A Comparison Of Multiple Imputation Methods For Categorical Data

by

Olanrewaju Michael Akande

Program in Statistical and Economic Modeling
Duke University

Date: _____

Approved:

_____
Jerome Reiter, Supervisor

_____
Fan Li, Co-Supervisor

_____
Michelle P. Connolly

Thesis submitted in partial fulfillment of the requirements for the degree of
Master of Science in the Program in Statistical and Economic Modeling
in the Graduate School of Duke University
2015

Abstract

A Comparison Of Multiple Imputation Methods For Categorical Data

by

Olanrewaju Michael Akande

Program in Statistical and Economic Modeling
Duke University

Date: _____

Approved:

_____
Jerome Reiter, Supervisor

_____
Fan Li, Co-Supervisor

_____
Michelle P. Connolly

An abstract of a thesis submitted in partial fulfillment of the requirements for the degree of Master of Science in the Program in Statistical and Economic Modeling
in the Graduate School of Duke University
2015

Abstract

This thesis evaluates the performance of several multiple imputation methods for categorical data, including multiple imputation by chained equations using generalized linear models, multiple imputation by chained equations using classification and regression trees and non-parametric Bayesian multiple imputation for categorical data (using the Dirichlet process mixture of products of multinomial distributions model). The performance of each method is evaluated with repeated sampling studies using housing unit data from the American Community Survey 2012. These data afford exploration of practical problems such as multicollinearity and large dimensions. This thesis highlights some advantages and limitations of each method compared to others. Finally, it provides suggestions on which method should be preferred, and conditions under which the suggestions hold.

To God Almighty, by whose grace I was able to complete this work, and to my family, my greatest supporters

Contents

List of Tables

List of Figures

List of Abbreviations and Symbols

Abbreviations

CART	Multiple Imputation by Classification and Regression Trees.
DPMPM	Dirichlet Process Mixture of Products of Multinomial Distributions Model.
FCS	Fully Conditional Specification.
GLM	Generalized Linear Model.
JM	Joint Modeling.
LCM	Latent Class Model
MAR	Missing at Random.
MCAR	Missing Completely At Random.
MCMC	Markov Chain Monte Carlo.
MI	Multiple Imputation.
MICE	Multiple Imputation by Chained Equations.
MNAR	Missing Not at Random.

Acknowledgements

I would like to thank everyone in the Department of Statistical Science. I thank my supervisor, Jerry Reiter, for all his immense support, patience, advice, mentor-ship and dedication over the past year. He has been pivotal in my overall development. Many thanks to my co-supervisor, Fan Li for her support, patience and guidance as well. She has been very instrumental with her timely advice and direction, both in research and non-research matters. Thanks to Quanli Wang for his assistance throughout the research. Many thanks to all members of the faculty, especially Mike West, Li Ma, Alan Gelfand, and David Dunson. Thanks to my academic advisor Michelle Connolly and to Charles Becker for all his advice as well. This work was supported by a grant from the National Science Foundation NSF(SES-1131897). I also acknowledge the financial support by the National University Commission (NUC) under the umbrella of the Federal Government of Nigeria, on whose sponsorship I am studying for my Masters at Duke University.

I would like to thank my parents for their encouragement. I will always cherish all your efforts and I will be forever grateful for your unending prayers and support. God bless you so much. My sincere thanks to my brothers, Ayodeji, Anuoluwapo and Oluwasegun for always being there and to my sister, Oluwafunmilayo for being the best. To my MSEM friends and classmates, especially Kirti and Amaze, thanks for working together with me at different times and on different projects. You really helped me find my feet in the program. To the PhD students who helped in one

way or the other, Nicole, Monika, thank you so much. To my Jesus City Family, thank you for all the prayers. To all my other friends, Tiffany, Tunde, Bola, Oseleye, Obafemi, Wale, Dayo, Gbolahan, Lanre, God bless you all for your support.

Most importantly, I thank God for His grace, mercy and favor through which I was able to successfully complete this thesis.

1

Introduction

Large scale censuses and sample surveys (such as the American Community Survey, American Housing Survey, Economic Census) commonly suffer from item nonresponse – partial or complete missing responses for some of the items or questions asked. Item nonresponse occurs for a number of different reasons, the most common of which is that units might respond to only some or none of the survey questions (either at random or in a systematic manner), causing some or all of the information about them to be missing. Another common reason is that respondents might provide invalid answers to some questions causing interviewers to exclude the values of their responses. Missing values in the data render statistical inference based on either the complete cases (all variables that are observed) or the available cases (all variables that are observed for a particular analysis) to be possibly biased and invalid, and thus, complete data methods cannot be directly used to analyze the data (Rubin, 1976). Using either available or complete cases implies tossing out cases with fully or partially missing data which means, sacrificing information that could be used to increase precision. Thus, it is imperative to develop methods that would reduce information loss due to missing data, given that complete data analysis is

also subject to information loss. In addition, for complex surveys, using available cases complicates all survey-weighted inference because original weights are no longer meaningful (Si and Reiter, 2013).

The missing data mechanism in any data with missing units, can be at random (missing completely at random – MCAR –, or missing at random – MAR), or not (missing not at random – MNAR) (Rubin, 1987; Little and Rubin, 2002). Data are MCAR when the probability of missing data on a variable is unrelated to its values and to the values of other measured variables. MAR is less stringent and implies that the probability of missing data on a variable is unrelated to its values but is related to the values of other measured variables. Finally, data are MNAR if the probability of missing data on a variable is systematically related to the underlying values of the variables that are missing. For real data, the exact missing data mechanisms are often unknown, especially in large data sets and thus, certain plausible assumptions have to be made accordingly. Some challenges to MI in large data are also highlighted in Li et al. (2012).

One common approach to dealing with item nonresponse is multiple imputation (MI) (Rubin, 1987). MI replaces missing values in a data set by sampling multiple (M) values from their predictive distributions, creating M complete data sets that can be analyzed using complete data analysis methods. Different MI methods have been proposed by researchers. MI by chained equations (MICE) (Raghunathan et al., 2001; Buuren and Groothuis-Oudshoorn, 2011; Royston and White, 2011; Su et al., 2011) is a fully conditional specification (FCS) approach to MI, which specifies univariate conditional distributions on a variable-by-variable basis, thus, sampling the missing values iteratively from the specified conditional distributions. The conditional distributions are specified to fit the type of data (discrete or continuous) and the preference of the researcher (for example, predictive mean matching versus linear regression for continuous data). Theoretically, the univariate conditional dis-

tributions specified using MICE can be potentially incompatible (Arnold and Press, 1989; Gelman and Speed, 1993). The imputation by ordered monotone blocks (IMB) method (Li et al., 2014) attempts to improve on this theoretical drawback by combining the advantages of the feasible MICE strategy and the theoretically-valid sequential imputation strategy for monotone missing data. The joint modeling (JM) specification is also often used for MI. JM involves specifying a joint distribution for a particular data set and drawing imputations from the derived conditional distributions using Markov Chain Monte Carlo (MCMC) methods. One method based on the JM specification is, MI using a Dirichlet process mixture of products of multinomial distributions model (DPMPM) (Manrique-Vallier and Reiter, 2013; Si and Reiter, 2013), which provides a fully Bayesian, non-parametric approach to MI for high dimensional categorical data with or without structural zeros. For other MI methods and modeling choices, see Schafer (1997); Shen (2000); Hopke et al. (2001); Rubin (2003); Yucel and Zaslavsky (2005); Schenker et al. (2006); Zhou et al. (2010); Schafer (2012).

MI has been implemented in different software packages, including PROC MI in SAS (Inc., 2008), the MI macro in ML-wiN (Rasbash et al., 2009), ICE in STATA (Royston, 2004; Royston and White, 2011), SOLAS (Ltd., 2001), IVEware in SAS (Raghunathan et al., 2002), *mice* in R (Buuren and Groothuis-Oudshoorn, 2011), *mi* in R (Su et al., 2011) and *pan* in R (Schafer, 2012). MI has been successfully applied to different data types (Schafer, 1997; Raghunathan and Paulin, 1998; Schafer et al., 1998; Shen, 2000; Hopke et al., 2001; Rubin, 2003; Yucel and Zaslavsky, 2005; Schenker et al., 2006; Zhou et al., 2010; Schafer, 2012) but few studies have attempted evaluations of competing MI methodology. In general, evaluations are few; even fewer for categorical data (Tang et al., 2005; Lee and Carlin, 2010; Hardt et al., 2013; Kropko et al., 2014; Donneau et al., 2015), motivating this thesis. Several performance measures have been used in the few evaluation studies including, mean

estimates, coverage rates, standardized percentage difference, within-imputation and between-imputation variability, root mean squared error, time to convergence and so on (Tang et al., 2005; Kropko et al., 2014).

In this thesis, we evaluate and compare the performance of three competing MI methods for categorical data including MICE using generalized linear models (GLMs), MICE using classification and regression trees (CART) and MI using a Dirichlet process mixture of products of multinomial distributions model (DPMPM). We conduct a Monte Carlo simulation-based comparison using data from the American Community Survey (ACS). This thesis focuses on the performance of each method as an MI engine, in comparison to others. Finally, we evaluate performance based on coverage rate, mean squared error, bias and interval length for marginal, bivariate and trivariate probability estimates.

The rest of this thesis is organized as follows: Section 2 provides an overview of MI, and discusses the three different methods for implementing MI evaluated in this thesis. Section 3 describes the simulation study – the data and the simulation design, and Section 4 presents the results. Section 5 concludes with a discussion.

2

Multiple Imputation

2.1 Background

Single imputation fills in a value (called an imputed value) for each missing value in the incomplete data set. The imputed value can be drawn from the distribution of the observed data. For example, for any of the variables, the imputed values can be drawn from some distribution with mean and standard deviation equal to the mean and standard deviation of the observed values. MI addresses the major disadvantage of the single imputation method – the failure to reflect sampling variability about the actual value of the missing item (Rubin, 1987) – by considering more than one possible replacement to the missing item. MI creates $M > 1$ (usually, $M \leqslant 10$) complete datasets from the original data, where the missing responses have been replaced by imputed responses simulated from predictive distributions. The M imputed responses represent plausible values from the distribution of possibilities for the missing values. These datasets are then analyzed and/or released to the public. When the imputation models meet certain conditions (Rubin, 1987, Chapter 4), valid inference about the population, using complete data statistical methods

and software, can be obtained. Each completed data set is analyzed and estimates for quantities of interest are calculated. These estimates are combined for all M data sets using the combining rules proposed by Rubin (1987). These rules serve to incorporate the uncertainty introduced by missing data and imputation into the inferences.

Suppose we seek inference about some estimand Q from the population (for example, a population proportion or probability), and suppose we wish to estimates its value with some estimator q and the variance of q with some estimator u (Q is often a multivariate vector). Rubin (1987) suggests obtaining inference about Q from the M completed data sets by calculating q and u in each of the completed data sets and combining the values in all M data sets. Specifically, we need to calculate:

$$\bar{q}_M = \sum_{l=1}^{M} q^{(l)}/M, \tag{2.1}$$

$$b_M = \sum_{l=1}^{M} (q^{(l)} - \bar{q}_M)^2/(M-1), \tag{2.2}$$

$$\bar{u}_M = \sum_{l=1}^{M} u^{(l)}/M. \tag{2.3}$$

We can use $\bar{q}_M$ to estimate Q and $T_M = (1 + \frac{1}{M})b_M + \bar{u}_M$ to estimate the variance of $\bar{q}_M$. When n and M are large, inferences for scalar Q can be based on normal distributions. Thus, a (1-α)% confidence interval for Q is $\bar{q}_M \pm z(\alpha/2)\sqrt{T_M}$. When M is moderate, a (1-α)% confidence interval for Q is $\bar{q}_M \pm t_{v_M}(\alpha/2)\sqrt{T_M}$, where v_M is the degrees of freedom of the t-distribution and $v_M = (M-1)(1 + r_M^{-1})^2$, where $r_M = (1 + M^{-1})b_M/\bar{u}_M$ (Rubin, 1987; Reiter et al., 2006). In this thesis, 95% confidence intervals are constructed based on these combining rules. For more in-depth reviews of MI, see Rubin (1996), Barnard and Meng (1999), Reiter and Raghunathan (2007), and Harel and Zhou (2007).

There are two well known approaches for imputing multivariate data: the JM specification and the FCS (Buuren, 2007). JM involves specifying a joint distribution for a particular data set and drawing imputations from the derived conditional distributions using Markov Chain Monte Carlo (MCMC) methods. JM is attractive when there is reason or evidence to believe that the joint model is a reasonable description of the data. It is also attractive because the conditional distributions guarantee the existence of a proper underlying joint distribution by construction. However, JM can be difficult to implement especially when the conditional distributions do not have desirable closed forms. This is one of the reasons why many researchers prefer FCS.

FCS specifies conditional distributions on a variable-by-variable basis, one distribution for each variable with missing entries. Imputations are drawn by iterating over each conditional distribution. FCS is attractive when no reasonable joint distribution can be used to adequately describe the data. It is also attractive when the conditional distributions have simple closed forms that can be sampled from easily. However, this type of specification also implies that an underlying joint distribution based on the conditional distributions might not exist; this issue is known as incompatibility of conditionals (Gelman and Speed, 1993). Nevertheless, FCS has been implemented in many contexts and is now widely accepted (Brand, 1999; Buuren et al., 1999; Raghunathan et al., 2001; Rubin, 2003; Buuren et al., 2006; Buuren, 2007; Burgette and Reiter, 2000). For the FCS, a low number of iterations is often sufficient (Buuren and Groothuis-Oudshoorn, 2011).

