DATA SCIENCE JOURNAL

Takahashi, M 2017 Statistical Inference in Missing Data by MCMC and Non-MCMC Multiple Imputation Algorithms: Assessing the Effects of Between-Imputation Iterations. *Data Science Journal,* 16: 37, pp. 1–17, DOI: https://doi.org/10.5334/dsj-2017-037

RESEARCH PAPER

Statistical Inference in Missing Data by MCMC and Non-MCMC Multiple Imputation Algorithms: Assessing the Effects of Between-Imputation Iterations

Masayoshi Takahashi
IR Office, Tokyo University of Foreign Studies, Tokyo, JP
mtakahashi@tufs.ac.jp

Incomplete data are ubiquitous in social sciences; as a consequence, available data are inefficient (ineffective) and often biased. In the literature, multiple imputation is known to be the standard method to handle missing data. While the theory of multiple imputation has been known for decades, the implementation is difficult due to the complicated nature of random draws from the posterior distribution. Thus, there are several computational algorithms in software: Data Augmentation (DA), Fully Conditional Specification (FCS), and Expectation-Maximization with Bootstrapping (EMB). Although the literature is full of comparisons between joint modeling (DA, EMB) and conditional modeling (FCS), little is known about the relative superiority between the MCMC algorithms (DA, FCS) and the non-MCMC algorithm (EMB), where MCMC stands for Markov chain Monte Carlo. Based on simulation experiments, the current study contends that EMB is a confidence proper (confidence-supporting) multiple imputation algorithm without between-imputation iterations; thus, EMB is more user-friendly than DA and FCS.

Keywords: MCMC; Markov chain Monte Carlo; Incomplete data; Nonresponse; Joint modeling; Conditional modeling

1 Introduction

Generally, it is quite difficult to obtain complete data in social surveys (King et al. 2001: 49). Consequently, available data are not only inefficient due to the reduced sample size, but also biased due to the difference between respondents and non-respondents, thus making statistical inference invalid. Since Rubin (1987), multiple imputation has been known to be the standard method of handling missing data (Graham 2009; Baraldi and Enders 2010; Carpenter and Kenward 2013; Raghunathan 2016).

While the theoretical concept of multiple imputation has been around for decades, the implementation is difficult because making a random draw from the posterior distribution is a complicated matter. Accordingly, there are several computational algorithms in software (Schafer 1997; Honaker and King 2010; van Buuren 2012). The most traditional algorithm is Data Augmentation (DA) followed by the other two new algorithms, Fully Conditional Specification (FCS) and Expectation-Maximization with Bootstrapping (EMB). Although an abundant literature exists on the comparisons between joint modeling (DA, EMB) and conditional modeling (FCS), no comparisons have been made about the relative superiority between the MCMC algorithms (DA, FCS) and the non-MCMC algorithm (EMB), where MCMC stands for Markov chain Monte Carlo. This study assesses the effects of between-imputation iterations on the performance of the three multiple imputation algorithms, using Monte Carlo experiments.

By way of organization, Section 2 introduces the notations in this article. Section 3 gives a motivating example of missing data analysis in social sciences. Section 4 presents the assumptions of imputation methods.

Section 5 shows the traditional methods of handling missing data. Section 6 introduces the three multiple imputation algorithms. Section 7 surveys the literature on multiple imputation. Sections 8 gives the results of the Monte Carlo experiments, showing the impact of between-imputation iterations on multiple imputation. Section 9 concludes with the findings and the limitations in the current research.

2 Notations

D is $n \times p$ data, where n is the sample size and p is the number of variables. The distribution of D is multivariate-normal with mean vector μ and variance-covariance matrix Σ, i.e., $D \sim N_p(\mu, \Sigma)$, where all of the variables are continuous. Let i refer to an observation index ($i = 1, \ldots, n$). Let j refer to a variable index ($j = 1, \ldots, p$). Let $D = \{Y_1, \ldots, Y_p\}$, where Y_j is the j-th column in D and Y_{-j} is the complement of Y_j, i.e., all columns in D except Y_j. Also, let Y_{obs} be observed data and Y_{mis} be missing data: $D = \{Y_{obs}, Y_{mis}\}$.

At the imputation stage, there is no concept of the dependent and independent variables, because imputation is not a causal model, but a predictive model (King et al. 2001: 51). Therefore, all of the variables are denoted Y_j with the subscript j indexing a variable number. However, at the analysis stage, one of the Y_j variables is the dependent variable and the remaining Y_{-j} are the independent variables. If the dependent variable is the p-th column in D, then the dependent variable is simply denoted Y and the independent variables are denoted $X_1, \ldots, X_{p-1}$.

Let R be a response indicator matrix that has the same dimension as D. Whenever D is observed, $R = 1$; otherwise, $R = 0$. Note, however, that non-italicized R refers to the R statistical environment. In the multiple imputation context, M refers to the number of imputations and T refers to the number of between-imputation iterations. In general, θ is an unknown parameter vector.

3 Motivating Example: Missing Economic Data

Social scientists have long debated the determinants of economic development across countries (Barro 1997; Feng 2003; Acemoglu, Johnson, and Robinson 2005). Using the data from the Central Intelligence Agency (CIA 2016) and Freedom House (2016), we may estimate a multiple regression model, in which the dependent variable is GDP per capita and the independent variables include social, economic, and political variables. The problem is that the data are incomplete (**Table 1**), where the median missing rate is 22.4% and the total missing rate is 62.3%.

