9

MULTIPLE IMPUTATION OF INCOMPLETE CATEGORICAL DATA USING LATENT CLASS ANALYSIS

*Jeroen K. Vermunt**
Joost R. van Ginkel†
*L. Andries van der Ark**
*Klaas Sijtsma**

We propose using latent class analysis as an alternative to log-linear analysis for the multiple imputation of incomplete categorical data. Similar to log-linear models, latent class models can be used to describe complex association structures between the variables used in the imputation model. However, unlike log-linear models, latent class models can be used to build large imputation models containing more than a few categorical variables. To obtain imputations reflecting uncertainty about the unknown model parameters, we use a nonparametric bootstrap procedure as an alternative to the more common full Bayesian approach. The proposed multiple imputation method, which is implemented in Latent GOLD software for latent class analysis, is illustrated with two examples. In a simulated data example, we compare the new method to well-established methods such as maximum likelihood

We would like to thank Paul Allison and Jay Magidson, as well as the editor and the three anonymous reviewers, for their comments, which very much helped to improve this article. We would also like to thank Greg Richards for providing the data of the ATLAS Cultural Tourism Research Project, 2003.

*Tilburg University
†Leiden University

estimation with incomplete data and multiple imputation using a saturated log-linear model. This example shows that the proposed method yields unbiased parameter estimates and standard errors. The second example concerns an application using a typical social sciences data set. It contains 79 variables that are all included in the imputation model. The proposed method is especially useful for such large data sets because standard methods for dealing with missing data in categorical variables break down when the number of variables is so large.

1. INTRODUCTION

Multiple imputation (MI) has become a widely accepted method for dealing with missing data problems (Rubin 1987:2–4). One of its attractive features is that it allows the handling of the missing data problem prior to the actual data analysis; that is, once the missing values are replaced by imputed values, the statistical analyses of interest can be performed using standard techniques. Another attractive feature of MI is that, contrary to single imputation, the multiply imputed versions of a data set reflect uncertainty about the imputed values, which is a requirement for obtaining unbiased standard errors in statistical analyses.

MI requires the specification of an imputation model, the exact choice of which will typically depend on the scale types of the variables in the data set. For (approximately) continuous variables, the most widely used imputation model is the multivariate normal model (Schafer 1997), which is available in SAS PROC MI (Yuan 2000) and in the missing-data library of S-plus (2006), as well as in stand-alone programs such as NORM (Schafer 1999) and AMELIA (King et al. 2001; Honaker, King, and Blackwell 2007). Graham and Schafer (1999) showed that MI under the multivariate normal model is rather robust to violations of normality. The most appropriate imputation model for categorical variables is the log-linear model (Schafer 1997), which is also implemented in the missing-data library of S-plus (2006). For data sets containing both categorical and continuous variables, Schafer (1997) proposed imputation using the general location model, a combination of a log-linear and a multivariate normal model implemented in the missing-data library of S-plus.

MI based on log-linear modeling provides an elegant and sound solution for many missing-data problems concerning categorical

variables. This was confirmed in simulation studies by Ezzati-Rice et al. (1995), Schafer et al. (1996), and Schafer (1997), who showed that log-linear imputation yields unbiased statistical inference, and it is robust against departures from the assumed imputation model. The main limitation of MI under the log-linear model is, however, that it can be applied only when the number of variables used in the imputation model is small—that is, only when we are able to set up and process the full multi-way cross-tabulation required for the log-linear analysis. Whereas social science data sets with 100 variables or more are very common, it is impossible to estimate a log-linear model for say a frequency table cross-classifying 100 trichotomous variables: The resulting table with 3^{100} (=5.15378e47) cell entries is much too large to be stored and processed. Note that the necessity to process each cell in the maximum likelihood estimation of log-linear models holds even if the specified model is very restricted—for example, if the model contains only two- and three-way association terms. An exception is the situation in which the log-linear model is collapsible (Agresti 2002), but it is unlikely that one will use such a model for imputation.

A first possible solution to the problem of a limited number of variables associated with the log-linear approach is to ignore the categorical nature of the variables and use an imputation model for continuous data instead, where discrete imputed values may be obtained by rounding the non-integer imputed values to the nearest feasible integer. Van Ginkel, Van der Ark, and Sijtsma (2007a; 2007b) found that MI under the multivariate normal model with rounding produces reliable results in discrete (ordinal) psychological test data—for example, in the estimation of Cronbach's alpha. Other authors, however, showed that rounding continuous imputations to the nearest admissible integer values may lead to serious bias (Allison 2005; Horton, Lipsitz, and Parzen 2003), especially if the variables concerned are used as independent variables in a regression analysis. This was confirmed by Bernaards, Belin, and Schafer (2007), who showed that the bias may be reduced by using a more sophisticated (adaptive) rounding procedure. Despite the fact that this approach may sometimes work well with dichotomous and ordinal categorical variables, it is clearly much more problematic when used with nominal variables.

A second possible solution is to use hot-deck imputation (Rubin 1987:9) rather than a statistical imputation model. This nonparametric imputation method involves a search for complete cases that have

(almost) the same values on the observed variables as the case with missing values, and then imputes the missing values of the latter by drawing from the empirical distribution defined by the former. Hot-deck imputation, which is available in the SOLAS program (2001), can be used for data sets containing large numbers of categorical variables. Whereas the standard hot-deck does not yield proper imputation, a variant called approximate Bayesian bootstrap (Rubin and Schenker 1986) does. However, Little and Rubin (2002:69) indicated the following about the nearest neighbor hot-deck procedure: "Since imputed values are relatively complex functions of the responding items, quasi-randomization properties of estimates derived from such matching procedures remain largely unexplored." This means that it is difficult to demonstrate that estimates will be unbiased under the missing at random (MAR) assumption. A simulation study by Schafer and Graham (2002) showed that hot-deck imputation may produce biased results irrespective of the missing data mechanism.

A third possible solution is to use one of the recently proposed sequential regression imputation methods, which include the MICE and ICE methods (Van Buuren and Oudshoorn 2000; Raghunathan et al. 2001; Van Buuren et al. 2006). Rather than specifying a model for the joint distribution of the variables involved in the imputation, the imputation model consists of a series of models for the univariate conditional distributions of the variables with missing values. For categorical variables, this will typically be a series of logistic regression equations. We found that for large numbers of variables the specification of the series of imputation models is rather difficult and time-consuming. Especially problematic is that, unlike log-linear imputation, sequential imputation does not pick up higher-order interactions, unless these are included explicitly in the imputation model. This means that it is likely that one misses important interactions that may seriously bias subsequent analyses. Also, unlike the log-linear approach, sequential regression imputation methods lack a strong statistical underpinning; that is, there is no guarantee that iterations will converge to the posterior distribution of the missing values.

Alternatively, we propose using the latent class (LC) model as an imputation model for categorical data. It is a statistically sound categorical data method that resolves the most important limitation of the log-linear approach; that is, that it can be applied to data sets containing more than a few variables. An LC model can be viewed as a mixture

model of independent multinomial distributions. Mixture models have been shown to be very flexible tools for density estimation because they can be used to approximate any type of distribution by choosing the number of mixture components (latent classes) sufficiently large (e.g., McLachlan and Peel 2000:11–14). Using the LC model as an imputation model resembles the use of LC models by Vermunt and Magidson (2003), who proposed using the LC model as a prediction or classification tool. As is explained in more detail below, the local independence assumption (Lazarsfeld 1950a, 1950b; Goodman, 1974) makes it possible to obtain maximum likelihood estimates of the parameters of LC models for large numbers of categorical variables with missing values.

Multiple imputations should reflect uncertainty about not only the missing values but also the unknown parameters of the imputation model. This parameter uncertainty is typically dealt with by using a full Bayesian approach: random imputations are based on random draws of the parameters from their posterior distribution (Rubin 1987; Schafer 1997). An alternative that allows staying within a frequentist framework is to use the nonparametric bootstrap, as is done in the Amelia II software (King et al. 2001; Honaker et al. 2007). This is also the approach used in this article and implemented in the syntax version of Latent GOLD program (Vermunt and Magidson 2008).

