A Comparative Study of Imputation Methods for Multivariate Ordinal Data

Chayut Wongkamthong
Social Science Research Institute,
Duke University
and
Olanrewaju Akande
Social Science Research Institute, and
Department of Statistical Science,
Duke University

November 26, 2020

Abstract

Missing data remains a very common problem in large datasets, including survey and census data containing many ordinal responses, such as political polls and opinion surveys. Multiple imputation (MI) is usually the go-to approach for analyzing such incomplete datasets, and there are indeed several implementations of MI, including methods using generalized linear models, tree-based models, and Bayesian non-parametric models. However, there is limited research on the statistical performance of these methods for multivariate ordinal data. In this article, we perform an empirical evaluation of several MI methods, including MI by chained equations (MICE) using multinomial logistic regression models, MICE using proportional odds logistic regression models, MICE using classification and regression trees, MICE using random forest, MI using Dirichlet process (DP) mixtures of products of multinomial distributions, and MI using DP mixtures of multivariate normal distributions. We evaluate the methods using simulation studies based on ordinal variables selected from the 2018 American Community Survey (ACS). Under our simulation settings, the results suggest that MI using proportional odds logistic regression models, classification and regression trees and DP mixtures of multinomial distributions generally outperform the other methods. In certain settings, MI using multinomial logistic regression models and DP mixtures of multivariate normal distributions, are able to achieve comparable performance, depending on the missing data mechanism and amount of missing data.

Keywords: Multiple imputation; Missing data; Mixtures; Nonresponse; Tree methods

1 Introduction

Even as analysts and data scientists continue to gain access to larger and more complex datasets than ever before, in most fields, missing data remains a crucial problem. In large surveys for example, data is usually missing because participants often do not provide answers to some questions. In fact, many surveys have seen steep declines in response rates (Pielsticker & Hiebl 2020, Saunders et al. 2015, Brick & Williams 2013, Curtin et al. 2005). It is well known that ignoring missing data can often lead to biased inference and a reduction in efficiency, particularly when the observed data and missing data are systematically different (Rubin 1976, Little & Rubin 2019). Analysts must therefore handle the missing data problem before or while analyzing such incomplete datasets.

A typical approach for handling missing data is multiple imputation (MI, Rubin (1987)). In MI, one creates multiple completed versions of the sample dataset by replacing the missing entries with plausible draws from predictive distributions of the variables with missing values. These predictions are usually generated either through (i) joint modeling (JM), where one first specifies a multivariate distribution for all the variables and then predicts the missing values using the implied conditional distributions, or (ii) a fully conditional specification (FCS, van Buuren et al. (2006)), where one specifies a univariate model for each incomplete variable conditional on other variables, and uses the estimated univariate models to predict the missing values, without first forming a proper joint distribution. Although the latter lacks the theoretical foundation of the former (Li et al. 2014), it is often preferred in practice because of its simplicity and flexibility.

Several authors have developed several imputation engines under the JM approach (Schafer 1997, Honaker et al. 2012, Yuan 2011). Bayesian nonparametric models in particular, including Dirichlet process (DP) mixture models, have shown great promise as flexible imputation engines. When dealing with multivariate categorical data for example, Si & Reiter (2013) and Manrique-Vallier & Reiter (2014*a*,*b*) use a DP mixture of products of

multinomial distributions (DPMPM). When dealing specifically with multivariate ordinal data, one can also use a DP mixture of multivariate normal distributions (DPMMVN, DeYoreo & Kottas (2014)) model, which uses latent variables with cut-offs to represent the ordinal variables. The DPMMVN model has an added advantage of being able to account for continuous covariates, when needed.

For the FCS approach, most analysts often use multiple imputation by chained equations (MICE, Buuren & Groothuis-Oudshoorn (2011)). MICE is usually implemented by specifying generalized linear models (GLMs) for the univariate conditional (predictive) distributions (Raghunathan et al. 2000). For ordinal data, the most common GLMs used for MICE are multinomial or polytomous logistic regression models, and proportional odds or cumulative logistic regression models. For high dimensional datasets, one major drawback of MICE using GLMs is that including many interaction effects in each univariate model can be tedious and computationally intensive (White et al. 2011). Thus, more recent implementations of MICE instead use machine learning and tree-based models as the conditional distributions. Two of the most popular tree-based methods for imputing multivariate categorical data, are classification and regression trees (CART, Breiman et al. (1984), Burgette & Reiter (2010), Doove et al. (2014)) and random forests (Breiman 2001, Shah et al. 2014). Recently, other machine learning methods such as GAIN (Yoon et al. 2018), have been proposed as potential solutions to some of the drawbacks of the other methods, including computational costs. GAIN uses generative adversarial networks (GAN) to create a framework for missing data imputation. GAIN has been shown to do very well, in terms of overall predictive performance, in several machine learning datasets. While GAIN itself is not a main focus of this article, we include it in many of the simulation scenarios included in the supplementary material, to provide some perspective on whether or not GAIN does provide an improvement on CART and random forests. Overall, imputation methods based on machine learning provide flexibility in model specification since they do not rely heavily on parametric assumptions and are usually less computationally intensive.

There have been many empirical comparisons of several MI methods for multivariate continuous data, multivariate nominal data, or both (Akande et al. 2017, Diaz-Ordaz et al. 2014, Buuren 2007, Laaksonen 2016). However, there are no rigorous empirical comparisons of MI methods for multivariate ordinal data. Specifically, to the best of our knowledge, there are no studies that provide practical guidance on potential advantages and disadvantages of using nominal models over the ordinal versions, when dealing with ordinal data. Also, recent studies comparing the performance of newer machine learning imputation methods to novel methods like MICE (Yoon et al. 2018, Stekhoven & Bühlmann 2012, Shah et al. 2014, Bertsimas et al. 2018) have often focused on metrics that focus on overall predictive power, for example, normalized root mean squared error (NRMSE) for continuous missing values, and proportion of falsely classified entries (PFC) for categorical missing values (Stekhoven & Bühlmann 2012, Yoon et al. 2018, Nazábal et al. 2020, Camino et al. 2019). Specifically, most have prioritized evaluating overall accuracy of the predictions, without extensive considerations of how the models preserve relationships in the data and properly account for uncertainty. Drawing conclusions on such empirical comparisons without fully accounting for the uncertainties, can be biased and misleading (Shah et al. 2014).

In this article, we compare six main MI methods for multivariate ordinal data, including (i) MICE using multinomial logistic regression models (MI-Multireg), (ii) MICE using proportional odds logistic regression models (MI-Polr), (iii) MICE using CART (MI-Cart), (iv) MICE using random forests (MI-Forest), (v) MI using DP mixtures of products of multinomial distributions (MI-DPMPM), and (vi) MI using DP mixtures of multivariate normal distributions (MI-DPMMVN). We compare all methods using evaluation metrics that assess how each method preserves and recovers both marginal features and joint relationships in the data. We base our empirical comparisons on hypothetical populations comprising data from the 2018 American Community Survey (ACS).

The remainder of this article is organized as follows. In section 2, we provide a review of the different imputation methods we assess in this study. In section 3, we describe the

ACS data, and the framework for our simulation study. We also define the evaluative metrics used to assess the quality of the imputed datasets under the methods. In section 4, we present the results on the simulation studies and discuss the performance of the different methods. In Section 5, we conclude with general takeaways and implementation considerations.

