Journal of Data Science **10**(2012), 465-481

Direct and Unbiased Multiple Imputation Methods for Missing Values of Categorical Variables

Yuanhui Xiao[1], Ruiguang Song[2*], Mi Chen[2] and H. Irene Hall[2]
[1] *Georgia State University and* [2] *Centers for Disease Control and Prevention*

Abstract: Missing data is a common problem in statistical analyses. To make use of information in data with incomplete observation, missing values can be imputed so that standard statistical methods can be used to analyze the data. Variables with missing values are often categorical and the missing pattern may not be monotone. Currently, commonly used imputation methods for data with a non-monotone missing pattern do not allow direct inclusion of categorical variables. Categorical variables are converted to numerical variables before imputation. For many applications, the imputed numerical values for those categorical variables must then be converted back to categorical values. However, this conversion introduces bias which can seriously affect subsequent analyses. In this paper, we propose two direct imputation methods for categorical variables with a non-monotone missing pattern: the direct imputation approach incorporated with the expectation-maximization algorithm and the direct imputation approach incorporated with a new algorithm: the imputation-maximization algorithm. Simulation studies show that both methods perform better than the method using variable conversion. An application to real data is provided to compare the direct imputation method and the method using variable conversion.

Key words: Bias, categorical variable, HIV, missing values, multiple imputation.

1. Introduction

In many survey and observational data systems, some critical variables may have missing values. These variables can be either numerical or categorical. For example, some demographic variables are categorical and they are important to identify population groups of interest. Although methods for handling or imputing missing values for numerical variables has been extensively discussed in the literature (e.g., Rubin, 1978; 1987; 1996), research on how to impute missing

*Corresponding author.

values in categorical variables is limited, particularly when the missing pattern is not monotone. A dataset with multiple variables is said to have a monotone missing pattern when variables in the dataset can be sorted in an order such that when a variable is missing for an observation, then all subsequent variables are missing for the observation.

There are many methods to handle data with a monotone missing pattern, for example, the propensity score method (Rosenbaum and Rubin, 1983) and the discriminant function method (Brand 1999, pp. 95-96). However, for data with a non-monotone missing pattern, the only method that has been widely used is the Markov Chain Monte Carlo (MCMC) method (Schafer, 1997) although other methods are available, for example, the multiple imputation using chained equations (van Buuren and Oudshoorn, 1999). When the variables with missing values are categorical, imputation of missing values has not been adequately addressed. Schafer (1997) introduced a saturated multinomial model from a Bayesian perspective. Allison (2001) proposed a method using MCMC with each categorical variable expressed by a group of dummy variables. During the imputation process, the dummy variables are treated as numerical variables. After imputation, the imputed numerical values for the dummy variables are converted back to categorical values for the original categorical variables. Although the MCMC method is unbiased under appropriate assumptions, the variable conversion introduces bias (Horton *et al.*, 2003, Ake, 2005, Song *et al.*, 2010). In this paper, we use the saturated multinomial model to impute data with a non-monotone missing pattern of categorical variables and compare the performance of this approach with the one based on the MCMC method.

Our work was motivated by a missing data problem in the US national HIV case surveillance system. The system collects many variables and most of them are categorical. Among the categorical variables is the case's HIV transmission category, which summarizes the multiple risks that the individual may have taken by selecting the one through which HIV was most likely to have been acquired. Transmission category is an important variable that identifies the populations at high risk of HIV infection. However, for a substantial proportion of the HIV cases reported to the Centers for Disease Control and Prevention (CDC) risk factor information is missing so their transmission category is unknown. The proportion of HIV cases with unknown transmission category has been increasing in recent years. In 1994, less than 20% of HIV cases were reported to the CDC without a transmission category, while in 2007, the proportion increased to approximately 40% (Harrison *et al.*, 2008). In addition, there are other numerical and categorical variables with missing values in the HIV surveillance database and the missing pattern of these variables is not monotone.

The data collected by the HIV surveillance system is the cornerstone for

monitoring and characterizing the epidemic and for planning and evaluating HIV-related prevention and care programs at the local, state and national level in the United States (CDC, 2010). Handling the missing data in this database is essential to reducing or controlling the biases in all subsequent analyses. In this paper, we introduce two imputation methods for categorical variables with a non-monotone missing pattern. The two methods use multinomial distributions to model categorical variables and impute missing values directly from multinomial distributions. Therefore, they reduce the bias introduced by the method that converts variables from categorical to numerical before imputation and then from numerical back to categorical after imputation. The proposed methods are described in the next section. To compare the performance of the proposed methods and the method for numerical variables with variable conversions (the MCMC method with variable conversion or simply the MCMC method), we conducted a simulation study. The simulation study design and results are presented in Section 3. We also applied one of the proposed imputation methods to the HIV surveillance database, which has plenty of missing data, and compared results with those derived from the MCMC method.

2. Imputation Methods

Let $\boldsymbol{X} = (X_1, X_2, \cdots, X_r)$ denote the set of all the variables of interest, where r is the total number of variables and all variables are categorical. For an observation of $\boldsymbol{X}$ (called an observed case), say $\boldsymbol{x}$, there are two possible outcomes: no value is missing on any of the r variables (called a complete case or observation) or at least one of the variables has a missing value (called an incomplete case or observation). Let $\boldsymbol{S}$ be the set of all observed cases with or without missing components, $\boldsymbol{S}_k$ the set of all observed complete cases, and $\boldsymbol{S}_m$ the set of all observed incomplete cases. For an incomplete observation $\boldsymbol{x}$ in $\boldsymbol{S}_m$, we denote the set of covariates with known values by $\boldsymbol{X}_k$, and the set of covariates with missing values by $\boldsymbol{X}_m$. If $\boldsymbol{Y}$ is a subset of $(X_1, X_2, \cdots, X_r)$, we denote by $\boldsymbol{x}(\boldsymbol{Y})$ the set of values taken by the variables in $\boldsymbol{Y}$. Also, we denote $\boldsymbol{x}(\boldsymbol{X}_k)$ by $\boldsymbol{x}_k$ and $\boldsymbol{x}(\boldsymbol{X}_m)$ by $\boldsymbol{x}_m$.

