LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

INSTITUT FÜR STATISTIK

Humera Razzak, Christian Heumann

THE ABILITY OF DIFFERENT IMPUTATION METHODS TO CAPTURE COMPLEX DEPENDENCIES IN HIGH DIMENSIONS

Technical Report Number 230, 2019
Department of Statistics
University of Munich

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

THE ABILITY OF DIFFERENT IMPUTATION METHODS TO CAPTURE COMPLEX DEPENDENCIES IN HIGH DIMENSIONS

Humera Razzak (Humera.Razzak@stat.uni-muenchen.de)
Department of Statistics, LMU Munich.

Christian Heumann (chris@stat.uni-muenchen.de)
Department of Statistics, LMU Munich.

ABSTRACT

Multiple-imputation (MI) is a method for treating the problem of missing data. There are various competing computational algorithms available in the *R* environment to address missing data problems of categorical and continuous variables. In the case of a high amount of missing information, large sample sizes and complex dependency structures among categorical variables, the utility of the provided *R* packages is somewhat limited. A computationally expedient, fully Bayesian, joint modeling (JM) approach known as "Dirichlet process mixtures of multinomial distributions" (DPMD), automatically models complex dependencies among variables. But this approach is limited to categorical variables only. We propose a simple and easy to implement combining algorithm which imputes continuous variables using various algorithms and uses the JM approach to detect complex dependency structures among categorical variables. We review, describe and evaluate software packages commonly available in *R* and compare the results with the proposed MI method by using as example an artificial data set. The results suggest that the MI approach which combines the JM approach and various algorithms based on generalized linear models dominates various algorithms when applied solely.

Keywords: Survey data; Multiple Imputation; Complex dependencies; Hybrid; Dirichlet process prior distributions, *R* - project.

1. INTRODUCTION

Item non response is a main problem in large scale surveys. Such surveys usually have a large number of categorical variables as compared to the number of continuous variables. Using only the available data results in decreased efficiency and possibly biased inference. Rubin (1987) has proposed multiple-imputation (MI), a method for handling missing data, more than 40 years ago. For more details, see Rubin (1987) and Schafer (1997).

MI requires random draws from the posterior distribution of the missing data given the observed data. Although this method is conceptually simple but can lead to potentially unsound imputations when there are mixed type variables (i.e. continuous and categorical variables with many categories). There exist various competing computational algorithms to impute data. There is a need to investigate which of these algorithms outperform the others with respect to MI in the presence of complex dependencies among categorical variables in large scale surveys. A fully Bayesian, joint modeling approach called "Dirichlet process mixtures of multinomial distributions" (DPMD) for multiple imputation (MI) for categorical data (Si and Reiter, 2013) in large scale surveys automatically models complex dependencies while being computationally efficient at the same time. Akande et al. (2017) have compared repeated sampling properties of various MI methods for categorical data. They found that chained equations using Classification and Regression Trees (CART), and a fully Bayesian approach based on Dirichlet Process

mixture models dominate the default chained equations approaches based on Generalized Linear Models (GLM's). The DPMD MI approach is limited to categorical variables; but it is possible to impute categorical variables with complex dependencies and high dimensions using DPMD and continuous variables with existing MI methods by combining two approaches. In this paper we propose a hybrid MI (HMI) approach which combines DPMD and existing MI approaches by imputing categorical variables with DPMD and use various imputation techniques to impute the continuous variables. In this paper, we compare the performance of existing and proposed MI methods in the presence of complex dependency structures among categorical variables. The judgment about the performance will be based on various dimensions, such as accuracy in comparison with the true values, point estimates and standard errors for the fitted GLM's and coverage rates of 95% confidence intervals.

2 NOTATIONS AND ASSUMPTIONS FOR THE MISSING MECHANISMS

Let *D* denote the incomplete data with sample size *n* and *p* variables. The distribution of *D* is an arbitrary multivariate distribution.
Also assume *i* and *j* refer to observations where *i=1,…,n* and variables *j=1,…,p,* respectively. There are two components of the data set $D= \{D^{obs}, D^{miss}\}$. A response indictor matrix with same dimensions as *D* is

$$\mathrm{R}_{ij} = \begin{cases} 0 \;\; if \; v_{ij} \; is \;\; missing \\ 1 \;\; if \; v_{ij} \; is \;\; observed \end{cases}$$

Note that we use R in atelic for the R environment in this article. Missing Completely At Random (MCAR) is one possible assumption where $Pr(\mathrm{R}|D^{miss}, D^{obs}) = Pr(\mathrm{R})$. The second possible assumption is Missing At Random (MAR) where $Pr(\mathrm{R}|D^{miss}, D^{obs}) = Pr(\mathrm{R}|D^{obs})$. Missing Not At Random (MNAR) is another possible assumption where $Pr(\mathrm{R}|D^{miss}, D^{obs}) \neq Pr(\mathrm{R}|D^{obs})$ and depends on D^{miss}. The third assumption is also called non-ignorable (NI) (Little and Rubin, 2002) and not further used in the paper.

3 IMPUTATION SOFTWARE

Various imputation algorithms are implemented in a variety of statistical packages to handle missing data and to perform MI. Many standard statistical packages i.e., *R*, S-Plus, SAS, SPSS, and STATA not only implement standard algorithms but also offer user-written programs to facilitate a variety of more elaborated methods to handle missing data. Readers who are interested in the comparison of the performances of these packages are suggested to read Yu et al. (2007) or Horton and Kleiman (2007). We take *R* under consideration in this paper due to its open source character and its popularity. NA's are used to indicate missing values in *R*. There are various statistical packages that use *R* environment to impute missing data. For example "Amelia II" implements MI by bootstrapping and Expectation Maximization (EM) algorithm, "Hmisc" implements MI using additive regression and bootstrapping, *R* package "mi" offers various features (e.g. choice of predictors, models, and transformations for chained imputation models etc.) for imputations, "mice" algorithm can impute mixed type data and offer various diagnostic functions to inspect the quality of the imputations,"yaImpute" performs nearest neighbor-based imputation, "mix" performs MI for mixed categorical and continuous data, "NPBayesImpute"

impute categorical data by using Dirichlet process mixtures of multinomial distributions, “norm” uses multivariate normal model for imputations, “pan” is a MI technique for multivariate panel or clustered data. The “mitools” is a useful package to combine the results from MI whereas the package “VIM” can be utilized for exploring data and the pattern of missing values. We use “Amelia II”, “Hmisc”, “mice” and "NPBayesImpute” in our examples.

4 REVIEW OF EXISTING APPROACHES

There is a wide range of imputation models available which are based on the missingness patterns. These approaches can be categorized according to the data types. In case of a monotone missing pattern, simple methods, i.e. “propensity” (Rosenbaum and Rubin, 1983) or “Predictive Mean Matching” (PMM) (Little, 1988), are used for continuous variables. Markov Chain Monte Carlo (MCMC) approaches use Markov chains to generate random draws from multidimensional probability distributions. One can obtain a sample of the desired distribution by recording states from the chain (Gilks, 1995). MCMC approaches are suggested for complicated missingness patterns. The MCMC approach has few downsides; it is complicated and usually requires more time. Statistical packages “SAS”, “S-Plus” and “*R*” etc. use the MCMC approach. Multivariate normality assumptions apply to both the predictive mean matching and MCMC approaches (Horton and Lipsitz, 2001). According to Schafer (1997), inferences based on this normality assumption can be robust for minor departures.