This thesis only focuses on categorical data and considers three MI methods for categorical data, two of which are based on FCS, and the third is a JM specification. The methods considered are multiple imputation by chained equations using generalized linear models (an FCS), multiple imputation by chained equations using classification and regression trees (also FCS), and multiple imputation using a

Bayesian Dirichlet Process Mixture of Products of Multinomial Distributions Model (a JM specification). MI for categorical data using any of these methods, has potential challenges, and the ACS data provides a good number of categorical variables to explore those challenges. First, the default modeling choices for categorical data using MICE for most of the previously mentioned packages ignores potential higher order interaction terms. We explore this to see it affects the estimation of bivariate and trivariate probabilities. We push the limits of each method on the maximum number of categories that any categorical variable can have. Finally, high order interactions between categorical variables can result in cells that have close to zero individuals in them. This can potentially affect the performance of each method because the number of observed units in this cells is approximately zero in the population. We conduct the evaluation with and without some of these variables and explore the differences.

2.2 Notation

We introduce the notation in the context of categorical data. Let $\mathbf{Y}$ represent the data of n individuals measured on p categorical variables, and $\mathbf{Y}_j$ (with $j = 1, \ldots, p$) be one of the p incomplete variables, that is, $\mathbf{Y} = (\mathbf{Y}_1, \ldots, \mathbf{Y}_p)$. Then each individual i, $i = 1, \ldots, n$ has an associated response vector $\mathbf{Y}_i = (Y_{i1}, \ldots, Y_{ip})$ whose elements can take on a set of L_j levels such as each $Y_{ij} \in \{1, \ldots, L_j\}$, where the levels have been labeled using consecutive numbers. Thus, $\mathbf{Y}_i \in \mathcal{C} = \prod_{j=1}^{p} \{1, \ldots, L_j\}$ and each combination $\mathbf{Y}_i$ can thus be viewed as a cell in the high dimensional contingency table formed by $\mathcal{C}$. Let the observed and missing parts of $\mathbf{Y}_j$ be $\mathbf{Y}_j^{\text{obs}}$ and $\mathbf{Y}_j^{\text{mis}}$ respectively, so that $\mathbf{Y}^{\text{obs}} = (\mathbf{Y}_1^{\text{obs}}, \ldots, \mathbf{Y}_p^{\text{obs}})$ and $\mathbf{Y}^{\text{mis}} = (\mathbf{Y}_1^{\text{mis}}, \ldots, \mathbf{Y}_p^{\text{mis}})$ represent the observed and missing data in $\mathbf{Y}$, respectively. Similarly, $\mathbf{Y}_i = (\mathbf{Y}_i^{\text{obs}}, \mathbf{Y}_i^{\text{mis}})$, where $\mathbf{Y}_i^{\text{obs}}$ and $\mathbf{Y}_i^{\text{mis}}$ again represents the observed and missing values in the $\mathbf{Y}_i$ vector respectively.

The number of imputations is $M > 1$ and the lth imputed data set is denoted as $\mathbf{Y}^{(l)}$, where $l = 1, \ldots, M$. Also, let $\mathbf{Y}_{-j} = (\mathbf{Y}_1, \ldots, \mathbf{Y}_{j-1}, \mathbf{Y}_{j+1}, \ldots, \mathbf{Y}_p)$ denote the collection of the $p-1$ variables in $\mathbf{Y}$ except $\mathbf{Y}_j$.

2.3 Multiple Imputation by Chained Equations using Generalized Linear Models

Multiple imputation by chained equations (MICE), also called sequential regression modeling (Raghunathan et al., 2001), is the most commonly used MI method. Suppose the hypothetically complete data $\mathbf{Y}$ is a partially observed random sample from a joint multivariate distribution $P(\mathbf{Y}|\boldsymbol{\theta})$ which is completely specified by $\boldsymbol{\theta}$, the vector of unknown parameters. If this joint distribution and the true value of $\boldsymbol{\theta}$ is known, random samples for $\mathbf{Y}^{\text{mis}}$ can be easily obtained. MICE can be used for both discrete and continuous variables. MICE proposes obtaining a posterior distribution for $\boldsymbol{\theta}$, and thus $\mathbf{Y}^{\text{mis}}$, by sampling iteratively from conditional distributions of the form (Buuren and Groothuis-Oudshoorn, 2011)

$$\begin{array}{c} P(\mathbf{Y}_1|\mathbf{Y}_{-1}, \boldsymbol{\theta}_1) \\ \vdots \\ P(\mathbf{Y}_p|\mathbf{Y}_{-p}, \boldsymbol{\theta}_p). \end{array} \tag{2.4}$$

The set of parameters $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_p$ are specific to respective conditional distributions and do not have to be a result of some factorization of the joint distribution $P(\mathbf{Y}|\boldsymbol{\theta})$. Starting from a simple random draw from observed marginal distributions for each variable (where all missing values are initially filled with plausible values), MICE suggests successive draws using an algorithm akin to a Gibbs sampler that iterates

over (Buuren and Groothuis-Oudshoorn, 2011)

$$
\begin{aligned}
\boldsymbol{\theta}_1^{\star(t)} &\sim P(\boldsymbol{\theta}_1|\mathbf{Y}_1^{\text{obs}}, \mathbf{Y}_2^{(t-1)}, \ldots, \mathbf{Y}_p^{(t-1)}) \\
\mathbf{Y}_1^{\star(t)} &\sim P(\mathbf{Y}_1|\mathbf{Y}_1^{\text{obs}}, \mathbf{Y}_2^{(t-1)}, \ldots, \mathbf{Y}_p^{(t-1)}, \boldsymbol{\theta}_1^{\star(t)}) \\
&\vdots \\
\boldsymbol{\theta}_p^{\star(t)} &\sim P(\boldsymbol{\theta}_p|\mathbf{Y}_p^{\text{obs}}, \mathbf{Y}_1^{(t)}, \ldots, \mathbf{Y}_{p-1}^{(t)}) \\
\mathbf{Y}_p^{\star(t)} &\sim P(\mathbf{Y}_p|\mathbf{Y}_p^{\text{obs}}, \mathbf{Y}_1^{(t)}, \ldots, \mathbf{Y}_{p-1}^{(t)}, \boldsymbol{\theta}_p^{\star(t)}),
\end{aligned}
\tag{2.5}
$$

where $\mathbf{Y}_j^{(t)} = (\mathbf{Y}_j^{\text{obs}}, \mathbf{Y}_j^{\star(t)})$ is the imputed variable at the tth iteration. Previous imputations for a variable $\mathbf{Y}_j^{\star(t-1)}$ only contribute to $\mathbf{Y}_j^{\star(t)}$ through other variables and not directly. Simply put, the MICE specification says, regress $\mathbf{Y}_1$ on $\mathbf{Y}_2, \ldots, \mathbf{Y}_p$ and estimate the parameters in the regression model using individuals with observed $\mathbf{Y}_1$. Then, replace missing $\mathbf{Y}_1$ values with simulated draws from the posterior predictive distribution of $\mathbf{Y}_1$. Do the same sequentially for $\mathbf{Y}_2, \ldots, \mathbf{Y}_p$, each time regressing $\mathbf{Y}_j$ on the other $p-1$ variables. Repeat the cycle for a number of iterations usually between 10 and 20 (Buuren and Groothuis-Oudshoorn, 2011). The result is one completed dataset.

Several authors have developed software packages for the implementation of MICE in STATA (Royston and White, 2011), SAS (Raghunathan et al., 2001) and R (Buuren and Groothuis-Oudshoorn, 2014). A commonly used package is the R package *mi* developed by Su et al. (2011). This thesis uses the *mice* package in R developed by Buuren and Groothuis-Oudshoorn (2014). For continuous variables, the model options for the conditional distribution models in the *mice* package are predictive mean matching (default), Bayesian linear regression, non-Bayesian linear regression, unconditional mean imputation, two-level linear model and simple repeated random sampling from the observed data. For categorical variables, the options are logistic regression (default for 2 level factor variables), polytomous unordered regression (default for > 2 level factor variables), linear discriminant analysis

and simple repeated random sampling from observed data. The default distributions for categorical variables are retained in this thesis. Also, the default arguments (maximum number of iterations = 10) for the *mice* package were used.

2.4 Multiple Imputation by Chained Equations using Classification and Regression Trees

Multiple imputation by classification and regression trees (CART) (Breiman et al., 1984) is MICE but with classification and regression trees as the set of conditional distributions (Burgette and Reiter, 2000). The MICE using CART algorithm (henceforth, CART) is a nonparametric approach to MICE using GLMs (henceforth, MICE) and it partitions (by recursive splits) the predictor space in a way that can be effectively represented by a tree structure, with leaves corresponding to the subsets of units. The values in each leaf represent the conditional distribution of the outcome for units in the data with predictors that satisfy the partitioning criteria that define that leaf. For example, consider a tree structure for an outcome variable $\mathbf{Y}$ which has 2 predictors gender (male – M, or female – F) and race (African-American – A, Caucasian – C, or Hispanic – H). A leaf L1 might contain female African-Americans, another leaf L2 might contain male African-Americans, L3 might contain female Caucasians, L4 might contain male Caucasians and L5, Hispanics irrespective of their gender. Thus, if we wanted to approximate any conditional distribution, we would use values in the corresponding leaf for that conditional distribution. In particular, if we want to approximate the distribution of $\mathbf{Y}$ for female Caucasians, we would use a random (usually bootstrap or Bayesian bootstrap) sample from the values of $\mathbf{Y}$ in L3. The tree structure is shown in Figure 2.1.

Even though large trees can be difficult to interpret, this is not a major concern when using CART for imputations. Burgette and Reiter (2000) point out that categorical predictors with many levels can cause computational difficulties for CART.

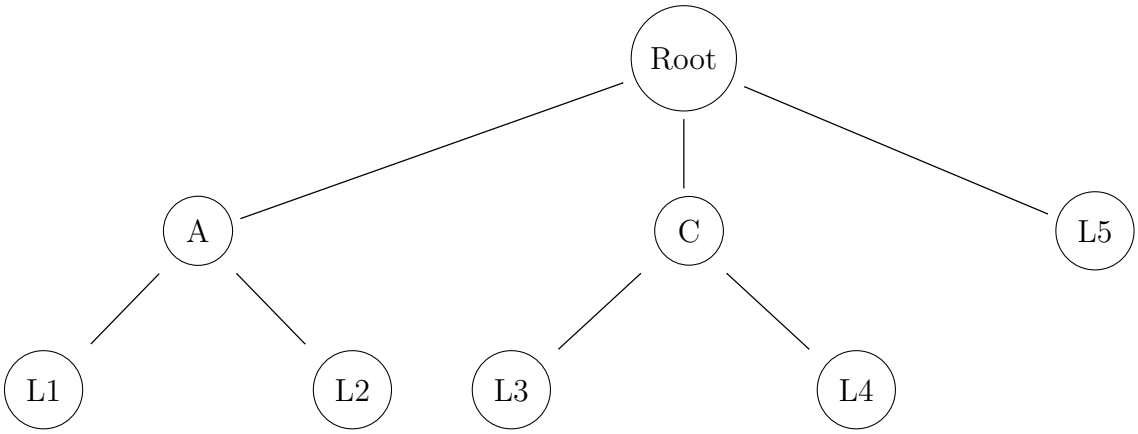

FIGURE 2.1: Illustrating the Tree Structure in CART

As an example, a categorical predictor with 32 levels results in over 2 billion potential partitions. CART is not a truly generating model (Burgette and Reiter, 2000), and this thesis seeks to explore the implications of this potential challenge for categorical data. Implementing sequential CART implies using CART models as the conditional distributions in (2.4). CART can be used for both discrete and continuous variables. For an extensive discussion about CART, see Burgette and Reiter (2000). In implementing MI by CART, this thesis uses the *cart* option in the *mice* package in R. Again, the default arguments (maximum number of iterations = 10 and the minimum number of observations in any terminal node – each leaf = 4) were retained.

2.5 Multiple Imputation using a Dirichlet Process Mixture of Products of Multinomial Distributions Model

Unlike MICE and CART, MI using a Dirichlet Process Mixture of Products of Multinomial Distributions Model (DPMPM) provides a fully Bayesian, non-parametric JM approach to MI for high dimensional categorical data with structural zeros (Manrique-Vallier and Reiter, 2013) or without structural zeros (Si and Reiter, 2013).

The DPMPM was proposed initially by Dunson and Xing (2009) as a nonparametric approach to inference on multivariate categorical data, and it provides full support on the space of all possible distributions. The latent class model (LCM) without any missing data is a finite mixture of product-multinomial distributions,

$$p(\mathbf{Y}|\boldsymbol{\lambda}, \boldsymbol{\pi}) = f^{\text{LCM}}(\mathbf{Y}|\boldsymbol{\lambda}, \boldsymbol{\pi}) = \sum_{k=1}^{K} \pi_k \prod_{j=1}^{p} \lambda_{jk}[Y_j], \tag{2.6}$$

where $\boldsymbol{\lambda} = \{\lambda_{jk}[l]\}$, all $\lambda_{jk}[l] > 0$ and $\sum_{l=1}^{L_j} \lambda_{jk}[l] = 1$. Here, $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_K)$, where $\sum_{k=1}^{K} \pi_k = 1$. Data under this model can be generated using,

$$Y_{ij}|z_i \overset{indep}{\sim} \text{Discrete}_1 : L_j(\lambda_{jz_i}[1], \ldots, \lambda_{jz_i}[L_j]) \quad \text{for all } i \text{ and } j, \tag{2.7}$$

$$z_i|\boldsymbol{\theta} \overset{iid}{\sim} \text{Discrete}_{1:K}(\theta_1, \ldots, \theta_K) \quad \text{for all } i. \tag{2.8}$$

For prior distributions, Si and Reiter (2013) and Manrique-Vallier and Reiter (2012) use

$$\lambda_{jk}[.] \overset{indep}{\sim} \text{Dirichlet}(\mathbf{1}_{L_j}), \tag{2.9}$$

$$\pi_k = V_k \prod_{h<k} (1 - V_h), \tag{2.10}$$

$$V_k \overset{iid}{\sim} \text{Beta}(1, \alpha) \text{ for } k = 1, \ldots, K-1; V_K = 1, \tag{2.11}$$

$$\alpha \sim \text{Gamma}(0.25, 0.25). \tag{2.12}$$

Estimation of parameters and posterior is straightforward through a Gibbs sampler as outlined in Si and Reiter (2013). Using the model as an MI engine simply includes the full conditional distribution of $\mathbf{Y}_i^{\text{mis}}$ in the Gibbs sampler. For an extensive description of the model, its justification as an MI engine, and the Gibbs Sampler, see Si and Reiter (2013). For implementation of the DPMPM, this thesis uses the *NPBayesImpute* package in R developed by Manrique-Vallier et al. (2014). K was set to 35 (after tuning with initial runs). The number of MCMC iterations was set to 10000 with burnin of 2000 and thining after every 20 iterations.