Table 1: Variables and Missing Rates.

Variables	Missing Rates
GDP per capita (purchasing power parity)	0.0%
Freedom House index	15.4%
Central bank discount rate	32.9%
Life expectancy at birth	2.6%
Unemployment rate	10.5%
Distribution of family income: Gini index	37.3%
Public debt	22.4%
Education expenditures	24.6%
Taxes and other revenues	6.1%
Military expenditures	43.0%

Data sources: CIA (2016) and Freedom House (2016).

Table 2 presents multiple regression models; however, the conclusions are susceptible to how we deal with missing data. The coefficients for central bank and public debt are statistically significant at the 5% error level using incomplete data, while they are not significant using multiply-imputed data. On the other hand, the coefficients for education and military are not significant using incomplete data, while they are significant using multiply-imputed data. Therefore, the issue of missing data is of grave concern in applied empirical research.

Table 2: Multiple Regression Analyses on GDP Per Capita.

	Incomplete Data		Multiply-Imputed Data	
Variables	**Coef.**	**Std. Err.**	**Coef.**	**Std. Err.**
Intercept	–7.323	3.953	–11.545*	3.495
Freedom	–0.321*	0.127	–0.362*	0.127
Central Bank	**–0.118***	**0.041**	–0.107	0.049
Life Expectancy	3.922*	0.794	4.908*	0.655
Unemployment	–0.205*	0.087	–0.214*	0.070
Gini	0.114	0.253	–0.018	0.363
Public Debt	**–0.198***	**0.092**	–0.002	0.093
Education	0.035	0.164	**–0.488***	**0.154**
Tax	0.357*	0.174	0.471*	0.151
Military	0.123	0.085	**0.299***	**0.109**
Number of obs.		86		228

Note: *significant at the 5% error level. Coef. stands for coefficient. Std. Err. stands for standard error. Since the distributions of these variables are skewed to the right (log-normal), the variables are log-transformed to normalize the distributions.

4 Assumptions of Imputation Methods

Missing data analyses always involve assumptions (Raghunathan 2016: 12). In order to judge the appropriateness of missing data methods, it is vital to comprehend the assumptions for the methods. Imputation involves the following four assumptions. These assumptions will play important roles in simulation studies (Section 8).

4.1 Assumptions of Missing Data Mechanisms

There are three common assumptions of missing data mechanisms in the literature (King et al. 2001: 50–51; Little and Rubin 2002; Carpenter and Kenward 2013: 10–21). The first assumption is Missing Completely At Random (MCAR), which is $Pr(R|D) = Pr(R)$. If respondents are selected to answer their income values by throwing dice, this is an example of MCAR. The second assumption is Missing At Random (MAR), which is $Pr(R|D) = Pr(R|Y_{obs})$. If older respondents are more likely to refuse to answer their income values and if the ages of the respondents are available in the data, this is an example of MAR. The third assumption is Not Missing At Random (NMAR), which is $Pr(R|D) \neq Pr(R|Y_{obs})$. If respondents with higher values of incomes are more likely to refuse to answer their income values and if the other variables in the data cannot be used to predict which respondents have high amounts of income, this is an example of NMAR.

4.2 Assumption of Ignorability

To be strict, the missing data mechanism is ignorable if both of the following conditions are satisfied: (1) The MAR condition; and (2) the distinctness condition, which stipulates that the parameters in the missing data mechanism are independent of the parameters in the data model (Schafer 1997: 11).

However, the MAR condition is said to be more relevant in real data applications (Allison 2002: 5; van Buuren 2012: 33). Thus, for all practical purposes, NMAR is Non-Ignorable (NI). The current study assumes that the missing data mechanism is MAR and thus ignorable.

4.3 Assumption of Proper Imputation

Imputation is said to be Bayesianly proper if imputed values are independent realizations of $Pr(Y_{mis}|Y_{obs})$, which means that successive iterates of Y_{mis} cannot be used because of the correlations between them (Schafer 1997: 105–106). Between-imputation convergence relies on a number of factors, but the fractions of missing information are one of the most influential factors (Schafer 1997: 84; van Buuren 2012: 113).

van Buuren (2012: 39) introduces a slightly simplified version of proper imputation, which he calls confidence proper. Let $\bar{\theta}$ be the multiple imputation estimate, $\hat{\theta}$ be the estimate based on the hypothetically complete data, $\bar{V}$ be the estimate of the sampling variance of the estimate based on the hypothetically complete data, and $\hat{V}$ be the sampling variance estimate based on the hypothetically complete data. An imputation procedure is said to be confidence proper if all of the following three conditions are satisfied: (1) $\bar{\theta}$ is equal to $\hat{\theta}$ when averaged over the response indicators sampled under the assumed response model; (2) $\bar{V}$ is equal to $\hat{V}$ when averaged over the response indicators sampled under the assumed response model; and (3) the extra inferential uncertainty due to missingness is correctly reflected. In order to check whether an imputation method is confidence proper, van Buuren (2012: 47) recommends to use bias, coverage, and confidence interval length as the evaluation criteria (See Section 8.2).

