs sampling: if this is not the case, then we suggest making $\alpha_k \; \forall \; k$ more informative by increasing its value (while maintaining a symmetric Dirichlet distribution) until full allocation is achieved.

Conditional Hyperparameter

In MI, the aim is to obtain imputations which resemble as much as possible the observed data, implying that the prior distributions should be dominated by the data likelihood (Schafer & Graham, 2002). For the conditional response probabilities, Si and Reiter (2013) proposed setting uniform priors for all variables and mixture components, that is, $\alpha_{kj} = (1, \ldots, 1) \; \forall \; k, j$. However, as will be shown in section Simulation Studies, this may still be too informative, leading to invalid imputations. Note that using such uniform priors for the conditional response probabilities is equivalent to adding $K \cdot s_j$ observations for each variable (see Step 3 of Algorithm 1). To prevent having too informative priors for this part of the model, we suggest making the conditional hyperparameters less influential by decreasing their values and setting them as low as $\alpha_{kjs} = 0.01$ or $0.05 \; \forall \; k, j, s$.[2]

Simulation Studies

Here we report the results of two simulation studies. In both studies, the performance of our method is compared to that of FLC, DPMM, and MICE. Study 1 concerns a situation with a large sample size and a small number of variables while Study 2 is based on data with a smaller sample size and a large number of variables. All analyses were performed with R version 3.3.0.

Study 1

Study Design

Population Model

The population model was specified for five predictor variables $Y_1, \ldots, Y_5$ and one outcome variable Y_6, all of which were trichotomous (coded with 0, 1, and 2). The relationships between the predictors were described by the log-linear model

$$\begin{aligned}\log \Pr(Y_1, Y_2, Y_3, Y_4, Y_5) \propto & \\ -0.5 \sum_{j=1}^{5} Y_j - \sum_{j=1}^{4} \sum_{k=j+1}^{5} Y_j Y_k - 0.2 Y_1 Y_3 Y_5 & \\ + 0.5 Y_2 Y_4 Y_5. & \qquad (2)\end{aligned}$$

Subsequently, the outcome was generated from a multinomial logistic model, defined for $\Pr(Y_6 = r|Y_1, \ldots, Y_5)$ ($r = \{1,2\}$), whose probabilities were specified through

$$\begin{aligned}\log(\Pr(Y_6 = 1)/\Pr(Y_6 = 0)) = & \\ -0.1 + Y_1 + \beta_{1,2} Y_2 + \beta_{1,3} Y_3 - 0.6 Y_4 & \\ + 0.5 Y_5 + \beta_{1,25} Y_2 Y_5 + \beta_{1,34} Y_3 Y_4 & \\ \log(\Pr(Y_6 = 2)/\Pr(Y_6 = 0)) = & \\ -0.6 + 1.8 Y_1 + \beta_{2,2} Y_2 + \beta_{2,3} Y_3 + Y_4 & \\ - 0.5 Y_5 + \beta_{2,25} Y_2 Y_5 + \beta_{2,34} Y_3 Y_4, & \qquad (3)\end{aligned}$$

where, as can be seen, the reference category is $Y_6 = 0$. The values of the β parameters are reported in Table 1. Based on models (2) and (3), we generated $N = 500$ datasets with $n = 5{,}000$ observations each.

Introducing Missingness

A low and a high missingness condition was created by introducing missing values in Y_2 and Y_3 according to MAR mechanisms. The total rate of missingness for both Y_2 and Y_3 was around 10% and 20% for the low and high missingness condition, respectively. Table 2 shows how the probability of a missing value depends on Y_1 and Y_4 for Y_2, and on Y_5 and Y_6 for Y_3.

[1] In LC models, the number of free parameters within each components is given by $p = \sum_j s_{j-1}$.
[2] This is equivalent to entering $0.01Ks_j$ or $0.05Ks_j$ imaginary observations for each variable.

Table 1. Parameter values under investigation in Study 1

Parameter	$\beta_{1,2}$	$\beta_{1,3}$	$\beta_{1,25}$	$\beta_{1,34}$	$\beta_{2,2}$	$\beta_{2,3}$	$\beta_{2,25}$	$\beta_{2,34}$
Value	−1.7	1.5	−0.25	0.1	−1.25	1.0	−0.5	0.2

Table 2. MAR mechanisms used in Study 1: The table reports the probability of nonresponses in Y_2 for each combination of Y_1,Y_4 and in Y_3 for each combination of Y_5,Y_6

Missingness rate	Y_1,Y_4	Pr(Y_2 is missing)	Y_5,Y_6	Pr(Y_3 is missing)
Low	0,0	.100	0,0	.125
	0,1	.025	0,1	.075
	0,2	.125	0,2	.100
	1,0	.150	1,0	.100
	1,1	.075	1,1	.150
	1,2	.050	1,2	.175
	2,0	.125	2,0	.150
	2,1	.200	2,1	.050
	2,2	.150	2,2	.125
Large	0,0	.200	0,0	.250
	0,1	.050	0,1	.150
	0,2	.250	0,2	.200
	1,0	.300	1,0	.200
	1,1	.150	1,1	.300
	1,2	.100	1,2	.350
	2,0	.250	2,0	.300
	2,1	.400	2,1	.100
	2,2	.300	2,2	.250