3

Simulation Study

3.1 Data

The data used was obtained from the United States Census Bureau and is available for general public access. The data represents housing unit data from the 2012 American Community Survey and contains about 1.5 million housing units (exactly 1477091) and 206 variables (44 flag variables, and 80 replicate weights). Obvious structural zeros (structural zeros represent impossible combination of categories – for example, a 12-year old female cannot be married in the U.S.) were eliminated by deleting variables. This thesis proceeds in this manner in an attempt to make the study as real as possible. Also, because interest is only in dealing with categorical variables, all continuous variables were removed. Thus, only a subset of the housing with 671153 housing units (individuals) and 37 categorical variables was used. A description of the data is presented in Appendix A.

3.2 Simulation Design

This thesis evaluates the performance of MICE, CART and the DPMPM using repeated sampling. Because of the inability of the *mice* package in R to handle categorical variables with more than 10 categories (for the implementing MICE but not CART – Buuren and Groothuis-Oudshoorn (2014)), 7 variables with more than 10 categories were first removed, leaving 28 variables. Those variables were then re-added for a comparison between CART and the DPMPM only. The subset of the data with 671153 housing units and 28 categorical variables was treated as a population from which 200 independent random samples of size $n = 10000$ (the sampling was also repeated for $n = 1000$) were taken from. For each sample, missing data was imposed by randomly deleting 30% (we also consider 45%) of the recorded item-level values of each variable (missing completely at random). In all 200 simulation runs, $M = 10$ completed datasets were created and all marginal one-way, and all joint two-way and three-way probabilities were calculated as estimands. Marginal one-way, and all joint two-way and three-way probabilities were calculated from the population and called "truth".

In each sample, 95% confidence intervals were calculated for each estimand using the combining rules of Rubin (1987). For each estimand, coverage rates (based on the 95% confidence intervals), mean squared error, average interval length and ratio of the average of T_M to the variance of $\bar{q}_M$ were calculated (see Section 2.1 for the definitions of T_M and $\bar{q}_M$) over all 200 simulations. All measures and estimands were also computed with each sample before introducing missing values; these are called "pre-missing data". The same was done with the imputed data sets using all three methods (MICE, CART and DPMPM). Although the missing completely at random missing data mechanism is adopted in this thesis in generating the missing data, an extension of this thesis will consider the MAR missing data mechanism.

3.3 Performance Measures

The comparison between the different MI methods was done using five different measures, all of which were calculated for marginal one-way, and all joint two-way and three-way probabilities.

1. *Coverage Rate.* This was calculated as the proportion of time, out of the 200 different samples, that the 95% confidence interval actually contained the "truth". Thus, high quality of the imputations generally implies coverage rates of at least 0.95 or close to the corresponding coverage rates in the "pre-missing data". For $i = 1, \ldots, 200$, we define

$$\text{Coverage}(\bar{q}_M) = \frac{\sum_{i=1}^{200} \mathbf{1}[Q \in \text{CI}(\bar{q}_{M(i)}, T_{M(i)})]}{200}, \tag{3.1}$$

   where $\mathbf{1}[Q \in \text{CI}(\bar{q}_{M(i)}, T_{M(i)})]$ is an indicator function which is equal to one when the confidence interval based on $\bar{q}_{M(i)}$ and $T_{M(i)}$ contains Q and equal to zero otherwise.

2. *Relative Mean Squared Error.* Mean squared error (MSE) was considered as a measure of the difference between $\bar{q}_M$ and Q. For easy interpretation and because MSE values are usually extremely low for bivariate and trivariate probability estimates, this thesis considered a rescaled version, relative MSE, defined as

$$\text{Rel.MSE}(\bar{q}_M) = \frac{\text{MSE } (\bar{q}_M) \text{ based on imputed data for MI method}}{\text{MSE } (q) \text{ based on pre-missing data}}. \tag{3.2}$$

   This is a measure of how inflated the MSE, based on any of the three MI methods, is compared to the MSE for the same sample with no missing data. Intuitively, if the imputations for a method are good, the MSE for that method

should be just as low as the MSE based on the pre-missing data. Quality imputations should thus result in values around 1.

3. *Normalized Bias.* The MSE was further broken down into its bias and variance components. We define a rescaled version of bias, normalized bias, as

$$\text{N.Bias}(\bar{q}_M) = \left| \frac{\text{Bias } (\bar{q}_M) \text{ based on imputed data for MI method}}{Q} \right|. \quad (3.3)$$

to provide perspective to how large the bias of an estimator is from the truth, when compared to that truth. Lower values would indicate higher quality of imputation.

4. *Variance Ratio.* As a check on the quality of the MI combining rules, we compute the variance ratio as

$$\text{Var.Rat}(\bar{q}_M) = \frac{\text{Average value of } T_M \text{ across all 200 simulations}}{\text{Variance of } \bar{q}_M \text{ across all 200 simulations}}. \quad (3.4)$$

According to Rubin (1987), one would expect $\mathbb{E}_{200}\left[(\bar{q}_M - Q)^2\right] \approxeq \sum_{i=1}^{200} T_{M(i)}$, and thus, the ratio between the two would provide insight into how far off they are. Reliable imputation procedures should result in values close to 1.

5. *Interval Length Ratio.* The length of the 95% confidence intervals for $\bar{q}_M$ (and q for the pre-missing data) were calculated and the average length of the interval across all 200 simulations was also calculated. The interval length ratio is defined as

$$\text{ILR}(\bar{q}_M) = \frac{\text{average interval length for } \bar{q}_M \text{ based on imputed data for MI method}}{\text{average interval length for } q \text{ based on pre-missing data}}. \quad (3.5)$$

4

Results

The key results are presented below in the next four subsections; more results are presented in Appendix B. In all box plots, MICE is abbreviated as MIC, CART as CAR, DPMPM as DP, and pre-missing data results as NO. Bivariate probabilities are cell counts for individuals in all possible joint two way cells for all the variables and similarly, trivariate probabilities are based on all possible three way cells in the large contingency table of all the variables. We present box plots of the log normalized bias in the results, in place of normalized bias, for clear presentation and discussion of findings. Due to the large number of bivariate and trivariate probabilities, this thesis further considers bivariate probabilities satisfying $np > 10$ and $n(1-p) > 10$, and trivariate probabilities satisfying $0.2 < p < 0.8$ to zoom in on results for cells with a reasonable number of individuals. These results are presented in Appendix B.

4.1 28 Variables: Sample Size of 10000 and 30% Missing

First, the results for all 28 variables with repeated sampling sample size of 10000 and 30% missing completely at random are presented. There are 98 marginal probabili-

ties, 4594 bivariate probabilities (3343 satisfying $np > 10$) and 137172 (3546 satisfying $0.2 < p < 0.8$) trivariate probabilities. For marginal probabilities, the DPMPM appears to have better coverage rates (greater than 80% coverage) than both MICE and CART, and really good coverage rates when compared to the coverage rates based on the complete data with no missing data (see Figure 4.1). For bivariate and trivariate probabilities, CART appears to have better coverage than both MICE and the DPMPM. This suggests that for data similar to this, if the analyst is particularly interested in higher order interactions, CART provides imputations that yield probability estimates with better coverage rates than MICE and the DPMPM. Comparing

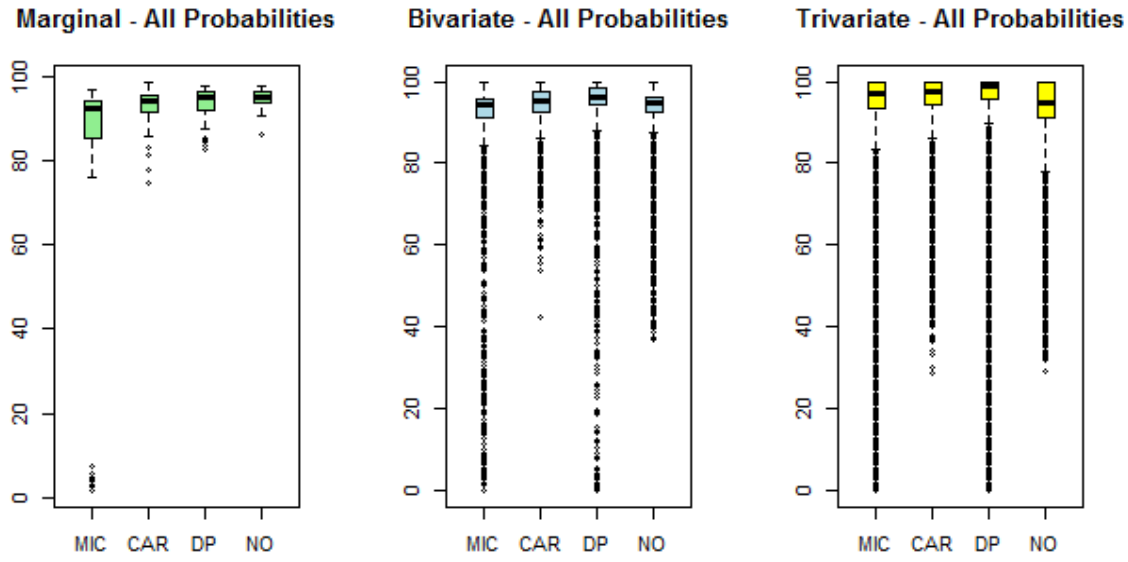

FIGURE 4.1: Coverage Rate For All Three Methods (28 Variables: Sample Size of 10000 & 30% Missing)

the distribution of coverage rates for all three methods to the distribution of coverage rates for the pre-missing data reveal that all three methods do have quality imputations and good performance even though their distributions are skewed to the left (long left tails). For the marginal probabilities, MICE in particular performs very well but does badly a few times. A closer look into the left tail reveal that

the results were being confounded by 7 variables. These are yes/no variables, with marginal probabilities for yes for all 7 variables, in the population, almost equal to 1. The variables are yes or no questions about the presence of bathtub, refrigerator, running water, sink, stove, telephone and flush toilet in the households (see Appendix A for the definition of the variables). These variables were also affecting the performance of the DPMPM for bivariate and trivariate probabilities and thus, they were removed and the simulation, rerun. The results are presented in the next section. CART gave estimates with the least relative mean squared error, although

Table 4.1: Relative Mean Squared Error For Marginal Probabilities For All Three Methods (28 Variables: Sample Size of 10000 & 30% Missing)

	MICE	CART	DP
Min.	1.04	1.01	1.01
1st Qu.	1.41	1.19	1.27
Median	1.61	1.35	1.46
Mean	688.40	676.80	2.17
3rd Qu.	2.56	1.46	1.95
Max.	27670.00	27550.00	6.71

Table 4.2: Relative Mean Squared Error For Bivariate Probabilities For All Three Methods (28 Variables: Sample Size of 10000 & 30% Missing)

	MICE	CART	DP
Min.	0.97	0.07	0.48
1st Qu.	1.46	1.06	1.21
Median	1.76	1.27	1.44
Mean	65.22	57.72	2.80
3rd Qu.	2.61	1.53	2.07
Max.	39530.00	38530.00	1236.00

this holds more clearly for bivariate and trivariate probabilities (Tables 4.1, 4.2 and 4.3). For marginal probabilities, CART and the DPMPM appear to be more similar than MICE. Due to the presence of outliers, the mean relative mean squared error is biased and is not used as a measure of performance. The outliers in Tables 4.1, 4.2 and 4.3 are due to bivariate and trivariate interactions of the 7 variables mentioned

Table 4.3: Relative Mean Squared Error For Trivariate Probabilities For All Three Methods (28 Variables: Sample Size of 10000 & 30% Missing)

	MICE	CART	DP
Min.	0.01	0.00	0.04
1st Qu.	1.21	0.80	0.97
Median	1.56	1.07	1.31
Mean	9.65	5.57	2.55
3rd Qu.	2.29	1.3890	1.80
Max.	49040.00	47150.00	5120.00

earlier which have low probabilities in the population. For example, the probability of having a flush toilet (TOIL = 1; see Table A.2) but no bathtub or shower (BATH = 2; see Table A.1) in the population is approximately 0. CART also has the least bias of all the three methods (see Figure 4.2). The interval length and variance ratios (see Appendix B) also reveal similar results to those already discussed.