The remainder of this article will focus on the following: First, we introduce the basic principles of MI. Second, we discuss MI using LC analysis (from here on abbreviated as LC MI) and also discuss issues such as dealing with parameter uncertainty, model selection, and model identifiability. Third, a constructed example is presented comparing LC MI with complete-case analysis, maximum likelihood estimation with incomplete data and log-linear MI, as well as demonstrating the validity of LC MI. Fourth, a large real-data example is discussed that illustrates LC MI for a situation in which log-linear MI is no longer feasible. We offer conclusions about the proposed LC MI approach and discuss possible extensions of this tool.

2. THE BASIC PRINCIPLE OF MULTIPLE IMPUTATION

This section describes the basic principles of MI. Let us first introduce the relevant notation. Let $\mathbf{Y}$ denote the $N \times J$ data matrix of interest with entries y_{ij}, where N is the number of cases and J the number of

variables, with indices i and j such that $1 \leq i \leq N$ and $1 \leq j \leq J$. In the presence of missing data, the data matrix has an observed part and a missing part. These two parts are denoted as $\mathbf{Y}_{obs}$ and $\mathbf{Y}_{mis}$, respectively, where $\mathbf{Y} = (\mathbf{Y}_{obs}, \mathbf{Y}_{mis})$. The unknown parameters that govern $\mathbf{Y}$ are collected in vector $\boldsymbol{\theta}$. Let $\mathbf{R}$ be a response-indicator matrix with entries r_{ij}, where $r_{ij} = 0$ if the value of y_{ij} is missing and 1 otherwise. The unknown parameters that govern $\mathbf{R}$ are collected in vector $\boldsymbol{\xi}$.

The basic idea of multiple imputation is to construct multiple, say M, complete data sets by random imputation of the missing values in $\mathbf{Y}_{mis}$. The researcher interested in a particular analysis model—say a linear regression model—can estimate the analysis model with each of these M complete data sets using standard complete data methods. The M results can be combined into a single set of estimates and standard errors reflecting the uncertainty about the imputed missing values (Rubin 1987:76–79).

What is needed to construct multiply imputed data sets is an imputation model: a model for the joint distribution of the response indicators and the survey variables in the data set, which is denoted as $P(\mathbf{R},\mathbf{Y}_{obs},\mathbf{Y}_{mis};\boldsymbol{\theta},\boldsymbol{\xi})$. When defining a model for $P(\mathbf{R},\mathbf{Y}_{obs},\mathbf{Y}_{mis};\boldsymbol{\theta},\boldsymbol{\xi})$, one typically separates the model for $\mathbf{Y}$ from the model for the missing data mechanism. This is achieved by the following decomposition:

$$P(\mathbf{R}, \mathbf{Y}_{obs}, \mathbf{Y}_{mis}; \boldsymbol{\theta}, \boldsymbol{\xi}) = P(\mathbf{Y}_{obs}, \mathbf{Y}_{mis}; \boldsymbol{\theta}) P(\mathbf{R}|\mathbf{Y}_{obs}, \mathbf{Y}_{mis}; \boldsymbol{\xi}), \quad (1)$$

where $P(\mathbf{Y}_{obs},\mathbf{Y}_{mis};\boldsymbol{\theta})$ is the marginal distribution of the survey variables and $P(\mathbf{R}|\mathbf{Y}_{obs},\mathbf{Y}_{mis};\boldsymbol{\xi})$ is the conditional distribution of the response indicators given the survey variables. It can easily be seen that the decomposition in equation (1) transforms the definition of a model for $P(\mathbf{R},\mathbf{Y}_{obs},\mathbf{Y}_{mis};\boldsymbol{\theta},\boldsymbol{\xi})$ into the definition of two submodels, one for $P(\mathbf{Y}_{obs},\mathbf{Y}_{mis};\boldsymbol{\theta})$ and one for $P(\mathbf{R}|\ \mathbf{Y}_{obs},\mathbf{Y}_{mis};\boldsymbol{\xi})$. Specification and estimation of the imputation model can be simplified further by making the additional assumption that the probability of having a certain pattern of missing values is independent of the variables with missing values conditionally on the observed; that is,

$$P(\mathbf{R}|\mathbf{Y}_{obs}, \mathbf{Y}_{mis}; \boldsymbol{\theta}) = P(\mathbf{R}|\mathbf{Y}_{obs}; \boldsymbol{\theta}). \quad (2)$$

If equation (2) holds, then the missing data are said to be missing at random (MAR; Rubin 1976; Little and Rubin 2002:12). When in addition

the submodels for $P(\mathbf{Y}_{obs},\mathbf{Y}_{mis};\boldsymbol{\theta})$ and $P(\mathbf{R}|\mathbf{Y}_{obs},\mathbf{Y}_{mis};\boldsymbol{\xi})$ do not have common parameters, the submodel for the missing data mechanism can be ignored when estimating the model for $P(\mathbf{Y}_{obs},\mathbf{Y}_{mis};\boldsymbol{\theta})$. The MAR assumption will be violated either when there are direct effects of variables with missing values on the response indicators after controlling for $\mathbf{Y}_{obs}$, or when there are variables that are not in the imputation model affecting both $\mathbf{Y}_{mis}$ and $\mathbf{R}$. In these cases, the missingness mechanism is not MAR (i.e., it is NMAR), and the validity of the results from likelihood-based methods cannot be guaranteed, unless the correct NMAR model for the missingness mechanism is specified.

It may be noted that the MAR assumption becomes more plausible when a larger number of variables are included in the imputation model (Schafer 1997:28). If the set $\mathbf{Y}_{obs}$ becomes larger, it becomes less likely that dependencies remain between $\mathbf{R}$ and $\mathbf{Y}_{mis}$ after conditioning on $\mathbf{Y}_{obs}$ (Schafer 1997:28). The main advantage of imputation methods compared to parameter estimation with incomplete data is that we can put more effort into building a model that is in agreement with the MAR assumption. Incomplete-data likelihood methods usually make use of a smaller set of variables, typically only the variables needed in the analysis model of interest.

To minimize the risk of bias, Schafer (1997:143) advocated using an imputation model that is as general as possible—for example, an unrestricted multivariate normal model or a saturated model. At worst, standard errors of the parameters derived from MI may slightly increase when the imputation model contains associations that can be attributed to sampling fluctuations (Schafer 1997:140–44). On the other hand, if the imputation model is too restrictive, results may be biased because the MAR assumption is violated (Schafer 1997:142–43). For this reason, Schafer recommended generating imputed values that are as much as possible in accordance with the observed data, so that the imputed values behave "neutral" in the subsequent statistical analyses.

The actual imputation of the missing values involves generating random draws from the distribution $P(\mathbf{Y}_{mis}|\mathbf{Y}_{obs})$, which is defined as follows (Rubin 1987; Schafer 1997):

$$\begin{aligned} P(\mathbf{Y}_{mis}|\mathbf{Y}_{obs}) &= \int P(\mathbf{Y}_{mis}|\mathbf{Y}_{obs};\boldsymbol{\theta})\, P(\boldsymbol{\theta}|\mathbf{Y}_{obs})\mathrm{d}\boldsymbol{\theta} \\ &= \int \frac{P(\mathbf{Y}_{obs},\, \mathbf{Y}_{mis};\boldsymbol{\theta})}{P(\mathbf{Y}_{obs};\, \boldsymbol{\theta})}\, P(\boldsymbol{\theta}|\mathbf{Y}_{obs})\mathrm{d}\boldsymbol{\theta}. \end{aligned} \tag{3}$$

The most popular way to perform this sampling is by Bayesian Markov chain Monte Carlo methods (Schafer 1997:105; Tanner and Wong 1987), which involves a two-step procedure. In the first step, values of $\boldsymbol{\theta}$ are drawn from the posterior distribution of the parameters $P(\boldsymbol{\theta}|\mathbf{Y}_{obs})$. These values of $\boldsymbol{\theta}$are used in the actual imputation step to obtain draws from $P(\mathbf{Y}_{mis} \mid \mathbf{Y}_{obs};\boldsymbol{\theta})$. Since each imputed data set is based on new draws from $P(\boldsymbol{\theta} \mid \mathbf{Y}_{obs})$, the multiple imputations will reflect not only uncertainty about $\mathbf{Y}_{mis}$ but also uncertainty about the parameters of the imputation model, yielding what Rubin referred to as proper imputations.

