Journal of Statistical Planning and Inference 140 (2010) 3252–3262

Contents lists available at ScienceDirect

Journal of Statistical Planning and Inference

journal homepage: www.elsevier.com/locate/jspi

# Latent class based multiple imputation approach for missing categorical data

Mulugeta Gebregziabher *, Stacia M. DeSantis

*Medical University of South Carolina, Department of Medicine, Division of Biostatistics and Epidemiology, 135 Cannon St., Charleston Suite 303, SC 29425, USA*

ARTICLE INFO

*Article history:*
Received 3 July 2009
Received in revised form
19 April 2010
Accepted 20 April 2010
Available online 24 April 2010

*Keywords:*
Bias
Case–control data
Latent class
Missing data
Multiple imputation

ABSTRACT

In this paper we propose a latent class based multiple imputation approach for analyzing missing categorical covariate data in a highly stratified data model. In this approach, we impute the missing data assuming a latent class imputation model and we use likelihood methods to analyze the imputed data. Via extensive simulations, we study its statistical properties and make comparisons with complete case analysis, multiple imputation, saturated log-linear multiple imputation and the Expectation–Maximization approach under seven missing data mechanisms (including missing completely at random, missing at random and not missing at random). These methods are compared with respect to bias, asymptotic standard error, type I error, and 95% coverage probabilities of parameter estimates. Simulations show that, under many missingness scenarios, latent class multiple imputation performs favorably when jointly considering these criteria. A data example from a matched case–control study of the association between multiple myeloma and polymorphisms of the Inter-Leukin 6 genes is considered.



## 1. Introduction

Missing covariate data are common in biomedical studies of categorical risk factors of disease. For example, unordered categorical demographic data such as race, gender, and marital status are frequently measured as risk factors, but may be missing for some study subjects. In the example that motivates the current investigation, we seek to investigate the association between multiple myeloma and a polymorphism of the IL-6 gene called the IL-6r$\alpha$ receptor SNP (-174 RA), which is missing in 40 pairs from a total of 112 case–control pairs due to problems related to assaying (Cozen et al., 2006). Moreover, two potential confounders body mass index (BMI) and education were missing 10% and 5% of values, respectively. The result is the omission of 27% of the study subjects if complete case analysis is used. For studies with small sample sizes like this, in the presence of a large proportion of missing covariate values, it can be difficult to obtain valid inference by analyzing the complete data using standard methods, irrespective of whether the missing data mechanism is ignorable or nonignorable. However, reasonable inference could be made by using principled missing data analysis techniques (Little and Rubin, 2002; Schafer, 1997a).

Despite the ubiquitousness of missing categorical data, there are not readily available principled methods for handling missing values for categorical variables. Current statistical methods for imputing missing categorical data have limited use in practice because of the concern about robustness and/or difficulty in implementation when the number of categorical

* Corresponding author. Fax: +1 843 876 1126.
*E-mail addresses:* gebregz@musc.edu (M. Gebregziabher), desantis@musc.edu (S.M. DeSantis).

0378-3758/$ - see front matter © 2010 Elsevier B.V. All rights reserved.
doi:10.1016/j.jspi.2010.04.020

variables is large. For example, two commonly used noncluster based approaches for missing categorical data are log-linear multiple imputation (LLMI) (Schafer, 1977b), and an ad hoc rounding approach (Bernaards et al., 2007). The former is computationally difficult as the number of categorical variables becomes moderately large (Schafer, 1997a). The latter ignores the categorical nature of the data by using procedures for continuous data with subsequent rounding to the nearest integer. While it has been shown that ignoring the categorical nature of the data by rounding to the nearest integer is robust to violations of normality (Bernaards et al., 2007), other authors have demonstrated that resulting parameter estimates may be biased (Allison, 2006; Horton et al., 2003).

As demonstrated in a recent study by Vermunt et al. (2008), latent class multiple imputation may be used to efficiently impute missing categorical data in the presence of a large number of observed categorical variables. Since latent class analysis explains the variation in observed variables typically using a small number of latent classes, and since it is directly applicable in the presence of many categorical variables, it is natural to apply in the missing categorical data setting. While latent class multiple imputation has been shown by Vermunt et al. (2008) to outperform log-linear imputation, these authors only made limited assessments of the method in terms of missing data mechanism scenarios. Moreover, they used the computationally less efficient (Xiang et al., 2006) nonparametric bootstrap approach instead of a full Bayesian Markov chain Monte Carlo (MCMC) for sampling from the posterior distribution of the missing data. They implemented their procedure in a specialized software called Latent GOLD.

Although a few other cluster based imputation methods have been developed, most consider imputation in extremely high-dimensional continuous data settings where the goal of the analysis is the actual clustering (i.e., where the number of variables is much larger than the number of subjects) (Fujikawa and Ho, 2002; Godfrey et al., 2002; Joernsten et al., 2007). But few, if any, reliable clustering based imputation methods have been proposed for studies whose goal is to measure association in the presence of missing categorical covariate data.

In summary, there are limitations with the above-mentioned nonclustered and cluster based approaches when applied to the imputation of data with a moderate to large number of covariates with missing data. Therefore, in this paper we evaluate a latent class based procedure for implementing multiple imputation of missing categorical covariates. Our aim is three fold. Firstly, we develop the method under a highly stratified data model assumption with one or more missing categorical covariate (e.g., for individually matched case–control data). Secondly, we investigate the statistical properties of this approach via an extensive simulation study under several missing data scenarios, and compare performance to existing imputation procedures. Thirdly, using a real data example, we demonstrate its applicability for missing genotypic data in matched case–control studies.

The outline of the rest of the paper is as follows. In Section 2 we briefly describe commonly used missing data techniques. In Section 3 we develop and describe latent class multiple imputation (LCMI). In Section 4, we assess its performance relative to other commonly used methods, and in Section 5, we apply the methods to the multiple myeloma data. Finally, in Section 6, we discuss our findings along with possibilities for future research.

## 2. Methods

### 2.1. Data, notation and model

While the problem can be developed under a general linear model framework, we considered a logistic model of the form

$$\text{logit}[\text{pr}(Y_{ij}=1|W_i,X_{ij},Z_{ij})]=q(W_i)+\beta_1 Z_{ij}+\beta_2 X_{ij}, \tag{1}$$

where $i=1,\ldots,K$ strata, $X_{ij}$ (subject to missing) and $Z_{ij}$ (not subject to missing) are the corresponding covariates. Let each strata, defined by $W_i$, have $n_i+1$ subjects and let $j=0,\ldots,n_i$. The coefficients $\beta_1$ and $\beta_2$ are the log-odds ratio and $q(W_i)$ is the random stratum-specific effect (or random baseline odds). This model is commonly known as a stratum-specific (random effects) logistic model. Suppose $S_{ij}$ indicates whether or not a subject is sampled from the population and $R_{ij}$ is a missing indicator for covariate $X_{ij}$ defined as $R_{ij}=I(X_{ij}$ is observed) which takes values $r_{ij}$. Assuming a missing at random (MAR) missing mechanism, $P(R=1|X,W,Y,Z)=P(R=1|W,Y,Z)$ and unbiased sampling of cases and controls, $P(S=1|X,W,Y,Z)=P(S=1|Y,W)$, we can re-write Eq. (1) as

$$\text{logit}[\text{pr}(Y_{ij}=1|W_i,X_{ij},Z_{ij},S_{ij}=1)]=q(W_i)+\beta_1 Z_{ij}+\beta_2 X_{ij}+\alpha(W_i)^*=q^*(W_i)+\beta_1 Z_{ij}+\beta_2 X_{ij},$$

where $q^*(W)=\alpha^*(W)+q(W)$, and $\alpha^*(W)=\log(P(S=1|Y=1,W)/P(S=1|Y=0,W))$. Further define the joint density of $X_{ij}$ and $Z_{ij}$ as $h_{ij}(y)=h(Z_{ij},X_{ij}|Y_{ij}=y,W_i,S_{ij}=1)$. The full likelihood is

$$L(\beta_1,\beta_2,q(W_1)\ldots q(W_K))=\prod_{i=1}^{K}\prod_{j=0}^{n_i}h_{ij}(y).$$

This likelihood can be estimated in several different ways based on how one handles the large number of nuisance parameters, $q(W_i)$, which could be a source of loss of efficiency. Two commonly used approaches are the random effects approach (Breslow and Clayton, 1993) and the fixed effects approach, which includes conditional likelihood (CL)

(Diggle et al., 1994). We used the CL method for matched case–control data (Breslow and Day, 1980). In fact, since sampling in matched case–control studies is made from the distribution of $(Z,X)$ conditional on $(Y,W)$, and because $\sum_j Y_{ij}$ is a complete sufficient statistic for $q(W_i)$, CL is a natural choice. The corresponding CL is

$$L_c(\beta_1,\beta_2)=\prod_{i=1}^{K}\frac{h_{i0}(1)\prod_{j=1}^{n_i}h_{ij}(0)}{\sum_{j=0}^{n_i}\{h_{ij}(1)\prod_{k\neq j}h_{ik}(0)\}}=\prod_{i=1}^{K}\frac{\exp(\beta_1 Z_{i0}+\beta_2 X_{i0})}{\sum_{j=0}^{n_i}\exp(\beta_1 Z_{ij}+\beta_2 X_{ij})}.$$

The maximum likelihood estimates from the CL score function are semi-parametric efficient estimators of $\beta_1$ and $\beta_2$ in the presence of the nuisance parameters $q(W_i)$ (Rathouz, 2003).

