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, 2020
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 it can become difficult and contentious, especially when there are categorical variables with complex dependencies and high dimensions to impute. 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 $i=1,\dots,n$ and variables $j=1,\dots,p$, 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 $Pr\left(\mathrm{R}|D^{miss},D^{obs}\right)= Pr(\mathrm{R})$. The second possible assumption is Missing At Random (MAR) where $Pr\left(\mathrm{R}|D^{miss},D^{obs}\right)= Pr\left(\mathrm{R}|D^{obs}\right)$. Missing Not At Random (MNAR) is another possible assumption where $Pr\left(\mathrm{R}|D^{miss},D^{obs}\right) \neq Pr\left(\mathrm{R}|D^{obs}\right)$ and depends on D^{miss}. The third assumption is also called non-ignorable (NI) and not further used in the paper. (Little and Rubin, 2002).

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*. "Amelia II", "Hmisc", "mi", "mice", "yaImpute", "mix", "cat", "NPBayesImpute", "norm", "pan", "monoman", "mvnml" etc., all use the *R* environment to impute missing data. "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. Short descriptions of some of these packages are presented in Table 1.

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. A sequence of random variables where the distribution of each element depends on the values of the previous ones is called a markov chain. 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 "mice" 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 is 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, er 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

The combination of the expectation-maximization algorithm and non-parametric bootstrapping is called the Expectation-Maximization with Bootstrapping (EMB) algorithm. First, starting values for the mean and variance and a multivariate distribution is assumed, the distribution is updated by calculating an expected value of the model likelihood followed by parameter estimation. Expectation and maximization steps are iterated until estimates converge.

However, below is a brief summary of the expectation-maximization (Schafer, 1997; Watanabe and Yamaguchi, 2000; Little and Rubin, 2002).

As far as the non-parametric bootstrapping method is concerned, it works by utilizing the observed sample as the pseudo-population and randomly drawing a subsample of size *n* with replacement from this observed sample. This process is repeated *M* times (Wooldridge, 2002). Assuming a data set with *q* observed and *n - q* missing values, the EMB algorithms consists of the following steps: First, bootstrap samples of size *n* are drawn from incomplete data *M* times by applying non-parametric bootstrapping method. Second, *M* point estimates of μ and Σ are calculated by applying the EM algorithm to each of these *M* bootstrap samples. Finally, *M* multiply-imputed data sets are constructed (Honaker and King, 2010). *R* package "Amelia II" (Honaker et al., 2011) version 1.6.1 implements this algorithm. EMB is computationally more efficient as compared to MCMC methods but is only an approximate Bayesian procedure.

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 Bayesian joint modeling. 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. $z_i \epsilon \ \{1, \ldots, K\}$ with probability $\pi_k = Pr\left(z_i = k\right)$. 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 $c_j \ \epsilon \ \left\{1, \ldots, d_j\right\}$, let $\yen_{klj}^{(j)} = Pr\left(C_{ij} = c_j | z_i = k\right)$. We can express the finite mixture model mathematically as $C_{ij} | z_i, \yen \ \widetilde{ind} \ Multinomial\left(\yen_{z_i 1}^{(j)}, \ldots, \yen_{z_i d_j}^{(j)}\right)$ for all *i* and *j* and $z_i | \pi \sim Multinomial\,(\pi_1, \ldots, \pi_k)$ for all *i*. For prior distributions on $\yen$ and π, we have $\pi_k = V_k\,(\prod_{l<k} 1 - V_l)$ 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 $\left(\yen_{k1}^{(j)}, \ldots, \yen_{kd_j}^{(j)}\right) \sim Dirichlet\ (a_{j1}, \ldots, a_{jd_j})$. 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, can include any possible distribution, is computationally efficient and easy to code. The *R* package, "NPBayesImpute" by Manrique-Vallier et al. (2014) implements this approach. Short coming of this package is that it only takes categorical variables into account.

4.3 FULLY CONDITIONAL SPECIFICATION (FCS): CHAINED EQUATIONS

The FCS approach is an alternative to MCMC (see Yu et al., 2007). 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 $t=1,\dots,T$ and $j=1,\dots,p$ 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.