Schafer (1997:289–331) proposed using log-linear analysis as an imputation tool for categorical data. The strong points of log-linear models are that they yield an accurate description of $P(\mathbf{Y}_{obs},\mathbf{Y}_{mis};\boldsymbol{\theta})$ and that they can easily be estimated with incomplete data. A serious limitation of the log-linear modeling approach is, however, that it can only be used with small numbers of variables, whereas imputation should preferably be based on large sets of variables. To overcome this drawback, we propose using LC analysis as a tool for the imputation of incomplete categorical data. The LC model is a categorical data model that can be used to describe the relationships between the survey variables as accurately as needed. Parameter estimation of LC models does not break down when the number of variables is large. Moreover, the model parameters can be easily estimated in the presence of missing data.

3. MULTIPLE IMPUTATION UNDER A LATENT CLASS MODEL

3.1. *Latent Class Analysis with Incomplete Data*

This section deals with MI using latent class models. We first introduce latent class models for incomplete categorical data, where special attention is paid to issues that are relevant when using these models in the context of MI. Then we discuss model selection and the exact implementation of the LC-based imputation procedure.

Let $\mathbf{y}_i$ denote a vector containing the responses of person i on J categorical variables, x_i a discrete latent variable, K the number of categories of x_i or, equivalently, the number of latent classes, and k a

particular latent class ($k = 1, \ldots, K$). The model we propose for imputation is an unrestricted LC model, in which $P(\mathbf{y}_i; \boldsymbol{\theta})$, the joint probability density of $\mathbf{y}_i$, is assumed to have the following well-known form (Lazarsfeld 1950a, 1950b; Goodman 1974; Vermunt and Magidson 2004):

$$\begin{aligned} P(\mathbf{y}_i; \boldsymbol{\theta}) &= \sum_{k=1}^{K} P(x_i = k; \boldsymbol{\theta}_x) P(\mathbf{y}_i | x_i = k; \boldsymbol{\theta}_y) \\ &= \sum_{k=1}^{K} P(x_i = k; \boldsymbol{\theta}_x) \prod_{j=1}^{J} P(y_{ij} | x_i = k; \boldsymbol{\theta}_{y_j}), \end{aligned} \tag{4}$$

where $\boldsymbol{\theta} = (\boldsymbol{\theta}_x, \boldsymbol{\theta}_y)$ or $\boldsymbol{\theta} = (\boldsymbol{\theta}_x, \boldsymbol{\theta}_{y_1}, \ldots, \boldsymbol{\theta}_{y_j}, \ldots, \boldsymbol{\theta}_{y_J})$. The indices in $\boldsymbol{\theta}_x$, $\boldsymbol{\theta}_y$, and $\boldsymbol{\theta}_{y_j}$ indicate to which set of multinomial probabilities the unknown parameters concerned belong. Equation (4) shows the two basic assumptions of the latent-class model:

1. The density $P(\mathbf{y}_i; \boldsymbol{\theta})$ is a mixture—or weighted average—of class-specific densities $P(\mathbf{y}_i \mid x_i = k; \boldsymbol{\theta}_y)$, where the unconditional latent class proportions $P(x_i = k; \boldsymbol{\theta}_x)$ serve as weights.
2. Responses are independent within latent classes, such that the joint conditional density $P(\mathbf{y}_i \mid x_i = k; \boldsymbol{\theta}_y)$ equals the product of the J univariate densities $P(y_{ij} \mid x_i = k; \boldsymbol{\theta}_{y_j})$. Note that $P(y_{ij} \mid x_i = k; \boldsymbol{\theta}_{y_j})$ is the probability that person i provides response y_{ij} to variable j conditional on membership of class k. This is generally referred to as the local independence assumption.

By choosing the number of latent classes sufficiently large, like any type of mixture model, an LC model will accurately pick up the first, second, and higher-order observed moments of the J response variables (McLachlan and Peel 2000:11–14). In the context of categorical variables, these moments are the univariate distributions, bivariate associations, and the higher-order interactions.

It is important to emphasize that we are using the LC model not as a clustering or scaling tool but as a tool for density estimation—that is, as a practical tool to obtain a sufficient exact representation of the true $P(\mathbf{y}_i; \boldsymbol{\theta})$, even for large J. This has the following implications:

1. Contrary to typical LC applications, there is no need to find interpretable latent classes or clusters. In fact, there is no need to interpret the parameters of the LC imputation model at all. This is not specific for LC analysis imputation. Also when using a multivariate normal or a log-linear imputation model, the parameters will not be interpreted.
2. Overfitting the data is less of a problem than underfitting; that is, picking up certain random fluctuations that are sample specific is less problematic than ignoring important association or interactions between the variables in the imputation model. Note that overfitting an LC model is similar to using a log-linear imputation model that includes nonsignificant parameters. This is likely to occur when using a saturated log-linear model. Underfitting an LC model is comparable to using a nonsaturated log-linear model in which important higher-order interactions are omitted.
3. It is well-known that LC models may be unidentified when the number of classes is large compared with the number of observed variables (for example, see Goodman 1974). Unidentifiability means that different values of $\boldsymbol{\theta}$ yield the same $P(\mathbf{y}_i; \boldsymbol{\theta})$, which makes the interpretation of the $\boldsymbol{\theta}$ parameters problematic. However, in the context of imputation, this is not a problem since we are interested only in $P(\mathbf{y}_i; \boldsymbol{\theta})$, which is uniquely defined even if the $\boldsymbol{\theta}$ parameters are not.
4. For large K, a solution may be obtained that is a local instead of a global maximum of the incomplete data log-likelihood function. Even if we increase the number of start sets to say 100—as we did in our analysis with the automated starting values procedure of Latent GOLD—there is no guarantee that we will find the global maximum likelihood solution. Whereas this is problematic if we wish to interpret the model parameters, in the context of MI this does not seem to be a problem, especially because a local maximum will typically give a representation of $P(\mathbf{y}_i; \boldsymbol{\theta})$ that is nearly as good as the global maximum.

We will use a simulated data example to illustrate these issues below.

Equation (4) describes the LC model, neglecting the missing data. However, for parameter estimation using maximum likelihood, as well as for the derivation of $P(\mathbf{Y}_{mis} | \mathbf{Y}_{obs}; \boldsymbol{\theta})$ needed for the actual imputation (see equation 3), the LC model must be expressed as a model for the observed data density $P(\mathbf{y}_{i,obs}; \boldsymbol{\theta})$; that is,

$$P(\mathbf{y}_{i,obs};\boldsymbol{\theta}) = \sum_{k=1}^{K} P(x_i = k;\boldsymbol{\theta}_x)\, P(\mathbf{y}_{i,obs}|x_i = k;\boldsymbol{\theta}_y)$$
$$= \sum_{k=1}^{K} P(x_i = k;\boldsymbol{\theta}_x) \prod_{j=1}^{J} \left[P(y_{ij}|x_i = k;\boldsymbol{\theta}_{y_j})\right]^{r_{ij}}, \qquad (5)$$

where again $r_{ij} = 0$ if the value of y_{ij} is missing and 1 otherwise. Note that for a respondent with a missing value on variable j (i.e., $r_{ij} = 0$), $[P(y_{ij}|x_i = k;\boldsymbol{\theta}_{y_j})]^{r_{ij}} = 1$, and thus the corresponding term cancels from equation (5). This illustrates how missing values are dealt with in the estimation of an LC model: Case i contributes to the estimation of the unknown model parameters only for the variables that are observed.

Maximum likelihood (ML) estimates of the parameters of an LC model can be obtained by maximizing the sum of the log of equation (5) across all cases—for example, by means of the EM algorithm (Goodman 1974; Dempster, Laird, and Rubin 1977). Because of the local independence assumption, the problem is collapsible, which implies that the M step of the EM algorithm involves processing J two-way x_i by y_{ij} cross-classifications. Thus, even for large J, ML estimation remains feasible. Except for very specific situations, log-linear analysis, on the contrary, requires processing the full J dimensional table.