Without loss of generality, we use positive integers to represent the levels of each variable and 0 for missing values. Thus, $\boldsymbol{x}(\boldsymbol{Y}) = \boldsymbol{0}$ means that the observation has missing values for all of the variables in $\boldsymbol{Y}$. Our goal is to substitute missing values in all incomplete observations with plausible values so that we have data sets with complete information on every observation for further statistical analyses.

Suppose that the joint probability distribution function $P_{\boldsymbol{X}}(\cdot)$ is known. For an observation $\boldsymbol{x}$ in $\boldsymbol{S}_m$, $\boldsymbol{X} = \boldsymbol{X}_k \cup \boldsymbol{X}_m$. The probability distribution of $\boldsymbol{X}_m$ conditional on $\boldsymbol{X}_k = \boldsymbol{x}_k$ is given by

$$P_{(\boldsymbol{X}_m|\boldsymbol{X}_k)}(\boldsymbol{X}_m = \boldsymbol{x}_m|\boldsymbol{X}_k = \boldsymbol{x}_k) = \frac{P_r(\boldsymbol{X}_m = \boldsymbol{x}_m, \boldsymbol{X}_k = \boldsymbol{x}_k)}{\sum_{\boldsymbol{x}:\boldsymbol{x}(\boldsymbol{X}_k)=\boldsymbol{x}_k} P_{\boldsymbol{X}}(\boldsymbol{x})}$$
$$= \frac{P_{\boldsymbol{X}}(\boldsymbol{x}_m, \boldsymbol{x}_k)}{\sum_{\boldsymbol{x}:\boldsymbol{x}(\boldsymbol{X}_k)=\boldsymbol{x}_k} P_{\boldsymbol{X}}(\boldsymbol{x})} \quad (1)$$

Our imputation method replaces the missing values of $\boldsymbol{x}(\boldsymbol{X}_m)$ with random values generated from the above conditional probability distribution, as follows. (1) Estimate the joint probability distribution function $P_{\boldsymbol{X}}(\cdot)$ using all observed cases with or without missing values. (2) Draw a random sample from the conditional distribution for each observation with missing values. (3) Repeat (1) and (2) to obtain multiple imputed values for each missing value. Note that the joint probability distribution function from (1) is only an estimate. Multiple samples from the same estimate will be correlated. Thus, one should not draw multiple samples at step (2) to obtain multiple imputed values for each missing value.

The joint distribution $P_{\boldsymbol{X}}$ of $\boldsymbol{X}$ is unknown. It has to be estimated from the observed data. The distribution of $\boldsymbol{X}$ can be considered as multinomial. The possible values of $\boldsymbol{X}$ are combinations of all possible values of all variables in $\boldsymbol{X}$. If the dataset has no missing values, then the maximum likelihood estimator (MLE) for $P_{\boldsymbol{X}}$ is the frequency distribution of the data. If the dataset contains missing values and these values are missing not completely at random, then the joint distribution $P_{\boldsymbol{X}}$ of $\boldsymbol{X}$ cannot be estimated by only using data with complete information on all variables. However, if missing values are missing not completely at random but conditionally at random based on the observed data, then we can use all observed data (including observations with missing values) to estimate the joint distribution.

We consider data with a non-monotone missing pattern and missing values are missing conditionally at random. We describe two iterative methods to estimate the joint distribution. The first method is based on the expectation-maximization (EM) algorithm (Dempster *et al.*, 1977, Little and Rubin, 1987), and the second method is based on an imputation-maximization (IM) algorithm described later in this section. The EM algorithm requires that the data are organized in the form of Table 1, where $\boldsymbol{x}_i$, $i = 1, 2, \cdots, n$, are all possible values of $\boldsymbol{X}$, and $f(\boldsymbol{x}_i)$ is the frequency of $\boldsymbol{x}_i$ or the total number of cases with the value $\boldsymbol{x}_i$. The EM algorithm is as follows:

Table 1: Data format

$\boldsymbol{X}$	Frequency
$\boldsymbol{x}_1$	$f(\boldsymbol{x}_1)$
$\boldsymbol{x}_2$	$f(\boldsymbol{x}_2)$
$\vdots$	$\vdots$
$\boldsymbol{x}_n$	$f(\boldsymbol{x}_n)$

1. Posit an initial estimate for the joint distribution $P_{\boldsymbol{X}}(\cdot)$. This can simply be a uniform distribution, or estimated from the complete cases.

2. **E-step**: For an incomplete case with $\boldsymbol{x} = (\boldsymbol{x}_k, \boldsymbol{x}_m)$, estimate the probability distribution of $\boldsymbol{X}_m$ conditional on $\boldsymbol{X}_k = \boldsymbol{x}_k$ by substituting the current estimate of $P_{\boldsymbol{X}}(\cdot)$ into (1). Make a table in the form of Table 1, where the $\boldsymbol{X}$-column contains all combinations of levels of the variables of $\boldsymbol{X}$ satisfying $\boldsymbol{y}(\boldsymbol{X}_k) = \boldsymbol{x}_k$ and the frequencies are the respective estimated conditional probabilities. This is equivalent to splitting the case into several "fractional" cases. (Each row in the table represents a "fractional" case since the frequency is a probability.)