Discriminant analysis or logistic regression are preferred for discrete variables for monotone missing pattern. There are a variety of imputation methods for categorical data in high dimensions. For details, see Vermunt et al. (2008). Log-linear models may be the preferred method for discrete variables, since arbitrary complex dependency structures can be modeled. But the implementation of this approach becomes difficult or impossible in high dimensions (Erosheva, et al., 2002). Naturally, there are a large number of possible models in high dimensions which makes model selection very challenging and makes it also impossible to select a model from all possible log-linear models as well. In this situation, implementation of automated model selection procedures becomes unavoidable. Moreover, model selection procedures become more complicated with missing data. Maximum likelihood estimates of the log-linear model coefficients can be biased in high dimensions (Bishop et al., 1975).

Imputation methods like fully normal (FN) imputation (Rubin and Schenker, 1986) convert categorical data to multivariate normal or continuous by applying rounding techniques. But there are evidences that the performance of these methods is limited. For example, an imputed value when made “plausible” using rounding, can tend to generate a bias and the method can fail even in low dimensions (Ake, 2005; Allison, 2000; Bernaards et al., 2007; Finch, 2010; Graham and Schafer, 1999; Horton et al., 2003; Yucel et al., 2011). Below we discuss in detail the MI algorithms we used for comparison purposes. Advantages and disadvantages of the algorithms are discussed as well.

4.1 EXPECTATION-MAXIMIZATION WITH BOOTSTRAPPING (EMB) USED BY AMELIA II

R package called ‘Amelia II’ by Honaker et al. (2011) implements imputation method. Amelia assumes that all variables in data set are distributed multivariate normally. ‘Amelia II’ combines the bootstrap (Efron, 1979) with the EM algorithm (Dempster, Laird, and Rubin, 1977). The combination of the expectation-maximization algorithm and bootstrapping is called

the Expectation-Maximization with Bootstrapping (EMB) algorithm. The bootstrapping method works by utilizing the observed sample as the pseudo-population and randomly drawing a subsample of size *n* with replacement from this observed sample. The EMB algorithm consists of the following steps: First: assuming a data set with q observed and n - q missing values**,** bootstrap samples of size *n* are drawn from incomplete data *M* times by applying bootstrapping method. Second: *M* point estimates of μ and Σ are calculated by applying the EM algorithm to each of these *M* bootstrap samples. Maximization steps are iterated until estimates converge. Finally, *M* multiply-imputed data sets are constructed by repeating this process *M* times (Wooldridge, 2002). For more details on the expectation maximization with bootstrapping (EMB) algorithm see (Schafer, 1997; Watanabe and Yamaguchi, 2000; Little and Rubin, 2002). Although EMB is computationally more efficient as compared to MCMC methods but is only an approximate Bayesian procedure (Lin, 2008).

4.2 MIXTURE MODELS FOR MULTIPLE IMPUTATION

To impute high-dimensional categorical data with significant item non-response, one has to face the challenges of model selection and estimation of log-linear models. Moreover, log-linear models and sequential regression techniques become computationally inefficient and potentially biased when the number of possible models becomes very large. Therefore, a MI technique is preferred that not only addresses these difficulties but also has a theoretical grounding as a coherent Bayesian joint model and tackles all sources of uncertainty, including parameter estimation and inference, see Rubin (1987). According to Si and Reiter (2013), Bayesian models incorporate such uncertainty automatically. They propose to use the fully Bayesian, joint modeling (JM) approach known as "Dirichlet process mixtures of products of multinomial distributions model" (DPMPM) which was originally proposed by Dunson and Xing (2009). DPMPM is a nonparametric Bayesian model for multivariate unordered categorical data. Below we describe categorical data imputation using DPMPM. A brief description is given how this approach can be combined with existing approaches through a flexible and easy to implement architecture.

Assume, we have item non-response in *n* individuals with *p* variables C_{ij} i.e. (value of variable *j* for individual *i*, where each *i* belongs to exactly one of $K < \infty$ latent classes). Further assume for $i = 1, \ldots, N$, we have the class z_i of individual i i.e. $z_i \in \{1, \ldots, K\}$ with probability $\pi_k = Pr(z_i = k)$. Let $\pi = \{\pi_1, \ldots, \pi_k\}$ be the same for all individuals. We suppose that within any class, each of the *p* variables independently follows a class-specific multinomial distribution. For any value $c_j \in \{1, \ldots, d_j\}$, let $\Upsilon_{klj}^{(j)} = Pr(C_{ij} = c_j | z_i = k)$. We can express the finite mixture model mathematically as $C_{ij}|z_i, \Upsilon \widetilde{ind} Multinomial\left(\Upsilon_{z_i1}^{(j)}, \ldots, \Upsilon_{z_i d_j}^{(j)}\right)$ for all *i* and *j* and $z_i|\pi \sim Multinomial(\pi_1, \ldots, \pi_k)$ for all i. For prior distributions on Υ and π, we have $\pi_k = V_k \left(\prod_{l<k} 1 - V_l\right)$ for $k = 1, \ldots, K$ and $V_k \sim Beta(1, \alpha)$ for $k = 1, \ldots, K-1, V_k = 1$. Finally we have $\alpha \sim Gamma(a_\alpha, b_\beta)$ and $\left(\Upsilon_{k1}^{(j)}, \ldots, \Upsilon_{kd_j}^{(j)}\right) \sim Dirichlet(a_{j1}, \ldots, a_{jd_j})$. In order to get complete data sets, first the latent class indicator for each individual is drawn from the full conditional and then, second, each missing C_{ij} is drawn from class-specific, independent categorical distributions.

This approach is consistent (i.e. any multivariate categorical data distribution can be approximated by DPMPM for a sufficiently large number of mixture (Dunson and Xing, 2009)), is computationally efficient and easy to code. The *R* package, "NPBayesImpute" by Manrique-Vallier et al. (2014) implements this approach. Shortcoming of this package is that it only takes categorical variables into account.

4.3 FULLY CONDITIONAL SPECIFICATION (FCS): CHAINED EQUATIONS

The FCS approach is another approach to multiple imputation. Multivariate missing data is imputed on a variable-by-variable basis. We specify a multivariate distribution $Pr(D, \mathrm{R} \mid \theta)$ using a series of conditional densities $Pr(D_j \mid D_{-j}, \mathrm{R}, \lambda_j)$ where λ is the unknown parameter of the imputation model. An imputation model is specified for each variable, depending on the observed values in the dataset and the response mechanism, i.e $Pr\left(D^{mis} | D^{obs}, \mathrm{R}\right)$ in our setting. A simple draw is made using the marginal distributions first. Then imputation is repeated over the conditionally specified imputation models (van Buuren, 2012). Imputations are created for each variable iteratively. Multivariate Imputation by Chained Equations (MICE) is a prominent conditionally specified imputation model. MICE works as follows.

1 Specify an imputation model for each variable D_j

$$Pr(D_{j,miss} \mid D_{j,obs}, D_{-j}, \mathrm{R}).$$

2 Let $\widetilde{D_{j,0}}$ be the starting imputation for each variable *j*. This value is e.g. obtained by making random draws from the observed values $D_{j,obs}$.