4.4 Assumption of Congeniality

Congeniality means that the imputation model is equal to the substantive analysis model. It is widely known that the imputation model can be larger than the substantive analysis model, but the imputation model cannot be smaller than the substantive analysis model (Enders 2010: 227–229; Carpenter and Kenward 2013: 64–65; Raghunathan 2016: 175–177).

5 Traditional Methods of Handling Missing Data

This section introduces listwise deletion, deterministic single imputation, and stochastic single imputation, which are used as baseline methods for comparisons in Section 8.

Listwise deletion (LD), also known as complete-case analysis, throws away any rows that have at least one missing value (Allison 2002: 6–8; Baraldi and Enders 2010: 10). Although it is simple and convenient, LD is less efficient due to the reduced sample size and may be biased if the assumption of MCAR does not hold (Schafer 1997: 23).

Deterministic single imputation (D-SI) replaces a missing value with a reasonable guess. The most straightforward version calculates predicted scores for missing values based on a regression model (Allison 2002: 11; Baraldi and Enders 2010: 12). If the goal of analysis is to estimate the mean of an incomplete variable, D-SI produces an unbiased estimate under the assumptions of MCAR and MAR. However, D-SI tends to underestimate the variation in imputed data (de Waal, Pannekoek, and Scholtus 2011: 231). D-SI is available as R-function `norm.predict` in MICE (van Buuren 2012: 57), where MICE stands for Multivariate Imputation by Chained Equations.

Stochastic single imputation (S-SI) also utilizes a regression model to predict missing values, but it adds to imputed values random components drawn from the residual distribution (Baraldi and Enders 2010: 13). S-SI is likely to recover the variation of an incomplete variable under the assumptions of MCAR and MAR; thus, compensating for the disadvantage of D-SI (de Waal, Pannekoek, and Scholtus 2011: 231). S-SI is available as R-function `norm.nob` in MICE (van Buuren 2012: 57).

However, both D-SI and S-SI tend to underestimate the standard error in imputed data because imputed values are treated as if they were real (Raghunathan 2016: 77).

6 Competing Multiple Imputation Algorithms

Multiple imputation was made widely known by Rubin (1987) and concise history can be found in Scheuren (2005). In theory, multiple imputation replaces a missing value by M simulated values ($M > 1$) independently and randomly drawn from the distribution of missing data. The variation among M simulated values reflects uncertainty about missing data; thus, making the standard error valid. In practice, missing data are by definition unobserved; therefore, the distribution of missing data is also unobserved. Instead, under the assumption of MAR (or MCAR), multiple imputation constructs the posterior predictive distribution of missing data, conditional on observed data. Then, a random draw is independently made from this posterior distribution (Rubin 1987: 75; King et al. 2001: 53–54; Carpenter and Kenward 2013: 38–39).

However, using the analytical methods, it is not easy to randomly draw sufficient statistics from the posterior distribution (Allison 2002: 33; Honaker and King 2010: 564). In order to solve this problem, three computational algorithms have been proposed in the literature.

6.1 Data Augmentation

The traditional algorithm of multiple imputation is the Data Augmentation (DA) algorithm, which is a Markov chain Monte Carlo (MCMC) technique (Takahashi and Ito 2014: 46–48). DA improves parameter estimates by repeated substitution conditional on the preceding value, forming a stochastic process called a Markov chain (Gill 2008: 379).

The DA algorithm works as follows (Schafer 1997: 72). Equation (1) is the imputation step that generates imputed values from the predictive distribution of missing values, given the observed values and the parameter values at iteration t. Equation (2) is the posterior step that generates parameter values from the posterior distribution, given the observed values and the imputed values at iteration $t+1$.

$$Y_{mis}^{(t+1)} \sim Pr\left(Y_{mis} \mid Y_{obs}, \theta^{(t)}\right) \tag{1}$$

$$\theta^{(t+1)} \sim Pr\left(\theta \mid Y_{obs}, Y_{mis}^{(t+1)}\right) \tag{2}$$

These two steps are repeated T times until convergence is attained. The convergence of MCMC is stochastic because it converges to probability distributions (Schafer 1997: 80). Therefore, it is hard to judge the convergence in MCMC.

There are two ways of generating multiple imputations by DA (Schafer 1997: 139; Enders 2010: 211–212). In the first method, a single chain is run for $M \times T$ iterations, taking every t-th iteration of Y_{mis}. In the second method, M parallel chains of length T are run, and the final values of Y_{mis} from M chains are taken as the imputations. The current study adopts the second method.

The software using this algorithm is R-Package NORM2, which was originally developed by Schafer (1997) and is currently maintained by Schafer (2016).

6.2 Fully Conditional Specification

An alternative algorithm to DA is the Fully Conditional Specification (FCS) algorithm, which specifies the multivariate distribution by way of a series of conditional densities, through which missing values are imputed given the other variables (Takahashi and Ito 2014: 50–53).