3. **M-step**: Combine all fractional cases from the allocation of incomplete cases in Step 2 with the original complete cases to form a new data set. The new dataset has no missing values and is used to update the current MLE of the joint distribution $P_{\boldsymbol{X}}(\cdot)$.

4. Repeat the E-step and M-step until there is no significant change in the estimate of the joint distribution $P_{\boldsymbol{X}}(\cdot)$.

The EM algorithm allocates fractions of incomplete (or partially classified) cases to form new data sets without missing values according to the estimated conditional distribution. However, we can also randomly allocate the partially classified cases to create a dataset for the next step. This is equivalent to filling in the missing values of an incomplete case by plausible values. So, instead of distributing a case in $\boldsymbol{S}_m$ into multiple fractional cases in Step 3 of the EM algorithm, we distribute the case as a whole into only one of the possible values of $\boldsymbol{X}_m$ based on a draw from the conditional distribution determined by (1). This idea is similar to the one used in imputation so we call it the Imputation-Maximization (IM) algorithm.

Both algorithms will converge if the estimate of the joint distribution of $\boldsymbol{X}$ computed from complete cases does not severely deviate from the true distribution. This is usually the case when there are a reasonable number of complete cases. Further, it is easy to verify that the IM algorithm creates a Markov chain (Schafer, 1997)

$$(\boldsymbol{X}_m^{(1)}, \boldsymbol{\theta}^{(1)}), (\boldsymbol{X}_m^{(2)}, \boldsymbol{\theta}^{(2)}), \cdots,$$

which converges to the conditional distribution of $\boldsymbol{X}_m$ given $\boldsymbol{X}_k$, where $\boldsymbol{\theta}$ is the parameter vector of the joint distribution of $\boldsymbol{X}$. Hence, the IM algorithm is essentially a MCMC type of method.

An extreme case is that, for an incomplete observation $\boldsymbol{x}$ there is no complete observation $\boldsymbol{y}$ with $\boldsymbol{y}(\boldsymbol{X}_k) = \boldsymbol{x}_k$, which implies that the set

$$\{\boldsymbol{y} : \boldsymbol{y}(\boldsymbol{X}_k) = \boldsymbol{x}_k \text{ and } \hat{P}_{\boldsymbol{X}}(\boldsymbol{y}) \neq 0\}$$

is empty, where $\hat{P}_{\boldsymbol{X}}$ refers to any estimate of $P_{\boldsymbol{X}}(\cdot)$ that appears in the computing process of the EM or IM algorithm. Consequently, the conditional probability distribution of $\boldsymbol{X}_m$ given $\boldsymbol{X}_k = \boldsymbol{x}_k$ cannot be estimated from (1) in the iterative process of the EM or IM algorithm. Therefore, this type of incomplete observation must be excluded from the computing process of the EM or IM algorithm for estimating the joint probability distribution of $\boldsymbol{X}$. Fortunately, in practice such observations are rare when the number of covariates is small but the total number of observations is not small. In such cases, we suggest using the following approximation method to estimate the conditional probability distribution of $\boldsymbol{X}_m$ given $\boldsymbol{X}_k = \boldsymbol{x}_k$.

Let $\hat{P}_{\boldsymbol{X}}$ be an estimate of the joint probability distribution of $\boldsymbol{X}$ obtained by the EM or IM algorithm. Let $\boldsymbol{Y}$ be the largest subset of $\boldsymbol{X}_k$ such that the set

$$\{\boldsymbol{y} : \boldsymbol{y}(\boldsymbol{Y}) = \boldsymbol{x}_k(\boldsymbol{Y}) \ \text{ and } \ \hat{P}_{\boldsymbol{X}}(\boldsymbol{y}) \neq 0\}$$

is not empty. Then

$$\hat{P}_{\boldsymbol{Y}}(\boldsymbol{X}_m = \boldsymbol{x}_{m_i} | \boldsymbol{X}_k = \boldsymbol{x}_k) = \frac{\sum_{\boldsymbol{y}:\boldsymbol{y}(\boldsymbol{Y})=\boldsymbol{x}_k(\boldsymbol{Y})} \text{ and } {}_{\boldsymbol{y}(\boldsymbol{X}_k)=\boldsymbol{x}_{m_i}} \hat{P}_{\boldsymbol{X}}(\boldsymbol{y})}{\sum_{\boldsymbol{y}:\boldsymbol{y}(\boldsymbol{Y})=\boldsymbol{x}_k(\boldsymbol{Y})} \hat{P}_{\boldsymbol{X}}(\boldsymbol{y})}$$

is an approximation to the conditional probability of $\boldsymbol{X}_m = \boldsymbol{x}_{m_i}$ given $\boldsymbol{X}_k = \boldsymbol{x}_k$ and can be used for imputing missing values. If the largest subset $\boldsymbol{Y}$ is not unique, then we estimate the conditional probability distribution of $\boldsymbol{X}_m$ given $\boldsymbol{X}_k = \boldsymbol{x}_k$ by the average of the probability distributions $\hat{P}_{\boldsymbol{Y}}(\cdot | \boldsymbol{X}_k = \boldsymbol{x}_k)$ and use this distribution to fill in the missing values.

3. Simulation Studies

We have proposed two direct imputation methods, one based on the EM algorithm and the other based on the IM algorithm. To evaluate the performance of the two methods, we conducted two simulation studies. The general design of the simulation studies is as follows. Given the joint distribution of a set of categorical variables, we simulate 1,000 random samples, each with 1,000 cases from the given joint distribution. For each simulated random sample, a number of cases were selected. For each selected case, some variables' values were removed and 10 plausible values were imputed for each missing value using each of the three imputation methods: the MCMC method with variable conversion and the two proposed direct imputation methods. Ten datasets with complete information were generated. Each data set was analyzed to estimate the joint and marginal distributions of the given categorical variables. Finally, estimates for each probability (p) from the 10 imputed data sets were combined by using the

standard multiple imputation procedure. The following statistics were computed from the 1,000 simulated random samples for each imputation method.