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

#Method	Acronym	Description
1	Amelia II	Uses a bootstrap +EM algorithm
2	mi	Uses multiple iterative regression algorithm
3	Hmisc	Uses Additive Regression, Bootstrapping and PMM algorithms
4	NPBayesImpute	Uses a fully Bayesian, joint modeling approach to multiple imputations for categorical data based on latent class models with structural zeros.
5	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.

5 MI METHOD FOR COMBINING ESTIMATES

For m = 1,…, M, assume q and u are complete-data estimators. 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.

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

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

$$\overline{u}_M = \sum_{m=1}^{M} \frac{u^{(m)}}{M}, \tag{3}$$

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

6 HYBRID MI (HMI) APPROACH

We propose an easy to implement hybrid MI (HMI) approach. HMI combines full Bayesian joint models (JM) MI with various MI algorithms commonly implemented in the *R* environment. Although a great number of algorithms have been designed to tackle missing data problems but many of them can only work with a specific amount of missing rate, variables and data. Many MI algorithms are specific for categorical variables, only, and cannot be implemented on continuous variables where as other accept both types of variables but fail to perform in presence of high dimensions and complex dependencies. These methods are computationally expensive and, in some cases, less accurate. Such complex structures are common in high dimension household surveys where categorical variables have lots of categories i.e. District, Country etc. These limitations motivate the use of DPMPM for MI of categorical data because this method has better capacity to work in high dimension and missing at random mechanism. The superiority of DPMPM MI is well established over default MICE, see Si and Reiter (2013). 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. To access 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 these methods are already provided in section 4 of this article. Algorithm 1 below explains HMI in detail.

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. $f(G^{z}_{num_m}\text{missing} \mid G^{z}_{cat_m}\text{imputed})$.
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:

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

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

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

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:
$logit\ Pr\ (y) = 0.5 - 0.1X_1 - 0.1X_2 - 0.1X_3 + 0.9X_4 - 0.5X_5 + 0.2X_6 - 0.1X7 - 0.5X_8 + \varepsilon$ 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 X_i to be missing (at random). For applying the DPMPM MI method, we set the prior specifications for a_α and b_α to $(a_\alpha = 0.05, b_\alpha = 0.01)$ with 80 as the maximum number of mixture components (*k*). We implement the DPMPM MI method using the *R* package "NPBayesImpute" version 0.6 (Manrique-Vallier et al., 2014). 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 (for convenience) iterations 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 (*M*) 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:

$$\text{RMSE} = \sqrt{E_\beta\left(\widehat{\beta}_j - \beta_j\right)^2},$$

$$ESE = \sqrt{E_{\beta}\,(\widehat{\beta_j} - \overline{\overline{\widehat{\beta_j}}}\,)^2},$$

where $\widehat{\beta_j}$ and β_j 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 $\overline{\beta_j}$. The smaller values for RMSEs and ESEs indicate better performance (Oba et al., 2003). Results for ESEs and RMSEs for the described standard and HMI techniques are presented in Tables 3, 4, 6 and 7 respectively. Tables 5 and 8 present the coverage rate of 95% confidence intervals for 10 and 20 imputed data sets.

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. Between imputations variation can be assessed by ESEs presented in Tables 3 and 6. We observe for the most part, that the ESEs are lower or equal for the "H.Hmisc" approach except for $ß_8$ which is slightly high as compared to the standard "Hmisc" approach. The ESEs for fitted generalized linear models under the "H.MICE" approach are smaller or equal for all cases as compared to its counterpart. The "H.Amelia" MI approach tends to have slightly higher ESEs, overall. The amount of bias can be assesed by RMSEs presented in Tables 4 and 7. All six methods are slightly downward biased for most of the cases. Estimates obtained from "H.MICE" tend to have lower bias for most of the cases as compared to the "MICE" MI approach. The coverage rates, for most estimands, are higher or at least similar between all three HMI methods. See tables 5 and 8. Pooled standard errors are often lower for the three HMI methods, see figures 2 and 4. The results suggest that the MI approach which combines the JM approach for categorical variables and various algorithms based on linear models for continuous variables is a useful alternative to the various algorithms when applied solely.