2 MI for Ordinal Data

We first present the notation for describing MI for multivariate ordinal data. Let $\boldsymbol{Y} = (\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_n)$ denote the $n \times p$ matrix containing the data for all n observations across p variables. For each observation i, $\boldsymbol{Y}_i = (Y_{i1}, \ldots, Y_{ip})$, with $i = 1, \ldots, n$, where each $Y_{ij} \in \{1, 2, \ldots, D_j\}$ represents the value of variable j for individual i with $j = 1, \ldots, p$. Also, let $\boldsymbol{Y}_j = (Y_{1j}, \ldots, Y_{nj})$. Let $\boldsymbol{R}$ denote the missing value indicator matrix where, for observation i and variable j, $R_{ij} = 1$ when Y_{ij} is missing, and $R_{ij} = 0$ otherwise. We partition $\boldsymbol{Y}_j$ as $\boldsymbol{Y}_j = (\boldsymbol{Y}_j^{obs}, \boldsymbol{Y}_j^{mis})$, where $\boldsymbol{Y}_j^{obs}$ represents all observed values for variable j, that is, data values in the $\boldsymbol{Y}_j$ vector corresponding to $R_{ij} = 0$, and $\boldsymbol{Y}_j^{mis}$ represents all missing values for variable j, that is, data values in the $\boldsymbol{Y}_j$ vector corresponding to $R_{ij} = 1$. We write $\boldsymbol{Y}^{obs} = (\boldsymbol{Y}_1^{obs}, \ldots, \boldsymbol{Y}_p^{obs})$, and $\boldsymbol{Y}^{mis} = (\boldsymbol{Y}_1^{mis}, \ldots, \boldsymbol{Y}_p^{mis})$. Finally, we write $\boldsymbol{Y} = (\boldsymbol{Y}^{obs}, \boldsymbol{Y}^{mis})$.

In MI, one creates $L > 1$ multiply-completed datasets $\mathbf{Z} = (\mathbf{Z}^{(1)}, \ldots, \mathbf{Z}^{(L)})$. Each $\mathbf{Z}^{(l)}$, with $l = 1, \ldots, L$, is generated by replacing missing values in the sample $\boldsymbol{Y}$, with draws from the predictive distributions of models estimated based on the observed data $\boldsymbol{Y}^{obs}$. Analysts then can compute sample estimates for estimands of interest in each completed dataset $\mathbf{Z}^{(l)}$, and combine them using MI inference rules developed by Rubin (1987). We review these rules in Section 3.2.

2.1 MICE

Under MICE, one imputes missing values by iteratively sampling plausible predicted values from a sequence of univariate conditional models, where the models are specified separately for each variable. First, one specifies an order for imputing the variables. The most popular choices are to (i) impute the variables in the order they appear in $\boldsymbol{Y}$, or (ii) impute the variables in increasing order of the number of missing cases. Next, one sets initial values for the missing entries. The most popular choice is to set the values by sampling from the marginal distribution of each $\boldsymbol{Y}_j^{obs}$. Alternatively, for each variable, one can also sample from its conditional distribution, given all other variables, where the distribution is constructed using only available cases.

After initialization, one cycles through the sequence of univariate conditional models, estimating and generating predictions from them, at each iteration. The process continues for T total iterations until the chain converges (Abayomi et al. 2008). We summarize the entire process in Algorithm 1.

Algorithm 1: MICE

Input: Data matrix $\boldsymbol{Y}$ with columns sorted as per imputation order
Output: Multiple completed datasets $\mathbf{Z}^{(1)}, \ldots, \mathbf{Z}^{(L)}$

for $l = 1, \ldots, L$ **do**
  Initialize $\boldsymbol{Y}_j^{mis}$
  for $t = 1, \ldots, T$ **do**
    for $j = 1, \ldots, p$ **do**
      Fit the conditional model $(\boldsymbol{Y}_j | \boldsymbol{Y}_j^{obs}, \{\boldsymbol{Y}_k^{(t)} : k < j\}, \{\boldsymbol{Y}_k^{(t-1)} : k > j\})$
      Generate $\boldsymbol{Y}_j^{mis(t)}$ from the implied
       $(\boldsymbol{Y}_j^{mis} | \boldsymbol{Y}_j^{obs}, \{\boldsymbol{Y}_k^{(t)} : k < j\}, \{\boldsymbol{Y}_k^{(t-1)} : k > j\})$
  Set $\mathbf{Z}^{(l)} = (\boldsymbol{Y}^{obs}, \boldsymbol{Y}^{mis(T)})$
return $\mathbf{Z}^{(1)}, \ldots, \mathbf{Z}^{(L)}$

2.1.1 MI-Multireg and MI-Polr

Most implementations of MICE use GLMs for the univariate conditional models in Step 5 of Algorithm 1. MI-Multireg and MI-Polr are the two most popular GLM models for implementing MICE for categorical data with more than two levels. MI-Multireg uses multinomial or polytomous logistic regression models (Engel 1989). Multinomial logistic regression models are the most common option for analyzing categorical data with more than two outcomes, making MI-Multireg a very popular MI method for both nominal and ordinal variables. MI-Polr on the other hand uses proportional odds logistic regression models (McCullagh 1980, 2005) for imputing missing data for ordinal variables. This results in a more parsimonious model than the multinomial logistic regression. Even though MI-Polr may be expected to be more appropriate for ordinal data than MI-Multireg, we include the latter in the scope of our study to see whether there are certain settings where MI-Multireg would in fact outperform MI-Polr, when dealing with ordinal data.

2.1.2 MI-Cart

In MI-Cart, one uses CART for the univariate conditional models in Step 5 of Algorithm 1, usually to provide more flexibility. CART for categorical data (Breiman et al. 1984) is a nonparametric approach for modeling the relationship between a categorical response variable and potential predictors. CART follows a decision tree structure, using recursive binary splits to partition the predictor space into distinct non-overlapping regions. The top of the tree usually corresponds to the root and each successive split divides the space into two new branches further down the tree. The splitting or partitioning criterion at each node is chosen to minimize the "information impurity", for example, the Gini index of the two child nodes. Also, any splits that do not decrease the lack of fit by some certain thresholds are pruned off. The tree is then built until a stopping criterion, like the minimum number of observations in each child node, is met.

To approximate the conditional distribution of any $\boldsymbol{Y}_j$, given a particular combination of the other variables, we use the proportion of values of $\boldsymbol{Y}_j^{obs}$ in the corresponding leaf. That is, to generate values for $\boldsymbol{Y}_j^{mis(t)}$ under MI-Cart, we traverse down the tree to the appropriate node using the combinations in $(\{\boldsymbol{Y}_k^{(t)} : k < j\}, \{\boldsymbol{Y}_k^{(t-1)} : k > j\})$, and sample from the $\boldsymbol{Y}_j^{obs}$ values in the leaf. For more details on MI-Cart, see Burgette & Reiter (2010).

2.1.3 MI-Forest

MI-Forest also uses a nonparametric approach for the univariate conditional models in Step 5 of Algorithm 1. MI-Forest uses random forests (Tin Kam Ho 1995), an ensemble tree method which builds multiple decision trees to the data. Specifically, random forest constructs many classification and regression trees using bootstrapped datasets, but only uses a sample of the predictors for each split in each tree. Doing so usually reduces the correlation among the trees since it prevents the same variables from dominating the splitting process across all trees. One key advantage of random forests is that it can handle mixed data types and the decorrelation should result in predictions with less variance.

In MI-Forest, we again follow the general outline in Algorithm 1. That is, for each $\boldsymbol{Y}_j$ with missing data, we train a random forest model using available cases, given all other variables. We then generate plausible values for $\boldsymbol{Y}_j^{mis(t)}$ under that model. Next, we cycle through all the variables to generate one completed dataset. Finally, we run the entire process L times to obtain L imputed datasets. For more details on MI-Forest, see Stekhoven & Bühlmann (2012), Shah et al. (2014).

2.2 MI-DPMPM

MI-DPMPM uses the DP mixture of products of multinomial distributions (DPMPM) model, to generate imputations for multivariate categorical data containing missing values (Hu et al. 2017). Intrinsically, the model treats each variable as being nominal. The DPMPM model assumes that each observation i in the data belongs to a latent class

$z_i \in \{1, \ldots, K\}$, and variables within each latent class $k = 1, \ldots, K$, follow independent multinomial distributions.

