LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

INSTITUT FÜR STATISTIK

Humera Razzak, Christian Heumann

A hybrid technique for the multiple imputation of survey data

Technical Report Number 229, 2019
Department of Statistics
University of Munich

http://www.statistik.uni-muenchen.de

A hybrid technique for the multiple imputation of survey data

HUMERA RAZZAK[1] CHRISTIAN HEUMANN[2]

Abstract

Most of the background variables in MICS (multiple indicator cluster surveys) are categorical with many categories. Like many other survey data, the MICS 2014 women's data suffers from a large number of missing values. Additionally, complex dependencies may be existent among a large number of categorical variables in such surveys. The most commonly used parametric multiple imputation (MI) approaches based on log linear models or chained equations (MICE) become problematic in these situations and often the implemented algorithms fail. On the other hand, nonparametric MI techniques based on Bayesian latent class models have worked very well if only categorical variables are considered. This paper describes how chained equations MI for continuous variables can be made dependent on categorical variables which have been imputed beforehand by using latent class models. Root mean square errors (RMSEs) and coverage rates of 95% confidence intervals (CI) for generalized linear models (GLM's) with binary response are estimated in a simulation study and a comparison is made among proposed and various existing MI methods. The proposed method outperforms the MICE algorithms in most of the cases with less computational time. The results obtained by the simulation study are supported by a real data example.

Keywords: Complex dependencies; MICE; Multiple Indicator Cluster Surveys

1 Introduction

Information on many variables is collected in different large-scale surveys like Multiple Indicator Cluster Surveys (MICS). The MICS provides opportunities to fill data gaps for monitoring the health situation of children and women in under developed countries. MICS collects data on various indicators like mortality, nutrition, child and reproductive health, etc. Face to face interviews with household members are conducted to collect data. Information based on background variables of the indicators mentioned

[1] humera.razzak@stat.uni-muenchen.de

[2] christian.heumann@stat.uni-muenchen.de

above is very important for data analysis, and for policy making (Corsi, Perkins and Subramanian, 2017). However, the problem of missing data is inevitable in such studies. For example, the data set of individual women from MICS 2014, which has been used in the real data example latter, has a high percentage of data missing on 200 background variables. This problem arises, for example, due to item non response (INR) or entry errors etc. Beside INR, general reasons for the missing datasets include data entry errors, system failures etc. There are three missing data mechanisms. Missing values in any data can be missing completely at random (MCAR), or missing at random (MAR), or missing not at random (MNAR) (Rubin, 1987; Little and Rubin, 2002). In MCAR, the probability of missing data on a variable is not correlated to itself and or other measured variables. In MAR, the probability of missing depends on other, observed, variables. Finally, data are MNAR if the probability of missing depends on the variable value itself. Practically all methods implemented in software assume MAR. MNAR is called "non-ignorable", if the parameters driving the missing data process and the parameters driving the data generating process are distinct (or independent in a Bayesian analysis), but this is not further considered in the paper. Exact missing data mechanisms are often unknown when dealing with large scale data sets. Therefore, most of the time, certain assumptions are made accordingly. Li et al. (2012) addresses some problems with missing large data. Little's MCAR test proposed by Little (1988) is used commonly for testing missing data being MCAR.

The representativeness of the sample can be reduced and inferences about the population can be distorted due to missing values. Moreover, ignoring missing data can lead to a bias of unknown direction and magnitude in the estimated parameters. Therefore, it is critical to impute the data, which usually provides more accurate inference compared to ad-hoc methods (e.g. complete case (CC) analysis or single imputation) in case of missing at random (MAR) (Abdella and Marwala, 2005; Little and Rubin, 2002). The CC analysis sacrifices all units where at least the value of one variable is missing. Such methods are still very popular in psychological research (Schlomer et al., 2010). However, the CC analysis (listwise deletion) can lead to biased estimates (Little and Rubin, 2002). The CC method also results in a loss of power, which can make the analysis inefficient (Little and Rubin, 2002). Despite of being the worst available method (Wilkinson and Task Force on Statistical Inference, 1999), CC is still the most applied technique due to the simplicity and availability as default option in statistical software packages (van Ginkel, 2007). The hot-deck method is another approach and belongs to the family of single-imputation approaches. This method replaces missing values with values from a "similar" responding unit (Andridge and Little, 2010) and the empirical distribution obtained is used to draw the

imputed values. In the case that the entire sample of respondents is being used as a single donor pool, this method produces consistent and unbiased estimates for missing completely at random (MCAR) data (Rubin, 1976; Little and Rubin, 2002). This method uses covariate information, avoids strong parametric assumptions and requires no careful modelling to develop selection criteria for imputing a value because it does not have any parametric model (Schafer and Graham, 2002). However, the problem with this method is that it lacks the clear criteria to guide the selection of the donor set of complete cases (Pérez et al., 2002). Bayesian bootstrap (Rubin, 1987) is a useful alternative when standard hot-deck becomes unsuitable to impute in the presence of a large number of variables (Andridge and Little, 2017). Other proposed methods for missing data use various statistical methods including self-organizing maps (SOM) (Kohonen, 1995; Oja and Kaski, 1999), k-nearest neighbour (kNN) (Batista and Monard, 2003), multi-layer perceptron (Sharpe and Solly, 1995), recurrent neural networks (Bengio and Gingras, 1995). Auto-associative neural network imputations with genetic algorithms are proposed by Pyle (1999), Narayanan et al. (2002), Chung and Merat (1996). Marseguerra and Zoia (2005) and Marwala and Chakraverty (2006) also implement some of the well-known methods used for handling missing data. Multi-task learning approaches are some other techniques based on machine learning methods (Ankaiah and Ravi, 2011). According to the studies of Horton and Kleinman (2007), Honaker et al. (2011), Royston and White (2011) and van Buuren and Groothuis-Oudshoorn (2011), over the last three decades, a wide range of variety and settings of multiple imputation (MI) techniques has been introduced for catering missing data problems in different research areas (Abdella and Marwala, 2005; Honaker et al., 2011; Little and Rubin, 2002; Schafer and Graham, 2002). MI, likelihood based analysis, and weighting approaches are alternatives to listwise and pairwise deletion methods. These methods usually make the assumption that the missing data is missing at random (MAR), hence making the estimates unbiased, consistent, and asymptotically normal (Allison, 2002; Barnard and Meng, 1999; Roth, 1994; Schafer and Graham, 2002) if that assumption holds. Model-based MI is currently considered the most popular method of addressing missing data problems. The true complete-data distribution and the missing-data mechanism form the basis of the imputation model which can be explicit or implicit by nature (Rubin, 1987). Draws from the posterior predictive distribution of the unobserved data given the observed data can be used to impute missing values. This process is repeated and M imputed data sets are created. By conducting the analysis on each of these data sets, the resulting M point and M variance estimates are then combined by a set of rules (Rubin, 1987). Missing values in continuous variables are often treated using a multivariate normal MI. These models are often robust to departure from normality by nature (Graham and Schafer, 1999; Schafer, 1997). Indicators in survey datasets are mostly categorical. Schafer (1997) describes that MI

with log-linear models can be used to generate imputed values for such indicators by capturing the associations in the joint distribution. A severe restriction is that the number of variables must in general be small (Vermunt et al., 2008). The fully conditional specification (FCS) (van Buuren, 2007), also known as MI by chained equations (MICE) (Raghunathan et al., 2001; van Buuren, 2007) is another important tool. Missing values are sequentially imputed by estimating a series of univariate conditional models. Normal regressions and logistic or multinomial logistic regressions are used for continuous and categorical dependent variables, respectively. Alternatively, a method called predictive mean matching (PMM) can be used. Newer implementations also allow classification and regression trees (CART). MICE is an iterative method and imputes missing values variable by variable. It uses the current regression estimates for the response variable, where the response variable in this context is the actual target variable in the iterative process for which missing values are imputed. MICE assumes that equivalent, or at least nearly as good, draws for the joint distribution of the variables can be approximated by the sequential draws from the univariate conditional models. There are three main limitations or difficulties in the implementation of MICE. First, there is a possible lack of compatibility among the set of univariate conditional regression models and the joint distribution of the variables being imputed (Arnold and Press, 1989; Gelman and Speed, 1993). Although an algorithm is proposed which selects the sequence of regression models such that they are assumed to be a good fit for the data, it is very complicated to establish exact conditions for convergence (Zhu and Raghunathan, 2016). Second, the risk of overlooking higher order interactions arises when MICE includes only the main effects in the univariate conditional regression models, although using CART may resolve this problem. Third, the procedure is very time consuming when higher-order interactions are included parametrically in the model (Vermunt et al., 2008). To resolve such complications, a fully Bayesian Joint Modelling (JM) approach, called Dirichlet process mixture of products of multinomial distributions (DPMPM), is proposed by Si and Reiter (2013). This approach uses nonparametric Bayesian versions of latent class models to multiply impute high-dimensional categorical data (Vermunt et al., 2008). This approach has two stages. In stage one, a mixture of independent multinomial distributions is modelled for a contingency table of the categorical variables. In the second stage, the mixture distributions are estimated non-parametrically with Dirichlet process prior distributions. Arbitrarily complex dependencies can be described by such mixtures of multinomials. Since the computation of these dependencies is practical and generally easy, they can serve as an effective general purpose MI engine. These models have been successfully used to impute missing values in up to 80 categorical variables (Si and Reiter, 2013).Murray and Reiter (2016) have also worked on combining Dirichlet process mixtures of multinomial and

multivariate normal distributions for categorical and continuous variables, but this approach involves complicated models to create the dependence structure between the continuous and the categorical variables. The R (R Core Team, 2018) package "NPBayesImputeCat" by Quanli et al. (2018) is a tool for non-parametric Bayesian JM MI, but the implementation of this package is restricted to categorical variables. Since categorical variables are internally represented as dummy variables which could easily double the actual number of predictors, the implementation of the FCS MI by chained equations algorithm becomes extremely slow or difficult in the presence of categorical variables with missing values. The R package "mice" by van Buuren and Groothuis-Oudshoorn (2011) implements MI by chained equations. Usually, household surveys based on health studies include data on a range of risk factors and health outcomes, including categorical variables with many categories mainly, and often the number of numeric variables is less as compared to categorical variables in such studies (Chandra et al., 2005; Gulliford et al., 1999). Therefore, one is limited in the choice of both MI methods, i.e. for using the former (JM), one has to sacrifice continuous variables in the analysis (or categorize them) and the latter (FCS) becomes problematic if many categorical variables are involved. Due to certain limitations, both approaches cannot be used together without correct modifications. An easy to implement hybrid technique is proposed in this paper which describes how FCS MI by chained equations for continuous variables can be blended with JM MI by latent class models for categorical variables.

