BAYESIAN LATENT CLASS MODELS FOR THE MULTIPLE IMPUTATION OF CROSS-SECTIONAL, MULTILEVEL AND LONGITUDINAL CATEGORICAL DATA

Davide Vidotto
Tilburg University

This research is funded by The Netherlands Organization for Scientific Research (NWO [grant project number 406-13-048]).
Printing was financially supported by Tilburg University.

ISBN: 978-94-6295-808-1

Printed by: Proefschriftmaken || www.proefschriftmaken.nl

Cover design: Faboosh design & art

BAYESIAN LATENT CLASS MODELS FOR THE MULTIPLE IMPUTATION OF CROSS-SECTIONAL, MULTILEVEL AND LONGITUDINAL CATEGORICAL DATA

PROEFSCHRIFT

ter verkrijging van de graad van doctor aan Tilburg University op gezag van de rector magnificus, prof. dr. E.H.L. Aarts, in het openbaar te verdedigen ten overstaan van een door het college voor promoties aangewezen commissie in de aula van de Universiteit

op vrijdag 2 maart 2018 om 14.00 uur

door

Davide Vidotto

geboren op 21 april 1988 te Treviso, Italië

Promotor:	Prof.dr. J. K. Vermunt
Copromotor:	Dr. K. Van Deun
Overige leden van de Promotiecommissie:	Prof.dr. S. van Buuren
	Prof.dr. F. Bassi
	Dr. A.O.J. Cramer
	Prof.dr. L.A. van der Ark

Alla mia famiglia

To my family

Voor mijn familie

TABLE OF CONTENTS

1

INTRODUCTION

This dissertation deals with the *multiple imputation* (MI; Rubin (1987)) of categorical data coming from different types of data collection and data analysis designs. In particular, the use of *latent class* (LC) models (Lazarsfeld, 1950) for the MI of data coming from cross-sectional study designs (as it was first proposed by Vermunt, Van Ginkel, Van der Ark and Sijtsma (2008)) will serve as a starting point to obtain imputation models that can deal with more complex designs, such as multilevel (i.e., when multiple individuals are nested within a group) and longitudinal (i.e., when multiple observations for each individual are observed across time) designs.

Latent Class models for Multiple Imputation

LC models are known among analysts and methodologists for their substantive use, in which the estimates provided by the model are used to define latent types (or profiles, clusters) of units. These profiles differ from each other for some characteristics, identified by the distribution of the scores on the indicator variables (usually categorical variables). Within each LC, the joint distribution of these features is described by a product of locally independent categorical (e.g., Multinomial) distributions by means of the *local independence* assumption. Local independence makes the model easily interpretable, and allows to take into account a large number of indicators for a specific theoretical construct. A graphical representation of the LC model is given in Figure 1.1, in which X represents the LC variable and the Y's represent the J indicators.

However, LC models - which are members of the family of mixture models - can be used in contexts other than latent groups identification. That is, since mixture

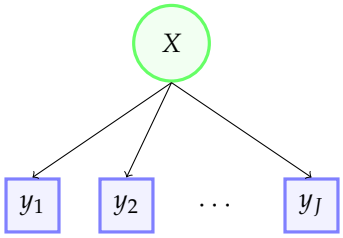


Figure 1.1: LC model, graphical representation. X: latent class variable; Y's: indicators (J in total).

models can correctly pick up unobserved heterogeneity and relevant relationships in the data if the number of specified LCs is large enough (McLachlan & Peel, 2000), LC models can be used as a density estimation tool. In density estimation, the goal is to estimate and describe the joint distribution of the variables present in a dataset, retrieving all possible associations which tie the variables to each other. Under this framework, interpretation of the model parameters is of little interest, and the main focus is on the *predictions* the model provides by means of these parameters. Furthermore, models used for density estimation are likely to require the estimation of a tremendous number of parameters, which would make them very hard to interpret. Thus, the model parameters are merely a device used to obtain predictions and/or an overall description of the joint distribution of the data.

Vermunt et al. (2008) exploited this feature of LC models, and proposed them for application in MI. In MI, the missing data of a dataset are replaced (or imputed, predicted) $M > 1$ times by different sets of values, the distribution of which is estimated with the imputation model. In particular, the task of the imputation model is to provide values sampled from $\Pr(D^{mis}|D^{obs})$, that is, the distribution of the missing data given the observed data. When the missing data mechanism is ignorable[1], MI can retrieve the correct distribution of the data (for some analysis model of interest), leading to proper substantive inferences. Furthermore, by obtaining M different imputations it becomes possible to quantify the uncertainty about the imputed missing values at the analysis stage. More specifically, substantive analyses are performed on each of the M imputed datasets, where for correct statistical inferences the results are pooled using Rubin (1987) 's rules.

LC models require a very easy model specification (the number of LCs), which makes them flexible and automatic, since relevant associations in the data need not to be specified a-priori. Concerning the model selection issue, in MI selecting a model that overfits the data (i.e., a model that capture sample-specific features) is

1 That is, the missing data generating mechanism is independent of the unobserved data and its parameter is distinct from the ones of the assumed data generating model (Rubin, 1976).

less problematic than an underfitting model (i.e., a model that ignores important relations in the data), as remarked by Vermunt et al. (2008). As a consequence, MI with LC models can be performed by selecting an arbitrarily large number of classes[2], which is similar to fitting a saturated log-linear model to the data, a strategy advocated by Schafer (1997). Unlike log-linear models, however, LC models can be used to impute datasets with a large number of variables. This is due to both the local independence assumption described above and the simple model specification required.

Chapter 2 reviews the different types of LC-based MI proposed in the literature. These include MI using the standard LC model (Vermunt et al., 2008), the Divisive LC model (Van der Palm, Van der Ark & Vermunt, 2014), the Bayesian version of the LC model and the Dirichlet process mixture of Multinomial distribution (Si & Reiter, 2013). A detailed description of how these LC models impute missing data is given, and benefits and drawbacks of the four approaches are discussed. Furthermore, a comparison of the four LC imputation methods is carried out by means of an empirical application.

Bayesian Latent Class models for Multiple Imputation

In Chapter 3 the use of Bayesian LC models for MI is investigated in more detail. As Schafer and Graham (2002) emphasized, Bayesian modeling for MI can directly lead to proper imputations, without resorting to bootstrap or other computational techniques (used for instance by Vermunt et al. (2008)). This advantage of Bayesian modeling is due to the fact that, when performing imputations with a statistical model, we need to account for two sources of uncertainty:

- uncertainty caused by the missing data D^{mis};
- uncertainty caused by the estimation of the imputation model parameter θ.

In particular, Bayesian models enable to embed all the uncertainty about θ in a single posterior distribution $\Pr(\theta|D^{obs})$, estimated conditioned on the observed data by means of the well-known Bayes' theorem:

$$\Pr(\theta|D^{obs}) = \frac{\pi(\theta)f(D^{obs}|\theta)}{\int_{\Theta}\pi(\theta)f(D^{obs}|\theta)d\theta}.$$

2 Working with a number of LCs larger than what is actually required by the data corresponds to overfitting in mixture modelling.

Here, $\pi(\cdot)$ represents a distribution which encloses the prior information we have about θ, while $f(\cdot)$ is the likelihood function which represents a probability distribution imposed on the observed data, identified by the value of θ.

Once the posterior distribution of the imputation model parameter is obtained, the imputations can be performed with the posterior predictive distribution of the missing data, given by

$$\Pr(D^{mis}|D^{obs}) = \int_{\Theta} f(D^{mis}|\theta) \Pr(\theta|D^{obs}) d\theta. \tag{1.1}$$

Equation (1.1) shows clearly how Bayesian modeling takes into account the two types of uncertainty described above: first, M samples from the posterior $\Pr(\theta|D^{obs})$ are drawn, and subsequently imputations are performed through $f(D^{mis}|\theta^{(m)})$, $m = 1, ..., M$. In order to estimate the Bayesian version of the LC model, a Gibbs sampler (Geman & Geman, 1984) with Data Augmentation (Tanner & Wong, 1987) configuration is needed (as proposed by Escobar and West (1995)), because we are dealing with latent variables. Data Augmentation is an algorithm that -at each iteration- first samples the latent variables for each unit in the dataset, and then updates the model parameters accordingly.

Bayesian estimation of unknown model parameters requires specification of prior knowledge about their values by means of the prior distribution $\pi(\cdot)$. In MI, it is common to perform imputations from a state of complete ignorance about the model parameters (most of the times the imputation model and the analysis model do not match, and imputer and analysts can be different entities), which suggests to use noninformative (or vague) priors for the imputation model parameters. A complication when running the Gibbs sampler for estimating LC models with a large number of classes is given by the fact that the prior distribution of the mixture weights strongly affects the number of classes allocated during the sampler iterations and, therefore, the quality of the imputations (Rousseau & Mergensen, 2011). Chapter 3 of the thesis therefore will examine the specification of different prior distributions for the imputation model parameter by means of a simulation study, providing useful guidelines about how to set them when performing MI.

Bayesian Multilevel Latent Class models for Multiple Imputation

In social sciences, researchers often have to deal with datasets containing more complex dependency structures. For instance, it is common to design a sampling mechanism in which data are collected for individuals (level-1 units) that come

from different groups (level-2 units); in this case, both level-2 and level-1 units are sampled and the data are said to be structured in a *multilevel* or *nested* form. A typical example is given by students' scores observed in different schools. Besides the associations between the various variables, other kinds of dependencies usually arise in this context, such as level-1 units coming from the same group that are correlated with each other (which, in substantive analysis, is usually accounted for by introducing one or more random effects in the model). Furthermore, variables can concern units at different levels of the hierarchy, such as measures obtained at the student-level and at the school-level. The presence of variables at different levels yields what is referred to as cross-level relationships.

When MI is performed on this kind of data with standard single-level imputation models, some of these relationships are probably lost in the imputed datasets. For instance, ignoring the nested structure of the data during the imputation stage is likely to produce too precise inferences for the substantive model parameters, inflating in this way the occurrence of Type-I errors. Furthermore, with single-level imputation models all variables would be treated as level-1 variables, thus disregarding the hierarchy of the sampling design.

To overcome these difficulties, a proper imputation model that accounts for the nested structure of the data must be used. In Chapter 4 the *Bayesian multilevel LC* (BMLC) *model* is proposed for this purpose. The BMLC model presented here is the Bayesian configuration of the frequentist non-parametric multilevel LC model proposed for substantive analysis by Vermunt (2003), in which the clustering occurs for units at both levels of the hierarchy. With such configuration, the model can take into account not only relevant associations among variables at both levels, but also within-group dependencies. These are picked up by means of the *conditional independence* assumption, according to which units at the lower-level become independent of each other, conditioned on the higher-level LC to which the units' group belongs. In Chapter 4 a simulation and a real data study are carried out to investigate the performance of the BMLC model as an imputation model.

Bayesian Latent Markov models for Multiple Imputation

Another common survey design involves collecting data over time for the same subjects, known as *longitudinal study* design. Analogous to the multilevel case, also in this scenario the imputation model must be tailored to take into account the specific types of dependencies that such data collection mechanism entails. These include auto-correlations and crossed-lagged relationships for the time-varying variables.

Moreover, time-constant variables are likely to be present in the dataset as well, implying that the imputation model should also be able to account for relationships between time-varying and time-constant variables.

In Chapter 5, the Bayesian *mixture latent Markov* (BMLM) model is proposed for the imputation of longitudinal data. Latent Markov models (Baum, Petrie, Soules & Weiss, 1970) represent a natural extension of LC models to longitudinal data, and involve the specification of dynamic latent states (i.e., LC membership that can vary in time) which follow a first-order Markov chain. Thus, the latent states can potentially capture the relevant relationships among the variables within each time point, as well as auto-correlations between adjacent time-points (by the first-order Markov assumption). Furthermore, the inclusion of a time-constant LC variable enables to capture dependencies across all time points, as well as enables including time-constant variables in the imputation model (which, in turn, should also be imputed, if missing). To evaluate how the BMLM model performs as an imputation model, two simulation studies and a real-data study were carried out, and their results are reported in Chapter 5.

The four main chapters of this dissertation can be read independently since they are written as articles for scientific journals. Because of this, the chapters contain some overlapping information and moreover notation is sometimes slightly different across chapters.

2

MULTIPLE IMPUTATION OF MISSING CATEGORICAL DATA USING LATENT CLASS MODELS: STATE OF THE ART

This chapter provides an overview of recent proposals for using latent class models for the multiple imputation of missing categorical data in large-scale studies. While latent class (or finite mixture) modeling is mainly known as a clustering tool, it can also be used for density estimation, i.e., to get a good description of the lower- and higher-order associations among the variables in a dataset. For multiple imputation, the latter aspect is essential in order to be able to draw meaningful imputing values from the conditional distribution of the missing data given the observed data.

We explain the general logic underlying the use of latent class analysis for multiple imputation. Moreover, we present several variants developed within either a frequentist or a Bayesian framework, each of which overcomes certain limitations of the standard implementation. The different approaches are illustrated and compared using a real-data psychological assessment application.

This chapter is published as Vidotto, D., Vermunt, J.K. & Kaptein, M.C. (2015). Multiple Imputation of Missing Categorical Data using Latent Class Models: State of the Art. *Psychological Test and Assessment Modeling, vol. 57, 4 pp. 542-576.*

2.1 Introduction

Social and behavioral science researchers often collect data using tests or questionnaires consisting of items which are supposed to measure one or more underlying constructs. In a psychology assessment study for example, this could be constructs such as anxiety, extraversion, or neuroticism. A very common problem is that a part of the respondents fail to answer all questionnaire items (Huisman, 1999), resulting in incomplete datasets. However, most of the standard statistical techniques can not deal with the presence of missing data. For example, computation of Cronbach's alpha requires that all variables in the scale of interest are observed.

Various methods for dealing with item nonresponse have been proposed (Little & Rubin, 2002; Schafer & Graham, 2002). Listwise and pairwise deletion, which simply exclude units with unobserved answers from the analysis, are the most frequently used in psychological research (Schlomer, Bauman & Card, 2010). These are, however, also the worst methods available (Wilkinson & Task Force on Statistical Inference, 1999): they result in loss of power and, unless the strong assumption that data are missing *completely at random* (MCAR)[1] is met, they may lead to severely biased results. Due to their simplicity and their widespread inclusion as standard options in statistical software packages, these methods are still the most common missing data handling techniques (Van Ginkel, 2007).

Methodological research on missing data handling has lead to two alternative approaches that overcome the problems associated with listwise or pairwise deletion: *maximum likelihood for incomplete data* (MLID) and *multiple imputation* (MI). Under the assumption that the missing data are *missing at random* (MAR), the estimates of the statistical model of interest (from here on also referred to as the substantive model) resulting from MLID or MI have the desirable properties to be unbiased, consistent, and asymptotically normal (Roth, 1994; Schafer & Graham, 2002; Allison, 2009; Baraldi & Enders, 2010). MLID involves estimation the parameters of the substantive model interest by maximizing the incomplete-data likelihood function. That is, the likelihood function consisting of a part for the units with missing data and a part for the units with fully observed data. While in MLID the missing data and the substantive model are the same, in MI (Rubin, 1987) the missing data handling model (or imputation model) and the substantive model(s) of interest can and will typically be different. Note that unlike single value imputation, MI replaces each missing value with $m > 1$ imputed values in order to be able to account for

1 According to Rubin (1976)'s classification, a missing data mechanism is said to be: (a) MCAR, when the probability of nonresponse in a variable is independent of the variable itself as well as of the other variables; (b) *missing at random* (MAR), when the probability of nonresponse in a variable depends only on the variables observed for the person concerned; (c) *missing not at random* (MNAR), when the probability of missingness is related to variables which are unobserved for the person concerned.

the uncertainty about the missing information. In practice, applying MI yields m complete datasets, each of which can be analyzed separately using the standard statistical method of interest, and where the m results should be combined in a specific manner. For more details on MI, we refer to Rubin (1987), Schafer (1997), and Little and Rubin (2002).

For continuous variables with missing values, Schafer (1997) proposed using the multivariate normal MI model, which has been shown to be quite robust to departures from normality (Graham & Schafer, 1999). Items of psychological assessment questionnaires, however, are categorical rather than continuous variables. For such categorical data, Schafer (1997) proposed MI with log-linear models, which can capture the relevant associations in the joint distribution of a set of categorical variables and can be used to generate imputation values. However, log-linear models for MI can only be applied when the number of variables is relatively small, as the number of cells in the multi-way cross-table that has to be processed increases exponentially with the number of variables (Vermunt et al., 2008).

An alternative MI tool is offered by the sequential regression modeling approach, which includes *multiple imputation by chained equation* (MICE) (Van Buuren & Oudshoorn, 1999). This is an iterative method that involves estimating a series of univariate regression models (e.g., a series of logistic or polytomous regressions in the case of categorical variables), where missing values are imputed (variable by variable) based on the current regression estimates for dependent variable concerned. The idea of MICE is that the sequential draws from the univariate conditional models are equivalent to or at least a good approximation of draws from the joint distribution of the variables in the imputation model. Despite of being an intuitive and practical method, also MICE has certain limitations. First, there is no statistical support that missing data draws converge to the posterior distribution of the missing data. Second, by default, MICE only includes the main effects in the regression equations, which risks to not pick up higher-order interactions among the variables. Furthermore, whereas the method allows including higher-order interactions, this can be a fairly difficult and time-consuming task when the number of variables in the imputation model is large (Vermunt et al., 2008).

Vermunt et al. (2008) proposed an imputation model for categorical data based on a maximum likelihood finite mixture or *latent class* (LC) model. LC models for MI seem to overcome various of the difficulties associated with log-linear models and MICE. LC models can efficiently be estimated also when the number of the variables is large (Si & Reiter, 2013). Also, with models containing a large enough number of latent classes, one can pick up both simple associations and complex higher-order interactions among the variables in the imputation (McLachlan & Peel,

2000). This makes the model appropriate for datasets coming from large-scale assessment studies, where the number of variables can be large and where association structures can be complex.

Recently, Van der Palm, Van der Ark and Vermunt (2016b) proposed a variant of the LC model called the *divisive latent class model*, which can be used for density estimation and MI. Compared to the standard LC model, this approach reduces computing time enormously. Instead of using frequentist maximum likelihood methods, LC analysis can also be implemented using a Bayesian approach as shown among others by Diebolt and Robert (1994). An interesting recent development concerns the use of Bayesian nonparametric methods for MI. More specifically, inspired by Dunson & Xing's (2009) *mixture of independent normal distribution with Dirichlet process prior*, Si and Reiter (2013) proposed using a nonparametric finite mixture model for MI in a Bayesian framework. In a recent work, Akande, Li and Reiter (2017) evaluated and compared the performance of MICE and DPMM for categorical data imputation by means of an empirical comparison, highlighting the ability of the latter to automatically find the relevant associations in the dataset at hand.

The aim of this chapter is to offer a state-of-the-art overview of MI using LC analysis in which we show similarities and differences and discuss pros and cons of the recently proposed frequentist and Bayesian approaches. The remainder of the chapter is structured as follows. In Section 2.2, the basic LC model is introduced and its use for MI is motivated. Section 2.3 describes the four different LC MI methods in more detail. Section 2.4 illustrates the use the four types LC MI methods in a real-data example, and also compares the obtained results with those obtained with listwise deletion and MICE. Section 2.5 discusses our main findings, gives recommendations for those who have to deal with missing data, and lists topics for further research.

2.2 Latent Class models and Multiple Imputation

2.2.1 Latent Class Analysis for Density Estimation

The latent class model (Lazarsfeld, 1950; Goodman, 1974) is a mixture model which describes the distribution of categorical data. Mixture models are flexible tools that allow modelling the association structure of a set of variables (their joint density) using a finite mixture of simpler densities (McLachlan & Peel, 2000). In LC analysis, each latent class (or mixture component) has its own specific multinomial density, defining the probability of having a specific response pattern. The estimated overall density is obtained as a weighted average of the class-specific densities. An impor-

tant assumption of LC analysis is *local independence* (Lazarsfeld, 1950), according to which the scores of different items are independent of each other within latent classes.

Before discussing the implications of using a LC model as a tool for density estimation, let us first briefly introduce its mathematical form with the aid of a small example. Let y_{ij} be the score of the i-th person on the j-th categorical item belonging to a $n \times J$ data-matrix Y ($i = 1, ..., n$, $j = 1, ..., J$), $\boldsymbol{y_i}$ the J-dimensional vector with all scores of person i, and x_i a discrete (unobserved) latent variable with K categories. In the LC model, the joint density $P(\boldsymbol{y_i}; \boldsymbol{\pi})$ has the following form:

$$\begin{aligned} P(\boldsymbol{y_i}; \boldsymbol{\pi}) &= \sum_{k=1}^{K} P(x_i = k; \boldsymbol{\pi}_x) P(\boldsymbol{y_i} | x_i = k; \boldsymbol{\pi}_y) \\ &= \sum_{k=1}^{K} P(x_i = k; \boldsymbol{\pi}_x) \prod_{j=1}^{J} P(y_{ij} | x_i = k; \boldsymbol{\pi}_{y_j}). \end{aligned} \tag{2.1}$$

The LC model parameters $\boldsymbol{\pi}$ can be partitioned into two sets: the latent class proportions ($\boldsymbol{\pi}_x$) and class-specific item response probabilities ($\boldsymbol{\pi}_y$), where the latter contains a separate set of parameters for each item ($\boldsymbol{\pi}_{y_j}$). The fact that we are dealing with a mixture distribution can be seen from the fact that the overall density is obtained as a weighted sum of the K class-specific multinomial densities $P(\boldsymbol{y_i} | x_i = k; \boldsymbol{\pi}_y)$, where the latent proportions serve as weights. Moreover, in (2.1) the local independence assumption becomes visible in the product over the J independent multinomial distributions (conditional on the k-th latent class).

By setting the number of latent classes large enough, LC models can capture the first, second, and higher-order moments of the J response variables (McLachlan & Peel, 2000), that is, univariate margins, bivariate associations, and higher-order interactions when dealing with categorical variables (Vermunt et al., 2008). Moreover, because of the local independence assumption, it is possible to obtain estimates of the model parameters also when J is very large.

A quantity of interest when using LC models is the units' *posterior class membership probabilities*, i.e., the probability that a unit belongs to the k-th class given the observed data pattern $\boldsymbol{y_i}$. It can be defined through the Bayes' theorem as follows:

$$P(x_i = k | \boldsymbol{y_i}; \boldsymbol{\pi}) = \frac{P(x_i = k; \boldsymbol{\pi}_x) P(\boldsymbol{y_i} | x_i = k; \boldsymbol{\pi}_y)}{P(\boldsymbol{y_i}; \boldsymbol{\pi})}.$$

As an example, suppose we have a data-matrix Y for $J = 5$ binary variables, where the first 3 observations have the observed patterns presented in Figure 2.1a. Suppose furthermore that we specified a 2-class model ($K = 2$) and obtained the

i	Item 1	Item 2	Item 3	Item 4	Item 5
1	1	1	1	1	2
2	2	2	2	2	1
3	1	2	1	2	1
...	...	...	...	...	...

(a)

Class	X = 1 (π_1 = 0.4)		X = 2 (π_2 = 0.6)	
Score	1	2	1	2
Item 1	0.9	0.1	0.1	0.9
Item 2	0.9	0.1	0.1	0.9
Item 3	0.9	0.1	0.1	0.9
Item 4	0.9	0.1	0.1	0.9
Item 5	0.9	0.1	0.1	0.9

(b)

Figure 2.1: *(a) Example of observed data-matrix Y for $J = 5$ dichotomous items and observed patterns y_i for $i = \{1, 2, 3\}$.*
(b) Example of 2-class LC model parameters: latent probabilities π_x (on the top) and conditional probabilities π_{y_j} (in the body of the table).

parameter estimates reported in Figure 2.1b. Looking at Figures 2.1a and 2.1b, it seems that the first observation is more likely to belong to class 1 and the second more likely to belong to class 2. Indeed, for the first observation the class 1 posterior probability $P(x_1 = 1|\boldsymbol{y_1}; \boldsymbol{\pi})$ equals 0.997, whereas for the second observation the class 2 posterior probability $P(x_2 = 2|\boldsymbol{y_2}; \boldsymbol{\pi})$ equals 0.999. The third unit has posteriors $P(x_3 = 1|\boldsymbol{y_3}; \boldsymbol{\pi}) = 0.86$ and $P(x_3 = 2|\boldsymbol{y_3}; \boldsymbol{\pi}) = 0.14$.

2.2.2 Multiple Imputation using LC Models

In a standard LC analysis, the aim is to find a meaningful clustering with a not too large number of well interpretable clusters. In contrast, when used for imputation purposes, the LC model is "just" a device for the estimation of $P(\boldsymbol{y_i}; \boldsymbol{\pi})$. In other words, in MI, LC models do not need to identify meaningful clusters, but instead should yield an as good as possible description for the joint density of the variables in the imputation model. This means that issues which are problematic in a standard LC analysis, such as nonidentifiability, parameter redundancy, overfitting, and boundary parameters, are less of an issue in a MI context. The main thing that counts is whether $P(\boldsymbol{y_i}; \boldsymbol{\pi})$ is approximated well enough in order to be able to generate as good as possible imputations based on $P(\boldsymbol{y_{i,mis}}|\boldsymbol{y_{i,obs}})$.

Specifically, Vermunt et al. (2008) motivate that when a LC model is used as a tool for estimating densities rather than clustering, some differences arise: (a) there is no need to interpret either the parameter estimates or the latent clusters of the latent class imputation model, (b) capturing some sample specific variability (namely overfitting the data) is not problematic in this context, because the aim is to reproduce a sample even with its specific fluctuation, while ignoring certain structures of the data (underfitting) can cause important associations between the variables to be ignored, (c) unidentifiability is not an issue either, inasmuch the quantity of interest $P(\boldsymbol{y_i}; \boldsymbol{\pi})$ is uniquely defined even when the values of $\boldsymbol{\pi}$ are not, and (d) obtaining a local maximum of the log-likelihood function, instead of a global maximum, is also not a problem since the former may provide a representation of $P(\boldsymbol{y_i}; \boldsymbol{\pi})$ that is approximately as good as the one provided by latter.

Once the LC model has been estimated using an incomplete dataset, it is possible to perform MI by randomly drawing m imputations for each nonresponse from the posterior distribution of the missing values given the observed data and the model parameters. To make this clearer, let us return to the small example introduced in the previous section. Suppose now we also have missing values as shown in Figure 2.2, and that under this new scenario the resulting LC 2-class model is again the one with the parameter values presented in Figure 2.1b. With $\boldsymbol{y_{i,obs}}$ we denote the observed part of the response pattern for person i, while the unknown part, marked with "?", is denoted by $\boldsymbol{y_{i,mis}}$. LC model parameter ($\boldsymbol{\pi}$) estimation and inference can be achieved with only the observed information, $\boldsymbol{y_{i,obs}}$. As shown among others by Vermunt et al. (2008), the probability $P(\boldsymbol{y_{i,mis}}|x_i = k; \boldsymbol{\pi}_y)$ cancels from the (incomplete data) log-likelihood function that is maximized, which implies that each subject contributes only to the parameters for the variables which are observed.[2]

Once the model has been estimated, the aim of MI is to generate an imputation for each "?" in the dataset by sampling from $P(\boldsymbol{y_{i,mis}}|\boldsymbol{y_{i,obs}}; \boldsymbol{\pi})$. This requires two draws: the first assigns a class to each unit using the posterior membership probabilities given $\boldsymbol{y_{i,obs}}$. Unit 1, for instance, has now a probability equal to $P(x_1 = 1|\boldsymbol{y_{1,obs}}; \boldsymbol{\pi}) = 0.98$ to belong to class 1 and $P(x_1 = 2|\boldsymbol{y_{1,obs}}; \boldsymbol{\pi}) = 0.02$ to belong to class 2. Once the class membership has been established, "?" in item j is replaced by drawing from the conditional multinomial distribution of j-th item in that class. If, in the previous step, the first unit was allocated to the first class, then the missing value of Item 4 will be replaced by the value 1 with probability 0.9 and by the value 2 with

2 In Vermunt et al. (2008) and Van der Palm et al. (2014) the procedure is given for maximum likelihood methods. For the Bayesian framework, Appendix B.1 shows how the model can be estimated conditional on $\boldsymbol{y_{i,obs}}$ only.

i	Item 1	Item 2	Item 3	Item 4	Item 5
1	1	1	1	?	2
2	2	2	2	2	?
3	1	2	?	2	1
...	...	...	...	...	...