Table 3. Simulated Data: ESEs for fitted generalized linear models to 10 imputed data sets

Variables	**H. Hmics**	**Hmics**	**H.Amelia**	**Amelia**	**H.MICE**	**MICE**
β_1	0.18	0.18	0.18	0.17	0.18	0.20
β_2	0.17	0.17	0.17	0.16	0.18	0.18
β_3	0.18	0.19	0.18	0.18	0.18	0.20
β_4	0.17	0.17	0.17	0.16	0.18	0.19
β_5	0.17	0.15	0.17	0.15	0.16	0.17
β_6	0.24	0.26	0.23	0.24	0.25	0.30
β_7	0.47	0.47	0.47	0.46	0.50	0.51
β_8	0.17	0.16	0.17	0.15	0.17	0.18

Table 4. Simulated Data: RMSEs for fitted generalized linear models to 10 imputed data sets

Variables	H. Hmics	Hmics	H.Amelia	Amelia	H.MICE	MICE
β_1	0.19	0.18	0.19	0.18	0.19	0.20
β_2	0.17	0.17	0.17	0.16	0.18	0.18
β_3	0.19	0.19	0.19	0.18	0.19	0.20
β_4	0.19	0.19	0.19	0.21	0.19	0.19
β_5	0.17	0.16	0.17	0.16	0.17	0.17
β_6	0.27	0.27	0.27	0.28	0.28	0.30
β_7	0.47	0.47	0.47	0.46	0.50	0.51
β_8	0.17	0.16	0.17	0.16	0.17	0.18

Table 5. Simulated Data: Coverage rate of 95% confidence intervals for 10 imputed data sets

Methods	β_1 β_2 β_3 β_4 β_5 β_6 β_7 β_8
Amelia	97 98 98 92 96 96 97 96
Hmics	97 97 96 94 96 97 97 97
MICE	96 95 95 95 96 96 95 95
H.Amelia	96 97 96 94 96 95 97 96
H.Hmics	96 97 96 94 96 95 97 96
H.MICE	96 97 96 94 96 95 96 96