3 Repeat this process for *t=1,…,T* and *j=1,…,p* as well.

4 Draw $\widetilde{\lambda_{j,t}} \sim Pr(\lambda_{j,t} \mid D_{j,obs}, \widetilde{D_{-j,t}}, \mathrm{R})$.

5 At the end draw imputations

$$\widetilde{D_{j,t}} \sim Pr(D_{j,miss} \mid D_{j,obs}, \widetilde{D_{-j,t}}, \mathrm{R}, \widetilde{\lambda_{j,t}}).$$

MICE uses logistic or multinomial logistic regression models for categorical variables. Similar to log-linear models, these conditional models suffer from model selection and estimation problems in high dimensions. Moreover, it is very time consuming to specify many conditional models when the number of variables is large. This can result in biased estimates if default settings are used for chained equations, i.e. when we are ignoring interaction effects in the conditional models and hence fail to capture complex dependencies (Vermunt et al., 2008). The *R* Package, "mice" 2.13 (van Buuren and Groothuis-Oudshoorn, 2011) implements the FCS algorithm.

4.4 ADDITIVE REGRESSIONS, BOOTSTRAPPING AND PREDICTIVE MEAN MATCHING TECHNIQUES

Additive regressions, bootstrapping and predictive mean matching techniques for MI are implemented in the "Hmisc" package using "aregImpute" functions. A brief summary of the steps used by the "aregImpute" algorithm is as follow:

Consider *p* variables containing *m* missing observations (NAs)

1 Initial values are assigned to the NAs by drawing a random sample of size *m* from observed values. Random samples are drawn with replacement if there exist a sufficient number of NAs.

2 The observations from the variable already imputed, i.e. having no missings, are used to draw a sample with replacement for a variable containing any missing value.

3 After transforming the variable, a flexible additive model is fitted to predict this target variable.

4 This semi-parametric fitted model is used to predict the target variable in all of the original observations.

5 The target variable can be imputed either by using the observed value whose predicted transformed value is closest to the predicted transformed value of the missing value or a drawn from a multinomial distribution with probabilities derived from distance weights.

6 Repeat this process whenever predicting other missing variables with current target variable by using random draws from imputations obtained.

This approach has few downsides. Many of the multiple imputations for an observation will be identical when the predicted transformed value is closest to the predicted transformed value of the missing value. This happens when less than three variables are used to predict the target variable and implementation of PMM fails. Moreover, PMM and Bayesian predicted values will always match to same donor observation when only monotonic transformations of left and right-side variables are allowed e.g., every bootstrap resample will give predicted values of the target variable that are monotonically related to predicted values from every other bootstrap resample.

5 MI METHOD FOR COMBINING ESTIMATES

For m = 1,…, *M,* assume *q* and *u* are complete-data estimates θ and its covariance matrix Σ. Let $q^{(m)}$ and $u^{(m)}$ be respectively the point estimates of quantity of interest, Q and variance estimates of $q^{(m)}$. Valid inferences for scalar Q by combining the $q^{(m)}$ and $u^{(m)}$, by Rubin (1987) are as follow.

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

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

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

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

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

with degrees of freedom $v_M = (M-1)(r^{-1})$, where $r = \frac{(1+M^{-1})b_M}{\overline{u}_M}$ represents the relative increase in the conditional variance due to the missing data (see Rubin, 1987). Confidence intervals can be constructed using standard multiple imputation confidence interval construction rules, possibly based on a t-distribution. For more details see Rubin (1996), Barnard and Meng (1999).

6 HYBRID MI (HMI) APPROACH

Implementations of fully conditional MI methods to tackle missing data can become problematic for high missing rates or when there exist complex dependencies structures among variables. For

example, implementation of MICE MI become challenging when incompatibility issue arises due to high dimensions in large scale complex data (White et al., 2011; Razzak and Heumann, 2019). Such complex structures are common in high dimension household surveys where categorical variables have lots of categories i.e. District, Country etc. Moreover these methods are computationally expensive and, in some cases, less accurate as compared to full Bayesian joint models for MI (Si and Reiter, 2013). Many MI algorithms are specific for categorical variables, only, and cannot be implemented with continuous variables or require transformations (other tricks) for continuous variables (Si and Reiter, 2013). Murray and Reiter (2016) implement Bayesian mixture models with local dependence to impute both categorical and continuous values. However, combining the Dirichlet process for multinomial (discrete) mixes with the ones for multivariate (continuous) normal mixes involves knowledge of complicated models to create the dependence structure between the continuous and the categorical variables. These limitations create serious problems for researchers to obtain complete datasets with mixed type variables. We propose an easy to implement hybrid MI (HMI) approach to handle incomplete complex datasets with mixed type variables. HMI combines full Bayesian joint models (JM) MI for categorical data with various MI algorithms commonly implemented in the *R* environment.

The proposed method consists of three stages: Firstly, data instances are separated into two different groups i.e. G_{cat} and G_{num}. All categorical variables are assigned to G_{cat} and numeric ones to G_{num}. Both groups may have missing information. We impute G_{cat} using the DPMPM MI method implemented in *R* package, "NPBayesImpute" (Manrique-Vallier et al., 2014) in the second stage. Then, we combine G_{cat} and G_{num} again but this time we have missing information in $G_{num,}$ only. Lastly, we apply different algorithms to impute G_{num} based on values already imputed by DPMPM. We investigate the ability of various approaches to detect complex dependency structures in high dimensions using the HMI approach. Algorithm 1 explains HMI in detail. To assess the efficiency, we applied three well known MI methods (*R*-packages: "mice", "Amelia" and "Hmisc") to both groups and contrast the results with our HMI methods ("H.Amelia", "H.MICE", "H.Hmics"). Details of all methods are already provided in section 4 of this article. However, short descriptions of existing and hybrid methods can be seen in Table 1 and Table 2 respectively.

Table 1. Basic information: Multiple Imputation in *R*

#Method	Acronym	Description
1	Amelia II	Uses a bootstrap + EM algorithm
2	Hmisc	Uses Additive Regression, Bootstrapping and PMM algorithms
3	NPBayesImpute	Uses a fully Bayesian, joint modeling approach to multiple imputations for categorical data based on latent class models with structural zeros.
4	mice	MI using FCP

Source: Based on Manuals available on http://www.r-project.org/

Table 2 .Basic information: Hybrid Multiple Imputation (HMI) in *R*

#Method	Acronym	Description
1	H.Amelia	Amelia+NPBayesImpute
2	H.Hmisc	Hmisc+NPBayesImpute
3	H.MICE	Mice+NPBayesImpute

Source: Self-prepared.

Algorithm 1: Hybrid MI

Require: *n x p* matrix with incomplete data.