- the average of the estimates for p: $\hat{p}$,
- the bias (average of $(\hat{p} - p)$),
- the relative bias (r-bias, average of $100 \times (\hat{p} - p)/p$),
- the average length (AveLen) and the coverage rate (CR) of the 95% confidence intervals.

The details of the two simulation studies are described in the next two subsections.

3.1 Simulation Study I: Single Variable with Missing Value

We first consider a simple missing data problem where there are only two variables in the data set, both categorical, and only one can have missing values. Call these two variables X and Y. See Table 2 for the joint and marginal probability distributions of X and Y. Suppose that the probability of missing Y is 0.35 if $X = 1, 0.25$ if $X = 2$, and 0.20 if $X = 3$. The missing value probabilities and distribution of X and Y are chosen to match those of AIDS cases among males with age $\geq$ 13 at AIDS diagnosis: X is the variable for race/ethnicity groups: non-Hispanic Blacks, Hispanics, and non-Hispanic Whites, distributed as 45%, 20% and 35%, respectively. Y represents the variable for transmission categories: men who have sex with men (MSM) only, injection drug use (IDU) only, MSM and IDU, and high-risk heterosexual contact. The percentages for each of the four groups are 65%, 10%, 5% and 20% for non-Hispanic Blacks, 65%, 15%, 5% and 15% for Hispanics, and 85%, 5%, 5% and 5% for non-Hispanic Whites, respectively.

Table 2: Joint and marginal distributions of X and Y in Simulation Study I

	Y				
X	1	2	3	4	Total
1	0.2925	0.0450	0.0225	0.0900	0.45
2	0.1300	0.0300	0.0100	0.0300	0.20
3	0.2975	0.0175	0.0175	0.0175	0.35
Total	0.7200	0.0925	0.0500	0.1375	1.00

3.2 Simulation Study II: Two Variables with Missing Value

Now, we consider a more complex missing data problem where several variables in the data could have missing values and the missing pattern is non-monotone. Suppose that there are three categorical variables: X, Y, and Z with the joint probability distribution specified in Table 3. We note that the variables Y and Z result from splitting the variable Y in the previous simulation study. We assume that Y and Z could have missing values, but X does not have any missing values. Similar to simulation study I, the probability of missing either Y or Z depends on X with the probability 0.35 when $X = 1, 0.25$ when $X = 2$, and 0.20 when $X = 3$. Among those with either Y or Z missing, the probability of missing Y only is 0.50, missing Z only is 0.30, and missing both Y and Z is 0.20.

Table 3: Joint and marginal distributions of X, Y, and Z in Simulation Study II

	$Y=1$		$Y=2$		
X	$Z=1$	$Z=2$	$Z=1$	$Z=2$	Total
1	0.2925	0.0450	0.0225	0.0900	0.45
2	0.1300	0.0300	0.0100	0.0300	0.20
3	0.2975	0.0175	0.0175	0.0175	0.35
Total	0.7200	0.0925	0.0500	0.1375	1.00
	$P(Y=1)=0.8125$		$P(Y=2)=0.1875$		
	$P(Z=1)=0.7700$		$P(Z=2)=0.2300$		

3.3 Simulation Results

In the analysis of the simulated data, we first ignored the incomplete or partially classified cases and used only the complete cases to estimate the joint and marginal distributions under each simulation setting as this approach is common practice. We then applied the MCMC method and the two direct imputation methods to impute missing values and used them to estimate the joint and marginal distributions. Results for missing values in a single variable are presented in Tables 4 (joint distribution) and 5 (marginal distribution). Results for missing values in two variables with a non-monotone missing pattern are presented in Tables 6 (joint distribution) and 7 (marginal distribution).

In both simulation settings, the estimates based on complete cases only are seriously biased. This is because missing value in both settings is not completely at random. This bias was removed or reduced by imputing values for data with incomplete information.

The biases associated with the complete case analysis are significantly reduced, but not completely eliminated by the MCMC method (Tables 4-7). The remaining bias is caused by the conversion from the imputed numerical values back to the categorical values. In Simulation Study I, the largest relative bias is

Table 4: Simulation Study I: Estimated joint probability distribution of (X, Y)