Figure 2.2: *Example of data-matrix Y for $J = 5$ dichotomous items and $i = \{1,2,3\}$, with both observed and missing data (the latter marked by "?").*

probability 0.1. The uncertainty about the imputations is accounted for by repeating this procedure $m > 1$ times for each unit with at least one missing value.

LC models can also be implemented within a Bayesian framework, which involves specifying prior distributions for the class proportions and the class-specific response probabilities. Two kinds of priors can be applied: a Dirichlet distribution or a Dirichlet Process prior. The Dirichlet distribution, used as prior for the multinomial conditional distributions or for the multinomial latent distribution of standard Bayesian LC models, is suited for modelling multivariate quantities that lie in the interval (0,1) and that sum to 1.[3] In the Dirichlet process approach, on the other hand, the number of latent classes becomes uncertain, and a baseline distribution is used as prior expectation density. A concentration parameter (α) rules the concentration of the prior for x_i around the baseline density: when α is large, the prior of x_i is highly concentrated around the expected baseline (the latent classes will tend to have equal sizes), while for small α there is a larger departure from the baseline (few classes will have most of the probability mass) (Congdon, 2006).

In a frequentist setting, maximum likelihood (ML) estimation is typically performed using an EM algorithm (Dempster, Laird & Rubin, 1977), whereas in a Bayesian framework, MCMC algorithms such as the Gibbs sampler are used (Geman & Geman, 1984; Gelfand & Smith, 1990). In mixture models, the Gibbs sampler iterations contain a Data Augmentation step in which units are allocated to latent classes. The Data Augmentation (DA) algorithm (Tanner & Wong, 1987) can be seen as a Bayesian version of the EM algorithm, which can be used for the estimation of Bayesian LC models. DA is particularly suitable also for MI computation as it also involves imputing the missing data given the current state of the model parameter as one of the steps. Tanner and Wong (1987) showed that under certain conditions, the algorithm converges to the true posterior distribution of the unknown quantities of interest. The m imputations are obtained by drawing m imputed scores from

3 For the mathematical formulation of the Dirichlet distribution, see the Appendix A.1.

the posterior distribution of the missing values. A description of both the Gibbs sampler and the DA algorithm is provided in Appendix A.2.

2.3 Four Different Implementations of Latent Class Multiple Imputation

In this section we present four different implementations of LC models for MI: the Maximum Likelihood LC model (MLLC), the standard Bayesian LC model (BLC), the Divisive LC model (DLC), and the Dirichlet Process Mixture of Multinomial distributions (DPMM). These four models share the characteristics of the LC model mentioned in the previous section, which make that each of them can serve an excellent tool for the MI of large datasets containing categorical variables.

These four types of LC models, however, also differ in a number of respects. First, they differ in the way in which they deal with the uncertainty about the model parameters. Note that taking into account this uncertainty during the imputation is a requirement for valid inference with a multiple imputed data set. The two frequentist models (MLLC and DLC) resort either on a nonparametric bootstrap or on different draws of class membership and missing scores, whereas the two Bayesian methods (BLC and DPMM) automatically embed parameter uncertainty by sampling the parameters from their posterior distribution.

Second, the four methods differ in the way they select the number of classes K. While the standard implementation of the LC model (MLLC and BLC) requires estimating and testing a series of models with different numbers of classes using some fit measure (e.g., the AIC), in DLC and DPMM the number of classes is determined in an automatic manner. In DPMM the number of latent classes is treated as a model parameter, while for the other three types of models K is fixed though unknown.

Lastly, the four methods differ in terms of computational efficiency. Note that the main factors affecting computation time are the sample size n, the number of classes K, and the number of variables J. While MLLC and BLC require estimating models with different numbers of classes to determine the required number of classes, DLC and DPMM have the advantage that a good fitting model is obtained in a single estimation run. For this reason, MLLC and BLC turn out to be the computationally most demanding methods, while DLC and DPMM are less demanding. In the remainder of this section, we provide a more detailed description of the four approaches.

2.3.1 Fixed K, Frequentist: the Maximum Likelihood LC Model

The MLLC approach uses a nonparametric bootstrap[4] in order to take into account the uncertainty about the imputation model parameter estimates, which is a requirement for valid post-imputation inference. Specifically, imputation using MLLC proceeds as follows: first, m nonparametric bootstrap samples Y_l^* ($l = 1, ..., m$) of size n are obtained from the original dataset Y; second, the LC model is estimated for each Y_l^*, providing m different sets of parameters $\boldsymbol{\pi}^l$; third, the original dataset is duplicated m times and for the l-th dataset the set of parameters $\boldsymbol{\pi}^l$ is used to impute the missing values from $P(\boldsymbol{y_{i,mis}}|\boldsymbol{y_{i,obs}}; \boldsymbol{\pi}^l)$.

To describe the joint distribution of the data as accurately as possible, K is selected based on penalized likelihood statistics, such as the AIC (Akaike Information Criterion) or the BIC (Bayesian Information Criterion) index. In MI, the AIC criterion is preferable over BIC since it yields a larger number of classes; nevertheless, an even higher K than the one indicated by the AIC index may be used, since, as already noticed, the risk of overfitting in the MI context is less problematic than the risk of underfitting.

Though Vermunt et al. (2008) showed that the performance of MLLC is similar to both *ML for incomplete data* and *MI using a log-linear model*, in terms of parameter bias, some issues with respect to the model-fit strategy remain; in order to select the optimal K value according to the AIC index, in fact, one needs to estimate a 1-class model, a 2-class and so on, until the best fitting model has been found.[5] It will be clear that this approach may be time-consuming, especially when used with large data sets.

MI through MLLC is available in software such as LatentGOLD (Vermunt & Magidson, 2013), which includes a special option for MI. In R, LC analysis can be performed with the package poLCA (Linzer & Lewis, 2014). This package could be used to implement the MI procedure described above.

4 The nonparametric bootstrap (Efron, 1979) is a technique that allows reproducing the distribution of some specific parameter by resampling observations from the original sample multiple times with replacement; in such a way, the original sample is treated as the population of interest. Through this procedure, which is useful when the theoretical distribution of the parameters of interest is difficult to derive, uncertainty about the model parameters can be inferred.

5 Rather than starting with a one class model and subsequently increasing the number of classes, alternative more efficient strategies may be used, such as starting with a large number of classes and both increasing and decreasing this number to see whether a larger number is needed or a smaller number suffices.

2.3.2 Fixed K, Bayesian: the Bayesian LC Model

While in the frequentist framework a nonparametric bootstrap is needed to account for parameter uncertainty, when using a Bayesian MCMC approach parameter uncertainty is automatically accounted for. More specifically, Rubin (1987) recommended using Bayesian methods in order to obtain proper imputations, which fully reflect the uncertainty about the model parameters and which are draws from the *posterior predictive distribution* of the missing data. Vermunt et al. (2008) mentioned the possibility to implement their approach using a Bayesian framework. Si and Reiter (2013) present the Bayesian LC (BLC) model as a natural step to go from the MLLC to the DPMM MI approach. Therefore, though the BLC model has not been proposed explicitly for MI, we present it here as one of the possible implementations of LC-based MI. As in the frequentist case, standard parametric BLC analysis requires that we first determine the value of K, for example, using the AIC index evaluated by ML estimation. Therefore, also with this approach, determining the number of classes may be rather time consuming in larger data sets.

For the distribution of $\boldsymbol{\pi}_x$ the prior will typically be a K-variate Dirichlet distribution (if K=2 this is equal to a Beta distribution), whereas for the conditional probabilities $\boldsymbol{\pi}_{y_j}$, a Dirichlet prior for each $j = 1, ..., J$ and $k = 1, ..., K$, with number of components equal to the number of categories of the j-th variable, is assumed. Setting weakly informative prior distributions helps the posterior distribution of $\boldsymbol{\pi}$ to be data dominated. For the Dirichlet distribution, an uniform prior is achievable by initializing all its parameters to 1.[6] Within the latent classes, the conditional probabilities are initialized to be equal to the observed marginal frequencies of the scores of each variable. Also, for MI, nonresponses are initialized with a random draw from the observed frequency distribution of the variables with missing values. Once the first set of $\boldsymbol{\pi}_x$ has been drawn from the Dirichlet prior, the Gibbs sampler proceeds as follows. First, each unit is assigned to a latent category by drawing from the posterior membership probabilities $P(x_i = k|\boldsymbol{y_i}; \boldsymbol{\pi})$; second, the parameters of the Dirichlet distribution for $\boldsymbol{\pi}_x$ are updated: this is done by adding the number of units dropped in the k-th latent class to the starting value of the k-th parameter (that is 1 in the case of a weakly informative prior). From this updating, a new value of $\boldsymbol{\pi}_x$ is extracted. Third, the parameters of the Dirichlet distributions of $\boldsymbol{\pi}_{y_j}$ are in turn updated in an analogue way: the number of units which take on one of the possible observed values of the j-th variable and dropped into the k-th latent class is added to the initial parameter value of the category concerned of the

6 This is equivalent to a prior sample size equal to the number of components of the Dirichlet distribution. Setting the Dirichlet prior with all its parameters equal to 1 is a common choice (Congdon, 2006) which yields an uniform, but not necessarily uninformative, distribution. Jeffrey's uninformative prior can be obtained by initializing all the parameters of the Dirichlet distribution equal to 1/2.

j-th Dirichlet prior of the k-th latent component (again, this is 1 in the case of a weak prior); after the updating, a new value of $\boldsymbol{\pi}_{y_j}$ is drawn. The fourth, and last, step is the imputation step: given the value $x_i = k$ of each unit (resulting from the first step), and the new set of probabilities $\boldsymbol{\pi}_{y_j}$, a new score for $y_{ij,mis}$ is drawn from $P(y_{ij}|x_i = k; \boldsymbol{\pi}_{y_j})$. Steps 1-4 are repeated until convergence is reached. Appendix B.1 gives a formal description of these steps.

A BLC model can be estimated in R through the package "BayesLCA" (A. White & Murphy, 2014).

2.3.3 Unknown K, Frequentist: the Divisive LC Model

The main problem of the standard LC approach is that it uses a substantial amount of computation time to estimate multiple models with increasing number of classes to determine the value of K. Divisive latent class (DLC) models (Van der Palm et al., 2016b) overcome this problem by breaking down the global estimation problem into a series of smaller local problems. The DLC model incorporates an algorithm that increases the number of latent classes within a single run until the possible improvements in model fit have been achieved. This implies that the best fitting model is found in a single estimation run. The DLC model has been developed by Van der Palm et al. (2016b, 2014) for density estimation and MI purposes, while a substantive interpretation of the resulting LC parameters is still unexplored.

The DLC algorithm involves evaluating a series of 1-class and 2-class models. At the start, a single LC assumed to contain the whole sample is split into two latent classes if the 2-class model improves the model fit sufficiently (for instance, in terms of log-likelihood). If this is the case, every unit will have a probability of belonging to each of the two latent classes, which corresponds to the posterior class membership probabilities. Using these posterior probabilities, two fuzzy subsamples are created. In the following step, these two new latent classes are checked separately to establish whether a further split into 2 classes, within each subsample improves the model fit. In the next steps, this operation is repeated for each newly formed latent class, until the best model fit is achieved for every fuzzy subsample. Since a DLC model is estimated sequentially, each submodel created at step s builds on the results of steps $1, ..., s-1$; in such a way an automatic estimate of the optimum K is obtained with much smaller computation time compared to the MLLC approach. Van der Palm et al. (2016b) discussed various decision rules to determine whether the improvement in model fit is large enough to accept a split of a latent class. Their advice is to use a stop-criterion based on the increase in the log-likelihood values

between the 1-class and 2-class model for a particular fuzzy subsample. For further technical details, we refer to Van der Palm et al. (2016b).

Van der Palm et al. (2014) observed that the DLC model in combination with the nonparametric bootstrap may yield biased parameter estimates in a subsequent substantive analysis. Therefore, they proposed implementing the actual MI procedure in slightly different from MLLC, while still taking into account the uncertainty about the imputation model parameters. First, the DLC model and its parameters $\boldsymbol{\pi}$ is estimated using the original dataset; second, the posterior membership probabilities $P(x_i = k|\boldsymbol{y_{i,obs}}; \boldsymbol{\pi})$ are computed; third, the original dataset is duplicated m times; fourth, the $P(x_i = k|\boldsymbol{y_{i,obs}}; \boldsymbol{\pi}^l)$ are used to assign m times a latent class to each respondent; last, for each missing value of unit i in item j, m missing scores are sampled using the conditional response probabilities $P(y_{ij}|x_i = k; \boldsymbol{\pi})$.

The Latent GOLD software allows performing DLC-based MI, while to our knowledge currently there is no R package that implements the DLC approach.

2.3.4 Variable K, Bayesian: Dirichlet Process Mixture of Products of Multinomial Distributions

Even if the AIC index provides a sufficiently large number of mixture components, once the value of K is determined uncertainty about K is ignored when generating the imputations. This counters Rubin (1987)'s suggestion to account for all possible uncertainties about the imputation model parameters in order to avoid underestimation of the variances of the substantive model parameters (Si & Reiter, 2013). The Dirichlet process mixture of products of multinomial distributions (DPMM) overcomes the need of an ad hoc selection of a fixed K and, moreover, automatically deals with the uncertainty about this parameter. This happens by assuming that in theory there is an infinite number of classes ($K = +\infty$), but letting the data fill only a smaller number of components that is actually needed. A simulation study by Si and Reiter (2013) showed that DPMM MI may outperform MICE in terms of bias and confidence interval coverage rates of the parameters of a substantive model.

DPMM offers a full Bayesian modeling approach for high-dimensional categorical data. Similarly to BLC, DPMM can be estimated through the Gibbs sampler. One of the possible conceptualization of the Dirichlet process which serves as a prior for the mixture proportions $\boldsymbol{\pi}_x$ is the *stick-breaking process* (Sethuraman, 1994; Ishwaran & James, 2001). In this formulation, an element of $\boldsymbol{\pi}_x$, say π_k ($k = 1, ..., +\infty$), is assumed to take on the form $\pi_k = V_k \prod_{h<k}(1 - V_h)$ for each k, where every V_k is drawn from a Beta distribution with parameters $(1, \alpha)$. Here, α, the concentration parameter of the process, is allowed to vary according to a Gamma distribution with

parameters (a, b). The conditional responses (and their prior) keep the same distributional form as in the BLC model, that is, multinomial densities with Dirichlet priors. Also in this case, it is possible to set weakly informative priors for the model parameters; Dunson & Xing's (2009) suggestion for weak priors is to initialize α to be equal to 1 and set the parameters of its Gamma distribution to $a = b = 0.25$. This allows each V_k to be uniformly distributed in the (0,1) range, whereas the Dirichlet priors of the conditional distributions can be made uniform by setting all their parameters to 1 (as we already saw for the BLC approach). Since the stick-breaking specification of the Dirichlet process incentivizes the size of each latent class π_k to decrease stochastically with k, this model tends to put meaningful posterior probabilities on a limited number of components automatically determined by the data. When the concentration parameter α is small, in fact, most of the probability mass is assigned to the first few components, with the number of significant components increasing as α increases. As a consequence, there will be a finite number of classes with a meaningful size, while the classes with a negligible probability mass will be ignored.

Since working with an infinite number of classes is impossible in practice, Si and Reiter (2013) proposed truncating the stick-breaking probabilities at an (arbitrarily) large K^*, but not so large as to compromise the computing speed. If, after running the MCMC chain, significant posterior masses are observed for all K^* components, the truncation limit should be increased. As for the BLC approach, conditional probabilities of the J variables within each latent class are initialized with the observed frequencies and, for MI, missing data are initialized too with draws from these frequency tables. The Gibbs sampler is then performed as follows. First, each unit is assigned to a latent category by drawing from the posterior membership probabilities $P(x_i = k|\boldsymbol{y_i}; \boldsymbol{\pi})$; second, V_k ($k = 1, ..., K^* - 1$ because of the truncation) are drawn from a Beta distribution, whose first parameter is updated by adding the number of units allocated in the k-th latent class to its initial value (set to 1), whereas α is updated by adding to it the number of units assigned to the latent classes which go from $k + 1$ to K^*; after setting $V_{K^*} = 1$, each π_k is calculated through the formula $\pi_k = V_k \prod_{h<k}(1 - V_h)$; in the third step the parameters of the conditional Dirichlet distributions of $\boldsymbol{\pi}_{y_j}$ are updated by adding the number of units, which take one of the possible observed values of the j-th variable and are dropped into the k-th latent class, to the initial parameter value of the related component of that distribution; after the updating, a new value of $\boldsymbol{\pi}_{y_j}$ is drawn; fourth, a new value for the concentration parameter α is drawn from the Gamma distribution with parameters updated as $a + K^* - 1$ and $b - log(\pi_{K^*})$; fifth, the imputation step analogue to BLC is performed. Steps 1-5 are repeated until convergence is reached. For a formal description of the algorithm, see Appendix B.2.

Table 2.1: Differing features of the four LC models for MI.

Method	Parameters uncertainty	K-handling	Time-consuming
MLLC	·nonparametric bootstrap	·fixed, determined a priori, AIC criterion	·Yes, estimation of multiple models
BLC	·embedded in the posterior distributions	·fixed, determined a priori, AIC criterion	·Yes, estimation of multiple models
DLC	·m different draws through the estimated model	·fixed, unknown a priori, automatically determined by algorithm	·No, best fitting model achieved in a single run
DPMM	·embedded in the posterior distributions	·uncertain, varying, ruled by the data	·No, best fitting model achieved in a single run

To our knowledge no off-the-shelf software is currently available that enables estimation of the DPMM model. We implemented a custom routine in R to fit the model. The R-code is available from the corresponding author upon request.[7]

Table 2.1 summarizes the main differences of the four models described in this section. In the next section, we are going to apply the LC MI models to a real-data example in order to show their working in the practice. We will examine similarities and differences between the four methods and also with listwise deletion and MICE.

2.4 Real-data Example

The KRISTA dataset (Van den Broek, Nyklicek, Van der Voort, Alings & Denollet, 2008) contains self-reported and interviewer-rated information from 748 patients aged between 18 and 80 years who got an Implantable Cardioverter Defibrillator (ICD) in two large Dutch referral hospitals between May 2003 and February 2009. The aim of the study was to determine whether personality factors affect the occurrence of anxiety as a result of the shocks the patients gets from the ICD. We selected the items of four scales to illustrate the application of MI: Eysenck Personality Questionnaire (EPQ, 24 binary items scored 0-1, 12 of which measure patient's neuroticism -EPQN- and the remaining 12 measure patient's extraversion -EPQE), Marlowe Crowne Scale (MC, 30 binary items scored 0-1), State-Trait Anxiety Inventory (STAI, 20 items on a 4-point Likert scale), and Anxiety Sensitivity Index (ASI, 16 items ranging from 0 to 4). We included in the analysis also the categorical

7 The code has been written and implemented to run with the example of Section 2.4, but it has been not validated with other data sets yet.

background variable Sex, yielding a total of $J = 91$ variables. After removing the persons without any observed score on the 90 questionnaire items, we have a sample size of $n = 706$ patients, $n_M = 555$ of which are men and $n_F = 151$ of which are women. Although in this reduced dataset the total percentage of missingness was very low (2.4%), it should be noticed that a method such as listwise deletion (LD) may cause a large amount of loss of power, since about 30% of the units contained at least one missing value, resulting in only $n^* = 494$ persons with fully observed information ($n_M^* = 400$ males and $n_F^* = 94$ women).

We also created a version of the same dataset with some extra MAR missingness.[8] The new total rate of missingness was about 22.5%. In this new case, only $n^{**} = 109$ units had a completely observed response pattern (of which $n_M^{**} = 96$ males and $n_F^{**} = 13$ women) while the remaining $n - n^{**} = 597$ cases (84.56 % of the units) had at least one missing value. This data set with a much larger percentage of missing values will be used to investigate whether and how the behaviour of the missing data models differs compared to the original low missingness situation.

Case I - Low missingness. We applied LD and MICE and the four LC MI methods to the original dataset. Subsequently, we computed the estimates of various quantities of interest for the resulting complete data sets. For the scales we selected, we obtained Cronbach's alpha ($\hat{\alpha}$), the means for males and females ($\hat{\mu}_M$ and $\hat{\mu}_F$) and their standard errors ($\hat{\sigma}_{\mu_M}$, and $\hat{\sigma}_{\mu_F}$), the t-value of the test for assessing the hypothesis of equality of means between men and women (against the alternative hypothesis $H_1 : \hat{\mu}_M \neq \hat{\mu}_F$) and the resulting p-value.

Note that the purpose of our example is to illustrate the use of the LC-based MI approaches with a real life application. Contrary to the controlled conditions of a simulation study, we do not know the true values of the quantities of interest. Instead, we will compare the estimates obtained with different imputation methods, as well as compare the estimates obtained in the low missingness condition (Case I) with those in the high missingness condition (Case II). For elaborate simulation studies on the behavior of the LC imputation models, we refer to Vermunt et al. (2008); Van der Palm, Van der Ark and Vermunt (2016a); Gebegziabher and DeSantis (2010); Si and Reiter (2013).

We applied MICE with its default setting using the R library (Van Buuren et al., 2014) and ran it for 15 iterations. For MLLC and BLC, we specified two kind of models, one resulting from the selection of K based on the AIC index and the other using an arbitrarily large value for K. Models specified through the AIC index will be denoted by MLLC(AIC) and BLC(AIC), while models with a large K will be

8 For the generation of the extra missingness, we followed Brand (1999) and Van Buuren, Brand, Groothuis-Oudshoorn and Rubin (2006). Appendix C details the procedure adopted.

denoted by MLLC(large) and BLC(large). For the former, we estimated a 1-class model, a 2-class model, and so on, up to a 70-class model. The best fitting model, according to the AIC index, was the 14-class model. MLLC(large) and BLC(large) were implemented with $K = 50$. Furthermore, we used the 1-class MLLC model (MLLC(1), an independence model), which is in fact a random version of mean (or mode) imputation. We used the MLLC(1) model to show the consequence of using an imputation model that does not correctly model the associations between the variables in the data file. The DLC model was estimated with a decision rule based on the improvement in log-likelihood larger than $0.6 \cdot J$, following Van der Palm et al.'s (2014) advice. This resulted in a model with $K = 111$ classes. DPMM, finally, was implemented with $K^* = 50$ truncated components. For BLC and DPMM, the Gibbs samplers were run with $B = 50000$ iterations and with the prior specifications described in Section 2.3.

Model-estimation and imputation was performed with LatentGOLD 5.0 (Vermunt & Magidson, 2013) for MLLC and DLC, while we implemented two routines in R 3.0.2 for the Gibbs samplers of BLC and DPMM. Following Graham, Olchowski and Gilreath (2007) we used $m = 20$ imputations for each method (including MICE). R 3.0.2 was used to obtain estimates for the parameters of interest with LD and the MI methods (pooled estimates for the latter).

Table 2.2 reports $\hat{\alpha}$, $\hat{\mu}_M$, $\hat{\mu}_F$,$\hat{\sigma}_{\mu_M}$, $\hat{\sigma}_{\mu_F}$, t-values, and p-values for each method. The $\hat{\sigma}_{\mu_M}$ and $\hat{\sigma}_{\mu_F}$ obtained with the MI methods reflect both the "within imputation" and the "between imputation" variability of the estimates of the population means. T-values were also calculated taking into account both the sources of variability. Null hypotheses rejected at the significance level of 5% are marked in boldface.

As can be seen, the estimates obtained with the different LC-MI implementations are all very similar. However, the estimates provided by the MI methods appear to differ systematically from the estimates of the LD method. For example, the $\hat{\alpha}$ estimates for the EPQN and ASI scales obtained with the MI approaches are always larger than the ones for LD, but the differences among the LC models (both frequentist and Bayesian) are very small. Also some differences between MICE and LC imputation methods can be observed. For example, the Cronbach's alpha of the EQPN and ASI scales of MICE are not only larger than those of LD, but also somewhat larger than those of the LC methods.

Also for $\hat{\mu}_M$ and $\hat{\mu}_F$, differences between the LC imputation models are very small. For instance, the mean of men's scores on the EPQN scale provided by DLC is only slightly larger than the ones provided by MLLC(AIC), MLLC(large), BLC(AIC), and BLC(large), the latter ones being very similar to one another, while DPMM yields an estimate that may appear somewhat different. Actually, it seems as if the LC-

Table 2.2: Final estimates of the quantities investigated on the KRISTA dataset (original missingness). p-values are based on the t-test with $n^* - 2 = 494$ df for the LD method, and with degrees of freedom calculated according to MI rules for MI methods. Significant 5% p-values are marked in boldface.