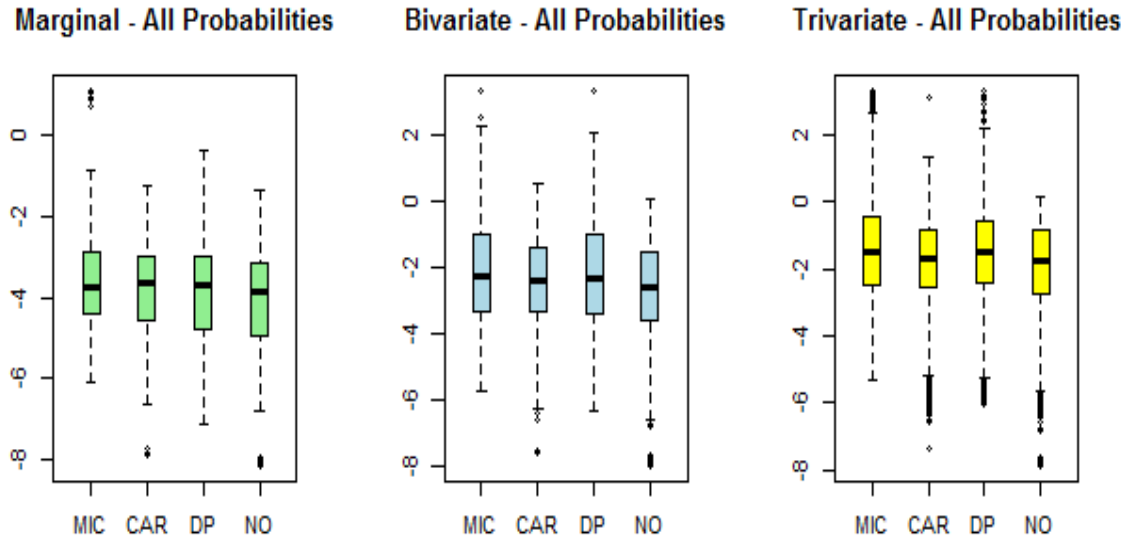

FIGURE 4.2: Log Normalized Bias For All Three Methods (28 Variables: Sample Size of 10000 & 30% Missing)

4.2 21 Variables: Sample Size of 10000 and 30% Missing

Removing all 7 variables mentioned above and repeating the simulation revealed similar results for coverage rates, although coverage rates appear to be better overall than before (Figure 4.3). There are now 83 marginal probabilities, 3253 bivariate probabilities (2590 satisfying $np > 10$) and 80069 (1360 satisfying $0.2 < p < 0.8$) trivariate probabilities. For coverage rates, the DPMPM still appears to have the

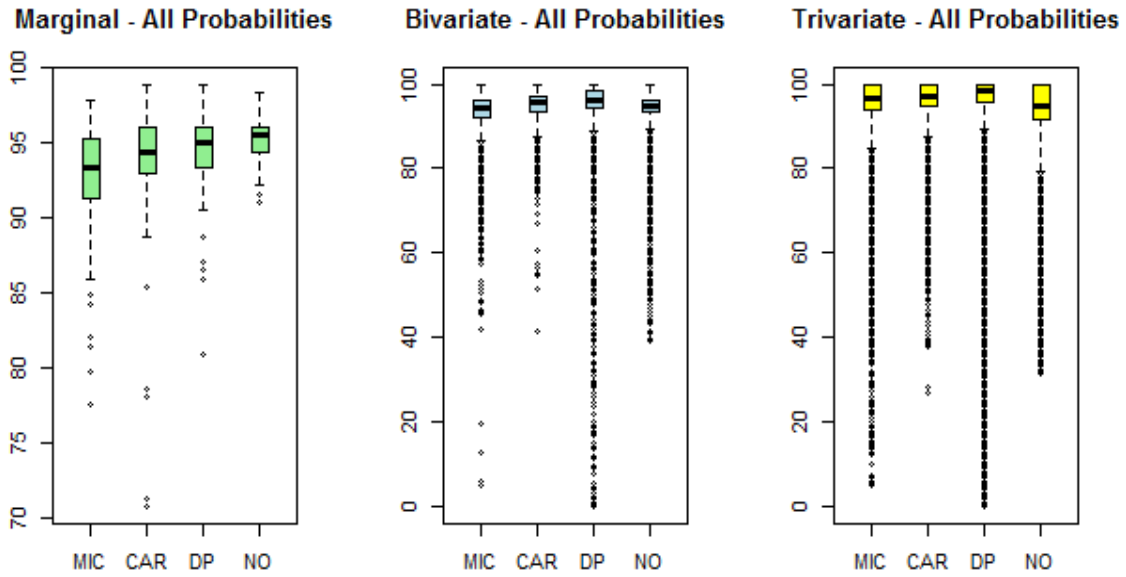

FIGURE 4.3: Coverage Rate For All Three Methods (21 Variables: Sample Size of 10000 & 30% Missing)

best performance for marginal probabilities (Figure 4.3). For bivariate and trivariate probabilities, the method with the best performance is not as clear. The distribution of coverage rates for the DPMPM has most of its density concentrated around 95% coverage while CART is mostly concentrated around 90% (see Figure 4.3). However, the DPMPM has a longer left tail than CART. Highly conservative researchers might prefer CART in this case while less conservative researches might prefer the DPMPM. CART still has the least relative mean squared error for marginal, bivariate and trivariate probabilities (see Tables 4.4, 4.5 and 4.6) but this again holds more

clearly for bivariate and trivariate probabilities. Removing the 7 variables improved distribution of the relative mean squared error in the right tail with less outliers than before. The outliers in Tables 4.4, 4.5 and 4.6 are still due to bivariate and trivariate interactions which have low probabilities in the population. For example, the probability of having no children in the household (HUPAC = 4; see Table A.2) and having one or more persons under 18 present in the household (R18 = 2; See Table A.2) in the population is approximately 0. As seen in Figure 4.4, CART again

Table 4.4: Relative Mean Squared Error For Marginal Probabilities For All Three Methods (21 Variables: Sample Size of 10000 & 30% Missing)

	MICE	CART	DP
Min.	1.00	1.01	1.00
1st Qu.	1.33	1.14	1.21
Median	1.49	1.34	1.38
Mean	1.71	1.38	1.68
3rd Qu.	1.67	1.48	1.67
Max.	4.19	2.96	8.05

Table 4.5: Relative Mean Squared Error For Bivariate Probabilities For All Three Methods (21 Variables: Sample Size of 10000 & 30% Missing)

	MICE	CART	DP
Min.	0.42	0.30	0.43
1st Qu.	1.39	1.09	1.20
Median	1.61	1.29	1.40
Mean	2.06	1.38	2.67
3rd Qu.	1.96	1.50	1.80
Max.	777.70	12.29	955.10

has the least (normalized) bias compared to both MICE and the DPMPM. Results for the other measures of performance are presented in Appendix B

4.3 21 Variables: Sample Size of 1000 and 30% Missing

The 7 variables were again left out and the sample size was reduced from 10000 to 1000 to examine the sensitivity of the results to varying sample size. The conclusion

Table 4.6: Relative Mean Squared Error For Trivariate Probabilities For All Three Methods (21 Variables: Sample Size of 10000 & 30% Missing)

	MICE	CART	DP
Min.	0.01	0.00	0.04
1st Qu.	1.09	0.85	0.91
Median	1.36	1.09	1.26
Mean	1.93	1.25	2.52
3rd Qu.	1.66	1.37	1.68
Max.	6868.00	1759.00	3373.00

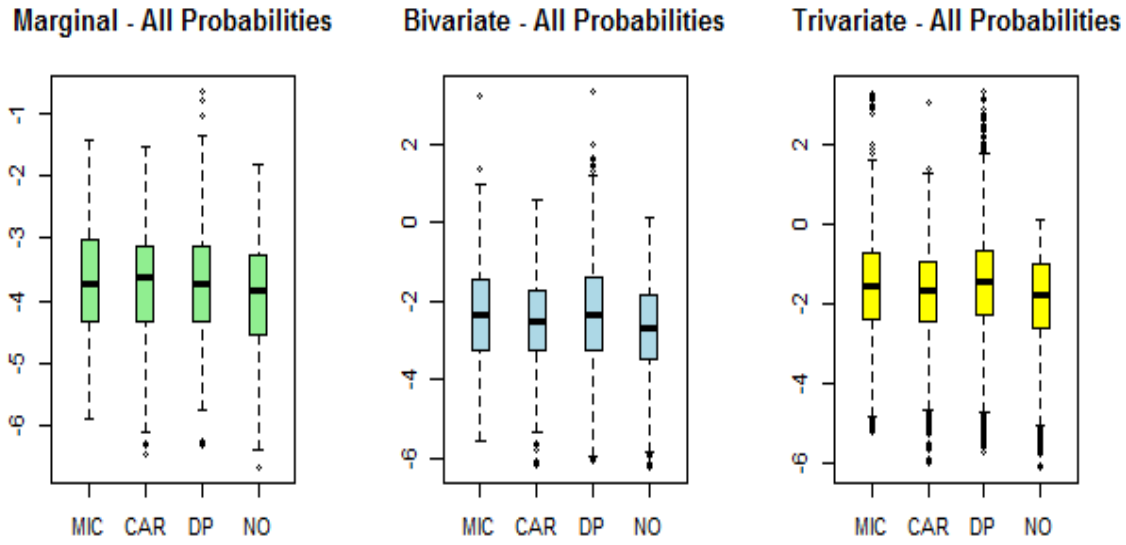

FIGURE 4.4: Log Normalized Bias For All Three Methods (21 Variables: Sample Size of 10000 & 30% Missing)

is the same as before although performance of each method is clearly reduced (Figures 4.5 and 4.6, Tables 4.7, 4.8 ad 4.9, and Appendix B). This is expected; precision is reduced as sample size decreases. We note here that relative mean squared error is more similar for CART and the DPMPM than before.

4.4 21 Variables: Sample Size of 10000 and 45% Missing

Leaving out all 7 variables and increasing the sample size back to 10000 but increasing the proportion of missing data to 45% yields the same result. It is important to note

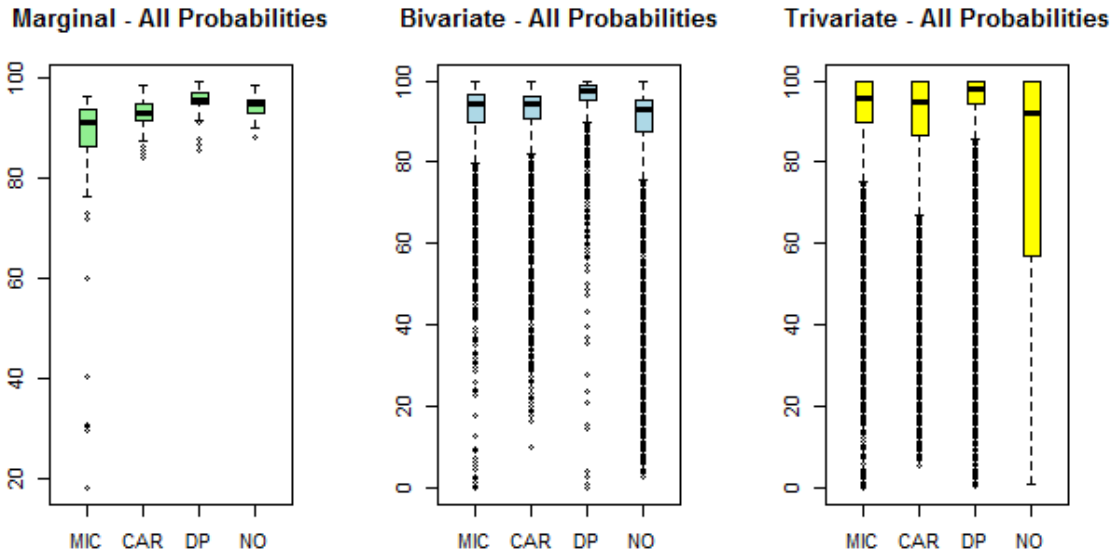

FIGURE 4.5: Coverage Rate For All Three Methods (21 Variables: Sample Size of 1000 & 30% Missing)

Table 4.7: Relative Mean Squared Error For Marginal Probabilities For All Three Methods (21 Variables: Sample Size of 1000 & 30% Missing)

	MICE	CART	DP
Min.	1.05	1.00	1.03
1st Qu.	1.44	1.15	1.21
Median	2.06	1.30	1.38
Mean	4.20	1.55	1.85
3rd Qu.	3.58	1.44	1.57
Max.	27.80	14.23	8.88

though that the performance of all methods is reduced significantly. Clearly, reducing sample size or increasing the proportion of missing data has similar effects on MI using these methods.