The paper is organized as follows: A detailed description of a fully Bayesian, JM approach for multiple imputations of large categorical datasets is given in Section 2. In Section 3, the measures of performance used for the comparisons are described. The hybrid algorithm is described in Section 4. Section 5 compares the performance of different imputation methods in simulation studies. In Section 6, the proposed method is applied to a real data set and results are discussed. Concluding remarks are given at the end.

2 Latent class models and multiple imputation

2.1 Bayesian latent class imputation model and MI

To understand a fully Bayesian, JM approach to multiply impute large categorical datasets, it is important to understand a few details regarding how mixture models are used for density estimation and MI. The distribution of categorical data can be described by a mixture model known as latent class model (Lazarsfeld, 1950). Mixture models are considered as flexible tools which model the association structure

of a set of variables (their joint density) by utilizing a finite mixture of simpler densities (McLachlan and Peel, 2000). The probability of having a specific response pattern is defined by each mixture component in a Latent Class Analysis (LCA). A weighted average of the class-specific densities generates the estimated overall density. As described by Lazarsfeld (1950), the scores of different items are independent of each other within latent classes due to local independence assumptions in LCA. A brief introduction to the mathematical form of an LC model as a tool for density estimation is given in the following: Let y_{ij} be the score of the i_{th} person on the j_{th} categorical item belonging to an $n \times J$ data-matrix Y *(i = 1, ..., n, j = 1, ..., J),* y_i the J-dimensional vector with all scores of person i, and x_i a discrete (unobserved) latent variable with K categories. In the LC model, the joint density P (y_i; $\boldsymbol{\pi}$) has the following form:

$$\mathrm{P}\,(\boldsymbol{y_i};\, \boldsymbol{\pi}) = \sum_{k=1}^{\mathrm{K}} P(x_i = k;\pi_{\boldsymbol{x}})\, \mathrm{P}\left(y_i | x_i = k;\pi_{\boldsymbol{y}}\right)$$

$$= \sum_{k=1}^{\mathrm{K}} P(x_i = k;\pi_{\boldsymbol{x}}) \prod_{j=1}^{J} \mathrm{P}\left(y_{ij} | x_i = k;\pi_{yj}\right) \qquad (1)$$

where $\boldsymbol{\pi}=(\pi_x, \pi_y)$ is a set of LC model parameters which can be partitioned into two parts. The first part contains the latent class proportions (π_x) and the second contains class-specific item response probabilities (π_y). A separate set of parameters for each of the J items ($\boldsymbol{\pi}_{yj}$) is assigned to the second part. Due to the fact that a mixture distribution is used, a weighted sum of the K class-specific multinomial densities $\mathrm{P}(y_i|x_i = k; \pi_y)$ generates the overall density. In this generation, the latent proportions are used as weights. From (1) it can be seen that the product over the J independent multinomial distributions (conditional on the *k-th* latent class) makes use of the local independence assumption. The first, second, and higher-order moments of the J response variables can be captured in LC models by setting the number of latent classes large enough (McLachlan and Peel 2000). The generated higher-order moments are actually the univariate margins, bivariate associations, and higher-order interactions when dealing with categorical variables (Vermunt et al., 2008). The unit's posterior class membership probabilities, i.e. the probability that a unit belongs to the k-th class given the observed data pattern y_i, is the quantity of interest when using LC models. According to the theorem of Bayes, we can define this quantity as follows:

$$\mathrm{P}(x_i = k | y_i\,;\boldsymbol{\pi}) = \frac{P(x_i = k;\pi_{\mathbf{x}})P(y_i|x_i = k;\pi_{\mathbf{y}})}{P(y_i;\boldsymbol{\pi})} \qquad (2)$$

2.2 Dirichlet process infinite mixtures of products of multinomials

The fully Bayesian, joint modeling (JM) approach known as "Dirichlet process mixtures of products of multinomial distributions model" (DPMPM) (Dunson and Xing, 2009) is described as:

1. Assume that each individual i belongs to exactly one of $K < \infty$ latent classes
2. For $i = 1, \ldots, n$, let $x_i \in \{1, \ldots, k\}$ indicate the class of individual i, and let $\pi_k = P(x_i = k)$. Assume further, that $\pi = \{\pi_1, \ldots, \pi_\infty\}$ is the same for all individuals. Within any class, we suppose that each of the j variables independently follows a class-specific multinomial distribution i.e. for any value $y_j \in \{1, \ldots, d_j\}$ let $\Psi_{kjy}^{(j)} = P(y_{ij} = y_j \mid x_i = k)$. Here, d_j is the the total number of categories for the variable j.

Mathematically expressing the finite mixture model as

$$y_{ij} | x_i, \Psi \overset{ind}{\sim} \text{Multinomial } (\Psi_{x_i 1}^{(j)}, \ldots, \Psi_{x_i d_j}^{(j)}) \text{ for all } i \text{ and } j \qquad (3)$$

$$x_i | \pi \sim \text{Multinomial } (\pi_1, \ldots, \pi_\infty) \text{ for all } i \qquad (4)$$

For prior distributions on Ψ and π, we have

$$\pi_k = V_k \left(\prod_{l<k} 1 - V_x \right) \text{ For } k = 1, \ldots, \infty$$

$$V_w \overset{iid}{\sim} \text{Beta } (1, \alpha)$$

$$\alpha \sim \text{Gamma } (a_\alpha, b_\alpha)$$

$$\Psi_{kj} \sim \text{Dirichlet } (a_{j1}, \ldots, a_{jd_j})$$

Here (a_α, b_α) and $(a_{j1}, \ldots, a_{jd_j})$ are analyst-supplied constants. Each element of $(a_{j1}, \ldots, a_{jd_j})$ is set to one in order to correspond to the uniform prior distribution. Following Dunson and Xing (2009), we set $a_\alpha = 0.25$ and $b_\alpha = 0.25$ and $k = 80, 150$ and 400 as numbers for the mixture components.

3 Evaluation of performance

In order to incorporate the uncertainty introduced by missing data and the imputations into the inferences, the estimates for quantities of interest obtained by analyzing each completed dataset are combined by utilizing rules proposed by Rubin (1987). Let Q be any quantity of interest (e.g. a population proportion or a probability or a regression coefficient). For $m = 1, \ldots, M$, let $q^{(m)}$ and $u^{(m)}$ be respectively the point estimate of Q in the m-th imputed data set with variance estimate $q^{(m)}$. Valid inferences for a scalar Q by combining the $q^{(m)}$ and $u^{(m)}$ according to Rubin (1987) are obtained as follows:

$$\overline{q}_M = \sum_{m=1}^{M} \frac{q^{(m)}}{M} \quad , \qquad (5)$$

$$b_M = \sum_{m=1}^{M} \frac{(q^{(m)} - \overline{q}_M)^2}{M-1} , \quad (6)$$

$$\overline{u}_M = \sum_{m=1}^{M} \frac{u^{(m)}}{M} , \quad (7)$$

$\overline{q}_M$ can be used to estimate Q and the variance of $\overline{q}_M$ can be estimated by

$$T_M = \left(1 + \frac{1}{M}\right) b_M + \overline{u}_M , \quad (8)$$

with degrees of freedom $v_M = (M-1)(1 + \frac{\overline{u}_M}{(\left(1+\frac{1}{M}\right)b_M)^2})$.

Confidence intervals can be constructed using standard multiple imputation confidence interval construction rules, which approximately follow a t-distribution. For more detail see Rubin (1996), Barnard and Meng (1999), Reiter et al. (2006), Harel and Zhou (2007).

4 Proposed hybrid architecture

Since the application of the package "NPBayesImputeCat" (Quanli et al., 2018) is limited to only categorical variables, the incomplete dataset is proposed to be partitioned into two sets, one consisting of categorical variables (*Miss.*$_{cat}$), (which MICE may not be able to impute due to reasons described in the introduction) and the other consisting of continuous variables (*Miss.*$_{num}$), where variables may be missing in both sets. A fully Bayesian JM (DPMPM) approach is used to fill in missing values by utilizing the package "NPBayesImputeCat" in *Miss.*$_{cat}$. This results in a complete version (*Imp.*$_{cat}$) of categorical variables independent of information available in the continuous variables. This complete version (*Imp.*$_{cat}$) of categorical variables can be used by MICE to construct chained equations based on categorical variables which have already been imputed by the fully Bayesian joint models to now impute the continuous variables. To achieve this, the dataset (*Miss.*$_{num}$) is added to the dataset (*Imp.*$_{cat}$) and MICE is run. This provides one completely imputed dataset where the imputations of the continuous variables obtained by FCS using chained equations depend on the information available in the imputed categorical variables. This process is repeated *M* times to obtain multiple imputed datasets using different algorithms offered by the R package "mice" (van Buuren and Groothuis-Oudshoorn, 2011) along with some prior

specifications and a number of mixture components used in the R package "NPBayesImputeCat" (Quanli et al., 2018). Algorithm 1 explains the proposed hybrid architecture in detail.