1. G_{cat} ,G_{num} ← Initial division of p variables into two factor and numeric groups
2. **for z= 1, … ,Z do**
3. **for *m= 1, … ,M* do**
4. $G^z_{cat_m}$ ← Imputation using NPBayesImpute.
5. $G^z_{cat_m}$ $G^z_{num_m}$ ← Combining $G^z_{cat_m}$ imputed and $G^z_{num_m}$ missing to generate partially imputed dataset.
6. G^z_m ← Imputing $G^z_{num_m}$ missing using mice |Amelia |Hmisc |i.e. $f(\ G^z_{num_m}\text{missing}\ |\ G^z_{cat_m}\text{imputed})$.
7. G^z_m ← Final imputed data set.
8. $\bar{q}^z_M \leftarrow \sum_{m=1}^{M} \frac{q^{(m)}}{M}$ Pooled point estimates[1].
9. $b^z_M \leftarrow \sum_{m=1}^{M} \frac{(q^{(m)}-\bar{q}_M)^2}{M-1}$
10. $\bar{u}^z_M \leftarrow \sum_{m=1}^{M} \frac{u^{(m)}}{M}$
11. $T^z_M \leftarrow \left(1+\frac{1}{M}\right) b^z_M + \bar{u}^z_M$ Pooled variances[2].
12. **end for**
13. $\bar{q} \leftarrow \sum_{z=1}^{Z} \frac{\bar{q}^z_M}{Z}$ Average of pooled point estimate[3].
14. $\bar{T} \leftarrow \sum_{z=1}^{Z} \frac{T^z_M}{Z}$ Average of pooled variance[4].

end for

[1] $\bar{q}^z_M$ are pooled point estimates over *M* imputed datasets across *z* simulations.
[2] T^z_M are pooled variances over *M* imputed datasets across *z* simulations.
[3] $\bar{q}$ is an average of pooled point estimates ($\bar{q}^z_M$) across *z* simulations.
[4] $\bar{T}$ is an average of pooled variances (T^z_M) across *z* simulations.

7 SIMULATION STUDIES

The simulation studies are inspired by Si and Reiter (2013). The data consists of $N = 1000$ observations. First, five binary variables (X_1, X_2, X_3, X_4, and X_5) are generated from a multivariate normal (MVN) distribution, followed by a categorization. The marginal distributions of X_1, X_2, X_3, X_4, X_5 are normal and we set the mean of each variable at 0 and the variance of each variable at 0.5. The correlation structure is given as:

$$\mathrm{H} = \begin{pmatrix} 1 & \cdots & \rho \\ \vdots & \ddots & \vdots \\ \rho & \cdots & 1 \end{pmatrix}$$

Where $\rho = 0.5$. Random variates are transformed into binary values using the following threshold:

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

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

We than define $\mu_6 = 5X_1 - 3X_2 + 5X_3 - 4X_4 + X_5$ and $\mu_7 = -2 + \mu_6$. Outcomes for two continuous covariates are generated from a normal distribution (ND) as described 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

$$logit\ Pr(X_8) = -1 - 1.5X_1 - 1.15X_2 + 1.25X_3 + 1.6X_4 + 2.88X_5 + 1.11X_6 - 1.5X7 - 1.9X_2X_3 + 2.3X_1X_3 - 1.5X_2X_6 - 2X_5X_6X7 + 1.21X_1X_5 - 2.7X_1X_2 + 1.2X_1X_2X_3 + 3X_6X_7.$$

We then define a co-variate dependent binary response generated from Bernoulli distributions with probabilities governed by the logistic regression as follow:

$logit\ Pr(y) = 0.5 - 0.1X_1 - 0.1X_2 - 0.1X_3 + 0.9X_4 - 0.5X_5 + 0.2X_6 - 0.1X7 - 0.5X_8$ and $\phi = \beta true = (0.5;-0.1;-0.1;-0.1;0.9;-0.5;0.2;-0.1;-0.5)$. We suppose that values in all covariates are MAR with the following probability

$$p = 1 - \frac{e^{(-0.001 - X7)}}{(1 + e^{(-0.001 - X7)})}.$$

This provides around 10% of the observations in Xi to be missing (at random). Since Si and Reiter (2013) did not observe noticeable differences in the posterior distributions of θ for higher values of prior specifications, we set relatively small prior specification values i.e. (a_α= 0.05, b_α= 0.01) in *R* package "NPBayesImpute" version 0.6 (Manrique-Vallier et al., 2014). Akande et al. (2017) suggest that the latent classes (k) less than 40 can appear sufficiently large based on tuning with initial runs. However, we follow Dunson and Xing (2009) who suggest that large enough k can make the latent class model consistent for any joint probability distribution in case of dependencies among variable. Therefore, we set the sufficiently large number of latent classes (k) 80 and run each MCMC chain for 1000 iterations using the first 200 as burn-in. We

implement a default version of chained equations using the "mice" software package in *R* version 2.12 (van Buuren and Oudshoorn, 1999). We implement bootstrap and PMM MI methods using 13 iterations (for convenience) with the "aregImpute" function in the "Hmisc" software package in *R* version 4.1 (Harrell, 2010). We also use the *R* package "Amelia II" version 1.6.1 (Honaker et al., 2011) with defaults as basic command. Various imputations are generated for each MI method. Five thousand sampling simulations are run.

Pooled point estimates and standard errors for the fitted GLM's with binary response are presented in figures 1, 2 ,3 and 4 for 10 and 20 imputed data sets, respectively. In order to get insight into the performance of the imputation algorithms, we make comparisons of different imputation methods using the root mean square error (RMSE) and empirical standard errors (ESE) indices, which are calculated using the following formulas:

$$\text{RMSE}\ \overline{q}_m = \sqrt{\frac{\sum_{z=1}^{Z}\left(\bar{q}_M^z - \Theta\right)^2}{Z}},$$

$$\text{ESE}_{\overline{q}_m} = \sqrt{\frac{\sum_{z=1}^{Z}\left(\bar{q}_M^z - \bar{q}\right)^2}{Z}}$$

where $\overline{q}_m$ and Θ denote the estimated parameter pooled over *M* imputed data sets and original parameters, respectively. The average values of the pooled estimated parameters over the 5000 simulations are presented by $\bar{q}$. The coverage rates of at least 95% are calculated as:

$$\text{Coverage rate}\ _{\overline{q}_m} = \frac{\sum_{z=1}^{Z} 1\ [\Theta \in \text{CI}\ (\bar{q}_M^z, T_M^z)]}{Z},$$

where $1\ [\Theta \in \text{CI}\ (\bar{q}_M^z,\ T_M^z)]$ is an indicator function whose value is equal to one when the confidence interval based on $\bar{q}_M^z$ and T_M^z contains Θ and equal to zero otherwise.