Assuming $X$ is categorical and is not fully observed, it is our goal to impute the missing values of $X$ in order to efficiently estimate parameters of the model given in Eq. (1). The details are given in Section 3.

### 2.2. Multiple imputation

In the first stage of multiple imputation (MI), $m$ simulated versions of the missing data are created under a data model using methods that incorporate appropriate variability across the $m$ imputations. In the second stage, the $m$ versions of the complete data are analyzed using standard analysis techniques, and the results are combined to produce the final results (Little and Rubin, 2002; Rubin, 1987). MI has been shown to result in valid statistical inferences that properly reflect the uncertainty due to missingness.

The imputation model is a joint distribution of the missing indicator, $R_{ij}$, and the observed and missing variables. This is defined as

$$\Pr(R_{ij},X_{ij},Z_{ij},W_i;\beta_1,\beta_2,\gamma)=\Pr(X_{ij},Z_{ij},W_i;\beta_1,\beta_2)\Pr(R_{ij}|X_{ij},Z_{ij},W_i;\gamma)=\Pr(X_{ij}|Z_{ij},W_i;\beta_1,\beta_2)\Pr(\beta_1|Z_{ij},W_i)\Pr(R_{ij}|X_{ij},Z_{ij},W_i;\gamma).$$

The method of choice for the first stage depends on the pattern of missingness, assumptions about the missing mechanism, and whether or not the distributions for $R_{ij}$ and $X_{ij}$ involve common parameters. Assuming data are MAR and that there are no common parameters between the two distributions, for data with monotone missing patterns, either a parametric method that assumes multivariate normality or a nonparametric method may be used. For data sets with arbitrary missing patterns, an MCMC approach that assumes multivariate normality is used to impute all missing values (or enough missing values such that the imputed data sets have monotone missing patterns). For missing categorical data, parametric models could be used (for instance, logistic or discriminant analysis) (Little and Rubin, 2002; Schafer, 1997a).

### 2.3. Log-linear imputation

Log-linear multiple imputation (LLMI) uses a saturated log-linear model to impute missing categorical data (Schafer, 1997a). As described in Agresti (2002), the saturated log-linear model can be reformulated as a logistic regression model with all two-way and higher order interaction terms included. We implemented LLMI in PROC MI by specifying the parametric logistic model with all main effects and higher order interactions included. The procedure uses MCMC sampling from the posterior distribution of the parameters of the log-linear model to get proper imputations (Little and Rubin, 2002). We note that since all higher order associations among categorical variables need to be considered, only a small number of variables may be included in the imputation model thereby rendering LLMI computationally infeasible even for a moderately large number of variables.

### 2.4. Expectation–Maximization method

The Expectation–Maximization (EM) algorithm iteratively finds the maximum likelihood estimate (MLE) of parameters (Dempster et al., 1977). The expected value of the complete data log likelihood is ascertained in the E-step. In the M-step, the complete data log likelihood is maximized with respect to parameter estimates. The two steps are iterated until a convergence criterion is met. We implemented the algorithm using the 'EM' estimation option in PROC MI assuming that the missing values constitute a parameter vector from a multivariate normal distribution.

## 3. Latent class multiple imputation

Following notation in Vermunt et al. (2008) we now use $Y$ to denote *all* of the observed covariate data. Suppose, for each subject $i$, $i=1,2,\ldots,N$, we observe $J$ categorical variables $Y_i=(Y_{i1},\ldots,Y_{ij})$. Let $\eta_k$ denote the probability of membership of a subject in unobserved latent class $k$ and let $K_i$ denote the latent class to which subject $i$ belongs, with $K_i$ taking values $k=\{1,\ldots,K\}$. The variables $Y_{ij}$ take values from $\{1,\ldots C_j\}$ where $C_j\geq 2$, thus $C_j$ represents the number of possible categories for categorical variable $j$. We denote the probability distribution of $Y_{ij}$ given latent class as $\pi_{jk}=P(Y_{ij}=c|K_i=k), c=1\ldots,C_j$.

For unordered categorical variables, we parameterize the $\pi_{jk}$ as

$$\pi_{jk}(c) = \frac{\exp(\beta_{jk})}{1+\sum_{l=1}^{C_j-1}\exp(\beta_{jkl})} \quad \text{and} \quad \pi_{jk}(C_j) = \frac{1}{1+\sum_{l=1}^{C_j-1}\exp(\beta_{jkl})}.$$

The $\beta$'s are unknown latent class-specific parameters whose collection is denoted as $\beta=(\beta'_{11},\ldots,\beta'_{kj})$ where each $\beta'_{kj}$ is a vector of length $C_j-1$. Based on the standard assumption that within latent class, variables are independent, the joint probability of $Y_i$ is expressed as

$$\phi(y_i) = P(Y_{i1},\ldots,Y_{ij}) = \sum_{k=1}^{K}\eta_k\prod_{j=1}^{J}\prod_{c=1}^{C_j}\pi_{jk}(c)^{Y_{ij}(c)}. \tag{2}$$

The latent class model can be fitted by maximizing the likelihood below with respect to parameter vectors $\eta$ and $\beta_{jk}$,

$$L(\eta_k,\pi_{jk}) = \sum_{i=1}^{N}\log[\phi(y_i)].$$

This proceeds using the EM algorithm as described for latent class models in Goodman (1974). The EM algorithm involves iterating between posterior probabilities of latent class membership as given by

$$P(K_i=k|Y_i=y_i) = \eta_k\prod_{j=1}^{J}\prod_{c=1}^{C_j}\pi_{jk}(c)^{y_{ij}(c)}/\phi(y_i),$$

where $\phi(y_i)$ is given in Eq. (2). To select the number of classes both the Akaike (AIC) and Bayesian Information Criteria (BIC) have been proposed; the latter has been shown to be superior in studies where the number of observed variables is moderately large (Houseman et al., 2006). In general, since imputation models are predictive models, the main purpose is not model parsimony; rather the purpose is to be able to create plausible imputations of the missing data (imputations that reflect the uncertainty in the observed data). Thus, it is expected that an imputation model should encompass the special features of the sample design and should be fitted in such a way that it is highly predictive of the missingness and without great emphasis on AIC or BIC values (Schafer, 1997a).

LCMI was implemented as follows (technical details are provided in Appendix A. First, we fit the latent class model to the observed data, $y_{i,obs}$. Second, we sampled from the posterior probability of latent class ($K_i$) given the observed data, $P(K_i=k|Y_{i,obs}=y_{i,obs})$. Third, we sampled from the distribution of the missing data conditional on class, $P(y_{i,miss}|K_i=k)$, via MCMC. Fourth, we used a within class posterior sampling via MCMC to impute the value.

The latent class model was fitted using PROC LCA Version 1.1.5 (Lanza et al., 2007, 2008). PROC LCA is a SAS procedure for latent class analysis developed for SAS Version 9.2 for Windows and is used to estimate latent classes measured by categorical indicators. Unlike Vermunt et al. (2008) who used the nonparametric bootstrap, we used the more computationally efficient and readily available full Bayesian MCMC approach to sample from the posterior distribution of the missing data model. Finally, after we imputed the missing categories we used conditional likelihood to estimate the parameters of the model presented in Eq. (1).

## 4. Simulation study

### 4.1. Missing data generation

We generated individually matched case–control sets using the paradigm of risk set sampling from cohort data in continuous time (Gebregziabher and Langholz, 2009; Langholz, 2007; Oakes, 1981; Thomas, 1977). Typically, in this paradigm, a random sample of controls of fixed size are sampled at each failure time of a cohort study independently from controls in each risk set. However, to evaluate the performance of estimators for bias as well as efficiency, it is sufficient and simpler to generate 'independent' risk sets rather than to generate failure time data and then form the risk sets. Details are in Langholz (2007).

As our primary interest in evaluating LCMI lies in determining how well the estimation of a few latent classes can improve upon standard multiple imputation techniques, we simulated data for both $K$=2 and 3 latent classes determined by $J$=5 observed variables. To generate data with a $K$-class model, we first set a vector for latent class prevalence, $(\eta_1,\ldots,\eta_K)$ = (0.35,0.65) for $K$=2 and (0.20,0.35,0.45) for $K$=3. The latent class membership for each individual, $i$, was sampled from a multinomial distribution with the above parameters. Based on latent class membership, we sampled $\beta_{jkc}$ values ($j$=1…$J$; $k$=1,…,$K$; $c$=1,2) from a uniform distribution with a class-dependent range. After parameters were simulated, class-specific probabilities were calculated and the data matrix was obtained conditional on latent class membership by sampling from the binomial distribution (see Eq. (2)).