		Missing data model								
	Scale	LD	MICE	MLLC(1)	MLLC(AIC)	MLLC(large)	DLC	BLC(AIC)	BLC(large)	DPMM
$\hat{\alpha}$	EPQN	0.833	0.861	0.845	0.850	0.850	0.850	0.850	0.850	0.850
	EPQE	0.873	0.865	0.860	0.864	0.863	0.864	0.863	0.863	0.862
	MC	0.759	0.763	0.732	0.735	0.736	0.736	0.734	0.735	0.736
	STAI	0.944	0.944	0.942	0.945	0.945	0.945	0.945	0.945	0.945
	ASI	0.886	0.900	0.890	0.894	0.892	0.894	0.893	0.894	0.894
$\hat{\mu}_M$	EPQN	8.802	8.480	8.612	8.593	8.606	8.610	8.598	8.597	8.589
	EPQE	4.832	4.939	4.867	4.903	4.866	4.878	4.881	4.879	4.885
	MC	20.467	20.269	20.469	20.474	20.468	20.470	20.447	20.445	20.448
	STAI	37.355	38.652	38.179	38.241	38.237	38.180	38.227	38.217	38.224
	ASI	12.847	13.547	13.260	13.328	13.314	13.351	13.337	13.375	13.367
$\hat{\mu}_F$	EPQN	8.010	7.352	7.535	7.521	7.517	7.509	7.510	7.524	7.520
	EPQE	5.223	5.272	5.138	5.113	5.176	5.148	5.133	5.117	5.126
	MC	22.032	21.467	21.759	21.732	21.807	21.756	21.736	21.742	21.736
	STAI	39.053	41.203	40.730	40.748	40.663	40.711	40.738	40.674	40.687
	ASI	13.979	15.591	15.108	15.272	15.150	15.185	15.261	15.252	15.276
$\hat{\sigma}_{\mu_M}$	EPQN	0.153	0.142	0.135	0.137	0.137	0.136	0.137	0.137	0.136
	EPQE	0.179	0.152	0.149	0.151	0.150	0.151	0.150	0.150	0.150
	MC	0.228	0.197	0.186	0.187	0.187	0.188	0.187	0.187	0.187
	STAI	0.567	0.510	0.491	0.499	0.496	0.497	0.498	0.497	0.497
	ASI	0.481	0.436	0.417	0.422	0.419	0.424	0.422	0.425	0.424
$\hat{\sigma}_{\mu_F}$	EPQN	0.336	0.284	0.274	0.278	0.277	0.277	0.278	0.279	0.278
	EPQE	0.356	0.273	0.262	0.264	0.264	0.266	0.266	0.265	0.263
	MC	0.415	0.365	0.324	0.327	0.324	0.327	0.327	0.325	0.329
	STAI	1.105	0.937	0.905	0.930	0.927	0.928	0.930	0.932	0.930
	ASI	0.936	0.885	0.845	0.864	0.849	0.859	0.859	0.861	0.862
t-value	EPQN	2.226	3.638	3.642	3.577	3.631	3.685	3.632	3.562	3.566
	EPQE	-0.957	-1.027	-0.851	-0.654	-0.972	-0.842	-0.785	-0.744	-0.754
	MC	-3.056	-2.809	-3.258	-3.171	-3.373	-3.222	-3.242	-3.272	-3.236
	STAI	-1.319	-2.333	-2.425	-2.341	-2.273	-2.368	-2.345	-2.297	-2.302
	ASI	-1.036	-2.140	-2.027	-2.096	-2.001	-1.977	-2.081	-2.018	-2.053
p-value	EPQN	**0.026**	**0.0003**	**0.0003**	**0.0004**	**0.0003**	**0.0002**	**0.0003**	**0.0004**	**0.0004**
	EPQE	0.339	0.305	0.395	0.513	0.332	0.400	0.433	0.457	0.451
	MC	**0.002**	**0.005**	**0.001**	**0.002**	**0.0008**	**0.001**	**0.001**	**0.001**	**0.001**
	STAI	0.188	**0.020**	**0.016**	**0.019**	**0.023**	**0.018**	**0.019**	**0.022**	**0.022**
	ASI	0.301	**0.033**	**0.043**	**0.036**	**0.046**	**0.048**	**0.038**	**0.044**	**0.040**

Note: LD = listwise deletion method; MICE = MI by chained equations method; MLLC(1) = MLLC imputation method with 1 latent class; MLLC(AIC) = MLLC imputation method with number of latent components determined by the AIC index; MLLC(large) = MLLC imputation method with an arbitrarily large number of latent components; DLC = DLC imputation method; BLC(AIC) = BLC imputation method with number of latent components determined by the AIC index; BLC(large) = BLC imputation method with an arbitrarily large number of latent components; DPMM = DPMM imputation method. / $\hat{\alpha}$: values for Cronbach's alpha; $\hat{\mu}_M$ = means of the total scores of the men; $\hat{\mu}_F$ = means of the totals scores of the women; $\hat{\sigma}_{\mu_M}$ and $\hat{\sigma}_{\mu_F}$ = standard errors of the means of the total scores of men and women; t-values and p-values refer to the test $H_0 : \hat{\mu}_M = \hat{\mu}_F$ vs. $H_1 : \hat{\mu}_M \neq \hat{\mu}_F$.

MI models produce estimates that differ mainly because randomness involved in the methods (parameter draws and imputation draws). Probably, if we ran these methods again, we would obtain slightly different estimates, but without important differences from the ones reported in Table 2.2. Differences between LD, MICE and LC-MI estimates are larger than the differences among the various LC MI methods. As far as MICE is concerned, it can be seen that the difference in estimated means between MICE and LD is usually larger than the difference between LC-MI and LD.

If we look at the SE estimates, the LC-MI procedures seem to yield somewhat smaller value than MICE and LD (which is disadvantaged by a smaller sample size). Furthermore, SEs are very similar across LC methods. Differences between LD and the MI methods turn out to be important for the t-tests: while we rejected only 2 null hypotheses (EPQN and MC) with LD, we have 4 out of 5 rejections (EPQN, MC, STAI and ASI) with all MI methods investigated.

It is also possible to see from Table 2.2 that the independence model, MLLC(1), does not produce very different results compared to the other LC MI models. The main difference occurred in the estimates of $\hat{\alpha}$, which are slightly lower than the Cronbach's alpha produced by the other LC-MI methods. The other quantities do not differ much from those obtained with the others LC imputation models. Seemingly, with this low rate of missingness, it is more important to prevent deleting cases with missing values than to have "correct" imputations for the missing values.

Given the similar results produced by the MI methods, a look at the computation times in Table 2.3 may be useful for a further comparison. For the MLLC approach, the required computation time to estimate models with fewer classes is also reported. Estimation of MLLC models with 1 up to 70 classes took almost 13 hours. For BLC and DPMM, we report the computation time required to run the Gibbs sampler for one model. The time spent on estimating all MLLC models should be added to the computation time to run the Gibbs sampler for BLC(AIC). Running the MICE with (only) 15 iterations required about 13 hours. Among the LC imputation methods, MLLC and BLC(AIC) are more time-consuming than DLC, BLC(Large), and DPMM, which are faster and took about the same computation time, as they do not require the estimation of multiple models to find the ideal number of classes.

Case II - High missingness. Table 2.4 reports the estimates obtained using the KRISTA dataset with extra (22.5%) missingness. The settings were the same as with Case I, except for the number of classes of MLLC(AIC) and BLC(AIC), which was $K = 10$, and the number of classes of DLC, which was $K = 106$. The LD method was applied with $n^{**} = 109$ persons with fully observed score patterns.

Table 2.3: Computation time for MI using MICE and the six different LC imputation models.

Imputation model	Model time	Total time
MICE*	/	13h05min
MLLC(AIC)**	0h58min	12h51min
MLLC(large)**	7h17min	12h51min
DLC	5h39min	5h39min
BLC(AIC)***	6h04min	18h55min
BLC(large)	6h41min	6h41min
DPMM	6h27min	6h27min

Note: *MICE was run for 15 iterations. **MLLC models were estimated from MLLC(1) to MLLC(70). The second column shows the required time to estimate the indicated model, while the third column shows the computation time taken to estimate all the 70 models. ***For BLC, in the second column the computation time needed to run the Gibbs sampler has been reported, while in the third column the computation time of MLLC for selecting the number of classes has been added.

As can be seen from Table 2.4, the contrast between LD and the MI methods, as well as the differences between MICE, the 1-class LC model, and the other LC models, are much clearer now. This shows that the way the imputation is performed matters with larger proportions of missing values. All LC imputation methods recover $\hat{\mu}_M$ and $\hat{\mu}_F$ well; that is, estimates of these means are similar or very close to those of the low-missingness case. Also the estimated standard errors of the means, $\hat{\sigma}_{\mu_M}$ and $\hat{\sigma}_{\mu_F}$, do not differ much from the previous case, though they are slightly smaller than for Case I. Notice, furthermore, that the MLLC(1) model yielded standard errors that are much smaller than the other methods, showing that an under-specified model will typically underestimate variability. The t-tests with MLLC(large), BLC(large) and DPMM yielded the same conclusions as with Case I, as 4 out of 5 tests are rejected at a significance level of 5 %. MLLC(AIC), DLC, and BLC(AIC) did not reject the hypothesis of equality of means for the ASI scale, which is result of the slightly lower power in the high missingness condition. LD seems to produce very much biased means and large standard errors (the latter resulting from the strongly reduced sample size). The MICE standard errors are similar those of the LC-MI methods, except for MLLC(1). However, the MICE estimated means are not only rather different from the LC-MI estimates, but also from MICE estimates for Case I. The largest differences are encountered for STAI and ASI.

As far as the LC-MI methods is concerned, larger differences between Case I and Case II occurred for Cronbach's alpha; that is, in the high missingness condition, the $\hat{\alpha}$ estimates are lower than in the low missingness condition. MLLC(1) produced the lowest values of $\hat{\alpha}$. The other methods are very similar to each other, but all

Table 2.4: Final estimates of the quantities investigated on the KRISTA dataset (extra missingness). p-values are based on the t-test with $n^{**} - 2 = 107$ df for the LD method, and with degrees of freedom calculated according to MI rules for MI methods. Significant 5% p-values are marked in boldface.

		Missing data model								
	Scale	LD	MICE	MLLC(1)	MLLC(AIC)	MLLC(large)	DLC	BLC(AIC)	BLC(large)	DPMM
$\hat{\alpha}$	EPQN	0.529	0.798	0.767	0.833	0.830	0.829	0.828	0.823	0.825
	EPQE	0.851	0.789	0.748	0.822	0.831	0.830	0.810	0.814	0.770
	MC	0.728	0.744	0.632	0.668	0.699	0.698	0.658	0.676	0.657
	STAI	0.919	0.903	0.892	0.941	0.937	0.938	0.935	0.935	0.936
	ASI	0.818	0.827	0.805	0.874	0.870	0.873	0.859	0.862	0.864
$\hat{\mu}_M$	EPQN	9.615	7.569	8.588	8.576	8.594	8.614	8.568	8.596	8.570
	EPQE	4.167	5.310	4.809	4.877	4.890	4.821	4.822	4.838	4.806
	MC	20.969	18.198	20.620	20.663	20.592	20.618	20.628	20.580	20.582
	STAI	36.365	40.890	38.485	38.211	38.390	38.106	38.330	38.314	38.324
	ASI	11.812	18.311	13.374	13.312	13.337	13.203	13.536	13.652	13.502
$\hat{\mu}_F$	EPQN	9.462	6.598	7.683	7.699	7.680	7.708	7.712	7.692	7.694
	EPQE	4.077	5.456	4.991	5.127	5.041	5.031	5.082	5.045	4.978
	MC	21.462	19.252	21.495	21.564	21.567	21.589	21.488	21.505	21.477
	STAI	37.615	43.520	40.598	40.592	40.650	40.638	40.786	40.592	40.656
	ASI	13.846	19.787	14.937	14.970	15.109	14.861	15.223	15.396	15.367
$\hat{\sigma}_{\mu_M}$	EPQN	0.190	0.137	0.120	0.135	0.133	0.133	0.134	0.132	0.135
	EPQE	0.334	0.141	0.124	0.146	0.142	0.145	0.140	0.143	0.135
	MC	0.428	0.216	0.167	0.169	0.180	0.183	0.175	0.179	0.173
	STAI	1.009	0.501	0.410	0.492	0.485	0.485	0.486	0.489	0.486
	ASI	0.761	0.452	0.353	0.408	0.401	0.409	0.407	0.400	0.407
$\hat{\sigma}_{\mu_F}$	EPQN	0.475	0.268	0.246	0.273	0.271	0.269	0.269	0.269	0.270
	EPQE	0.866	0.258	0.220	0.245	0.252	0.256	0.253	0.251	0.227
	MC	1.169	0.376	0.299	0.316	0.319	0.328	0.320	0.318	0.319
	STAI	1.950	0.918	0.771	0.914	0.893	0.915	0.906	0.920	0.908
	ASI	2.292	0.904	0.756	0.845	0.854	0.841	0.830	0.848	0.868
t-value	EPQN	0.281	3.247	3.462	2.972	3.131	3.138	2.948	3.151	3.001
	EPQE	0.093	-0.484	-0.685	-0.830	-0.498	-0.673	-0.872	-0.684	-0.619
	MC	-0.398	-2.308	-2.475	-2.436	-2.539	-2.516	-2.317	-2.475	-2.405
	STAI	-0.441	-2.470	-2.397	-2.258	-2.174	-2.429	-2.366	-2.167	-2.237
	ASI	-0.911	-1.531	-2.011	-1.859	-1.980	-1.857	-1.933	-1.972	-2.066
p-value	EPQN	0.779	**0.001**	**0.0006**	**0.0003**	**0.002**	**0.002**	**0.003**	**0.002**	**0.003**
	EPQE	0.926	0.629	0.494	0.407	0.618	0.501	0.383	0.494	0.536
	MC	0.692	**0.021**	**0.014**	**0.015**	**0.011**	**0.012**	**0.021**	**0.014**	**0.016**
	STAI	0.660	**0.014**	**0.017**	**0.024**	**0.030**	**0.015**	**0.018**	**0.031**	**0.026**
	ASI	0.364	0.126	**0.045**	0.063	**0.048**	0.064	0.054	**0.049**	**0.039**

Note: LD = listwise deletion method; MICE = MI by chained equations method; MLLC(1) = MLLC imputation method with 1 latent class; MLLC(AIC) = MLLC imputation method with number of latent components determined by the AIC index; MLLC(large) = MLLC imputation method with an arbitrarily large number of latent components; DLC = DLC imputation method; BLC(AIC) = BLC imputation method with number of latent components determined by the AIC index; BLC(large) = BLC imputation method with an arbitrarily large number of latent components; DPMM = DPMM imputation method. / $\hat{\alpha}$: values for Cronbach's alpha; $\hat{\mu}_M$ = means of the total scores of the men; $\hat{\mu}_F$ = means of the totals scores of the women; $\hat{\sigma}_{\mu_M}$ and $\hat{\sigma}_{\mu_F}$ = standard errors of the means of the total scores of men and women; t-values and p-values refer to the test $H_0 : \hat{\mu}_M = \hat{\mu}_F$ vs. $H_1 : \hat{\mu}_M \neq \hat{\mu}_F$.

Table 2.5: Comparsion of $\hat{\alpha}$ (MC scale) estimated after performing MI only on items of MC scale.

Imputation model	$\hat{\alpha}(MC)$
MICE	0.743
MLLC(1)	0.631
MLLC(AIC)	0.729
MLLC(large)	0.727
DLC	0.725
BLC(AIC)	0.725
BLC(large)	0.728
DPMM	0.728

smaller than in Case I. Especially the alpha value for the MC scale is quite a bit lower. The fact that $\hat{\alpha}$ seems to be underestimated indicates that the LC MI models have some difficulties in capturing and describing the complex associations among the 91 variables used in the imputation model. MICE provides a Cronbach's alpha value closer to the estimates of Case I than the LC methods for the MC scale, but for the other scales MICE seems to yield larger downward biased alpha values than the LC-MI methods.

In order to see whether focusing on a single scale improves the estimate of Cronbach's alpha, we performed a separate MI with MICE and the LC methods for the 30 items of the MC scale (the scale with the worst results in terms of $\hat{\alpha}$, compared with the results of Table 2.2). From Table 2.5 it can be seen that MLLC(AIC), MLLC(large), DLC, BLC(AIC), BLC(large), and DPMM are doing much better now, their estimates being much closer to those of Table 2.2. MLLC(1), on the other hand, is still doing badly, which confirms that it is an inadequate imputation model. MICE produced a Cronbach's alpha identical to the one with all 91 variables.

2.5 Discussion

This chapter offered a state-of-the-art overview on the use of LC models as tools for MI. One feature that makes LC models attractive imputation tools for psychological assessment studies is that they do not require complex model specification, since only the specification of the number of classes, K, is needed. Second, LC models can efficiently be computated even when dealing with a large number of

variables. Third, by selecting a large enough number of classes, LC models can pick up complex associations in high-dimensional datasets.

Four possible LC implementation for MI were described: the Maximum Likelihood LC (MLLC), the Bayesian LC (BLC), the Divisive LC (DLC), and the Dirichlet Process Mixture of Multinomial Distributions (DPMM) approaches. While sharing the attractive features of LC modeling for MI, these methods differ in various ways. One is how they account for the uncertainty about the imputation model parameters: whereas MLLC uses a nonparametric bootstrap and DLC draws m unit class-memberships from the estimated model, the Bayesian methods (BLC and DPMM) draw parameters from their posterior distribution. Second, the decision regarding the number of classes K is handled differently by the four approaches. MLLC and BLC require model comparison through for example the AIC, DLC determines K in a single run of its sequential algorithm, and DPMM leaves the number of classes unspecified. In MLLC and BLC, it is also possible to set K to an arbitrary large value, which makes them more similar to DLC and DPMM, also in terms of computation time.

We illustrated the use of the LC imputation methods and compared them with listwise deletion and MICE using a dataset with 91 categorical variables from a psychological assessment study. We looked at two situations: the original situation with a low rate of missingness and a situation with a much higher rate of missingness obtained by creating additional missing values. In the first situation, the various types of LC imputation models yielded very similar results; that is, similar Cronbach's alpha values, means for men and women, standard errors and t-tests. However, the fact that the results obtained with the 1-class imputation model were also similar but those obtained with listwise deletion different indicated that in this low missingness case it was more important to keep the records with missing values than to have a correct imputation model. MICE imputation supplied estimates very similar to those of the LC models, although minor systematic differences appeared between these two different types of imputation methods.

The differences between LC imputation with both the under-specified model and MICE were much larger in the high missingness situation. The estimates for Cronbach's alpha and the standard errors of the means were smaller (too small) in the 1-class model, showing that the imputation model matters. Furthermore, LC imputation methods introduced less bias in the estimates of means and standard errors than MICE in Case II, whereas MICE appeared to better recover the alpha for one scale but worse for the other four scales.

When comparing the LC-MI estimates of Cronbach's alpha between the low and high missingness condition, we saw that alpha was underestimated with more miss-

ing values, where the degree of this underestimation varied per scale. This shows that also the LC-MI methods do not pick up perfectly the variability in and the associations between the variables in the dataset. When performing the imputation for a single subscale rather than for all 91 variables simultaneously, the LC MI models yielded much better estimates for Cronbach's alpha. Capturing the associations among the variables turns out to be easier with a smaller more homogeneous set of items, showing that in practice it may be a good idea to perform the imputation per subset. Whether this is generally the case, is something that needs future research.

Despite of their favorable features for missing data imputation in large-scale studies, various issues concerning the implementation of the LC imputation models need further research. The first, and most important, is their moderate performance in capturing all the associations with high rates of missingness. It may be that we need an even larger number of classes than we used in our application. In the Bayesian specification, we have to specify the prior distributions for the parameters, and it is well known that the choice of priors may affect the results. Therefore, also the specification of the priors in the context of LC-MI needs further study.

Moreover, LC models can easily be extended with a regression model in which the latent classes are predicted using background variables, such as sex, age, and education level. Such an approach has not been used for MI yet, but it may be interesting to investigate whether inclusion of explanatory variables may improve the obtained imputations.

While we focused on the LC-based imputation methods for cross-sectional categorical data, the methods may also be applied with mixed categorical and continuous data, as well as with more complex longitudinal or multilevel designs. This reflects the wide range of applications in which LC models can be used. For instance, LC models for multilevel data (both for continuous and categorical variables) are described by Vermunt (2008), while latent Markov models for longitudinal data are among others described by Baum et al. (1970). Such more advanced LC models may also be used for MI. A possible Bayesian (DPMM) implementation of LC MI for longitudinal panel studies is provided by Si (2012).

APPENDIX

A Bayesian Tools

A.1 The Dirichlet Distribution

In Appendices A and B, $f(\cdot)$ will denote a generic probability distribution or density. The Dirichlet Distribution will be denoted with $\text{Dir}(\lambda)$, where $\lambda = (\lambda_1, ..., \lambda_d)$ is a multi-dimensional parameter.

Suppose we have a random variable U with d components ($d \geq 2$) such that $U = (U_1, ..., U_d)$; then $U \sim \text{Dir}(\lambda)$, or equivantly,

$$f(U|\lambda) = \frac{1}{B(\lambda)} \prod_{i=1}^{d} u_i^{d_i - 1}$$

in the d-dimensional simplex $\{(u_1, ..., u_d) : u_i \in \mathbb{R}^+ \ \forall \ i, u_1 + ... + u_d = 1\}$. Here, each u_i is a realization of U_i and $B(\lambda)$ is the multivariate Beta function. When $d = 2$, the Dirichlet distribution becomes a Beta distribution.

This density can be used to model sets of probabilities of mutually exclusive and exhaustive events. This property, as well as its functional form, makes the Dirichlet distribution a conjugate candidate for the Multinomial distribution, thus forming the Dirichlet-Multinomial conjugate. According to this model, if the prior distribution of the set of parameters of the Multinomial distribution with d categories, say π, follows a Dirichlet distribution with parameter $\lambda = (\lambda_1, ..., \lambda_d)$ and the data $Y = (Y_1, ..., Y_d)$ are assumed to be distributed according to a Multinomial distribution with d components, then the resulting posterior is $\lambda|Y \sim Dir(\lambda_1 + Y_1, ..., \lambda_d + Y_d)$. In case of $d = 2$ the Dirichlet-Multinomial conjugate corresponds to the Beta-Binomial.

A.2 Bayesian Computation

A.2.1 The Gibbs sampler

Consider a L-dimensional random variable $\theta = (\theta_1, ..., \theta_L)$ and suppose that we want to compute the marginal densities $f(\theta_i)$, $i = 1, ..., L$. Furthermore, suppose that these marginal densities are obtainable by integration, $f(\theta_i) = \int f(\theta_1, ..., \theta_L) d(\theta_{-i})$, in which $\theta_{-i} = (\theta_1, ..., \theta_{i-1}, \theta_{i+1}, ..., \theta_L)$, is difficult to compute due to its analytical complexity, but that a series of conditional distributions $f(\theta_i|\theta_{-i})$ is available for each $i = 1, ..., L$. The Gibbs sampler, after initializing the variables with some value $\theta^{(0)} = (\theta_1^{(0)}, ..., \theta_L^{(0)})$, proceeds as follows:

1. Draw $\theta_1^{(t+1)} \sim f(\theta_1|\theta_2^{(t)}, ..., \theta_L^{(t)})$
2. Draw $\theta_2^{(t+1)} \sim f(\theta_2|\theta_1^{(t+1)}, \theta_3^{(t)} ..., \theta_L^{(t)})$

   $\vdots$

L. Draw $\theta_L^{(t+1)} \sim f(\theta_L|\theta_1^{(t+1)}, \theta_2^{(t+1)} ..., \theta_{L-1}^{(t+1)})$

for $t = 1, ..., T$, where T is the total number of iterations of the sampler. Under mild conditions, the Gibbs sampler converges to the stationary distributions $f(\cdot)$. For further technical details, we refer to Gelfand and Smith (1990). Liu (1994) argued that the efficiency of the Gibbs sampler can be further improved by considering blocks of correlated components together. For instance, it is possible to group θ into two blocks, $G_1 = (\theta_1, ..., \theta_{d'})$ and $G_2 = (\theta_{d'+1}, ..., \theta_L)$. The result is a two-blocks Gibbs sampler:

1. Draw $G_1^{(t+1)} \sim f(G_1|G_2^{(t)})$
2. Draw $G_2^{(t+1)} \sim f(G_2|G_1^{(t+1)})$

for $t = 1, ..., T$.

A.3 The Data Augmentation Algorithm

The Data Augmentation (DA) Algorithm (Tanner & Wong, 1987) is a special case of the Gibbs sampler. It exploits the fact that

$$f(\theta|Y) = \int_{\mathcal{Z}} h(\theta, Z|Y) dz,$$

that is, the *completion* or *data augmentation* of f. Here, Z are unobserved or latent data whose support is denoted by $\mathcal{Z}$, whereas Y denotes a set of observed variables

and $h(\cdot)$ is a probability density function. The aim of the DA algorithm is to simplify the sampling from the joint distribution $f(\theta|Y)$ through a simpler conditional distribution $f(\theta|Z,Y)$. For the DA algorithm, both $f(\theta|Z,Y)$ and $f(Z|\theta,Y)$ must be available. After initializing the unobserved Z and the values of θ with some arbitrary values $Z^{(0)}$ and $\theta^{(0)}$, the algorithm consist of two steps:

- Imputation Step: Draw $Z^{(t+1)} \sim f(Z|Y,\theta^{(t)})$
- Posterior Step: Draw $\theta^{(t+1)} \sim f(\theta|Y,Z^{(t+1)})$

for $t = 1,...,T$. In fact, this a version of the DA algorithm in which a single Z-value is drawn at each step. The original DA algorithm with multiple Z-value draws, as well as the conditions for convergence to the target distribution, can be found in Tanner and Wong (1987).

The DA algorithm can be seen as the Bayesian counterpart of the EM algorithm. Since both latent variables and missing data can be treated as unobserved values, this algorithm is of particular interest in applications such as LC-MI.

B Bayesian Multiple Imputation via Mixture Modeling

The notation and the model specification are the same as described in Section 2.2.1. The parameters of a specific class k (i.e., $\boldsymbol{\pi}_{\boldsymbol{y_j}}$ when $x = k$) will be denoted by $\boldsymbol{\pi}_{\boldsymbol{j}}^{\boldsymbol{k}}$. For parameters initialization and implementation of the algorithms, we follow Si and Reiter (2013). In order to simplify notation, a dot in the condition sign, i.e. $P(\ |\cdot)$, will indicate a conditioning on all the data and other parameters included in the model.

B.1 The Bayesian Latent Class Multiple Imputation Model

a. Distributional assumptions.
-Data likelihoods:

- $x_i \sim \text{Multinom}(\boldsymbol{\pi_x})$ where $\text{Multinom}(\boldsymbol{\pi_x})$ is the Multinomial distribution with parameter $\boldsymbol{\pi_x} = (\pi_1,...,\pi_k,...,\pi_K)\ \forall\ i$;
- $Y_{ij}|x_i = k \sim \text{Multinom}(\boldsymbol{\pi}_{\boldsymbol{j}}^{\boldsymbol{k}})$ with $\boldsymbol{\pi}_{\boldsymbol{j}}^{\boldsymbol{k}} = (\pi_{j1}^k,...,\pi_{jd}^k,...,\pi_{jd_j}^k)$ where d_j is the number of categories of the variable $Y_j\ \forall\ i,j$.

-Parameters priors:

- $\boldsymbol{\pi_x} \sim \text{Dir}(\boldsymbol{\alpha_x})$ with $\boldsymbol{\alpha_x} = (\alpha_1,...,\alpha_k,...,\alpha_K)$;
- $\boldsymbol{\pi}_{\boldsymbol{j}}^{\boldsymbol{k}} \sim \text{Dir}(\boldsymbol{\alpha}_{\boldsymbol{j}}^{\boldsymbol{k}})$ with $\boldsymbol{\alpha}_{\boldsymbol{j}}^{\boldsymbol{k}} = (\alpha_{j1}^k,...,\alpha_{jd}^k,...,\alpha_{jd_j}^k)$.

b. Implementation.

-Parameters initialization:

- set $\boldsymbol{\alpha}_{\boldsymbol{x}}^{(0)} = (\alpha_1^{(0)}, ..., \alpha_k^{(0)}, ..., \alpha_K^{(0)}) = (1, ..., 1)$;
- set $\boldsymbol{\alpha}_{\boldsymbol{j}}^{\boldsymbol{k}(0)} = (\alpha_{j1}^{k(0)}, ..., \alpha_{jd}^{k(0)}, ..., \alpha_{jd_j}^{k(0)}) = (1, ..., 1)\ \forall\ j, k$;
- initialize $\boldsymbol{\pi}_{\boldsymbol{x}}^{(0)}$ with a draw from the Dirichlet distribution with parameters $\boldsymbol{\alpha}_{\boldsymbol{x}}^{(0)}$
- set $P(y_{ij} = d|x_i = k; \boldsymbol{\pi}_{y_j})^{(0)} = \pi_{jd}^{k(0)} = \hat{f}(y_{j,obs} = d)\ \forall\ i, j, k$, where $\hat{f}(y_{j,obs} = d)$ is the marginal observed empirical probability that $y_{ij} = d$;
- sample a value for $Y_{ij,mis}$ from $\hat{f}(y_{j,obs})\ \ \forall\ i$ in $Y_{j,mis}$.