Algorithm 1: Proposed hybrid architecture

Require: *P nxp* matrix with incomplete data

1. $Miss._{cat}$, $Miss._{num}$ ← Initial division of p variables into factor and numeric subsets.
2. **for** $z = 1, \ldots, Z$ **do**
3. **for** $m = 1, \ldots, M$ **do**
4. $Imp._{cat_m}^{z}$ ← Imputing $Miss._{cat}$ using R package "NPBayesImputeCat".
5. $Imp._{cat_m}^{z}$ $Miss._{num_m}^{z}$ ← Combining $Imp._{cat_m}^{z}$ and $Miss._{num_m}^{z}$ to generate partially imputed dataset.
6. Imp_m^z ← Imputing $Imp._{cat_m}^{z}$ $Miss._{num_m}^{z}$ using R package "mice" i.e. $f(Miss._{num_m}^{z} \mid Imp._{cat_m}^{z})$
7. Imp_m^z ← Final imputed data set.
8. $\bar{q}^{(z)} \leftarrow \sum_{m=1}^{M} \frac{q^{(m)}}{M}$ Pooled point estimates[1].
9. $b^{(z)} \leftarrow \sum_{m=1}^{M} \frac{(q^{(m)} - \bar{q}^{(z)})^2}{M-1}$
10. $\bar{u}^{(z)} \leftarrow \sum_{m=1}^{M} \frac{u^{(m)}}{M}$
11. $T^{(z)} \leftarrow \left(1 + \frac{1}{M}\right) b^{(z)} + \bar{u}^{(z)}$ Pooled variances[2].
12. **end for**
13. $\bar{q} \leftarrow \sum_{z=1}^{Z} \frac{\bar{q}^{(z)}}{Z}$ Average of pooled point estimate[3].
14. $\bar{T} \leftarrow \sum_{z=1}^{Z} \frac{T^{(z)}}{Z}$ Average of pooled variance[4].

end for

1: $\bar{q}^{(z)}$ are pooled point estimates over M imputed datasets across z simulations.

2: $T^{(z)}$ are pooled variances over M imputed datasets across z simulations.

3: $\bar{q}$ is an average of pooled point estimates $(\bar{q}^{(z)})$ across z simulations.

4: $\bar{T}$ is an average of pooled variances $(T^{(z)})$ across z simulations.

5 Simulation studies

Simulation studies are conducted to examine the impact of MI by our proposed method. The incomplete data is generated as MAR with (known) effects and the number of categorical variables is kept more than the number of continuous variables, aiming to compare strategies in a realistic data situation.

We generate a sample of size $n=\{1000\}$ for five $(X_1, X_2, X_3, X_4, X_5)$ dimensional correlated random covariates from a multivariate normal distribution MVN. The marginal distributions of X_1, X_2, X_3, X_4, X_5 are normal and we set the mean and variance of each variable to 0 and 0.5 respectively. The correlation structure is given as:

$$R = \begin{pmatrix} 1 & \cdots & \rho \\ \vdots & \ddots & \vdots \\ \rho & \cdots & 1 \end{pmatrix},$$

where $\rho = 0.5$. The following component-wise threshold is used to transform random covariates into binary values.

$$X_i = \begin{cases} 0 \ if \ X_i \leq 0.5 \ , \\ 1 \ if \ X_i > 0.5 \ , \end{cases}$$

where $i=1, 2, 3, 4, 5$.

We than define $\mu_6 = -0.2X_1-0.3X_2+0.5X_3 -0.2X_4+ 0.22X_5$ and $\mu_7 = -2+\mu_6$. Outcomes for two continuous covariates are generated from normal distributions (ND) described as below:

$$X_6 \sim N(\mu_6; \sqrt{2}),$$

$$X_7 \sim N(\mu_7; \sqrt{2}).$$

We generate X_8 from Bernoulli distributions with probabilities governed by the logistic regression with $logit\ Pr(X_8) = -3+1.5X_1 -2.15X_2+2.25 X_3+1.6X_4-1.88X_5 +1.11X_6 -0.96 X_2X_3 +2.3 X_1X_3+0.5 X_2X_6-2 X_5 X_6+1.21 X_1 X_5-2.7 X_1X_2+1.2 X_1X_2X_3+3 X_6X_7$.

A covariate dependent binary response y is generated from Bernoulli distributions with probabilities governed by the logistic regression with

$logit\ Pr(y) = 0.2-0.1X_1 -0.1 X_2-0.1 X_3+0.3X_4-0.5X_5 +0.2 X_6-0.1 X_7-0.1 X_8$ and $\beta true = (0.2;-0.1 ;-0.1 ;-0.1 ;0.3 ;-0.5;0.2 ;-0.1 ;-0.1)$. We suppose that values in all covariates are missing at random with the following probabilities

$$p = 1 - \frac{e^{(-\tau - X_7)}}{(1 + e^{(-\tau - X_7)})} \ ,$$

where τ is a constant. The probabilities defined above yield about 10% to 15 % of the observations in X_j to be missing (at random) for τ=-1.5 and τ= -0.5 respectively. We repeat the process 1000 times, each time generating new binary response variables and new missing patterns. We use three purely MICE based MI methods, namely classification and regression trees (CART) (Breiman, 2001), ppredictive mean matching (PMM) (Morris et al., 2014) and the Default (which uses logistic models for categorical and PMM for continuous variables). We use two Hybrid Multiple Imputation (HMI) methods e.g. H.CART and H.DEF depending on various combinations with MICE algorithms (Default and CART) and different tuning parameters (a_α, b_α; k). We further define $H.CART_1$ which is a combination of $MICE._{CART}$ and ($a_\alpha = 0.25, b_\alpha = 0.25, k = 80$), $H.CART_2$ which is a combination of $MICE._{CART}$ and ($a_\alpha = 0.25, b_\alpha = 0.25, k = 150$) and $H.CART_3$ which is a combination of $MICE._{CART}$ and ($a_\alpha = 0.25, b_\alpha = 0.25, k =$

400). Also we define $H.DEF_1$ which is a combination of $MICE._{DEF}$ and (a_α= *0.25,* b_α= *0.25,k = 80),* $H.DEF_2$ which is a combination of $MICE._{DEF}$ and (a_α= *0.25,* b_α= *0.25,k = 150)* and $H.DEF_3$ which is a combination of $MICE._{DEF}$ and (a_α= *0.25,* b_α= *0.25,k = 400).* In order to achieve convergence and estimates from simulations in a reasonable time, a Gibbs sampler with 100 Markov-Chain-Monte-Carlo (MCMC) iterates is used. Two hundred iterations are run to insure convergence and to have the results of the simulations in a reasonable time when using the HMI methods. The R (R Core Team, 2018) version 3.0.1 is used to perform all calculations. The packages "mice" (van Buuren and Groothuis-Oudshoorn, 2011), version 2.17 and "NPBayesImputeCat" (Quanli et al., 2018), version 0.1 are used to perform MICE for continuous data and non-parametric Bayesian MI for categorical variables, respectively. Three sets of *M=10* imputed datasets are generated using MICE methods, i.e. $MICE._{PMM}$, $MICE._{DEF}$ and $MICE._{CART}$, three sets of (*M=10)* imputed datasets are generated using $H.CART_1$, $H.CART_2$ and $H.CART_3$ and three sets of *(M=10)* imputed datasets are generated using $H.DEF_1$, $H.DEF_2$ and $H.DEF_3$. The number of multiple imputations *(M=10)* is large in order to get better estimates of standard errors. Even a higher number of *M* would have been desirable but would have led to further increased computing times. Simulated root mean square errors (RMSEs), empirical standard errors (ESEs) and coverage rates of 95% confidence intervals for generalized linear models (GLM's) with binary response and mixed covariates are estimated via combining rules described above and a comparison is made among the proposed and various existing MI methods. Tables 1-2 and Tables 3-4 display the coverage rates of 95% confidence intervals (CI) and RMSEs (ESEs) for the 10% and 15% MAR datasets, respectively, across *1000* simulations. Figures 1-2 and Figures 3-4 show boxplots of the pooled point estimates and standard errors for 10% and 15% MAR datasets, across *1000* simulations respectively.