3.2. *Model Selection*

As was already mentioned, mixture models can approximate a wide variety of distributions (McLachlan and Peel 2000:11–14). For LC-based MI, this means that the imputation model will accurately approximate the distribution of $\mathbf{y}_i$ by choosing K sufficiently large. Following Schafer's (1997:140–44) advise to use an imputation model that describes the data as accurately as possible, we should thus use a large number of latent classes.

There is one particular question of interest: What is a sufficiently large value for K? We propose using the Bayesian information criterion (BIC; Schwarz 1978), Akaike's information criterion (AIC; Akaike 1974), and a variant of AIC called AIC3 (Bozdogan 1993; Andrews and Currim 2003), which are typically used for model selection in LC analysis with sparse tables. These measures have in common that they combine model fit (log-likelihood value, LL) and parsimony (number of

parameters, *Npar*) into a single value. The model with the lowest BIC, AIC, or AIC3 value is the preferred one. The three information criteria are defined as follows:

$$BIC = -2LL + \log(N) \times Npar, \tag{6}$$

$$AIC = -2LL + 2 \times Npar, \tag{7}$$

$$AIC3 = -2LL + 3 \times Npar. \tag{8}$$

Equations (6), (7), and (8) show that the three criteria differ only with respect to the weight attributed to parsimony. Because the log of the sample size—log(N)—is usually larger than 3, BIC tends to select a model with fewer latent classes than AIC and AIC3. Simulation studies by Lin and Dayton (1997), Andrews and Currim (2003), and Dias (2004) showed that BIC tends to underestimate the number of classes, whereas AIC tends to select a model with too many classes. Andrews and Currim (2003) and Dias (2004) also showed that for selecting K in LC models, AIC3 provides a good compromise between BIC and AIC.

3.3. *Imputation Procedure*

Multiple imputation involves obtaining M draws from $P(\mathbf{y}_{i,mis} \mid \mathbf{y}_{i,obs})$ (see equation 3). This requires obtaining draws from $P(\boldsymbol{\theta} \mid \mathbf{y}_{i,obs})$ and subsequently from $P(\mathbf{y}_{i,mis} \mid \mathbf{y}_{i,obs}; \boldsymbol{\theta})$. As is also done in the AMELIA II software (King et al. 2001; Honaker et al. 2007), we propose obtaining M sets of parameters $\boldsymbol{\theta}$ using a nonparametric bootstrap procedure. (For a general introduction in the bootstrap procedure and an application in LC analysis see Efron and Tibshirani [1993] and Dias and Vermunt [forthcoming], respectively.) First, M nonparametric bootstrap samples from $\mathbf{Y}$ are obtained and denoted as $\mathbf{Y}^*_1, \ldots, \mathbf{Y}^*_m, \ldots, \mathbf{Y}^*_M$. Second, for each bootstrap sample an LC model is estimated resulting in M sets of parameters $\boldsymbol{\theta}^*_1, \ldots, \boldsymbol{\theta}^*_m, \ldots, \boldsymbol{\theta}^*_M$. For imputed data set m, we sample from $P(\mathbf{y}_{i,mis} \mid \mathbf{y}_{i,obs}; \boldsymbol{\theta} = \boldsymbol{\theta}^*_m)$ for $m = 1, \ldots M$. In this way, we take the uncertainty of the parameter estimates into account.

In an LC model, the conditional distribution $P(\mathbf{y}_{i,mis} \mid \mathbf{y}_{i,obs}; \boldsymbol{\theta})$ can be written as

$$P(\mathbf{y}_{i,mis}|\mathbf{y}_{i,obs};\boldsymbol{\theta}) = \sum_{k=1}^{K} P(x_i = k, \mathbf{y}_{i,mis}|\mathbf{y}_{i,obs};\boldsymbol{\theta})$$
$$= \sum_{k=1}^{K} P(x_i = k|\mathbf{y}_{i,obs};\boldsymbol{\theta})\, P(\mathbf{y}_{i,mis}|x_i = k;\boldsymbol{\theta}_y)$$
$$= \sum_{k=1}^{K} P(x_i = k|\mathbf{y}_{i,obs};\boldsymbol{\theta}) \prod_{j=1}^{J} \left[P(y_{ij}|x_i = k;\boldsymbol{\theta}_{y_j}\right]^{1-r_{ij}}. \quad (9)$$

In the first row of equation (9), we introduce the discrete latent variable x_i. The second row is obtained using the local independence assumption between $\mathbf{y}_{i,mis}$ and $\mathbf{y}_{i,obs}$ given x_i. The last row uses again the local dependence assumption, but now among the variables with missing values. It should be noted that $P(x_i = k \mid \mathbf{y}_{i,obs};\boldsymbol{\theta})$ is a subject's latent classification probability for class k, which is part of the standard output of an LC analysis. These probabilities, which are also referred to as posterior class membership probabilities, can be obtained as follows:

$$P(x_i = k|\mathbf{y}_{i,obs};\boldsymbol{\theta}) = \frac{P(x_i = k;\boldsymbol{\theta}_y)\, P(\mathbf{y}_{i,obs}|x_i = k;\boldsymbol{\theta}_y)}{P(\mathbf{y}_{i,obs};\boldsymbol{\theta})}, \quad (10)$$

The terms in equation (10) were defined in equation (5).

Equation (9) suggests how to sample from $P(\mathbf{y}_{i,mis} \mid \mathbf{y}_{i,obs};\boldsymbol{\theta})$ in LC MI; that is, in a first step, assigning a person randomly to one of the K latent classes using $P(x_i = k \mid \mathbf{y}_{i,obs};\boldsymbol{\theta})$ (see Goodman 2007) and subsequently, in a second step, sampling $\mathbf{y}_{i,mis}$ conditionally on the assigned class using $P(\mathbf{y}_{i,mis} \mid x_i = k;\boldsymbol{\theta})$. This second step can be performed separately for each variable with a missing value; that is, by means of independent draws from univariate multinomial distributions with probabilities $P(y_{ij} \mid x_i = k;\boldsymbol{\theta}_{y_j})$.

The possibility to perform the second step for each variable separately shows that the LC imputation method is applicable to large data sets and has the additional advantage that it can be implemented using standard software for LC analysis with missing data, such as Latent GOLD 4.0 (Vermunt and Magidson 2005), LEM (Vermunt 1997), Mplus (Muthén and Muthén 2006), and the R library poLCA (Linzer and Lewis 2007). Each of these packages provides both the latent classification probabilities $P(x_i = k \mid \mathbf{y}_{i,obs};\boldsymbol{\theta})$ and the response probabilities $P(y_{ij} \mid x_i = k;\boldsymbol{\theta}_{y_j})$ as output. Using these probabilities one can

draw random values for the missing data with pseudo random number generators that are readily available in general statistical packages. The new syntax module of the Latent GOLD 4.5 implements the procedure described above, including the nonparametric bootstrap, fully automatically. We provide more details about the implementation of LC MI in Latent GOLD 4.5 in the appendix.

4. EXAMPLES

In this section, LC MI is illustrated and compared with other methods using two examples. The first uses a simulated data set, which makes it possible to compare the obtained results with the known truth. The second example uses a large real data set with many categorical variables with missing values.

4.1. *A Simulated Data Example*

In the introduction we already described various simulation studies showing that (proper) MI—say, under a correctly specified log-linear model—is able to yield unbiased parameter estimates and unbiased asymptotic standard errors for the analysis model of interest (e.g., Ezzati-Rice et al. 1995; Schafer et al. 1996; Schafer 1997). Also, the effect of sample size and the effect of different types of violations of the MAR assumption have been studied. Therefore, rather than performing an extended simulation study in which the same issues are investigated for LC MI, we concentrate on those aspects that are specific for the LC MI method proposed in this paper.

The main question is this: How does the proposed LC MI method compare to ML estimation with incomplete data and MI using a (saturated) log-linear model? There are also two more specific questions: What happens if we select too few latent classes? What happens if we select too many latent classes?