The FCS algorithm works as follows (van Buuren and Groothuis-Oudshoorn 2011: 6–7; van Buuren 2012: 110; Zhu and Raghunathan 2015). Equation (3) draws the unknown parameters of the imputation model, given the observed values and the t-th imputations, where $\tilde{Y}_{-j}^{(t)} = \left(\tilde{Y}_1^{(t)}, \ldots, \tilde{Y}_{j-1}^{(t)}, \tilde{Y}_{j+1}^{(t-1)}, \ldots, \tilde{Y}_p^{(t-1)}\right)$, where tilde denotes a random draw. Equation (4) draws imputations, given the observed values, the t-th imputations, and the t-th parameter estimates. These two steps are repeated for $j = 1, \ldots, p$.

$$\tilde{\theta}_j^{(t)} \sim Pr\left(\theta_j^{(t)} \mid Y_{j,obs}, \tilde{Y}_{-j}^{(t)}\right) \tag{3}$$

$$\tilde{Y}_j^{(t)} \sim Pr\left(Y_{j,mis} \mid Y_{j,obs}, \tilde{Y}_{-j}^{(t)}, \tilde{\theta}_j^{(t)}\right) \tag{4}$$

The entire process is repeated for $t = 1, \ldots, T$ until convergence is attained. FCS can be considered an MCMC method, because FCS is a Gibbs sampler under the compatible conditionals (van Buuren and Groothuis-Oudshoorn 2011: 6; van Buuren 2012: 109). This means that the convergence of FCS is stochastic. Therefore, it is hard to judge the convergence in FCS.

The software using this algorithm is R-Package MICE (van Buuren and Groothuis-Oudshoorn 2011), which stands for Multivariate Imputation by Chained Equations and is currently maintained by van Buuren et al. (2015). The FCS algorithm is also known as Sequential Regression Multivariate Imputation (Raghunathan 2016: 76).

6.3 Expectation-Maximization with Bootstrapping

Another emerging algorithm is the Expectation-Maximization with Bootstrapping (EMB) algorithm, which combines the Expectation-Maximization (EM) algorithm with the nonparametric bootstrap to create multiple imputation (Takahashi and Ito 2014: 55–57).

The EMB algorithm works as follows (Honaker and King 2010: 565; Honaker, King, and Blackwell 2011: 4). Suppose that a random sample of size n is drawn from a population, where some values are missing in the sample. Bootstrap resamples of size n are randomly drawn from the sample data with replacement M times (Horowitz 2001: 3163–3165; Carsey and Harden 2014: 215). The variation among the M resamples represents uncertainty about estimation. The EM algorithm is applied to each of these M bootstrap resamples to refine M point estimates of parameter θ. Equation (5) is the expectation step that calculates the Q-function by averaging the complete-data log-likelihood over the predictive distribution of missing data. Equation (6) is the maximization step that finds parameter values at iteration $t+1$ by maximizing the Q-function.

$$Q\left(\theta \mid \theta^{(t)}\right) = \int l\left(\theta \mid Y\right) Pr\left(Y_{mis} \mid Y_{obs}, \theta^{(t)}\right) dY_{mis} \tag{5}$$

$$\theta^{(t+1)} = \arg\max_{\theta} Q\left(\theta \mid \theta^{(t)}\right) \tag{6}$$

These two steps are repeated until convergence is attained, where the converged value is a Maximum Likelihood Estimate (MLE) under well-behaved conditions (Schafer 1997: 38–39; Do and Batzoglou 2008). The convergence of EM is deterministic because it converges to a point in the parameter space (Schafer 1997: 80). Therefore, it is straightforward to judge the convergence in EM. The substitution of MLEs from bootstrap resamples is asymptotically equal to a sample from the posterior distribution (Little and Rubin 2002: 216–217).

The software using this algorithm is R-Package AMELIA II (Honaker, King, and Blackwell 2011), which was originally developed by King et al. (2001) and is currently maintained by Honaker, King, and Blackwell (2016).

6.4 Relationships among the Three Algorithms

The three algorithms share certain characteristics with each other, but not exactly the same as summarized in **Table 3**.

Table 3: Relations among DA, EMB, and FCS.

	Joint Modeling	Conditional Modeling
MCMC	DA	FCS
Non-MCMC	EMB	

DA and EMB are joint modeling while FCS is conditional modeling (Kropko et al. 2014). Joint modeling specifies a multivariate distribution of missing data while conditional modeling specifies a univariate distribution on a variable-by-variable basis (van Buuren 2012: 105–108). Conditional modeling is more flexible and joint modeling is computationally more efficient (van Buuren 2012: 117; Kropko et al. 2014).

DA and FCS are different versions of MCMC techniques. On the other hand, EMB is not an MCMC technique. It is said that DA and FCS require between-imputation iterations to be confidence proper (Schafer 1997: 106; van Buuren 2012: 113) while EMB does not need iterations to be confidence proper (Honaker and King 2010: 565). However, as is clear in Section 7, whether EMB is confidence proper when DA and FCS are improper, this is an open question that has not been tested in the literature.

7 Comparative Studies on Multiple Imputation in the Literature

Table 4 presents the literature that compared imputation methods. Nine studies compared multiple imputation with other missing data methods, such as listwise deletion, single imputation, and maximum likelihood. Among these nine studies, four studies focused on DA (Schafer and Graham 2002; Abe and Iwasaki 2007; Lee and Carlin 2012; von Hippel 2016), four studies on FCS (Donders et al. 2006; Stuart et al. 2009; Cheema 2014; Deng et al. 2016), and one study on an unknown algorithm (Shara et al. 2015).