5.1 Results

As discussed, we used two HMI methods i.e. ("H.CART" and "H.DEF") for comparison with three MICE based MI methods, i.e. ("$MICE._{DEF}$", "$MICE._{CART}$" and "$MICE._{PMM}$"). In the simulation study in section 5, we generated datasets with two missing rates, i.e. 10% and 15%, using a MAR process. The HMI method "$H.DEF_1$" provides almost equal 95% CI coverage rates for the most parts and the remaining two "H.DEF" methods, i.e. ("$H.DEF_2$" and "$H.DEF_3$") provide better results for the most parts as compared to the "$MICE._{DEF}$" and "$MICE._{PMM}$" MI methods. This may imply that larger values for *k* have an effect on the overall performance of the "H.DEF" MI methods. All three MI methods based on "H.CART" provide better 95% CI coverage rates for the most parts as compared to "$MICE._{DEF}$" and "$MICE._{PMM}$", but slightly worse coverage than "$MICE._{CART}$" for some of the simulations. Surprisingly,

the coverage rates for the regression coefficient β_8 of all three "H.CART" based MI methods are higher for the 10% MAR datasets, indicating a better ability to detect complex dependency structure as compared to "MICE.$_{CART}$". See Tables 1-2. However, we observe no such real differences in the monte carlo errors (Koehler et al., 2009). This can be due to the limited number of simulation runs used. We observe for the most parts that the between imputation variations (i.e. ESEs) for all HMI MI methods are smaller compared to "MICE.$_{DEF}$" and "MICE.$_{PMM}$" and almost equal compared to "MICE.$_{CART}$". The amount of bias is also relatively low for the proposed HMI methods, see Tables 3-4. The average point estimates based on the proposed HMI methods are close to the corresponding true values in most of the cases, see Figures 1-2. Average standard errors based on the proposed HMI methods are also smaller for all cases as compared to the three MICE based MI methods, see Figures 3-4.

Table 1. Simulated data: 95% confidence intervals (CI) coverage rates for 10% MAR.

Method	β_1	β_2	β_3	β_4	β_5	β_6	β_7	β_8
MICE.$_{PMM}$	95	95	96	95	95	94	95	96
MICE.$_{CART}$	97	96	97	96	96	96	95	96
MICE.$_{DEF}$	95	95	96	96	95	96	95	95
H.DEF$_1$	96	96	96	94	95	95	96	96
H.CART$_1$	95	96	97	94	96	96	97	97
H.DEF$_2$	96	96	96	95	95	95	95	97
H.CART$_2$	95	96	96	94	97	95	96	96
H.DEF$_3$	96	96	96	94	95	94	95	97
H.CART$_3$	96	96	96	95	97	96	96	97

Table 2. Simulated data: 95% confidence intervals (CI) coverage rates for 15% MAR.

Method	β_1	β_2	β_3	β_4	β_5	β_6	β_7	β_8
MICE.$_{PMM}$	97	95	95	95	95	96	95	97
MICE.$_{CART}$	98	96	97	95	94	95	95	97
MICE.$_{DEF}$	94	95	95	96	96	96	96	96
H.DEF$_1$	97	97	97	96	96	96	95	98
H.CART$_1$	96	97	97	95	96	96	96	96
H.DEF$_2$	98	96	97	96	95	96	95	97
H.CART$_2$	96	96	96	95	96	96	96	97
H.DEF$_3$	98	96	96	96	95	96	96	97
H.CART$_3$	96	96	97	96	96	96	96	97

Table 3. Simulated data: RMSEs (ESEs) for 10% MAR**.**

Variables	Mice.$_{PMM}$	MICE.$_{DEF}$	MICE.$_{CART}$	H.DEF$_1$	H.CART$_1$	H.DEF$_2$	H.CART$_2$	H.DEF$_3$	H.CART$_3$
β_1	0.16(0.16)	0.16(0.16)	0.14(0.14)	0.15(0.15)	0.15(0.15)	0.15(0.15)	0.15(0.15)	0.15(0.15)	0.15(0.15)
β_2	0.16(0.16)	0.16(0.15)	0.15(0.15)	0.15(0.15)	0.15(0.15)	0.15(0.15)	0.15(0.15)	0.15(0.15)	0.15(0.15)
β_3	0.16(0.16)	0.16(0.16)	0.15(0.15)	0.16(0.16)	0.15(0.15)	0.16(0.16)	0.15(0.15)	0.16(0.16)	0.15(0.15)
β_4	0.16(0.16)	0.16(0.16)	0.15(0.15)	0.16(0.16)	0.16(0.16)	0.16(0.16)	0.16(0.16)	0.16(0.16)	0.16(0.16)
β_5	0.16(0.16)	0.16(0.16)	0.16(0.16)	0.16(0.15)	0.15(0.15)	0.16(0.15)	0.15(0.15)	0.16(0.15)	0.15(0.15)
β_6	0.08(0.08)	0.08(0.08)	0.08(0.08)	0.08(0.08)	0.08(0.08)	0.08(0.08)	0.08(0.08)	0.08(0.08)	0.08(0.08)
β_7	0.05(0.05)	0.04(0.04)	0.04(0.04)	0.04(0.04)	0.04(0.04)	0.05(0.05)	0.04(0.04)	0.05(0.04)	0.04(0.04)
β_8	0.19(0.19)	0.19(0.19)	0.17(0.17)	0.17(0.17)	0.17(0.16)	0.17(0.17)	0.17(0.16)	0.17(0.17)	0.17(0.16)

Table 4. Simulated data: RMSEs (ESEs) for 15% MAR**.**

Variables	Mice.$_{PMM}$	MICE.$_{DEF}$	MICE.$_{CART}$	H.DEF$_1$	H.CART$_1$	H.DEF$_2$	H.CART$_2$	H.DEF$_3$	H.CART$_3$
β_1	0.15(0.15)	0.17(0.17)	0.14(0.14)	0.14(0.14)	0.15(0.15)	0.14(0.14)	0.15(0.15)	0.14(0.14)	0.15(0.15)
β_2	0.16(0.16)	0.17(0.17)	0.15(0.15)	0.15(0.15)	0.15(0.15)	0.15(0.15)	0.15(0.15)	0.15(0.15)	0.15(0.15)
β_3	0.17(0.17)	0.17(0.17)	0.15(0.15)	0.15(0.15)	0.16(0.16)	0.15(0.15)	0.16(0.16)	0.16(0.15)	0.16(0.16)
β_4	0.16(0.16)	0.16(0.16)	0.16(0.16)	0.16(0.15)	0.16(0.16)	0.16(0.15)	0.16(0.16)	0.16(0.15)	0.16(0.16)
β_5	0.17(0.17)	0.17(0.17)	0.17(0.16)	0.16(0.16)	0.16(0.16)	0.17(0.16)	0.16(0.16)	0.17(0.16)	0.16(0.16)
β_6	0.08(0.08)	0.08(0.08)	0.08(0.08)	0.08(0.08)	0.08(0.08)	0.08(0.08)	0.08(0.08)	0.08(0.08)	0.08(0.08)
β_7	0.05(0.05)	0.05(0.05)	0.05(0.05)	0.05(0.05)	0.05(0.05)	0.05(0.05)	0.05(0.05)	0.05(0.05)	0.05(0.05)
β_8	0.20(0.20)	0.21(0.21)	0.18(0.17)	0.17(0.17)	0.18(0.17)	0.18(0.17)	0.18(0.17)	0.18(0.17)	0.18(0.17)

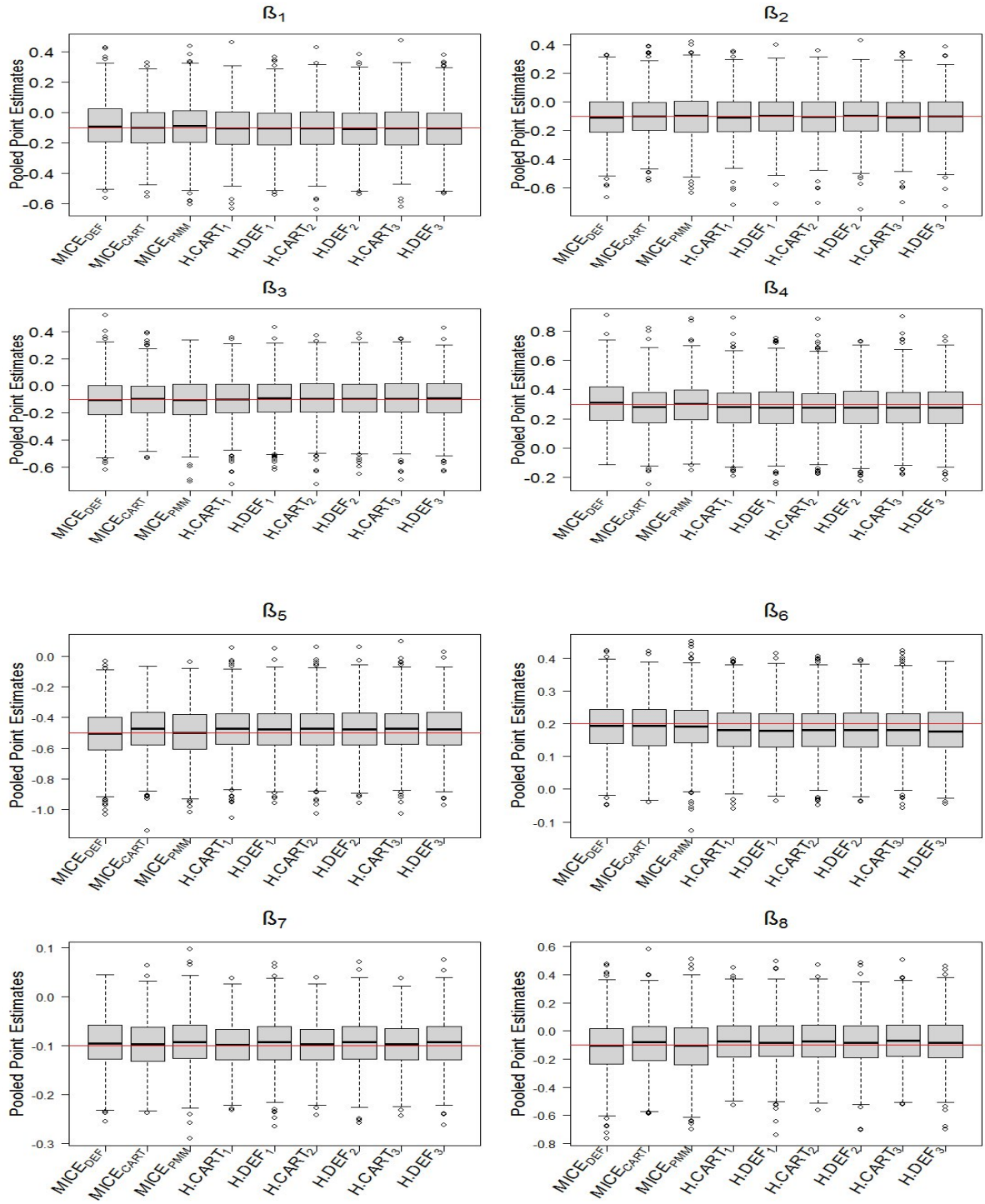

Figure 1. Simulated data: Boxplots of the pooled point estimates for 10% MAR (1000 simulations).