Specifically, let $\pi_k = P(z_i = k)$, for $k = 1, \ldots, K$, be the probability that observation i belongs to latent class k. Let $\lambda_{kjd} = P(Y_{ij} = d | z_i = k)$ be the probability that Y_{ij} takes value d, given that it belongs to latent class k. Finally, let $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_K)$ and $\boldsymbol{\lambda} = \{\lambda_{kjd} : k = 1, \ldots, K; j = 1, \ldots, p; d = 1, \ldots, D_j\}$. We write the generative model as

$$Y_{ij} | z_i, \boldsymbol{\lambda} \overset{ind}{\sim} \text{Discrete}(\lambda_{z_i j 1}, \ldots, \lambda_{z_i j D_j}) \quad \text{for all } i \text{ and } j \tag{1}$$

$$z_i | \boldsymbol{\pi} \overset{iid}{\sim} \text{Discrete}(\pi_1, \ldots, \pi_K) \quad \text{for all } i. \tag{2}$$

Under this model specification, averaging over the K latent classes induces marginal dependence between the variables. With a large enough number of classes K, the model is consistent for any joint probability distribution (Dunson & Xing 2012).

For prior distributions, we follow Schifeling & Reiter (2016), Si & Reiter (2013), Manrique-Vallier & Reiter (2014*b*). We use independent uniform Dirichlet distributions for each probability vector in $\boldsymbol{\lambda}$, and the truncated stick breaking representation of the DP for the mixture probabilities. We have

$$(\lambda_{kj1}, \ldots, \lambda_{kjD_j}) \overset{ind}{\sim} \text{Dirichlet}(\mathbf{1}_{D_j}) \quad \text{for all } j \text{ and } k \tag{3}$$

$$\pi_k = V_k \prod_{h<k} (1 - V_h) \tag{4}$$

$$V_k | \alpha \overset{iid}{\sim} \text{Beta}(1, \alpha), \quad \text{for } k = 1, \ldots, K-1; \quad V_K = 1 \tag{5}$$

$$\alpha \sim \text{Gamma}(0.25, 0.25). \tag{6}$$

We obtain posterior samples of all parameters in the model using Gibbs sampling. We also follow Si & Reiter (2013) when setting K, starting with a relatively modest value during initial runs. Whenever the number of occupied classes reaches K across any of the iterations, we gradually increase K until that is no longer the case. Finally, we handle

missing values directly within the Gibbs sampler. At any iteration t, we sample each $Y_{ij}^{mis(t)}$ value in $\boldsymbol{Y}_j^{mis(t)}$ from (1), conditional on the current draw of the parameters, and the latent class z_i. To obtain L multiple imputed datasets, we randomly select L posterior predictive samples spread over the iterations, after the chain converges. For more details on the DPMPM model, including further discussions on prior specifications and setting K, see Si & Reiter (2013), Manrique-Vallier & Reiter (2014*b*).

2.3 MI-DPMMVN

MI-DPMMVN uses the DP mixture of multivariate normal distributions (DPMMVN) model (DeYoreo & Kottas 2014), to generate imputations for multivariate ordinal data. Like the DPMPM, the DPMMVN model also assumes that each observation i in the data belongs to a latent class $z_i \in \{1, \ldots, K\}$. However, unlike the DPMPM, the DPMMVN model is primarily designed for modeling ordinal response variables. The DPMMVN model assumes that the ordinal variables arise from continuous latent variables. The latent variables (as well as continuous predictors when available) are assumed to jointly follow a multivariate normal distribution.

Specifically, suppose each $Y_{ij} \in \{1, 2, \ldots, D_j\}$ as before. We introduce continuous latent random variables $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{ip})$ for all i, and cut-offs $\boldsymbol{\gamma_j} = (\gamma_{j0}, \ldots, \gamma_{jD_j})$ for each j. Each $Y_{ij} = d$ if and only if the corresponding $x_{ij} \in (\gamma_{j(d-1)}, \gamma_{jd}]$, where $d = 1, 2, \ldots, D_j$. These latent variables follow a DP mixture of multivariate normal distributions. We have

$$\boldsymbol{x}_i | z_i, \boldsymbol{\mu}_{z_i}, \boldsymbol{\Sigma}_{z_i} \overset{ind}{\sim} N(\boldsymbol{\mu}_{z_i}, \boldsymbol{\Sigma}_{z_i}) \quad \text{for all } i \text{ and } j \tag{7}$$

$$z_i | \boldsymbol{\pi} \overset{iid}{\sim} \text{Discrete}(\pi_1, \ldots, \pi_K) \quad \text{for all } i, \tag{8}$$

combined with the truncated stick breaking DP prior in (4) to (6). For prior distributions

on the remaining parameters, we follow DeYoreo & Kottas (2014) and specify

$$\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k | \boldsymbol{m}, \boldsymbol{V}, \boldsymbol{S} \overset{iid}{\sim} N(\boldsymbol{\mu} | \boldsymbol{m}, \boldsymbol{V}) \cdot IW(\boldsymbol{\Sigma} | \nu, \boldsymbol{S}) \quad \text{for all } k = 1, \ldots, K \tag{9}$$

$$\boldsymbol{m} \sim N(\boldsymbol{a}_m, \boldsymbol{B}_m); \quad \boldsymbol{V} \sim IW(a_V, \boldsymbol{B}_V); \quad \boldsymbol{S} \sim W(a_S, \boldsymbol{B}_S). \tag{10}$$

Here, $W(a_S, \boldsymbol{B}_S)$ is a Wishart distribution with mean $a_S \boldsymbol{B}_S$, and $IW(a_V, \boldsymbol{B}_V)$ is an inverse-Wishart distribution with mean $(a_v - p - 1)^{-1} \boldsymbol{B}_V$. We set the parameters in (4) to (10) to represent vague prior specifications following recommendations by DeYoreo & Kottas (2014). To set K, we can use a similar strategy as in DPMPM. The cut-offs for each variable can be fixed beforehand at arbitrary values and the covariance matrices will be identifiable (DeYoreo & Kottas 2014).

We again use Gibbs sampling to generate posterior draws of the parameters, and handle missing values directly within the sampler, just as we did with the DPMPM. Specifically, for each missing entry, we sample the latent variables using (7), and use the cut-offs to impute the corresponding ordinal level. For more details, see DeYoreo & Kottas (2014).

3 Empirical Study

We conduct our empirical comparison of the methods using data from the 2018 American Community Survey (ACS) Public Use Microdata Sample (PUMS) files (U.S. Census Bureau 2018). As is generally the case for single-year ACS PUMS, the 2018 ACS PUMS contains sample data from about 1 in every 40 households in the United States. The data includes variables collected both at the individual level (e.g., income, age, sex, educational attainment) and at the household level (e.g., lot size, and number of rooms, vehicles, and persons). The data can be downloaded from the United States Bureau of the Census *(https://www2.census.gov/programs-surveys/acs/data/pums/2018/)*.

For this study, we combine housing unit records with the household head records from

the individual level data, and we treat this as our unit of analysis. We do so to create a richer set of ordinal variables than what is available at the household level alone. Also, we only retain individual level data for the household head alone because individuals are nested within households by design, which violates the independence-across-observations assumption needed for all the imputation methods. We remove all empty housing units and all identification variables, including serial number, state, area code, and division. This results in data containing 1,257,501 records and 11 ordinal variables, which we treat as our population data, from which we repeatedly sample from. The 11 ordinal variables comprise of 1 variable with two levels, 3 variables with three to five levels, 4 variables with six to ten levels, and 3 variables with more than ten levels. We describe the variables in more detail in the supplementary material.

We follow the approach in Akande et al. (2017) and use repeated sampling to evaluate the statistical properties of the imputation methods. Here, we randomly sample 10,000 units from the population, without replacement, and create varying proportions of missing data, under both the missing completely at random (MCAR, Little & Rubin (2019)) and missing at random (MAR, Little & Rubin (2019)) scenarios. First, we set five of the variables to be fully observed. We do so to mimic real applications where some of the variables are often fully observed. Next, we set either 30% or 45% of the values of the remaining six variables to be missing according to either MCAR or MAR mechanisms. Across all datasets, the missing data scenarios tend to result in complete cases ranging from 1% (for 45% MAR) to 12% (for 30% MAR), making complete case analysis untenable.