Four studies investigated specialized situations for multiple imputation, such as small-sample degrees of freedom in DA (Barnard and Rubin 1999), Likert-scale data in DA (Leite and Beretvas 2010), non-parametric multiple imputation (Cranmer and Gill 2013), and variance estimators (Hughes, Sterne, and Tilling 2016).

Seven studies compared different multiple imputation algorithms (King et al. 2001; Horton and Lipsitz 2002; Horton and Kleinman 2007; Lee and Carlin 2010; Hardt, Herke, and Leonhart 2012; Kropko et al. 2014; McNeish 2017). The comparative perspective in most of the seven studies, except King et al. (2001), is based on the difference between joint modeling and conditional modeling. Thus, the perspective from MCMC vs. non-MCMC is generally lacking in the literature.

Ten studies did not explicitly state the number of iterations *T*. Furthermore, Horton and Kleinman (2007) used the default setting in software for *T*, and the information in Kropko et al. (2014) can be only found in their computer codes, not in the article.

Thus, no studies in **Table 4** have systematically investigated the effects of convergence on the three multiple imputation algorithms.

8 Monte Carlo Simulation

Section 4 introduced MAR, proper imputation, and congeniality as crucial assumptions. To make the assumptions of MAR and congeniality realistic, an inclusive analysis strategy is recommended in the literature (Enders 2010: 16–17; Raghunathan 2016: 73), which contains any auxiliary variables that can increase the predictive power of the imputation model or any variables that may be related to the missing data mechanism. What complicates the matter, however, is that auxiliary variables themselves are often incomplete. This creates a dilemma in multiple imputation. Including many auxiliary variables makes it more likely for MAR and congeniality to be satisfied, but including many incomplete variables leads to a higher total missing rate, which further makes it more difficult for convergence in MCMC to be attained.

Table 4: Summary of the 20 Studies on Multiple Imputation.

Authors	MI Algorithms	Sample Size	Number of Variables	Number of Imputations	Number of Iterations	Missing Rate
Barnard and Rubin (1999)	DA	10, 20, 30	2	3, 5, 10	Unknown	10%, 20%, 30%
Horton and Lipsitz (2001)	DA, FCS	10000	3	10	200	50%
Schafer and Graham (2002)	DA	50	2	20	Unknown	73%
Donders et al. (2006)	FCS	500	2	10	Unknown	40%
Abe and Iwasaki (2007)	DA	100	4	5	100	20%, 30%
Horton and Kleinman (2007)	**DA, EMB, FCS**	133774	10	10	5	41%
Stuart et al. (2009)	FCS	9186	400	10	10	18%
Lee and Carlin (2010)	DA, FCS	1000	8	20	10	33%
Leite and Beretvas (2010)	DA	400	10	10	Unknown	10%, 30%, 50%
Hardt, Herke, and Leonhart (2012)	**DA, EMB, FCS**	50, 100, 200	3, 13, 23, 43, 83	20	**Unknown**	20%, 50%
Lee and Carlin (2012)	DA	1000	8	20	Unknown	10%, 25%, 50%, 75%, 90%
Cranmer and Gill (2013)	EMB, MHD	500	5	Unknown	NA	20%, 50%, 80%
Cheema (2014)	FCS	10, 20, 50, 100, 200, 500, 1000, 2000, 5000, 10000	4	Unknown	Unknown	1%, 2%, 5%, 10%, 20%
Kropko et al. (2014)	**DA, EMB, FCS**	1000	8	5	**30**	25%
Shara et al. (2015)	Unknown	2246	8	Unknown	Unknown	20%, 30%, 40%
Deng et al. (2016)	FCS	100	200, 1000	10	20	40%
von Hippel (2016)	DA	25, 100	2	5	Unknown	50%
Hughes, Sterne, and Tilling (2016)	Unknown	100, 1000	5	50	Unknown	40%, 60%
McNeish (2017)	DA, FCS	20, 50, 100, 250	4	5, 25, 100	Unknown	10%, 20%, 30%, 50%

Note: DA stands for Data Augmentation, EMis for Expectation-Maximization with Importance Sampling, FCS for Fully Conditional Specification, EMB for Expectation-Maximization with Bootstrapping, and MHD for Multiple Hot Deck. Unknown means that information is unavailable. NA means Not-Applicable.

When assumptions do not hold in statistical methods, analytical mathematics does not often provide answers about the properties of the methods (Mooney 1997: 1). Monte Carlo simulation converts the computer into an experimental laboratory, where the researcher can control various conditions in the environment to observe the outcomes (Carsey and Harden 2014: 4). Thus, Monte Carlo simulation is a powerful method of assessing the performance of statistical methods under various settings especially when assumptions are violated.

8.1 Monte Carlo Simulation Designs

The current study prepares two versions of simulation data, (1) theoretical and (2) realistic. Auxiliary variables X are generated by R-Function `mvrnorm`. All of the computations are done in R version 3.2.4. The computer used in the current study is HP Z440 Workstation (Windows 7 Professional, processor: Intel Xeon CPU E5-1603 v3), with the processor speed of 2.80 GHz and the memory (RAM) of 32.0 GB under the 64 bit operating system. The number of Monte Carlo simulation runs is set to 1000.