-The algorithm:

For $t = 1, ..., T$:

1. sample $x_i^{(t)} \in \{1, ..., K\}\ \forall\ i = 1, ..., n$ from a Multinomial distribution with *posterior membership probabilities* as parameters:

$$P(x_i^{(t)} = k|\cdot) = \frac{\pi_k^{(t-1)} \prod_{j=1}^{J} \left(\prod_{d=1}^{d_j} \left(\pi_{jd}^{k(t-1)} \right)^{\mathcal{I}(y_{ij}=d)} \right)}{\sum_{h=1}^{K} \pi_h^{(t-1)} \prod_{j=1}^{J} \left(\prod_{d=1}^{d_j} \left(\pi_{jd}^{h(t-1)} \right)^{\mathcal{I}(y_{ij}=d)} \right)}$$

   where $\mathcal{I}(y_{ij} = d) = 1$ if $y_{ij} = d$ and 0 otherwise;

2. sample

$$(\boldsymbol{\pi}_{\boldsymbol{x}}^{(t)}|\cdot) \sim \text{Dir}\left(\alpha_1^{(0)} + \sum_{i=1}^{n} \mathcal{I}(x_i^{(t)} = 1), ..., \alpha_K^{(0)} + \sum_{i=1}^{n} \mathcal{I}(x_i^{(t)} = K) \right)$$

   where $\mathcal{I}(x_i^{(t)} = k)$ is an indicator variable which is equal to 1 if $x_i^{(t)} = k$ and 0 otherwise;

3. draw

$$(\boldsymbol{\pi}_{\boldsymbol{j}}^{\boldsymbol{k}(\boldsymbol{t})}|\cdot) \sim \text{Dir}\left(\alpha_{j1}^{k(0)} + \sum_{i:x_i^{(t)}=k} \mathcal{I}(y_{ij} = 1), ..., \alpha_{jd_j}^{k(0)} + \sum_{i:x_i^{(t)}=k} \mathcal{I}(y_{ij} = d_j) \right)$$

   $\forall\ i, j, k$;

4. (*imputation step*): given the value $x_i^{(t)} = k$ of each unit, for each $\{i, j\}$ in Y_{mis} sample from

$$(Y_{ij}^{(t)}|\cdot) \sim \text{Multinom}(\boldsymbol{\pi}_{\boldsymbol{j}}^{\boldsymbol{k}(\boldsymbol{t})}).$$

Once the MCMC chain has completed its iterations, the m imputations are obtained by selecting m draws from the sampled values Y_{mis} from Step 4 in such a way that the draws are sufficiently independent. The thinning of the chain by taking the draws far enough from one another reduces the autocorrelations within the chain.

Furthermore, with a few minor adjustments, it is possible to let the algorithm estimate a LC model using only the observed information Y_{obs}. This involves ignoring Step 4 (the imputation step) and setting $\mathcal{I}(y_{ij} = d) = 0$ for every $Y_{ij} \in Y_{mis}$ in Step 1, in order to ignore the missing values at every iteration of the MCMC chain. In this way, each subject contributes to the model estimation and parameters updating only through his observed values.

B.2 The Dirichlet Process Mixture of Multinomial Distributions Imputation Model

This section gives a short description of the implementation of the algorithm. For a more detailed information on the Dirichlet Process Mixture, see Congdon (2006) and Escobar and West (1995).

a. Distributional assumptions.

-Data likelihood:

- $x_i \sim \text{Multinom}(\boldsymbol{\pi_x})$ where $\boldsymbol{\pi_x} = (\pi_1, ..., \pi_k, ..., \pi_\infty)\ \forall\ i$;
- in the practice, the process is truncated at some K^*, with K^* choosen to be arbitrarily high, so that $\boldsymbol{\pi_x} = (\pi_1, ..., \pi_k, ..., \pi_{K^*})$;
- $Y_{ij}|x_i = k \sim \text{Multinom}(\boldsymbol{\pi_j^k})$ with $\boldsymbol{\pi_j^k} = (\pi_{j1}^k, ..., \pi_{jd}^k, ..., \pi_{jd_j}^k)$ where d_j is the number of categories of the variable Y_j.

-Parameters priors:

- $\pi_k = V_k \prod_{h<k}(1 - V_h)$ for $h \in \{1, ..., K^*\}$;[9]
- $V_k \sim \text{Beta}(1, \alpha)$ where α is the *concentration parameter* of the process;
- $\alpha \sim \text{Gamma}(a, b)$;
- $\boldsymbol{\pi_j^k} \sim \text{Dir}(\boldsymbol{\alpha_j^k})$ with $\boldsymbol{\alpha_j^k} = (\alpha_{j1}^k, ..., \alpha_{jd}^k, ..., \alpha_{jd_j}^k)$.

b. Implementation.

-Parameters initialization:

- set $\alpha^{(0)} = 1$;
- initialize V_k with a draw from the Beta(1,1) distribution $\forall\ k$;

9 This is the stick-breaking representation of the Dirichlet Process. For technical details, see Sethuraman (1994) and Ishwaran and James (2001).

- initialize the parameters of the Gamma distribution $(a, b) = (0.25, 0.25)$;[10]
- set $\boldsymbol{\alpha}_{\boldsymbol{j}}^{\boldsymbol{k}(0)} = (\alpha_{j1}^{k(0)}, ..., \alpha_{jd}^{k(0)}, ..., \alpha_{jd_j}^{k(0)}) = (1, ..., 1)\ \forall\ j, k$;
- set $P(y_{ij} = d|x_i = k; \boldsymbol{\pi}_{y_j})^{(0)} = \pi_{jd}^{k(0)} = \hat{f}(y_{j,obs} = d)\ \forall\ i, j, k$, where $\hat{f}(y_{j,obs} = d)$ has been defined in Appendix B.1;
- sample a value for $Y_{ij,mis}$ from $\hat{f}(y_{j,obs})\ \ \forall\ i$ in$Y_{j,mis}$.

-The algorithm:
For $t = 1, ..., T$:

1. sample $x_i^{(t)} \in \{1, ..., K\}\ \forall\ i = 1, ..., n$ from a Multinomial distribution that has the *posterior membership probabilities* as parameters:

$$P(x_i^{(t)} = k|\cdot) = \frac{\pi_k^{(t-1)} \prod_{j=1}^{J} \left(\prod_{d=1}^{d_j} \left(\pi_{jd}^{k(t-1)} \right)^{\mathcal{I}(y_{ij}=d)} \right)}{\sum_{h=1}^{K^*} \pi_h^{(t-1)} \prod_{j=1}^{J} \left(\prod_{d=1}^{d_j} \left(\pi_{jd}^{h(t-1)} \right)^{\mathcal{I}(y_{ij}=d)} \right)}$$

   where $\mathcal{I}(y_{ij} = d) = 1$ if $y_{ij} = d$ and 0 otherwise;

2. sample $V_k^{(t)}$ for each $k \in \{1, ..., K^* - 1\}$ from

$$(V_k^{(t)}|\cdot) \sim Beta\left(1 + \sum_{i=1}^{n} \mathcal{I}(x_i^{(t)} = k), \alpha^{(t-1)} + \sum_{h=k+1}^{K^*} \left(\sum_{i=1}^{n} \mathcal{I}(x_i^{(t)} = h) \right)\right)$$

   where $\mathcal{I}(x_i^{(t)} = k)$ is an indicator variable which is equal to 1 if $x_i^{(t)} = k$ and 0 otherwise; set $V_{K^*} = 1$ and calculate each $\pi_k^{(t)} = V_k^{(t)} \prod_{h<k}(1 - V_h^{(t)})$;

3. draw

$$(\boldsymbol{\pi}_j^{\boldsymbol{k(t)}}|\cdot) \sim \text{Dir}\left(\alpha_{j1}^{k(0)} + \sum_{i:x_i^{(t)}=k} \mathcal{I}(y_{ij} = 1), ..., \alpha_{jd_j}^{k(0)} + \sum_{i:x_i^{(t)}=k} \mathcal{I}(y_{ij} = d_j)\right)$$

   $\forall\ i, j, k$;

4. update the value of $\alpha^{(t)}$ according to

$$(\alpha^{(t)}|\cdot) \sim \text{Gamma}\left(a + K^* - 1, b - log(\pi_{K^*}^{(t)})\right)$$

10 In consistent with Dunson and Xing (2009) and Si and Reiter (2013)'s guidelines.

5. (*imputation step*): given the value $x_i^{(t)} = k$ of each unit, for each $\{i, j\}$ in Y_{mis} sample from

$$(Y_{ij}^{(t)}|\cdot) \sim \text{Multinom}(\boldsymbol{\pi}_j^{\boldsymbol{k(t)}}).$$

Once the chain has completed its iterations, m imputation are obtained in the same way as seen for the Bayesian LC imputation model (see Appendix B.1). A DPMM (without imputation purposes) can also be estimated using only the observed part of the dataset, Y_{obs}, in a fashion similar as indicated for the Bayesian LC analysis (see Appendix B.1).

C Generating the Extra Missingness for the Real-data Example

In the Case II condition of the real-data example, missing values were generated by a MAR mechanism following Brand (1999) and Van Buuren et al. (2006). After selecting sample units with fully observed data ($n^* = 494$, from here on referred as Y^*), we specified γ (i.e., the sought new proportion of incomplete cases) and P, the number of new missing data patterns that should be created. The missing data patterns were randomly generated through a series of binomial distributions, yielded P vectors $R_p = (r_{p1}, ..., r_{pJ})$ of length J (in our example application $J = 91$), where each $r_{pj} = 0$ if variable Y_j^* is missing in pattern p and $r_{pj} = 1$ otherwise. Moreover, the relative frequency of each pattern, $f = (f_1, ..., f_P)$, is such that $\sum_p f_p = 1$.

Subsequently, each person was randomly allocated to one of the P patterns according with probabilities f: in this way, $\gamma\, n^*\, f_p$ units were made incomplete for each pattern according to the following probabilistic model. First, a linear combination of the observed variables for each case in block p was calculated. For instance, we calculated $c_{ip} = \sum_j w_{pj} r_{pj} Y_{ij}^*$, where w_{pj} are the regression coefficient resulting from the linear regression of Y_p^* on $Y_{-p}^* = (Y_1^*, ..., Y_{p-1}^*, Y_{p+1}^*, ..., Y_J^*)$. Second, for each p the c_{ip} were categorized into 3 categories by specifying 2 cutoff points. For this purpose, we used the 0.33 and 0.66 quantiles of c_{ip} within each block p. Third, odds ratio of having response pattern R_p were specified for the second and the third category (the first category is the reference category), were determined. For simplicity, in every block we created a MARTAIL MAR mechanism with the same odds ratio for each p. A MARTAIL mechanism is more likely to generate missingness for the lowest and highest c_{ip} scores. More specifically, we set the missingness odds ratios to 0.25 and 1.00 for the medium and high c_{ip} categories, respectively, yielding the missingness probabilities as in Brand (1999), equation 5.5. Finally, a random draw u_{ip} from the uniform distribution determined whether the data for

unit i should be set as missing according to pattern R_p; that is, if the value of u_{ip} is not larger than the corresponding probability of missingness for case i, missing values were created in agreement with response pattern R_p. Once this operation was accomplished, the resulting sub-dataset of 494 units was merged with the 212 incomplete cases of the original dataset, yielding the dataset used for our Case II analyses.

3

BAYESIAN LATENT CLASS MODELS FOR THE MULTIPLE IMPUTATION OF CATEGORICAL DATA

Latent class analysis has been recently proposed for the multiple imputation of missing categorical data, using either a standard frequentist approach or a non-parametric Bayesian model called Dirichlet process mixture of multinomial distributions. 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 a different number of classes should be estimated to determine the required number of classes, and subsequently the selected model is re-estimated for multiple bootstrap samples, to take into account parameter uncertainty during the imputation stage. Whereas the Bayesian Dirichlet process model performs 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 a number of clusters during the Gibbs sampling, leading to an underfitting model yielding invalid imputations. In this chapter, 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.

This chapter is submitted for publication.

3.1 Introduction

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 (1987)'s 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 (Little & Rubin, 2002; Schafer & Graham, 2002; Allison, 2009). 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 missing values 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 and 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 (Vermunt et al., 2008; Si & Reiter, 2013).

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 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 non-parametric 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 EM algorithm, first for determining the number of classes using a model selection criterion like the AIC, and subsequently for obtaining the m imputations, which involves re-estimating 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 chapter 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 feature of our approach is that the number of classes needed for the imputation model is determined using a single, 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 chapter is as follows. In Section 3.2, the BLC model for the MI of categorical data is introduced, along with its estimation and set-up. Section 3.3 describes two simulation studies which compare the BLC model with different prior specifications, as well as with the DPMM, FLC, and MICE approaches. Section 3.4 reports the results of a real-data experiment. Section 3.5 concludes with final remarks by the authors.

3.2 Bayesian Latent Class Imputation

Bayesian imputations are derived from the posterior predictive distribution of the missing data given the observed data, i.e. $\Pr(Y_{mis}|Y_{obs}) = \int \Pr(Y_{mis}|\pi)\Pr(\pi|Y_{obs})d\pi$, 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 $\Pr(\pi|Y_{obs})$, and then by sampling from the predictive distribution $\Pr(Y_{mis}|\pi^{*(l)})$, $l = 1, ..., m$. The posterior $\Pr(\pi|Y_{obs})$ is estimated via Gibbs sampling and derived from two quantities: a probabilistic model for the data (the likelihood) and a prior distribution for π.

3.2.1 The data model

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

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

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

$$\Pr(y_i; \pi) = \sum_{k=1}^{K} \pi_k \prod_{j=1}^{J} \prod_{s=1}^{s_j} (\pi_{kjs})^{\mathcal{I}_{ijs}}, \tag{3.1}$$

where $\mathcal{I}_{ijs}$ is an indicator variable equal to 1 when $y_{ij} = s$ and 0 otherwise. Below, we will use the symbols π_x and π_{kj} to refer to the two sets of model parameters, i.e. $\pi_x = (\pi_1, ..., \pi_K)$ and $\pi_{kj} = (\pi_{kj1}, ..., \pi_{kjs_j})$, while $\pi = (\pi_x, \pi_{11}, ..., \pi_{KJ})$.

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 k-th 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).

3.2.2 The prior distributions

Model (3.1) 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

$$\pi_x \sim Dir(\alpha_x)$$

and

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

$\forall\ k, j$. 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

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

and

$$\alpha_{kj} = (\alpha_{kj1}, ..., \alpha_{kjs}, ..., \alpha_{kjs_j}),$$

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, $\alpha_x = (c_1, ..., c_1)$ and $\alpha_{kj} = (c_2, ..., c_2)$. 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 0 only some of the classes will have a non-negligible posterior probability mass.

3.2.3 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 auto-correlations among the draws. Subsequently, m imputed data sets 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 and 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 j-th variable given the k-th latent class, $Y_j|X = k$, can be initialized through random draws from uniform Dirichlet distributions: $\pi_x^0 \sim Dir(1, ..., 1)$ and $\pi_{kj}^0 \sim Dir(1, ..., 1)\ \forall\ k, j$, 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 $\Pr(\pi|Y_{obs})$. Since a BLC imputation model may consist of a large number of parameters and since the quantity of interest in MI is the likelihood $\Pr(Y_{obs}|\pi)$, convergence is assessed by inspecting the traceplot of the log-likelihood function calculated at each iteration, as suggested by Schafer (1997).

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

Algorithm 3.1:

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

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

   in which $\mathcal{I}_{ijs}^*$ equals 1 when $y_{ij} = s$ and $y_{ij} \in Y_{obs}$, and 0 otherwise;

2. sample

$$(\pi_x^{(t)} | Y_{obs}, x^{(t)}, \alpha_x) \sim Dir \left(\alpha_1 + \sum_{i=1}^{n} \mathcal{I}(x_i^{(t)} = 1), \cdots, \alpha_K + \sum_{i=1}^{n} \mathcal{I}(x_i^{(t)} = K) \right)$$

   where $\mathcal{I}(x_i^{(t)} = k)$ is equal to 1 if $x_i^{(t)} = k$ and 0 elsewhere;

3. draw

$$(\pi_{kj}^{(t)} | Y_{obs}, x^{(t)}, \alpha_{kj}) \sim Dir \left(\alpha_{kj1} + \sum_{i: x_i^{(t)} = k} \mathcal{I}_{ij1}^*, \cdots, \alpha_{kjs_j} + \sum_{i: x_i^{(t)} = k} \mathcal{I}_{ijs_j}^* \right),$$

   $\forall \; k, j$.

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 $\Pr(\pi | Y_{obs})$. For the imputations, at each dth iteration we store the sampled parameters and class memberships, yielding $\pi^{*(1)}, ..., \pi^{*(m)}$ from $\Pr(\pi | Y_{obs})$ and $x_i^{(1)}, ..., x_i^{(m)}$. The imputed values are subsequently drawn from the posterior predictive distribution of the missing data, denoted by $\Pr(Y_{mis}^{*(l)} | Y_{obs}, \pi^{*(l)})$, $l = 1, ..., m$. 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, $l = 1, ..., m$, given the sampled value $x_i^{(l)} = k$ of each unit, and for each $\{i, j\} \in Y_{mis}$, sample from

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

and store the imputed values.

In the experiments described in Sections 3.3 and 3.4, Algorithm 3.1 is run with a routine we implemented in R, which is available upon request from the first author.

3.2.4 **Setting up the model**

3.2.4.1 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 an 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 3.1 (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 3.1) and obtain the posterior distribution $K|Y_{obs}$; (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 is usually required (Vermunt et al., 2008). Therefore, we suggest using the posterior maximum of the distribution $K|Y_{obs}$, that is, the largest K^* such that $Pr(K = K^*|Y_{obs}) > 0$. Afterwards, it is possible to perform the imputations (Algorithm 3.1 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 Sections 3.3 and 3.4 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.

3.2.4.2 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 $max(\alpha_1, ..., \alpha_K) < p/2$, 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 0, while with $min(\alpha_1, ..., \alpha_K) > p/2$, the possibly redundant classes will be given a non-negligible 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 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, $\alpha_{kj} = (1, ..., 1) \; \forall \; k, j$. However, as will be shown in Section 3.3, 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 3.1). 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]

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.

3.3 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.

3.3.1 Study 1

3.3.1.1 Study Design

Population model. The population model was specified for five predictor variables $Y_1, ..., Y_5$ 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

$$\log \Pr(Y_1, Y_2, Y_3, Y_4, Y_5) \propto -0.5 \sum_{j=1}^{5} Y_j - \sum_{j=1}^{4} \sum_{j'=j+1}^{5} Y_j Y_{j'} - 0.2 Y_1 Y_3 Y_5 + 0.5 Y_2 Y_4 Y_5. \tag{3.2}$$

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

$$\begin{aligned} \log(\Pr(Y_6 = 1)/\Pr(Y_6 = 0)) &= -0.1 + Y_1 + \beta_{1,2} Y_2 + \beta_{1,3} Y_3 - 0.6 Y_4 + 0.5 Y_5 + \\ &\quad \beta_{1,25} Y_2 Y_5 + \beta_{1,34} Y_3 Y_4 \\ \log(\Pr(Y_6 = 2)/\Pr(Y_6 = 0)) &= -0.6 + 1.8 Y_1 + \beta_{2,2} Y_2 + \beta_{2,3} Y_3 + Y_4 - 0.5 Y_5 + \\ &\quad \beta_{2,25} Y_2 Y_5 + \beta_{2,34} Y_3 Y_4, \end{aligned} \tag{3.3}$$

where, as can be seen, the reference category is $Y_6 = 0$. The values of the β parameters are reported in Table 3.1. Based on models (3.2) and (3.3), we generated $N = 500$ datasets with $n = 5000$ 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

Table 3.1: 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	-0.5	0.2

high missingness condition, respectively. Table 3.2 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.

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. (2013)'s method exposed in Section 3.2.4.1. In particular, for each simulated datasets we ran steps 1-3 of Algorithm 3.1 with 20 components for $T = 3000$ iterations, of which $b = 1000$ served as burn-in. The remaining 2000 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 $\alpha_k = 20$ 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 $T = 5000$ total and $b = 1000$ 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 (3.3) 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 partic-

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 et al. (2006), to reach convergence the number of iterations does not need to be large, and we decided to set it equal to 20.

Table 3.2: MAR mechanisms used in Study 1: the table reports the probability of missingness 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

ular, we considered the estimates of the parameters reported in Table 3.1: these parameters correspond to the main and interaction effects of the variables with missing values (Y_2 and Y_3).

3.3.1.2 Results

Tables 3.3 and 3.4 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.

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.

Table 3.3: Relative bias, stability and coverage rate observed for the estimates of eight multinomial logistic model parameters in (3.3) 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

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.

Table 3.4: Relative bias, stability and coverage rate observed for the estimates of eight multinomial logistic model parameters in (3.3) 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

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.

Table 3.5: Probability of observing 1 for the independently generated variables of Study 2.

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

3.3.2 **Study 2**

3.3.2.1 Study Design

Population model. In Study 2 we used $J = 21$ binary variables $Y_1, ..., Y_{21}$ (coded with 0 and 1), 20 predictors and 1 outcome. The first 15 predictors were generated from the following log-linear model:

$$\begin{aligned}\log \Pr(Y_1, ..., Y_{15}) \propto & -0.15 \sum_{j=1}^{15} Y_j + 0.5 \sum_{j=1}^{4} \sum_{j'=j+1}^{5} Y_j Y_{j'} - 0.1 \sum_{j=6}^{10} \sum_{j'=j+1}^{11} Y_j Y'_j \\ & + 0.15 \sum_{j=12}^{14} \sum_{j'=j+1}^{15} Y_j Y'_j + 0.3 Y_1 Y_2 Y_7 + 0.6 Y_3 Y_4 Y_8 - 0.4 Y_6 Y_9 Y_{10},\end{aligned} \tag{3.4}$$

while the remaining 5 predictors were assumed to be independent of the rest, with marginal probabilities $\Pr(Y_j = 1),\ j = 16, ..., 20$, as reported in Table 3.5.

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

$$\begin{aligned}\text{logit}(Y_{21}) = & -1.9 + \beta_1 Y_1 + 1.8 Y_2 - 0.95 Y_3 - 0.9 Y_4 + .8 Y_5 + \beta_6 Y_6 - 0.5 Y_7 + 0.6 Y_8 + Y_9 \\ & + 0.55 Y_{10} - 0.6 Y_{11} + 0.75 Y_{12} - 1.2 Y_{13} + 0 Y_{14} + 0 Y_{15} + \beta_{16} Y_{16} - 0.85 Y_{17} \\ & + 0.55 Y_{18} + 0 Y_{19} + \beta_{20} Y_{20} + \beta_{1.5} Y_1 Y_5 + \beta_{1.17} Y_1 Y_{17} + \beta_{1.5.17} Y_1 Y_5 Y_{17}.\end{aligned} \tag{3.5}$$

Besides the two- and three-way interaction terms, in model (3.5) we also specified some null effects (coefficients equal to 0) in order to assess how the imputation models deal with irrelevant variables. The values of the β parameters are shown in Table 3.6. From models (3.4) and (3.5) (and the variables described in Table 3.5), we generated $N = 200$ datasets with $n = 2000$ observations.

Table 3.6: 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	1.3	-.85	0.45

Table 3.7: MAR mechanisms used in Study 2.

Variable with missingness	Condition	Pr(Variable 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

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 3.7) 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,

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 3.8: Relative bias, stability and coverage rate observed for the estimates of seven logistic model parameters in (3.5) after applying three different imputation models. For the null effect β_{20} absolute bias is reported.

		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

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.

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 3.8). For the null effect β_{20}, we considered the absolute bias.

3.3.2.2 Results

The results reported in Table 3.8 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(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 was the interval for β_6, which resulted in too low coverage after imputing with FLC, BLC(1), or DPMM.

3.4 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 two years to a random sample of households resident in the Unites 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.

3.4.1 Study Design

The data. From the original dataset (which consisted of $n = 2538$ 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 3.9. [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 3.9) was the outcome and the $Y_1, ..., Y_{11}$ of Table 3.9 were the predictors. More specifically, the model we estimated was

$$log\left(\frac{\Pr(Y_0 \leq s)}{\Pr(Y_0 > s)}\right) \propto \sum_{j=1}^{11} \beta_j Y_j + \beta_{57} Y_6 Y_7 + \beta_{48} Y_4 Y_8. \tag{3.6}$$

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

Table 3.9: Variables used in the real-data application. Top: variables of the analysis model (3.6). Bottom: variables used to generate missingness.

Variables for the analysis model		
Variable Label	Variable Description	Values (range)
Y_0	Respondent's happiness	1 Not Happy - 3 Very Happy
Y_1	Respondent's opinion about his/her life	1 Dull - 3 Exciting
Y_2	Respondent's job satisfaction	1 Very dissatisfied - 4 Very satisfied
Y_3	Respondent's health status	1 Poor - 5 Excellent
Y_4	Respondent's marital status	0 Not married - 1 Married
Y_5	Respondent's employment status	1 Self employed - 2 Work for someone else
Y_6	Respondent's political view	1 Liberal - 3 Conservative
Y_7	Respondent's gender	0 Female - 1 Male
Y_8	Respondent's working status	1 Full time - 4 Not working
Y_9	Respondent's employer	1 Government - 2 Private
Y_{10}	Respondent's family income	1 <5000 - 4 >25000
Y_{11}	Respondent's time spent with friends	1 Almost every day - 7 Never
Variables used to generate missingness		
Variable Label	Variable Description	Values (range)
Y_{12}	Respondent's education	0 <Highschool - 4 Graduate
Y_{13}	Respondent's working contract	1 Full time - 2 Part-time
Y_{14}	Respondent's occupation prestige (score)	1 10/19 - 8 80/89

Table 3.10: MAR mechanisms used to generate missing data in the real-data application.

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

The first columns of Table 3.11 (below) 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 3.10. 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 $b = 5000$ iterations for the burn-in, and 5000 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 $\alpha_k = 40$, 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 $\alpha_{kjs} = 0.05$. The DPMM model was implemented with $K = 20$. The Gibbs sampler for both BLC and DPMM was run with $T = 55000$ and $b = 5000$. For MICE, 20 iterations were used for each imputation.

Outcomes. After imputing the data, model (3.6) 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]

3.4.2 **Results**

The results reported in Table 3.11 show that MICE performed badly: its point estimates for both main and interaction effects were rather far from those obtained

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

Table 3.11: Results of the real-data application. The table shows the point estimates and the standard errors for the ordered logistic regression model (3.6) estimated on the complete data ($n = 477$) and on the incomplete datasets, imputed with the MICE, FLC, BLC and DPMM methods. The * indicates the 5% significant parameter estimates.

			Imputation method							
Parameter	Complete Data		MICE		FLC		BLC		DPMM	
	Est.	S.E.	Est.	S.E.	Est.	S.E.	Est.	S.E.	Est.	S.E.
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
β_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

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 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.

3.5 Discussion

In this chapter, 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 non-parametric 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 ($n = 5000$) 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 ($n = 2000$) 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 chapter. However, 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*.

4

BAYESIAN MULTILEVEL LATENT CLASS MODELS FOR THE MULTIPLE IMPUTATION OF NESTED CATEGORICAL DATA

With this chapter, we propose using a Bayesian multilevel latent class (or mixture) model for the multiple imputation of nested categorical data. Unlike recently developed methods that can only pick-up associations between pairs of variables, the multilevel mixture model we propose is flexible enough to automatically deal with complex interactions in the joint distribution of the variables to be estimated. 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 available R-routine. Results indicate that the Bayesian Multilevel Latent Class model is able to recover unbiased parameter estimates of the analysis models considered in our studies, as well as to correctly reflect the uncertainty due to missing data.