For each random sample of 10,000 observations, and under each combination of missingness mechanism and proportion of missing data, we use each imputation method to create $L = 50$ multiple completed datasets. Figure 1 describes the process for each random sample of 10,000 observations. We repeat the entire process 500 times, each time generating new samples from the population and new missingness patterns.

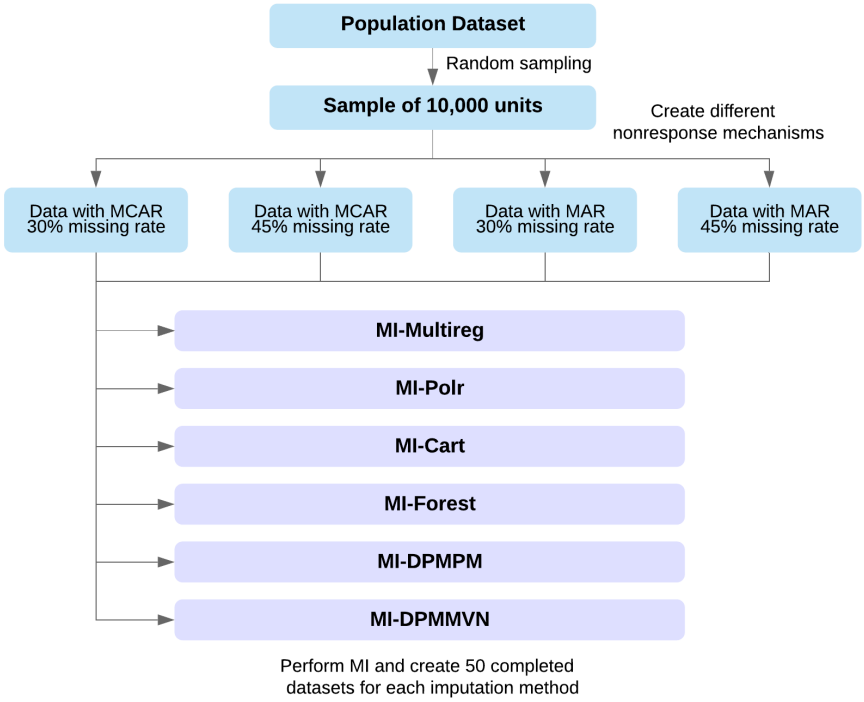


Figure 1: Imputation process for a single random sample from the population.

3.1 Software Implementation

Across all methods, we keep most of the default options and arguments in the packages and code, to mimic the default option most analysts often use. For MI-Multireg, MI-Polr, and MI-Cart, we use the mice package in R (Buuren & Groothuis-Oudshoorn 2011). For MI-Multireg, we use the "polyreg" option, and only consider main effects in the models. For MI-Polr, we use the "polr" option, and once again only consider main effects in the models. For MI-Cart, we use the "cart" option and keep all default arguments, including the requirement that at least four observations must be in each of the terminal nodes.

For MI-Forest, we use the missForest package in R (Stekhoven 2013). We grow 100 trees in each forest. Based on our investigations, as long as the number of trees is large enough, the results across the sample datasets are usually insensitive to it. We set the number of variables being considered for each split to the default value, that is, the square root of the total number of possible predictors. As a sensitivity check, we increase this number to half

the number of possible predictors. Qualitatively, the results are very similar.

For MI-DPMPM, we use the NPBayesImputeCat package in R (Manrique-Vallier & Reiter 2014*a*). We set the number of latent classes $K = 40$, which we find to be sufficiently large based on initial runs on sample datasets of size 10,000. For each simulation run, we run the Gibbs sampler for 15,000 iterations, and set the first 5,000 as burn-in. We assess the convergence of the sampler by examining trace plots of combinations of the model parameters, such as random samples of marginal probabilities, that are insensitive to label switching. For MI-DPMMVN, we adapt the R code in the supplementary material for DeYoreo & Kottas (2014). We set the number of latent classes $K = 50$ based on initial runs. Similar to MI-DPMPM, we run the Gibbs sampler for each simulation run for 15,000 MCMC iterations, and set the first 5,000 as burn-in. Again, we assess convergence using parameters and estimands that are insensitive to label switching.

To generate $L = 50$ multiple completed datasets for any given sample, under our simulation scenarios, using a standard notebook computer, MI-Multireg typically runs for approximately 1 hour and 30 minutes on average, MI-Polr runs for about 50 minutes, MI-Cart and MI-Forest run for about 40 minutes, MI-DPMPM runs for about 5 minutes, while MI-DPMMVN typically runs for about 1 hour and 30 minutes.

3.2 Performance metrics

To compare the performances for all MI methods, we use metrics that evaluate accuracy in estimating joint relationships within the data, rather than one-number metrics that only probe overall accuracy. We evaluate each MI method using coverage rates, relative MSE, and bias, based on the multiply-imputed datasets. We evaluate these metrics on sets of marginal probabilities, bivariate probabilities, and trivariate probabilities separately. We compute estimates of these probabilities in each completed dataset and combine them using the MI rules developed by Rubin (1987).

As a quick review, let q be the completed-data point estimator of estimand Q, and let

u be the estimator of the variance associated with q. For $l = 1, \ldots, L$, let $q^{(l)}$ and $u^{(l)}$ be the values of q and u in completed dataset $\mathbf{Z}^{(l)}$. Then, the MI point estimate of Q is $\bar{q}_L = \sum_{l=1}^{L} q_l/L$, with corresponding variance given by $T_L = (1 + 1/L)b_L + \bar{u}_L$, where $b_L = \sum_{l=1}^{L}(q^{(l)} - \bar{q}_L)^2/(L-1)$, and $\bar{u}_L = \sum_{l=1}^{L} u^{(l)}/L$. Inferences about Q can be made using $(\bar{q}_L - Q) \sim t_\nu(0, T_L)$, where t_v is a t-distribution with $\nu = (L-1)(1 + \bar{u}_L/[(1 + 1/L)b_L])^2$ degrees of freedom. MI inferences require that the central limit theorem applies to the complete-data estimate and that the sampling distribution of the complete-data estimate is approximately normal. We therefore eliminate combinations with Q equal to or near zero in the population. For coverage calculation, we also remove combinations that result in point estimates $\bar{q}$ very close to zero in any of the datasets.

To compute coverage rates, we first construct 95% confidence intervals for each estimand, using the MI inference rules. We then compute the proportion of the one hundred 95% confidence intervals that contain the corresponding Q from the full population data. High quality of the imputations would generally imply coverage rates close to the nominal rate of 0.95 or close to the corresponding coverage rates in the "pre-missing data". For relative MSE, we compare the MSE of each estimator from the imputed datasets, to the MSE from the original sample without missing data. That is,

$$\text{Relative MSE} = \frac{\text{MSE based on imputed data}}{\text{MSE based on pre-missing data}} = \frac{\sum_{h=1}^{500}[\bar{q}_L^{(h)} - Q]^2}{\sum_{h=1}^{500}[\hat{q}^{(h)} - Q]^2}. \tag{11}$$

Here, $\bar{q}_L^{(h)}$ is the value of $\bar{q}_L$ from sampled dataset h while $\hat{q}^{(h)}$ is the point estimate of Q from the same dataset before introducing missing values. Lower values of relative MSE, especially values closer to one, would indicate higher quality of imputations. We further break the MSE for each MI method down into the bias and variance components. This allows us to allocate the differences in the relative MSE values to one of the two components.

We have

$$\text{Bias} = \left[\frac{1}{500} \sum_{h=1}^{500} \bar{q}^{(h)} \right] - Q. \tag{12}$$

Lower values of bias would clearly indicate higher quality of imputations.