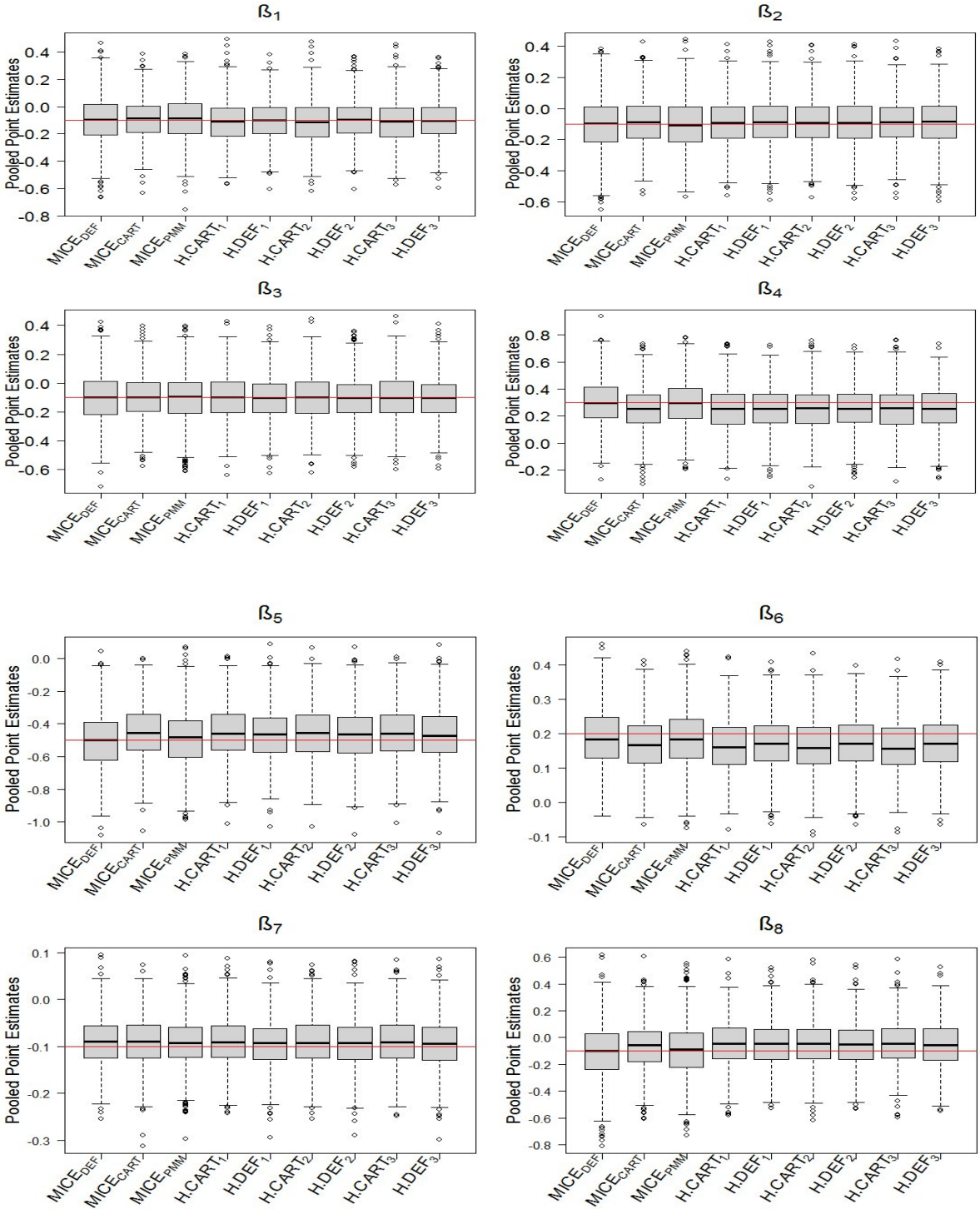

Figure 2. Simulated data: Boxplots of the pooled point estimates for 15% MAR (1000 simulations).

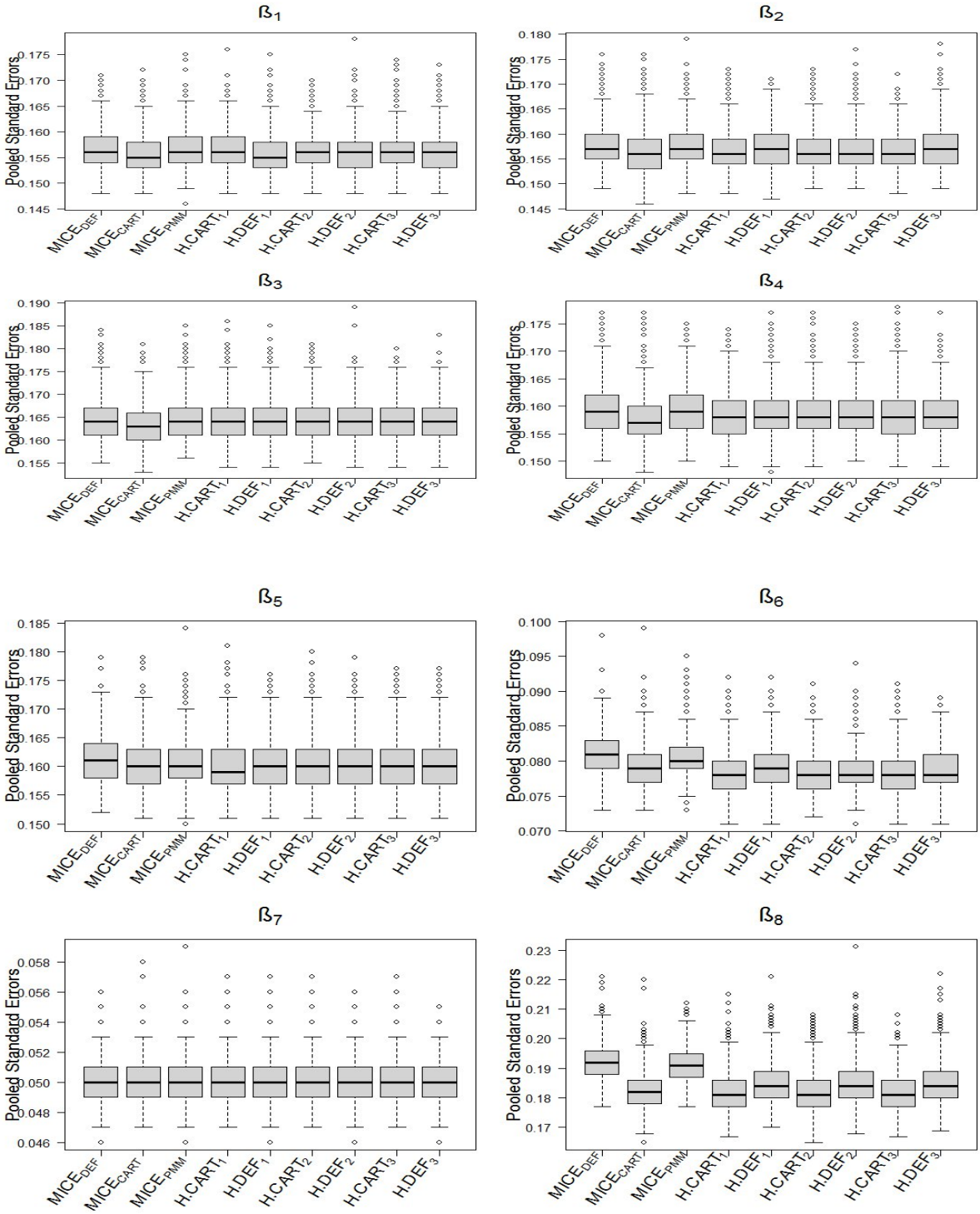

Figure 3. Simulated data: Boxplots of the pooled standard errors for 10% MAR (1000 simulations).