We simulated one large data set ($N = 10{,}000$) from a population model with the following characteristics:

- Six dichotomous variables: y_1 to y_6; y_1 to y_5 were the independent variables and y_6 was the dependent variable.

- For the relationship among the independent variables, we assumed a log-linear model with, under dummy coding, one-variable terms equal to -2.0 and two-variable associations equal to 1.0; that is,

$$\log P(y_1, y_2, y_3, y_4, y_5) \propto \sum_{s=1}^{5} -2.0 y_s + \sum_{s=1}^{5} \sum_{t=s+1}^{5} 1.0 y_s y_t.$$

  This is equivalent to a model defined under effect coding with all one-variable terms equal to 0.0 and all two-variable terms equal to .25.
- For the dependent variable, we assumed a logit model containing main effects of the independent variables as well as a two-way interaction between y_2 and y_3. Using dummy coding, the population logit equation was defined as follows:

$$\text{logit } P(y_6) = -3.0 + 1.0y_1 + 2.0y_2 + 2.0y_3 + 1.0y_4 + 1.0y_5 - 2.0y_2 y_3.$$

- Both y_1 and y_2 were assumed to have MAR missing values, where the missingness on y_1 depended on y_3 and y_4—missingness probabilities were .1, .4, .4, and .7, respectively, for the four possible combination of y_3 and y_4—and the missingness on y_2 depended on y_5 and y_6—with probabilities equal to .7, .4, .4, and .1, respectively. In total almost 70 percent of 10,000 cases had at least one missing value.

Note that we used a large sample because we were not interested in assessing the effect of sampling fluctuations. We assumed a MAR model with a large proportion of missing values in the predictor variables because this is the kind of situation in which LC MI work should work. The key element in the population model specification is the inclusion of a large interaction term in the regression model for y_6, which in fact implies that there is a three-variable association between y_2, y_3, and y_6. While such an association is automatically picked up by a saturated log-linear model, it should be investigated whether an LC model picks it up as well.

Table 1 shows the log-likelihood, BIC, AIC, and AIC3 values for 1- to 10-class models and for the saturated log-linear model estimated with the simulated data set. Based on BIC, we should select the 3-class model and based on either the AIC or AIC3 criterion one should select the 6-class model. This difference in suggested number of latent classes is larger than usually encountered in standard applications of LC analysis,

TABLE 1
Log-Likelihood, BIC, AIC, and AIC3 Values for the Latent Class Models and the Saturated Loglinear Model Estimated with the Simulated Data Set

Model	Log-Likelihood	BIC	AIC	AIC3
$K = 1$	−35578	71211	71168	71174
$K = 2$	−28380	56880	56786	56799
$K = 3$	−28114	**56411**	56267	56287
$K = 4$	−28096	56441	56246	56273
$K = 5$	−28073	56459	56213	56247
$K = 6$	−28051	56479	**56183**	**56224**
$K = 7$	−28044	56530	56184	56232
$K = 8$	−28038	56582	56186	56241
$K = 9$	−28033	56637	56190	56252
$K = 10$	−28029	56694	56197	56266
Saturated	−28025	56630	56176	56239

which can be explained from the fact that the data are not agreement with a clean latent structure. What is of interest for the current application is whether the selected imputation model strongly affects the quality of the multiple imputations—that is, the ability to recover the parameters of the logit model for y_6.

Table 2 reports the obtained logit coefficients and their standard errors for an extended set of analyses. In order to reduce Monte Carlo errors as much as possible, in the MI-based methods we always used 50 imputed data sets. Complete case analysis served as the worst case scenario: The LC MI procedure should clearly not perform worse than this method. The ML with incomplete data (using the LEM program; Vermunt 1997) and saturated log-linear model MI served as the best case scenarios; that is, as the golden standards with which LC MI is compared. It should be noted that because we are dealing with a sample, the parameter estimates obtained with these golden standards are also not exactly equal to their population values (see again Table 2).

Complete case analysis performed rather well. The largest biases occurred in the constant and the effect of y_5, the independent variable that is related to missingness in y_2. As can be expected when 70 percent of the sample is excluded from the analysis, standard errors were much larger than they were for ML with incomplete data and log-linear MI. As could be expected, these latter two methods produced very similar results: the only difference is the larger standard error for the interaction

TABLE 2
Parameter Estimates and Standard Errors of the Logit Model Estimated with the Simulated Data Using Complete Case Analysis, Maximum Likelihood with Incomplete Data, and Multiple Imputation Under a Saturated Log-linear Model and Under Latent Class Models with 1, 2, 3, 6, and 10 Classes.

	Predictor													
	Constant		y_1		y_2		y_3		y_4		y_5		$y_2\,y_3$	
Method	Coef.	S.E.	Coef.	S.E.	Coef.	S.E.	Coef.	S.E.	Coef.	S.E.	Coef.	S.E.	Coef.	S.E.
True value	−3.00		1.00		2.00		2.00		1.00		1.00		−2.00	
Complete case analysis	−2.33	0.099	1.00	0.105	1.82	0.132	1.99	0.155	1.15	0.107	0.76	0.106	−2.09	0.208
ML with incomplete data	−3.00	0.074	1.03	0.082	1.86	0.117	1.96	0.103	1.08	0.063	1.07	0.064	−1.99	0.157
Saturated log-linear MI	−3.01	0.076	1.02	0.083	1.89	0.113	1.99	0.112	1.10	0.065	1.05	0.065	−2.04	0.174
LC MI with $K = 1$	−2.70	0.076	0.58	0.062	0.68	0.092	1.63	0.103	1.32	0.056	1.34	0.056	−0.62	0.136
LC MI with $K = 2$	−2.81	0.067	0.93	0.069	1.48	0.105	1.57	0.094	1.06	0.061	1.05	0.062	−1.16	0.141
LC MI with $K = 3$	−2.86	0.068	0.95	0.074	1.66	0.107	1.69	0.097	1.04	0.062	1.02	0.063	−1.43	0.144
LC MI with $K = 6$	−2.99	0.077	1.05	0.080	1.83	0.123	1.94	0.110	1.07	0.062	1.07	0.066	−1.95	0.172
LC MI with $K = 10$	−2.99	0.075	0.99	0.083	1.87	0.118	1.97	0.112	1.09	0.064	1.06	0.065	−2.00	0.184

term under log-linear MI. Although we did not apply sequential regression and approximate Bayesian bootstrap MI, it can be expected that these methods will also work well in this application, where for sequential regression imputation it is, of course, crucial that the interaction term is included in the imputation model.

MI under a 1-class model performed badly. By increasing the number of classes to 2 and 3, most parameter estimates and standard errors improved. However, for the 3-class model, which was suggested by the BIC, the interaction term and the main effects of the two predictors involved in the interaction term obtained with the 3-class model were still rather far from the population values. The 6-class model—the model selected by AIC and AIC3—recovered the parameters and the standard errors very well, and the same applied to the 10-class model. These results confirm our initial idea that it is important to select K sufficiently large, and that overfitting is less problematic than underfitting. Whereas the 3-class does not seem to pick up the three-variable association well, the 6-class does. Moreover, using a 10-class model, an unidentified LC model which clearly overfits the data does not seem to be problematic.

4.2. *A Real Data Set Example*

In this second example, LC-based MI was applied to a data set from the *ATLAS Cultural Tourism Research Project 2003* (ATLAS 2004), a study on the motivations, activities, and impressions of visitors of cultural sites and events. The data set consists of 4292 observations and 79 categorical variables: 52 with 2 categories, 1 with 3, 19 with 5, 2 with 6, and 1 with 7, 8, 9, 10, and 17 categories, respectively. Complete information is available for only 794 respondents. The aim of our application of LC MI to this large data set is to illustrate the main merit of this method compared with log-linear MI; that is, whereas log-linear MI cannot be used as an imputation model for 79 variables, LC MI can without any problem. Note that other alternatives to log-linear MI, such as sequential regression and approximate Bayesian bootstrap MI, would also be difficult to apply in an imputation model with 79 variables.