This chapter is submitted for publication.

4.1 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 these dependencies to be taken 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 missingess 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 variable 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 substitutes the original incomplete dataset with $M > 1$ completed datasets, in which the missing values have been replaced by means of an impu-

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

tation model. Good performance of MI is obtained when the imputation model preserves the original relationships present among the variables (reflected in the imputed data), while the imputation model parameters are not of primary interest: the imputation model is only used to draw 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). In particular, in this chapter we deal with MI of missing level-1 and level-2 predictors of the analysis model.

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 (FCS) (Van Buuren et al., 2006) and joint modeling (Schafer, 1997). While the former is based on a variable-by-variable imputation, and requires specification of separate conditional models for each variable with missing observations, the latter only needs specification of a joint multivariate model of the variables 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 under which the imputations have to be performed.

Survey data often record categorical variable 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). Recently developed FCS approaches for multilevel data are the one-step FCS (Jolani, Debray, Koffijberg, Van Buuren & Moons, 2015) and the two-steps FCS (Resche-Rigon & White, 2016): the former uses a homoscedastic covariance matrix for the level-1 errors, while the second assumes heteroscedastic matrices. These methods cannot handle more than two categories for each categorical variable, and have not been extended yet to the imputation of level-2 predictors. 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 the lower- and higher-level variables is then estimated, and the imputations are based on the normal variable 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 the imputations. While representing a further step in the literature, JOMO still has some major limitations. By working with multivariate normal distributions, imputations yielded by JOMO can correctly reflect only pairwise linear relationships in the data, i.e., important relationships that may occur between pairs of variables. Possible higher-orders of associations, such as interactions and non-linearities, are disregarded by JOMO, making it less flexible and possibly leading to less optimal imputations if more complex dependencies are present which are of interest in the subsequent analysis of the MI dataset. Furthermore, the default prior distributions for the covariance matrices used by JOMO can become very informa-

tive in case of small (level-1) sample sizes, leading to biased parameter estimates and/or standard errors, as observed through a simulation study by Audigier et al. (2017).

Vermunt et al. (2008) proposed performing single-level MI of categorical data with frequentist *Latent Class* (LC) or mixture models, while Si and Reiter (2013) implemented the same model under a non-parametric Bayesian framework. The attractive part of using LC models for MI is their flexibility, since mixture models 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 (Vermunt et al., 2008). Furthermore, the model works with the original scale type of the data, preventing the risk of rounding bias (Horton et al., 2003). 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 it 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.

Multilevel MI of categorical data with LC models can be performed by estimating single-level LC models separately for each higher-level unit, performing in this way the imputations independently for each higher-level unit. However, this approach has some disadvantages. First, by focusing on a single higher-level unit it becomes impossible to either use or impute values of higher-level variables since these are constants within a higher-level unit. Therefore, this method cannot be used when missingness is present also in the higher-level variables. Second, this method can be applied only when the number of level-2 units is small, and the number of level-1 units for each group is large. When this method is run with a large number of higher-level units, model estimation (and selection) becomes time-consuming, because a larger number of LC models (and, therefore, parameters) must be implemented. Furthermore, small level-1 sample sizes for (some of the) level-2 units will make the LC model extremely unstable (Vermunt, 2003), leading to overly uncertain imputations.

With this chapter, we propose the use of a LC imputation model which is more naturally tailored for multilevel data: the *Bayesian Multilevel Latent Class* (BMLC) model. The BMLC imputation model we propose corresponds to 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 of groups within the same level-2 LC are assumed

to be independent from each other. The BMLC model extends the work of Vermunt (2003) to include also level-2 indicators, allowing for correct imputations at both levels of the dataset.

The outline of the chapter is as follows. In Section 4.2, the BMLC model is introduced, along with model and prior selection and model estimation issues. In Section 4.3, a simulation study is performed with two different sample size conditions. Section 4.4 shows an application to a real-data situation. Finally, Section 4.5 concludes with final remarks by the authors.

4.2 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 $\boldsymbol{\pi}$, that is $\Pr(D^{mis}|D^{obs}, \boldsymbol{\pi})$, 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 $\boldsymbol{\pi}$. Since $\Pr(\boldsymbol{\pi}|D^{obs}) \propto \Pr(\boldsymbol{\pi}) \Pr(D^{obs}|\boldsymbol{\pi})$, we need to specify a data model - $\Pr(D^{obs}|\boldsymbol{\pi})$ - and a prior distribution - $\Pr(\boldsymbol{\pi})$ - in order to obtain the posterior of $\boldsymbol{\pi}$. Model estimation, as well as the imputation step, is performed through Gibbs sampling.

4.2.1 The Data Model

We now introduce the BMLC models as if there were no missing data in the dataset ($D^{obs} = D$). 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 variables $Z_1, ..., Z_t, ..., Z_T$, each with R_t observed categories ($t = 1, ..., T$) and S level-1 categorical variables $Y_1, ..., Y_S$, 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, one for the level-2 (or higher-level) units and one 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

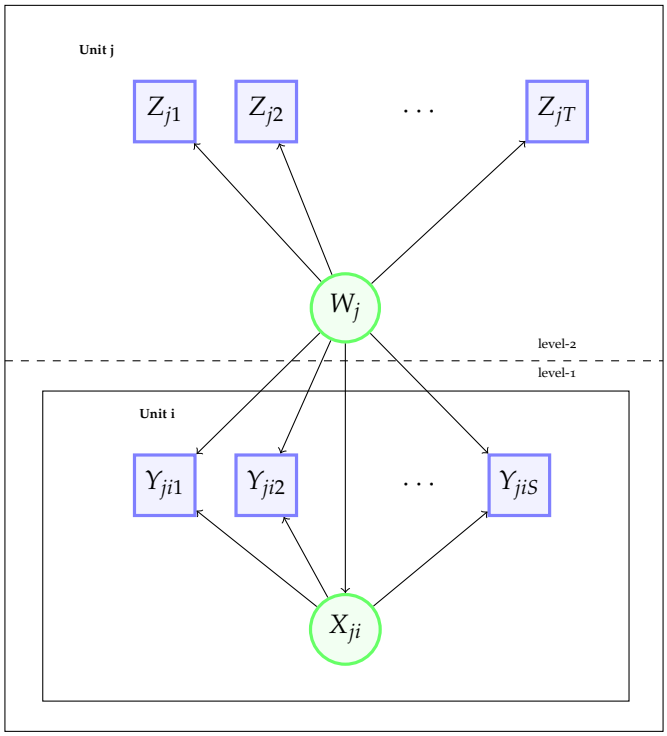


Figure 4.1: Graphical representation of the multilevel LC model with observed variables at both levels of the hierarchy.

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

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

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 4.1 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,
$W \sim \text{Multinom}(\boldsymbol{\pi}_W)$
$X|W=l \sim \text{Multinom}(\boldsymbol{\pi}_{lX})$ for $l=1,...,L$
$Z_t|W=l \sim \text{Multinom}(\boldsymbol{\pi}_{lt})$ for $t=1,...,T,\ l=1,...,L$
$Y_s|W=l,X=k \sim \text{Multinom}(\boldsymbol{\pi}_{lks})$ for $s=1,...,S,\ l=1,...,L,\ k=1,...,K$.
The parameters denote a vector containing the probabilities of each category of the corresponding Multinomial distribution. That is, $\boldsymbol{\pi}_W = (\pi_1,...,\pi_l,...,\pi_L)$, $\boldsymbol{\pi}_{lX} = (\pi_{l1},...,\pi_{lk},...,\pi_{lK})$, $\boldsymbol{\pi}_{lt} = (\pi_{lt1},...,\pi_{ltr},...\pi_{ltR_t})$, $\boldsymbol{\pi}_{lks} = (\pi_{lks1},...,\pi_{lksu},...,\pi_{lksU_s})$. The whole parameter vector is $\boldsymbol{\pi} = (\boldsymbol{\pi}_W, \boldsymbol{\pi}_{lX}, \boldsymbol{\pi}_{lt}, \boldsymbol{\pi}_{lks})$.

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

$$\Pr(\mathbf{Z}_j = \mathbf{z}_j, \mathbf{Y}_j = \mathbf{y}_j; \pi) = \sum_{l=1}^{L} \pi_l \prod_{t=1}^{T} \prod_{r=1}^{R_t} (\pi_{ltr})^{\mathcal{I}_{jt}^r} \prod_{i=1}^{n_j} \pi_{jil}, \tag{4.1}$$

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

$$\pi_{jil} = \sum_{k=1}^{K} \pi_{lk} \prod_{s=1}^{S} \prod_{u=1}^{U_s} (\pi_{lksu})^{\mathcal{I}_{jis}^u} \tag{4.2}$$

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

The model is capable of capturing between- and within-level-2 unit 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:

2 Vermunt (2003) denoted with 'non-parametric' the version of the multilevel LC model that uses a categorical random effect, for which a Multinomial distribution is assumed. This is opposed to the 'parametric' version of the model, which uses a (normally distributed) continuous random effect.

- the *local independence* assumption, according to which variables at level-2 are independent from each other within each LC W_j and variables 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.

By virtue of 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 variables. Both equations (4.1) and (4.2) incorporate these assumptions through their product terms.

It is also noteworthy that, by excluding the last product (over i) in equation (4.1) we obtain the standard LC model for the level-2 units, while, by excluding the product over t in equation (4.1) and setting $L = 1$, we obtain the standard LC model for the level-1 units.

In Bayesian MI, the quantity $\Pr(\mathbf{Z}_j, \mathbf{Y}_j; \boldsymbol{\pi})$ tends to dominate the (usually non-informative) prior distribution of the parameter, because the primary interest of an imputation model is the estimation of the joint 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 distribution of the variables. 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 variables 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).

4.2.2 The Prior Distribution

In order to obtain a Bayesian estimation of the model defined by equations (4.1) and (4.2), a prior distribution for $\boldsymbol{\pi}$ 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 parameters represent *pseudo-count* artificially added by the analyst in the model. Thus, for the BMLC model we assume as priors:

- $\boldsymbol{\pi}_W \sim \text{Dir}(\boldsymbol{\alpha}_W)$,
- $\boldsymbol{\pi}_{lX} \sim \text{Dir}(\boldsymbol{\alpha}_{lX})$,
- $\boldsymbol{\pi}_{lt} \sim \text{Dir}(\boldsymbol{\alpha}_{lt})$,
- $\boldsymbol{\pi}_{lks} \sim \text{Dir}(\boldsymbol{\alpha}_{lks})$.

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

Because in our MI application we will work with symmetric Dirichlet priors[3], in the remainder of the chapter 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.

In MI a large number of LCs is usually required when performing the imputations. The probability of empty clusters increases with the number of classes L or K when standard priors (such as the uniform Dirichlet prior) are used (Hoijtink & Notenboom, 2004). This causes the Gibbs sampler (described in Section 4.2.4) to sample from the prior distributions of the empty components, hence becoming unstable (Fruhwirth-Schnatter, 2006). In turn, this can lead to imputations that produce poor inferences, especially in terms of bias and coverage rate for some of the parameter estimates in the analysis model, as shown in Chapter 3. Better inferences can be obtained by setting the hyperparameters of the mixture components in such a way that units are distributed across all the LCs during the Gibbs sampler iterations. This is achievable by increasing the values of α_l and α_{lk} (maintaining symmetric Dirichlet distributions) until all the LCs are filled throughout the sampler iterations. Whether the selected values are large enough can easily be assessed with MCMC graphical output.[4] With such priors, the Gibbs sampler is able to draw from the equilibrium distribution $\pi|\mathbf{Z}, \mathbf{Y}$ and, accordingly, it can produce imputations that lead to correct inferences, since the model exploits all the selected classes. Because the imputation model parameter values do not need be interpreted in MI, more informative priors do not represent a problem here.

About the prior distribution of the conditional response probabilities, in Chapter 3 we advocated using hyperparameters which influence the imputations as

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

4 The value of the pseudo-counts for the LC proportions hyperparameter should be at least equal to half times the number of free parameters to be estimated within each LC, in order to cause the sampler to give non-zero weights to the extra components. See Rousseau and Mergensen (2011) for technical details.

little as possible. Their results indicated that uniform Dirichlet priors lead to biased parameter estimates of the analysis model, especially interaction terms (when present). However, decreasing the hyperparameter of the variables' conditional distribution probabilities to 0.01 (or 0.05) led the imputation model to obtain unbiased terms. Making the prior distribution of the conditional response probabilities as non-informative as possible is effective because it helps to identify the LCs and create 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 (4.1) 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 4.3. For the lower-level model (model (4.2) in the previous section), we will assume that the findings of Chapter 3 hold.[5] Therefore, we will set informative values for $\alpha_{lk}\ \forall\ l,k$ and non-informative values for $\alpha_{lksu}\ \forall\ l,k,s,u$.

4.2.3 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. In mixture modeling, over-fitting corresponds to selecting a number of classes larger than what is required by the data.

For the BMLC model in MI applications, model selection can be performed similar to Gelman et al. (2013)' s method (chapter 22). The procedure requires running the Gibbs sampler described in Algorithm 4.1 (without Step 7) of Section 4.2.4 with arbitrarily large L^* and K^*, and setting hyperparameters for the LC probabilities that can favor empty superfluous components. Following Gelman et al. (2013)' s guidelines[6], these values could be equal to $\alpha_l = 1/L^*$ and $\alpha_k = 1/K^*$. At the end of every iteration of the preliminary Gibbs sampler, we keep track of the number

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

6 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 large enough number of LCs for the imputations. Therefore, Gelman et al. (2013) determined the number of classes based on the posterior mode, while we perform model selection based on the posterior maximum. 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.

of LCs that are allocated, in order to obtain a distribution for L and K when the algorithm terminates. If the posterior maxima L_{max} and K_{max} of such distributions are smaller than the proposed L^* and K^*, in the next step the imputations can be performed with L_{max} and K_{max}. However, if either L_{max} or K_{max} (or both of them) equals L^* 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. This method corresponds to the multilevel extension of the model selection proposed in Chapter 3 for single-level LC MI. The method for BMLC models will be tested in the simulation study of Section 4.3 and in the real-data experiment of Section 4.4.

4.2.4 Estimation and Imputation

Since we are dealing with unobserved variables (W and X), model estimation is performed through a Gibbs sampler with Data Augmentation configuration (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 4.1 (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 $\Pr(\boldsymbol{\pi}|\mathbf{Y}^{obs}, \mathbf{Z}^{obs})$.

After fixing K, L and $\boldsymbol{\alpha}$, 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 $\Pr(\boldsymbol{\pi}|\mathbf{Y}^{obs}, \mathbf{Z}^{obs})$, 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 $\boldsymbol{\pi}^{(0)}$ through draws from uniform Dirichlet distributions (that is, Dirichlet distributions with all their parameter values set equal to 1), in order to obtain $\boldsymbol{\pi}_W^{(0)}$, $\boldsymbol{\pi}_{lX}^{(0)}$, $\boldsymbol{\pi}_{lt}^{(0)}$ and $\boldsymbol{\pi}_{lks}^{(0)}$ $\forall$ l, k, t, s. After all these preliminary steps are performed, the Gibbs sampler is run as shown in Algorithm 4.1.

Algorithm 4.1:

(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

$$\Pr(W_j^{(h)} = l | \mathbf{Y}_j^{obs}, \mathbf{Z}_j^{obs}, \boldsymbol{\pi}^{(h-1)}) =$$

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

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

$$\left(\boldsymbol{\pi}_W^{(h)} | W^{(h)}, \boldsymbol{\alpha}_W \right) \sim Dir \left(\alpha_1 + \sum_{j=1}^{J} \mathcal{I} \left(W_j^{(h)} = 1 \right), ..., \alpha_L + \sum_{j=1}^{J} \mathcal{I} \left(W_j^{(h)} = L \right) \right)$$

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

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

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

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

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

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

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

(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 m-th draw ($m = 1, ..., M$) must include $w_j^{(m)}$, $x_{ji}^{(m)}$, $\boldsymbol{\pi}_{lt}^{(m)}$ and $\boldsymbol{\pi}_{lks}^{(m)}$ $\forall j, i, t, s \in \{\mathbf{Y}^{mis}, \mathbf{Z}^{mis}\}$, the missing part of the dataset. Perform the m-th imputation for the variables at level-2 by drawing

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

and the m-th imputation for the variables at level-1 by drawing

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

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

Clearly, the M parameter values obtained in 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.

4.3 Study 1: Simulation Study

4.3.1 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 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 was no missingness in each dataset) were taken as benchmark.

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

$$log\Pr(\mathbf{Z}_j) = -.1\sum_{t=1}^{5} Z_{jt} + .1\sum_{t=1}^{4}\sum_{t'=(t+1)}^{5} Z_{jt}Z_{jt'} + .8Z_{j1}Z_{j2}Z_{j4}.$$

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

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

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

$$\begin{aligned} logit\Pr(Y_{ji6}|\mathbf{Y}_{ji}, \mathbf{Z}_j) =& \beta_{j0} + \beta_1 Y_{ji1} + \beta_2 Y_{ji2} + \beta_3 Y_{ji3} + \beta_4 Y_{ji4} + (\beta_5 + \gamma_{35}Z_{j3})Y_{ji5} \\ &+ \beta_{24}Y_{ji2}Y_{ji4} \end{aligned} \tag{4.3}$$

was the level-1 response model and

$$\beta_{j0} = \beta_{00} + \gamma_1 Z_{j1} + \gamma_2 Z_{j2} + \gamma_3 Z_{j3} + \gamma_4 Z_{j4} + \gamma_5 Z_{j5} + u_j, \text{ with } u_j \sim N(0, \tau^2) \tag{4.4}$$

was the level-2 model. Table 4.1 shows the numerical values of the level-1 parameters $\beta_{00}, ..., \beta_{24}$, the level-2 parameters $\gamma_1, ..., \gamma_5$, and the cross-level interaction γ_{35}. Table 4.1 also reports the value of the variance of the random effects, τ^2. Model (4.3)-(4.4) was the analysis model of our study, in which the main goal was recovering its parameter estimates after generating missingness.

Table 4.1: Parameter values for model (4.3)-(4.4).

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

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 variable 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. We applied Gelman et al. (2013) 's method described in Section 4.2.3 for model selection, by running a preliminary Gibbs sampler (with 1000 burn-in and 2000 estimation iterations) and obtaining a posterior distribution for L and K for each incomplete dataset. From these distributions, we selected the posterior maxima as the number of components to be used in the imputation stage. This led to an average number of classes equal to $L = 8.53$ at level-2 and $K = 10.83$ at level-1 when $J = 50$, $n_j = 20$ and $L = 9.70$ at level-2 and $K = 10.80$ at level-1 when $J = 200$, $n_j = 5$. Hyperparameters of the level-1 LCs and conditional responses (namely $\boldsymbol{\alpha}_{lx}$ and $\boldsymbol{\alpha}_{lks} \; \forall \; l, k, s$) were set following the guidelines of Section 4.2.2, that is, with informative prior distributions[7] for the parameters $\boldsymbol{\pi}_{lX}$ and with a non-informative prior distribution for the parameters $\boldsymbol{\pi}_{lks}$. In order to assess 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

7 We set $\alpha_{lk} = (\sum_s (U_s - 1)) \; \forall \; l, k$, i.e., the number of free parameters within each level-1 LC; this value was sufficiently large to ensure units' allocation across all the level-1 LCs.

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[8] 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 performed 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 non-informative priors and cluster-specific random covariance matrices for the lower-level errors.

In order to have a benchmark for results comparison, we also estimated model (4.3)-(4.4) to the complete data, before generating the missingness.

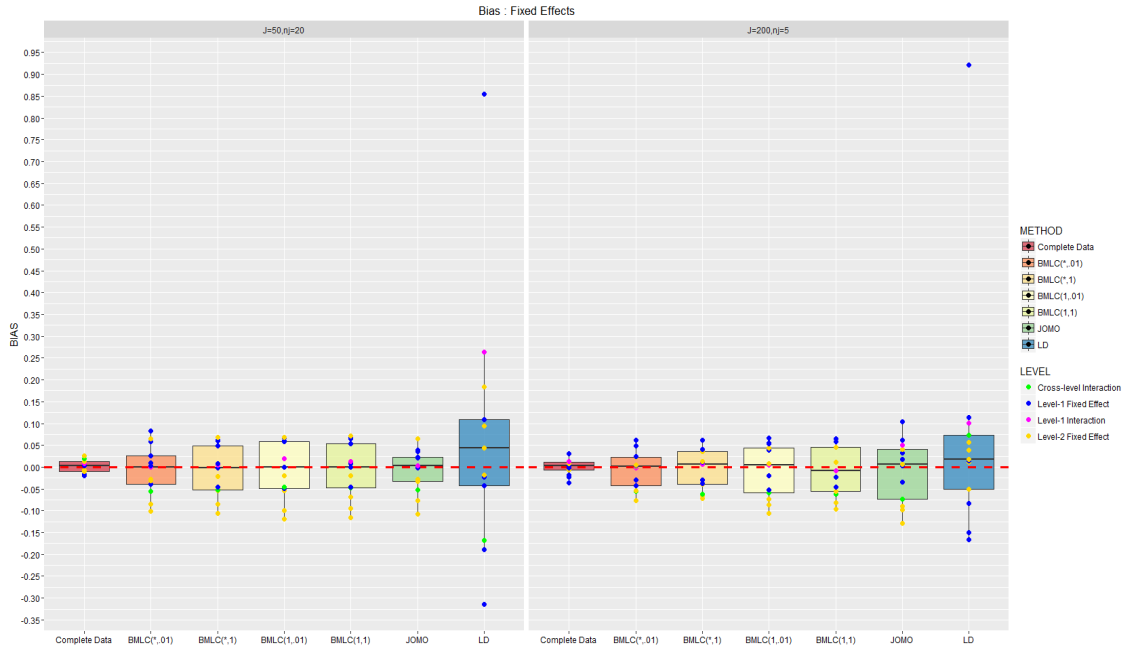
Study outcomes. For each parameter of model (4.3)-(4.4), we compared the bias of the estimates, along with their standard deviation (to assess stability) and coverage rate of the 95% confidence intervals. Analyses were performed with R version 3.3.0. JOMO was run from the `jomo` R-library. For each dataset, the analysis model (4.3)-(4.4) was estimated with the `lme4` package in R.

4.3.2 Study Results

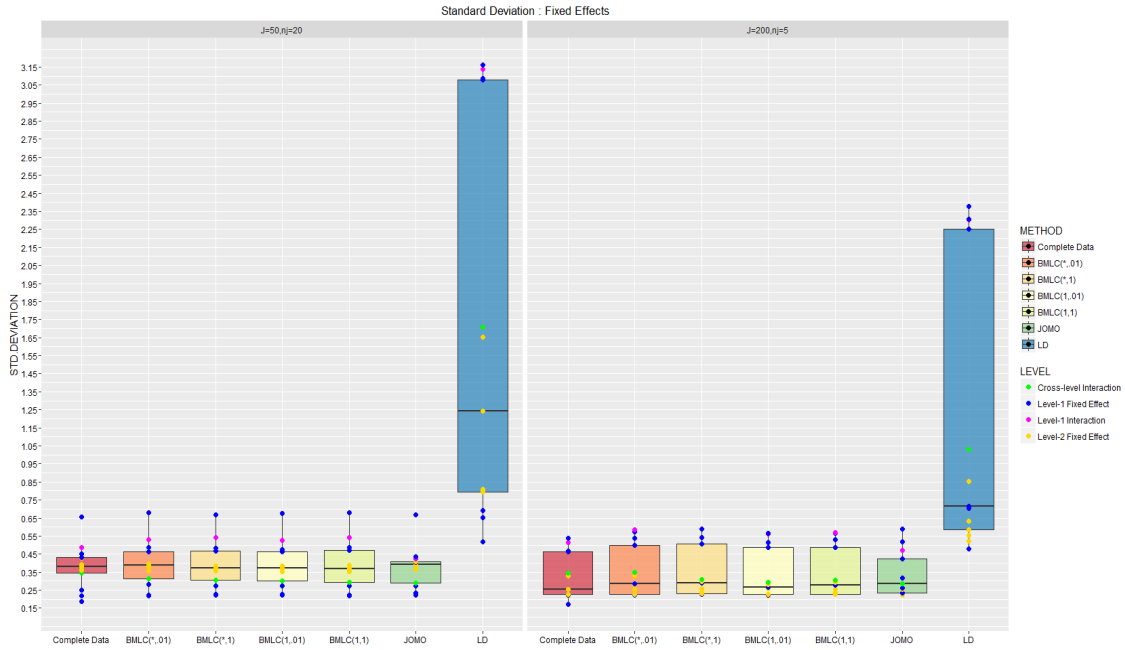
Figures 4.2a, 4.2b and 4.3 show the bias, standard deviations and coverage rates of the 95% confidence intervals for the thirteen fixed effect coefficients of model (4.3)-(4.4), averaged over the 500 datasets. The figures also show point estimates of each coefficient, distinguishing between level-1, level-2 and cross-level interaction fixed effects.

Figure 4.2a reports the bias of the fixed-effects estimates. Under both scenarios, BMLC and JOMO imputation appeared as the missing-data methods which pro-

8 Calculated through $\alpha_l = (\sum_t (R_t - 1) + (K - 1) + K(\sum_s U_s - 1)) \ \forall \ l$.



(a)



(b)

Figure 4.2: 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: $J = 200$, $n_j = 5$.

duced the least biased estimates, producing similar results (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 terms of bias. In this first condition BMLC imputation seem to be slightly outperformed by JOMO imputation, which also produced, on average, unbiased 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-wards. In the $J = 200$, $n_j = 5$ condition, the bias for the BMLC models was reduced with respect to the previous scenario. In this condition, the specification of the prior distribution seemed to have an effect in the final estimates produced by the BMLC model. In particular, models with priors that favored a full allocation of the level-2 units across all the L classes, as the BMLC(*,.01) and the BMLC(*,1), resulted with a slightly smaller bias than models with priors that did not favor full allocation, namely the BMLC(1,.01) and the 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 main effects (β_2, β_4) and the two interaction terms resulted in a larger bias than in the the previous case. This was mainly due to the prior distributions for the level-1 covariance matrices specified by the JOMO method, which are more influential with smaller group sizes (Audigier et al., 2017). In addition, in the second scenarios the BMLC imputation model under all prior specifications could correctly retrieve the level-1 interaction term and yield the least biased cross-level interaction term among all missing data techniques.

Figure 4.2b shows the stability of the estimates produced by all models, represented by their standard 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, by ignoring complex relationships, was an imputation model simpler than what was required by the data, and produced estimates that did not vary as they should.