The outcome variable, $Y$, for each subject was generated according to the model, $\text{logit}[\Pr(Y_{ij}=1)]=\alpha_i+\sum_{i=1}^{5}\beta_jX_{ij}$ where $\beta$ represents the collection of logistic regression parameters $\{\beta_1,\ldots,\beta_5\}$. In a given stratum, case–control status for the $i$ th individual was determined with probability proportional to $\exp(\alpha_i+\sum_{i=1}^{5}\beta_jX_{ij})$. We considered a binary exposure variable, $X_1$ and four covariates $X_2$, $X_3$, $X_4$, and $X_5$ that were generated jointly with $X_1$; these covariates are potential binary

confounders of the relationship between $X_1$ and $Y$. For simulations under the null case, we set $\beta_1 = 0$. For the nonnull case, we set $\beta_1 = 0.69$ to yield an odds ratio of 2.0 (consistent with that expected in the multiple myeloma data example). Further, two scenarios $\beta_2 = 0$ and 0.69 with $\beta_3, \ldots, \beta_5 = 0$ were considered.

After generating complete data according to the above model, data sets with missing exposure ($X_1$) were generated from the cohort with a 10%, 30% and 50% missing proportion. We considered a wide range of missing scenarios broadly classified into missing completely at random (MCAR), missing at random (MAR), and missing not at random (MNAR), in the sense of Little and Rubin (2002). Further specification of the missingness within MAR was based on the dependence of the probabilities of missing $X_1$ on another covariate or on the outcome. That is, missing $X_1$ may depend on $X_2$ (MAR($X_2$)), on both $X_2$ and $Y$ (MAR($X_2$, $Y$)), or on $Y$ only (MAR($Y$)). Missingness within the MNAR setting was based on the dependence of the probabilities of missing $X_1$ on $X_1$ (MNAR($X_1$)), on both $X_1$ and $X_2$ (MNAR($X_1$, $X_2$)), or on both $X_1$ and $Y$ (MNAR($X_1$, $Y$)). We made the assumption that the missingness model was logistic with all the variables as covariates,

$$\text{logit}[\text{pr}(M = 1|X_1, \ldots X_5, Y)] = \gamma_0 + \gamma_1 X_1 + \gamma_2 X_2 + \gamma_3 X_3 + \gamma_4 X_4 + \gamma_5 X_5 + \gamma_6 Y,$$

where $M = 1 - R$ is a binary indicator that takes a value of 1 if $X_1$ is missing and 0 if $X_1$ is observed. The intercept of the model $\gamma_0$ determines the overall proportion of missingness while the other $\gamma$ parameters are the corresponding log odds ratio of missingness for each variable. The parameters, $\gamma$, that led to those missingness mechanisms are reported in the Web Appendix Table B7.

### 4.2. Simulation results

The imputed data sets from each scenario were analyzed using all five methods described in the paper. From 1000 replicates we computed the mean of both $\hat{\beta}_1$ and $\hat{\beta}_2$, their corresponding asymptotic standard errors, and the empirical coverage probability of the 95% confidence interval. To ensure that simulations were performing adequately, we also computed the type-I error rate and power of the Wald test for the parameters of interest. We report results for the model generated under the assumption of three classes; results for the 2 class model are provided in the Appendix.

The results of the simulation study are tabulated into five tables. Tables 1 and 2 show the means and ASEs of the estimated log odds ratio parameters for the null and nonnull simulation scenarios, respectively. Table 3 shows the means and ASEs of the estimated log odds ratio parameters, $\hat{\beta}_2$ for the completely observed confounder under the null scenario. Table 4 shows the 95% coverage probabilities for these parameters and the type-I error rate of the Wald test for testing $H_0 : \beta_1 = 0$. Finally, the power of the test is reported in the Appendix. Column labels LCMI-2, LCMI-3, and LCMI-4 refer to estimates from LCMI based on fitting a 2, 3, or 4 class imputers model.

As mentioned, our primary interests were two-fold. The first was to extensively assess the performance of LCMI. The second was to compare LCMI to other data imputation procedures, namely LLMI. While we do not expect any one method to perform best in every situation, we examined these methods across a very broad array of missing data scenarios in order to observe the overall behavior of LCMI and provide some general advice.

As expected, the bias and ASE of the parameters of interest increase as the percent missingness increases, as well as when the missing data mechanism is MNAR. Tables 1 and 2 show that other than for some MNAR scenarios, the bias for $\beta_1$ resulting from the imputation procedures, as well as for the potentially confounding covariate, $\beta_2$, is reasonably small across the missing data mechanisms. Under the null, LCMI results in close to zero bias (as do most of the imputation methods) and notably smaller ASEs than observed for all of the comparison methods, including LLMI. The only exception is EM, which has its own shortcomings as elaborated below. LCMI provides estimates of $\beta_1$ with either little or no bias and high efficiency, even when the number of latent classes is misspecified as 2 or 4.

Although EM gives the smallest ASE in general, Table 4 shows that EM has very poor coverage probability and highly inflated type I error rates (as large as 0.33 for MNAR and as large as 0.20 for MAR scenarios), thus we would not advocate this method for ignorable or nonignorable missing data and therefore will not discuss it further.

It is clear in our simulation scenario that LCMI results in smaller ASEs than both LLMI and MI with respect to the parameter of interest, $\beta_1$ (Tables 1 and 2) as well as with respect to the parameter for the potential confounder, $\beta_2$ (Table 3). In general, the more efficient performance of LCMI over LLMI and MI can be seen in both the null and nonnull scenarios, and for every type of missingness considered within these scenarios. In addition, LCMI maintains a type I error rate near 0.05 and a coverage probability near 0.95 across an array of MNAR and MAR scenarios (see Table 4). Most importantly, the correctly specified 3-class model either rivals or outperforms LLMI and MI with respect to coverage and type 1 error rate (see Table 4). There are no scenarios for which LLMI consistently outperforms LCMI with respect to the four criteria considered. The improved performance of LCMI over LLMI is the most pronounced for 50% missingness, which is arguably the most important scenario as accounting for missing data is usually recommended for such high proportions of missingness.

While all of the imputation methods studied give unbiased estimates for 10% missing data and when the missing mechanism is MCAR, they all break down for the MNAR case where missingness depends on both $X_1$ and $Y$. Interestingly, LCMI and LLMI perform reasonably well when the MNAR data did not depend on the outcome $Y$ (MNAR($X_1$)) and LCMI maintains the lowest ASE of the two.

As information criteria often select the wrong number of optimal latent classes, it is important to assess the effect of class misspecification on the subsequent analysis. It is clear under the columns LCMI-2 and LCMI-4 in each of the tables

**Table 1**
Estimates of $\beta_1$ and SE($\hat{\beta}_1$) in a three class LC model, where $\beta_1 = 0$, $\beta_2 = 0$, pr($X_1$=1)=0.5 and pr($X_2$=1)=0.5 for 1:1 matched study with $n$=200, 1000 simulations.