Settings of the Imputation Models

For all the imputation models, we performed $m = 20$ imputations. For the BLC and the FLC models, we performed model selection with the Gelman et al.'s (2013) method exposed in section Model Selection: Number of Classes. In particular, for each simulated datasets we ran Steps 1–3 of Algorithm 1 with 20 components for T = 3,000 iterations, of which b = 1,000 served as burn-in. The remaining 2,000 iterations were used to determine the distribution of the number of LCs. This led to an average (maximum) number of classes equal to $\bar{K} = 15.94$ in the low missingness condition and to $\bar{K} = 15.41$ in the high missingness condition. The FLC imputation model was run with LatentGOLD 5.1 (Vermunt & Magidson, 2013) with the settings given in Vermunt et al. (2008). We imputed the data with the BLC model using different prior specifications. In particular, we manipulated α_k to be equal to 1 and to 20 (we found out that α_k = 20 was sufficiently large to ensure full allocation of the units across all the LCs), and α_{kjs} to be equal either to 1 or to 0.01. The BLC models we used will be denoted with BLC(α_k, α_{kjs}); for instance, BLC(1,1) indicates the BLC model with uniform priors for both the latent proportions and the conditional response probabilities. We ran the DPMM model with K = 20 and hyperparameters of the Dirichlet Process prior set as in Si and Reiter (2013); α_{kjs} was handled as done for the BLC model. Therefore, we will denote the two DPMM models we implemented with DPMM(1) and DPMM(.01). The Gibbs sampler for both the BLC and the DPMM methods were run with self-implemented routines,[3] with T = 5,000 total and b = 1,000 burn-in iterations. Lastly, the MICE method was run with its standard settings and with 20 iterations for each imputation[4] using the mice library (Van Buuren et al., 2014).

Outcomes

After applying the imputation models, estimating model (3) on each imputed dataset, and applying the pooling rules for MI, we compared relative bias, stability (i.e., the standard deviation of the estimates across the 500 replications), and coverage rates of the 95% confidence intervals of the MI estimates. In particular, we considered the estimates of the parameters reported in Table 1: these parameters correspond to the main and interaction effects of the variables with missing values (Y_2 and Y_3).

Results

Tables 3 and 4 show the results for the Low and High missingness condition, respectively.

Low Missingness Condition

In the first condition, the largest bias was observed for the two interaction terms $\beta_{1,25}$ (MICE) and $\beta_{1,34}$ (MICE, FLC, BLC(1,1), BLC(20,1), DPMM(1)). The interaction term $\beta_{2,34}$ recovered by BLC(1,1) and DPMM(1) was also biased. Parameter estimates produced by all the LC methods tended to be similar in terms of stability, but the most stable parameter estimates were provided by MICE. The coverage rate of the 95% confidence intervals was close to the nominal level for all the parameters estimated after processing the data with any of the considered imputation methods, except for the confidence intervals of the main effects $\beta_{1,2}$ and $\beta_{1,3}$ produced by MICE, which were too short.

[3] We implemented the DPMM model as described in Si and Reiter (2013).

[4] MICE produces m imputations by starting from m different (independently drawn) values for the missing data. Subsequently, the imputation model parameters and the missing data are iteratively updated in parallel for a number of specified iterations. Following Van Buuren, Brand, Groothuis-Oudshoorn, and Rubin (2006), to reach convergence the number of iterations does not need to be large, and we decided to set it equal to 20.

Table 3. Relative bias, stability, and coverage rate observed for the estimates of eight multinomial logistic model parameters in model (3) after applying three different imputation models