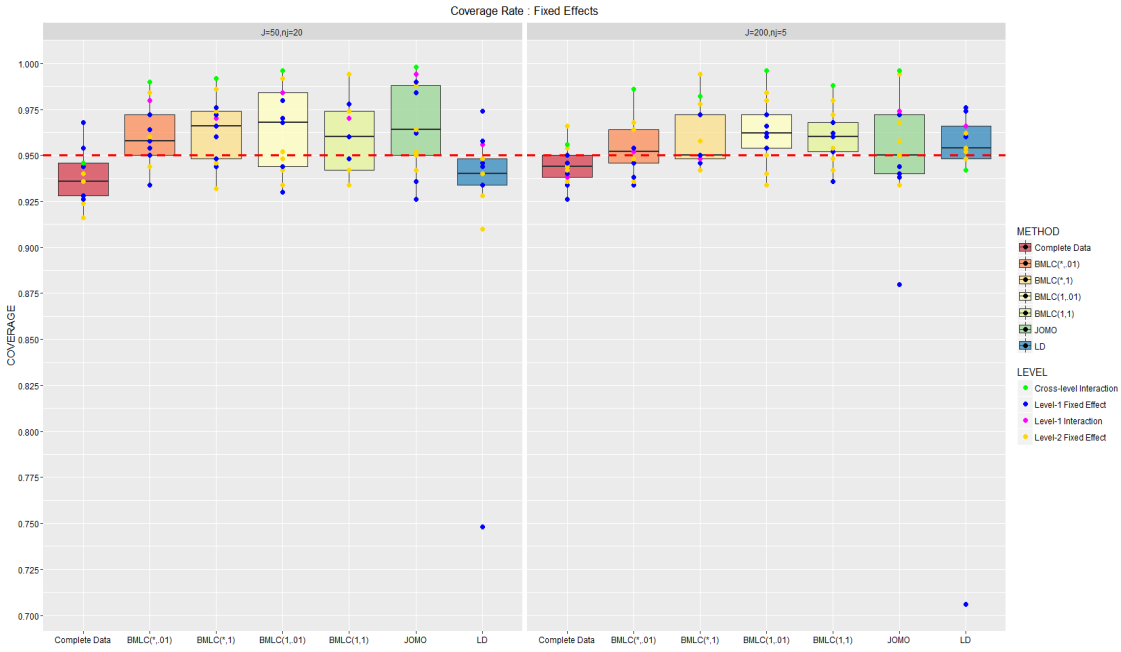


Figure 4.3: Coverage rates observed for the confidence intervals of the thirteen fixed multilevel logistic regression level-1, level-2, and cross-level coefficients obtained with complete data and the missing data methods BMLC(*,.01), BMLC(*,1), BMLC(1,.01), BMLC(1,1), JOMO and LD. Left: J = 50; nj = 20. Right: J = 200; nj = 5.

Table 4.2: 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.

$\tau^2 = 1$: 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)	-0.15	-0.05
BMLC(1,.01)	-0.16	-0.07
BMLC(1,1)	-0.15	-0.06
JOMO	-0.11	-0.01
LD	**-0.31**	0.05

Figure 4.3 displays the coverage rates of the 95% confidence intervals obtained with each method. For the Complete Data, the asymptotic confidence intervals had (on average) too low coverage due to the finite sample sizes used for our simulations. This is in line with the results of other multilevel simulation studies such as the ones performed by Maas and Hox (2005) and Paccagnella (2011). In the light of these results, LD produced overall coverage rates rather close 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 LD method generated coefficients for one of the parameters (β_3) with a too low coverage (about 0.7). The BMLC and JOMO imputation methods produced more conservative confidence intervals with $J = 50$ than with $J = 200$. In this latter case, their intervals appeared closer to the 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, most of the confidence intervals produced by the two imputation methods (BMLC and JOMO), had a coverage rate larger than their nominal level. This can be the consequence of the large amount of missingness (about 30% for each variable) entered in the data. Moreover, for the BMLC model, this can also be attributed to the over-fitting strategy pursued in this chapter. JOMO also led to intervals that were too short for a level-1 fixed effect (β_1) when $J = 200$. This was probably caused by the too small uncertainty captured by the model (see figure 4.2b), or by the stronger influence of the prior distribution in the second condition.

Table 4.2 reports the results obtained for the variance of the random effects, in term of bias. All the BMLC models yielded a random effect variance very close to the Complete Data case under both scenarios, while the JOMO method - which uses continuous random effects for the imputations - led to the least biased estimates for such parameter. Interestingly, in both conditions the variance estimated by JOMO was less biased than the Complete Data estimator. 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 $J = 50$.

4.4 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 variables 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 variables as outcome. 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 5 days of computation on a normal calculator (Intel Core i7), JOMO had not completed the burn-in iterations yet, and we decided to stop the process. This highlights another issue of the JOMO method (as implemented in the `jomo` package): when dealing with large datasets, the routine must handle 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).

4.4.1 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 variable was the country indicator. At level-1, variables consisted either of social, political, economical and behavioral questions, which the respondents 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) indicators 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 variables) into integer valued categories[9]. This enabled us to run the BMLC model on this dataset.

After removing level-1 variables 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 variables 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 ($n_j = 1948$), 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 looked 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 variables that could significantly explain the variability of the outcome 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

9 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 variable.

$$\begin{aligned} logit\,\Pr(Y_{ji0}|\mathbf{Y}_{ji},\mathbf{Z}_j) =& \beta_{j0} + (\beta_1 + \gamma_{11}Z_{j1})Y_{ji1} + \beta_2 Y_{ji2} + \beta_3 Y_{ji3} + \beta_4 Y_{ji4} + \beta_5 Y_{ji5} \\ &+ \beta_6 Y_{ji6} + \beta_{j7} Y_{ji7} + \beta_8 Y_{ji8} + \beta_9 Y_{ji9} \end{aligned} \tag{4.5}$$

at level-1 and

$$\begin{aligned} \beta_{j0} =& \beta_{00} + \gamma_1 Z_{j1} + u_{j0}, \text{ with } u_{j0} \sim N(0, \tau_0^2 = 0.29), \\ \beta_{j7} =& \beta_{70} + u_{j1}, \text{ with } u_{j1} \sim N(0, \tau_1^2 = 0.02) \end{aligned} \tag{4.6}$$

at level-2. A description of the 11 variables used in the model can be found at the top of Table 4.3, while the values of the coefficients (both fixed and random) are reported in the second column of Table 4.5 below. Furthermore, columns 5 and 8 of Table 4.5 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 missingness indicator. We did not only use the variables in model (4.5)-(4.6) 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 4.3. Table 4.4 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 variables. 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 missing values. The BMLC was run with all the 23 variables listed in Table 4.3, and was set as follows. We performed model selection using the method exposed in Section 4.2.3 based on Gelman et al. (2013)'s technique. A preliminary run of the Gibbs sampler with $L^* = 6$ and $K^* = 30$ indicated that running Algorithm 4.1 with $L = 2$ (the posterior maximum of L) and $K = 26$ (the posterior maximum of K) was sufficient to perform the 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 which guaranteed full allocation were $\alpha_l = 1500$ for each l at level-2 and $\alpha_{lk} = 50$ for each l, k at level-1. $M = 100$ 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.

Table 4.3: ESS data variables used in the Study 2.

Variable 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_1Z_1	Cross-level interaction between Y_1 and Z_1	-
	Other Variables used to generate missingess:	
Variable 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 4.4: Missingness generating mechanism for the variables of the ESS dataset.

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 4.5: 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 (4.5)-(4.6), attained after applying each method to the (fully or partially) observed data. Non-significant 5% p-values are marked in boldface.

	Estimates			Standard errors			p-values		
Parameter	Complete Data	LD	BMLC	Complete Data	LD	BMLC	Complete Data	LD	BMLC
β_{00}	-3.45	-2.72	-3.29	0.33	0.44	0.36	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	0.06	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	0.01
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	0.40	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	0.62	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**	0.03
τ_0^2	0.29	0.42	0.32						
τ_1^2	0.02	0.03	0.01						

4.4.2 Study Results

Table 4.5 shows 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 columns of Table 4.5, 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 4.5), 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.

4.5 Discussion

In this chapter 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 4.2, 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 compared with the LD method. 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 too influential default prior distributions (JOMO), the BMLC model offers a flexible imputation technique, able to pick up complex orders of associations among the variables 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 missing multilevel categorical data (e.g., coming from surveys), 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 chapter.

Despite the proven utility of the BMLC imputation model, some issues still need to be better crystallized by further studies. First, the current chapter aimed to give a general introduction of the BMLC model as a tool for MI, highlighting some of its strengths. Therefore, the simulation study in Section 4.3 was carried out under two sample size conditions typical of multilevel analysis (i.e., few large or several small level-2 units) and a moderately large proportion of missing data (about 30% per variable). The performance of the BMLC imputation model may be investigated further with other more extensive simulation studies, in which the model is tested

against more extreme missingness rates and sample size conditions (e.g., with few small or several large higher-level units). Second, the setting of the prior distribution for the higher-level mixture weights must be better examined, especially when the level-2 sample size is small and the number of classes selected with the method of Section 4.2.3 is (relatively) large. In these cases, achieving full allocation of the higher-level units across all the level-2 LCs is problematic, no matter how large the value of α_l. For instance, in the condition with $J = 50$ groups in the simulation study of Section 4.3, in which we selected an average number of level-2 LCs equal to $L = 8.53$ and a value for the hyperparameter α_l equal to the number of free parameters within each higher-level LC, the number of classes filled by the Gibbs sampler was on average roughly equal to $L = 5$. We tried to re-run the experiment by increasing the value of α_l, always obtaining similar results (in terms of classes allocated and MI inferences). It is possible that, because of the small sample size J, the Gibbs sampler reached the maximum possible number of classes that could be filled, and the groups could not be allocated to any new LC. We noticed, however, that the informative values used for α_l could help the Gibbs sampler to stabilize the number of occupied classes at that possible maximum. That is, for a maximum number of classes $\bar{L}$ that the sampler could occupy with informative hyperparameter α_l, the posterior distribution of the occupied number of classes during the imputation stage was $\Pr(L = \bar{L}|\mathbf{Z}, \mathbf{Y}) = 1$. Therefore, it is possible that in order for the Gibbs sampler to work correctly in presence of a small number of higher-level groups it is more important to have the level-2 units allocated to a stable number of classes, rather than to reach the full allocation of all the specified LCs. This can be the reason of the good results obtained in the simulation study of Section 4.3 with $J = 50$. However, in order to confirm our intuition, a more comprehensive study with different settings for the number of higher-level units and LCs, as well as for the value of the level-2 mixture weights hyperparameter α_l, should be carried out in future research.

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 variable 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. This can be achieved, for instance, by assuming mixture of univariate Normal (for the continuous data) and Multinomial (for the categorical data) distributions. In this case, Gelman et al. (2013)'s method might still be used for the model selection. Third, the model can be easily extended to

deal with three or more levels of the 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). Fourth, the proposed BMLC imputation model with LCs at two levels can easily be generalized to situations with more levels, where there is no need that the multiple levels are mutually nested. For example, one could deal with children nested within both schools and neighborhoods, where schools and neighborhoods form crossed rather than nested levels. These extensions are straightforward by making sure that the Gibbs sampler gets the LCs at one level conditioning on the sampled LCs for all other levels.

5

MULTIPLE IMPUTATION OF LONGITUDINAL CATEGORICAL DATA THROUGH BAYESIAN MIXTURE LATENT MARKOV MODELS

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

5.1 Introduction

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

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

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

1. capture dependencies among variables within measurement occasions;
2. capture overall dependencies between time points resulting from the fact that individuals differ from one another in a systematic way;
3. capture potential stronger relationships between adjacent time points;

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

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

4. automatically (i.e., without explicit specification) capture complex relationships in the data, such as higher-order interactions and non-linear associations;
5. respect the measurement scale of the variables (continuous/categorical).

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

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

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

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

An alternative solution for categorical data is represented by mixture or latent class (LC) models (Lazarsfeld, 1950), proposed and shown to provide good results as imputation models by Vermunt et al. (2008). Mixture modeling allows for flexible joint-density estimation of the categorical variables in the dataset, and requires only the specification of the number of LCs K. When K is set large enough, the model can automatically capture the relevant associations of the joint distribution of the variables (McLachlan & Peel, 2000; Vermunt et al., 2008), achieving requirement 4. However, standard LC models are better suited for cross-sectional datasets, because they do not account for the longitudinal architecture of the data, and, accordingly, do not satisfy requirement 3 above.

A natural extension of the LC model to longitudinal categorical data, which in addition accounts for unobserved heterogeneity between units, is represented by the *mixture Latent Markov* (MLM) model (Vermunt, 2010). With the MLM model subjects are clustered at two levels. At the higher level, a time-constant LC variable groups the units with similar time-varying patterns with each other, meeting in this way requirement 2. At the within-subject level, dynamic latent states (LSs; i.e., LCs that can vary over time) are specified for each time point, and -with the first-order Markov assumption- the LS distribution at time t depends only on the LS occupied at time $t-1$. From a MI point of view, the dynamic LSs help accounting for stronger dependencies across adjacent time points, satisfying requirement 3 above. Furthermore, the distribution of the observed variables at a specific time point depends non only on the time-constant LCs but also on the dynamic LSs, allowing to take dependencies within time points into account, thus meeting requirements 1 and 4. Lastly, the model respects the data scale (requirement 5) by assuming Multinomial distributions for all variables in the measurement model. As a further advantage, the MLM model can produce imputations also for time-constant variables with missing values, when present in the dataset at hand.

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

The outline of the chapter is as follows. In Section 5.2, the model is formally introduced, and the model selection issue is addressed. Sections 5.3 and 5.4 describe a simulation and a real-data study evaluating the performance of the Bayesian MLM (BMLM) imputation model. The authors provide final remarks in Section 5.5.

5.2 The Bayesian mixture Latent Markov Model for Multiple Imputation

Bayesian estimation of the MLM model requires defining the exact data generating model, such as the number of classes for the mixture part and the number of states for the latent Markov chain, as well as the prior distribution of the model parameters. This allows obtaining $\Pr(\boldsymbol{\theta}|\mathbf{y}^{obs})$, the posterior distribution of the unknown model parameters given the observed data $\mathbf{y}^{obs}$. In MI, the M sets of imputations are obtained from the posterior predictive distribution of the missing data, i.e. $\Pr(\mathbf{y}^{mis}|\mathbf{y}^{obs}) = \int \Pr(\mathbf{y}^{mis}|\boldsymbol{\theta}) \Pr(\boldsymbol{\theta}|\mathbf{y}^{obs}) d\boldsymbol{\theta}$. To achieve this, M parameter values $\boldsymbol{\theta}^{(m)}$ $(m = 1, ..., M)$ are first sampled from $Pr(\boldsymbol{\theta}|\mathbf{y}^{obs})$, and subsequently the imputations are drawn from $Pr(\mathbf{y}^{mis}|\boldsymbol{\theta}^{(m)})$.

5.2.1 Data generating model and prior distribution

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

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

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

The MLM model is composed of four parts:

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

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

$$\Pr(\mathbf{z}_i = \mathbf{u}, \mathbf{y}_i = \mathbf{r}^*) = \sum_l \omega_l \Lambda_{\mathbf{u}l} \pi_{\mathbf{r}^* l} \tag{5.1}$$

where, at the within-subject level,

$$\pi_{\mathbf{r}^* l} = \Pr(y_i = \mathbf{r}^*|w = l) = \sum_{s_1,...,s_T} \nu_{\kappa l} \Phi_{\mathbf{r}1kl} \prod_{t>1} \xi_{q,k(t)l} \Phi_{\mathbf{r}tkl}. \tag{5.2}$$

Figure 5.1 represents the path diagram of the data generating model. The picture stresses the double task executed by the subject-level mixture component w: capturing dependencies among the time constant variables and overall dependencies between all time points. Figure 5.1 also shows how the LS s_t at time t affects the distribution of both s_{t+1} and $\mathbf{y}_{it}$, capturing dependencies between variables within time point t (by means of the emission probabilities) as well as relationships between adjacent time points (by means of the transition probabilities). With such a model configuration, requirement 2 of Section 5.1 is satisfied with the time-constant latent variable w, while requirements 1 and 3 are met by means of the latent Markov structure assumed upon the time-varying variables.

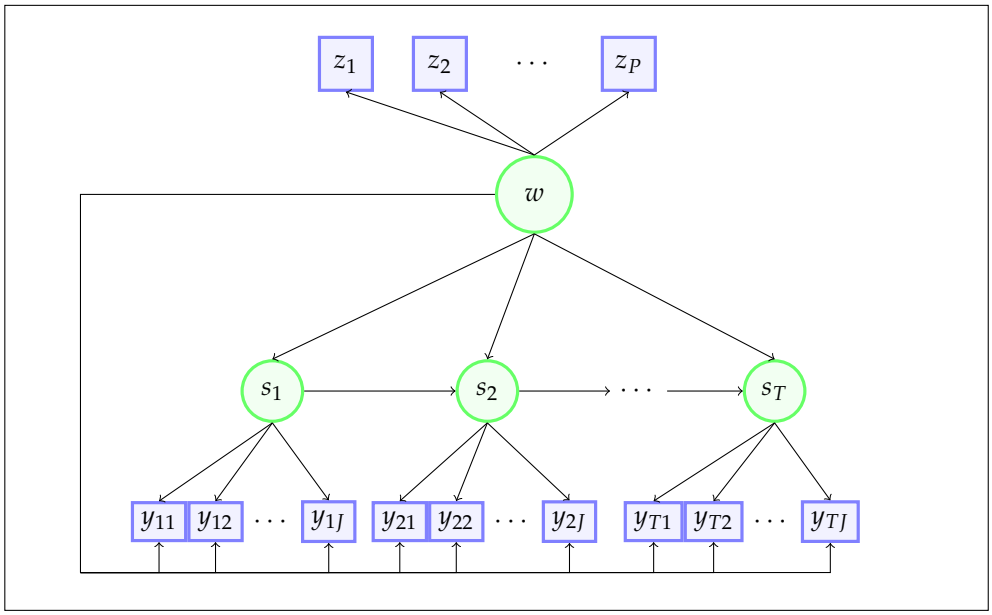


Figure 5.1: MLM model, graphical representation. w: time-constant latent class variable; z: time-constant variables; s: dynamic latent variable; y: time-varying variables.

Importantly, the model can also be implemented in absence of the time-constant variables, which involves dropping the term $\Lambda_{\mathbf{u}l}$ from equation (5.1) and the nodes representing the time-constant variables $z_{i1}, ..., z_{iP}$ from Figure 5.1.

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

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

For the Bayesian specification of the model, distributional assumptions must be made for all variables and parameters in model (5.1)-(5.2). Since all (latent and observed) variables in the model are categorical, a Multinomial distribution will be adopted for each of them. Formally:

- $w \sim Multinomial(\boldsymbol{\omega})$, with $\boldsymbol{\omega}$ the latent weights vector $(\omega_1, ..., \omega_L)$;

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

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

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

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

5.2.2 Model Selection

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

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

Bayesian modeling offers a simple solution to detect the number of LSs to be used in the imputation model. The method is described by Gelman et al. (2013), chapter

22 for standard mixture models (i.e., for $T = 1$). Their method consists of preliminarily processing the data by estimating a LC model (by means of the Gibbs sampler) with an arbitrarily large number of classes ($K = K^*$) and prior distributions for the latent variable parameter that favor the occurrence of empty components (e.g., with $\alpha_k = 1/K^* \; \forall \; k$) during the iterations of the Gibbs sampler. Counting the number of latent clusters (at each time point) occupied by the units during every iteration leads to a probability distribution for K once the Gibbs sampler is terminated. Gelman et al. (2013), who developed the method for substantive analysis, suggested to use the posterior mode of such distributions to perform inference and obtain interpretable classes. For MI purposes, in Chapter 3 we recommended using the posterior maximum of the resulting posterior distribution.[4] Once K has been chosen, the mixture model can be re-run (with prior distributions set as described in Appendix A) and the imputations can then be performed.

For the BMLM model, this method can be used to detect the number of states K when $L = 1$ (with hyperparameters $\alpha_k = \gamma_k = 1/K^* \; \forall \; k$ and $K = K^*$), as is shown using the first simulation study presented in Section 5.3 and the BMLM model (case $L = 1$) in the application of Section 5.4. More specifically, K can be set equal to the largest posterior maximum across all time points. When the number of latent clusters L is larger than 1, Gelman et al. (2013)'s method can be used to determine both L and K (as shown in the second simulation study of Section 5.3 and in the application of Section 5.4, case $L > 1$), by setting arbitrarily large values for the number of latent classes and states ($L = L^*$ and $K = K^*$) when running the preliminary Gibbs sampler, and hyperparameters for the latent classes proportions and transition probabilities equal to $\eta_l = 1/L^* \; \forall \; l$ and $\alpha_k = \gamma_k = 1/K^* \; \forall \; k$. The number of clusters to be used for the mixture components can then be chosen to be equal to the posterior maximum of the resulting distribution for L. The number of latent states can be chosen to be the largest among the L (smallest) posterior maxima observed across time points. That is, we would first consider the smallest posterior maxima of the number of latent states occupied at each time point (within each latent cluster $l = 1, ..., L$), and subsequently we would choose K as the maximum of the resulting L-dimensional vector. We opt for the smallest posterior maxima across time points, rather than for the largest ones, in order not to incur to the risk of leaving some of the latent states empty during the imputation stage, which could make the Gibbs sampler unstable, as explained in Appendix A.

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

5.2.3 Model Estimation and Imputation Step

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

Appendix B reports the Gibbs sampler (Algorithm 5.1) used to estimate model (5.1)-(5.2). For MI, model estimation is performed only on $\mathbf{z}^{obs}, \mathbf{y}^{obs}$, as in Vermunt et al. (2008). During one iteration, units are first allocated to the time-constant classes according to the *posterior membership probabilities* $\Pr(w|\boldsymbol{\theta}, \mathbf{z}_i, \mathbf{y}_i)$ and then, conditioned on the sampled w, units are assigned to the states of the LM chain at each time point. For each subject, the sequence $s_1, ..., s_T$ is drawn via *multi-move sampling* (Chib, 1996; Fruhwirth-Schnatter, 2006) through their posterior distribution $\Pr(s_1, ..., s_T|w = l, \boldsymbol{\theta}, \mathbf{y}^{obs})$. Multi-move sampling requires to store the *filtered state probabilities* $\Pr(s_t|\mathbf{y}_{it}, \boldsymbol{\theta})$ for each time point. How to perform multi-move sampling and compute the filtered-state probabilities is reported in Algorithms 5.2 and 5.3 of Appendix B. After units have been allocated to the LSs, the model parameters are updated using subsequent steps of Algorithm 5.1.

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

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

and

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

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

5.3 Simulation Studies

Performance of the BMLM imputation model was assessed by means of two simulation studies. In the first study, four time-varying variables were used, while in the

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

second four time-constant variables were added. In this way, with Study 1 we assessed the performance of the Bayesian LM model (that is, the BMLM model with $L = 1$) and in Study 2 we checked the performance of the BMLM model with a larger number of subject-level LCs. In both studies, analyses were carried out with R version 3.3.0.

5.3.1 **Study 1: time-varying variables**

In the first simulation study, we compared the BMLM imputation method with MICE and with *complete case* (CC) analysis method.[6] We followed this approach: first, $N = 200$ datasets were generated from a population model (which was not a MLM!). Second, MAR (and MCAR) missing values were created for some of the observations in the datasets. Third, the missing-data techniques (BMLM, MICE and CC analysis) were applied to deal with the missingness and it was checked whether the population model parameters of main interest were recovered. In order to highlight the ability of the BMLM model to *automatically* capture interaction terms (requirement 4 of Section 5.1), the MICE method was run with its default, main-effect only, settings. Last, results (in terms of bias, stability and coverage rate of the estimates) were compared across methods.

5.3.1.1 Set-up

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

$$log\,\Pr(Y_{11}, Y_{12}, Y_{13}) \propto -0.5 \sum_{j} Y_{1j} + \sum_{j=1}^{2} \sum_{j'=j+1}^{3} Y_{1j} Y_{1j'} - 0.5 Y_{11} Y_{12} Y_{13}. \tag{5.3}$$

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

$$\text{logit}\,\Pr(Y_{tj}) = 0.5 Y_{(t-1)j} - 0.15 \sum_{j' \neq j} Y_{(t-1)j'}, \tag{5.4}$$

6 CC analysis is a commonly used missing-data method which simply discards all units with at least one missing value from the dataset.

Table 5.1: Parameter values for model (5.5).

Parameter	β_0	β_1	β_2	β_3	β_{12}	ρ	τ
Value	-0.4	0.6	-1	0.8	-0.8	0.75	0.2

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

$$\operatorname{logit}\Pr(Y_{t4}) = \begin{cases} \beta_0 + \beta_1 Y_{t1} + \beta_2 Y_{t2} + \beta_3 Y_{t3} + \beta_{12} Y_{t1} Y_{t2} & \text{if } t = 1 \\ \beta_0 + \beta_1 Y_{t1} + \beta_2 Y_{t2} + \beta_3 Y_{t3} + \beta_{12} Y_{t1} Y_{t2} & \\ \qquad + \rho Y_{(t-1)4} + \tau Y_{(t-1)3} & \text{if } t > 1. \end{cases} \tag{5.5}$$

The values for the model parameters are reported in Table 5.1. These parameters were chosen in order to assess how the missing data techniques could capture different aspects of the data:

- $\beta_0, \beta_1, \beta_2, \beta_3, \beta_{12}$ were used to assess how the techniques recovered relationships among variables at the same time point;
- ρ was used to assess how the models could recover auto-correlations in Y_4 at lag-1;
- τ served to determine whether the models could recover crossed-lagged associations (between Y_3 and Y_4) at lag-1.

From this population model, we generated $N = 200$ datasets for $n = 200$ subjects and $T = 10$ time points. The datasets were created and stored in long format, so that each dataset was composed of 2000 rows and 6 column (Y_1,...,Y_4 as well as the subject and time indicators).

Generating missingness. For each dataset, we generated missing data in the predictor Y_2 and the outcome Y_4 according to the following mechanism. Let R_{tj} be an indicator function equal to 1 when Y_{tj} was missing ($j \in \{2, 4\}$) and 0 when Y_{tj} was observed. For each $t = 1, ..., 10$ Y_2 was made missing through

$$\Pr(R_{t2} = 1) = \begin{cases} 0.45 & \text{if } Y_{t3} = 0 \\ 0.20 & \text{if } Y_{t3} = 1. \end{cases} \tag{5.6}$$

Missingness in the outcome was instead produced with

$$\Pr(R_{t4}=1)=\begin{cases}0.30 & \text{if } t=1\\ 0.35 & \text{if } Y_{(t-1)1}=0 \text{ and } t>1\\ 0.25 & \text{if } Y_{(t-1)1}=1 \text{ and } t>1.\end{cases} \tag{5.7}$$

While for Y_{t2} missingness was fully MAR and dependent on present values of Y_{t3}, for the outcome the missingness mechanism depended on the time indicator t. In particular, at $t=1$ missing values were entered according to a MCAR mechanism. For $t>1$, missingness for Y_{t4} was MAR with a probability depending on the value of $Y_{(t-1)1}$. In such a way, we allowed the missingness mechanism of the outcome to depend also on past values. Mechanisms (5.6)-(5.7) led to about 30% missingness (across and within each time point) for both Y_2 and Y_4.

Missing data methods. After missingness was generated, we implemented three missing data techniques on the dataset. The first one was CC analysis. The second was the BMLM imputation technique presented in this chapter. Study 1 was mainly carried out to assess the performance of the standard Bayesian LM imputation model for longitudinal data. Therefore, in Study 1 we set $L=1$ for the time-constant mixture component. Using Gelman et al. (2013) 's method in Section 5.2.2 for model selection (running a preliminary Gibbs sampler with $K^*=25$ for 1000 burn-in and 2000 estimation iterations) led to an average (posterior maximum) number of states for the imputations equal to $K=15.38$ across the 200 datasets. The setting of the prior distribution is reported in Appendix A. For the imputation step, the Gibbs sampler was run for $B=3000$ iterations, in which $I=1000$ served as burn-in. For each dataset, $M=20$ imputations were performed.