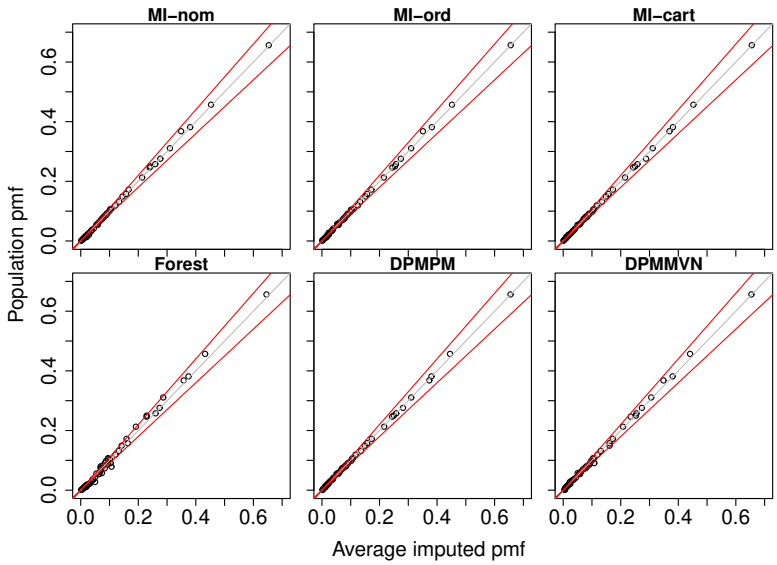
4 Simulation Results

We summarize the results for all the MI methods using graphical displays and tables. We present the most important findings for the MAR and MCAR scenarios, with missingness rates of 30% and 45% for the six variables with missing data. We also focus primarily on results for marginal and bivariate probabilities. We present additional results and findings, including the results for trivariate probabilities, in the supplementary material. In all graphical displays and tables, we write MI-Multireg as MI-nom, MI-Polr as MI-ord, MI-Forest as Forest, MI-DPMPM as DPMPM, and MI-DPMMVN as DPMMVN.

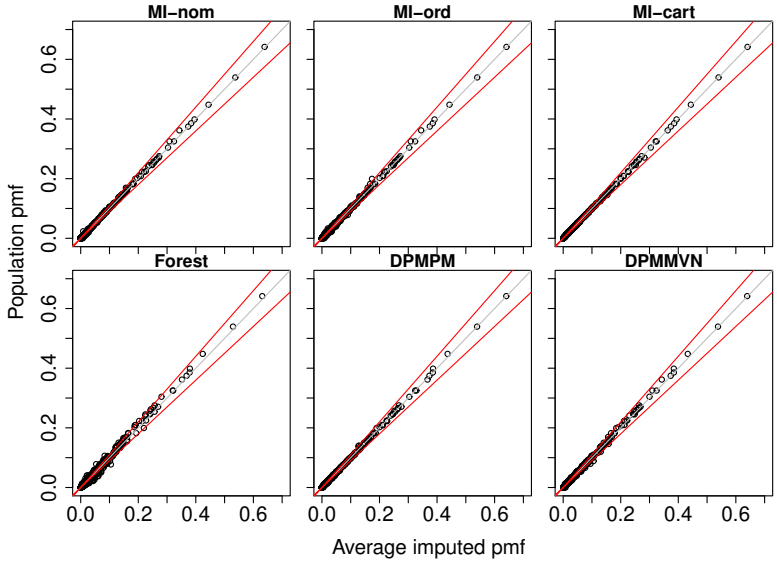
4.1 MAR Scenarios

Since MAR is often the most plausible assumption in practice, we begin with an evaluation of all the MI methods under MAR. First, we create a simulation scenario with $n = 10,000$ and 30% missingness rate. To incorporate MAR, we set missing values independently for two variables (VEH and WKL), and then generate missingness indicators, for four other variables (NP, SCHL, AGEP, and PINCP), from logistic regression models, conditional on the remaining five fully observed variables (MV, RMSP, ENG, MARHT, and RACNUM). For simplicity, we only include main effects in the logistic regression models, and we tune the coefficients to result in approximately 30% missing data, per variable. We present the specified logistic regression models in the supplementary material.

Figure 2 displays the estimated marginal and bivariate probabilities for each MI method,



(a) Results for marginal probabilities



(b) Results for bivariate probabilities

Figure 2: Estimated marginal and bivariate probabilities based on the multiply-imputed datasets from all MI methods, compared to true values in the population dataset, under the 30% MAR scenario. The red lines represent 10% error lines for the 45-degree line.

compared to the true values of the same probabilities in the full population dataset. For the most part, all the MI methods, except MI-Forest, tend to result in point estimates of

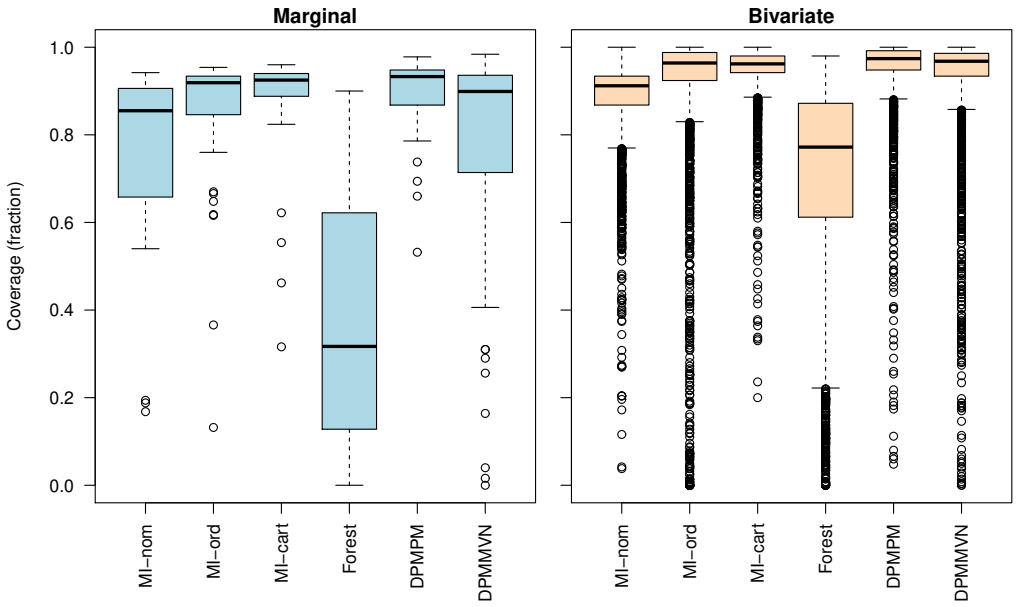


Figure 3: Distributions of coverage rates for all MI methods, under the 30% MAR scenario.

the marginal and bivariate estimands that are very close to the population values, with most of the points falling almost entirely on the 45-degree line. Looking more closely, MI-Multireg, MI-Cart, and MI-DPMPM tend to result in the most accurate point estimates, while MI-Polr, and MI-DPMMVN often yield comparable results, especially for marginal probabilities. On the other hand, MI-Forest appears to result in the least accurate estimates for both marginal and bivariate estimands. This is relative however, since most of the points still appear to be mostly close to the 45-degree line.

Figure 3 displays the estimated coverage rates of the 95% confidence intervals for the marginal and bivariate probabilities, under each method. The distributions of coverage rates for most of the methods are visibly left skewed, especially for the bivariate probabilities. The extremely low coverage rates often correspond to combinations of the variables with very low probabilities in the full population. MI-Polr, MI-Cart, MI-DPMPM, all tend to result in coverage rates that are generally very close to the nominal 95% level, with comparable right tails. In fact, the median coverage rates for all three methods are all above 0.90. The coverage rates for MI-Cart in particular tend to have to have the least skew

Table 1: Distributions of relative MSEs for marginal and bivariate probabilities across all MI methods, under the 30% MAR scenario.