| | CCA | | MI | | EM | | LLMI | | LCMI-2 | | LCMI-3 | | LCMI-4 | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | $\beta_1$ | ASE | $\beta_1$ | ASE | $\beta_1$ | ASE | $\beta_1$ | ASE | $\beta_1$ | ASE | $\beta_1$ | ASE | $\beta_1$ | ASE |
| Pr(Missing=10%) | | | | | | | | | | | | | | |
| MCAR[a] | −0.009 | 0.341 | −0.010 | 0.325 | −0.014 | 0.304 | −0.012 | 0.324 | −0.011 | 0.321 | −0.010 | 0.316 | −0.011 | 0.313 |
| MAR(Y)[b] | −0.012 | 0.349 | −0.008 | 0.328 | 0.000 | 0.303 | −0.013 | 0.327 | 0.001 | 0.325 | 0.003 | 0.316 | −0.001 | 0.314 |
| MAR(X2)[c] | −0.013 | 0.342 | −0.008 | 0.326 | −0.021 | 0.305 | −0.011 | 0.325 | −0.007 | 0.321 | −0.003 | 0.316 | 0.000 | 0.312 |
| MAR(Y,X2)[d] | −0.020 | 0.362 | −0.005 | 0.336 | −0.009 | 0.304 | −0.002 | 0.336 | −0.002 | 0.329 | 0.039 | 0.322 | 0.056 | 0.322 |
| MNAR(X1)[e] | −0.003 | 0.344 | −0.003 | 0.325 | −0.004 | 0.300 | −0.004 | 0.324 | −0.009 | 0.320 | −0.009 | 0.312 | −0.008 | 0.312 |
| MNAR(X1,X2)[f] | −0.019 | 0.344 | −0.005 | 0.324 | −0.007 | 0.301 | −0.002 | 0.322 | −0.009 | 0.320 | −0.010 | 0.315 | −0.013 | 0.312 |
| MNAR(X1,Y)[g] | 0.027 | 0.348 | 0.017 | 0.328 | 0.018 | 0.300 | 0.017 | 0.325 | 0.018 | 0.321 | 0.030 | 0.311 | 0.033 | 0.311 |
| Pr(Missing=30%) | | | | | | | | | | | | | | |
| MCAR[a] | 0.001 | 0.455 | −0.011 | 0.382 | −0.011 | 0.306 | −0.006 | 0.379 | −0.006 | 0.357 | −0.009 | 0.34 | −0.003 | 0.333 |
| MAR(Y)[b] | −0.021 | 0.496 | −0.002 | 0.400 | −0.005 | 0.305 | −0.006 | 0.393 | 0.001 | 0.381 | 0.001 | 0.347 | 0.004 | 0.343 |
| MAR(X2)[c] | 0.002 | 0.473 | 0.002 | 0.394 | 0.011 | 0.307 | −0.009 | 0.391 | 0.002 | 0.374 | 0.006 | 0.355 | 0.009 | 0.348 |
| MAR(Y,X2)[d] | −0.032 | 0.503 | 0.001 | 0.397 | 0.010 | 0.305 | −0.013 | 0.399 | −0.003 | 0.38 | −0.066 | 0.348 | −0.099 | 0.341 |
| MNAR(X1)[e] | −0.027 | 0.474 | −0.021 | 0.379 | −0.037 | 0.295 | −0.007 | 0.375 | −0.023 | 0.363 | −0.015 | 0.334 | −0.024 | 0.332 |
| MNAR(X1,X2)[f] | −0.042 | 0.484 | −0.012 | 0.387 | −0.011 | 0.300 | −0.002 | 0.383 | −0.012 | 0.37 | −0.011 | 0.343 | −0.009 | 0.337 |
| MNAR(X1,Y)[g] | −0.151 | 0.490 | −0.141 | 0.390 | −0.151 | 0.298 | −0.132 | 0.391 | −0.139 | 0.37 | −0.164 | 0.337 | −0.187 | 0.334 |
| Pr(Missing=50%) | | | | | | | | | | | | | | |
| MCAR[a] | −0.054 | 0.721 | −0.024 | 0.475 | −0.009 | 0.314 | −0.011 | 0.478 | −0.012 | 0.432 | −0.002 | 0.383 | 0.001 | 0.365 |
| MAR(Y)[b] | −0.032 | 0.812 | 0.005 | 0.490 | 0.020 | 0.311 | 0.005 | 0.500 | 0.013 | 0.451 | −0.003 | 0.390 | −0.006 | 0.373 |
| MAR(X2)[c] | 0.015 | 0.839 | 0.015 | 0.507 | −0.004 | 0.311 | −0.066 | 0.568 | 0.017 | 0.470 | 0.011 | 0.430 | 0.015 | 0.413 |
| MAR(Y,X2)[d] | −0.082 | 0.817 | −0.014 | 0.489 | 0.001 | 0.309 | −0.020 | 0.516 | −0.016 | 0.453 | −0.077 | 0.400 | −0.117 | 0.386 |
| MNAR(X1)[e] | 0.009 | 0.744 | −0.025 | 0.458 | −0.037 | 0.295 | −0.027 | 0.472 | −0.028 | 0.429 | −0.033 | 0.370 | −0.037 | 0.365 |
| MNAR(X1,X2)[f] | −0.035 | 0.800 | −0.001 | 0.480 | −0.008 | 0.303 | −0.005 | 0.488 | 0.014 | 0.436 | −0.001 | 0.379 | −0.006 | 0.371 |
| MNAR(X1,Y)[g] | −0.186 | 0.790 | −0.144 | 0.470 | −0.167 | 0.296 | −0.153 | 0.495 | −0.127 | 0.431 | −0.172 | 0.381 | −0.194 | 0.369 |

[a] *m*issing *c*ompletely *a*t *r*andom.
[b] *m*issing *a*t *r*andom conditional on *Y*.
[c] *m*issing *a*t *r*andom conditional on *X*2.
[d] *m*issing *a*t *r*andom conditional on *Y*,*X*2.
[e] *m*issing *n*ot *a*t *r*andom conditional on *X*1.
[f] *m*issing *n*ot *a*t *r*andom conditional on *X*1,*X*2.
[g] *m*issing *n*ot *a*t *r*andom conditional on *X*1,*Y*.

that the benefit of LCMI holds for whether or not the number of classes is misspecified (for example, using a 2-class or 4-class imputers model). While bias does not appear to improve by overfitting with respect to the number of latent classes, the ASE does decrease monotonically as one moves from the 2-class to 4-class imputers model. While our simulation findings are generally consistent with reports from Vermunt et al. (2008), our simulations do not necessarily imply a benefit to arbitrarily increasing the number of classes.

## 5. Data example

The motivating data example is a case–control study of the association between multiple myeloma and polymorphisms in the IL-6 region. The details of the study are reported in Cozen et al. (2006). Briefly, cases were residents of Los Angeles County diagnosed with primary multiple myeloma or plasmacytoma (ICD-03 9731-9734) from October 1, 1999 through December 31, 2002, and under age 75 years at diagnosis. Cases were ascertained by the University of Southern California Cancer Surveillance Program (USC-CSP), the population-based cancer registry for Los Angeles County. A total of 150 cases and two groups of controls (117 relative controls and 126 population controls identified by random digit dialing) were recruited. DNA was extracted and together with other SNPs, the IL-6r$\alpha$ receptor SNP in exon 9 was amplified to identify the ala (A) and asp (D) coding alleles. However, in this SNP, the coding allele was missing in a total of 40 (29 cases and 11 population controls) and in at least one of 40 pairs (29 cases and 11 relative controls) due to problems related to assaying (Cozen et al., 2006). Our primary interest is to estimate the association between IL-6r$\alpha$ and the risk of multiple myeloma in the matched case–control study, adjusting for covariates. We focus on the case to relative control comparison.

In addition to missing genotypic data, two cases and four controls were missing information on potential confounding covariates, body mass index (BMI) and education. While we cannot be certain about the missing data mechanism, it is reasonable to assume that it does not depend on case–control status (since blood serum was collected before case–control status was ascertained). Table 5 displays the odds ratio estimates after applying the methods. Analyses show that, as expected, CCA resulted in odds ratio with the largest standard errors. Using complete cases only, the odds ratio [95%

**Table 2**
Estimates of $\beta_1$ and SE($\hat{\beta}_1$) in a three class LC model, where $\beta_1 = 0.69$, $\beta_2 = 0$, pr($X_1$=1)=0.5 and pr($X_2$=1)=0.5 for 1:1 matched study with $n$=200, 1000 simulations.