The first setting is theoretical. The number of observations is 1000, which is equivalent to the 75th percentile of the sample sizes found in the studies listed in **Table 4**. The number of variables p is changed from 2, 3, 4, 5, 6, 7, 8, 9, to 10, which is equivalent to the 70th percentile of the number of variables found in the studies listed in **Table 4**. Note that in another simulation run, not reported here, p was changed to 20, and the conclusions were similar. As was assumed in Section 2, auxiliary variables x_j are multivariate-normal with the mean of 0 and the standard deviation of 1, i.e., $X \sim N_{p-1}(0, 1)$, where the number of auxiliary variables is $p - 1$. The correlation among x_j is randomly generated in R as follows: `r<-matrix(runif(9^2,-1,1), ncol=9)` and `Cor<-cov2cor(r%*%t(r))`. The generated correlation matrix is shown in equation (7). The p-th variable y_i is a linear combination of x_j such that $y_i = \beta_0 + \beta_1 x_{1i} + \ldots + \beta_{p-1} x_{p-1i} + \varepsilon_i$, where $\beta_j \sim U(-2.0, 2.0)$ and $\varepsilon_i \sim N(0, \sigma)$. Note that β_j includes β_0 and $\sigma \sim U(0.5, 2.0)$.

$$Cor_1 = \begin{bmatrix} 1.000 & -0.231 & 0.335 & 0.401 & -0.276 & 0.247 & -0.120 & 0.327 & -0.068 \\ -0.231 & 1.000 & 0.074 & -0.761 & 0.041 & -0.623 & -0.083 & -0.432 & -0.183 \\ 0.335 & 0.074 & 1.000 & 0.183 & -0.323 & 0.254 & -0.458 & 0.434 & -0.801 \\ 0.401 & -0.761 & 0.183 & 1.000 & 0.007 & 0.639 & -0.094 & 0.676 & 0.169 \\ -0.276 & 0.041 & -0.323 & 0.007 & 1.000 & -0.547 & 0.357 & -0.025 & 0.081 \\ 0.247 & -0.623 & 0.254 & 0.639 & -0.547 & 1.000 & 0.024 & 0.204 & 0.023 \\ -0.120 & -0.083 & -0.458 & -0.094 & 0.357 & 0.024 & 1.000 & -0.486 & 0.373 \\ 0.327 & -0.432 & 0.434 & 0.676 & -0.025 & 0.204 & -0.486 & 1.000 & -0.153 \\ -0.068 & -0.183 & -0.801 & 0.169 & 0.081 & 0.023 & 0.373 & -0.153 & 1.000 \end{bmatrix} \quad (7)$$

The second setting is realistic. The number of observations is 228, which is the full sample size of the real data in **Table 2**. The number of variables p is again changed from 2, 3, 4, 5, 6, 7, 8, 9, to 10. Auxiliary variables x_j are multivariate-normal with the means and standard deviations based on the empirical data (log-transformed), where x_j consist of the nine independent variables in **Table 2** (CIA 2016; Freedom House 2016). Note that, as was explained in **Table 2**, the raw empirical data are log-normal; therefore, the input data are log-transformed. Furthermore, the correlation matrix is based on the empirical data (log-transformed) as in equation (8). The p-th variable y_i is a linear combination of x_j such that $y_i = \beta_0 + \beta_1 x_{1i} + \ldots + \beta_{p-1} x_{p-1i} + \varepsilon_i$, where β_j (including β_0) reflects the coefficients in multiple regression models using the empirical data and $\varepsilon_i \sim N(0, \sigma_{resid})$, where σ_{resid} is the residual standard deviation from the empirical regression model.

$$Cor_2 = \begin{bmatrix} 1.000 & 0.646 & -0.500 & -0.007 & 0.376 & -0.354 & -0.378 & -0.534 & 0.312 \\ 0.646 & 1.000 & -0.531 & 0.021 & 0.371 & -0.305 & -0.150 & -0.427 & 0.049 \\ -0.500 & -0.531 & 1.000 & -0.474 & -0.512 & 0.278 & 0.092 & 0.280 & -0.086 \\ -0.007 & 0.021 & -0.474 & 1.000 & 0.205 & 0.079 & 0.014 & 0.086 & 0.161 \\ 0.376 & 0.371 & -0.512 & 0.205 & 1.000 & -0.204 & -0.089 & -0.370 & 0.220 \\ -0.354 & -0.305 & 0.278 & 0.079 & -0.204 & 1.000 & 0.106 & 0.212 & -0.180 \\ -0.378 & -0.150 & 0.092 & 0.014 & -0.089 & 0.106 & 1.000 & 0.578 & -0.128 \\ -0.534 & -0.427 & 0.280 & 0.086 & -0.370 & 0.212 & 0.578 & 1.000 & -0.134 \\ 0.312 & 0.049 & -0.086 & 0.161 & 0.220 & -0.180 & -0.128 & -0.134 & 1.000 \end{bmatrix} \quad (8)$$