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

Outcomes. Bias, stability (in terms of standard deviation of the produced estimates) and coverage rates of the 95% confidence intervals of the parameters in model (5.5) were used in order to evaluate the performance of each method.

5.3.1.2 Results

Results for bias, stability and coverage rates are reported in Table 5.2.

Overall, the BMLM outperformed both CC analysis and MICE in terms of bias. Not only relationships between the outcome and the predictors ($\beta_0, ..., \beta_3, \beta_{12}$) could be well recovered by the BMLM model, but also auto- and cross-lagged associations (ρ and τ). On the other hand, CC analysis was one of the best performing methods in terms of bias observed in the regression slopes of the predictors Y_1, Y_2 and Y_3, that is β_1, β_2, β_3, as well as the interaction β_{12}. However, the intercept β_0 and the auto-correlation parameter ρ and the crossed-lagged association parameter τ resulted somewhat biased with the CC analysis method. This means that these relationships were either weaker (ρ) or lost (τ) in the complete cases of the datasets. Lastly, the method that produced most biased estimates in Study 1 was MICE. While it could retrieve the estimates of τ successfully, it failed to capture other relevant relationships. In particular, due to the fact that MICE was run with its standard settings, the interaction effect resulted in extremely large bias, and the main effect parameter β_1 was far from its true value.

As far as the stability is concerned, the estimates yielded by the BMLM model were placed in an intermediate position between the ones resulting from CC analysis and MICE. In particular, Table 5.2 shows that the BMLM model estimates were more stable than the CC analysis estimates (due to the smaller sample size in the latter) and slightly (except β_{12}) less stable than the MICE estimates. The extra uncertainty (with respect to the MICE method) in the BMLM imputation estimates contributed -along with unbiasedness- to produce confidence intervals with coverage rates fairly close to their nominal level. Conversely, due to biased estimates (CC analysis) as well as too modest variability (MICE), most of the confidence intervals produced in the competing methods were too short. CC analysis yielded confidence intervals with coverage rate close to the exact ones for β_1, β_2, β_3, β_{12}.

5.3.2 **Study 2: time-constant and time-varying items**

In Study 2 we used the same approach for data and missingness generation and imputation of Study 1, with the addition of subject-level variables which will be used as new predictors in the substantive model under consideration. In Study 2 the BMLM imputation was implemented with $L > 1$, and its performance was again confronted with the CC analysis and MICE methods (the latter run with its default settings, as in Study 1). As a further challenge for the missing-data techniques, we generated *missing visits* (or complete missingness) for the time-varying variables, which often occur in longitudinal analysis.

Table 5.2: Study 1: results observed for the estimates of the AR logistic regression coefficients in model (5.5) for three missing data methods: CC (complete case analysis), BMLM (Bayesian mixture Latent Markov model with $L = 1$) imputation, MICE imputation. Large bias (in absolute value) and too low coverage rates are marked in boldface.

		Missing data method		
	Parameter	CC	BMLM	MICE
Bias	$\beta_0 = -0.40$	**0.34**	0.03	0.12
	$\beta_1 = 0.60$	-0.01	-0.07	**-0.26**
	$\beta_2 = -1$	-0.02	0.01	0.00
	$\beta_3 = 0.80$	-0.02	-0.03	-0.09
	$\beta_{12} = -0.80$	-0.03	0.08	**0.42**
	$\rho = 0.75$	**-0.17**	0.04	-0.15
	$\tau = 0.20$	**-0.22**	-0.05	0.00
Stability	$\beta_0 = -0.40$	0.17	0.16	0.15
	$\beta_1 = 0.60$	0.21	0.19	0.16
	$\beta_2 = -1$	0.20	0.20	0.17
	$\beta_3 = 0.80$	0.15	0.12	0.11
	$\beta_{12} = -0.80$	0.29	0.28	0.23
	$\rho = 0.75$	0.20	0.14	0.13
	$\tau = 0.20$	0.19	0.12	0.12
Coverage	$\beta_0 = -0.40$	**0.46**	0.94	**0.89**
Rate	$\beta_1 = 0.60$	0.95	0.94	**0.66**
	$\beta_2 = -1$	0.96	0.96	1.00
	$\beta_3 = 0.80$	0.94	0.92	**0.88**
	$\beta_{12} = -0.80$	0.96	0.97	**0.58**
	$\rho = 0.75$	**0.88**	0.97	**0.80**
	$\tau = 0.20$	**0.78**	0.98	0.95

Table 5.3: Values of the parameters in model (5.9).

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

5.3.2.1 Set-up

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

$$log\,\Pr(Z_1, Z_2, Z_3, Z_4) \propto 0.5.\sum_p Z_p - \sum_{p=1}^{3}\sum_{p'=p+1}^{4} Z_p Z_{p'} + 2.8 Z_1 Z_2 Z_3 \tag{5.8}$$

The population model for the three binary predictors Y_1, Y_2, Y_3 was the same as in (5.3)-(5.4). The AR logistic model for the time-varying outcome Y_4 was specified as

$$logit\,\Pr(Y_{t4}) = \begin{cases} \beta_0 + \beta_1 Y_{t1} + \beta_2 Y_{t2} + \beta_3 Y_{t3} + \beta_{12} Y_{t1} Y_{t2} + \mu_1 Z_1 + \mu_2 Z_2 & \\ \qquad + \mu_3 Z_3 + \mu_4 Z_4 & \text{if } t = 1 \\ \beta_0 + \beta_1 Y_{t1} + \beta_2 Y_{t2} + \beta_3 Y_{t3} + \beta_{12} Y_{t1} Y_{t2} + \mu_1 Z_1 + \mu_2 Z_2 & \\ \qquad + \mu_3 Z_3 + \mu_4 Z_4 + \rho Y_{(t-1)4} + \tau Y_{(t-1)3} & \text{if } t > 1. \end{cases} \tag{5.9}$$

Table 5.3 shows the parameter values chosen for β_0,...,β_{12}, ρ, τ, and $\mu_1, ..., \mu_4$. The new set of regression parameters $\mu_1, ..., \mu_4$ served to monitor how the missing data models could retrieve the relationships between the time-varying outcome and the time-constant variables. From the population model (5.3)-(5.4)-(5.8)-(5.9), we generated $N = 200$ datasets with $n = 200$ units and $T = 10$ time points.

Generating missingness. MAR missingness was generated in Z_1, Z_2, Y_1 and Y_3. Defining R_p equal to 1 when Z_p was missing and 0 otherwise for $p \in \{1, 2\}$, and R_{tj} as in Section 5.3.1.1, missingness was entered as follows. For the subject-level variable Z_1,

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

and Z_2

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

For the time-varying items, we generated missing values in Y_{t3} with a system analogous to mechanism (5.6) (conditioned on Y_{t2}) of Section 5.3.1.1, while for Y_{t1} the mechanism was

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

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

Missing data methods. As for Study 1, the BMLM model was compared with the CC analysis and MICE. Gelman et al. (2013)'s method described in Section 5.2.2 was used for the selection of L and K. Running a preliminary Gibbs sampler for each datasets led to select an average number of latent clusters equal to $L = 7.76$ and average number number of LSs equal to $K = 10.54$ (starting with $L^* = 10$ and $K^* = 15$, with 3000 iterations for the Gibbs sampler, of which 1000 for the burn-in). Appendix A reports how the prior distributions for the BMLM model were set. $B = 3000$ iterations were run for the imputation step, including $I = 1000$ of burn-in. MICE was implemented in the same way as in Study 1, with the only difference that now also the time-constant variables are included in the imputation model, serving as predictors of (and predicted by) the time-varying variables.

Outcomes. As done in Study 1 bias, stability, and coverage rates of the parameter estimates in model (5.9) were used to determine the behavior of the missing-data methods under investigation.

5.3.2.2 Results

Results for Study 2 are shown in Table 5.4. The BMLM imputation method could, overall, retrieve approximately unbiased parameter estimates not only for the predictors of the time-varying variables, but also for the parameters of the time-constant variables, $\mu_1, ..., \mu_4$. Similar to Study 1, CC analysis retrieved unbiased parameter

Table 5.4: Study 2: results observed for the estimates of the AR logistic regression coefficients in model (5.9) for three missing data methods: CC (complete case analysis), BMLM (Bayesian Mixture Latent Markov model) imputation, MICE imputation. Large bias (in absolute value) and too low coverage rates are marked in boldface.

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

estimates for the main effects parameters of the time-varying variables (as well as the main effects of the subject-specific variables), but retrieved biased intercept and lagged-relationships. The MICE imputation technique could not picked up the estimates of the main and interaction effects of time-varying variables (specially β_1 and β_{12}), but could recover unbiased lagged relationships (ρ and τ) and parameters of the time-constant effects.

As observed in Study 1, CC analysis produced the most unstable estimates among the three methods. Estimates yielded by the BMLM technique and MICE had, overall, similar stability for all types of regression coefficients, although the main and interaction effects of time-varying predictors produced by the BMLM model tended to vary more. The BMLM method yielded confidence intervals that were mostly close to their nominal level. MICE produced confidence intervals for the time-constant and lagged effects with coverage rates rather close to their nominal level, but intervals with too low coverage for main and interaction effects of the time-varying items. The confidence intervals computed after CC analysis were close to their nominal coverage level, excluding the intervals of β_0, ρ and τ, which resulted in a too low coverage.

5.4 Real-data Study

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

We used data collected by CentERData through their LISS panel, which consists of a (representative) sample of Dutch individuals, who participate in monthly Internet surveys. Key topics surveyed once per year include work, education, income, housing, time use, political views, values, and personality.[7] For our experiment, we selected the first 4 yearly waves ($T = 4$, from June 2008 until June 2011) of the Housing questionnaire.

7 More information about the LISS panel can be found at www.lissdata.nl.

Table 5.5: Real-data experiment: variables used in the panel regression model (5.10) (top part) and to generate missingness (bottom part). Type of variables: TV = time-varying; TC = time-constant. R = respondent.

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

5.4.1 Study set-up

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

Next, using this dataset, we estimated a panel regression model with random intercept and auto-regressive errors for the outcome variable 'House Satisfaction'[8]; this variable is denoted by Y_{t0} in Table 5.5. Among the remaining variables, we detected 7 (time-varying) predictors ($Y_{t1}, ..., Y_{t7}$ in Table 5.5) that were significant at the 5% level, yielding a total of $J = 8$ variables in the analysis model. Descriptions of these variables, including the time indicator t, are given in Table 5.5 (top part). Some of these were re-coded (transformed from continuous to categorical) and for others we collapsed some categories (so that their frequencies were not too small).

The panel regression model we estimated was

$$Y_{it0} = \beta_0 + \sum_{j=1}^{6} \beta_j Y_{itj} + \beta_{16} Y_{it1} Y_{it6} + \tau_1 Y_{i(t-1)1} + \tau_7 Y_{i(t-1)7} + u_{i0} + \epsilon_{it} \qquad (5.10)$$

8 The name of the variable was `cd08a001` in the original dataset.

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

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

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

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

The values of the model parameters $\beta_0, ...\beta_6, \beta_{16}, \tau_1, \tau_7, \sigma_1^2, \sigma_2^2, \rho$ estimated on the complete data are reported in the first columns of Table 5.6 below, along with their standard errors. All predictor effects were signficant at 5% level as highlighted, except for Y_{t6}, one of the variables yielding the significant interaction term β_{16}.

Generating missingness. Apart from the variables $Y_{t0}, ..., Y_{t7}$, we used the time-constant variable gender denoted with Z_1 in Table 5.5, to generate MAR missingness in the variable Y_{t1} (Z_1 was thus also included in the imputation models as a time-constant variable). In particular, by denoting the missingness of Y_{t1} with R_{t1}, we created missing values for Y_{t1} with the logistic model

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

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

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

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

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

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

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

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

Missing data methods. As in the studies reported in Section 5.3, we compared the performance of three missing data methods to retrieve the parameters of model (5.10): CC analysis, BMLM MI and MICE.

With CC analysis we estimated model (5.10) on the dataset with only complete observations, i.e., excluding all cases with missing data. This left a dataset with 591 rows, with sample sizes ranging from $n = 129$ at wave four to $n = 171$ at wave one.

We decided to run the BMLM model using two settings: the first with only one subject-level LC, i.e. with $L = 1$, and the second with $L > 1$. For both settings we performed model selection (for the number of LSs K) with Gelman et al. (2013) 's method reported in Section 5.2.2. In the first scenario, running the preliminary Gibbs sampler (for 5000 iterations, 2000 of which served as burn-in) with $K^* = 80$ led us to select $K = 55$. For the second scenario, we ran the preliminary Gibbs sampler with $L^* = 20$ and $K^* = 20$, and the same number of iterations as the previous case. This led us to choose $L = 18$ and $K = 9$. In what follows, the BMLM model with $L = 1$ and $K = 55$ will be denoted by BMLM(1), and the BMLM model with $L = 18$ and $K = 9$ by BMLM(2). In the subsequent step, $M = 50$ imputations were performed during 50000 iterations (plus 10000 iterations for the burn-in) for both BMLM(1) and BMLM(2).

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

Outcomes. We compared the results provided by each missing data method with the results observed for the complete-data case. In particular, we focused on the point estimates of all parameters in model (5.10) as well as the standard errors for the fixed effects $(\beta_0, ..., \tau_7)$. We also examined which fixed effect estimates were significant at a 5% level.

5.4.2 Results

The results are reported in Table 5.6. Both CC analysis and the two versions of the BMLM imputation model retrieved point estimates of the fixed effects rather close to those of the complete-data analysis. Exceptions for the CC analysis were the main effects β_1 and β_6 and the interaction term β_{16}, which were slightly different from the corresponding values obtained with the complete data. Some of the standard errors yielded by CC analysis were inflated because of the limited sample

Table 5.6: Real-data experiment: results for the parameters in model (5.10). Est. = point estimate. S.E. = standard error. 5% significant predictors are denoted with a '*' next to the point estimates obtained with each method.

			Missing Data Method							
Parameter	Complete Data		CC analysis		BMLM(1)		BMLM(2)		MICE	
	Est.	S.E.	Est.	S.E.	Est.	S.E.	Est.	S.E.	Est.	S.E.
β_0	0.86*	0.23	1.03*	0.30	0.99*	0.27	0.99*	0.29	1.04*	0.27
β_1	0.73*	0.08	0.67*	0.11	0.67*	0.10	0.66*	0.10	0.65*	0.11
β_2	0.12*	0.02	0.12*	0.03	0.08*	0.03	0.09*	0.03	0.10*	0.03
β_3	-0.05*	0.02	-0.06	0.03	-0.06*	0.03	-0.06*	0.03	-0.06*	0.03
β_4	-0.52*	0.16	-0.49*	0.22	-0.51*	0.18	-0.48*	0.20	-0.40*	0.19
β_5	-0.09*	0.03	-0.12*	0.04	-0.08*	0.04	-0.08*	0.04	-0.08	0.04
β_6	0.07	0.04	0.03	0.06	0.09	0.05	0.11	0.05	0.07	0.05
β_{16}	-0.05*	0.02	-0.03	0.02	-0.05*	0.02	-0.05*	0.02	-0.05*	0.02
τ_1	0.11*	0.02	0.12*	0.03	0.12*	0.03	0.11*	0.03	0.12*	0.03
τ_7	-0.10*	0.03	-0.11*	0.05	-0.11*	0.04	-0.09*	0.04	-0.12*	0.04
σ_1^2	0.19	-	0.20	-	0.20	-	0.21	-	0.21	-
σ_2^2	0.25	-	0.26	-	0.28	-	0.28	-	0.30	-
ρ	0.13	-	0.07	-	0.12	-	0.11	-	0.10	-

size exploited by this method, which made some parameter estimates no longer significant at the 5% level (in Table 5.6, some fixed effects are no longer marked with a '*'). Conversely, despite a couple of values being slightly off (the intercept β_0 and the main effect β_1), both BMLM(1) and BMLM(2) could exploit the original sample size, causing the standard errors to be only slightly larger than those of complete-data analysis (reflecting in this way the imputation step uncertainty). As a result, all parameters that were significant with the full data were also significant after imputing the missing values with the BMLM model. The MICE method did not manage to recover well all parameter estimates; for instance, the intercept β_0 and the main effects β_1 and β_4 were (in a more or less pronounced manner) far from the estimates of the complete-data condition, while the standard errors observed after imputing the data with MICE were close to the BMLM MI estimates. Nevertheless, the parameter β_5 which was significant with the complete data and the BMLM imputation method, was no longer significant with the MICE.

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

5.5 Discussion

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

The performance of BMLM-based MI approach was evaluated and compared with other two missing data methods, CC analysis and MICE, by means of two simulations studies and a real-data experiment. In the simulation studies, the analysis model used was a logistic model including an auto-regression term and a crossed-lagged relationship coefficient (Study 1), as well as main effects of time-constant predictors (Study 2). Results showed a good (overall) performance of the BMLM imputation model compared with the competing methods, since it could retrieve (approximately) unbiased estimates for all types of parameters specified in the sub-

stantive models, with coverage rates of the confidence intervals that were never too small compared to their 95% nominal level. The good performance of the BMLM model in Study 2 showed that the model can also cope with missing visits when these are present at any time point. Conversely, CC analysis could not recover well the lagged relationships in terms of both bias and confidence intervals, with coverage rates that were too low for their nominal level, while MICE provided biased time-varying main and interaction effects, with corresponding confidence intervals that tended to be too narrow.

In the real-data experiment we estimated a panel regression model using data from the LISS panel. The model included main and interaction fixed effects, along with crossed-lagged relationships and random intercept. Furthermore, the distribution of the residuals was described by a variance and an auto-correlation coefficient. As done in Study 2, we also included cases with missing visits in the LISS dataset as a further challenge for the missing data methods. Results demonstrated the superiority of the BMLM model (run with two different conditions of number of clusters and states) when compared to competing methods; in particular, the same conclusions (i.e., the same terms were statistically significant) were drawn for the complete-data analysis and the BMLM imputation method. This did not happen with the CC and the MICE techniques, for which some terms were not significant anymore. In addition, the BMLM method retrieved variance and error components close to the complete-data analysis.

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

Furthermore, the proposed imputation model itself can be extended in various useful ways. Firstly, while we dealt with categorical (both ordinal and nominal) variables, the BMLM model can be extended to accommodate mixed types of data, i.e., it can be implemented on datasets containing both categorical and continuous variables. This can be achieved, for instance, by specifying mixtures of univariate Normal and Multinomial distributions. Second, although we assumed the BMLM

model to have a Markov chain of order 1, it is possible to consider lags of higher orders by conditioning the distribution of the dynamic LSs at time t on the configuration of the states at earlier time points, e.g. $t-2$, $t-3$, etc., if these kinds of lags are needed in the substantive analysis. Third, when the measurement may occur at different continuous time points rather than at fixed discrete occasions, imputations of the missing data can be provided by assuming a continuous-time latent Markov chain for the distribution of the LSs. Last, for applications in which the subjects observed across time are coming from different groups (e.g., patients coming from different hospitals), the model can be moved towards a multilevel framework, for instance, by adding a further LC variable at the group-level.

APPENDIX

A Setting the prior distribution

As outlined in Section 5.2.1, independent Dirichlet distributions can be specified for each Multinomial in model (5.1)-(5.2). In a MI context, in which the imputation model does not necessarily match the analysis model, it is common to have no previous knowledge about the imputation model parameters. In such a case, symmetric Dirichlet priors can be chosen: $Dirichlet(c_1, c_2, ..., c_D)$ where $c_1 = c_2 = ... = c_D$. This is the approach we used in all the experiments of the chapter, and implied in the remaining of the current section.

Rousseau and Mergensen (2011) found out that when a Bayesian mixture model is overfitting the data (as our model selection approach of Section 5.2.2 implies), units are allocated by the Gibbs sampler to some of the extra LCs if each component of the latent probabilities hyperparameter is at least as large as half times the number of free parameters within each components. For the BMLM model, this means that each pseudo-count of the LSs $\alpha_k \ \forall \ k$ should be set at least equal to $\sum_j (R_j - 1)/2$. Following the guidelines of Chapter 3, who examined the behavior of the prior distribution for standard Bayesian LC models (for the MI of cross-sectional missing data), we suggest increasing α_k and $\gamma_k \ \forall \ k$ in such a way that as many states $s_1, ..., s_T$ as possible are occupied during the imputation stage, which can be assessed with the MCMC output. By manipulating with trial-and-error (before the imputation step) the hyperparameters in the priors of the latent states probabilities, we decided to set $\alpha_k = \gamma_k = 5$ in Study 1 and 2 of Section 5.3, while in the real-data experiment of Section 5.4 - in which the number of within-state free parameters was equal to 27 - we arbitrarily set $\alpha_k = \gamma_k = 100$ (for both the BMLM(1) and the BMLM(2)). As reported in Chapter 3, full allocation of the latent classes/states helps to capture all relevant associations in the data, preventing the sampler from becoming unstable; in fact, in this way the states are identified by the data, rather than by the prior distribution of the emission probabilities.

In Study 1 and in the empirical study we found out by means of pre-imputation inspections that reinforcing the prior persistence probabilities caused the Gibbs sampler to produce higher likelihood values (on average) during its iterations. In turn, this could help the BMLM model to better recover the lagged relationships specified

for that study. Persistence probabilities are represented by the diagonal elements of the matrix $\boldsymbol{X}_l$. These probabilities can be reinforced by manipulating the hyperparameter vector of the q-th row of $\boldsymbol{X}_l$, by setting it equal to $\boldsymbol{\gamma} = (\gamma_1, ..., \gamma_q^*, ..., \gamma_K)$ with $\gamma_q^* > \gamma_k \; \forall \; k \neq q$. In Study 1 we achieved this by setting $\gamma_q^* > \sum_{k \neq q} \gamma_k$, with $\gamma_k = 5$ and $\gamma_q^* = K\gamma_k = 5K$, while in the empirical study this was done with $\gamma_k = 100$ and $\gamma_q^* = K\gamma_k = 100K$ (in this study, K was equal to 55 for the BMLM(1) and to 9 for the BMLM(2)). Reinforcing the persistence probabilities in Study 2 was not necessary, since increasing this did not entail any increase in the (averaged) likelihood values produced during the Gibbs sampler iterations.

Concerning the hyperparameters for the weights of the time-constant LCs, we decided to perform the imputations of Study 2 in Section 5.3 and the real-data experiment in Section 5.4 (for the BMLM(2) model) by setting η_l equal to the number of free parameters within each time-constant component, i.e., we set $\eta_l = \{(K-1)(K+1) + K(\sum_j R_j - 1) + \sum_p U_p - 1\} \; \forall \; l$.

Lastly, for the time-constant conditional and the time-varying emission probabilities we follow the guidelines of Chapter 3 and set $\zeta_{upl} = \delta_{rjkl} = 0.01$ or 0.05 $\forall \; u, p, r, j, k, l$ (final results are usually similar for these two values). This setting helps to make the prior pseudo-counts of the parameters ruling the conditional distribution of the observed data less influential in the imputation step.

B BMLM model estimation

In this section, the Gibbs sampler for the BMLM model estimation is described. It is assumed that L, K, and the model hyperparameters have been established already according to the guidelines of Section 5.2.2 and Appendix A. Furthermore, also the total number of Gibbs sampler iterations B should be chosen. I of these B iterations will be used as burn-in (such that model estimation is performed on the last $B - I$ iterations). I should be large enough to make the sampler attain the equilibrium distribution of the model parameter, which can be assessed by typical MCMC output inspection, e.g., by considering the traceplot of the log-likelihood functions generated at each iterations (as suggested in Chapter 3). Additionally, $\boldsymbol{\theta}^{(0)}$ is initialized by sampling all model parameters from uniform Dirichlet distributions, in such a way to increase the likelihood of initializing the sampler in the interior of the parameter space, speeding up convergence.

Algorithm 5.1 reports the steps for the Gibbs sampler. In order to sample the states of the Markov chain for each subject, multi-move sampling is used. The steps necessary to perform multi-move sampling are shown in Algorithm 5.2. Multi-

move sampling, in turn, requires the calculation of the filtered state probabilities $\Pr(s_t = k|\boldsymbol{\theta}, w = l, \mathbf{y}_{it})$, the computation of which is described in Algorithm 5.3.

B.1 The Gibbs sampler

Algorithm 5.1

For b=1,...,B:

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

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

   for each $l = 1, ..., L$, and where $\pi_{\mathbf{r}^* l} = \Pr(\mathbf{y}_i = \mathbf{r}^*|w = l)^{(b-1)}$ (equation 5.2);

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

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

   This can be achieved with multi-move sampling (see Algorithm 5.2 below);

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

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

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

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

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

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

5. for $l = 1, ..., L$ compute $\pi^{(b)}_{\mathbf{r}^* l}$ conditioned on $w^{(b)} = l$ after updating the parameter values of each within-class LM model:
   - for $t = 1$, update the initial state probabilities

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

   where $\mathcal{I}_{it}(s_t^{(b)} = k) = 1$ if for unit i $s_t^{(b)} = k$ and o otherwise;
   - for $q = 1, ..., K$ and $\forall\ t \geq 2$ update the transition probabilities

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

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

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

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

B.2 Multi-move sampling

Algorithm 5.2:

1. For i=1,...,n calculate and store the filtered state probabilities $\Pr(s_t^{(b)} | \boldsymbol{\theta}^{(b-1)}, w^{(b)} = l, \mathbf{y}_{it})$ for $t = 1, ..., T$ (see Algorithm 5.3);
2. for $i = 1, ..., n$ sample $s_T^{(b)}$ from $\Pr(s_T^{(b)} | \boldsymbol{\theta}^{(b-1)}, w^{(b)} = l, \mathbf{y}_{iT})$;

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

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

B.3 Filtered State Probabilities

Algorithm 5.3:

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

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

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

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

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

2. for $t = 2, ..., T$:
   - for $i = 1, ..., n, k = 1, ..., K$ compute
   $\Pr(s_t^{(b)} = k | \boldsymbol{\theta}^{(b-1)}, \mathbf{y}_{i(t-1)}) = \sum_q \xi_{q,kl}^{(b-1)} \Pr(s_{t-1}^{(b)} = q | \boldsymbol{\theta}^{(b-1)}, w^{(b)} = l, \mathbf{y}_{i(t-1)})$;
   - for $i = 1, ..., n, k = 1, ..., K$ compute the filtered state probabilities through
   $\Pr(s_t^{(b)} = k | \boldsymbol{\theta}^{(b-1)}, w^{(b)} = l, \mathbf{y}_{it} = \mathbf{r}_t) =$

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

where

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

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

6

DISCUSSION

This dissertation proposed and investigated the use of Bayesian mixture models for the multiple imputation (MI) of categorical data for different study designs. Chapters 2 and 3 dealt with MI of cross-sectional categorical data. Chapter 2 gave an overview of different (frequentist and Bayesian) latent class (LC) imputation models present in the literature, and highlighted pros and cons of each method. In Chapter 3 a closer investigation of Bayesian LC analysis for MI was presented, and the effect of different prior distributions on the substantive inferences was shown by means of a simulation study. Chapter 4 was about MI of multilevel categorical data, and for this purpose a Bayesian multilevel LC (BMLC) model was proposed. A simulation study and a real-data study showed the good performance of the model at retrieving correct inferences for the substantive analysis model considered (random intercept and random slope logistic regression models). Last, Chapter 5 dealt with the imputation of longitudinal data, proposing the use of Bayesian mixture latent Markov (BMLM) models for this task. The good performance of the BMLM model as an imputation model was established by means of two simulation studies and a real-data application. Overall, it has been shown that mixture models represent a flexible imputation method, which can capture complex relationships in the data with a very simple model specification. Furthermore, the imputation models are tailored for the sampling design used at the data collection stage, an aspect that allows them to capture all the demanded variability to perform the imputations.