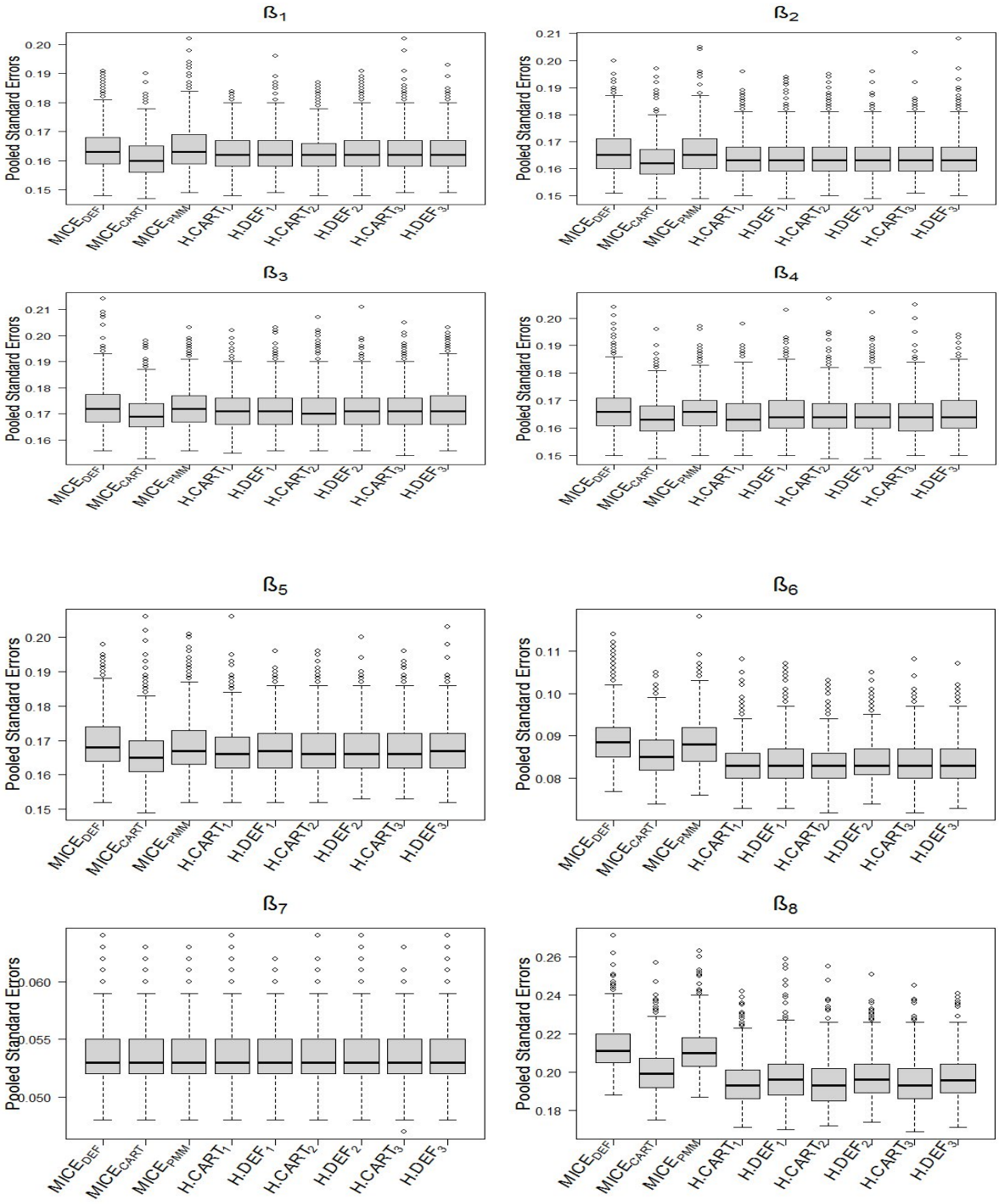

Figure 4. Simulated data: Boxplots of the pooled standard errors for 15% MAR (1000 simulations).

6 Real data-based example

6.1 Motivation

The Bureau of Statistics Punjab has conducted the Multiple Indicator Cluster Survey (MICS) Punjab, 2014, Pakistan, in collaboration with the United Nations Children's Fund (UNICEF). The Government of the Punjab has provided the major funding through the Annual Development Program 2014-15 and UNICEF has provided the annual report. The documents related to MICS Punjab consisting of the final report, key findings, survey plan, list of indicators and questionnaires can be found on the MICS website (www. http://bos.gop.pk). UNICEF in the 1990s has developed the global MICS program as an international household survey program. MICS provides support to the countries in gathering universal comparable data consisting of a wide range of indicators on the health and socio economic situation of children and women. We have used the MICS 2014 women's data that comprises more than *200* background variables on 61286 observations from 36 districts of Punjab. The data contains information of women's background characteristics like demographics, age, education, motherhood and recent births etc. Most of the background variables are categorical with lots of categories whereas few variables like age are numeric. The health benefits of breastfeeding are no longer in doubt (WHO 2003). Breastfeeding does not only contribute to the early development of a child but is also crucial for the wellbeing of the mother as well. MICS 2014 women's data can be used to determine the effect of various factors affecting the feeding practices in Punjab. This analysis could be very helpful in decision making policies regarding women and child health.

6.2 Imputation of MICS background variables

Since MICS data for women contains data with a possibly complex dependency structure, the application of the package “mice” can become problematic due to various limitations, e.g. non-convergence of the Gibbs sampler in special cases, large amount of missing values, tedious work required for specification of imputation models and interaction terms in presence of large data bases with hundreds of variables and multicollinearity problems (van Buuren and Oudshoorn, 1999). It was not possible to have a proper comparison of the proposed and existing MI approaches in such cases. Therefore, it was decided to select a subset containing *7* continuous and *37* categorical variables. The selection of variables is made according to the evidence from demographical and behavioral risk factors effecting inclination towards breastfeeding. Some of the selected categorical variables, i.e. district, has lots of categories *(k=36),* hence keeping the analysis comparable and challenging at the same time. Among these *43* variables, *5* variables have less than *14%* missing values; *16* variables have between *32* to *68* per cent missing values; *20* variables have between *80* to *95* per cent missing values. Only *2* variables are completely observed. All

variables are included in the imputation model. The reasons of missing observations in MICS data are typical, i.e. nonresponse, don't know, not reached, etc. For the sake of multiple imputations, all reasons for item nonresponse are treated as MAR.

The whole process of creating imputations is repeated twenty times and *M=10* completed datasets are generated for each MI method. The binary response (Ever Breastfeed), which compromises two categories (Yes / No), is finally modeled using a GLM analysis model depending on four categorical variables (Mother Ever Attended School: two categories, Delivery by C Section: two categories, Satisfaction from Health: two categories, Area: two categories) and two continuous covariates (Age of Mother and Freq. of Mother Reads New). The R package "VIM" (Templ et al., 2012) is utilized to explore the pattern of missing values. Figure 5 displays the proportion of missing values and the missing data pattern for the variables used in the analysis model. Since there are no true values to compare for in the real data example, we calculated complete case (CC) estimates for comparison purposes (Table 5). The time taken by each MI method is shown in Table 6. Boxplots of the pooled point estimates and standard errors for the real data are shown in Figures 6 and 7 respectively.

6.3 Results

Figure 5 in the real data example displays the bar plot on the left side which shows the proportions of missing values in the predictors. The categorical predictor "Delivery By C Section" has the highest amount of missing values (i.e. more than 80%) followed by "Ever Breastfeed" (about 80%), "Satisfaction From Health" (about 60%) and "Freq. of Mother Reads New" (about 40%). The amount of missing values is rather small for "Mother Ever Attended School" and "Age" (i.e. less than 20%). The categorical predictor "Area" has no missing values. An aggregation plot on the right side shows all existing combinations of missing (red) and imputed observed (blue) values. The frequencies of different combinations can be seen by a small bar plot on the right side (Templ et al., 2012). The aggregation plot reveals that if missing values occur in the variable "Ever Breastfeed", they most often also occur in the variables "Satisfaction From Health", "Freq. of Mother Reads New" and "Delivery By C Section". We note, that the standard errors for most of the coefficients are smaller relative to the (absolute) point estimates under all MI methods (see Figures 6-7). We noticed that point estimates in $MICE._{CART}$ are nearer to the estimates in complete case analysis for most of the cases as compared to the hybrid methods (see Table 5). In the real data example, the HMI methods tend to show smaller pooled standard errors for most of the co-variates as compared to the MICE methods. We see, that when HMI MI methods are applied to the real data set, the pooled standard errors are comparatively smaller for all covariates as compared to the "$MICE._{DEF}$" MI method and smaller for most the covariates (i.e. "Age", "Freq. of Mother

Reads New", "Delivery By C Section" and "Area") as compared to the "MICE.$_{PMM}$" MI method. "H.CART" tends to show smaller pooled standard errors for the covariates (i.e. "Age" and "Delivery by C Section") as compared to its counterparts. For the rest of the covariates, the differences are also not so high, which suggests a reasonable performance compared to MICE, see Figures 6-7. The computational burden is significantly reduced for most of the settings using the proposed HMI methods, see Table 6.

Table 5. Real data: complete case (CC) estimates

Variables	est	se
Age	0.14	0.06
Mother Attended School	-0.59	0.77
Freq. Mother Reads News	-0.09	0.15
Delivery by C Section	0.43	0.25
Satisfaction From Health	0.27	0.27
Area	0.16	0.25

The "est" and "se" denote the point estimates and standard errors of the coefficients of the GLM, respectively.

Table 6. Real data: Time taken by various MI methods.

Method	Time
MICE.$_{CART}$	4.20_d
MICE.$_{PMM}$	3.52_d
MICE.$_{DEF}$	3.14_d
H.DEF$_1$	1.70_d
H.CART$_1$	1.62_d
H.DEF$_2$	1.68_d
H.CART$_2$	1.64_d
H.DEF$_3$	1.82_d
H.CART$_3$	1.77_d

Note: Time = the time to complete 10 multiple imputation by variants of MI across 20 simulations and d = days.