8 SIMULATION RESULTS

As discussed, we used three software package in *R* i.e. ("Amelia","MICE" and "Hmisc") for comparison with our proposed HMI methods, i.e. ("H.Amelia","H.MICE" and "H.Hmisc"). We limited the simulation study to low missingness rates and consider 10% of values MAR, only. We also increased the number of imputations from *M=10* to *M=20* for eventually better estimates. Table 3 shows the performance of various MI methods based on estimated means RMSEs, ESEs (top) and coverage rates of 95% confidence intervals (bottom) over 5000 simulation runs. The estimated amount of bias and between imputations variation can be assesed by RMSEs and ESEs respectively. Overall, "MICE" tends to result in the most mean coverage rates concentrated around 95% and fewest high rates. The mean coverage rates for "H.MICE" tend to be larger than the mean coverage rates for "MICE", although both tend to be close to 95%. Standard "Amelia" results in coverage rates above 95% for most of the covariates. Sometimes it reaches very high rates for categorical covariates (i.e. *M* = 10: ß$_2$ and ß$_3$= 98) except one binary covariate where it reaches very low rates (i.e. *M* = 10, 20: ß$_4$ = 92). "H. Amelia" results in mean coverage rates for all covariates that are concentrated slightly above

95%, but its lower and upper tails are comparable to that of “Amelia”. “Hmisc” results in the mean coverage rates for most of the covariates that are concentrated very above 95%, it has the longest upper tail, sometimes reaching very high rates (i.e. M = 20: $ß_2$ = 98). Across the simulations, the mean coverage rates for “H.Hmisc” tend to be similar to the mean coverage rates for “Hmisc” but its upper tail is comparable to that of “Hmisc” (i.e. M = 20: $ß_2$ = 97). We observe that the estimated mean ESEs for “H.MICE” MI method are smaller for all types of covariates as compared to “MICE”, whereas “H.Hmisc” shows similar or smaller mean ESEs as compared to “Hmisc” and “H. Amelia” shows similar or slightly higher mean ESEs as compared to “Amelia” for most of the covariates. The estimated mean RMSEs for “H.MICE” MI method are smaller for most of the covariates as compared to “MICE”, whereas “H.Hmisc” have similar or slightly higher mean RMSEs as compared to “Hmisc” and Amelia” have the similar or smaller mean RMSEs as compared to Amelia” for most of the covariates. There seem to be similarities in structure among all MI methods i.e. all methods are slightly upward biased for most of the binary covariates e.g. $ß_1$, $ß_2$, $ß_3$, $ß_5$, $ß_8$ and downward biased for continues covariates and one binary covariates e.g. $ß_4$. The point estimates based on “MICE” and “H.MICE” methods are closer to the corresponding true values as compared to other methods (see Figures 1-2). Hybrid MI methods (i.e. “H.MICE”, “H.Hmisc”, “H. Amelia”) tend to have smaller standard errors as compared to their counterparts (i.e. “MICE”, “Hmisc”, “Amelia”) for most of the covariates except three binary covariates i.e. $ß_2$, $ß_5$, $ß_8$ where “H.Amelia” shows similar or slightly higher standard errors as compared to “Amelia” (see Figures 3-4).

Table 3**.** The performance of methods for MI

				M =10			
	Coef.	**H. Hmics**	**Hmics**	**H.Amelia**	**Amelia**	**H.MICE**	**MICE**
RMSEs(ESEs)	β_1	0.19(0.17)	0.18(0.18)	0.19(0.17)	0.18(0.16)	0.19(0.18)	0.20(0.20)
	β_2	0.17(0.17)	0.17(0.16)	0.17(0.17)	0.16(0.16)	0.18(0.17)	0.18(0.18)
	β_3	0.19(0.18)	0.19(0.18)	0.19(0.18)	0.18(0.17)	0.19(0.18)	0.20(0.20)
	β_4	0.19(0.18)	0.19(0.17)	0.19(0.17)	0.21(0.16)	0.19(0.18)	0.19(0.19)
	β_5	0.17(0.16)	0.16(0.16)	0.17(0.16)	0.16(0.15)	0.17(0.16)	0.17(0.17)
	β_6	0.27(0.23)	0.27(0.26)	0.27(0.23)	0.28(0.24)	0.28(0.24)	0.30(0.30)
	β_7	0.47(0.46)	0.47(0.47)	0.47(0.46)	0.46(0.46)	0.50(0.49)	0.51(0.51)
	β_8	0.17(0.17)	0.16(0.15)	0.17(0.17)	0.16(0.15)	0.17(0.17)	0.18(0.18)
Coverage %	β_1	96	97	96	97	96	96
	β_2	97	97	97	98	97	95
	β_3	96	96	96	98	96	95
	β_4	94	94	94	92	94	95
	β_5	96	96	96	96	96	96
	β_6	95	97	95	96	95	96
	β_7	97	97	97	97	96	95
	β_8	96	97	96	96	96	95

				M = 20			
	Coef.	**H. Hmics**	**Hmics**	**H.Amelia**	**Amelia**	**H.MICE**	**MICE**
RMSEs(ESEs)	β_1	0.18(0.17)	0.18(0.18)	0.18(0.17)	0.18(0.16)	0.19(0.18)	0.20(0.20)
	β_2	0.17(0.17)	0.16(0.16)	0.17(0.17)	0.16(0.16)	0.17(0.17)	0.18(0.18)
	β_3	0.18(0.18)	0.18(0.18)	0.18(0.18)	0.18(0.17)	0.19(0.18)	0.20(0.20)
	β_4	0.19(0.18)	0.19(0.17)	0.19(0.17)	0.20(0.16)	0.19(0.18)	0.19(0.19)
	β_5	0.16(0.16)	0.16(0.16)	0.17(0.16)	0.16(0.15)	0.16(0.16)	0.17(0.17)
	β_6	0.28(0.23)	0.27(0.26)	0.27(0.23)	0.28(0.24)	0.28(0.25)	0.30(0.30)
	β_7	0.46(0.46)	0.47(0.47)	0.46(0.46)	0.46(0.46)	0.49(0.49)	0.51(0.51)
	β_8	0.17(0.17)	0.16(0.15)	0.17(0.17)	0.17(0.15)	0.17(0.17)	0.18(0.18)
coverage %	β_1	96	97	96	97	96	95
	β_2	97	98	96	98	97	96
	β_3	97	97	97	97	96	95
	β_4	94	95	94	92	94	96
	β_5	96	96	96	96	96	96
	β_6	95	97	96	96	95	95
	β_7	97	97	97	97	96	95
	β_8	96	96	96	96	96	96