X	Y	p	$\hat{p}$	bias	r-bias	AveLen	CR
		No imputation (using only the complete cases)					
1	1	0.2925	0.2636	-0.0289	-9.9%	0.064	56.2%
1	2	0.0450	0.0404	-0.0046	-10.3%	0.029	86.5%
1	3	0.0225	0.0202	-0.0023	-10.2%	0.020	87.7%
1	4	0.0900	0.0806	-0.0094	-10.5%	0.040	80.7%
2	1	0.1300	0.1351	0.0051	3.9%	0.050	93.2%
2	2	0.0300	0.0311	0.0011	3.6%	0.025	94.3%
2	3	0.0100	0.0101	0.0001	1.2%	0.014	92.4%
2	4	0.0300	0.0309	0.0009	2.9%	0.025	95.3%
3	1	0.2975	0.3295	0.0320	10.7%	0.069	54.6%
3	2	0.0175	0.0194	0.0019	10.8%	0.020	93.7%
3	3	0.0175	0.0195	0.0020	11.5%	0.020	93.8%
3	4	0.0175	0.0198	0.0023	13.0%	0.020	94.2%
		Imputation based on MCMC method					
1	1	0.2925	0.2970	0.0045	1.5 %	0.063	93.9%
1	2	0.0450	0.0444	-0.0006	-1.3 %	0.031	91.6%
1	3	0.0225	0.0198	-0.0027	-12.0 %	0.020	85.4%
1	4	0.0900	0.0891	-0.0009	-1.0 %	0.042	91.3%
2	1	0.1300	0.1313	0.0013	1.0 %	0.044	93.4%
2	2	0.0300	0.0287	-0.0013	-4.4 %	0.023	91.6%
2	3	0.0100	0.0090	-0.0010	-10.0 %	0.013	89.8%
2	4	0.0300	0.0305	0.0004	1.5 %	0.024	95.8%
3	1	0.2975	0.2954	-0.0021	-0.7 %	0.058	93.8%
3	2	0.0175	0.0183	0.0008	4.5 %	0.019	95.0%
3	3	0.0175	0.0160	-0.0015	-8.8 %	0.016	88.3%
3	4	0.0175	0.0208	0.0032	18.6 %	0.021	95.9%
		Direct imputation incorporated with EM algorithm					
1	1	0.2925	0.2923	-0.0002	-0.1%	0.062	93.5%
1	2	0.0450	0.0450	0.0000	0.0%	0.030	92.9%
1	3	0.0225	0.0225	0.0000	0.2%	0.022	92.7%
1	4	0.0900	0.0899	-0.0001	-0.1%	0.041	93.3%
2	1	0.1300	0.1299	-0.0001	-0.1%	0.044	94.4%
2	2	0.0300	0.0300	0.0000	-0.1%	0.024	92.7%
2	3	0.0100	0.0101	0.0001	0.7%	0.014	91.2%
2	4	0.0300	0.0299	-0.0001	-0.3%	0.024	93.4%
3	1	0.2975	0.2978	0.0003	0.1%	0.058	94.1%
3	2	0.0175	0.0175	0.0000	0.1%	0.018	92.7%
3	3	0.0175	0.0175	0.0000	0.1%	0.018	93.0%
3	4	0.0175	0.0175	0.0000	0.0%	0.018	93.3%
		Direct imputation incorporated with IM algorithm					
1	1	0.2925	0.2930	0.0005	0.2%	0.062	93.9%
1	2	0.0450	0.0448	-0.0002	-0.3%	0.030	93.6%
1	3	0.0225	0.0223	-0.0002	-1.0%	0.021	91.6%
1	4	0.0900	0.0900	0.0000	0.0%	0.041	93.3%

Table 4: (continued) Simulation Study I: Estimated joint probability distribution of (X, Y)

X	Y	p	$\hat{p}$	bias	r-bias	AveLen	CR
		Direct imputation incorporated with IM algorithm					
2	1	0.1300	0.1298	-0.0002	-0.2%	0.044	94.0%
2	2	0.0300	0.0300	0.0000	0.0%	0.024	93.2%
2	3	0.0100	0.0100	0.0000	0.2%	0.014	91.7%
2	4	0.0300	0.0302	0.0002	0.6%	0.024	93.5%
3	1	0.2975	0.2976	0.0001	0.0%	0.058	95.3%
3	2	0.0175	0.0174	-0.0001	-0.4%	0.018	93.2%
3	3	0.0175	0.0173	-0.0002	-0.9%	0.018	93.9%
3	4	0.0175	0.0175	0.0000	-0.1%	0.018	92.9%

Table 5: Simulation Study I: Estimated marginal probability distributions of X and Y

	p	$\hat{p}$	bias	r-bias	AveLen	CR
	No imputation (using only the complete cases)					
$X = 1$	0.4500	0.4047	-0.0453	-10.1%	0.072	30.9%
2	0.2000	0.2071	0.0071	3.6%	0.059	91.7%
3	0.3500	0.3881	0.0381	10.9%	0.071	44.9%
$Y = 1$	0.7200	0.7281	0.0081	1.1%	0.065	91.4%
2	0.0925	0.0908	-0.0017	-1.8%	0.042	92.2%
3	0.0500	0.0498	-0.0002	-0.4%	0.032	95.0%
4	0.1375	0.1312	-0.0063	-4.6%	0.049	90.3%
	Imputation based on MCMC method					
$X = 1$	0.4500	0.4503	0.0003	0.1%	0.062	94.7%
2	0.2000	0.1994	-0.0006	-0.3%	0.049	93.8%
3	0.3500	0.3504	0.0004	0.1%	0.059	94.2%
$Y = 1$	0.7200	0.7236	0.0036	0.5%	0.066	92.4%
2	0.0925	0.0914	-0.0011	-1.2%	0.042	88.3%
3	0.0500	0.0448	-0.0052	-10.5%	0.029	79.1%
4	0.1375	0.1403	0.0028	2.0%	0.052	91.2%
	Direct imputation incorporated with EM algorithm					
$X = 1$	0.4500	0.4498	-0.0002	0.0%	0.062	94.3%
2	0.2000	0.1998	-0.0002	-0.1%	0.050	94.6%
3	0.3500	0.3503	0.0003	0.1%	0.059	94.4%
$Y = 1$	0.7200	0.7200	0.0000	0.0%	0.064	93.7%
2	0.0925	0.0925	0.0000	0.0%	0.041	93.9%
3	0.0500	0.0502	0.0002	0.3%	0.031	93.6%
4	0.1375	0.1374	-0.0001	-0.1%	0.049	93.9%
	Direct imputation incorporated with IM algorithm					
$X = 1$	0.4500	0.4501	0.0001	0.0%	0.062	95.0%
2	0.2000	0.2000	0.0000	0.0%	0.050	94.4%
3	0.3500	0.3499	-0.0001	0.0%	0.059	95.0%
$Y = 1$	0.7200	0.7204	0.0004	0.1%	0.064	94.5%
2	0.0925	0.0923	-0.0002	-0.2%	0.041	93.4%
3	0.0500	0.0496	-0.0004	-0.8%	0.031	93.3%
4	0.1375	0.1377	0.0002	0.1%	0.049	92.9%