4.5 CART vs DPMPM comparison with 28 Variables: Sample Size of 10000 and 30% Missing

As previously mentioned, because of the inability of the *mice* package in R to handle categorical variables with more than 10 categories, 7 variables which had more than

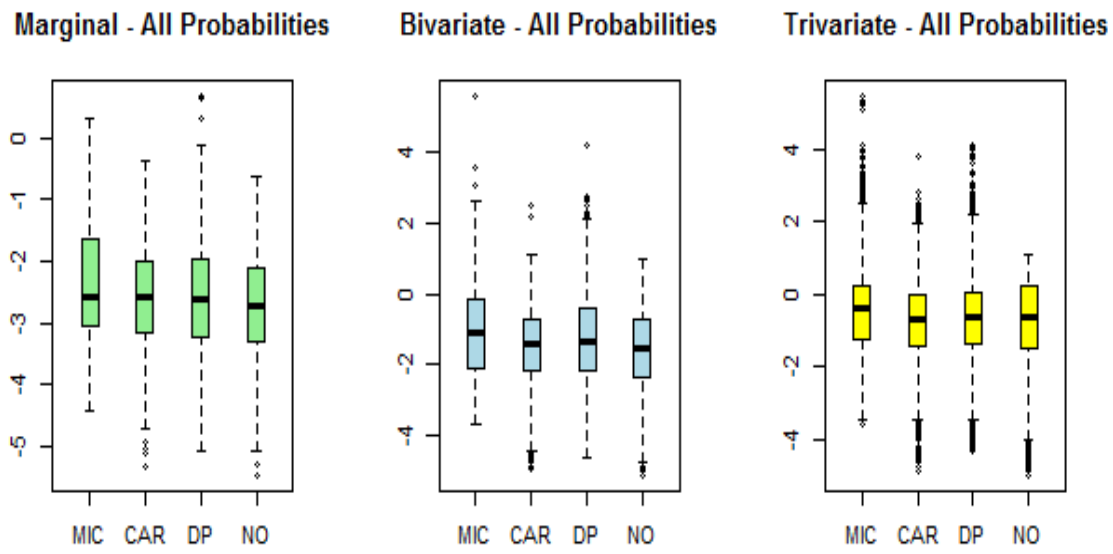

FIGURE 4.6: Log Normalized Bias For All Three Methods (21 Variables: Sample Size of 1000 & 30% Missing)

Table 4.8: Relative Mean Squared Error For Bivariate Probabilities For All Three Methods (21 Variables: Sample Size of 1000 & 30% Missing)

	MICE	CART	DP
Min.	0.94	0.62	0.48
1st Qu.	1.48	1.04	1.12
Median	1.80	1.20	1.27
Mean	2.99	1.34	1.62
3rd Qu.	2.48	1.39	1.51
Max.	139.60	21.06	58.28

10 categories were removed. We re-added them for a comparison between CART and the DPMPM only. There are 196 marginal probabilities, 18058 bivariate probabilities and 1041532 trivariate probabilities in this comparison. Due to the large number of bivariate and trivariate probabilities, and to speed up computation time, we took a random sample of 5000 bivariate probabilities and 20000 trivariate probabilities. We conducted our comparison between CART and the DPMPM using these samples. The results mimic those already discussed and so we only include results for coverage rates and relative mean squared error in Appendix B (Figure B.1 and Table B.1).

Table 4.9: Relative Mean Squared Error For Trivariate Probabilities For All Three Methods (21 Variables: Sample Size of 1000 & 30% Missing)

	MICE	CART	DP
Min.	0.91	0.94	0.98
1st Qu.	2.40	1.14	1.23
Median	3.80	1.26	1.41
Mean	5.34	2.40	1.82
3rd Qu.	6.54	2.35	1.82
Max.	33.65	21.49	29.95

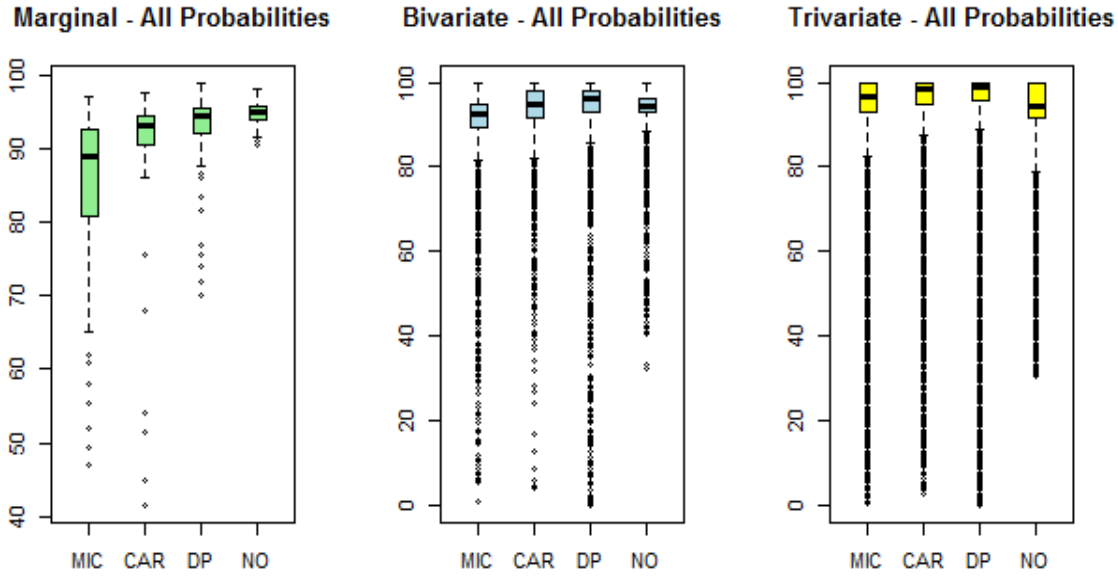

FIGURE 4.7: Coverage Rate For All Three Methods (21 Variables: Sample Size of 10000 & 45% Missing)

Table 4.10: Relative Mean Squared Error For Marginal Probabilities For All Three Methods (21 Variables: Sample Size of 10000 & 45% Missing)

	MICE	CART	DP
Min.	1.19	1.07	1.03
1st Qu.	1.81	1.40	1.38
Median	2.36	1.72	1.75
Mean	3.23	1.93	2.57
3rd Qu.	3.20	2.00	2.14
Max.	15.06	7.12	17.24

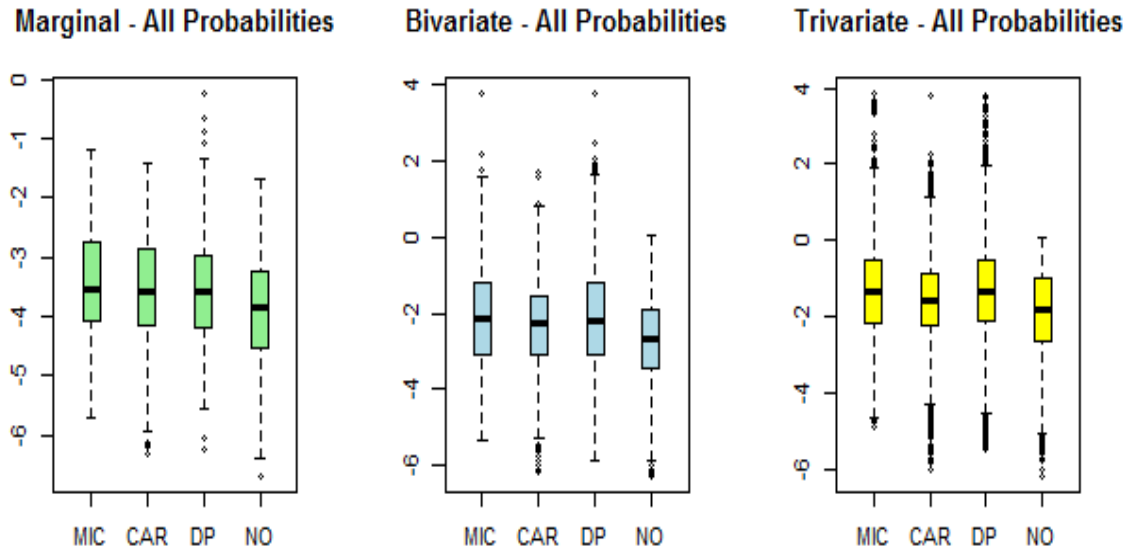

FIGURE 4.8: Log Normalized Bias For All Three Methods (21 Variables: Sample Size of 10000 & 45% Missing)

Table 4.11: Relative Mean Squared Error For Bivariate Probabilities For All Three Methods (21 Variables: Sample Size of 10000 & 45% Missing)

	MICE	CART	DP
Min.	0.98	0.25	0.33
1st Qu.	1.92	1.24	1.52
Median	2.50	1.67	1.85
Mean	4.07	2.15	5.06
3rd Qu.	3.33	2.22	2.89
Max.	2420.00	79.08	2665.00

Table 4.12: Relative Mean Squared Error For Trivariate Probabilities For All Three Methods (21 Variables: Sample Size of 10000 & 45% Missing)

	MICE	CART	DP
Min.	0.01	0.00	0.08
1st Qu.	1.37	0.81	1.07
Median	1.91	1.26	1.66
Mean	3.74	2.04	4.74
3rd Qu.	2.62	1.90	2.59
Max.	10500.00	5970.00	9799.00

5

Discussion

5.1 Findings and Conclusions

Overall, all three methods appear to produce reasonably high quality imputations. Any one can be used to impute categorical data, according to the preference of the researcher or analyst. However, relative to one another, CART appears to have better performance than MICE and the DPMPM (arguably, only slightly better than the DPMPM). This might be a somewhat surprising result given that one might expect the DPMPM to have the best performance, because a JM approach is usually more robust than the FCS. But the FCS using CART is very simple while the JM specification is in fact relatively complex. Thus, for categorical data, if the data has a simple data structure, with very few or no higher order interactions and no complex dependency structure between the variables, CART has a lower probability of giving false positive imputations. For data similar to the ACS data used in this thesis, based on the results presented in the previous section, we suggest MI using CART. An extension of this study would consider a dataset that has more complex structures between the variables than the ACS. Also, looking at each model closely, one would observe that the design of this study directly aligns with the strength of

CART. The recursive splits in the tree structure allows CART to effectively capture higher order interaction more accurately, as long as the number of individuals in each leaf is large enough.

While this thesis concludes that CART has the best performance of all three methods for categorical variables without structural zeros, we note a few different points mentioned in the results section. Researchers have to make a decision on the importance of variables when one of the categories has probability almost equal to 1. The DPMPM especially, struggles with such variables. For better performance overall, we suggest that researchers and analysts should consider removing such variables (such as the 7 variables removed from section 4.2 onwards). Quality of imputations for any of the three methods will be significantly lower for data sets with higher proportion of missing data and/or lower sample size.

5.2 Extensions and Future Work

A number of extensions to this study have already been mentioned in different sections so far. In an attempt to improve on the performance of the DPMPM, an extension of this study will look into the possibility of varying the initialization of the MCMC algorithm (initializing at the truth for a start). Future work will consider the MAR missing mechanism rather than the missing completely at random mechanism considered here. Lastly, as previously mentioned, data sets with more complex structures between the variables will be considered.

Appendix A

Data Description

The data used is a subset of 28 variables from the 2012 American Community Survey housing unit data.

ACS variable	Description	Observed Categories
ACR	Lot Size	1=GQ/not a one-family house or mobile home, 2=house less than one acre, 3=house on one to less than ten acres, 4=house on ten or more acres
AGS	Yearly Sales of Agricultural Products	1=GQ/vacant/not a one-family house or mobile home/less than 1 acre, 2=none, 3=$1-$999, 4=$1000-$2499, 5=$2500-$4999, 6=$5000-$9999, 7=$10000+
BATH	Bathtub or shower	1=yes, 2=no
BUS	Business or medical office on property	1=GQ/not a one-family house or mobile home, 2=yes, 3=no
REFR	Refrigerator	1=yes, 2=no
RWAT	Hot and cold running water	1=yes, 2=no
SINK	Sink with a faucet	1=yes, 2=no
STOV	Stove or range	1=yes, 2=no
TEL	Telephone	1=yes, 2=no, 3=suppressed

⋆ GQ = Group Quarters HH = Household

Table A.1: Subset of variables from ACS 2012 (part 1)

ACS variable	Description	Observed Categories
TEN	Tenure	1=owned with mortgage or loan (include home equity loans), 2=owned free and clear, 3=rented, 4=occupied without payment of rent
TOIL	Flush toilet	1=yes, 2=no
VEH	Vehicles available	1...6=number of vehicles, 7=none
HHL	Household language	1=English only, 2=Spanish, 3=other Indo-European languages, 4=Asian and Pacific Island languages, 5=other language
HHT	Household/family type	1=married-couple family household, 2=male householder, no wife present, 3=female householder, no husband present, 4=not living alone
HUGCLNPP	Grandparent Headed Household With No Parent Present	1=household without grandparent living with grandchildren, 2=household with grandparent living with grandchildren, 3=household with grandparent living with grandchildren and Grandparent headed household with no parent present
HUPAC	HH presence and age of children	1=with children under 6 years only, 2=with children 6 to 17 years only, 3=with children under 6 years and 6 to 17 years, 4=no children
HUPAOC	HH presence and age of own children	1=presence of own children under 6 years only, 2=presence of own children 6 to 17 years only, 3=presence of own children under 6 years and 6 to 17 years, 4=no own children present
HUPARC	HH presence and age of related children	1=presence of related children under 6 years only, 2=presence of related children 6 to 17 years only, 3=presence of related children under 6 years and 6 to 17 years, 4=no related children present
LNGI	The number of household members who are 14 and over speaks English only or speaks a language other than English and speaks English 'very well' in households	1=at least one, 2=no one
MULTG	Multigenerational household	1=no, 2=yes

⋆ GQ = Group Quarters HH = Household

Table A.2: Subset of variables from ACS 2012 (part 2)

ACS variable	Description	Categories
MV	When moved into this house or apartment	1=12 months or less, 2=13 to 23 months, 3=2 to 4 years, 4=5 to 9 years, 5=10 to 19 years, 6=20 to 29 years, 7=30 years or more
NR	Presence of nonrelative in household	1=none, 2=1 or more nonrelatives
PARTNER	Unmarried partner household	1=no unmarried partner in household, 2=male householder, male partner, 3=male householder, female partner, 4=female householder, female partner, 5=female householder, male partner
PSF	Presence of subfamilies in household	1=no subfamilies, 2=1 or more subfamilies
R18	Presence of persons under 18 years in household (unweighted)	1=no person under 18 in household, 2=1 or more persons under 18 in household
R65	Presence of persons 65 years and over in household (unweighted)	1=no person 65 and over, 2=1 person 65 and over, 3=2 or more persons 65 and over
SRNTVAL	Specified Rent or Value Owner Unit	1=not specified rent or value owner unit, 2=specified rent unit, 3=specified value owner unit
WIF	Workers in family during the past 12 months	1=1 worker, 2=2 workers, 3=3 or more workers, 4=GQ/vacant/non-family household, 5=no workers