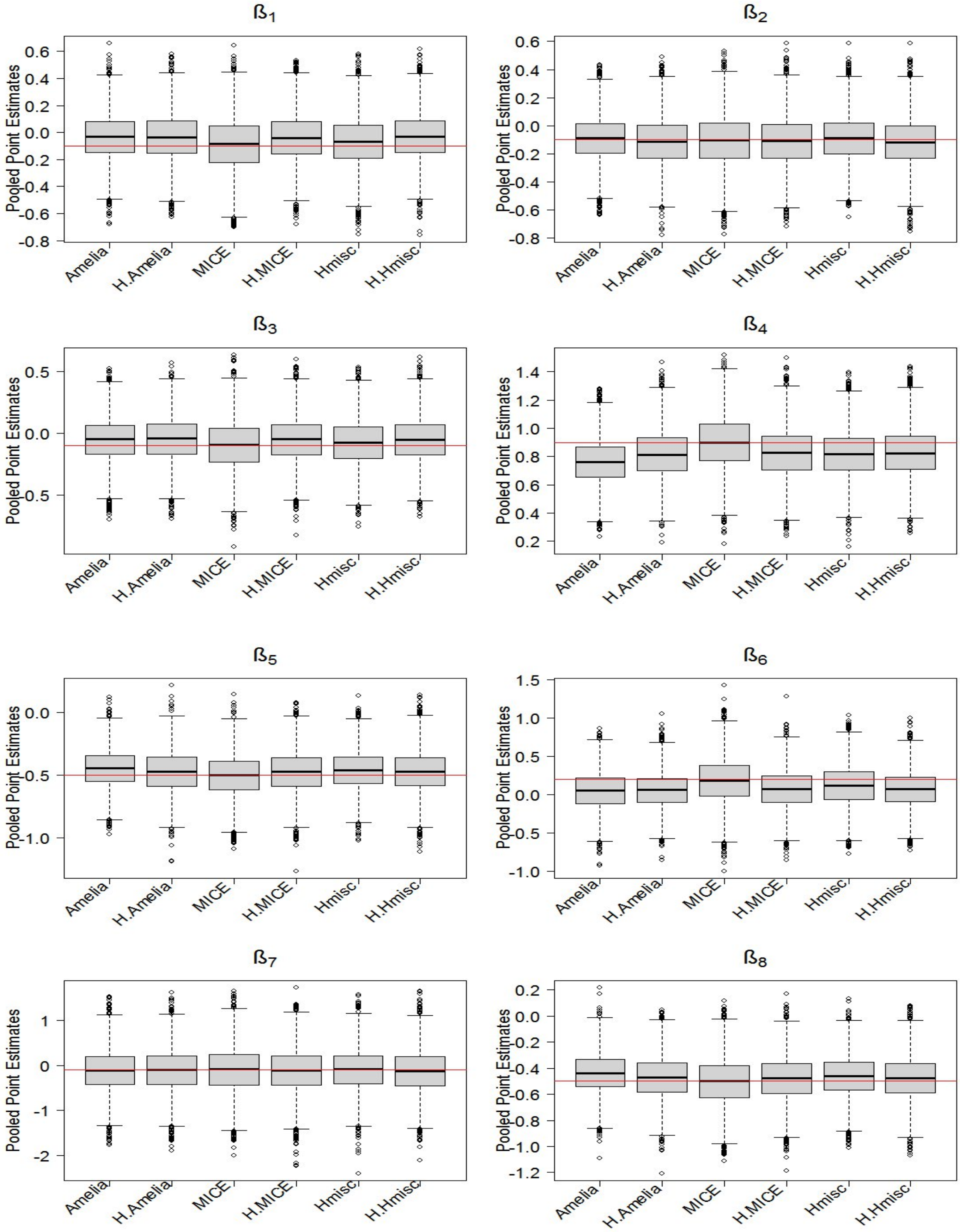

Figure 1. Simulated data: Pooled Point Estimates for 10 imputed data sets

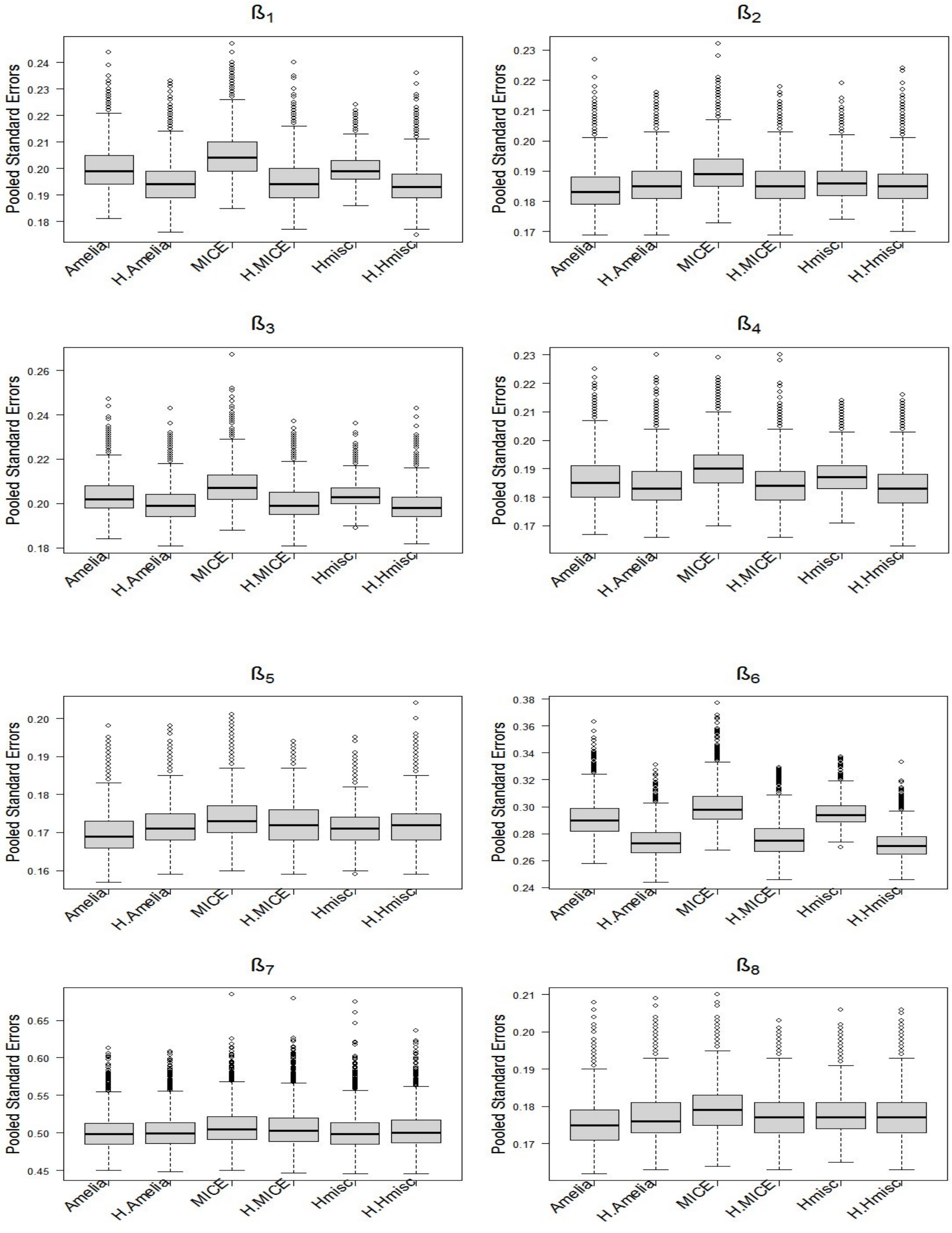

Figure 2. Simulated data: Pooled Standard Errors for 10 imputed data sets

Table 6. Simulated Data: ESEs for fitted generalized linear models to 20 imputed data sets

Variables	**H.Hmics.**	**Hmics.**	**H.Amelia**	**Amelia**	**H.MICE**	**MICE**
β_1	0.17	0.18	0.17	0.16	0.18	0.20
β_2	0.17	0.16	0.17	0.16	0.17	0.18
β_3	0.18	0.18	0.18	0.17	0.18	0.20
β_4	0.18	0.17	0.17	0.16	0.18	0.19
β_5	0.16	0.16	0.16	0.15	0.16	0.17
β_6	0.23	0.26	0.23	0.24	0.25	0.30
β_7	0.46	0.47	0.46	0.46	0.49	0.51
β_8	0.17	0.15	0.17	0.15	0.17	0.18