Table 6: Simulation Study II: Estimated joint probability distribution of (X, Y, Z)

X	Y	Z	p	$\hat{p}$	bias	r-bias	AveLen	CR
			No imputation (using only the complete cases)					
1	1	1	0.2925	0.2643	-0.0282	-9.7%	0.064	59.4%
1	1	2	0.0450	0.0404	-0.0046	-10.2%	0.029	87.3%
1	2	1	0.0225	0.0202	-0.0023	-10.3%	0.020	87.3%
1	2	2	0.0900	0.0811	-0.0089	-9.9%	0.040	83.0%
2	1	1	0.1300	0.1347	0.0047	3.6%	0.050	94.2%
2	1	2	0.0300	0.0311	0.0011	3.5%	0.025	95.3%
2	2	1	0.0100	0.0102	0.0002	1.8%	0.014	92.1%
2	2	2	0.0300	0.0307	0.0007	2.4%	0.025	95.6%
3	1	1	0.2975	0.3289	0.0314	10.5%	0.069	55.8%
3	1	2	0.0175	0.0196	0.0021	12.0%	0.020	94.2%
3	2	1	0.0175	0.0194	0.0019	10.6%	0.020	93.9%
3	2	2	0.0175	0.0196	0.0021	11.9%	0.020	95.1%
			Imputation based on MCMC method					
1	1	1	0.2925	0.2889	-0.0036	-1.2%	0.059	94.3%
1	1	2	0.0450	0.0480	0.0029	6.5%	0.031	95.8%
1	2	1	0.0225	0.0260	0.0035	15.8%	0.024	95.6%
1	2	2	0.0900	0.0874	-0.0026	-2.9%	0.038	93.4%
2	1	1	0.1300	0.1294	-0.0006	-0.5%	0.043	93.1%
2	1	2	0.0300	0.0295	-0.0005	-1.6%	0.023	94.5%
2	2	1	0.0100	0.0101	0.0001	1.2%	0.014	96.2%
2	2	2	0.0300	0.0303	0.0003	1.1%	0.023	96.3%
3	1	1	0.2975	0.2947	-0.0028	-0.9%	0.057	94.1%
3	1	2	0.0175	0.0187	0.0012	6.9%	0.019	96.1%
3	2	1	0.0175	0.0183	0.0008	4.7%	0.019	96.5%
3	2	2	0.0175	0.0186	0.0011	6.4%	0.018	97.2%
			Direct imputation incorporated with EM algorithm					
1	1	1	0.2925	0.2927	0.0002	0.1%	0.058	94.9%
1	1	2	0.0450	0.0450	0.0000	0.1%	0.029	93.0%
1	2	1	0.0225	0.0224	-0.0001	-0.3%	0.021	91.2%
1	2	2	0.0900	0.0899	-0.0001	-0.1%	0.039	93.8%
2	1	1	0.1300	0.1303	0.0003	0.2%	0.043	94.9%
2	1	2	0.0300	0.0300	0.0000	0.1%	0.023	93.5%
2	2	1	0.0100	0.0100	0.0000	0.4%	0.014	91.4%
2	2	2	0.0300	0.0299	-0.0001	-0.3%	0.023	93.7%
3	1	1	0.2975	0.2972	-0.0003	-0.1%	0.057	95.0%
3	1	2	0.0175	0.0175	0.0000	0.0%	0.017	92.9%
3	2	1	0.0175	0.0174	-0.0001	-0.3%	0.018	92.6%
3	2	2	0.0175	0.0175	0.0000	0.1%	0.017	93.5%
			Direct imputation incorporated with IM algorithm					
1	1	1	0.2925	0.2925	0.0000	0.0%	0.058	94.4%
1	1	2	0.0450	0.0449	-0.0001	-0.3%	0.029	92.8%
1	2	1	0.0225	0.0226	0.0001	0.4%	0.022	92.2%
1	2	2	0.0900	0.0900	0.0000	0.0%	0.039	93.9%

Table 6: (continued) Simulation Study II: Estimated joint probability distribution of (X, Y, Z)

X	Y	Z	p	$\hat{p}$	bias	r-bias	AveLen	CR
			Direct imputation incorporated with IM algorithm					
2	1	1	0.1300	0.1301	0.0001	0.1%	0.043	94.8%
2	1	2	0.0300	0.0298	-0.0002	-0.6%	0.023	93.1%
2	2	1	0.0100	0.0099	-0.0001	-0.5%	0.014	91.1%
2	2	2	0.0300	0.0302	0.0002	0.6%	0.023	93.5%
3	1	1	0.2975	0.2975	0.0000	0.0%	0.057	94.9%
3	1	2	0.0175	0.0175	0.0000	0.1%	0.017	93.3%
3	2	1	0.0175	0.0174	-0.0001	-0.4%	0.018	92.8%
3	2	2	0.0175	0.0175	0.0000	0.1%	0.017	93.1%