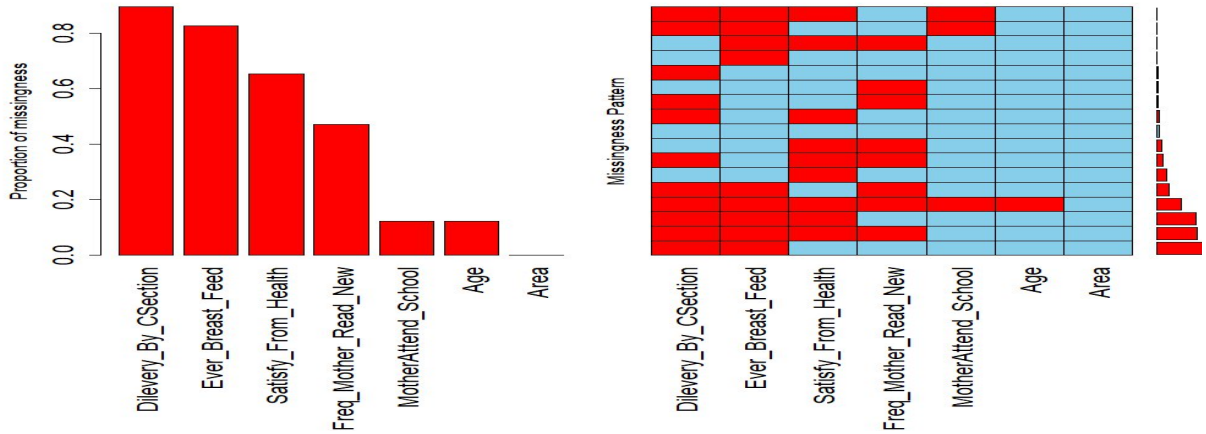

Figure 5. Aggregation graphic of the incomplete covariates.

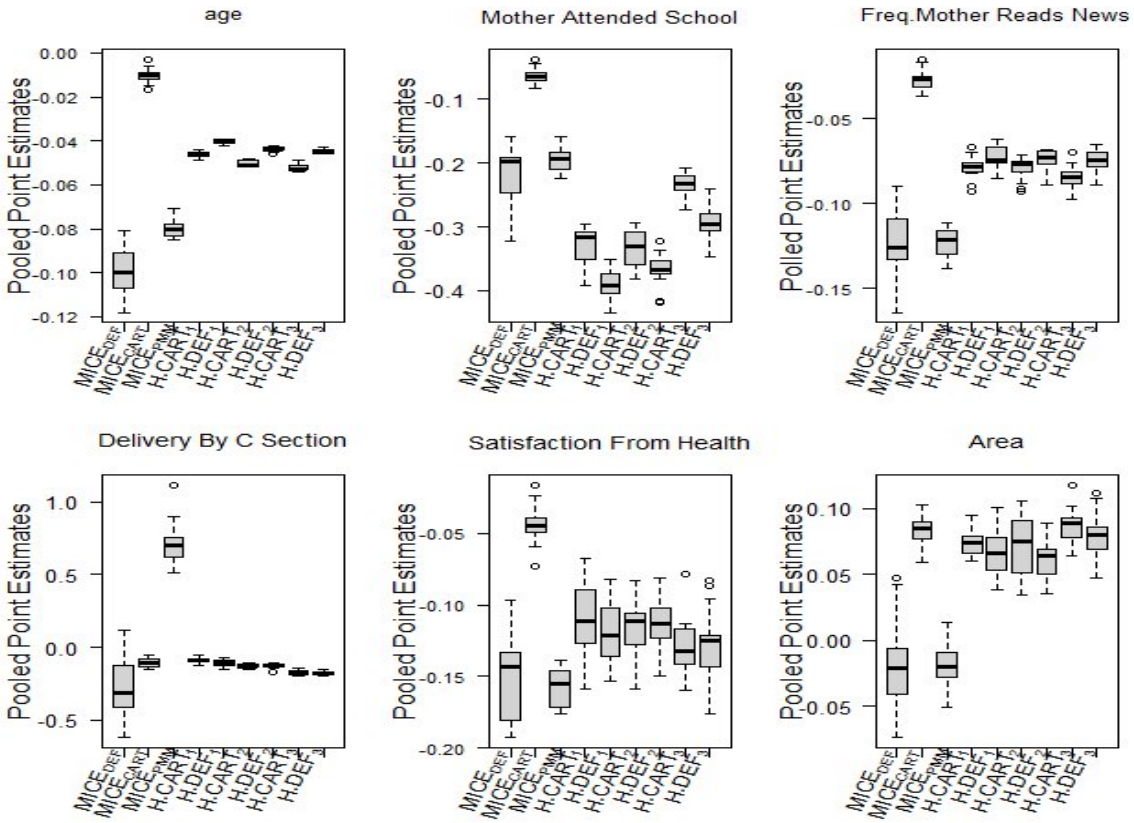

Figure 6. Real data: Boxplots of the pooled point estimates.

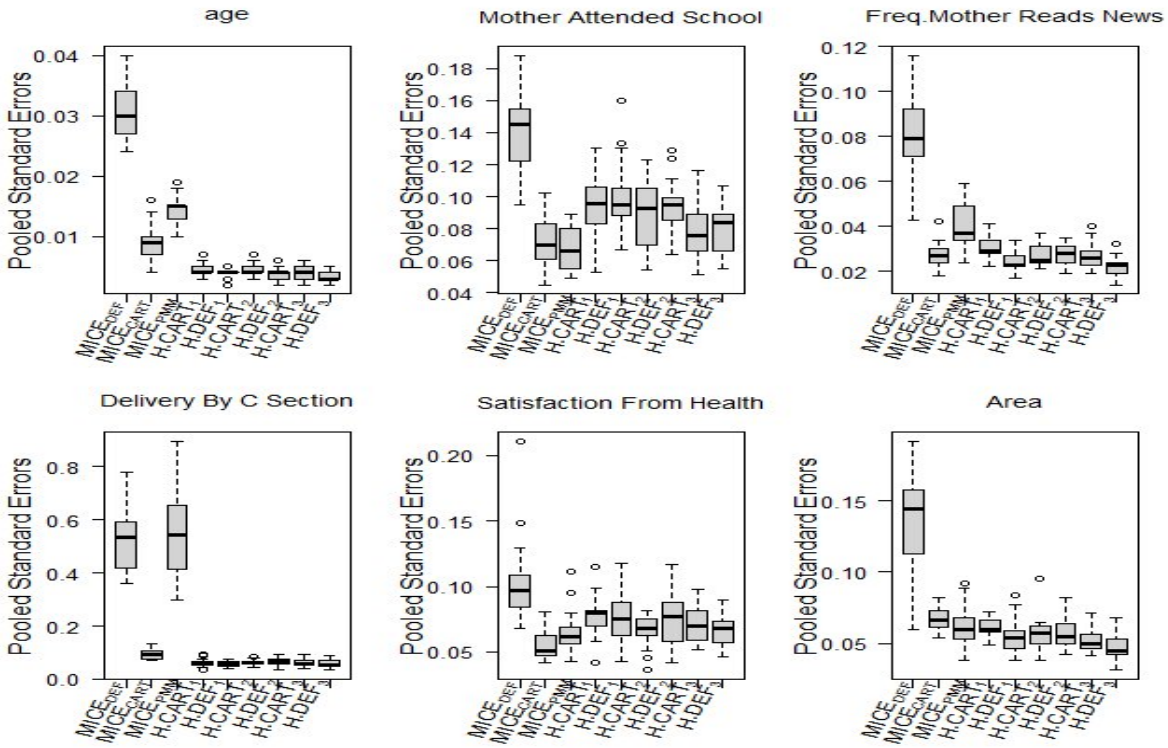

Figure 7. Real data: Boxplots of the pooled standard errors.

7 Concluding remarks

The superiority of CART and the JM technique DPMPM over the default MI methods in MICE is already established in Akande et al. (2017). Results from simulations and a real data example show that for most of the cases, hybrid techniques tend to perform better not only than the default MI methods in MICE, but also than the remaining MICE options in the presence of mixed type variables, at least for the considered GLM analysis model with binary response. The statistical properties of the proposed approach can be further studied for continuous response with mixed type covariates. In this method, chained equations used to multiply impute continuous variables are made dependent on categorical variables which have been already multiply imputed by DPMPMs through a conceptually simple method. The user can choose a set of incomplete categorical covariates that the regular MICE can sometimes fail to impute due to various restrictions, i.e. large datasets, complex dependencies, high percentage of missing data, specification of higher order interactions, multicolinearity and other instability problems. Missing values in categorical variables can be imputed by a nonparametric MI approach called DPMPMs. After filling

the categorical variables, these variables are replaced in the original dataset in order to perform regular MICE. This method combines MI by chained equations and mixtures of multinomial distributions. This approach could be very appropriate for a large number of variables with complex association structures, especially coming from large sample surveys. To implement this method, no knowledge of complicated models is required. Various combinations of prior specifications and the maximal number of mixture components can be chosen together with the appropriate MICE algorithms to achieve better coverage rates and point estimates. We have observed that increasing the maximal number of mixture components tends to result in better coverage rates compared to most of the MICE methods in many cases. The proposed method is more flexible in specifying higher order interactions in the model. It also eliminates the use of predictor selection beforehand. Further comparisons can be made for data with ordinal nature and more categories with large values of prior specifications. Our proposed method is also computationally inexpensive and requires less time even when performed with a large number of iterations. Since most of the educational and health surveys contain lots of categorical and comparatively less continuous variables, various organizations can use this imputation method to create completed datasets without understanding the complexity of the dependency and model structures. However, of note, one limitation of the proposed method is, that the information available in the continuous variables is not used for imputing the categorical variables. Therefore, it is too early to make any final conclusion until unless experiments with diversity of settings are conducted.