| | CCA | | MI | | EM | | LLMI | | LCMI-2 | | LCMI-3 | | LCMI-4 | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | $\beta_1$ | ASE | $\beta_1$ | ASE | $\beta_1$ | ASE | $\beta_1$ | ASE | $\beta_1$ | ASE | $\beta_1$ | ASE | $\beta_1$ | ASE |
| Pr(Missing=10%) | | | | | | | | | | | | | | |
| MCAR[a] | 0.738 | 0.378 | 0.730 | 0.361 | 0.725 | 0.337 | 0.726 | 0.359 | 0.727 | 0.355 | 0.691 | 0.350 | 0.680 | 0.348 |
| MAR(Y)[b] | 0.729 | 0.386 | 0.726 | 0.363 | 0.722 | 0.334 | 0.724 | 0.362 | 0.720 | 0.357 | 0.690 | 0.353 | 0.685 | 0.349 |
| MAR(X2)[c] | 0.747 | 0.379 | 0.733 | 0.363 | 0.71 | 0.335 | 0.722 | 0.361 | 0.732 | 0.357 | 0.704 | 0.356 | 0.699 | 0.356 |
| MAR(Y,X2)[d] | 0.731 | 0.402 | 0.730 | 0.368 | 0.739 | 0.335 | 0.717 | 0.369 | 0.717 | 0.362 | 0.720 | 0.353 | 0.738 | 0.351 |
| MNAR(X1)[e] | 0.727 | 0.382 | 0.735 | 0.359 | 0.728 | 0.331 | 0.725 | 0.360 | 0.726 | 0.354 | 0.700 | 0.346 | 0.691 | 0.345 |
| MNAR(X1,X2)[f] | 0.745 | 0.383 | 0.730 | 0.361 | 0.728 | 0.333 | 0.721 | 0.358 | 0.725 | 0.355 | 0.690 | 0.350 | 0.680 | 0.349 |
| MNAR(X1,Y)[g] | 0.755 | 0.388 | 0.755 | 0.362 | 0.745 | 0.331 | 0.749 | 0.359 | 0.742 | 0.355 | 0.726 | 0.349 | 0.727 | 0.349 |
| Pr(Missing=30%) | | | | | | | | | | | | | | |
| MCAR[a] | 0.775 | 0.512 | 0.741 | 0.422 | 0.733 | 0.338 | 0.716 | 0.423 | 0.724 | 0.400 | 0.654 | 0.376 | 0.639 | 0.371 |
| MAR(Y)[b] | 0.766 | 0.553 | 0.744 | 0.433 | 0.741 | 0.336 | 0.721 | 0.434 | 0.716 | 0.414 | 0.66 | 0.384 | 0.665 | 0.380 |
| MAR(X2)[c] | 0.789 | 0.528 | 0.734 | 0.439 | 0.705 | 0.333 | 0.681 | 0.454 | 0.718 | 0.42 | 0.657 | 0.409 | 0.649 | 0.398 |
| MAR(Y,X2)[d] | 0.789 | 0.566 | 0.720 | 0.443 | 0.707 | 0.331 | 0.687 | 0.448 | 0.695 | 0.425 | 0.575 | 0.393 | 0.550 | 0.386 |
| MNAR(X1)[e] | 0.788 | 0.543 | 0.740 | 0.421 | 0.715 | 0.324 | 0.721 | 0.432 | 0.712 | 0.402 | 0.646 | 0.370 | 0.645 | 0.367 |
| MNAR(X1,X2)[f] | 0.743 | 0.538 | 0.722 | 0.421 | 0.708 | 0.328 | 0.696 | 0.427 | 0.709 | 0.407 | 0.659 | 0.374 | 0.648 | 0.368 |
| MNAR(X1,Y)[g] | 0.645 | 0.557 | 0.605 | 0.428 | 0.586 | 0.320 | 0.557 | 0.432 | 0.589 | 0.409 | 0.502 | 0.376 | 0.483 | 0.376 |
| Pr(Missing=50%) | | | | | | | | | | | | | | |
| MCAR[a] | 0.852 | 0.804 | 0.755 | 0.518 | 0.762 | 0.346 | 0.675 | 0.541 | 0.728 | 0.468 | 0.639 | 0.430 | 0.635 | 0.414 |
| MAR(Y)[b] | 0.888 | 0.928 | 0.770 | 0.538 | 0.765 | 0.342 | 0.699 | 0.572 | 0.717 | 0.486 | 0.632 | 0.430 | 0.657 | 0.413 |
| MAR(X2)[c] | 0.866 | 0.926 | 0.757 | 0.579 | 0.682 | 0.335 | 0.644 | 0.678 | 0.718 | 0.544 | 0.639 | 0.503 | 0.633 | 0.489 |
| MAR(Y,X2)[d] | 0.860 | 0.924 | 0.725 | 0.538 | 0.750 | 0.334 | 0.658 | 0.598 | 0.701 | 0.499 | 0.556 | 0.454 | 0.541 | 0.435 |
| MNAR(X1)[e] | 0.906 | 0.912 | 0.759 | 0.521 | 0.760 | 0.323 | 0.694 | 0.543 | 0.711 | 0.468 | 0.646 | 0.410 | 0.640 | 0.407 |
| MNAR(X1,X2)[f] | 0.799 | 0.879 | 0.736 | 0.519 | 0.728 | 0.330 | 0.684 | 0.566 | 0.708 | 0.476 | 0.641 | 0.421 | 0.641 | 0.415 |
| MNAR(X1,Y)[g] | 0.697 | 0.940 | 0.583 | 0.541 | 0.562 | 0.318 | 0.529 | 0.589 | 0.576 | 0.486 | 0.488 | 0.421 | 0.476 | 0.418 |

[a] *m*issing completely *a*t *r*andom.
[b] *m*issing *a*t *r*andom *c*onditional on *Y*.
[c] *m*issing *a*t *r*andom *c*onditional on *X*2.
[d] *m*issing *a*t *r*andom *c*onditional on *Y*,*X*2.
[e] *m*issing *n*ot *a*t *r*andom conditional on *X*1.
[f] *m*issing *n*ot *a*t *r*andom conditional on *X*1,*X*2.
[g] *m*issing *n*ot *a*t *r*andom conditional on *X*1,*Y*.

confidence interval] in those with one versus no copies of the A allele of the IL-6r$\alpha$ polymorphism was 1.60 [0.66,3.86], and in those with one versus two copies was 1.12 [0.28,4.52]. Imputation was made based on a two and three class LC models and results are reported for both LCMI-2 and LCMI-3, respectively. The odds ratio from this LCMI for the 2 class model (which had favorable BIC/AIC values) were 1.95 [1.27,2.97] and 1.51 [0.73,3.13], respectively. Unlike CCA, this represents a significantly increased odds of multiple myeloma in those carrying just one allele, as well as tighter confidence intervals for these estimates. We note that MI and LLMI resulted in similar inference as LCMI, but the asymptotic standard error for LCMI was the smallest among all imputation procedures, which is consistent with the simulation findings. Interestingly, MI and LLMI result in almost the exact same inference, thus the methods appear to be interchangeable in the current context.

Even though we do not recommend it, some researchers analyze missing covariate data from matched case–control studies using unconditional logistic regression after breaking the match. We analyzed the unmatched data with adjustment for matching covariates and the comparative advantage of LCMI was still apparent (see bottom panel of Table 5).

## 6. Discussion

In this paper, we proposed and examined a latent class based multiple imputation (LCMI) of missing covariate data. We provided a comprehensive assessment of LCMI as well as a broader systemic comparison of LCMI with other methods, which has been lacking in the literature (Vermunt et al., 2008). Specifically, we met the goals outlined in the Introduction by demonstrating improved efficiency of this method over existing imputation methods and complete case analysis for estimation of covariate effects (for covariate data that are both missing and nonmissing) in studies with highly stratified data. We used both simulations and a real data example to exemplify the applicability of LCMI to missing covariate data in matched case control studies. Our simulations demonstrated that LCMI leads to unbiased parameter estimates with smaller standard errors than other commonly used approaches, under many missing data scenarios. LCMI also leads to confidence intervals with the nominal 95% coverage. In addition, both reported and unreported sensitivity analyses

**Table 3**
Estimates of $\beta_2$ and SE($\hat{\beta}_2$) in a three class LC model, where $\beta_1 = 0$, $\beta_2 = 0$, pr($X_1$=1)=0.5 and pr($X_2$=1)=0.5 for 1:1 matched study with $n$=200, 1000 simulations.

| | CCA | | MI | | EM | | LLMI | | LCMI-2 | | LCMI-3 | | LCMI-4 | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | $\beta_2$ | ASE | $\beta_2$ | ASE | $\beta_2$ | ASE | $\beta_2$ | ASE | $\beta_2$ | ASE | $\beta_2$ | ASE | $\beta_2$ | ASE |
| Pr(Missing=10%) | | | | | | | | | | | | | | |
| MCAR[a] | 0.004 | 0.341 | −0.006 | 0.305 | −0.006 | 0.306 | −0.006 | 0.306 | −0.005 | 0.306 | −0.006 | 0.305 | −0.006 | 0.304 |
| MAR(Y)[b] | −0.004 | 0.350 | −0.009 | 0.305 | −0.007 | 0.307 | −0.005 | 0.307 | −0.009 | 0.306 | −0.008 | 0.305 | −0.008 | 0.305 |
| MAR(X2)[c] | −0.004 | 0.343 | −0.004 | 0.305 | −0.006 | 0.306 | −0.005 | 0.306 | −0.006 | 0.305 | −0.007 | 0.304 | −0.007 | 0.304 |
| MAR(Y,X2)[d] | −0.004 | 0.364 | −0.006 | 0.305 | −0.007 | 0.307 | −0.009 | 0.307 | −0.008 | 0.306 | −0.016 | 0.305 | −0.019 | 0.305 |
| MNAR(X1)[e] | −0.008 | 0.350 | −0.008 | 0.305 | −0.008 | 0.307 | −0.007 | 0.306 | −0.007 | 0.306 | −0.006 | 0.304 | −0.006 | 0.304 |
| MNAR(X1,X2)[f] | −0.010 | 0.347 | −0.005 | 0.305 | −0.007 | 0.306 | −0.007 | 0.306 | −0.006 | 0.305 | −0.005 | 0.304 | −0.005 | 0.304 |
| MNAR(X1,Y)[g] | −0.004 | 0.355 | −0.010 | 0.305 | −0.010 | 0.307 | −0.010 | 0.306 | −0.011 | 0.306 | −0.012 | 0.304 | −0.013 | 0.304 |
| Pr(Missing=30%) | | | | | | | | | | | | | | |
| MCAR[a] | −0.005 | 0.457 | −0.006 | 0.306 | −0.006 | 0.311 | −0.007 | 0.310 | −0.006 | 0.308 | −0.007 | 0.305 | −0.008 | 0.303 |
| MAR(Y)[b] | 0.007 | 0.499 | −0.008 | 0.307 | −0.008 | 0.314 | −0.005 | 0.312 | −0.008 | 0.310 | −0.008 | 0.305 | −0.008 | 0.305 |
| MAR(X2)[c] | −0.015 | 0.474 | −0.009 | 0.306 | −0.008 | 0.312 | −0.005 | 0.312 | −0.008 | 0.309 | −0.008 | 0.305 | −0.008 | 0.304 |
| MAR(Y,X2)[d] | 0.017 | 0.509 | −0.010 | 0.307 | −0.007 | 0.314 | −0.005 | 0.314 | −0.010 | 0.309 | 0.002 | 0.305 | 0.007 | 0.305 |
| MNAR(X1)[e] | 0.007 | 0.495 | −0.001 | 0.306 | −0.003 | 0.312 | −0.007 | 0.312 | −0.003 | 0.309 | −0.005 | 0.305 | −0.004 | 0.304 |
| MNAR(X1,X2)[f] | −0.028 | 0.493 | −0.006 | 0.306 | −0.007 | 0.312 | −0.008 | 0.312 | −0.005 | 0.309 | −0.006 | 0.305 | −0.006 | 0.304 |
| MNAR(X1,Y)[g] | −0.020 | 0.510 | 0.024 | 0.306 | 0.023 | 0.314 | 0.021 | 0.314 | 0.021 | 0.309 | 0.020 | 0.306 | 0.022 | 0.305 |
| Pr(Missing=50%) | | | | | | | | | | | | | | |
| MCAR[a] | 0.013 | 0.803 | −0.007 | 0.308 | −0.004 | 0.322 | −0.008 | 0.323 | −0.012 | 0.314 | −0.002 | 0.306 | 0.001 | 0.304 |
| MAR(Y)[b] | −0.027 | 2.344 | −0.014 | 0.310 | −0.009 | 0.327 | −0.010 | 0.326 | 0.013 | 0.316 | −0.003 | 0.306 | −0.006 | 0.305 |
| MAR(X2)[c] | 0.082 | 1.920 | −0.002 | 0.309 | −0.009 | 0.327 | 0.008 | 0.335 | 0.017 | 0.316 | 0.011 | 0.308 | 0.015 | 0.306 |
| MAR(Y,X2)[d] | 0.102 | 1.699 | −0.013 | 0.309 | −0.007 | 0.326 | −0.005 | 0.328 | −0.016 | 0.315 | −0.077 | 0.307 | −0.117 | 0.307 |
| MNAR(X1)[e] | −0.020 | 2.001 | 0.002 | 0.308 | −0.004 | 0.324 | −0.004 | 0.323 | −0.028 | 0.314 | −0.033 | 0.305 | −0.037 | 0.305 |
| MNAR(X1,X2)[f] | −0.083 | 1.216 | −0.012 | 0.308 | −0.010 | 0.325 | −0.009 | 0.326 | 0.014 | 0.314 | −0.001 | 0.305 | −0.006 | 0.305 |
| MNAR(X1,Y)[g] | −0.100 | 1.388 | 0.030 | 0.308 | 0.023 | 0.325 | 0.019 | 0.324 | −0.127 | 0.314 | −0.172 | 0.307 | −0.194 | 0.306 |