As the first step we estimated LC models with 1 to 35 latent classes to select a model for MI. The obtained BIC, AIC, and AIC3 values pointed at different possible imputation models: BIC selected

the model with 8 latent classes, AIC3 with 31 latent classes, and AIC still did not reach its minimum value with 35 latent classes. We generated multiply imputed versions of the ATLAS data set using 8, 31, and 50 latent classes and $M = 10$.

After performing the MI, six variables were selected from the data file for a statistical analysis. A central survey question in this study on respondents' motivations for visiting cultural attractions is "I want to find out more about the local culture," answered on a five-point scale ranging from 1 (totally disagree) to 5 (totally agree). This variable was used as the dependent variable in an (adjacent-category) ordinal regression model (Agresti 2002:286–88). Table 3 provides detailed information on the variables used in the analysis, among others on the number of cases with a missing value. We estimated two regression equations, one without and one with "Admission expenditure," a predictor with a very large proportion of missing values. Inclusion of "Admission expenditure" reduces the number of cases with complete information from 3950 to 1424.

Tables 4 and 5 present the coefficients of the two ordinal regression models obtained using complete case analysis as well as the three multiple imputed data sets. For the first analysis, the differences between the four sets of estimates are rather small, which confirms the finding from the simulated data example that it does not matter so much how many latent classes are used in the imputation model as long as their number is large enough. Moreover, because of the rather small proportion of missing values, it is not surprising that complete case analysis and MI gave similar results, although there are some differences in the parameter estimates for education.

Whereas the three imputed data sets give very similar results for the second analysis, complete case analysis results are rather different (see Table 5). Not only are the standard errors much larger, also the effect of education seems to have been distorted by the fact that such a large portion of the sample should be excluded from the analysis.

Although, contrary to the simulated data example, it is not possible to compare the obtained estimates with their population values, LC MI seems to work well in this application. It is reassuring that estimates are similar to complete case analysis when the proportion of missing values is small, are similar across the two regression equations, and are not strongly dependent on the number of classes used in the imputation model.

TABLE 3
Information on the Variables Used in the Ordinal Regression for the ATLAS Cultural Tourism Research Project 2003 Data (ATLAS 2004)

Variable		Categories	Number of Missing Values ($N = 4292$)
I want to find out more about the local culture	1	Totally disagree	154
	2	Disagree	
	3	Neutral	
	4	Agree	
	5	Totally agree	
Gender	1	male	41
	2	female	
Age	1	15 or younger	28
	2	16–19	
	3	20–29	
	4	30–39	
	5	40–49	
	6	50–59	
	7	60 or older	
Highest level of educational qualification	1	Primary school	62
	2	Secondary school	
	3	Vocational education	
	4	Bachelor's degree	
	5	Master's or doctoral degree	
Is your current occupation (or former) connected with culture?	1	Yes	149
	2	No	
Admission expenditure	1	0 – < 25 euro	2801
	2	25 – < 50 euro	
	3	50 – < 75 euro	
	4	75 – < 100 euro	
	5	$\geq$ 100 euro	

5. CONCLUSION

This paper dealt with MI of missing values in data sets containing large numbers of categorical variables. More specifically, for situations in which the standard log-linear modeling imputation approach is no longer feasible, we proposed using an unrestricted LC model as an alternative MI tool. The LC model is not only a flexible categorical data model that is able to pick up complex dependencies between the variables included in the imputation model, but it can also be easily estimated with

TABLE 4
Parameter Estimates ànd Standard Errors of the First Ordinal Regression for the ATLAS Cultural Tourism Research Project 2003 Data Using Complete Case Analysis and LC MI with 8, 31, and 50 Classes

	Complete Cases ($N = 3950$)		$K = 8$		$K = 31$		$K = 50$	
Predictor	Coef.	S.E.	Coef.	S.E.	Coef.	S.E.	Coef.	S.E.
Gender	0.11	0.03	0.09	0.03	0.09	0.03	0.09	0.03
Age	0.04	0.01	0.04	0.01	0.04	0.01	0.04	0.01
Primary school	0.00		0.00		0.00		0.00	
Secondary school	−0.13	0.11	−0.17	0.11	−0.19	0.11	−0.18	0.11
Vocational education	−0.08	0.11	−0.13	0.11	−0.13	0.11	−0.12	0.11
Bachelor's degree	0.06	0.11	0.00	0.11	−0.01	0.11	0.00	0.11
Master's or doctoral degree	0.14	0.11	0.07	0.11	0.07	0.11	0.08	0.11
Occupation and culture	−0.11	0.04	−0.11	0.04	−0.11	0.04	−0.12	0.04

large numbers of partially observed categorical variables. The necessary steps for obtaining the actual imputations are easy to program using the standard output from LC analysis software. Parameter uncertainty with respect to the LC imputation model was dealt with using a non-parametric bootstrap procedure, which made it possible to perform LC MI within the well-developed maximum likelihood framework.

TABLE 5
Parameter Estimates ànd Standard Errors of the Second Ordinal Regression for the ATLAS Cultural Tourism Research Project 2003 Data Using Complete Case Analysis and LC MI with 8, 31, and 50 Classes

	Complete Cases ($N = 1424$)		$K = 8$		$K = 31$		$K = 50$	
Predictor	Coef.	S.E.	Coef.	S.E.	Coef.	S.E.	Coef.	S.E.
Gender	0.11	0.05	0.09	0.03	0.09	0.03	0.09	0.03
Age	0.03	0.02	0.04	0.01	0.04	0.01	0.04	0.01
Primary school	0.00		0.00		0.00		0.00	
Secondary school	−0.39	0.22	−0.18	0.11	−0.20	0.11	−0.19	0.11
Vocational education	−0.36	0.22	−0.14	0.11	−0.15	0.11	−0.13	0.11
Bachelor's degree	−0.14	0.22	−0.01	0.11	−0.02	0.11	−0.01	0.11
Master's or doctoral degree	−0.16	0.22	0.06	0.11	0.05	0.11	0.07	0.11
Occupation and culture	−0.10	0.06	−0.11	0.04	−0.11	0.04	−0.12	0.04
Admission expenditure	0.05	0.02	0.03	0.01	0.04	0.01	0.04	0.01

In this research, we did not compare the proposed LC MI imputation model with hot-deck imputation, imputation using a multivariate normal model, or sequential imputation methods. A systematic comparison of LC MI with these methods requires a complex and extended simulation study, which is outside the scope of this paper. Therefore, we do not claim that LC MI is a better approach than these methods. Instead we demonstrated that LC models allow us to construct an MI method that (1) respects the categorical nature of the variables, (2) is flexible in the sense that it can pick up complex associations, (3) is easy to apply and neutral in the sense that no detailed *a priori* content knowledge is needed to build an imputation model, and (4) is applicable to large data sets.

Our simulated data example showed that LC MI with a sufficiently large number of latent classes yields parameter estimates and standard errors that are almost identical to the ones obtained using either ML for incomplete data or log-linear MI. In order to make sure that the number of latent classes is sufficiently large, we recommended the use of AIC3 or AIC to select the number of classes instead of BIC: The harm caused by possibly selecting a model with too many classes turns out to be negligible. The presented real data application, which was mainly meant to show that it is possible to apply LC MI to such large problems, confirmed that after a certain point increasing the number of classes makes little or no difference anymore.

In the present study, we restricted ourselves to the use of the simple unrestricted LC model. The proposed LC MI method has the potential to be expanded to more general situations using readily available more advanced LC models. We offer a few examples of possible extensions:

- Whereas we did not make a distinction between independent variables and dependent variables in our LC MI models, this would be possible using LC models with covariates. Vermunt and Magidson (2003) showed that such a structure yields a better prediction for the dependent variable, and this may also be the case for variables that should be imputed. See also Von Hippel (2007) for an extended discussion on the different roles that dependent and independent variables may play in the context of MI.
- The LC model may be restricted, for example, to account for the ordinal nature for the variables included in the imputation model

that were now all treated as nominal. Moreover, restrictions can be imposed on the latent classes themselves—for example, to yield latent classes that are in agreement with a particular multidimensional latent structure (see Magidson and Vermunt 2001).