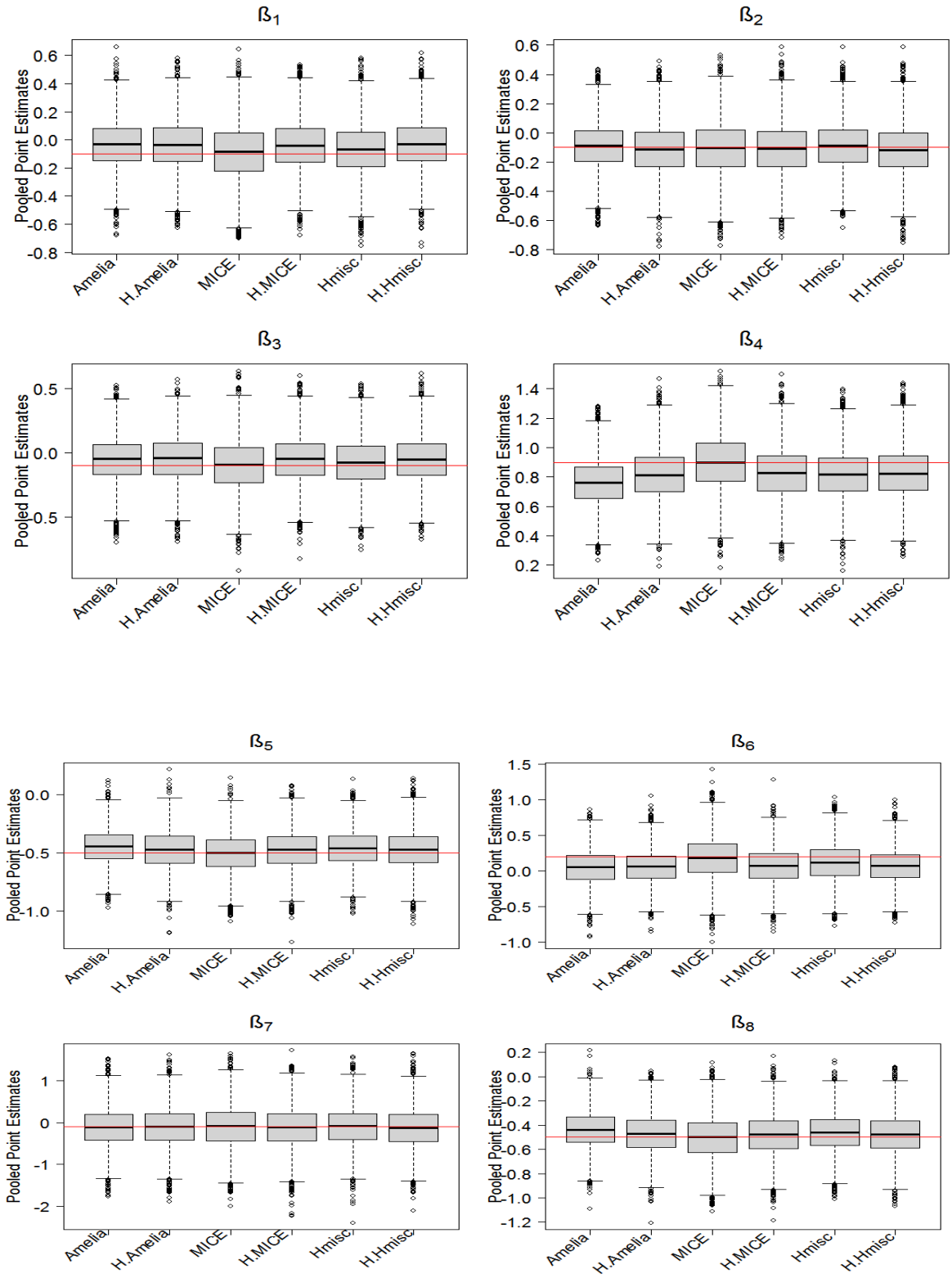

Figure 1. Boxplots for the point estimates across 5000 simulations and 10 imputations by various imputation methods.

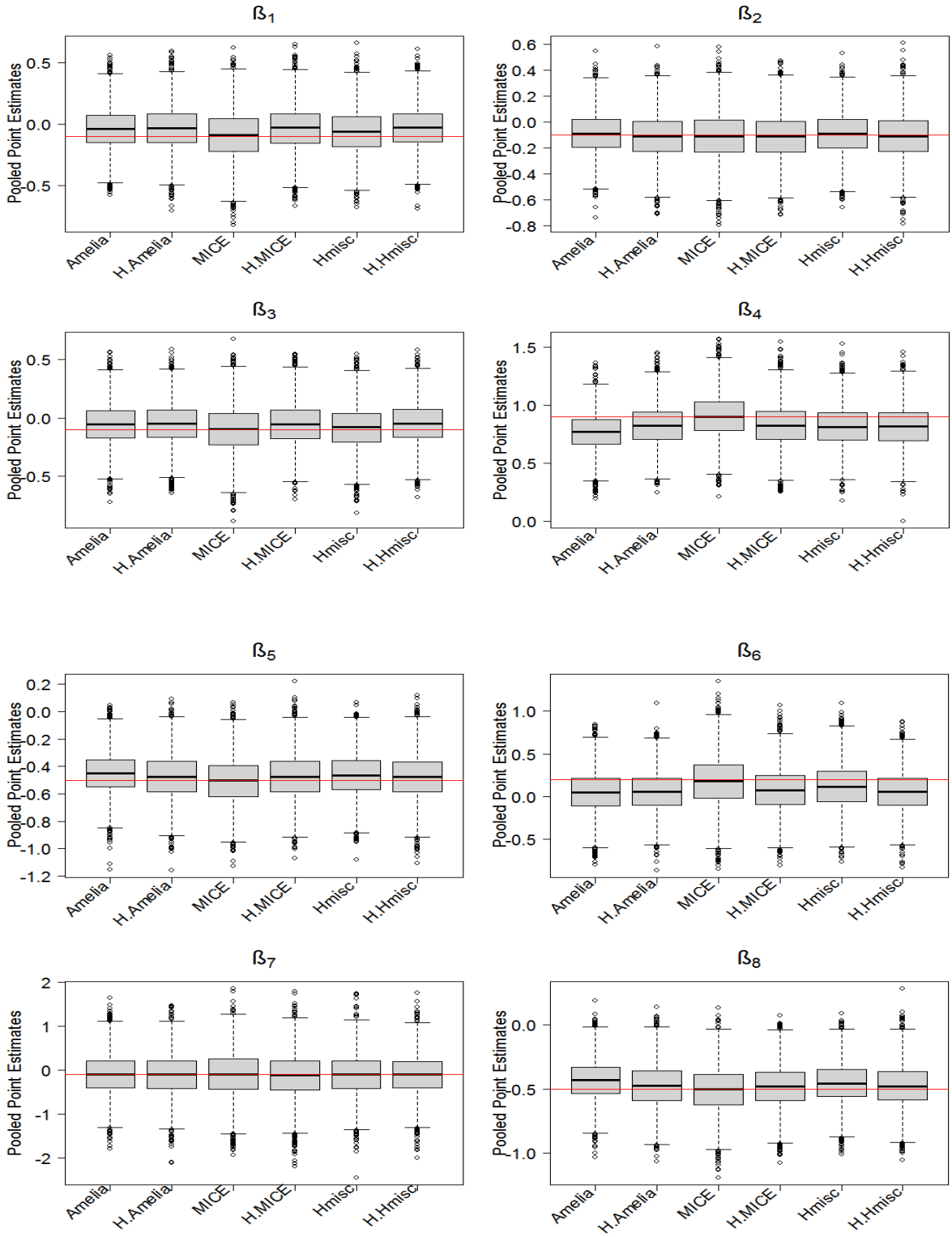

Figure 2. Boxplots for the point estimates across 5000 simulations and 20 imputations by various imputation methods.