Table 7: Simulation Study II: Estimated marginal probability distributions of X, Y, and Z

	p	$\hat{p}$	bias	r-bias	AveLen	CR
	No imputation (using only the complete cases)					
$X = 1$	0.4500	0.4059	-0.0441	-9.8%	0.072	31.8%
2	0.2000	0.2066	0.0066	3.3%	0.059	94.5%
3	0.3500	0.3874	0.0374	10.7%	0.071	43.8%
$Y = 1$	0.8125	0.8189	0.0064	0.8%	0.056	91.0%
2	0.1875	0.1811	-0.0064	-3.4%	0.056	91.0%
$Z = 1$	0.7700	0.7775	0.0075	1.0%	0.061	91.7%
2	0.2300	0.2225	-0.0075	-3.3%	0.061	91.7%
	Imputation based on MCMC method					
$X = 1$	0.4500	0.4503	0.0003	0.1%	0.062	94.7%
2	0.2000	0.1994	-0.0006	-0.3%	0.049	93.8%
3	0.3500	0.3504	0.0004	0.1%	0.059	94.2%
$Y = 1$	0.8125	0.8091	-0.0034	-0.4%	0.054	94.9%
2	0.1875	0.1909	0.0034	1.8%	0.054	94.9%
$Z = 1$	0.7700	0.7675	-0.0025	-0.3%	0.056	95.2%
2	0.2300	0.2325	0.0025	1.1%	0.056	95.2%
	Direct imputation incorporated with EM algorithm					
$X = 1$	0.4500	0.4501	0.0001	0.0%	0.062	94.8%
2	0.2000	0.2002	0.0002	0.1%	0.050	94.6%
3	0.3500	0.3497	-0.0003	-0.1%	0.059	95.2%
$Y = 1$	0.8125	0.8127	0.0002	0.0%	0.054	93.4%
2	0.1875	0.1873	-0.0002	-0.1%	0.054	93.4%
$Z = 1$	0.7700	0.7701	0.0001	0.0%	0.054	94.1%
2	0.2300	0.2299	-0.0001	0.0%	0.054	94.1%
	Direct imputation incorporated with IM algorithm					
$X = 1$	0.4500	0.4500	0.0000	0.0%	0.062	94.6%
2	0.2000	0.2001	0.0001	0.0%	0.050	94.2%
3	0.3500	0.3499	-0.0001	0.0%	0.059	94.6%
$Y = 1$	0.8125	0.8123	-0.0002	0.0%	0.054	93.6%
2	0.1875	0.1877	0.0002	0.1%	0.054	93.6%
$Z = 1$	0.7700	0.7700	0.0000	0.0%	0.054	95.3%
2	0.2300	0.2300	0.0000	0.0%	0.054	95.3%

18.6%, which occurred when $X = 3$ and $Y = 4$ while in Simulation Study II, the largest relative bias is 15.8%, which occurred when $X = 1$, $Y = 2$ and $Z = 1$. The method for estimating a proportion by the corresponding sample proportion is less efficient when the proportion is small so the bias introduced by variable conversion is in general more severe in this case. However, in both simulation studies, all the relative biases are less than one percent by the direct imputation methods. In fact, the simulation results show that the direct methods are almost unbiased. This is expected since the direct imputation methods bypass variable conversion, an important source for biases.

From Simulation Study I, we see that the confidence intervals perform comparably for the three methods (Tables 4-5). However, from Simulation Study II, we see that the coverage probabilities of the 95% confidence intervals based on the MCMC method are higher when the probability p is low (Tables 6-7). The computation of the confidence intervals is based on a normal approximation, which is poor when the probability of interest is close to zero or one. Thus, we computed the average lengths and coverage rates from 1,000 simulated random samples, each with 1,000 cases, without any missing values. Results from the two direct methods are similar to those observed for simulated random samples without missing values. So, it seems that the direct methods would create imputed data sets which are closer to samples from the correct distribution (results not shown).

Finally, as noted, the variable Y in Study I is split into two variables in Study II (Y and Z). When the variable Y in Study I is missing, then one of the two variables Y or Z in Study II must be missing, but they need not both be missing. Therefore, more information is lost in study I than in study II. This explains why the MCMC approach is not as biased in Study II.

4. Application

In this section, we apply both the direct IM-imputation method and the MCMC method to imputing missing values of categorical variables in the US national HIV case surveillance data. This dataset contains all AIDS cases diagnosed since the early 1980s and reported to CDC by June of 2009. We examined only AIDS cases diagnosed between 2000 through 2007. There were about 300,000 adults and adolescents (aged 13 years or older) diagnosed with AIDS during this time period. Among them 21% had unknown transmission category (19% among males and 29% among females).

Having a missing value for transmission category was correlated with many variables. They include, but are not limited to, sex, race/ethnicity, birth country, age at diagnosis, geographic region of residence at diagnosis, population size of metropolitan statistical area (MSA) of residence, year of diagnosis, and the type

of facility where the person was diagnosed. In addition to missing transmission category, other variables could be missing. For example, 14% of AIDS cases had missing birth country and 18% had missing diagnosis facility type. Both variables are correlated with the transmission category variable. Also, the missing pattern of these variables is not monotone.

Note that except for age, year of diagnosis, and population size of MSA of residence, all the above variables are categorical. Population size is reported as one of a small number of categories in the surveillance database, and for presentation and analysis purposes, we have likewise used age groups and time intervals to categorize these two variables into a small number of categories. With all other variables being categorical, the distribution of AIDS cases stratified by these variables can be viewed as multinomial. Therefore, missing values of these variables can be imputed using the methods proposed in this paper.

Based on the proportion of data with missing values, we choose to impute 10 plausible values for each missing value to achieve a 95% relative efficiency (Rubin 1987, p. 114). Because the numbers of transmission categories for males and females are different, we performed the imputation separately for males and females. The relative differences between the two methods in estimating the distribution probabilities ranged from -5.3% to 2.9%. The difference is significant for both female transmission category proportions and for the male heterosexual proportion. The direct method estimate is higher for the heterosexual proportion in both male and female population groups (Table 8).