⋆ GQ = Group Quarters HH = Household

Table A.3: Subset of variables from ACS 2012 (part 3)

Appendix B

Other Results

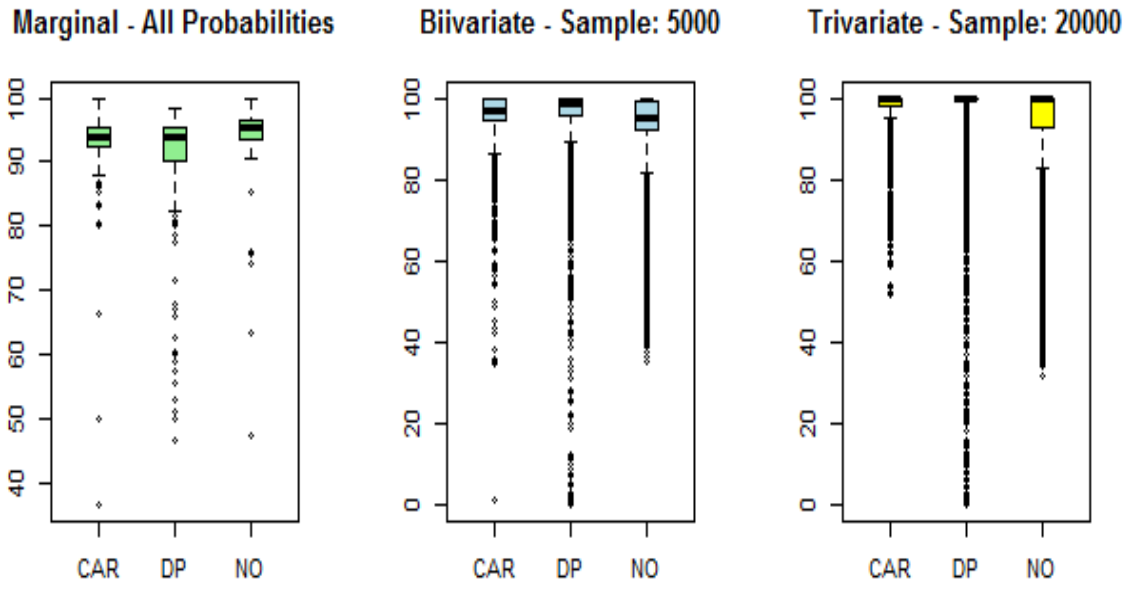

FIGURE B.1: Coverage Rate For CART vs DPMPM Comparison (28 Variables: Sample Size of 10000 & 30% Missing)

Table B.1: Relative Mean Squared Error For CART vs DPMPM Comparison (28 Variables: Sample Size of 10000 & 30% Missing)

	Marginal		Bivariate		Trivariate	
	CART	DP	CART	DP	CART	DP
Min.	0.72	0.99	0.00	0.29	0.00	0.00
1st Qu.	1.20	1.32	0.88	1.05	0.64	0.65
Median	1.34	1.59	1.10	1.38	0.82	0.94
Mean	1.36	147.50	1.20	21.78	0.89	5.61
3rd Qu.	1.47	2.70	1.31	2.21	1.05	1.48
Max.	2.19	9852.00	266.80	8720.00	52.71	6675.00

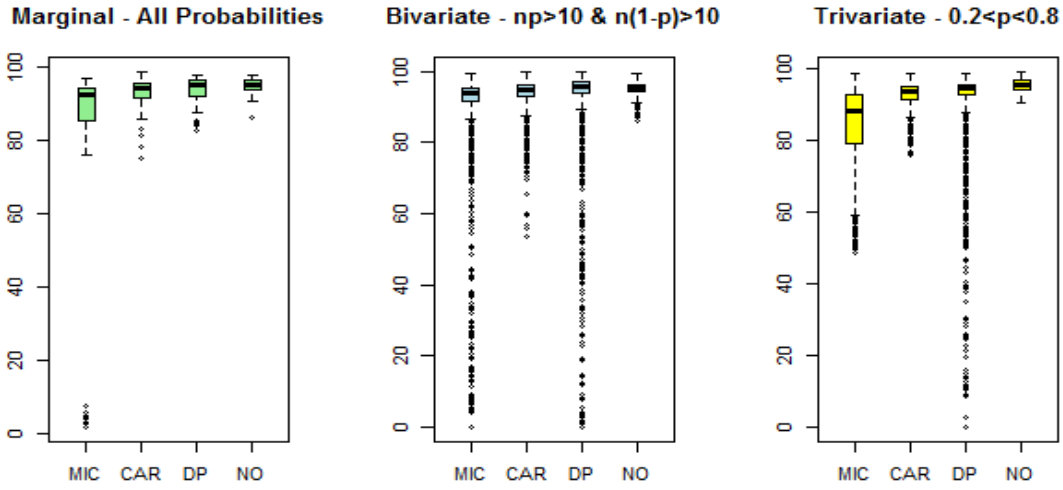

FIGURE B.2: Coverage Rate For All Three Methods (28 Variables: Sample Size of 10000 & 30% Missing)

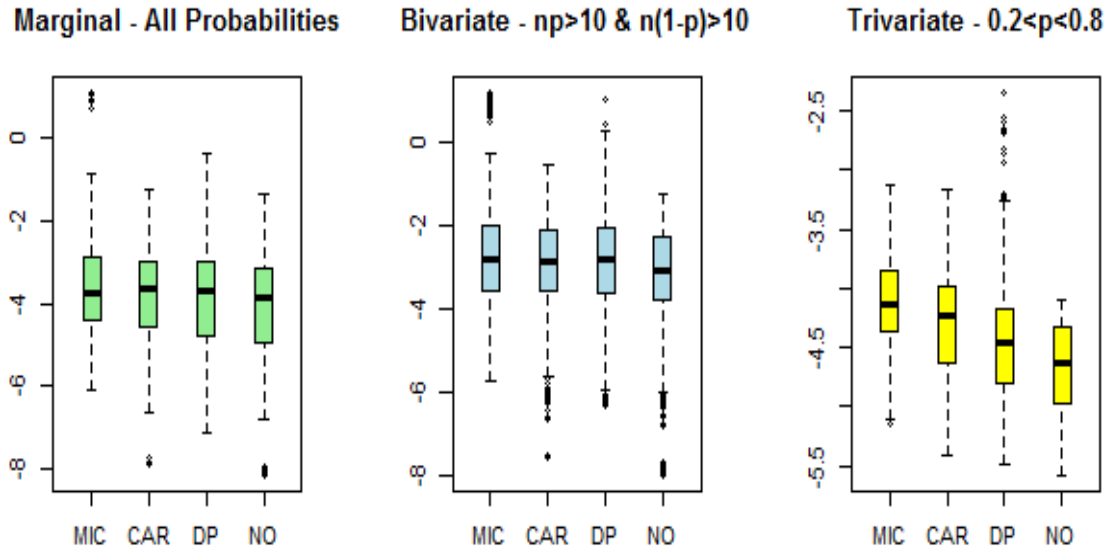

FIGURE B.3: Log Normalized Bias For All Three Methods (28 Variables: Sample Size of 10000 & 30% Missing)

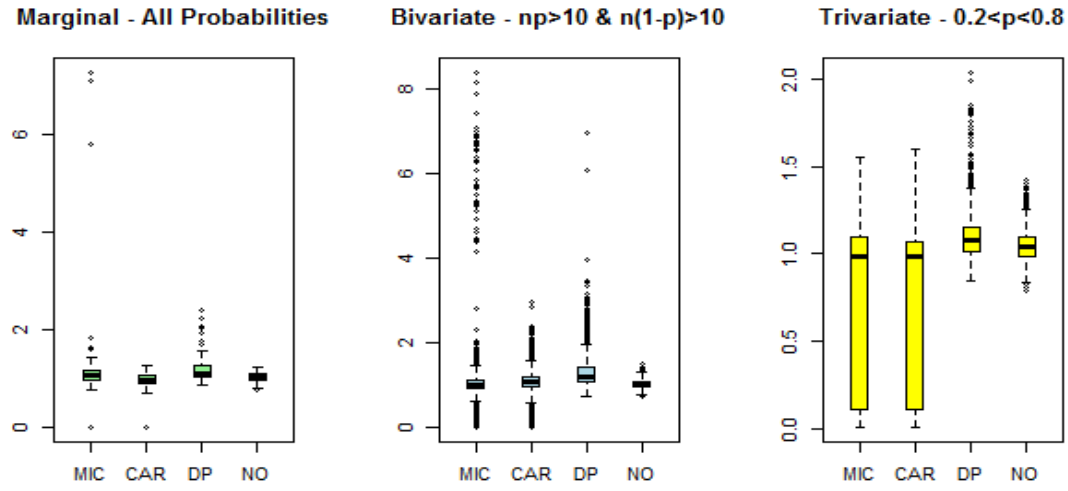

FIGURE B.4: Variance Ratio For All Three Methods (28 Variables: Sample Size of 10000 & 30% Missing)

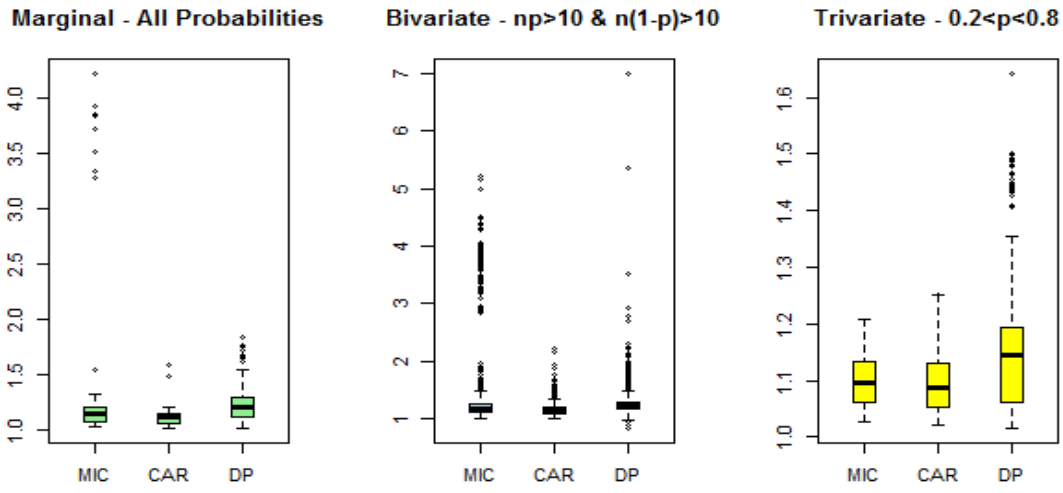

FIGURE B.5: Interval Length Ratio For All Three Methods (28 Variables: Sample Size of 10000 & 30% Missing)

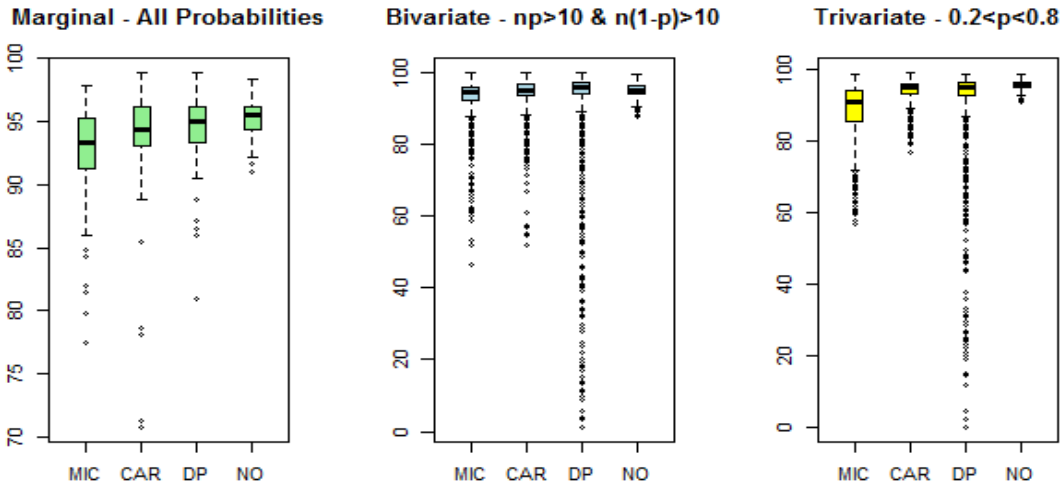

FIGURE B.6: Coverage Rate For All Three Methods (21 Variables: Sample Size of 10000 & 30% Missing)