	MI-nom	MI-ord	MI-cart	Forest	DPMPM	DPMMVN
				Marginal		
Min.	1.4	1.2	1.3	1.4	1.3	1.0
1st Qu.	1.9	1.6	1.5	6.8	1.6	1.5
Median	2.6	1.8	1.7	16.1	1.9	2.2
3rd Qu.	6.8	2.3	2.3	27.7	2.8	4.6
Max.	56.7	22.7	8.6	223.7	8.9	234.3
				Bivariate		
Min.	0.0	0.0	0.0	0.3	0.0	0.0
1st Qu.	1.4	0.9	0.9	1.6	0.9	0.9
Median	2.0	1.1	1.1	3.6	1.0	1.0
3rd Qu.	3.1	1.9	1.5	8.5	1.5	1.7
Max.	1440.0	9180.3	71.0	679.7	208.5	971.9

overall. MI-Multireg on the hand tends to result in coverage rates that are generally lower than MI-Polr, MI-Cart, MI-DPMPM. That said, we once again note that MI-Multireg still generally tends to result in more accurate point estimates than MI-Polr, according to Figure 2. Finally, MI-Forest tends to result in coverage rates that are much farther off from the nominal 95% level, than all the other methods.

Table 1 displays the distributions of the estimated relative MSE for the marginal and bivariate probabilities, under each method. For marginal probabilities, MI-Polr, MI-Cart, and MI-DPMPM tend to yield point estimates with the lowest MSE, relative to the MSE of the pre-missing data. Overall, across both the marginal and bivariate estimands, MI-Cart and MI-DPMPM generally result in the smallest relative MSE values, when looking at the entire distributions of the relative MSE values, thus outperforming all other methods under this simulation setting. MI-Multireg, and MI-DPMMVN tend to result in relative MSE values that are higher than, but mostly comparable to MI-Polr, MI-Cart, and MI-DPMPM. However, we note that while MI-DPMMVN generally results in relative MSE values that are higher than MI-Polr for marginal estimands, it actually results in relative MSE values that are lower than the latter, when looking at bivariate estimands. This suggests that when looking at measures of accuracy such as MSE, MI-DPMMVN may be more likely to yield

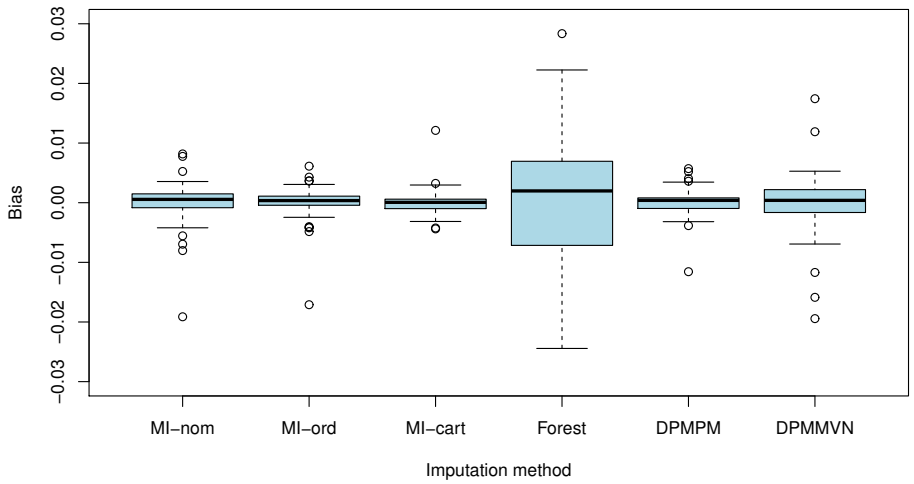


Figure 4: Distributions of estimated bias for marginal probabilities across all MI methods, under the 30% MAR scenario.

more accurate estimates for joint relationships, than the other MI using GLM methods. MI-Forest on the other hand tends to yield estimates with the highest relative MSE values across the board. This is consistent with the conclusions from the other metrics.

To provide additional context for the trends in the relative MSE values, we also examine the estimated bias for all the methods. Figure 4 displays the distributions of estimated bias for the marginal probabilities under each method. The overall trends appear to mimic the trends in the estimated relative MSE, at least for marginal estimands. That is, MI-Cart, MI-Polr, and MI-DPMPM, once again appear to generally outperform all other methods, with MI-Multireg, and MI-DPMMVN not too far behind. MI-Forest tends to yield point estimates with the largest bias across the board. We present the estimated bias values for bivariate and trivariate probabilities, as well as numerical summaries of all the estimated bias values in the supplementary material.

As a sensitivity analysis, we examine whether the overall conclusions from our MAR scenario change as the missingness rate increases. We create the same MAR scenario as before but tune all parameters to result in 45% missingness rate, for each of the six variables

Table 2: Distributions of relative MSEs for marginal and bivariate probabilities across all MI methods, under the 45% MAR scenario.

	MI-nom	MI-ord	MI-Cart	Forest	DPMPM	DPMMVN
				Marginal		
Min.	1.7	1.5	1.6	3.9	1.6	1.1
1st Qu.	2.9	2.0	1.9	11.5	2.3	2.3
Median	5.4	2.4	2.4	31.1	2.8	3.7
3rd Qu.	14.2	3.8	3.7	67.5	4.3	9.0
Max.	284.1	1101.2	22.8	277.7	21.9	590.6
				Bivariate		
Min.	0.0	0.0	0.0	0.1	0.0	0.0
1st Qu.	1.8	0.9	0.9	2.3	0.9	0.9
Median	3.1	1.3	1.2	6.4	1.1	1.3
3rd Qu.	5.9	3.0	2.0	17.4	2.1	2.9
Max.	5334.4	28220.4	204.4	1698.9	360.4	1533.9

with missing data. Table 2 displays the distributions of the estimated relative MSE for the marginal and bivariate probabilities, under each method. As expected with a higher missingness rate, all relative MSE values are generally larger across all methods. For all the methods and for most of the estimands, the relative MSE values for the marginal estimands increase by a multiplicative factor of approximately 2, compared to the 30% MAR scenario. The overall trends for the marginal and bivariate probabilities are very similar to the 30% MAR scenario, with MI-Multireg, MI-Polr, MI-Cart, MI-DPMPM, and MI-DPMMVN, once again yielding more accurate point estimates than MI-Forest. Although, the largest differences in the relative MSE values, when comparing this scenario to the 30% MAR scenario, correspond to MI-Forest, MI-Polr actually has the largest maximum relative MSE values across all the methods. Across all methods, the maximum values result from one of the four variables with missing values under MAR. Two levels of this variable have very low probabilities in the original data. With 45% missing data, there is very little data in combinations involving that level, and the proportional odds model appears to struggle the most from this. We also present additional results for this scenario, including the estimated marginal, and joint probabilities for each MI method, as well as the distributions of estimated coverage rates, and bias, in the supplementary material. Qualitatively, the

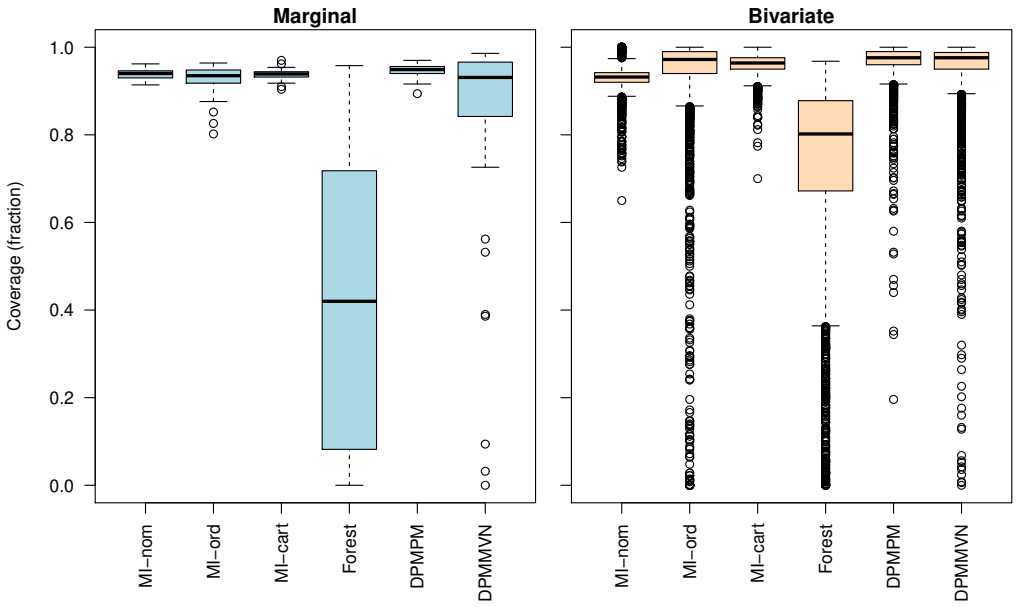


Figure 5: Distributions of coverage rates for all MI methods, under the 30% MCAR scenario.

overall trends are the same when we increase the missingness rate from 30% to 45%.