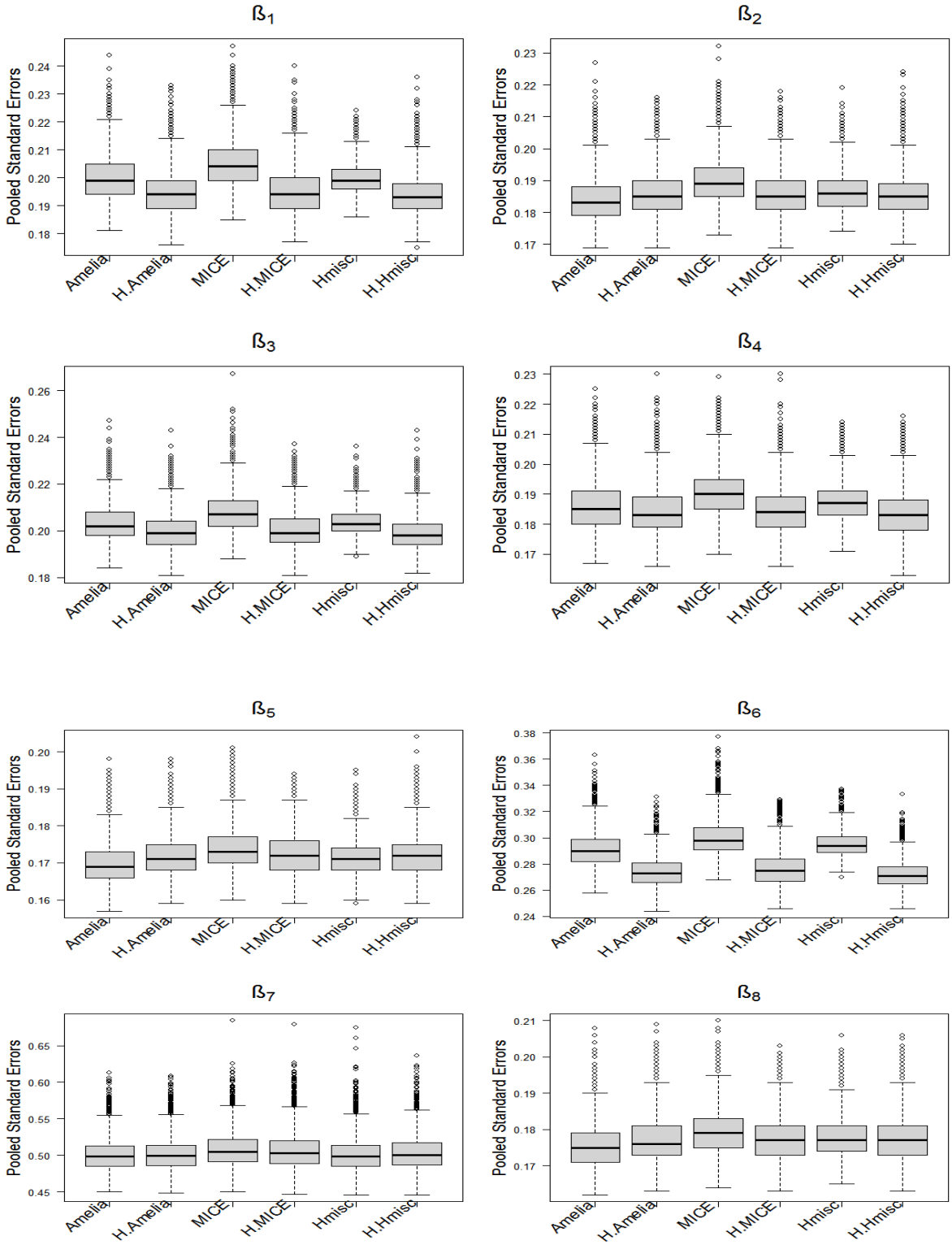

Figure 3. Boxplots for the standard errors across 5000 simulations and 10 imputations by various imputation methods.

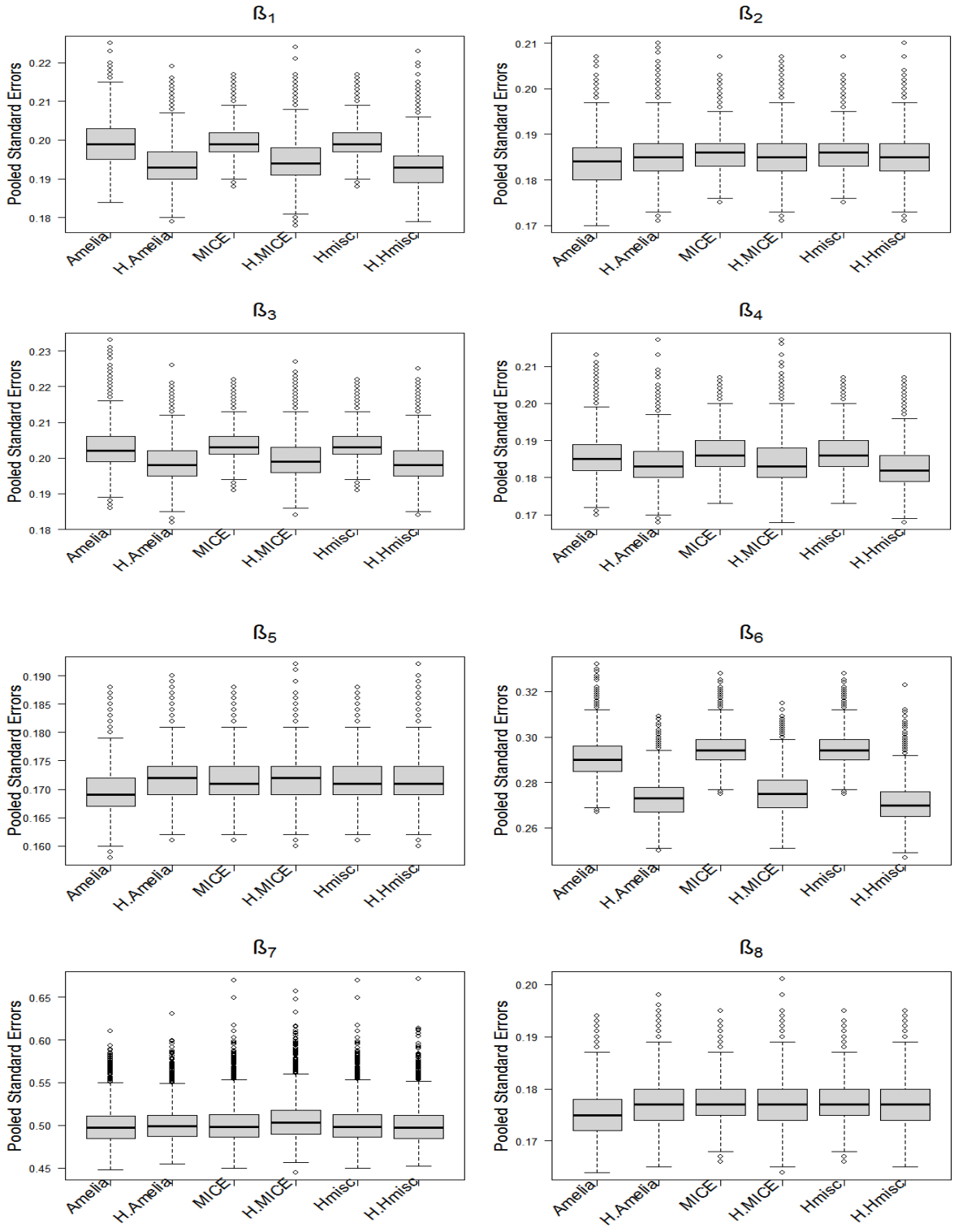

Figure 4. Boxplots for the standard errors across 5000 simulations and 20 imputations by various imputation methods.