	Low missingness condition							
	Parameter							
Method	$\beta_{1,2}$	$\beta_{1,3}$	$\beta_{1,25}$	$\beta_{1,34}$	$\beta_{2,2}$	$\beta_{2,3}$	$\beta_{2,25}$	$\beta_{2,34}$
Relative bias								
MICE	−0.06	−0.09	**−0.22**	**0.22**	0.02	−0.06	−0.04	0.03
FLC	0.00	0.01	−0.02	**0.22**	0.01	0.01	0.02	0.06
BLC(1,1)	0.00	0.00	−0.08	**−0.21**	0.01	0.00	−0.11	**−0.18**
BLC(20,1)	0.00	0.00	−0.07	**−0.20**	0.01	−0.01	−0.09	−0.15
BLC(1,.01)	0.00	0.00	−0.04	−0.03	0.01	0.00	−0.05	−0.08
BLC(20,.01)	0.00	0.00	−0.02	0.05	0.00	0.00	−0.02	−0.02
DPMM(1)	0.00	0.00	−0.10	**−0.52**	0.02	0.00	−0.14	**−0.40**
DPMM(.01)	0.00	0.00	−0.04	−0.06	0.01	0.00	−0.06	−0.09
Stability								
MICE	0.09	0.08	0.11	0.16	0.08	0.10	0.19	0.15
FLC	0.10	0.10	0.13	0.19	0.08	0.11	0.20	0.17
BLC(1,1)	0.10	0.10	0.13	0.18	0.08	0.11	0.18	0.16
BLC(20,1)	0.10	0.10	0.13	0.18	0.08	0.11	0.18	0.16
BLC(1,.01)	0.10	0.10	0.13	0.19	0.08	0.11	0.19	0.17
BLC(20,.01)	0.10	0.10	0.13	0.19	0.08	0.11	0.19	0.17
DPMM(1)	0.10	0.10	0.13	0.17	0.08	0.11	0.17	0.16
DPMM(.01)	0.10	0.10	0.13	0.19	0.08	0.11	0.19	0.17
Coverage rate								
MICE	**0.82**	**0.72**	0.96	0.98	0.94	0.92	0.97	0.98
FLC	0.93	0.95	0.95	0.96	0.95	0.95	0.97	0.95
BLC(1,1)	0.94	0.96	0.95	0.97	0.95	0.95	0.95	0.96
BLC(20,1)	0.93	0.96	0.94	0.97	0.96	0.95	0.96	0.95
BLC(1,.01)	0.94	0.95	0.95	0.95	0.96	0.95	0.96	0.95
BLC(20,.01)	0.94	0.96	0.94	0.97	0.94	0.95	0.97	0.95
DPMM(1)	0.94	0.95	0.95	0.96	0.95	0.95	0.95	0.94
DPMM(.01)	0.93	0.95	0.95	0.96	0.95	0.95	0.96	0.95

Notes. MICE = MICE imputation technique; FLC = frequentist LC imputation model; BLC(1,1) = Bayesian LC imputation model with $\alpha_k = 1$, $\alpha_{kjs} = 1$; BLC(20,1) = Bayesian LC imputation model with $\alpha_k = 20$, $\alpha_{kjs} = 1$; BLC(1,.01) = Bayesian LC imputation model with $\alpha_k = 1$, $\alpha_{kjs} = .01$; BLC(20,.01) = Bayesian LC imputation model with $\alpha_k = 20$, $\alpha_{kjs} = .01$; DPMM(1) = DPMM imputation model with $\alpha_{kjs} = 1$; DPMM(.01) = DPMM imputation model with $\alpha_{kjs} = .01$. Largest values in relative bias and too low coverage rates are marked in boldface.

High Missingess Condition

With a larger rate of missingness more pronounced relative bias was observed across a larger number of estimates and for more imputation methods. All methods, with the exception of BLC(20,.01), retrieved a biased estimate of the parameter $\beta_{1,34}$. Furthermore, the interaction terms $\beta_{2,25}$ and $\beta_{2,34}$ provided by all the Bayesian LC models (excluding BLC(20,.01)) were also biased. The remaining interaction term ($\beta_{1,25}$) was correctly recovered by all methods, except for MICE and DPMM(1). As with low missingness, all LC methods retrieved similarly stable estimates, although now the BLC(1,1), BLC(20,1), and DPMM(1) models tended to produce relatively more stable estimates for some of the parameters. As in the previous condition, the confidence intervals for all parameters produced by most methods were close to their 95% nominal level. The only exceptions were the much too low coverage for the main effects $\beta_{1,2}$, $\beta_{1,3}$, and $\beta_{2,3}$ produced by MICE and the slightly too low coverage for the interaction terms $\beta_{2,25}$ and $\beta_{2,34}$ by various of the LC-based methods.

Study 2

Study Design

Population Model

In Study 2 we used $J = 21$ binary variables $Y_1, \ldots, Y_{21}$ (coded with 0 and 1), 20 predictors and 1 outcome. The first 15

Table 4. Relative bias, stability, and coverage rate observed for the estimates of eight multinomial logistic model parameters in model (3) after applying three different imputation models