Table 7. Simulated Data: RMSEs for fitted generalized linear models to 20 imputed data sets

Variables	**H.Hmics.**	**Hmics.**	**H.Amelia**	**Amelia**	**H.MICE**	**MICE**
β_1	0.18	0.18	0.18	0.18	0.19	0.20
β_2	0.17	0.16	0.17	0.16	0.17	0.18
β_3	0.18	0.18	0.18	0.18	0.19	0.20
β_4	0.19	0.19	0.19	0.20	0.19	0.19
β_5	0.16	0.16	0.17	0.16	0.16	0.17
β_6	0.28	0.27	0.27	0.28	0.28	0.30
β_7	0.46	0.47	0.46	0.46	0.49	0.51
β_8	0.17	0.16	0.17	0.17	0.17	0.18

Table 8. Simulated Data: Coverage rate of 95% confidence intervals for 20 imputed data sets

Methods	β_1	β_2	β_3	β_4	β_5	β_6	β_7	β_8
Amelia	97	98	97	92	96	96	97	96
Hmics	97	98	97	95	96	97	97	96
MICE	95	96	95	96	96	95	95	96
H.Amelia	96	96	97	94	96	96	97	96
H.Hmics	96	97	97	94	96	95	97	96
H.MICE	96	97	96	94	96	95	96	96