In both settings, x_j are incomplete variables for imputation, y_i is completely observed in all of the situations, and u_{ij} are a set of $p - 1$ continuous uniform random numbers ranging from 0 to 1 for the missing data mechanism. As was introduced in Section 4.1, under the assumption of MAR, the missingness of x_{ji} depends on the values of y_i and u_{ij}, i.e., x_{ji} is missing if $y_i <$ median(y_i) and $u_{ij} < 0.5$, and x_{ji} is missing if $y_i >$ median(y_i) and $u_{ij} > 0.9$. This creates approximately 30% missing values in each x_j. This is realistic, because the average missing rates of income and earnings are 30% on a variable basis in the National Health Interview Survey (Schenker et al. 2006: 925) and the median missing rate is 30.0% in **Table 4**. Note that the above setting may be translated into the following statement. Variable y_i is age and x_{1i} is income. The missingness of income depends on age and some random components. Income is missing if age is less than the median of age and uniform random numbers are less than 0.5. Also, income is missing if age is larger than the median of age and uniform random numbers are larger than 0.9.

Although the literature (Graham, Olchowski, and Gilreath 2007; Bodner 2008; Takahashi and Ito 2014: 68–71) recommends to use relatively large M, the simulation studies in **Table 4** use relatively small M. This is due to the computational burden of Monte Carlo simulation for multiple imputation. Considering this practical issue, the current study sets M to 20, which is equivalent to the 75th percentile of the number of multiply-imputed data found in the studies listed in **Table 4**.

As for T, there is no consensus in the literature (**Table 4**). There are no clear-cut rules for determining whether MCMC algorithms attained convergence (Schafer 1997: 119; King et al. 2001: 59; van Buuren and Groothuis-Oudshoorn 2011: 37). Though not perfect, doubling the number of EM iterations is a rule of thumb for a conservative estimate about convergence speed for MCMC (Schafer and Olsen 1998; Enders 2010: 204). Since it is not possible to check convergence in each of the 1000 simulation runs, the current study relies on the rule of thumb to set T.

8.2 Criteria for Judging Simulation Results

The estimand in all of the simulation runs is β_1 in $y_i = \beta_0 + \beta_1 x_{1i} + \ldots + \beta_{p-1} x_{p-1i} + \varepsilon_i$. The purpose of multiple imputation is to find an unbiased estimate of the population parameter that is confidence valid (van Buuren 2012: 35–36).

Unbiasedness can be assessed by equation (9), because an estimator $\hat{\theta}$ is an unbiased estimator of θ if the expected value of $\hat{\theta}$ is equal to the true θ (Mooney 1997: 59; Gujarati 2003: 899).

$$\text{Bias}\left(\hat{\theta}\right) = E\left(\hat{\theta}\right) - \theta \tag{9}$$

Unbiasedness and efficiency can be simultaneously assessed by the Root Mean Square Error (RMSE), defined as equation (10). The RMSE measures the spread around the true value of the parameter, placing slightly more emphasis on efficiency than bias (Gujarati 2003: 901; Carsey and Harden 2014: 88–89).

$$\text{RMSE}\left(\hat{\theta}\right) = \sqrt{E\left(\hat{\theta} - \theta\right)^2} \tag{10}$$

Confidence validity can be assessed by the coverage probability of the nominal 95% confidence interval (CI), which 'is the proportion of simulated samples for which the estimated confidence interval includes the true parameter' (Carsey and Harden 2014: 93). The formula of the standard error for proportions is equation (11), where π is the proportion and s is the number of simulation runs.

$$\text{SE}\left(\pi\right) = \sqrt{\frac{\pi\left(1-\pi\right)}{s}} \tag{11}$$

The standard error of the 95% CI coverage over 1000 iterations is $\sqrt{0.95 \times 0.05 / 1000} \approx 0.007$ which is 0.7%. Therefore, with 95% confidence, the estimated coverage probability should be between 93.6% and 96.4% (Abe and Iwasaki 2007: 10; Lee and Carlin 2010: 627; Carsey and Harden 2014: 94–95; Hughes, Sterne, and Tilling 2016).

8.3 Results of the Simulation

Abbreviations in this section are explained in **Table 5**, where MI stands for multiple imputation and SI for single imputation.

8.3.1 Theoretical Case

This section presents the results of the Monte Carlo simulation for the theoretical case, where the correlation matrix and the regression coefficients are randomly generated.

Table 6 shows the Bias and RMSE values for the regression coefficient β_1. The Bias and RMSE values for listwise deletion and single imputation methods indicate that these methods are not recommended at all. All of the Bias and RMSE values from EMB, DA1, DA2, and FCS2 are almost identical, showing that they are generally unbiased. However, FCS1 is rather biased, quite similar to S-SI. Therefore, when between-imputation iterations are ignored, there are no discernible effects on bias and efficiency in EMB and DA, but FCS may suffer from some bias.

Table 5: Abbreviations and the Missing Data Methods.

Abbreviations	**Missing Data Methods**
CD	Complete data without missing values
LD	Listwise deletion
EMB	MI by AMELIA II
DA1	MI by NORM2 with no iterations
DA2	MI by NORM2 with 2*EM iterations
FCS1	MI by MICE with no iterations
FCS2	MI by MICE with 2*EM iterations
D-SI	Deterministic SI by `norm.predict` in MICE
S-SI	Stochastic SI by `norm.nob` in MICE

Table 6: Bias and RMSE (Theoretical Data).