	High missingness condition							
	Parameter							
Method	$\beta_{1,2}$	$\beta_{1,3}$	$\beta_{1,25}$	$\beta_{1,34}$	$\beta_{2,2}$	$\beta_{2,3}$	$\beta_{2,25}$	$\beta_{2,34}$
Relative bias								
MICE	−0.12	**−0.18**	**−0.38**	**0.34**	0.04	−0.13	−0.13	0.02
FLC	0.00	0.01	−0.02	**0.35**	0.02	0.02	−0.02	0.09
BLC(1,1)	0.01	−0.01	−0.14	**−0.56**	0.03	−0.01	**−0.28**	**−0.41**
BLC(20,1)	0.00	−0.01	−0.13	**−0.55**	0.03	−0.02	**−0.25**	**−0.37**
BLC(1,.01)	0.00	0.00	−0.05	**−0.23**	0.02	0.00	**−0.16**	**−0.23**
BLC(20,.01)	0.00	0.00	−0.02	−0.04	0.01	0.00	−0.10	−0.09
DPMM(1)	0.01	−0.01	**−0.17**	**−0.99**	0.05	−0.01	**−0.33**	**−0.75**
DPMM(.01)	0.00	0.00	−0.05	**−0.32**	0.02	0.00	**−0.17**	**−0.28**
Stability								
MICE	0.08	0.08	0.09	0.16	0.09	0.10	0.18	0.14
FLC	0.11	0.11	0.13	0.21	0.09	0.12	0.20	0.20
BLC(1,1)	0.11	0.10	0.13	0.19	0.09	0.11	0.16	0.16
BLC(20,1)	0.11	0.10	0.13	0.19	0.08	0.11	0.17	0.17
BLC(1,.01)	0.11	0.11	0.13	0.21	0.09	0.12	0.18	0.19
BLC(20,.01)	0.11	0.11	0.14	0.21	0.09	0.12	0.19	0.19
DPMM(1)	0.10	0.10	0.13	0.19	0.08	0.11	0.16	0.17
DPMM(.01)	0.11	0.11	0.13	0.20	0.09	0.12	0.19	0.18
Coverage rate								
MICE	**0.48**	**0.17**	0.96	0.98	0.93	**0.80**	0.96	0.99
FLC	0.94	0.94	0.96	0.95	0.95	0.91	0.96	0.94
BLC(1,1)	0.95	0.95	0.95	0.96	0.94	0.95	0.92	0.95
BLC(20,1)	0.95	0.96	0.96	0.96	0.96	0.96	0.92	0.95
BLC(1,.01)	0.95	0.95	0.96	0.95	0.95	0.94	0.94	0.93
BLC(20,.01)	0.93	0.95	0.96	0.95	0.94	0.94	0.95	0.95
DPMM(1)	0.95	0.95	0.96	0.92	0.93	0.96	**0.89**	**0.87**
DPMM(.01)	0.94	0.95	0.96	0.95	0.95	0.95	0.93	0.94

Notes. MICE = MICE imputation technique; FLC = frequentist LC imputation model; BLC(1,1) = Bayesian LC imputation model with $\alpha_k = 1$, $\alpha_{kjs} = 1$; BLC(20,1) = Bayesian LC imputation model with $\alpha_k = 20$, $\alpha_{kjs} = 1$; BLC(1,.01) = Bayesian LC imputation model with $\alpha_k = 1$, $\alpha_{kjs} = .01$; BLC(20,.01) = Bayesian LC imputation model with $\alpha_k = 20$, $\alpha_{kjs} = .01$; DPMM(1) = DPMM imputation model with $\alpha_{kjs} = 1$; DPMM(.01) = DPMM imputation model with $\alpha_{kjs} = .01$. Largest values in relative bias and too low coverage rates are marked in boldface.

predictors were generated from the following log-linear model:

$$\begin{aligned}\log \Pr(Y_1, \ldots, Y_{15}) \propto & \\ & -0.15\sum_{j=1}^{15} Y_j + 0.5\sum_{j=1}^{4}\sum_{k=j+1}^{5} Y_j Y_k \\ & -0.1\sum_{j=6}^{10}\sum_{k=j+1}^{11} Y_j Y_k + 0.15\sum_{j=12}^{14}\sum_{k=j+1}^{15} Y_j Y_k \\ & + 0.3 Y_1 Y_2 Y_7 + 0.6 Y_3 Y_4 Y_8 - 0.4 Y_6 Y_9 Y_{10}, \end{aligned} \quad (4)$$

while the remaining five predictors were assumed to be independent of the rest, with marginal probabilities $\Pr(Y_j = 1)$, $j = 16, \ldots, 20$, as reported in Table 5.

Given $Y_1, \ldots, Y_{20}$ the outcome Y_{21} was generated from the following binary logistic model:

$$\begin{aligned}\text{logit}(Y_{21}) = & -0.9 + \beta_1 Y_1 + 1.8 Y_2 - 0.95 Y_3 - 0.9 Y_4 \\ & + 0.8 Y_5 + \beta_6 Y_6 - 0.5 Y_7 + 0.6 Y_8 + Y_9 + 0.55 Y_{10} \\ & - 0.6 Y_{11} + 0.75 Y_{12} - 1.2 Y_{13} + 0 Y_{14} + 0 Y_{15} \\ & + \beta_{16} Y_{16} + 0.85 Y_{17} + 0.55 Y_{18} + 0 Y_{19} + \beta_{20} Y_{20} \\ & + \beta_{1.5} Y_1 Y_5 + \beta_{1.17} Y_1 Y_{17} + \beta_{1.5.17} Y_1 Y_5 Y_{17}. \end{aligned} \quad (5)$$

Besides the two- and three-way interaction terms, in model (5) we also specified some null effects (coefficients equal to zero) in order to assess how the imputation models deal with irrelevant variables. The values of the β parameters are shown in Table 6. From models (4) and (5) (and the items described in Table 5), we generated $N = 200$ datasets with $n = 2{,}000$ observations.