4.2 MCAR Scenarios

Since analysts still sometimes have to conduct analyses under MCAR, we create a simulation scenario with $n = 10,000$ and 30% missingness rate. To incorporate MCAR, we set missing values independently for the six variables. Figure 5 displays the estimated coverage rates of the 95% confidence intervals for the marginal and bivariate probabilities, under each method. Overall, the coverage rates for all the methods are generally higher than the coverage rates from the MAR scenarios. MI-Multireg, MI-Polr, MI-Cart, MI-DPMPM, and MI-DPMMVN, once again yield more accurate point estimates than MI-Forest. Interestingly, under this scenario, the median coverage rate (and overall distribution of the rates) for MI-Multireg is actually higher than the median coverage rate (and overall distribution of the rates) for MI-Polr for marginal estimands. For bivariate estimands, the trend reverses, with MI-Polr having the higher median coverage rate. However, MI-Polr

Table 3: Distributions of relative MSEs for marginal and bivariate probabilities across all MI methods, under the 30% MCAR scenario.

	MI-nom	MI-ord	MI-Cart	Forest	DPMPM	DPMMVN
				Marginal		
Min.	1.2	1.2	1.2	1.2	1.2	0.8
1st Qu.	1.4	1.4	1.4	3.7	1.4	1.1
Median	1.4	1.5	1.4	8.5	1.4	1.6
3rd Qu.	1.5	1.7	1.4	20.3	1.5	2.4
Max.	2.0	3.0	1.5	78.0	5.2	86.6
				Bivariate		
Min.	0.3	0.3	0.3	0.6	0.4	0.4
1st Qu.	1.4	0.8	0.9	1.6	0.8	0.8
Median	1.5	1.0	1.1	2.8	1.0	0.9
3rd Qu.	1.8	1.4	1.3	5.5	1.3	1.3
Max.	11.7	3259.6	25.7	115.5	105.6	329.4

has a much longer lower tail, sometimes reaching very low rates. This finding is consistent with the result of MCAR scenario with 45% missing rate reported in the supplementary material. Even though MI-Forest generally improves from the MAR scenario, the median coverage rate for the marginal estimands remain lower than 50%.

Table 3 displays the distributions of the estimated relative MSEs for the marginal and bivariate probabilities, under each method. The overall trend is consistent with the trend from the coverage rates. The results also suggest that MI-Multireg can yield much more comparable results to MI-Polr, when dealing with MCAR, rather than MAR.

As a final sensitivity analysis, we examine whether the overall conclusions from our first MCAR scenario change as the missingness rate increases, as we did under MAR. We do so by increasing the missingness rates independently for the six variables. Table 4 displays the distributions of the estimated relative MSE for the marginal and bivariate probabilities, under each method. As expected with a higher missingness rate, all relative MSE values are generally larger across all methods. The overall trends remain the same as the 30% MCAR scenario, with MI-Multireg, MI-Polr, MI-Cart, MI-DPMPM, and MI-DPMMVN, once again yielding more accurate point estimates than MI-Forest. We present additional results for this scenario in the supplementary material.

Table 4: Distributions of relative MSEs for marginal and bivariate probabilities across all MI methods, under the 45% MCAR scenario.

	MI-nom	MI-ord	MI-Cart	Forest	DPMPM	DPMMVN
			Marginal			
Min.	1.6	1.4	1.4	2.3	1.4	0.7
1st Qu.	1.8	1.8	1.7	7.7	1.8	1.4
Median	1.9	2.0	1.8	22.6	1.9	2.6
3rd Qu.	2.1	2.4	1.9	43.1	2.0	4.3
Max.	4.6	5.0	2.1	161.9	11.6	198.3
			Bivariate			
Min.	0.0	0.1	0.0	0.4	0.2	0.3
1st Qu.	1.7	0.8	0.9	2.2	0.8	0.8
Median	1.9	1.1	1.2	5.2	1.1	1.0
3rd Qu.	2.8	2.0	1.6	11.5	1.7	1.8
Max.	39.5	7145.5	53.9	221.3	263.3	1127.6

5 Discussion

Our simulation results suggest several general conclusions about the performances of the MI methods for ordinal data. Overall, MI-Cart and MI-DPMPM arguably outperform all other methods, resulting in coverage rates around 0.90 across all scenarios, the lowest relative MSE values, and the least biased estimates altogether. MI-Polr often yields comparable estimates especially when the missingness rate is 30%, and is only slightly worse when the missingness rate increases to 45%. Since MI-Multireg requires more parameters compared to MI-Polr, it can yield slightly worse estimates when the data is missing under MAR, particularly as the missingness rate increases. While MI-DPMMVN generally yields less accurate estimates than MI-Cart, and MI-DPMPM, the estimates are not often far off particularly for the bivariate and trivariate estimands.

When doing MI for ordinal data, clearly analysts are able to choose between treating the variables as either ordinal or nominal. Across most of the simulation scenarios, the results suggest that models developed specifically for ordinal data can provide very accurate estimates of marginal, bivariate and even trivariate probabilities, comparable to MI-Cart and MI-DPMPM. However, there are also scenarios where models developed for ordinal

data may not always provide clear advantages over other alternatives. Specifically, if one generally prefers GLM models for example, the results suggest that one might actually favor MI-Multireg over MI-Polr when dealing with MCAR scenarios with low missingness rates. Indeed, it is well known that the proportional odds model in MI-Polr may not always match the distributions in the data, so that one must check the model assumptions carefully before selecting MI-Polr over MI-Multireg. Overall, when dealing with MI using GLM models, there can be advantages to treating the variable as nominal instead of ordinal, in certain scenarios. Similarly, MI-DPMPM generally outperforms MI-DPMMVN across all simulation scenarios, again suggesting that in the context of MI, there may not always be clear advantages for flexible ordinal models over other alternatives. However, when choosing between MI-DPMPM and MI-DPMMVN, we again suggest that analysts do extensive model checking to ascertain that the model assumptions are met.

MI-Forest and other machine learning methods, for example, GAIN (Yoon et al. 2018), are developed primarily to be generally more accurate prediction engines. However, this predictive power may not always guarantee correct distributional characteristics that preserves relationships among variables in the data. Moreover, by always imputing the most probable values, they are likely to underestimate the uncertainty in the missing entries. Indeed, we also include two implementations of GAIN in many of the simulation scenarios included in the supplementary material. Across all the simulation scenarios, all other MI methods generally outperform MI-Forest and GAIN. For example, the 95% confidence intervals based on MI-Forest and the two implementations of GAIN are generally much narrower than the intervals based on the other methods, so that they often fail to capture the population estimands very often. Thus, the coverage rates are often far from the nominal level for all the simulation results both in this article and the supplementary material. We do note that our second implementation of GAIN, shows some improvement over the baseline GAIN method in Yoon et al. (2018), however, the results are still generally less accurate than MI-Multireg, MI-Polr, MI-Cart, MI-DPMPM, and MI-DPMMVN. When the

goal of the analysis focuses on overall accuracy of the predictions, averaged across all the variables, we note that these two machine learning methods can still be useful.