9 CONCLUDING REMARKS

Based on results obtained by simulations, we can make several general conclusions about various MI procedures. First, the default application of "MICE", appears to be inferior to "H.MICE", overall. "H.MICE" utilizes the JM approach to identify complex dependency structures among categorical variables where missing continuous variables are imputed using the PMM technique. Of course, one could use various applications offered by MICE, (e.g. CART). Second, analysts may prefer "H.Amelia" for high coverage rates for most estimands with slight bias and due to its fastness[5]. Third, identification of a clear winner between "Hmisc" and "H.Hmisc" is little difficult. "H.Hmisc" tends to result in slightly higher mean RMSEs than "Hmisc" does, but its coverage rates are comparable that of "Hmisc". Based on results obtained by simulations, we can also make some general conclusions about three HMI procedures. Analysts concerned with getting at least nominal coverage rates for most estimands at the expense of some high mean RMSEs and ESEs, may prefer "H.MICE" over "H.Hmisc" and "H.Amelia". Simulation studies indicate that "H.Hmisc" and "H.Amelia" tend to perform in most cases. Further evaluations with diversity of experimental settings will undoubtedly be needed to account for this behavior. Increasing the number of imputed data sets improves results by reducing RMSEs. Since now, we have considered small numbers of prior specifications (a_α , b_β) and mixture components (*k)* in simulations, extensive comparisons are required for increased levels of a_α , b_β and *k.* We considered only binary response with binary and continuous covariables. Of course, statistical properties of the HMI approach can be studied for continuous response with mixed type covariates, also. Additionally, data with ordinal nature and more categories can be included for further comparisons. Real data applications can prove to be useful to see potential of proposed methods.

References

Allison, P.D. 2000. *Multiple imputation for missing data: A cautionary tale,* Sociological Methods and Research, 28, 301-309.

Ake, C.F. 2005. Rounding after multiple imputation with non-binary categorical covariates. In *Proceedings of the 13th Annual SAS Users Group International Conference.* SAS Institute Inc.

Akande, O., Li, F. and Reiter, J.P. 2017. *An Empirical Comparison of Multiple Imputation Methods for Categorical Data,* The American Statistician, 71, 162-170.

Bishop, Y., Feinberg, S. and Holland, P. 1975. *Discrete multivariate analysis: Theory and practice.* Cambridge, MA: MIT Press.

Barnard, J. and X, Meng. 1999. *Applications of multiple imputation in medical studies: From aids to nhanes.* Statistical Methods in Medical Research, 8(1), 17–36.

5 The time taken by hybrid methods may vary depending on number of iterations and mixture components assigned i.e. it takes more time for large values of *k* and iterations. Therefore, "H.Amelia" is slower than "Amelia" but fastest then all the remaining MI methods used in analysis.

Bernaards, C.A., Belin, T.R. and Schafer, J.L. 2007. *Robustness of a multivariate normal approximation for imputation of binary incomplete data,* Statistics in Medicine, 26, 1368-1382.

Dempster, A.P., Laird, N.M. and Rubin D.B. 1977. *Maximum likelihood from incomplete data via the EM algorithm.* Journal of the Royal Statistical Society, series B, 39, 1–38.

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

Efron, B. 1979. *Bootstrap Methods: Another Look at the Jackknife*, The Annals of Statistics, 7, 1–26.

Erosheva, E. A. Fienberg, S. E. and Junker, B. W. 2002. *Alternative statistical models and representations for large sparse multi-dimensional contingency tables*, Annales de la Faculte´ des Sciences de Toulouse, 11, 485-505.

Finch, W.H. 2010. *Imputation methods for missing categorical questionnaire data: A comparison of approaches*, Journal of Data Science, 8, 361-378.

Gilks, W., Richardson, S., Spiegelhalter, D. 1996. *Markov Chain Monte Carlo in Practice.* New York: Chapman and Hall/CRC.

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

Horton, N.J. and Lipsitz, S.R. 2001. *Multiple imputation in practice: comparison of software packages for regression models with missing variables*, The American Statistician, 55(3), 244-254.

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

Horton, N.J. and Kleinman, K.P. 2007. *Much Ado About Nothing:A Comparison of Missing Data Methods and Software to Fit Incomplete Data Regression Models*, The American Statistician, 6 (1), 79-90.

Honaker, J. and Gary, K. 2010. *What to do About Missing Values in Time Series Cross-Section Data*, American Journal of Political Science, 54(2), 561-581.

Honaker, J., King, G. and Blackwell, M. 2011. *Amelia II: A Program for Missing Data*, Journal of Statistical Software, 45(7), 1-47.

Little, R.J.A. 1988. *Missing-Data Adjustments in Large Surveys*, Journal of Business and Economic Statistics, 6, 287-296.

Little, R.J.A. and Rubin, D.B. 2002. *Statistical Analysis with Missing Data*. Hoboken, NJ: John Wiley Sons.

Lin, T.H. 2008. *A comparison of multiple imputation with EM algorithm and MCMC method for quality of life missing data*, Quality & Quantity, 44(2), 277–287.

Manrique-Vallier, D., Reiter, J.P., Hu, J. and Quanli, W. 2014. *NPBayesImpute: Non-parametric Bayesian multiple imputation for categorical data*, The Comprehensive *R* Archive Network.

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 American Statistical Association, 111, 1466–1479.

Oba, S., Sato, M.A., Takemasa, I., Monden, M., Matsubara, K.I. and Ishii, S.A. 2003. *A bayesian missing value estimation method for gene expression profile data,* Bioinformatics, 19, 2088-96.

Rosenbaum, P.R. and Rubin, D.B. 1983. *Assessing Sensitivity to an Un-observed Binary Covariate in an Observational Study with Binary Outcome*, Journal of the Royal Statistical Society, Series B, 45, 212-218.

Rubin, D.B. and Schenker, N. 1986. *Multiple Imputation for Interval Estimation from Simple Random Samples with Ignorable Nonresponse*, Journal of the American Statistical Association, 81, 366–374.

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

Rubin, D.B. 1996. *Multiple Imputation After 18+ Years*. Journal of the American Statistical Association, 91(434), 473–89.

Razzak, H. and Heumann, C. 2019. *Hybrid multiple imputation in a large scale complex survey,* Statistics in Transition New Series, 20(4), 33-58.

Schafer, J.L. 1997. *Analysis of Incomplete Multivariate Data*, London: Chapman & Hall/CRC.

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.

van Buuren, S. and Oudshoorn, C. 1999. *Flexible multivariate imputation by MICE (Tech. rep. TNO/VGZ/PG 99.054)*, Leiden: TNO Preventie en Gezondheid.

Vermunt, J.K., Ginkel, J.R.V., der Ark, L.A.V. 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 in R, Journal of Statistical Software, in press.

van Buuren, S. 2012. *Flexible Imputation of Missing Data*, London: Chapman & Hall/CRC.

Watanabe, M. and Kazunori Y. 2000. *EM Algorithm to Fukanzen Data no Shomondai (EM Algorithm and the Problems of Incomplete Data)*, Tokyo: Taga Shuppan.

Wooldridge, J.M. 2002. E*conometric Analysis of Cross Section and Panel Data*, Cambridge, MA: MIT Press.

White, I.R., Royston, P. and Wood, A.M. 2011. Multiple imputation using chained equations: issues and guidance for practice, *Statistics in Medicine*, 30(4), 377–399.

Yu, L.-M., Burton, A. and Rivero-Arias, O. 2007. *Evaluation of software for multiple imputation of semi-continuous data*, Statistical Methods in Medical Research, 16, 243-258.

Yucel, R.M., He, Y. and Zaslavsky, A.M. 2011. *Gaussian-based routines to impute categorical variables in health surveys*, Statistics in Medicine, 30, 3447-3460.