Table 5. Probability of observing 1 for the independently generated items of Study 2

$Pr(Y_{16} = 1) = 0.7$
$Pr(Y_{17} = 1) = 0.6$
$Pr(Y_{18} = 1) = 0.55$
$Pr(Y_{19} = 1) = 0.6$
$Pr(Y_{20} = 1) = 0.7$

Table 6. Parameter values under investigation in Study 2

Parameter	β_1	β_6	β_{16}	β_{20}	$\beta_{1.5}$	$\beta_{1.17}$	$\beta_{1.5.17}$
Value	0.8	1.1	−0.45	0.0	1.3	−0.85	0.45

Introducing Missingness

Missingness was entered in Y_1 (involved in all the interaction terms), Y_6, Y_{16}, and Y_{20} (an irrelevant predictor). The marginal rate of missingness (generated with the MAR mechanism reported in Table 7) was equal to 25% for each variable with missing values.

Settings of the Imputation Models

The specifications used for the imputation models were similar to Study 1. For FLC and BLC, our model selection procedure gave an average (maximum) number of classes of $\bar{K} = 16.31$, while we increased the number of classes for the DPMM, specifying for the latter 20 more classes than the FLC and BLC models.[5] Based on the results of Study 1, we decided not to vary α_{kjs} anymore, but instead fixed it to 0.01 for both BLC and DPMM. The latent hyperparameter of the BLC model α_k was set to be equal to either 1 or 80, where the latter was chosen to be sufficiently large to ensure full allocation of the latent classes. This is indicated with BLC(1) and BLC(80).

Outcomes

To assess the performance of the imputation models, we looked at relative bias, stability, and coverage rates for the coefficients of the variables with missing values (see Table 8). For the null effect β_{20}, we considered the absolute bias.

Results

The results reported in Table 8 show that the null effect β_{20}, the three-way interaction term $\beta_{1.5.17}$, and the main effects β_6 and β_{16} were well retrieved by all methods. The two-way interaction terms resulting from MICE, BLC(1), and DPMM were remarkably biased, while FLC and BLC

Table 7. MAR mechanisms used in Study 2

Item with missingness	Condition	Pr(Item is missing)
Y_1	$Y_3 = 0, Y_4 = 0$	.15
	$Y_3 = 0, Y_4 = 1$	.05
	$Y_3 = 1, Y_4 = 0$	.25
	$Y_3 = 1, Y_4 = 1$	.30
Y_6	$Y_5 = 0, Y_{21} = 0$	.30
	$Y_5 = 0, Y_{21} = 1$	.20
	$Y_5 = 1, Y_{21} = 0$	.10
	$Y_5 = 1, Y_{21} = 1$	.35
Y_{16}	$Y_9 = 0, Y_{10} = 0$	.30
	$Y_9 = 0, Y_{10} = 1$	.25
	$Y_9 = 1, Y_{10} = 0$	.10
	$Y_9 = 1, Y_{10} = 1$	.40
Y_{20}	$Y_{14} = 0, Y_{15} = 0$	.35
	$Y_{14} = 0, Y_{15} = 1$	.10
	$Y_{14} = 1, Y_{15} = 0$	.10
	$Y_{14} = 1, Y_{15} = 1$	.45

Table 8. Relative bias, stability, and coverage rate observed for the estimates of seven logistic model parameters in model (5) after applying three different imputation models

	Parameter						
Method	β_1	β_6	β_{16}	β_{20}	$\beta_{1.5}$	$\beta_{1.17}$	$\beta_{1.5.17}$
Relative bias							
MICE	**0.20**	−0.01	0.00	0.01	−**0.22**	−**0.16**	−0.06
FLC	−0.05	−0.09	−0.10	0.00	−0.11	−0.14	−0.05
BLC(1)	0.01	−0.12	−0.13	0.00	−**0.21**	−**0.16**	−0.06
BLC(80)	−0.04	−0.08	−0.08	0.00	−0.09	−0.12	−0.05
DPMM	0.02	−0.12	−0.13	0.00	−**0.22**	−**0.16**	−0.06
Stability							
MICE	0.41	0.14	0.15	0.14	0.38	0.40	0.35
FLC	0.44	0.13	0.13	0.13	0.42	0.42	0.35
BLC(1)	0.40	0.14	0.13	0.13	0.40	0.39	0.35
BLC(80)	0.44	0.14	0.14	0.14	0.43	0.42	0.36
DPMM	0.40	0.14	0.13	0.13	0.39	0.40	0.35
Coverage rate							
MICE	0.98	0.93	0.92	0.96	0.96	0.96	0.94
FLC	0.94	**0.88**	0.95	0.96	0.96	0.96	0.96
BLC(1)	0.97	**0.84**	0.94	0.98	0.96	0.97	0.95
BLC(80)	0.94	0.91	0.94	0.96	0.96	0.96	0.94
DPMM	0.96	**0.87**	0.94	0.98	0.96	0.97	0.94

Notes. For the null effect β_{20} absolute bias is reported. MICE = MICE imputation technique; FLC = frequentist LC imputation model; BLC(1) = Bayesian LC imputation model with $\alpha_k = 1$; BLC(80) = Bayesian LC imputation model with $\alpha_k = 80$; DPMM = DPMM imputation model. Largest values in relative bias and too low coverage rates are marked in boldface.