References

Arnold, B. C. and Press, S. J. 1989. "Compatible Conditional Distributions". *Journal of the American Statistical Association* 84:152-156.

Allison P. D. 2002. *Missing Data*. Thousand Oaks. CA: Sage Publications.

Abdella, M. and Marwala, T. 2005. "The use of genetic algorithms and neural networks to approximate missing data in database". In Proceedings of the IEEE 3rd International Conference on Computational Cybernetics, 2005. 24: 207-212.

Ankaiah, N. and Ravi, V. 2011. "A novel soft computing hybrid for data imputation". In Proceedings of the 7th International Conference on Data Mining (DMIN). Las Vegas. USA.

Akande, O., Li, F. and Reiter, J. 2017. "An empirical comparison of multiple imputation methods for categorical data". *The American Statistician* 71: 162–170.

Andridge, R.R. and Little, R.J.A. 2017. "A Review of Hot Deck Imputation for Survey Non-response". *International statistical review* 78(1): 40-64.

Bengio, Y. and Gingras, F. 1995. "Recurrent neural networks for missing or asynchronous data. In Touretzky, D.S., Mozer, M.C. and Hasselmo, M.E. editors". *Advances in Neural Information Processing Systems* 8: 95–401. MIT Press, Cambridge, MA.

Barnard, J. and Meng, X. 1999. "Applications of multiple imputation in medical studies: From AIDS to NHANES". *Statistical Methods in Medical Research* 8:17-36.

Breiman, L. 2001. "Random Forests". *Machine Learning* 45(1): 5-32.

Batista, G. and Monard, M.C. 2003. *Experimental comparison of K-nearest neighbour and mean or mode imputation methods with the internal strategies used by C4.5 and CN2 to treat missing data*. University of Sao Paulo.

Chung, D. and Merat, F.L. 1996. Neural network based sensor array signal processing. In: Proc Int Conf Multisens Fusion Integr Intell Syst. Washington. USA 757–764.

Chandra, A., Martinez, G.M., Mosher, W.D., Abma, J.C. and Jones, J. 2005. "Fertility, family planning, and reproductive health of U.S. women: data from the 2002 National Survey of Family Growth". *Vital Health Stat* 23: 1-160.

Corsi, D.J., Perkins, J.M., Subramanian, S.V. 2017. "Child anthropometry data quality from Demographic and Health Surveys, Multiple Indicator Cluster Surveys, and National Nutrition Surveys in the West Central Africa region: are we comparing apples and oranges?". *Global Health Action* 10:1328185.

Dunson, D. B. and Xing, C. 2009. "Nonparametric Bayes modeling of multivariate categorical data". *Journal of the American Statistical Association* 104:1042-1051.

Gelman, A. and Speed, T. P. 1993. "Characterizing a joint probability distribution by conditionals". *Journal of the Royal Statistical Society Series B: Statistical Methodology* 55: 85-188.

Graham, J. W. and Schafer, J. L. 1999. "On the performance of multiple imputation for multivariate data with small sample size. In R. Hoyle (Ed.)". *Statistical strategies for small sample research* 1-29.

Gulliford, M.C., Ukoumunne, O.C. and Chinn, S. 1999. "Components of Variance and Intra class Correlations for the Design of Community-based Surveys and Intervention Studies: Data from the Health Survey for England". *American Journal of Epidemiology* 149(9): 876-883.

Harel, O. and Zhou, X.H. 2007. "Multiple imputation: Review of theory, implementation and Software". *Statistics in Medicine* 26: 3057-3077.

Horton, N.J. and Kleinman, K.P. 2007. "Much ado about nothing: a comparison of missing data methods and software to fit incomplete regression models". *The American Statistician* 61: 79-90.

Honaker, J., King, G., and Blackwell, M. 2011. "Amelia II: A program for missing data". *Journal of Statistical Software* 45(7):1-47.

Kohonen, T. 1995. *Self-Organizing Maps.* Springer. Heidelberg.

Koehler, E., Brown, E. and Haneuse, S.J. 2009. "On the Assessment of Monte Carlo Error in Simulation-Based Statistical Analyses". *The American Statisticians* 63(2):155-162.

Lazarsfeld, P. F. 1950. *The logical and mathematical foundation of latent structure analysis. In S. A. Stouffer, L. Guttman, E. A. Suchman, P. F. Lazarsfeld, S. A. Star, & J. A. Clausen, Studies in social psychology in World War H: Vol. 4. Measurement and prediction (chap. 10, pp. 362-412).* Princeton, NJ: Princeton University Press.

Little, R. J. A. 1988. "A Test of Missing Completely at Random for Multivariate Data with Missing Values". *Journal of the American Statistical Association* 83(404): 1198-1202.

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

Li, F., Yu, Y., Rubin, D. B. 2012. *Imputing missing data by fully conditional models: some cautionary examples and guidelines*. Duke University Department of Statistical Science Discussion Paper 11–24.

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

Marseguerra, M. and Zoia, A. 2005. "The autoassociative neural network in signal analysis. II. Application to on-line monitoring of a simulated BWR component". *Annals of Nuclear Energy* 32(11):1207–1223.

Marwala, T. and Chakraverty, S. 2006. "Fault classification in structures with incomplete measured data using auto associative neural networks and genetic algorithm". *Current Science India* 90(4):542–548.

Morris, T.P., Ian, R.W. and Patrick, R. 2014. "Tuning Multiple Imputation by Predictive Mean Matching and Local Residual Draws. *BMC Medical Research Methodology* 14 (1): 75.

Murray, J. S. and Reiter, J. P. 2016. "Multiple imputation of missing categorical and continuous values via Bayesian mixture models with local dependence". *Journal of the American Statistical Association* 111: 1466 - 1479.

Narayanan, S., Vian, J. L., Choi, J., El-Sharkawi, M. and Thompson, B.B. 2002. *Set constraint discovery: missing sensor data restoration using auto-associative regression machines*. In Proceedings of the international Joint Conference on Neural Networks (IJCNN). 2872–2877. Honolulu.

Oja E. and Kaski, S. 1999. *Kohonen Maps*. Elsevier. Amsterdam.

Pyle, D. 1999. *Data preparation for data mining*. Morgan Kauf- mann Publishers Inc. San Francisco.

Pérez, A., Dennis, R.J., Gil, J.F., Rondón, M.A. and López, A. 2002. "Use of the mean, hot deck and multiple imputation techniques to predict outcome in intensive care unit patients in Colombia". Statistics in Medicine 21:3885-3896.

Quanli, W., Danial, M.V., Reiter, J.P. and Jigchen, H. 2018. *NPBayesImputeCat: Non-Parametric Bayesian Multiple Imputation for Categorical Data.* R package version 0.1, https://CRAN.R-project.org/package=NPBayesImputeCat.

Rubin, D. B. 1976. "Inference and Missing Data". *Biometrika* 63: 581-590.

Rubin, D.B. 1987. *Multiple Imputation for Nonresponse in Surveys*. Wiley, New York.

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

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

Raghunathan, T.W., Lepkowksi, J.M., Van Hoewyk, J. and Solenbeger, P. A. 2001. "Multivariate technique for multiply imputing missing values using a sequence of regression models". *Survey Methodology* 27: 85-95.

Reiter, J. P., Raghunathan, T. E. and Kinney, S. 2006. "The importance of modeling the survey design in multiple imputation for missing data". *Survey Methodology* 32: 143-149.

Royston, P. and White, I.R. 2011. "Multiple imputation by chained equations (mice): Implementation in Stata". *Journal of Statistical Software* 45(4): 1-20.

R Core Team. 2018. *R: A Language and Environment for Statistical Computin.* Vienna, Austria: R Foundation for Statistical Computing.

Sharpe, P. K. and Solly, R. J. 1995. "Dealing with missing values in neural network-based diagnostic systems". *Neural Computing and Applications* 3(2):73–77.

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

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

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

Si, Y. and 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.

Templ, M., Andreas, A., Alexander, K. and Bernd, P. 201. *VIM: Visualization and Imputation of Missing Values.* http://cran.r-project.org/web/packages/VIM/VIM.pdf.

van Buuren, S. and Groothuis-Oudshoorn, K. 1999. *Flexible multivariate imputation by MICE*. TNO Prevention and Health. Leiden.

van Buuren, S. 2007. “Multiple imputation of discrete and continuous data by fully conditional specification”. *Statistical Methods in Medical Research* 16: 219-242.

van Ginkel, J. R. 2007. *Multiple imputation for incomplete test, questionnaire and survey data*. Ph.D. dissertation. Tilburg University. Dept. of Methodology and Statistics.

Vermunt, J. K., van Ginkel, J. R., van der Ark, L. A. and Sijtsma, K. 2008. “Multiple imputation of incomplete categorical data using latent class analysis”. *Sociological Methodology* 38: 369-397.

van Buuren, S., and Groothuis-Oudshoorn, K. 2011. “mice: Multivariate imputation by chained equations”. *R. Journal of Statistical Software* 45(3):1-67.

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

World Health Organization (WHO). 2003. *Community-based Strategies for Breastfeeding Promotion and Support in Developing Countries, 2003. Dept. of child and adolescent health and development.* Geneva.

Zhu, J. and Raghunathan, T. E. 2016. “Convergence Properties of a Sequential Regression Multiple Imputation Algorithm”. *Journal of the American Statistical Association* 110(511): 1112-1124.