[a] *m*issing *c*ompletely *at* random.
[b] *m*issing *at* random conditional on *Y*.
[c] *m*issing *at* random conditional on *X*2.
[d] *m*issing *at* random conditional on *Y*,*X*2.
[e] *m*issing *n*ot *a*t random conditional on *X*1.
[f] *m*issing *n*ot *a*t random conditional on *X*1,*X*2.
[g] *m*issing *n*ot *a*t random conditional on *X*1,*Y*.

showed that parameter estimates were robust to latent class model misspecification. While all of the methods break down for the MNAR($X_1$,$Y$) situation, LCMI and LLMI performed reasonably well under MNAR($X_1$) and LCMI maintained the lowest ASE. This is of great practical value in case–control studies, as missingness is likely to depend on case control status especially in settings where data are collected after disease occurrence (Gebregziabher and Langholz, 2009).

Our imputation of the IL-6r$\alpha$ receptor SNP in order to assess its association with multiple myeloma upheld our simulation findings. Firstly, imputation of this SNP based on 2 and 3 class models was consistent thereby supporting the simulation findings that the method is robust to misspecification of the number of classes. This is important since the use of different information criteria often result in conflicting model selection decisions (e.g., Vermunt et al., 2008). Secondly, asymptotic standard errors resulting from LCMI were smaller than those obtained from MI and LLMI, while odds ratio estimates were slightly larger providing evidence for more precise inference.

Our comprehensive assessment of the five methods under numerous missingness mechanisms was focused on a case in which the data were able to be clustered. While we chose a basic simulation scenario (i.e., based on the supposition of a small number of underlying latent classes), it should also be noted that if the data do not lend themselves to clustering, then LCMI would likely be equivalent to MI. In other words, one would impute conditional on a one-class model. In general, latent class imputation is similar to multiple imputation but allows for imputation based only on like observations and leads to improved performance when dealing with missing categorial data. Thus, in situations where log-linear imputation is difficult to implement (e.g., when the number of observed variables is large), LCMI is the preferred method since LCA can be performed when there are a large number of observed categorical variables.

We note that there are several ways in which one can impute missing data based on the posterior probability for the individual with incomplete observations on a categorical covariate (Steps 2 and 4 in Section 2.2). One can sample from the observed data using bootstrap (Vermunt et al., 2008) or from the posterior distribution of the missing data using full Bayesian MCMC, as we have done. On the other hand, one could more simply assign the subject to the class with the highest probability. Another option is to sample from as many classes as are estimated, and use a weighted imputation from the distribution of the missing data given latent class. It is not clear if the imputation procedure holds under other

**Table 4**
95% confidence interval coverage (95CI) and Type-I error rate (TE) of Wald test in a three class LC model for testing $H_0 : \beta_1 = 0$, where $\beta_1 = 0$, $\beta_2 = 0$, pr($X_1$=1)=0.5 and pr($X_2$=1)=0.5 for 1:1 matched study with $n$=200, 1000 simulations.

| | CCA | | MI | | EM | | LLMI | | LCMI-2 | | LCMI-3 | | LCMI-4 | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 95CI | TE | 95CI | TE | 95CI | TE | 95CI | TE | 95CI | TE | 95CI | TE | 95CI | TE |
| Pr(Missing=10%) | | | | | | | | | | | | | | |
| MCAR[a] | 0.94 | 0.04 | 0.94 | 0.04 | 0.91 | 0.07 | 0.94 | 0.05 | 0.95 | 0.04 | 0.95 | 0.03 | 0.95 | 0.04 |
| MAR(Y)[b] | 0.93 | 0.05 | 0.94 | 0.04 | 0.92 | 0.07 | 0.94 | 0.05 | 0.95 | 0.04 | 0.95 | 0.03 | 0.95 | 0.02 |
| MAR(X2)[c] | 0.94 | 0.03 | 0.95 | 0.05 | 0.91 | 0.08 | 0.94 | 0.05 | 0.96 | 0.04 | 0.96 | 0.03 | 0.95 | 0.03 |
| MAR(Y,X2)[d] | 0.94 | 0.05 | 0.94 | 0.03 | 0.90 | 0.08 | 0.95 | 0.03 | 0.94 | 0.04 | 0.95 | 0.03 | 0.96 | 0.04 |
| MNAR(X1)[e] | 0.95 | 0.04 | 0.94 | 0.03 | 0.91 | 0.06 | 0.96 | 0.04 | 0.96 | 0.03 | 0.95 | 0.02 | 0.95 | 0.03 |
| MNAR(X1,X2)[f] | 0.95 | 0.04 | 0.95 | 0.04 | 0.91 | 0.07 | 0.94 | 0.05 | 0.95 | 0.05 | 0.96 | 0.03 | 0.96 | 0.03 |
| MNAR(X1,Y)[g] | 0.95 | 0.04 | 0.95 | 0.04 | 0.91 | 0.06 | 0.96 | 0.03 | 0.96 | 0.04 | 0.96 | 0.03 | 0.97 | 0.03 |
| Pr(Missing=30%) | | | | | | | | | | | | | | |
| MCAR[a] | 0.95 | 0.06 | 0.95 | 0.06 | 0.83 | 0.15 | 0.96 | 0.05 | 0.96 | 0.06 | 0.95 | 0.04 | 0.95 | 0.05 |
| MAR(Y)[b] | 0.95 | 0.03 | 0.93 | 0.06 | 0.84 | 0.18 | 0.95 | 0.05 | 0.94 | 0.05 | 0.95 | 0.04 | 0.94 | 0.04 |
| MAR(X2)[c] | 0.94 | 0.05 | 0.94 | 0.05 | 0.79 | 0.18 | 0.95 | 0.04 | 0.95 | 0.04 | 0.95 | 0.03 | 0.95 | 0.03 |
| MAR(Y,X2)[d] | 0.95 | 0.07 | 0.95 | 0.05 | 0.83 | 0.18 | 0.93 | 0.06 | 0.94 | 0.05 | 0.94 | 0.04 | 0.92 | 0.05 |
| MNAR(X1)[e] | 0.95 | 0.04 | 0.96 | 0.04 | 0.83 | 0.19 | 0.96 | 0.05 | 0.96 | 0.04 | 0.97 | 0.03 | 0.96 | 0.03 |
| MNAR(X1,X2)[f] | 0.96 | 0.04 | 0.95 | 0.04 | 0.84 | 0.16 | 0.95 | 0.04 | 0.96 | 0.05 | 0.97 | 0.04 | 0.96 | 0.03 |
| MNAR(X1,Y)[g] | 0.92 | 0.07 | 0.93 | 0.07 | 0.78 | 0.20 | 0.93 | 0.08 | 0.92 | 0.07 | 0.91 | 0.08 | 0.92 | 0.08 |
| Pr(Missing=50%) | | | | | | | | | | | | | | |
| MCAR[a] | 0.96 | 0.05 | 0.94 | 0.06 | 0.74 | 0.27 | 0.93 | 0.06 | 0.95 | 0.05 | 0.95 | 0.05 | 0.93 | 0.07 |
| MAR(Y)[b] | 0.97 | 0.03 | 0.94 | 0.08 | 0.73 | 0.32 | 0.92 | 0.07 | 0.93 | 0.06 | 0.94 | 0.05 | 0.93 | 0.06 |
| MAR(X2)[c] | 0.96 | 0.04 | 0.93 | 0.06 | 0.70 | 0.28 | 0.91 | 0.04 | 0.93 | 0.06 | 0.94 | 0.05 | 0.94 | 0.05 |
| MAR(Y,X2)[d] | 0.96 | 0.05 | 0.94 | 0.05 | 0.71 | 0.25 | 0.91 | 0.05 | 0.93 | 0.06 | 0.93 | 0.05 | 0.93 | 0.06 |
| MNAR(X1)[e] | 0.97 | 0.03 | 0.93 | 0.07 | 0.73 | 0.30 | 0.94 | 0.06 | 0.94 | 0.05 | 0.94 | 0.04 | 0.93 | 0.05 |
| MNAR(X1,X2)[f] | 0.96 | 0.05 | 0.94 | 0.07 | 0.74 | 0.28 | 0.95 | 0.05 | 0.95 | 0.04 | 0.95 | 0.02 | 0.94 | 0.05 |
| MNAR(X1,Y)[g] | 0.95 | 0.05 | 0.93 | 0.07 | 0.67 | 0.33 | 0.92 | 0.08 | 0.90 | 0.08 | 0.91 | 0.07 | 0.88 | 0.09 |