[5] With the DPMM model superfluous classes are given weights equal to zero during the Gibbs sampling. Hence, with such an imputation model any selected number of classes leads to similar inferences provided that this number is large enough.

Table 9. Variables used in the real-data application

Item label	Item description	Values (range)
Items for the analysis model		
Y_0	Respondent's happiness	1 = *Not happy* to 3 = *Very happy*
Y_1	Respondent's opinion about his/her life	1 = *Dull* to 3 = *Exciting*
Y_2	Respondent's job satisfaction	1 = *Very dissatisfied* to 4 = *Very satisfied*
Y_3	Respondent's health status	1 = *Poor* to 5 = *Excellent*
Y_4	Respondent's marital status	0 = *Not married* to 1 = *Married*
Y_5	Respondent's employment status	1 = *Self employed* to 2 = *Work for someone else*
Y_6	Respondent's political view	1 = *Liberal* to 3 = *Conservative*
Y_7	Respondent's gender	0 = *Female* to 1 = *Male*
Y_8	Respondent's working status	1 = *Full time* to 4 = *Not working*
Y_9	Respondent's employer	1 Government – 2 Private
Y_{10}	Respondent's family income	1 <5,000 to 4 >25,000
Y_{11}	Respondent's time spent with friends	1 = *Almost every day* to 7 = *Never*
Items used to generate missingness		
Y_{12}	Respondent's education	0 <*Highschool* to 4 = *Graduate*
Y_{13}	Respondent's working contract	1 = *Full time* to 2 = *Part-time*
Y_{14}	Respondent's occupation prestige (score)	1 = 10/19 to 8 = 80/89

Notes. Top: items of the analysis model (6). Bottom: items used to generate missingness.

(80) provided good estimates for these parameters. The β_1 coefficient was also correctly recovered by all methods, except for MICE. FLC and BLC(80) produced the least stable estimates, probably due to the fact that a larger number of LCs was exploited by these two methods. DPMM and BLC(1) returned similarly stable estimates: their standard deviations were overall smaller than those of the other two LC imputation methods. MICE provided the least varying estimates across all the imputation methods. All methods yielded confidence intervals with acceptable coverage (close to the 95% nominal level). The only exceptions were the interval for β_6, which resulted in too low coverage after imputing with FLC, BLC(1), or DPMM.

Real-Data Study

The *General Social Survey* (GSS; National Opinion Research Center, 1972) is a survey conducted by the National Opinion Research Center and administered every 2 years to a random sample of households resident in the United States. Here we use data from this study to evaluate the imputation models in a situation where the associations between variables are as encountered in real data. Our experiment was carried out with the GSS cross-sectional wave of 2014. Analyses were again performed with R 3.3.0.

Study Design

The Data

From the original dataset (which consisted of n = 2,538 units and J = 895) we removed all records with missing data and "Don't know" and "Not applicable" answers. The resulting dataset had a sample size equal to n = 477. Subsequently, we selected a subset of J = 15 variables, of which the first 12 were the possible outcome and the predictors of a potential analysis model, and the remaining 3 were used to generate the missingness (and therefore included in the imputation models). The variables names and the description of their categories are listed in Table 9.[6]

The Substantive Model

The analysis was performed with an ordered logistic model estimated on the complete dataset (with n = 477), in which the variable Happiness (Y_0 in Table 9) was the outcome and the $Y_1, \ldots, Y_{11}$ of Table 9 were the predictors. More specifically, the model we estimated was

$$\log\left(\frac{\Pr(Y_0 \leq s)}{\Pr(Y_0 > s)}\right) \propto \sum_{j=1}^{11} \beta_j Y_j + \beta_{57} Y_5 Y_7 + \beta_{48} Y_4 Y_8. \quad (6)$$

[6] For some variables the categories were reversed, while for others some categories were combined.

Table 10. Results of the real-data application

			Imputation method							
	Complete data		MICE		FLC		BLC		DPMM	
Parameter	Estimate	*SE*	Estimate	*SE*	Estimate	*SE*	Estimate	*SE*	Estimate	*SE*
β_1	1.12*	0.21	1.06*	0.46	1.12*	0.22	1.14*	0.21	1.19*	0.21
β_2	0.82*	0.15	0.56	0.35	0.95*	0.17	0.87*	0.18	0.68*	0.17
β_3	0.80*	0.16	1.27*	0.37	0.79*	0.16	0.77*	0.16	0.79*	0.16
$\beta 4$	−0.24	0.42	−0.05	0.92	−0.51	0.48	−0.35	0.51	−0.27	0.50
β_5	0.62	0.46	0.72	2.21	0.69	0.62	0.56	0.61	0.62	0.63
β_6	0.25	0.13	0.40	0.29	0.36*	0.16	0.28	0.16	0.25	0.16
β_7	3.02*	1.24	5.50	5.03	3.72*	1.58	3.18*	1.59	3.20*	1.59
β_8	−0.47*	0.19	−0.05	0.41	−0.52*	0.21	−0.46*	0.22	−0.40	0.22
β_9	−0.22	0.27	−0.50	0.65	−0.15	0.28	−0.21	0.27	−0.22	0.27
β_{10}	0.13	0.21	0.05	0.48	0.14	0.23	0.14	0.22	0.14	0.22
β_{11}	−0.17*	0.07	−0.09	0.17	−0.20*	0.08	−0.18*	0.08	−0.17*	0.07
β_{57}	−1.70*	0.65	−3.11	2.55	−2.08*	0.82	−1.81*	0.83	−1.81*	0.83
β_{48}	0.69*	0.28	0.29	0.62	0.84*	0.32	0.74*	0.35	0.72*	0.35