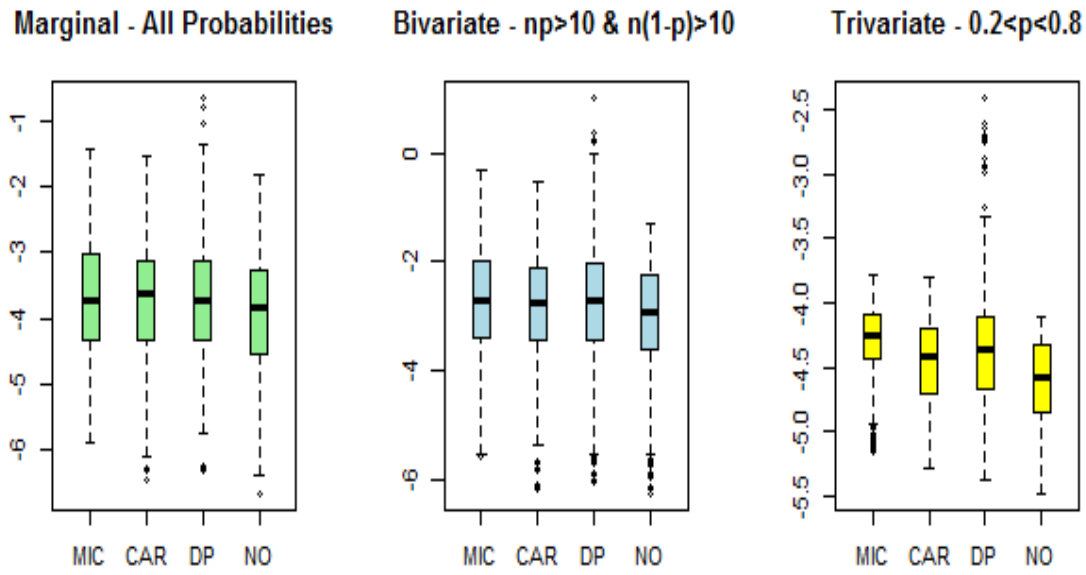

FIGURE B.7: Log Normalized Bias For All Three Methods (21 Variables: Sample Size of 10000 & 30% Missing)

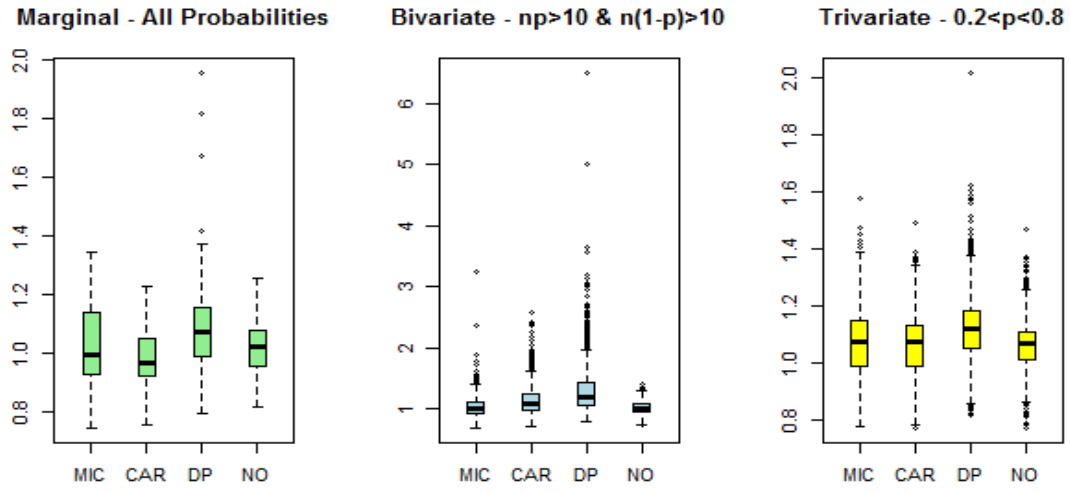

FIGURE B.8: Variance Ratio For All Three Methods (21 Variables: Sample Size of 10000 & 30% Missing)

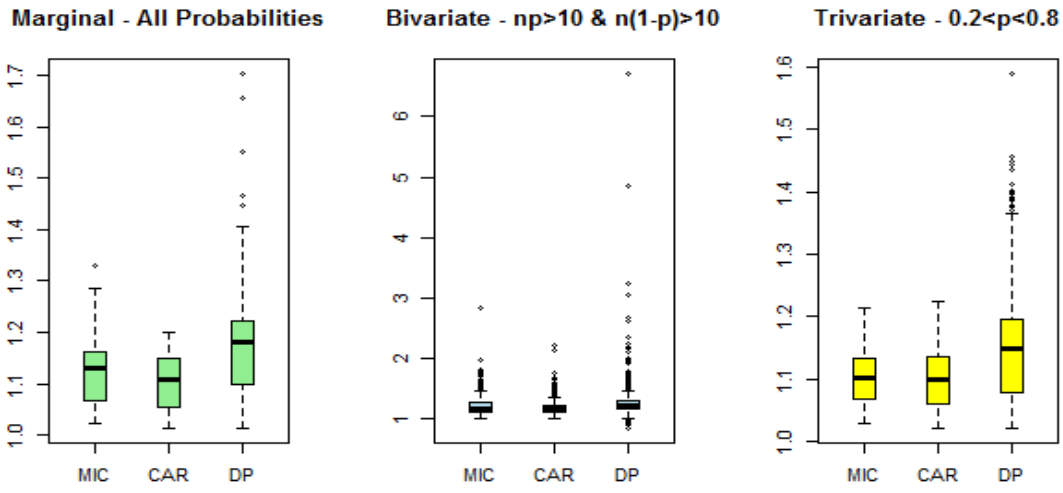

FIGURE B.9: Interval Length Ratio For All Three Methods (21 Variables: Sample Size of 10000 & 30% Missing)

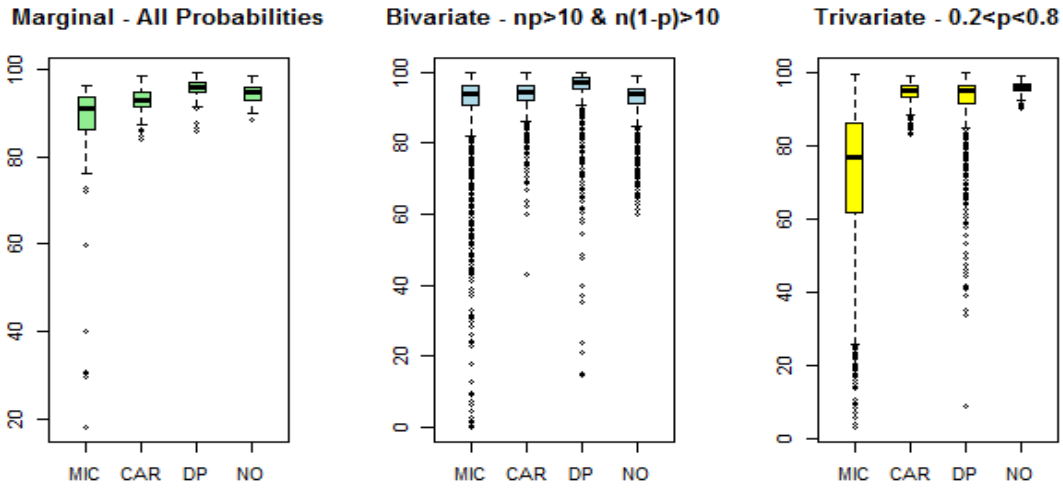

FIGURE B.10: Coverage Rate For All Three Methods (21 Variables: Sample Size of 1000 & 30% Missing)

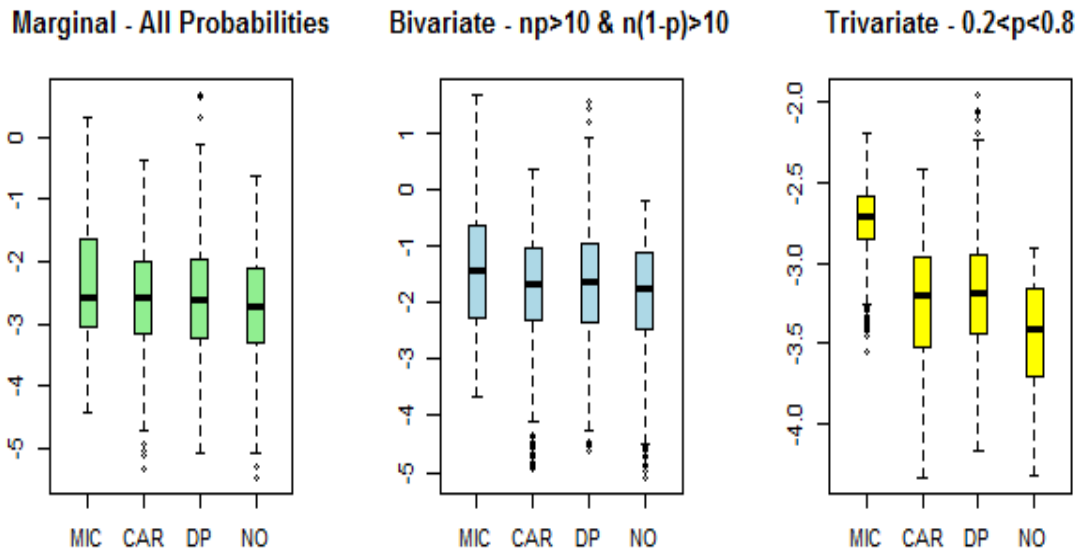

FIGURE B.11: Log Normalized Bias For All Three Methods (21 Variables: Sample Size of 1000 & 30% Missing)

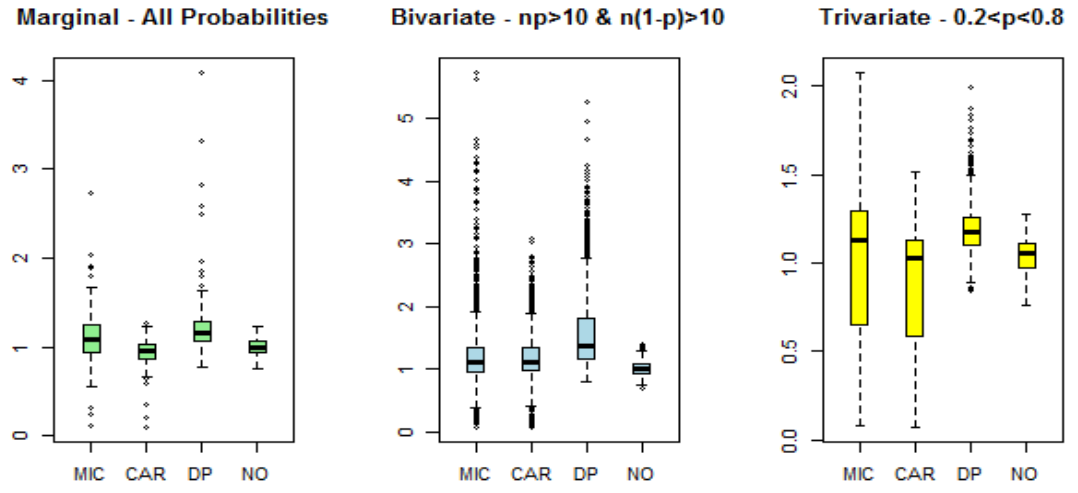

FIGURE B.12: Variance Ratio For All Three Methods (21 Variables: Sample Size of 1000 & 30% Missing)

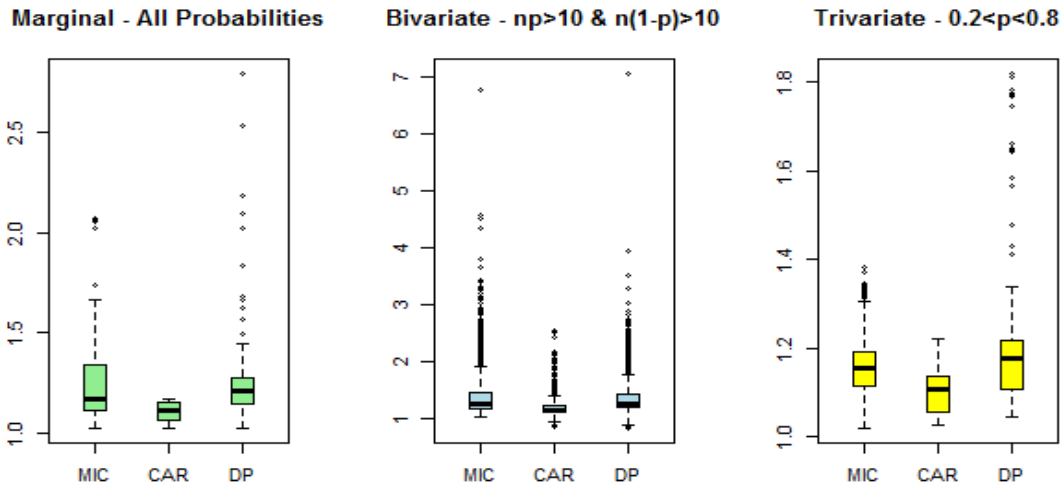

FIGURE B.13: Interval Length Ratio For All Three Methods (21 Variables: Sample Size of 1000 & 30% Missing)

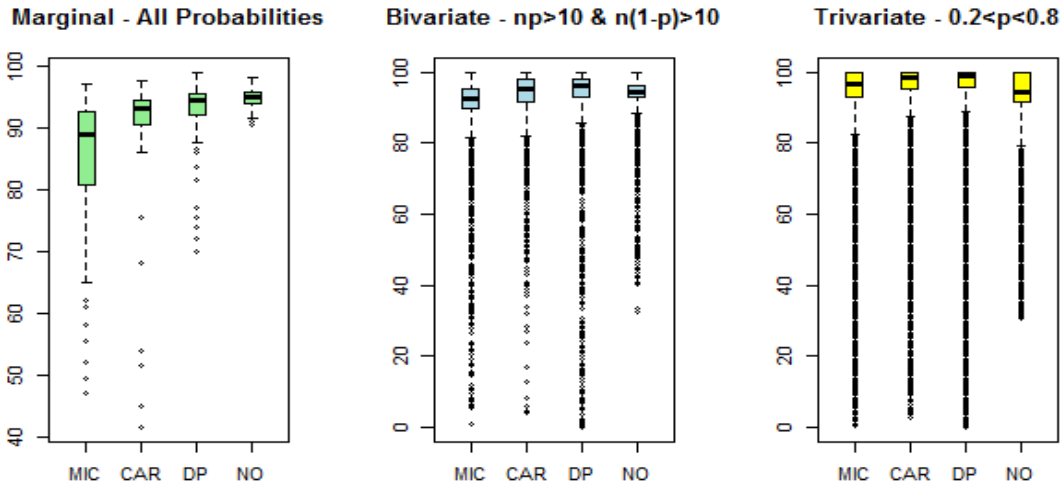

FIGURE B.14: Coverage Rate For All Three Methods (21 Variables: Sample Size of 10000 & 45% Missing)