[a] *m*issing completely *a*t *r*andom.
[b] *m*issing *a*t *r*andom *c*onditional on *Y*.
[c] *m*issing *a*t *r*andom *c*onditional on *X*2.
[d] *m*issing *a*t *r*andom *c*onditional on *Y*,*X*2.
[e] *m*issing *n*ot *a*t *r*andom conditional on *X*1.
[f] *m*issing *n*ot *a*t *r*andom conditional on *X*1,*X*2.
[g] *m*issing *n*ot *a*t *r*andom conditional on *X*1,*Y*.

**Table 5**
Odds ratio (standard error) estimates for the association between multiple myeloma and IL-6rα using conditional likelihood for the individually matched and unconditional likelihood after breaking the match, Los-Angeles County, 1999–2002.

| Variable | Values | CCA | EM | MI | LLMI | LCMI-3 | LCMI-2 |
|---|---|---|---|---|---|---|---|
| *Conditional likelihood*: *odds ratio and SE estimates* | | | | | | | |
| IL6174R | DD | 1.00(–) | 1.00(–) | 1.00(–) | 1.00(–) | 1.00(–) | 1.00(–) |
| | AD | 1.60(0.45) | 1.93(0.37) | 1.92(0.23) | 1.91(0.23) | 2.03(0.22) | 1.95(0.22) |
| | AA | 1.12(0.71) | 1.70(0.62) | 1.25(0.39) | 1.25(0.40) | 1.23(0.37) | 1.51(0.36) |
| Gender | Male | 2.23(0.44) | 2.59(0.38) | 4.44(0.21) | 4.40(0.22) | 4.48(0.21) | 4.48(0.21) |
| Age | Age > 60 | 1.15(0.64) | 2.25(0.56) | 4.66(0.33) | 4.60(0.37) | 4.48(0.32) | 4.48(0.33) |
| BMI | BMI | 1.25(0.26) | 1.27(0.21) | 1.39(0.13) | 1.38(0.13) | 1.40(0.13) | 1.39(0.12) |
| EDUC | EDUC | 1.38(0.27) | 1.23(0.22) | 1.46(0.13) | 1.46(0.13) | 1.43(0.13) | 1.43(0.13) |
| *Unconditional likelihood*: *odds ratio and SE estimates* | | | | | | | |
| IL6174R | DD | 1.00(–) | 1.00(–) | 1.00(–) | 1.00(–) | 1.00(–) | 1.00(–) |
| | AD | 1.46(0.32) | 1.48(0.28) | 1.39(0.14) | 1.39(0.14) | 1.40(0.13) | 1.38(0.13) |
| | AA | 0.79(0.49) | 0.97(0.45) | 0.75(0.23) | 0.72(0.24) | 0.71(0.22) | 0.79(0.21) |
| Gender | Male | 1.67(0.29) | 1.72(0.26) | 1.63(0.12) | 1.63(0.12) | 1.63(0.12) | 1.65(0.12) |
| Race | Black | 1.01(0.36) | 1.04(0.32) | 0.99(0.15) | 1.00(0.16) | 0.97(0.14) | 0.98(0.14) |
| Age | Age > 60 | 1.09(0.29) | 1.12(0.26) | 1.15(0.12) | 1.16(0.12) | 1.15(0.12) | 1.15(0.12) |
| BMI | BMI | 1.12(0.18) | 1.17(0.17) | 1.16(0.08) | 1.16(0.09) | 1.16(0.08) | 1.16(0.07) |
| EDUC | EDUC | 1.12(0.18) | 1.08(0.16) | 1.05(0.07) | 1.06(0.07) | 1.05(0.07) | 1.04(0.07) |

SE=asymptotic standard error; CCA=complete case analysis using conditional logistic; MI=multiple imputation with discriminant method; LLMI=log-linear multiple imputation; EM= Expectation Maximization algorithm; LCMI-2=latent class multiple imputation with two classes; LCMI-3=latent class multiple imputation with three classes.

common missing data scenarios, or whether the bootstrap approach offers any advantages. However, head to head comparisons of nonparametric bootstrapping and a Bayesian approach to sampling from posterior probability distributions using an MCMC method, have showed that the more computationally intensive bootstrap method may give unstable estimates than the full Bayesian MCMC (Alfaro, 2003; Xiang et al., 2006). On the other hand, in the weighted imputation approach, the imputed value for the individual will likely be distributed across classes and hence that might improve efficiency by directly accounting for the uncertainty in latent class estimation. One caveat that needs to be mentioned is the concern that the covariates ($X_1,\ldots,X_5$) are generated under the assumed LC model and therefore might favor LCMI over other methods. However, we clarify that neither the model used to generate the $Y$ based on $X_1,\ldots,X_5$, nor the missing data model $\Pr(M=1|X_1,\ldots,X_5)$, depended on the latent class. Thus, the simulated data provides a fair comparison of the other methods with LCMI.

One limitation of our data example that may have masked some of the advantage of LCMI over MI is that the myeloma data set consists of only a moderate number of variables. Further applications of this method should consider its behavior for a very large number of observed variables. Vermunt et al. (2008) applied the method to impute missing categorical data in the ATLAS Cultural Tourism Research Project. These authors fitted models with up to 26 latent classes from data containing 79 observed variables. It would be interesting to compare the performance of MI to LCMI in these very high-dimensional settings. As few, if any, data imputation techniques for ordinal variables exist, a potential extension would be to apply the method to ordinal data. It could also be extended to missing categorical data in longitudinal settings using latent transition analysis instead of LCA.

In conclusion, this study is the first to implement and assess LCMI in matched case–control data with missing categorical covariates under the highly stratified logistic model framework. It is also the first to consider LCMI under many missing data scenarios. We have demonstrated that this method provides estimates that exhibit high statistical efficiency, little to no bias, and can be implemented using standard statistical software. Even though the method is developed for a highly stratified logit model (i.e., Eq. (1)), it can easily be extended into any generalized linear model framework.

## Acknowledgments

This work is partially supported by SC EPSCoR/IDeA and MUSC office of the Provost. We would like to thank Dr. Wendy Cozen for allowing us to use the myeloma data as a data example in this paper.