*Notes.*The table shows the point estimates and the standard errors for the ordered logistic regression model (6) estimated on the complete data (n = 477) and on the incomplete datasets, imputed with the MICE, FLC, BLC, and DPMM methods. "*" indicates the 5% significant parameter estimates.

The first column of Table 10 reports the estimates and the standard errors of the β's parameters obtained with the complete data, where significant predictors at 5% are highlighted.

Introducing Missingness

We artificially created missing values for the variables Y_2, Y_5, Y_6, and Y_8. MAR missingness was generated with the four different logistic models described in Table 11. The parameters of these logistic models were set such that the rate of missingness was between 25% and 33% per variable.

Imputation Model Settings

For each MI method m = 50 imputations were performed. For the model selection, we ran the BLC model with 50 components and b = 5,000 iterations for the burn-in, and 5,000 to estimate the distribution of K. The resulting posterior maximum for the number of classes was equal to 16. Therefore, we performed the imputations with the FLC and BLC models with K = 16. The latent hyperparameter for the BLC model was set equal to α_k = 40, which was large enough to ensure full allocation of the LCs, while the conditional hyperparameter for the BLC and the DPMM models was set equal to α_{kjs} = 0.05. The DPMM model was implemented with K = 20. The Gibbs sampler for both BLC and DPMM was run with T = 55,000 and b = 5,000. For MICE, 20 iterations were used for each imputation.

Table 11. MAR mechanisms used to generate missing data in the real-data application

Item with missingness	Missingness generating model
Y_2	$1 - 1.5\ Y_{12}$
Y_5	$-2.2 + 1.2\ Y_{13}$
Y_6	$1.3 - 1.25\ Y_9$
Y_8	$2.1 - 0.8\ Y_{14}$

Outcomes

After imputing the data, model (6) was estimated for each completed dataset. We focused on the point estimates and the standard errors obtained after applying the MI pooling rules. We also assessed which estimates were significant at 5% after calculating their MI p-values[7].

Results

The results reported in Table 10 show that MICE performed badly: its point estimates for both main and interaction effects were rather far from those obtained with the Complete Data. Furthermore, MICE produced very large standard errors, causing most of the estimates to be no longer significant (except for β_1 and β_3). In contrast, the LC imputation models (FLC, BLC, and DPMM) yielded parameter estimates close to those of the Complete Data, and the extra uncertainty due to the presence of missing data (reflected in the standard errors) was much smaller

[7] The degrees of freedom were calculated as in Van Buuren (2012).

than with the MICE. Because of this, most of the parameters that were significant with the Complete Data were also significant (at the 5% level) after imputing the data using the LC-based imputation techniques. The only exceptions were β_6, which became significant with FLC, and β_8, which was no longer significant with DPMM. The significant parameters according to the BLC imputation were the same as those by the Complete Data.

Discussion

In this paper, we proposed using a BLC model for the MI of categorical data. As any LC model, this model is automatically able to capture the dependencies present in the data – including complex interactions – with the simple specification of the needed number of classes. We also highlighted the advantages of performing the imputations with the BLC model, rather than with the FLC or the DPMM method. Compared to the FLC model, the BLC model offers a very fast and intuitive model selection step, which makes use of the posterior distribution of the number of LCs required by the data and which can be obtained with an extra (preliminary) run of the Gibbs sampler. Another computational advantage is that parameter uncertainty is automatically accounted for, whereas the FLC requires using a nonparametric bootstrap procedure. Compared to the DPMM approach, the BLC model offers important additional flexibility through the specification of the hyperparameter for the latent class proportions. By setting its value large enough, one guarantees the allocation of units across all LCs, which is a way to avoid the risk of underfitting associated with the DPMM model.