- The LC model can be also extended to include continuous variables in addition to categorical variables (McLachlan and Peel 2000; Vermunt and Magidson 2002). This may provide an alternative to MI under the general location model (Schafer 1997:289–331) when the number of variables is large.
- For the imputation of longitudinal data, we may use special types of LC models that have been developed for such situations, such as the discrete-state latent Markov model (Van de Pol and Langeheine 1990; Vermunt, Bac, and Magidson 2008).
- Similarly, when the data set has a multilevel structure, we may choose to impute the missing values using a multilevel LC model (Vermunt 2003).
- Whereas the typical MI model assumes MAR missing data, NMAR models may be specified by including the response indicators matrix as an additional set of observed variables in the LC imputation model (Moustaki and Knott 2000). Setting up NMAR MI model is, however, not at all straightforward (for example, see Allison 2000).

Each of these extensions would provide worthwhile topics for future research.

APPENDIX: USING LATENT GOLD 4.5 FOR MULTIPLE IMPUTATION

As was illustrated using the two examples in this paper, three steps have to be considered for multiple imputation. Each step can be easily performed using the Latent GOLD 4.5 software (Vermunt and Magidson 2008).

In step 1 an LC model must be selected. This is most easily achieved using the Latent GOLD graphical point and click user interface, because it allows estimation of a series of models—say, LC models with 1 to 10 classes—in a single run. In the technical settings, it should be noted that missing values should be included in analysis. Moreover, to make the occurrence of local maxima less likely, we may

set the number of random start sets to 100 and the number of initial EM iterations per set to 250. In large models with many parameters (as in our second example), it is wise to suppress the use of the Newton Raphson algorithm and the computation of standard errors. When the Newton Raphson is suppressed, the maximum number of EM iteration should be increased to, for example, 5000 to ensure convergence.

In step 2, the selected LC model is used to generate *M* completed data sets. Once a particular LC model is selected, the "Generate Syntax" option should be used to create a syntax version of the selected LC model. One line should be added to this syntax file—that is, outfile filename imputation=M; where M indicates the requested number of imputations. As an illustration, we show the syntax of the 6-class imputation model used for our simulated data example:

```
options
    algorithm tolerance=1e-008 emtolerance=0.01
      emiterations=5000 nriterations=0;
    startvalues seed=0 sets=100
      tolerance=1e-005 iterations=250;
    bayes categorical=1 variances=1 latent=1
      poisson=1;
    missing includeall;
    output profile;
    outfile data='imputedlca6.dat'
      imputation=50;
variables
    dependent Y1 nominal, Y2 nominal,
      Y3 nominal, Y4 nominal, Y5 nominal,
      Y6 nominal;
    latent Class nominal 6;
equations
    Class <- 1;
    Y1 <- 1 + Class;
    Y2 <- 1 + Class;
    Y3 <- 1 + Class;
    Y4 <- 1 + Class;
    Y5 <- 1 + Class;
    Y6 <- 1 + Class;
```

When running this syntax, a new data file is created called imputedlca6.dat containing 50 stacked imputed data files, as well as new variable imputation˙# containing the data set number.

An alternative to using the "Generate Syntax" option is to store the selected LC model using the save "Syntax with Parameters" option, in which case reestimation of the imputation model uses the stored parameters as starting values.

In step 3, the statistical analyses are conducted on all M completed data sets, and the results are combined. With the Latent GOLD syntax, it is possible to run these statistical analyses (for example, fitting a logit model) on all data sets created in step 2 and combine the results. The entire process is automated. As an example, we give the syntax used for the statistical analysis of the the simulated data set:

```
options
    algorithm tolerance=1e-008
      emtolerance=0.01 emiterations=250
      nriterations=50;
    startvalues seed=0 sets=10
      tolerance=1e-005 iterations=50;
    bayes categorical=1 variances=1
      latent=1 poisson=1;
    missing excludeall;
    output parameters=first standarderrors
      estimatedvalues;
variables
    imputationid imputation˙#;
    independent Y1 nominal, Y2 nominal,
      Y3 nominal, Y4 nominal, Y5 nominal;
      dependent Y6 nominal;
equations
    Y6 <- 1 + Y1 + Y2 + Y3 + Y4 + Y5 + Y2 * Y3;
```

The only difference with a standard analysis is the inclusion of the line imputationid imputation˙#; (printed in boldface) in the variables section of the syntax file. The program will analyze each of the imputed data sets and combine the results using the well-known formulas.

REFERENCES

Agresti, Alan. 2002. *Categorical Data Analysis*. New York: Wiley.

Akaike, H. 1974. "A New Look at Statistical Model Identification." *IEEE Transactions on Automatic Control* 19:716–23.

Allison, Paul. 2000. "Multiple Imputation for Missing Data: A Cautionary Tale." *Sociological Methods and Research* 28:301–9.

———. 2005. "Imputation of Categorical Variables with PROC MI." Presented at the SAS Users Group International Conference, April 10–13, Philadelphia, PA.

Andrews, Rick L., and Imran S. Currim. 2003. "A Comparison of Segment Retention Criteria for Finite Mixture Logit Models." *Journal of Marketing Research* 40:235–43.

ATLAS. 2004. *The ATLAS Cultural Tourism Research Project, 2003*. Retrieved December 2006 (www.geocities.com/atlasproject2004).

Bernaards, Coen A., Thomas R. Belin, and Joseph L. Schafer. 2007. "Robustness of a Multivariate Normal Approximation for Imputation of Binary Incomplete Data." *Statistics in Medicine* 26:1368–82.

Bernaards, Coen A., and Klaas Sijtsma. 2000. "Influence of Imputation and EM Methods on Factor Analysis When Item Nonresponse in Questionnaire Data Is Nonignorable." *Multivariate Behavioral Research* 35:321–64.

Bozdogan, Hamparsum. 1993. "Choosing the Number of Component Clusters in the Mixture Model Using a New Informational Complexity Criterion of the Inverse-Fisher Information Matrix." Pp. 40–54 in *Information and Classification, Concepts, Methods and Applications*, edited by O. Opitz, B. Lausen, and R. Klar. Berlin: Springer.

Dempster, Arthur P., Nan M. Laird, and Donald B. Rubin. 1977. "Maximum Likelihood from Incomplete Data via the EM Algorithm." *Journal of the Royal Statistical Society*, *Series B*, 39:1–38.

Dias, José G. 2004. "Finite Mixture Models: Review, Applications, and Computer-Intensive Methods." PhD dissertation, Groningen University.

Dias, José G., and Jeroen K. Vermunt. Forthcoming. "A Bootstrap-Based Aggregate Classifier for Model-Based Clustering." *Computational Statistics*.

Efron, Bradley, and Robert J. Tibshirani. 1993. *An Introduction to the Bootstrap*. London: Chapman and Hall.

Ezzati-Rice, Trena M., Wayne Johnson, Meena Khare, Roderick J. A. Little, Donald B. Rubin, and Joseph L. Schafer. 1995. "A Simulation Study to Evaluate the Performance of Model-Based Multiple Imputations in NCHS Health Examination Surveys." Pp. 257–266 in *Proceedings of the Annual Research Conference*. Washington, DC: Bureau of the Census.

Fuchs, Camil. 1982. "Maximum Likelihood Estimation and Model Selection in Contingency Tables with Missing Data." *Journal of the American Statistical Association* 77:270–78.

Goodman, Leo A. 1974. "Exploratory Latent Structure Analysis Using Both Identifiable and Unidentifiable Models." *Biometrika* 61:215–31.

———. 2007. "On the Assignment of Individual to Latent Classes." Pp. 1–22 in *Sociological Methodology*, vol. 37, edited by Yu Xie. Boston, MA: Blackwell Publishing.

Graham, John W., and Joseph L. Schafer. 1999. "On the Performance of Multiple Imputation for Multivariate Data with Small Sample Size." Pp. 1–28 in *Statistical Strategies for Small Sample Research*, edited by R. Hoyle. Thousand Oaks, CA: Sage Publications.