5. Summary

In this paper, we evaluated two direct approaches for filling in missing values of categorical variables, which were motivated by the problem of imputing missing HIV transmission category in the national HIV case surveillance database. The methods described here are applicable to any situation where only categorical data are involved. The MCMC method often assumes that the joint distribution of the variables under consideration is multivariate normal, which may not be true in practice. In contrast, direct imputation approaches impose no assumption on the joint distribution of a set of categorical variables. However, the natural distribution for a discrete variable is multinomial. Hence, one may consider that the joint distribution is multinomial. This is another attractive feature of the proposed methods.

Note that both IM and EM algorithms perform equally well and their differences are hardly distinguishable. In addition, as they are almost equally efficient in terms of computing time and both algorithms are fast, using either one of them should be practical.

Table 8: Estimated distributions of transmission categories among AIDS cases diagnosed in the United States, 2000-2007, based on the MCMC and the direct IM-imputation methods

Sex	Transmission Category	n	p	Missing Proportion	MCMC			Direct IM Imputation			Difference	
					p	L95	U95	p	L95	U95	Absolute	Relative
Male												
	MSM	107586	60.0%		58.8%	58.6%	59.1%	58.6%	58.3%	58.8%	-0.3%	-0.5%
	IDU	32525	18.1%		18.9%	18.7%	19.1%	18.6%	18.5%	18.8%	-0.3%	-1.3%
	MSM&IDU	13419	7.5%		7.1%	6.9%	7.2%	7.2%	7.1%	7.3%	0.1%	1.3%
	HRH	25831	14.4%		15.2%	15.0%	15.4%	15.6%	15.5%	15.8%	0.4%	2.9%
	Subtotal	179361										
	Missing	41335		18.7%								
	Total	220696										
Female												
	IDU	16698	29.7%		30.2%	29.9%	30.6%	28.6%	28.1%	29.1%	-1.6%	-5.3%
	HRH	39594	70.3%		69.8%	69.4%	70.1%	71.4%	70.9%	71.9%	1.6%	2.3%
	Subtotal	56292										
	Missing	23028		29.0%								
	Total	79320										

MSM, men who have sex with men; IDU, injection drug use; HRH, High-risk heterosexual contact.

Acknowledgements

The research is supported by the Research Initial Grant of Georgia State University, Atlanta, Georgia, USA. The authors thank the staff who collected and maintained the HIV surveillance data used herein. The authors also would like to thank Timothy A. Green, Ph.D., Lillian S. Lin, Ph.D., and Marie Morgan of CDC for their helpful comments and editorial suggestions, which led to a significant improvement of the original manuscript.

References

Ake, C. F. (2005). Rounding after multiple imputation with non-binary categorical covariates. Paper presented at the 30th Annual Conference of SAS Users Group International, Philadelphia, Pennsylvania. http://www2.sas.com/proceedings/sugi30/112-30.pdf.

Allison, P. D. (2001). *Missing Data.* Thousand Oaks, Sage Publications, California.

Brand, J. P. L. (1999). *Development, Implementation and Evaluation of Multiple Imputation Strategies for the Statistical Analysis of Incomplete Data Sets.* Ph.D. Dissertation, Erasmus University, Rotterdam.

Centers for Disease Control and Prevention. (2010). *HIV Surveillance Report, 2008*; vol. 20. http://www.cdc.gov/hiv/topics/surveillance/ resources/reports/.

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

Harrison, K. M., Kajese, T., Hall, H. I. and Song, R. (2008). Risk factor redistribution of the national HIV/AIDS surveillance data: an alternative approach. *Public Health Reports* **123**, 618-627.

Horton, N. J., Lipsitz, S. R. and Parzen, M. (2003). A potential for bias when rounding in multiple imputation. *American Statistician* **57**, 229-232.

Little, R. J. A. and Rubin, D. B. (1987). *Statistical Analysis with Missing Data.* John Wiley Sons, Inc., New York.

Rubin, D. B. (1978). Multiple imputations in sample surveys. In *Proceedings of the Survey Research Methods*: *Section of the American Statistical Association*, 20-34.

Rubin, D. B. (1987). *Multiple Imputation for Nonresponse in Surveys.* John Wiley Sons, Inc., New York.

Rubin, D. B. (1996). Multiple imputation after 18+ years. *Journal of the American Statistical Association* **91**, 473-489.

Rosenbaum, P. R. and Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika* **70**, 41-55.

Schafer, J. L. (1997). *Analysis of Incomplete Multivariate Data.* Chapman and Hall, New York.

Song, R., McDavid Harrison, K., Hanson, D. and Hall, I. H. (2010). Correction of bias in imputing missing values of categorical variables. *Communications in Statistics - Theory and Methods* **39**, 350-362.

Van Buuren, S. and Oudshoorn, C. G. M. (1999). Flexible multivariate imputation by MICE. *Technical Report.* TNO Preventie en Gezondheid, TNO/VGZ /PG 99.054, Leiden.

Received October 17, 2011; accepted February 22, 2012.

Yuanhui Xiao
Department of Mathematics and Statistics
Georgia State University
Atlanta, GA 30303, USA
yxiao@gsu.edu

Ruiguang Song
Division of HIV/AIDS Prevention
Centers for Disease Control and Prevention
Atlanta, GA 30333, USA
RSong@cdc.gov

Mi Chen
Division of HIV/AIDS Prevention
Centers for Disease Control and Prevention
Atlanta, GA 30333, USA
MChen2@cdc.gov

H. Irene Hall
Division of HIV/AIDS Prevention
Centers for Disease Control and Prevention
Atlanta, GA 30333, USA
IHall1@cdc.gov