Two simulation studies and a real-data experiment were carried out in which the BLC model was contrasted with the FLC, DPMM, and MICE methods. In the first study, we used a large sample size (n = 5,000) and a small number of variables (J = 6), and we manipulated the total rate of missingness in the variables with nonresponses. In the second study, a smaller sample size (n = 2,000) and a larger J (= 21) were considered. In both studies, the latent hyperparameter of the BLC model was also manipulated, in order to emphasize the influence of this value on the final imputations. In the real-data study, the sample size was n = 477 and the number of variables (used for the imputations) was equal to J = 15. In all studies, the BLC imputation model (with large values for the latent hyperparameter and small values for the conditional hyperparameter) provided the best results in terms of bias, stability, and coverage rates for the main and interaction effects of the substantive model. In the real-data study, the BLC model also detected the same set of significant parameters as with the Complete Data analysis. The FLC method (implemented with the same number of classes of the BLC model) also yielded good results, although worse than the BLC method (e.g., the bias of one of the interaction terms in Study 1 was remarkable). This was probably due to the fact the FLC model, unlike the BLC model with a large value of the latent hyperparameter, gave too small weights to LCs that were important for the imputations. The DPMM model and the BLC model with uniform prior for the latent proportions both failed to correctly retrieve the estimates of some interaction terms. Lastly, the MICE method was not flexible enough to be able to capture all important features of the data in most situations.

Based on our results, our recommendation for researchers that need to deal with (MAR) missing categorical data is to use our BLC MI approach combined with the model selection and prior specifications described in this paper. A limitation of this new MI approach is that it can be used only with cross-sectional categorical data. However, in future research, we will extend it to deal with combinations of categorical and continuous variables, as well as with data from multilevel and longitudinal designs in which more complex dependencies may arise. Another challenge for future research is to develop a version of the BLC imputation model for situations in which the missing data are *missing not at random*.

Declaration of Conflicting Interests

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

Acknowledgment

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. *The Sage Handbook of Quantitative Methods in Psychology, 4*, 72–89.

Gelfand, A. E., & Smith, A. F. M. (1990). Sampling based approaches to calculating marginal densities. *Journal of the American Statistical Association, 85*, 398–409. https://doi.org/10.1080/01621459.1990.10476213

Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (2013). *Bayesian data analysis* (3rd ed.). London, UK: Chapman & Hall.

Goodman, L. A. (1974). Exploratory latent structure analysis using both identifiable and unidentifiable models. *Biometrika, 61*, 215–231. https://doi.org/10.1093/biomet/61.2.215

Hoijtink, H., & Notenboom, A. (2004). Model based clustering of large data sets: Tracing the development of spelling ability. *Psychometricka, 69*, 481–498. https://doi.org/10.1111/j.1467-9868.2011.00781.x

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, NJ: Princeton University Press.

Little, R. J. A., & Rubin, D. B. (2002). *Statistical analysis with missing data* (2nd ed.). New York, NY: Wiley.
McLachlan, G. J., & Peel, D. (2000). *Finite mixture models*. New York, NY: Wiley.
National Opinion Research Center. (1972). *General Social Survey. GSS (1972–2014) Release 4. Cross-sectional wave 2014,*. Chicago, IL: University of Chicago. Retrieved from http://www3.norc.org/GSS+Website/Download/SPSS+Format
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. https://doi.org/10.1111/j.1467-9868.2011.00781.x
Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys*. New York, NY: Wiley.
Schafer, J. L. (1997). *Analysis of incomplete multivariate data*. London, UK: Chapman & Hall.
Schafer, J. L., & Graham, J. W. (2002). Missing data: Our view of the state of the art. *Psychological Methods, 7*, 147–177. https://doi.org/10.1037/1082-989X.7.2.147
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. https://doi.org/10.3102/1076998613480394
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. https://doi.org/10.1080/10629360600810434
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., Groothuis-Oudshoorn, K., Robitzsch, A., Vink, G., Doove, L., & Jolani, S. (2014). *mice: Multivariate Imputation by Chained Equations*. (R package version 2.22) [Computer software manual]. Retrieved from http://cran.r-project.org/web/packages/mice/index.html
Vermunt, J. K., & Magidson, J. (2013). *LatentGOLD 5.0 upgrade manual*. Belmont, MA: Statistical Innovations Inc.
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. https://doi.org/10.1111/j.1467-9531.2008.00202.x

Received May 13, 2016
Revision received February 21, 2017
Accepted December 15, 2017
Published online June 21, 2018

Davide Vidotto
Department of Methodology and Statistics
Tilburg University
Warandelaan 2
5037 AB Tilburg
The Netherlands
d.vidotto@uvt.nl

Davide Vidotto received a Master in Statistical Sciences at the University of Padova (Italy) in 2013, and is currently a PhD candidate at Tilburg University (NL). His research is focused on missing data imputation.

Jeroen K. Vermunt received his PhD degree in social sciences research methods from Tilburg University in the Netherlands in 1996, where he is currently a full professor in the Department of Methodology and Statistics. His research interests include latent class and finite mixture models, IRT modeling, longitudinal and event history data analysis, multilevel analysis, and generalized latent variable modeling.

Katrijn Van Deun is assistant professor in Methodology and Statistics at Tilburg University. She obtained a Master in Psychology and in Statistics and a PhD in Psychology. Her main area of expertise is scaling, clustering and component analysis techniques, which she applies in the fields of psychology, chemometrics, and bioinformatics.