Finally, as with any simulation study, the conclusions from our simulation scenarios may not generalize to other different settings. For example, we only consider data containing ordinal variables, whereas data from many other applications often include mixed data types. In those settings, it is possible that the trends we observe based on our results may change. Of all the methods we consider in our study, only MI-Cart, MI-Forest and MI-DPMMVM can be applied directly to general mixed data types. GAIN can also be applied but only after modifications to the loss functions. Future work may focus on evaluating the performance of these methods in settings with mixed data types.

6 Supplementary Materials

The supplementary material contains the data dictionary for the ACS dataset, the logistic regressions for the MAR scenarios, as well as additional tables and displays for all the simulation scenarios in Section 4, and histograms showing the marginal distributions of all variables in the ACS dataset.

References

Abayomi, K., Gelman, A. E. & Levy, M. (2008), 'Diagnostics for multiple imputations', *Journal of the Royal Statistical Society, Series C* **57**, 273–291.

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

Bertsimas, D., Pawlowski, C. & Zhuo, Y. (2018), 'From predictive methods to missing data imputation: An optimization approach', *Journal of Machine Learning Research* **18**, 1–39.

Breiman, L. (2001), 'Random forests', *Machine Learning* **45**, 5–32.

Breiman, L., Friedman, J. H., Olshen, R. A. & Stone, C. J. (1984), *Classification and Regression Trees*, Belmont, CA: Wadsworh, Inc.

Brick, J. M. & Williams, D. (2013), 'Explaining rising nonresponse rates in cross-sectional surveys', *The ANNALS of the American Academy of Political and Social Science* **645**, 36–59.

Burgette, L. & Reiter, J. P. (2010), 'Multiple imputation via sequential regression trees', *American Journal of Epidemiology* **172**, 1070–1076.

Buuren, S. (2007), 'Multiple imputation of discrete and continuous data by fully conditional specification', *Statistical methods in medical research* **16**, 219–42.

Buuren, S. & Groothuis-Oudshoorn, C. (2011), 'Mice: Multivariate imputation by chained equations in r', *Journal of Statistical Software* **45**.

Camino, R. D., Hammerschmidt, C. A. & State, R. (2019), 'Improving missing data imputation with deep generative models', *ArXiv* **abs/1902.10666**.

Curtin, R., Presser, S. & Singer, E. (2005), 'Changes in telephone survey nonresponse over the past quarter century', *Public Opinion Quarterly* **69**, 87–98.

DeYoreo, M. & Kottas, A. (2014), 'Bayesian nonparametric modeling for multivariate ordinal regression', *Journal of Computational and Graphical Statistics* .

Diaz-Ordaz, K., Kenward, M. G., Gomes, M. & Grieve, R. (2014), 'A comparison of multiple imputation methods for bivariate hierarchical outcomes'.

Doove, L., Buuren, S. & Dusseldorp, E. (2014), 'Recursive partitioning for missing data imputation in the presence of interaction effects', *Computational Statistics & Data Analysis* **72**, 92–104.

Dunson, D. & Xing, C. (2012), 'Nonparametric bayes modeling of multivariate categorical data', *Journal of the American Statistical Association* **104**, 1042–1051.

Engel, J. (1989), Polytomous logistic regression: Methods and application, *in* A. Decarli, B. J. Francis, R. Gilchrist & G. U. H. Seeber, eds, 'Statistical Modelling', Springer New York, New York, NY, pp. 135–143.

Honaker, J., King, G. & Blackwell, M. (2012), 'Amelia ii: A program for missing data', *Journal of Statistical Software* **45**.

Hu, J., Reiter, J. P. & Wang, Q. (2017), 'Dirichlet process mixture models for modeling and generating synthetic versions of nested categorical data', *Bayesian Analysis* p. forthcoming.

Laaksonen, S. (2016), 'Multiple imputation for a continuous variable', *Journal of Mathematics and Statistical Science* **2016**, 624–643.

Li, F., Baccini, M., Mealli, F., Zell, E., Frangakis, C. & Rubin, D. (2014), 'Multiple imputation by ordered monotone blocks with application to the anthrax vaccine research program', *Journal of Computational and Graphical Statistics* **23**.

Little, R. & Rubin, D. B. (2019), *Statistical Analysis with Missing Data: Third Edition*, New York: John Wiley & Sons.

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

Manrique-Vallier, D. & Reiter, J. P. (2014*b*), 'Bayesian multiple imputation for large-scale categorical data with structural zeros', *Survey Methodology* **40**, 125–134.

McCullagh, P. (1980), 'Regression models for ordinal data', *Journal of the Royal Statistical Society. Series B* **42**.

McCullagh, P. (2005), *Proportional-Odds Model*, American Cancer Society, chapter 5, pp. 3560–63.
URL: *https://onlinelibrary.wiley.com/doi/abs/10.1002/0470011815.b2a10049*

Nazábal, A., Olmos, P., Ghahramani, Z. & Valera, I. (2020), 'Handling incomplete heterogeneous data using vaes', *Pattern Recognition* **107**, 107501.

Pielsticker, D. & Hiebl, M. (2020), 'Survey response rates in family business research', *European Management Review* **17**, 327–346.

Raghunathan, T., Lepkowski, J., Hoewyk, J. & Solenberger, P. (2000), 'A multivariate technique for multiply imputing missing values using a sequence of regression models', *Survey Methodology* **27**.

Rubin, D. (1976), 'Inference and missing data', *Biometrika* **63**, 581–592.

Rubin, D. B. (1987), *Multiple Imputation for Nonresponse in Surveys*, Wiley.

Saunders, C., Elliott, M., Lyratzopoulos, G. & Abel, G. (2015), 'Do differential response rates to patient surveys between organizations lead to unfair performance comparisons?', *Medical Care* **54**, 1.

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

Schifeling, T. S. & Reiter, J. P. (2016), 'Incorporating marginal prior information into latent class models', *Bayesian Analysis* **11**, 499–518.

Shah, A., Bartlett, J., Carpenter, J., Nicholas, O. & Hemingway, H. (2014), 'Comparison of random forest and parametric imputation models for imputing missing data using mice: A caliber study', *American journal of epidemiology* **179**, 764–74.

Si, Y. & Reiter, J. P. (2013), 'Nonparametric Bayesian multiple imputation for incomplete categorical variables in large-scale assessment surveys', *Journal of Educational and Behavioral Statistics* **38**, 499–521.

Stekhoven, D. & Bühlmann, P. (2012), 'Missforest?non-parametric missing value imputation for mixed-type data', *Bioinformatics (Oxford, England)* **28**, 112–8.

Stekhoven, D. J. (2013), *missForest: Nonparametric Missing Value Imputation using Random Forest*. R package version 1.4.

Tin Kam Ho (1995), Random decision forests, *in* 'Proceedings of 3rd International Conference on Document Analysis and Recognition', Vol. 1, pp. 278–282 vol.1.

U.S. Census Bureau (2018), 'American community survey, 2018 american community survey 1-year estimates public use microdata samples', retrieved from.
URL: *https://www2.census.gov/programs-surveys/acs/data/pums/2018/*

van Buuren, S., Brand, J. P. L., Groothuis-Oudshoorn, C. G. M. & Rubin, D. B. (2006), 'Fully conditional specification in multivariate imputation', *Journal of Statistical Computation and Simulation* **76**(12), 1049–1064.

White, I. R., Royston, P. & Wood, A. M. (2011), 'Multiple imputation using chained equations: Issues and guidance for practice', *Statistics in Medicine* **30**(4), 377–399.
URL: *http://dx.doi.org/10.1002/sim.4067*

Yoon, J., Jordon, J. & Schaar, M. (2018), Gain: Missing data imputation using generative adversarial nets, *in* 'Proceedings of the 35th International Conference on Machine Learning', Stockholm, Sweden.

Yuan, Y. (2011), 'Multiple imputation using sas software', *J Stat Software* **45**.