Honaker, James, Gary King, and Matthew Blackwell. 2007. "Amelia II: A Program for Missing Data." Institute for Quantitative Social Science, Harvard University.

Horton, Nicholas J., Stuart P. Lipsitz, and Michael Parzen. 2003. "A Potential for Bias When Rounding in Multiple Imputation." *American Statistician* 57(4):229–32.

King, Gary, James Honaker, Anne Joseph, and Kenneth Scheve. 2001. "Analyzing Incomplete Political Science Data: An Alternative Algorithm for Multiple Imputation." *American Political Science Review* 95(1):49–69.

Lazarsfeld, Paul F. 1950a. "The Logical and Mathematical Foundation of Latent Structure Analysis." Pp. 361–412 in *Measurement and Prediction*, edited by S. A. Stouffer et al. Princeton, NJ: Princeton University Press.

———. 1950b. "The Interpretation and Mathematical Foundation of Latent Structure Analysis." Pp. 413–72 in *Measurement and Prediction*, edited by S. A. Stouffer et al. Princeton, NJ: Princeton University Press.

Lin, Ting H., and C. Mitchell Dayton. 1997. "Model Selection Information Criteria for Non-nested Latent Class Models." *Journal of Educational and. Behavioral Statistics*, 22:249–64.

Linzer, Drew A., and Jeffrey Lewis. 2007. *poLCA: Polytomous Variable Latent Class Analysis. R Package Version 1.1*. Retrieved December 2007 http://userwww.service.emory.edu/~dlinzer/poLCA.

Little, Roderick J. A., and Donald B. Rubin. 2002. *Statistical Analysis with Missing Data*. 2nd ed. New York: Wiley.

Magidson, Jay, and Jeroen K. Vermunt. 2001. "Latent Class Factor and Cluster Models, Bi-plots and Related Graphical Displays." Pp. 223–64 in *Sociological Methodology*, Vol. 31, edited by Mark P. Becker and Michael E. Sobel. Boston, MA: Blackwell Publishing.

McLachlan, Geoffrey J., and David Peel. 2000. *Finite Mixture Models*. New York: Wiley.

Moustaki, Irini, and Martin Knott. 2000. "Weighting for Item Non-response in Attitude Scales by Using Latent Variable Models with Covariates." *Journal of the Royal Statistical Society,* Series A, 163:445–59.

Muthén, Linda K., and Bengt O. Muthén. 2006. *Mplus 4.2 User's Guide*. Los Angeles: Muthén and Muthén.

Raghunathan, Trivellore E., James M. Lepkowski, John H. Van Hoewyk, and Peter W. Solenberger. 2001. "A Multivariate Technique for Multiply Imputing Missing Values Using a Sequence of Regression Models." *Survey Methodology*, 27(1):85–95.

Rubin, Donald B. 1976. "Inference and Missing Data." *Biometrika,* 63:581–92.

———. 1987. *Multiple Imputation for Nonresponse in Surveys*. New York: Wiley.

Rubin, Donald B., and Nathaniel Schenker. 1986. "Multiple Imputation for Interval Estimation from Simple Random Samples with Ignorable Nonresponse." *Journal of the American Statistical Association*, 90:822–28.

Schafer, Joseph L. 1997. *Analysis of Incomplete Multivariate Data*. London: Chapman and Hall.

———. 1999. "NORM: Multiple Imputation of Incomplete Multivariate Data Under a Normal Model, Version 2." Software for Windows 95/98/NT, available from http://www.stat.psu.edu/~jls/misoftwa.html.

Schafer, Joseph L., Trena M. Ezzati-Rice, Wayne Johnson, Meena Khare, Roderick J. A. Little, and Donald B. Rubin. 1996. "The NHANES III Multiple Imputation Project." Pp. 28–37 in *Proceedings of the Survey Research Methods Section of the American Statistical Association*. Retrieved May 29, 2006 (http://www.amstat.org/sections/srms/Proceedings/papers/1996_004.pdf).

Schafer, Joseph L., and John W. Graham. 2002. "Missing Data: Our View of the State of the Art." *Psychological Methods* 7:147–77.

Schwarz, Gideon. 1978. "Estimating the Dimension of a Model." *Annals of Statistics*, 6:461–64.

SOLAS. 2001. *SOLAS for Missing Data Analysis 3.0* [computer software]. Cork, Ireland: Statistical Solutions.

S-Plus 8 for Windows. 2006. *S-Plus 8 for Windows* [computer software]. Seattle, WA: Insightful Corporation.

Tanner, Martin A., and Wing H. Wong. 1987. "The Calculation of Posterior Distributions by Data Augmentation." *Journal of the American Statistical Association,* 82:528–40.

Van Buuren, Stef, Jaap P. L. Brand, Karin Groothuis-Oudshoorn, and Donald B. Rubin. 2006. "Fully Conditional Specification in Multivariate Imputation." *Journal of Statistical Computation and Simulation* 76(12):1049–64.

Van Buuren, Stef, and C. G. M. Oudshoorn. 2000. *Multivariate Imputation by Chained Equations: MICE V1.0 User's Manual*. Leiden, Netherlands: Toegepast Natuurwetenschappelijk Onderzoek (TNO) Report PG/VGZ/00.038.

Van de Pol, Frank, and Rolf Langeheine. 1990. "Mixed Markov Latent Class Models." Pp. 213–47 in *Sociological Methodology*, vol. 20, edited by Clifford C. Clogg. Cambridge, MA: Blackwell Publishing.

Van Ginkel, Joost R., L. Andries Van der Ark, and Klaas Sijtsma. 2007a. "Multiple Imputation of Item Scores in Test and Questionnaire Data, and Influence on Psychometric Results." *Multivariate Behavioral Research*, 42:387–414.

Van Ginkel, Joost R., L. Andries Van der Ark, and Klaas Sijtsma. 2007b. "Multiple Imputation of Item Scores When Test Data Are Factorially Complex." *British Journal of Mathematical and Statistical Psychology*, 60:315–37.

Von Hippel, Paul T. 2007. "Regression with Missing Ys: An Improved Strategy for Analyzing Multiple Imputed Data." Pp. 83–117 in *Sociological Methodology*, Vol. 37, edited by Yu Xie. Boston, MA: Blackwell Publishing.

Vermunt, Jeroen K. 1997. "LEM: A General Program for the Analysis of Categorical data." Department of Methodology and Statistics, Tilburg University.

———. 2003. "Multilevel Latent Class Models." Pp. 213–39 in *Sociological Methodology*, vol. 33, edited by Ross M. Stolzenberg. Boston, MA: Blackwell Publishing.

Vermunt, Jeroen K., and Jay Magidson. 2002. "Latent Class Cluster Analysis." Pp. 89–106 in *Applied Latent Class Analysis*, edited by Jacques A. Hagenaars and Allan McCutcheon. Cambridge, England: Cambridge University Press.

———. 2003. "Latent Class Models for Classification." *Computational Statistics and Data Analysis* 41: 531–37.

———. 2004. "Latent Class Analysis." Pp. 549–53 in *The Sage Encyclopedia of Social Science Research Methods*, edited by M. S. Lewis-Beck, A. E. Bryman, and T. F. Liao. Thousand Oaks, CA: Sage Publications.

———. 2005. *Latent GOLD 4.0 User's Guide*. Belmont, MA: Statistical Innovations.

———. 2008. *LG-syntax User's Guide: Manual for Latent GOLD 4.5 Syntax Module*. Belmont, MA: Statistical Innovations.

Vermunt, Jeroen K., Bac Tran, and Jay Magidson. 2008. "Latent Class Models in Longitudinal Research." Pp. 373–85 in *Handbook of Longitudinal Research: Design, Measurement, and Analysis*, edited by S. Menard. Burlington, MA: Elsevier.

Yuan, Yang C. 2000. "Multiple Imputation for Missing Data: Concepts and New Development." *Proceedings of the Twenty-Fifth Annual SAS Users Group International Conference* (Paper No. 267). Cary, NC: SAS Institute. Retrieved May 29, 2006 (http://www.ats.ucla.edu/stat/sas/library/multipleimputation.pdf).