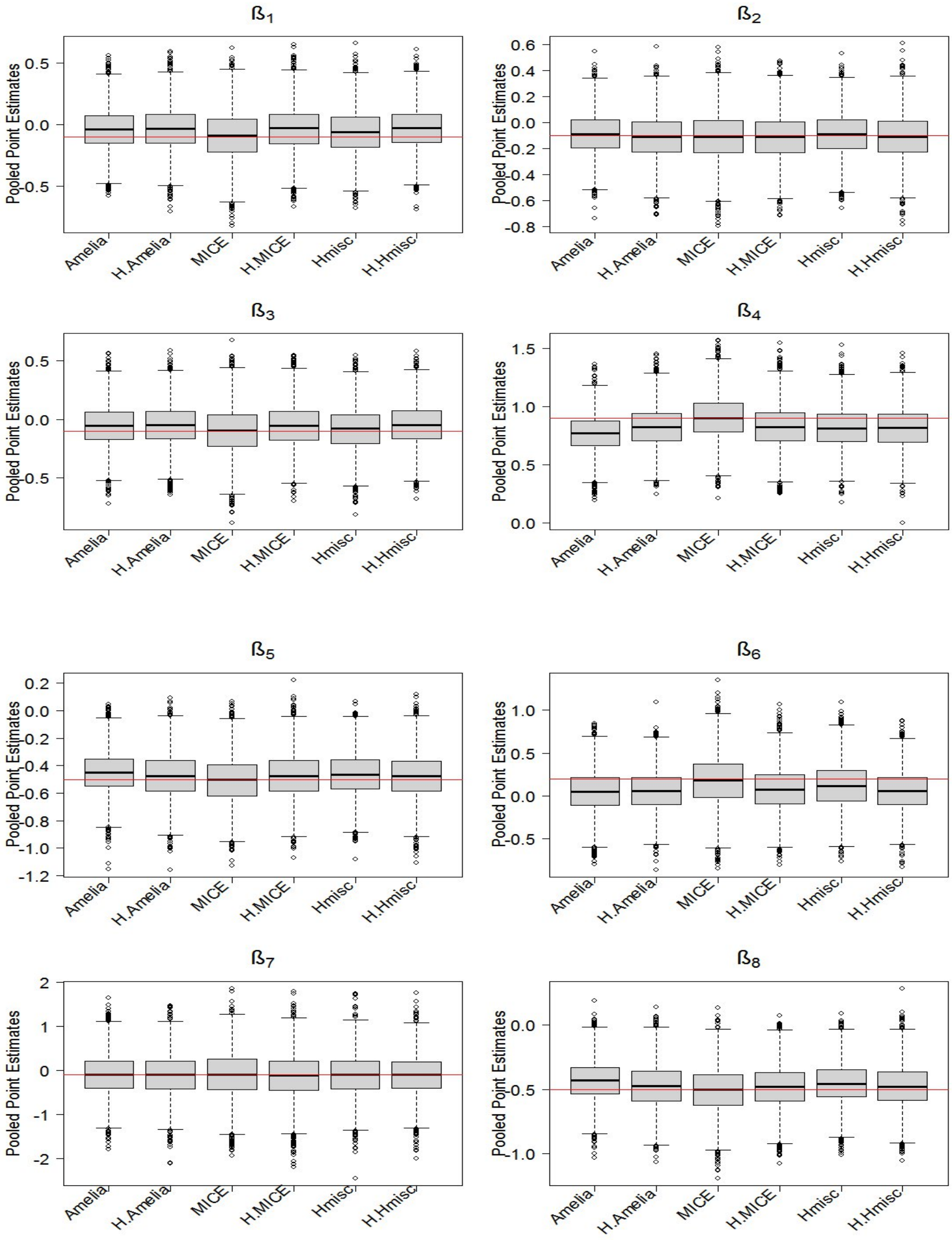

Figure 3. Simulated data: Pooled Point Estimates for 20 imputed data sets

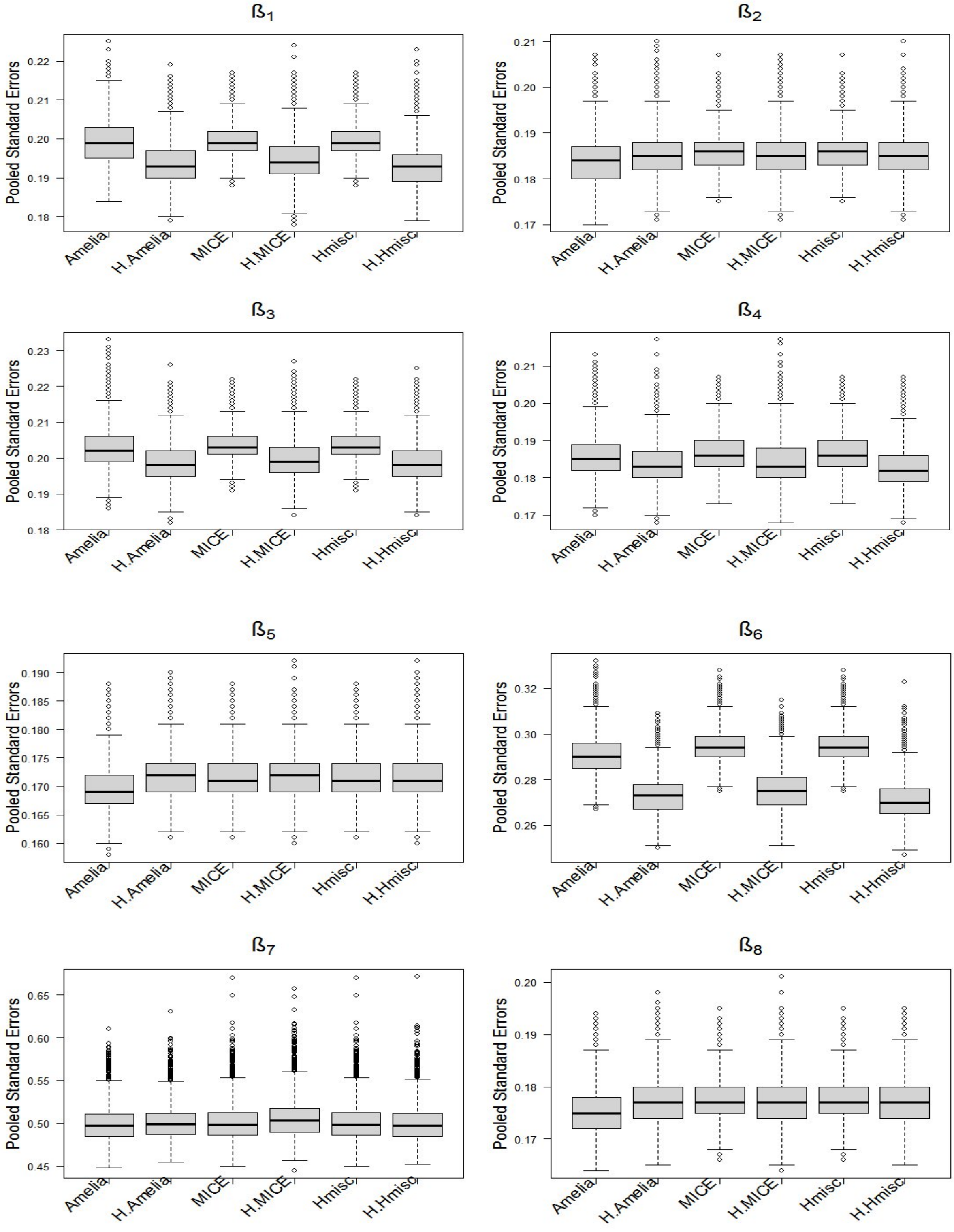

Figure 4 . Simulated data: Pooled Standard Errors for 20 imputed data sets

9 CONCLUDING REMARKS

Based on results obtained by simulations, we can make several general conclusions about the three HMI procedures. The three HMI methods use the JM approach for categorical data having complex dependency structure and three different algorithms for continuous variables. First, the default application of "MICE" (which uses logistic models for categorical and PMM for continuous variables), 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 as used beforehand. Of course, one could use various applications offered by MICE, i.e. classification and regression trees (CART) or PMM to automatically find and model important dependence structures which were not possible by default MICE. But it would be problematic to select appropriate sets of interaction effects to include in the conditional models in the high dimensional case. Second, identification of a clear winner between "Hmisc" and "H.Hmisc" is little difficult. The coverage rates of the 95% confidence intervals for "Hmisc" tend to be larger than the coverage rates for "H.Hmisc". The "H.Hmisc" tends to result in lower standard errors than "Hmisc". It tends to result in smaller RMSEs as well. Third, analysts may prefer "H.Amelia" for high coverage rates for most estimands with slight bias and due to its fastness[5]. 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. The generalization of these results may not be possible to all other settings. Extensive comparison is 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.

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.

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.

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.

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

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.

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.

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.

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.

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.

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.