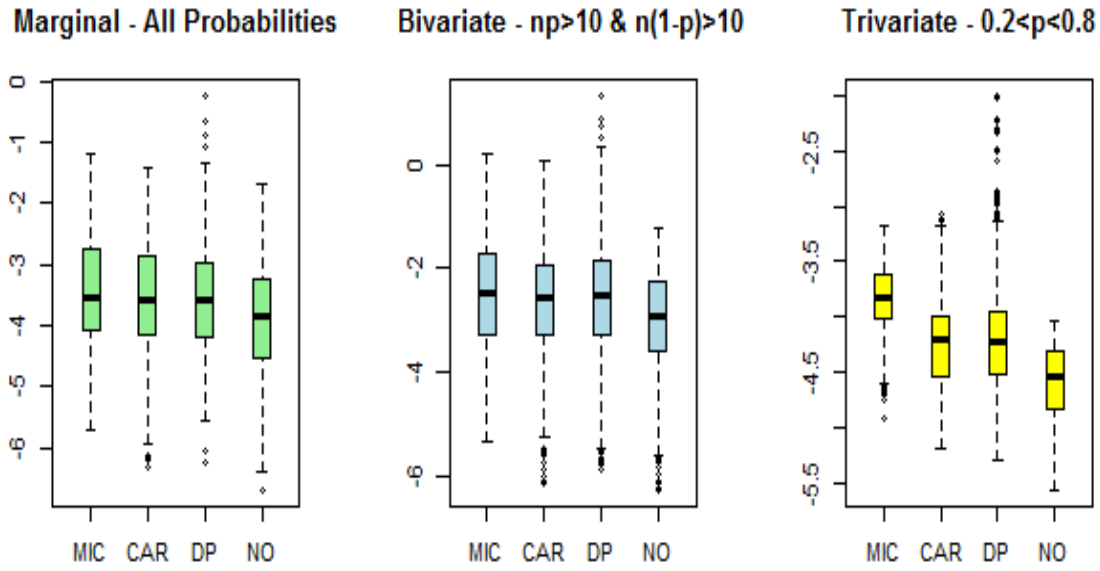

FIGURE B.15: Log Normalized Bias For All Three Methods (21 Variables: Sample Size of 10000 & 45% Missing)

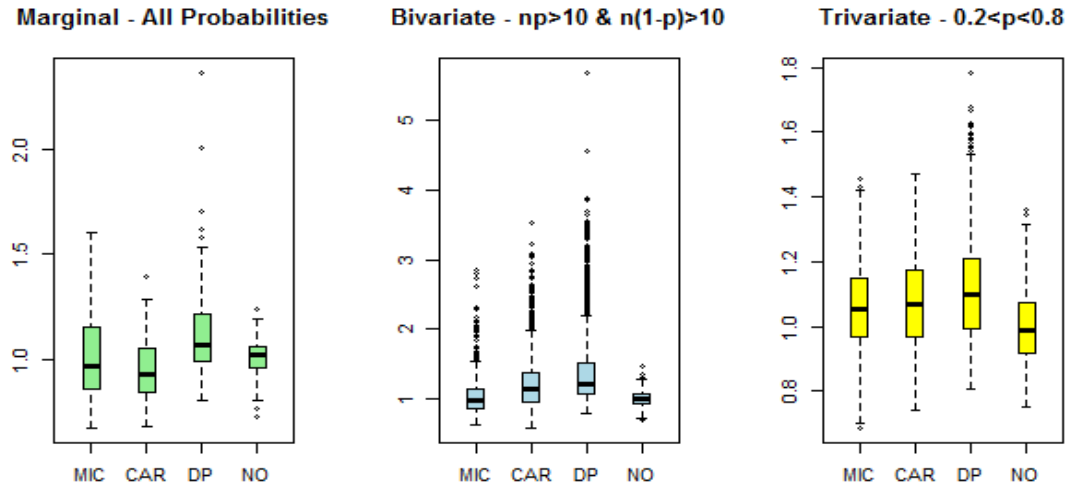

FIGURE B.16: Variance Ratio For All Three Methods (21 Variables: Sample Size of 10000 & 45% Missing)

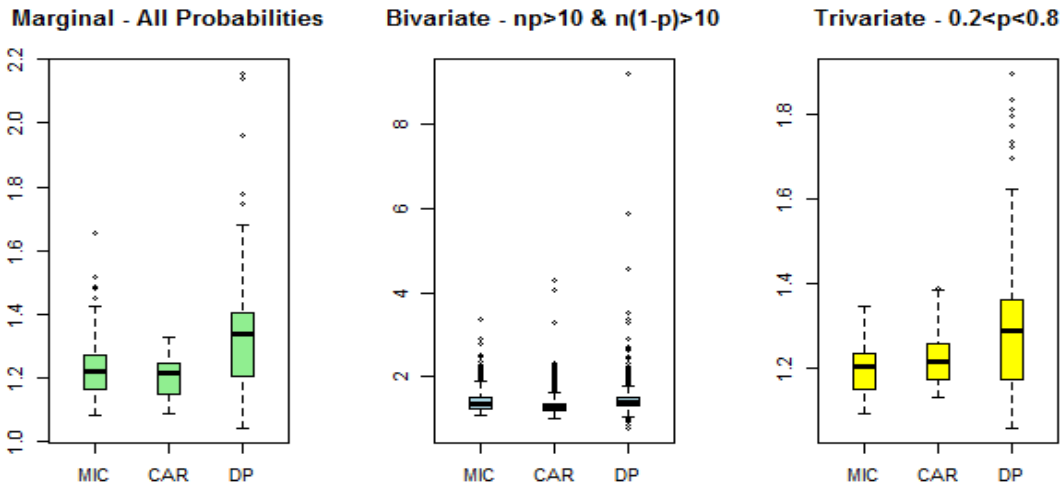

FIGURE B.17: Interval Length Ratio For All Three Methods (21 Variables: Sample Size of 10000 & 45% Missing)

Bibliography

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

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

Brand, J. P. L. (1999), *Development, implementation and evaluation of multiple imputation strategies for the statistical analysis of incomplete data sets*, Erasmus University.

Breiman, L., Friedman, J. H., Olshen, R. A., and Stone, C. I. (1984), *Classification and Regression Trees*, Chapman and Hall/CRC.

Burgette, L. F. and Reiter, J. P. (2000), "Multiple imputation for missing data via sequential regression trees," *American Journal of Epidemiology*, 172, 1070–1076.

Buuren, S. V. (2007), "Multiple imputation of discrete and continuous data by fully conditional specification," *Statistical Methods in Medical Research*, 16, 219–242.

Buuren, S. V. and Groothuis-Oudshoorn, K. (2011), "mice: Multivariate imputation by chained equations in R," *Journal of Statistical Software*, 45, 1–67.

Buuren, S. V. and Groothuis-Oudshoorn, K. (2014), "mice: Multivariate imputation by chained equations," *The Comprehensive R Archive Network*.

Buuren, S. V., Boshuizen, H. C., and Knook, D. L. (1999), "Multiple imputation of missing blood pressure covariates in survival analysis," *Statistics in Medicine*, 18, 681–694.

Buuren, S. V., Brand, J. P. L., Groothius-Oudshoorn, C. G. M., and Rubin, D. B. (2006), "Fully conditional specifications in multivariate imputation," *Journal of Statistical Computation and Simulation*, 76, 1049–1064.

Donneau, A. F., Mauer, M., Molenberghs, G., and Albert, A. (2015), "A simulation study comparing multiple imputation methods for incomplete longitudinal ordinal data," *Communications in Statistics: Simulation & Computation*, 44, 1311–1338.

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, 185–188.

Hardt, J., Herke, M., Brian, T., and Laubach, W. (2013), "Multiple Imputation of Missing Data: A Simulation Study on a Binary Response," *Open Journal of Statistics*, 3, 370–378.

Harel, O. and Zhou, X. H. (2007), "Multiple imputation: review of theory, implementation and software," *Statistics in Medicine*, 26, 3057–3077.

Hopke, P. K., Liu, C., and Rubin, D. B. (2001), "Multiple imputation for multivariate data with missing and below-threshold measurements: time-series concentrations of pollutants in the Arctic," *Biometrics*, 57, 22–33.

Inc., S. I. (2008), *SAS 9.2 Reference*, SAS Institute Inc., Cary, NC.

Kropko, J., Goodrich, B., Gelman, A., and Hill, J. (2014), "Multiple imputation for continuous and categorical data: Comparing joint multivariate normal and conditional approaches," *Political Analysis*, 22, 497–519.

Lee, K. J. and Carlin, J. B. (2010), "Multiple Imputation for Missing Data: Fully Conditional Specification Versus Multivariate Normal Imputation," *American Journal of Epidemiology*, 171, 624–632.

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

Li, F., Baccini, M., Mealli, F., Zell, E. Z., Frangakis, C. E., and Rubin, D. B. (2014), "Multiple imputation by ordered monotone blocks with application to the Anthrax Vaccine Adsorbed Trial," *Journal of Computational and Graphical Statistics*, 23, 877–892.

Little, R. J. A. and Rubin, D. B. (2002), *Statistical analysis with missing data*, Wiley, New Jersey.

Ltd., S. S. (2001), *SOLAS for missing data analysis*, Statistical Solutions, Cork, Ireland.

Manrique-Vallier, D. and Reiter, J. P. (2012), "Bayesian estimation of discrete multivariate truncated latent structure models," *Journal of the American Statistical Association*, 107, 1385–1394.

Manrique-Vallier, D. and Reiter, J. P. (2013), "Bayesian multiple imputation for large-scale categorical data with structural zeros," *Survey Methodology*, 40, 125–134.

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

Raghunathan, T. E. and Paulin, G. S. (1998), "Multiple imputation of income in the Consumer Expenditure Survey: Evaluation of statistical inference," *Proceedings of the Section on Business and Economic Statistics of the American Statistical Association*, pp. 1–10.

Raghunathan, T. E., Lepkowski, J. M., van Hoewyk, J., and Solenberger, P. (2001), "A multivariate technique for multiply imputing missing values using a sequence of regression modelss," *Survey Methodology*, 27, 85–96.

Raghunathan, T. E., Solenberger, P. W., and Hoewyk, V. J. (2002), "IVEware: imputation and variance estimation software user guide." *Survey Research Center, Institute for Social Research, University of Michigan*.

Rasbash, J., Charlton, C., Browne, W. J., Healy, M., and Cameron, B. (2009), "MLwiN Version 2.1," *Centre for Multilevel Modeling, University of Bristol*.

Reiter, J. P. and Raghunathan, T. E. (2007), "The multiple adaptations of multiple imputation," *Journal of the American Statistical Association*, 102, 1462–1471.

Reiter, J. P., Raghunathan, T. E., and Kinney, S. (2006), "The importance of modeling the sampling design in multiple imputation for missing data," *Survey Methodology*, 32, 143–150.

Royston, P. (2004), "Multiple imputation of missing values," *The Stata Journal*, 3, 227–241.

Royston, P. and White, I. R. (2011), "Multiple Imputation by Chained Equations (MICE): Implementation in Stata," *Journal of Statistical Software*, 45, 1–20.

Rubin, D. B. (1976), "Inference and missing data (with discussion)," *Biometrika*, 63, 581–592.

Rubin, D. B. (1987), *Multiple imputation for nonresponse in surveys*, John Wiley & Sons, New York.

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

Rubin, D. B. (2003), "Nested multiple imputation of NMES via partially incompatible MCMC," *Statistica Neerlandica*, 57, 3–18.

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

Schafer, J. L. (2012), "Pan: multiple imputation for multivariate panel or clustered data," *The Comprehensive R Archive Network*.

Schafer, J. L., Ezzati-Rice, T. M., Johnson, W., K. M., Little, R. J. A., and Rubin, D. B. (1998), "The NHANES III multiple imputation project," *Proceedings of the survey Research Methods Section of the American Statistical Association*, pp. 28–37.

Schenker, N., Raghunathan, T. E., C. P. L., Makuc, D. M., Zhang, G., and Cohen, A. J. (2006), "Multiple imputation of missing income data in the National Health Interview Survey," *Journal of the American Statistical Association*, 101, 924–933.

Shen, Z. (2000), "Nested multiple imputation," *Ph.D. dissertation. Harvard University, Department of Statistics*.

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, 199–521.

Su, Y. S., Gelman, A., Hill, J., and Yajima, M. (2011), "Multiple imputation with diagnostics (mi) in R: Opening Windows into the black box," *Journal of Statistical Software*, 45.

Tang, L., Song, J., Belin, T. R., and Unuetzer, J. (2005), "A comparison of imputation methods in a longitudinal randomized clinical trial," *Statistics in Medicine*, 24, 2111–2128.

Yucel, R. M. and Zaslavsky, A. M. (2005), "Imputation of binary treatment variables with measurement error in administrative data," *Journal of American Statistical Association*, 100, 1123–1132.

Zhou, Y., Little, R. J. A., and Kalbfleisch, J. D. (2010), "Block-conditional missing at random models for missing data," *Statistical Science*, 25, 517–532.