Despite the promising results observed for the Bayesian LC imputation models, not all issues related to their use have been addressed in this dissertation. A first problem, common to all the models presented in this thesis, is related to model selection. As constantly remarked in the thesis, when dealing with mixture imputation models the main concern is about data underfitting, rather than overfitting. This

means that capturing features that are specific for the sample is of little concern here, since these also include the relevant associations present at the population level. On the other hand, ignoring relevant associations of the population leads to biased post-imputation inferences. We decided, as a strategy, to select the number of classes by exploiting a characteristic of the Gibbs sampler when used in combination with mixture models. That is, by manipulating the prior distribution of the latent class weights it is possible to determine beforehand whether units should be allocated to all the LCs, or only to those classes that are non-redundant. Our proposed method was used for model selection by Gelman et al. (2013) in a substantive analysis framework, and it consists of running a preliminary Gibbs sampler in order to assess how many classes are filled by units during model estimation, with a prior distribution that encourages the emptying of extra components. This procedure leads to a posterior distribution of the number of classes given the data, and the approach suggested in this dissertation is to pick the maximum of such distribution and perform the imputations with this number of classes. Despite of the fact that the method was shown to work rather well in the simulation studies and real-data applications carried out and reported in the dissertation, we did not investigate the performance of our imputation models with different choices for the number of classes. For instance, selecting the posterior mode - rather than the posterior maximum - would theoretically lead to faster computations and to larger stability of the post-imputation parameter estimates, but possibly also to an increase in the bias of such estimates. On the other hand, increasing the value of the posterior maximum (by some arbitrary constant) would produce most likely a decrease in bias, but not without a loss in stability and slower computations at the imputation stage. Therefore, the choice of the number of classes should depend on the main focus of the analysis (bias/stability) and should be inspected in more detail in future research. For the BMLM model of Chapter 5, an alternative option for the choice of the number of latent states could be represented by their maximum posterior mode observed across all time points, which should also be tested in future research.

A problem related to the Bayesian specification of the model is the choice of the prior distribution. The approach used in the thesis was to use independent (symmetrical) Dirichlet distributions for each set of multinomial probabilities involved in the imputation models. As a general approach, we recommend using hyperparameters for the mixture(s) weights that prevent the occurrence of empty components during the imputations, and very small pseudo-counts for the probabilities of the variables conditioned on the LCs. In this way, the Gibbs sampler can perform model estimation without sampling from the prior distribution of empty clusters, and the distribution of the observed variables conditioned on the LCs is determined by the data. With such a configuration, the resulting imputations are

more accurate than the imputations performed with other types of prior distributions (e.g., with uniform priors for both the mixture weights and the conditional response probabilities). However, while in Chapter 3 (for standard LC models) and 4 (for multilevel LC models) a number of possible prior conditions was compared to assess their effect on the final imputations, we did not provide a full overview of all possible prior specifications. For instance, we did not consider the effect of Jeffrey's priors for the conditional response probabilities, nor we considered possible data-driven priors (i.e., priors that can be derived from some characteristic of the observed data, such as their empirical distribution). Furthermore, it is possible that the data analyst and/or the data imputer is in possession of some prior information concerning the imputation model parameters, or data relationships. In these cases, informative (non-symmetrical) Dirichlet priors can be specified for the model, and they should make the imputations more precise. Therefore, a more complete overview on the specification of the hyperparameters of the model is needed in the literature. Morevoer, the BMLM model of Chapter 5 was not tested with all possible combinations of prior specifications, and it is possible that better settings can be found for such model, also with symmetric priors.

The imputation models described in this dissertation can be extended in a number of ways. An issue not covered is the scale type of the variables in the dataset. While survey data are mostly observed on a categorical (nominal/ordinal) scale, it is common to have to deal with datasets that contain both continuous and categorical variables. The LC models proposed in this thesis are designed only to describe associations under the first scenario, but they can easily be extended to accommodate for both types of scales. A possible solution can be given by using mixtures of independent univariate Multinomial (for the categorical variables in the dataset) and Normal (for the continuous part of the dataset) distributions. Similar to the Bayesian LC models encountered in this dissertation, model selection can still be performed with Gelman et al. (2013)'s method. However, LC imputation models with such mixed data type configuration have not been tested yet in the literature. Therefore, their functioning in this context must be properly assessed.

In this project, we did not deal with impossible score combinations in the data (also known as structural zeroes), which means that the 'unconstrained' imputation models proposed in the dissertation can produce with nonzero probability imputations with -for example- pregnant fathers, or married kids. These issues can be overcome by accounting for structural zeroes within the imputation model, as proposed for instance by Manrique-Vallier and Reiter (2014) and Hu, Reiter and Wang (2017) for the estimation of Dirichlet processes in cross-sectional and multilevel contexts, correspondingly. In particular, they assumed that the observed data are sampled from a restricted set in which structural zeroes are impossible. This

restricted set is in turn a subset of a larger, augmented space, in which also the impossible combinations can occur. With appropriate modifications of the Gibbs sampler, Manrique-Vallier and Reiter (2014) showed how this technique allows to sample from the correct distribution of the parameter space accounting for impossible combinations (by integrating over the structural zeroes in the augmented space), while Hu et al. (2017) exploited this method to generate multilevel synthetic data. In a similar fashion, this approach can be used to perform the imputations with the LC models of this thesis, and its effectiveness should be explored in the future. Another approach could be given by collapsing the variables with structural zeroes and eliminate those categories corresponding to the impossible combinations. Another method to deal with this issue might be to allow for local dependencies between variables that comprise structural zeroes, and restraining their parameter space with the prior distribution. These alternative approaches should also be investigated in future studies.

The BMLC imputation model of Chapter 4 was developed for two-level hierarchies (e.g., students within schools). Nevertheless, more complex sampling designs can occur in scientific research: for instance, when students within different schools are observed from multiple regions in a country, or even from multiple countries. In the former case, the hierarchy is composed of three levels (students, schools, regions) while in the latter case a fourth level (countries) is needed to explain all the necessary variability in the data. Fortunately, it is possible to generalize the BMLC model of Chapter 4 to a larger number of hierarchical levels, which involves having LCs at each of the levels and conditioning lower-level model probabilities on higher-level class memberships. The Gibbs sampler can easily be modified to account for such more complex structures. The adequacy of the BMLC model for contexts with larger number of levels in the hierarchy should therefore be considered and inspected in the future.

The BMLM model of Chapter 5 assumes a first-order Markov chain for the distribution of the latent states to capture lagged relationships in the time-varying variables. While it makes sense to assume that adjacent time points carry over dependencies among each other, it is possible for some applications that relationships at higher lags occur in the data, and that these are needed in the analysis model. In this case, the BMLM model should take into account relationships between more distant time points. The BMLM model can be extended to accommodate for such relationships. In order to achieve this, a second- (or higher-) order Markov chain must be assumed for the latent states, the distribution of which must then be conditioned not only on the previous time point, but also to earlier points in time. As a proposal for future research, it would be interesting to compare the performance of the first-order mixture latent Markov model proposed in this thesis with a higher-

order mixture latent Markov model when in the analysis model dependencies with higher-order lags are needed.

Lastly, the BMLM model proposed in this dissertation deals with individuals (subjects) observed over time. However, these individuals in turn may come from multiple groups (higher-level units such as schools, hospitals, and so on) and also variables for the higher-level units may be observed over time. In these cases, the BMLM model must be framed into a multilevel setting in order to take into account the new form of dependencies present in the data. This can be done, for instance, by adding a level of hierarchy to the simple BMLM model and introducing a LC variable at this level (similar to the BMLC models with new higher-levels of hierarchy) which picks up possible heterogeneity between the clusters from which the lower-level subjects are observed.

BIBLIOGRAPHY

Akande, O., Li, F. & Reiter, J. (2017). An empirical comparison of multiple imputation methods for categorical data. *The American Statistician, 71(2)*, 162-170.

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(1)*, 57-74.

Audigier, V., White, I. R., Jolani, S., Debray, T., Quartagno, M., Carpenter, J., ... Resche-Rigon, M. (2017). *Multiple imputation for multilevel data with continuous and binary variables.* Retrieved from https://arxiv.org/abs/1702.00971

Baraldi, A. N. & Enders, C. K. (2010). An introduction to modern missing data analyses. *Journal of School Psychology, 48(1)*, 5-37.

Baum, L. E., Petrie, T., Soules, G. & Weiss, N. (1970). A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains. *Annals of Mathematical Statistics, 41(1)*, 164-171.

Brand, J. P. L. (1999). *Development, implementation and evaluation of multiple imputation strategies for the statistical analysis of incomplete data sets* (Chapter 5). Dissertation. Erasmus University Rotterdam, The Netherlands.

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

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

Congdon, P. (2006). *Bayesian statistical modelling* (Second ed.). Chichester: Wiley.

Dempster, A. P., Laird, N. M. & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society Series B (Methodological), 39(1)*, 1-38.

Diebolt, J. & Robert, C. (1994). Estimation of finite mixture distributions through Bayesian sampling. *Journal of the Royal Statistical Society Series B (Methodological), 56(2)*, 363-375.

Drechsler, J. (2015). Multiple imputation of multilevel missing data - Rigor versus simplicity. *Journal of Educational and Behavioral Statistics, 40(1)*, 69-95.

Dunson, D. B. & Xing, C. (2009). Nonparametric Bayes modeling of multivariate categorical data. *Journal of the American Statistical Association, 104(487)*, 1042-

1051.

Efron, B. (1979). Bootstrap methods: another look at the Jakknife. *The Annals of Statistics, 7(1)*, 1-26.

Escobar, M. D. & West, M. (1995). Bayesian density estimation and inference using mixtures. *Journal of the American Statistical Association, 90(430)*, 577-588.

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

Gebegziabher, M. & DeSantis, S. (2010). Latent class based multiple imputation approach for missing categorical data. *Journal of Statistical Planning and Inference, 140(11)*, 3252-3262.

Gelfand, A. E. & Smith, A. F. M. (1990). Sampling based approaches to calculating marginal densities. *Journal of the American Statistical Association, 85(410)*, 398-409.

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

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

Goldstein, H., Carpenter, J. R., Kenward, M. & Levin, K. (2009). Multilevel models with multivariate mixed response types. *Statistical modelling, 9(3)*, 173-197.

Goodman, L. A. (1974). Exploratory latent structure analysis using both identifiable and unidentifiable models. *Biometrika, 61(2)*, 215-231.

Graham, J. W., Olchowski, A. E. & Gilreath, T. D. (2007). How many imputations are really needed? Some practical clarifications of multiple imputation theory. *Prevention Science, 8(3)*, 206-213.

Graham, J. W. & Schafer, J. L. (1999). On the performance of multiple imputation for multivariate data with small sample size. In R. Hoyle (Ed.), *Statistical strategies for small sample research*, pp. 1-29. Thousand Oaks, CA: Sage.

Hoijtink, H. & Notenboom, A. (2004). Model based clustering of large data sets: tracing the development of spelling ability. *Psychometricka, 69(3)*, 481-498.

Horton, N. J., Lipsitz, S. & Parzen, M. (2003). A potential for bias when rounding in Multiple Imputation. *The American Statistician, 57(4)*, 229-232.

Hu, J., Reiter, J. P. & Wang, Q. (2017). Dirichlet process mixture models for modeling and generating synthetic versions of nested categorical data. *Bayesian Analysis. doi: 10.1214/16-BA1047.*. Retrieved from http://projecteuclid.org/euclid.ba/1485227030

Huisman, M. (1999). *Item nonresponses: Occurrence, causes, and imputation of missing answers to test items*. Leiden, The Netherlands: DSWO Press.

Ishwaran, H. & James, L. F. (2001). Gibbs sampling for stick-breaking priors. *Journal*

of the American Statistical Association, 96(453), 161-173.

Jolani, S., Debray, T., Koffijberg, H., Van Buuren, S. & Moons, K. (2015). Imputation of systematically missing predictors in an individual participant data meta-analysis: a generalized approach using MICE. *Statistics in Medicine, 34(11)*, 1841-1863.

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

Linzer, D. & Lewis, J. (2014). poLCA: Polytomous variable Latent Class Analysis [Computer software manual]. Retrieved from http://cran.r-project.org/web/packages/poLCA/index.html (R package version 1.4.1.)

Little, R. J. A. & Rubin, D. B. (2002). *Statistical analysis with missing data* (2nd ed.). New York: Wiley.

Liu, J. S. (1994). The collapsed Gibbs sampler in Bayesian computation with application to a gene-regulation problem. *Journal of the American Statistical Association, 89(427)*, 958-966.

Maas, C. & Hox, J. (2005). Sufficient sample sizes for multilevel modeling. *Methodology: European journal of Research methods for the Behavioral and Social sciences, 1(3)*, 86-92.

Manrique-Vallier, D. & Reiter, J. P. (2014). Bayesian estimation of discrete multivariate latent structure models with strutural zeros. *Journal of Computational and Graphical Statistics, 23(4)*, 1061-1079.

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

National Opinion Research Center. (1972). General Social Survey. GSS (1972-2014) Release 4. Cross-sectional wave 2014. Retrieved from http://www3.norc.org/GSS+Website/Download/SPSS+Format/. University of Chicago.

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.

Paccagnella, O. (2011). Sample size and accuracy of estimates in multilevel models. *Methodology 7(3)*, 111-120.

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(3)*, 143-149.

Resche-Rigon, M. & White, I. (2016). Multiple imputation by chained equations for

systematically and sporadically missing multilevel data. *Statistical Methods in Medical Research*, 1-16. doi: https://doi.org/10.1177/0962280216666564

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

Roth, P. L. (1994). Missing data: A conceptual review for applied psychologysts. *Personnel Psychology, 47(3)*, 537-560.

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

Rubin, D. B. (1976). Inference and missing data. *Biometrika, 63(3)*, 581-592.

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

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

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

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(2)*, 437-457.

Schlomer, G. L., Bauman, S. & Card, N. (2010). Best Practices for Missing Data Management in Counseling Psychology. *Journal of Counseling Psychology, 57(1)*, 1-10.

Sethuraman, J. (1994). A constructive definition of Dirichlet priors. *Statistica Sinica, 4(2)*, 639-650.

Si, Y. (2012). *Nonparametric Bayesian methods for Multiple Imputation of large scale incomplete categirical data in panel studies*. Ph.D. Thesis. Duke University, USA.

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(5)*, 499-521.

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

Van Buuren, S. (2011). Multiple imputation of multilevel data. In Eds, Hox J, J. & Roberts J, K. (eds.), *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(12), 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.

Van Buuren, S., Groothuis-Oudshoorn, K., Robitzsch, A., Vink, G., Doove, L. & Jolani, S. (2014). mice: Multivariate Imputation by Chained Equations [Computer software manual]. Retrieved from http://cran.r-project.org/web/packages/mice/index.html (R package version 2.22)

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

Van den Broek, K., Nyklicek, I., Van der Voort, P., Alings, M. & Denollet, J. (2008). Shocks, personality, and anxiety in patients with an implantable defibrillator. *Pacing and Clinical Electrophysiology, 31*, 850-857.

Van der Palm, D. W., Van der Ark, L. A. & Vermunt, J. K. (2014). Divisive latent class modeling as an incomplete-data method for categorical data. *Manuscript submitted for publication*.

Van der Palm, D. W., Van der Ark, L. A. & Vermunt, J. K. (2016a). A comparison of incomplete-data methods for categorical data. *Statistical Methods in Medical Research, 25(2)*, 754-774.

Van der Palm, D. W., Van der Ark, L. A. & Vermunt, J. K. (2016b). Divisive latent class modeling as a density estimation method for categorical data. *Journal of Classification, 33(1)*, 52-72.

Van Ginkel, J. R. (2007). *Multiple imputation for incomplete test, questionnaire and survey data*. Ph.D. Thesis. Tilburg University, The Netherlands.

Vermunt, J. K. (2003). Multilevel latent class models. *Social methodology, 33(1)*, 213-239.

Vermunt, J. K. (2008). Latent class and finite mixture models for multilevel datasets. *Statistical Methods in Medical Research, 17(1)*, 33-51.

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

Vermunt, J. K. & Magidson, J. (2013). *LatentGOLD 5.0 Upgrage 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(1)*, 369-397.

White, A. & Murphy, B. (2014). BayesLCA: Bayesian Latent Class Analysis [Computer software manual]. Retrieved from http://cran.r-project.org/web/

packages/BayesLCA/index.html (R package version 1.5)

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

Wilkinson, L. & Task Force on Statistical Inference. (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist, 54(8)*, 594-604.

Yucel, R. (2008). Multiple imputation inference for multivariate multilevel continuous data with ignorable non-response. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, 366(1874)*, 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)

SUMMARY

This dissertation investigates the use of *latent class* (or *mixture*) models for Multiple Imputation (MI). MI is a technique that enables the retrieval of parameter estimates and the performance of statistical inference in the presence of missing data in a dataset. While missing data may represent an issue for standard statistical analysis (e.g., they can introduce bias and loss of power in the final or substantive analysis), MI seeks to fix the problem by replacing the missing data with plausible imputed data, predicted by means of an imputation model. Repeating the replacements (or imputations) several times allows the uncertainty of the imputed values to be taken into account, and leads to valid inferences.

In this context, the choice of the imputation model is crucial: it should not only preserve all the relevant relationships needed for a specific analysis of interest (e.g., the main effects of a regression reflect the relationships between the outcome and the predictors), but it should be able also to reflect overall relationships present in the data, in such a way to allow to carry out further analyses with other (more complex) kinds of associations (e.g., interaction terms represent the simultaneous relationship between two predictors and the outcome). Thus, in MI we are interested in the predictions produced by the imputation model - and how they reflect relationships among variables - rather than in interpreting its parameter values. The broader the imputation model, the better it can capture important relationships in the data. As a consequence, overfitting the data with the imputation model is of smaller concern than underfitting: while an underfitting model might ignore important relationships of the data, an overfitting one takes into account all relevant relationships, as well as sample-specific fluctuations. As a result, in the former case the model could produce too poor imputations, while in the latter case the relevant relationships are preserved by the model.

The thesis deals in particular with the MI of missing categorical data; while methods for continuous data have been extensively explored, in the literature there is a lack of MI models for categorical data. With categorical data, the focus is on retrieving relevant associations in the joint distribution of the categorical variables of a dataset. The saturated log-linear model, which takes into account all theoretically possible associations of the data, is a typical choice in this context. However, saturated log-linear models are computationally appealing only with a small number of items. As a solution, recent proposals for the MI of categorical data include the use of either latent class analysis (frequentist framework) or the Dirichlet Process Mixture of Multinomial Distributions (Bayesian framework) as imputation models, which both belong to the family of mixture models. Unlike MI via saturated log-

linear models, MI through latent class models can be performed on datasets containing a large number of variables by means of the *local independence* assumption, which assumes independence between variables once their distribution is conditioned on the latent classes.

In order to reflect all the necessary variability for the imputations, the imputation model should be tailored for the design used to collect and analyze the data. For instance, cross-sectional data need a model that takes all relevant associations among items into account; with multilevel data, in which several lower-level units are nested within higher-level units (such as students nested within schools), correlations and dependencies arising from units of the same group must be also accounted for; with longitudinal data, variables are observed over time for the same units, and auto-correlations and lagged relationships are likely to arise. Ignoring these aspects of the data may lead to underfitting and, as a consequence, to biased (and/or too stable) post-imputation inferences. The purpose of this thesis is to propose and investigate different types of latent class models for the MI of categorical data; each of these types of models are tailored for the design chosen for the data collection and analysis. Thus, Chapter 2 of the thesis offered a review of the latent class models present in the literature for the MI of cross-sectional categorical data. Chapter 3 investigated in detail the behavior of Bayesian latent class models for the MI of cross-sectional data. Chapter 4 examined the behavior of Multilevel latent class models for the MI of multilevel data. Lastly, Chapter 5 assessed the performance of the Mixture latent Markov model for the imputation of longitudinal data. All models presented in the thesis have been developed under a Bayesian framework and estimated by means of the Gibbs sampler. Bayesian analysis is well-suited for MI, since it automatically accounts for the variability caused by both the missing data distribution and the parameter uncertainty. Another purpose of the thesis was to find a way to perform model selection which is suitable for MI. With mixture models, model selection is equivalent to detecting the number of components (or classes) to be used at the imputation stage. To achieve this, we exploited a feature of the Gibbs sampler run in combination with mixture models: with a preliminary run of the sampler (and with a particular setting of the prior distribution of the mixture components), it is possible to obtain a (posterior) distribution of the number of classes actually occupied by the data. As a general approach, we chose the maximum of this distribution in order to perform the imputations, in such a way to use the broadest possible imputation model.

As mentioned above, Chapter 2 provided an overview of latent class models for the MI of missing categorical data present in the literature. Latent class modeling, mainly known as a clustering tool, can be used for density estimation, i.e., to get a good description of the lower- and higher-order associations among the variables in

a dataset. For MI, the latter aspect is essential in order to be able to draw meaningful imputing values from the conditional distribution of the missing data given the observed data. In this chapter, we explained the general logic underlying the use of latent class analysis for MI, and presented several variants developed within either a frequentist or a Bayesian framework. In this chapter, the different approaches were illustrated and compared using a real-data psychological assessment application.

The great advantage of using latent class models for MI is represented by their flexibility, which allows to capture complex relationships in the data, given that the number of specified latent classes is large enough. However, the frequentist latent class model and the Dirichlet Process Mixture of Multinomial Distributions both have certain disadvantages. The frequentist approach is computationally demanding because it requires estimating many latent class models: first models with a different number of classes should be estimated to determine the required number of classes, and subsequently the selected model is re-estimated for multiple bootstrap samples, to take into account parameter uncertainty during the imputation stage. Whereas 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 too small a number of clusters during the Gibbs sampling, leading to an underfitting model yielding invalid imputations. In Chapter 3, we proposed an alternative approach which combined the strengths of the two methods; that is, we used the Bayesian standard latent class model as an imputation model. We showed how model selection can be performed prior to the imputation step, using a single run of the Gibbs sampler and, moreover, showed 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 indicated that with a proper setting of the prior distributions, the Bayesian latent class model yields valid imputations and outperforms competing methods.

With Chapter 4, we proposed using a Bayesian multilevel latent class model for the MI of nested categorical data. Unlike recently developed methods that can only pick-up associations between pairs of variables, the multilevel mixture model we proposed is flexible enough to automatically deal with complex interactions in the joint distribution of the variables to be estimated. After formally introducing the model and showing how it can be implemented, we carried out a simulation study and a real-data study in order to assess its performance, and compared it with the commonly used listwise deletion and an available R-routine. Results indicated that the Bayesian Multilevel latent class model is able to recover unbiased parameter estimates of the analysis models considered in our studies, as well as to correctly reflect the uncertainty due to missing data, outperforming the other methods.

Chapter 5 introduced an MI tool for longitudinal studies: MI using the Bayesian mixture Latent Markov models. Besides retaining the benefits of latent class models, i.e., respecting the (categorical) measurement scale of the variables and preserving possibly complex relationships between variables within a measurement occasion, the Markov dependence structure of the Bayesian mixture Latent Markov model allows capturing lagged dependencies between adjacent time points, while the time-constant mixture structure allows capturing dependencies across all time points, as well as retrieving associations between time-varying and time-constant variables. The performance of the BMLM model for MI was evaluated by means of two simulation studies and a real-data experiment, in which it was compared with complete case analysis and MI by chained equations. Results showed good performance of the proposed method in retrieving the parameters of the analysis model. In contrast, competing methods could provide correct estimates only for some aspects of the data.

Several extensions of the models proposed in this dissertation are possible. The main one concerns the measurement scale of the variables assumed by the models: while in social and behavioral sciences categorical scales are frequently used in questionnaires, variables measured with mixed types of scales (i.e., continuous and categorical) can be frequently found in different contexts. The mixture models described above can be easily modified to accommodate for both kinds of measurement scales (e.g., by assuming mixtures of Normal and Multinomial distributions), but their performance must be evaluated in future research. Multilevel latent class models can also be adjusted to account for more than two levels in the hierarchy, while mixture latent Markov models can be extended to include second or higher-level orders of lagged relationships.

ACKNOWLEDGMENTS

During the years of my Ph.D. I have grown a lot both as a person and professionally. There are several people I would like to thank that contributed to this.

First of all, I would like to thank my promoter Jeroen and co-promoter Katrijn. Jeroen, you did not only gave me the chance to become a Ph.D. student in The Netherlands, but you also helped me to better grasp the functioning of latent class models and categorical data analysis during my research. Furthermore, you provided me useful guidelines and insights making things look easy and funny (although they were not); in this way, you helped me to approach scientific research with the right spirit. Katrijn, you helped me with the computational aspects of my dissertation, as well as with the writing style of my papers and letters for the reviewers. Thank you also for giving me the chance to continue my academic career in Tilburg. Jeroen and Katrijn, I want to thank you also for the patience you showed when reading and improving my writings. I would also like to thank Maurits Kaptein (il Capitano!), who was my co-promoter during the first year of my Ph.D., and with whom I wrote the second chapter of the dissertation.

Next, I am grateful to the members of my committee defense, who immediately accepted our invitation. Furthermore, I would also like to help them for the positive comments provided concerning my dissertation.

All my MTO colleagues have been a very nice company during the last four years, and I am thankful to all of them for the great time spent together. Guys, academic life with you was an amazing experience! In particular, I would like to thank all my office roomies: Zsuzsa and Margot, you helped me a lot with all kinds of bureaucracy issues I came across in the beginning of my experience, and treated me almost like a small brother; Eva, Laura and Sara, we had great times together, and I will miss hearing your Dutch and not understanding anything of it in the office. Thank you to all members of the VICI and Bayesian meetings, who always provided great comments and feedback to improve my papers. A special mention goes also to all my colleagues that came with me to the ISBA conference in Cagliari (Maurits, Joris, Reza, Dino, Florian): I will never forget that amazing trip together! Among the MTOers, I would like to express my gratitude also to John, who helped me with my first teaching experiences, and to Marieke, who has always been helpful any time I needed.

I would have never worked so gladly on my project without good company in life. Thus, I want to thank my girlfriend Paula: you have always been by my side and always done everything to make my life feel (and actually become) better.

Thanks also to my crew of friends in Tilburg; in the moments of need, they were always ready to support me, and in the moments of leisure, they never missed the chance to make me laugh. In particular, I would like to mention Tomás, Jorg, Sarah, Pablo, Tulin, Nicolooó (my brother), Marti, Rodi, Cana, Minoo, Alex & Alex, Malte, Brains, and Diego, as well as my housemates Zhivko and Dre. Buddies, you are like a second family for me. Of course some people are missing and I might have forgotten others. To apologize, I send this 'thank you' to them, for the good times and the nice journey together!

In addition, I want to thank my family - my mom, my dad, my brothers - for the support they gave me at any point of my life, whatever choice and path I decided to undertake. I am specially grateful also to the professors of the Statistical Sciences department in Padova (in particular Bruno Scarpa, Lorenzo Bernardi, Francesca Bassi). Most likely, I would not be here without their important teachings. My former colleagues Daniele Durante and Sabrina Vettori have always been an important reference to me, and constantly (still now!) inspired my understanding of statistics. Lastly, I want to thank my friends in Italy. Despite the distance, they have always made me feel close to them, as I never had left Italy!