## Appendix A. Technical Details of the LCMI method

Following notation in Vermunt et al (2008) we now use $Y$ to denote *all* of the observed covariate data. Suppose, for each subject $i$, $i=1,2,\ldots,N$, we observe $J$ categorical variables $(Y_{i1},\ldots,Y_{ij})$. Let $\eta_k$ denote the probability of membership of a subject in unobserved latent class $k$ and let $K_i$ denote the latent class to which subject $i$ belongs, with $K_i$ taking values $k=\{1,\ldots,K\}$. The variables $Y_{ij}$ take values from $\{1,\ldots C_j\}$ where $C_j\geq 2$, thus $C_j$ represents the number of possible categories for categorical variable $j$. We denote the probability distribution of $Y_{ij}$ given latent class as $\pi_{jk}=P(Y_{ij}=c|K_i=k), c=1,\ldots,C_j$. For unordered categorical variables, we parameterize the $\pi_{jk}$ as,

$$\pi_{jk}(c)=\frac{exp(\beta_{jk})}{1+\sum_{l=1}^{C_j-1}exp(\beta_{jkl})} \quad \text{and} \quad \pi_{jk}(C_j)=\frac{1}{1+\sum_{l=1}^{C_j-1}exp(\beta_{jkl})}$$

The $\beta$'s are unknown latent class-specific parameters whose collection is denoted as $\beta=(\beta'_{11},\ldots,\beta'_{kj})$ where each $\beta'_{kj}$ is a vector of length $C_j-1$. Based on the standard assumption that within latent class, variables are independent, the joint probability of $Y_i$ is expressed as

$$\phi(y_i)=P(Y_{i1},\ldots,Y_{ij})=\sum_{k=1}^{K}\eta_k\prod_{j=1}^{J}\prod_{c=1}^{C_j}\pi_{jk}(c)^{Y_{ij}(c)}. \tag{A.1}$$

The latent class model can be fitted by maximizing the likelihood below with respect to parameter vectors $\eta$ and $\beta_{jk}$,

$$L=\sum_{i=1}^{N}log\sum_{k=1}^{K}\eta_k\prod_{j=1}^{J}\prod_{c=1}^{C_j}\pi_{jk}(c)^{Y_{ij}(c)}.$$

This proceeds using the EM algorithm as described for latent class models in Goodman (1974). The EM algorithm involves iterating between posterior probabilities of latent class membership as given by,

$$P(K_i=k|Y_i=y_i)=\eta_k\prod_{j=1}^{J}\prod_{c=1}^{C_j}\pi_{jk}(c)^{y_{ij}(c)}/\phi(y_i),$$

where $\phi(y_i)$ is given in equation (A.1). To select the number of classes both the Akaike and Bayesian Information Criteria had been used even though BIC was shown to be superior in studies where the number of observed variables is moderately large Houseman et al., 2006.

In the presence of missing covariate data, our goal is to improve upon the standard multiple imputation technique for missing categorical data by first clustering like observations and then imputing based on latent class. That is, the latent class model is being used as a tool for the estimation of $P(y_i;\eta,\beta)$ as a means for data imputation based on the true

distribution of the observed data but not to obtain interpretable latent classes. We are therefore not concerned with identifiability of parameter estimates or the well known label-switching problem Stephens, 2000. Even though any permutation of latent classes results in several maximum likelihood solutions, $P(y_i;\beta,\eta)$ is uniquely identified.

The latent class model is expressed as a model for the observed data density, $p(y_{i,obs};\eta,\beta)$,

$$p(y_{i,obs};\eta,\beta)=\sum_{k=1}^{K}\eta_k\prod_{j=1}^{J}\{p(y_{ij}|K_i=k;\beta)\}^{r_{ij}}$$

where $r_{ij}=0$ if the value of $y_{ij}$ is missing and 1 otherwise. Note, $r_{ij}$ represents a realization of the missing data indicator, $R_{ij}$, so only variables $j=1,\ldots,J$ that do not have missing values contribute to the estimation of the model. This results in unbiased parameter estimates due to the assumption of conditional independence of variables given latent class assignment, leading to a straightforward strategy for class-based multiple imputation. Once the latent class model is fitted via the EM algorithm, one can easily obtain draws from the distribution of the missing data conditional on the observed data (eg. via MCMC or Bootstrap) and with the conditional independence assumption we get,

$$p(y_{i,mis}|y_{i,obs};\eta,\beta)=\sum_{k=1}^{K}\frac{p(K_i=k)p(y_{i,obs}|K_i=k)}{p(y_{i,obs})}p(y_{i,mis}|K_i=k). \tag{A.2}$$

Since the first part of equation (A.2) is the posterior probability of latent class membership given the observed data, then the distribution of $y_{i,mis}|y_{i,obs}$ can be rewritten as $p(y_{i,mis}|y_{i,obs})=\sum_{k=1}^{K}p(K_i=k|y_{i,obs};\eta,\beta)p(y_{i,mis}|K_i=k)$. Recall that as only the observed data are used to fit the latent class model, $p(y_{i,mis}|K_i=k)$ is equivalent to $\prod_{j=1}^{J}p(y_{ij}|K_i=k)^{1-r_{ij}}$, where $y_{ij}$ are the complete data and $r_{ij}$ the missing data indicator.

## Appendix B. Supplementary data

Supplementary data associated with this article can be found in the online version at doi: 10.1016/j.jspi.2010.04.020.

## References


Agresti, A., 2002. Categorical Data Analysis, second ed. John Wiley & Sons Inc., Hoboken, NJ.

Alfaro, K., 2003. Bayes or bootstrap? A simulation study comparing the performance of Bayesian Markov chain Monte Carlo sampling and bootstrapping in assessing phylogenetic confidence. Molecular Biology and Evolution 20, 255–266.

Allison, P.D., 2006. Multiple imputation of categorical variables under the multivariate normal model. Paper presented at the annual meeting of the American Sociological Association, Montreal Convention Center. Montreal, Quebec, Canada, Aug 11, 2006, http://www.allacademic.com/meta/p102543_index.html.

Bernaards, C., Belin, T., Schafer, J., 2007. Robustness of a multivariate normal approximation for imputation of binary incomplete data. Statistics in Medicine 26, 1368–1382.

Breslow, N., Clayton, D., 1993. Approximate inference in generalized linear mixed models. Journal of the American Statistical Association 88, 9–25.

Breslow, N., Day, N., 1980. Statistical Methods in Cancer Research II: The Analysis of Case–Control Studies. IARC Scientific Publications, Lyon, France.

Cozen, W., Gebregziabher, M., Conti, D., et al., 2006. Interleukin-6 related genotypes, bodymass index and risk of multiple myeloma and plasmacytoma. Cancer Epidemiology, Biomarkers and Prevention 15, 1–7.

Dempster, A., Laird, N., Rubin, D., 1977. Maximum likelihood for incomplete data via the em algorithm. Journal of the Royal Statistical Society, Series B 39, 1–38.

Diggle, P., Liang, K., Zeger, S., 1994. Analysis of Longitudinal Data. Oxford University Press, Oxford, UK.

Fujikawa, Y., Ho, T., 2002. Cluster-based Algorithms for Dealing with Missing Values. Springer, Berlin, Germany.

Gebregziabher, M., Langholz, B., 2009. A semiparametric missing-data-induced method for missing covariate data in individually matched case–control studies. Biometrics, in press, doi:10.1111/j.1541-0420.2009.01322.x.

Godfrey, A., Wood, G., Ganesalingam, S., Nichols, M., Qiao, C., 2002. Two-stage clustering in genotype-by-environment analyses with missing data. Journal of Agricultural Science 139, 67–77.

Goodman, L., 1974. Exploratory latent structure analysis using both identifiable and unidentifiable models. Biometrika 61, 215–231.

Horton, N., Lipsitz, S., Parzen, M., 2003. A potential for bias when rounding in multiple imputation. The American Statistician 57, 229–232.

Houseman, E., Coull, B., Betensky, R., 2006. Feature-specific constrained latent class analysis for medium throughput genomic data. Biometrics 9, 1062–1070.

Joernsten, R., Ouyang, M., Wang, H., 2007. A meta-data based method for dna microarray imputation. Bioinformatics 8.

Langholz, B., 2007. Use of cohort information in the design and analysis of case–control studies. Scandinavian Journal of Statistics 34, 120–136.

Lanza, S.T., Collins, L.M., Lemmon, D.R., Schafer, J.L., 2007. Proc LCA: a SAS procedure for latent class analysis. Structural Equation Modeling 14, 671–694.

Lanza, S.T., Lemmon, D.R., Schafer, J.L., Collins, L.M., 2008. Proc LCA & proc LTA user's guide version 1.1.5 beta URL: 〈methodology.psu.edu/index.php/downloads/proclcalta.html〉.

Little, R., Rubin, D., 2002. Statistical Analysis with Missing Data. John Wiley & Sons Inc, New York, NY.

Oakes, D., 1981. Survival times: aspects of partial likelihood (with discussion). International Statistical Review 49, 235–264.

Rathouz, P., 2003. Likelihood methods for missing covariate data in highly stratified studies. Journal of the Royal Statistical Society, Series B 65, 711–723.

Rubin, D., 1987. Multiple Imputation for Nonresponse in Surveys. John Wiley & Sons Inc, New York, NY.

Schafer, J., 1997a. Analysis of Incomplete Multivariate Data. Chapman & Hall, London, UK.

Schafer, J., 1997b. Imputation of missing covariates under a general linear mixed model. Technical Report, Department of Statistics, Penn State University.

Stephens, M., 2000. Dealing with label switching in mixture models. Journal of the Royal Statistical Society, Series B 62, 795–809.

Thomas, D., 1977. Addendum to the paper by F.D.K Liddel, J.C. McDonald and D.C. Thomas. Journal of the Royal Statistical Society, Series A 140, 483–485.

Vermunt, J., van Ginkel, J., van der Ark, L., Sijtsma, K., 2008. Multiple imputation of incomplete categorical data using latent class analysis. Sociological Methodology 38, 369–397.

Xiang, Q., Edward, J., Gadbury, G., 2006. Interval estimation in a finite mixture model: modeling p values in multiple testing applications. Computational Statistics and Data Analysis 5, 570–586.