		Number of Variables								
		2	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
CD	Bias	0.001	0.003	0.001	0.002	0.001	0.001	0.001	0.002	0.001
	RMSE	0.040	0.047	0.038	0.039	0.058	0.026	0.046	0.039	0.047
LD	Bias	**0.032**	**0.135**	**0.105**	**0.104**	**0.332**	**0.085**	**0.129**	**0.210**	**0.116**
	RMSE	0.059	0.153	0.122	0.121	0.349	0.103	0.160	0.228	0.155
EMB	Bias	0.000	0.004	0.002	0.000	0.005	0.001	0.005	0.005	0.002
	RMSE	0.046	0.053	0.050	0.051	0.075	0.041	0.069	0.059	0.072
DA1	Bias	0.001	0.002	0.003	0.001	0.001	0.000	0.003	0.003	0.002
	RMSE	0.046	0.053	0.050	0.051	0.074	0.041	0.069	0.058	0.072
DA2	Bias	0.002	0.001	0.005	0.002	0.001	0.000	0.001	0.003	0.000
	RMSE	0.046	0.053	0.050	0.051	0.074	0.041	0.069	0.058	0.072
FCS1	Bias	0.002	0.001	**0.082**	**0.040**	**0.090**	**0.047**	**0.093**	**0.027**	**0.233**
	RMSE	0.047	0.053	0.097	0.062	0.116	0.065	0.109	0.052	0.239
FCS2	Bias	0.001	0.002	0.004	0.002	0.001	0.000	0.001	0.002	0.001
	RMSE	0.046	0.053	0.050	0.051	0.075	0.041	0.069	0.058	0.071
D-SI	Bias	**0.186**	**0.242**	**0.174**	**0.093**	**0.187**	**0.098**	**0.231**	**0.070**	**0.163**
	RMSE	0.192	0.248	0.182	0.110	0.207	0.109	0.248	0.099	0.189
S-SI	Bias	0.002	0.000	**0.081**	**0.038**	**0.090**	**0.047**	**0.091**	**0.029**	**0.230**
	RMSE	0.050	0.057	0.102	0.066	0.124	0.076	0.119	0.062	0.241

Note: Biased results are in boldface, i.e., Bias > 0.010.

Table 7 gives the coverage probability of the 95% CI for β_1. The CIs for listwise deletion and single imputation methods are not confidence valid. When the number of auxiliary variables is small (and hence the overall missing rate is small), the between-imputation iterations may be ignored, where all of the multiple imputation CIs are confidence valid. However, as the number of auxiliary variables becomes large, DA1 and FCS1 drift away from the confidence validity. EMB, DA2, and FCS2 are confidence valid regardless of the number of variables and the missing rate. This shows that EMB is confidence proper even if it does not iterate. This is an important finding in the current study.

Table 8 shows the CI lengths. The CI length by listwise deletion is generally too long, reflecting inefficiency due to the reduced sample size. The CI lengths by single imputation methods are 'correct' in the sense that they are quite similar to those of complete data analysis; however, this means that single imputation methods ignore estimation uncertainty associated with imputation. This is the cause of confidence invalidity of single imputation methods in **Table 7**. The CI length by DA1 is too short and the CI length by FCS1 is too long. The CI lengths by EMB, DA2, and FCS2 are essentially equal, reflecting the correct level of estimation uncertainty associated with imputation.

Table 9 displays the computational time required to generate multiple imputations. When the number of auxiliary variables is small (and hence the overall missing rate is small), DA2 is fastest among the three confidence proper multiple imputation algorithms. On the other hand, as the number of auxiliary variables becomes large, EMB becomes fastest. As is known in the literature (van Buuren 2012: 117; Kropko et al. 2014), FCS2 is at least 5 times slower and can be more than 50 times slower than EMB and DA2. However, the difference in computational time is not substantial, given that all of the computations can be done within a few minutes.

Table 7: Coverage of the 95% CI (Theoretical Data).

	Number of Variables								
	2	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
CD	95.3	94.9	94.2	94.0	96.0	96.0	95.3	94.9	94.6
LD	**88.5**	**47.9**	**54.6**	**56.7**	**10.8**	**65.1**	**69.2**	**32.1**	**78.1**
EMB	95.0	95.1	94.2	95.5	94.9	94.4	94.3	94.1	95.0
DA1	94.6	94.9	**93.2**	**93.1**	94.1	**91.8**	**92.9**	**92.4**	**92.9**
DA2	94.3	95.8	95.1	94.1	94.8	94.3	94.2	**93.2**	94.9
FCS1	94.2	95.0	**75.0**	**91.6**	**84.4**	95.5	**84.5**	**96.8**	**6.8**
FCS2	94.7	95.6	94.4	93.9	95.4	94.5	94.2	95.0	95.0
D-SI	**0.8**	**0.2**	**2.2**	**37.8**	**22.2**	**16.9**	**8.3**	**51.0**	**22.5**
S-SI	**88.9**	**89.6**	**47.8**	**75.0**	**62.3**	**64.4**	**48.9**	**76.0**	**3.7**

Note: Confidence invalid results are in boldface, i.e., outside of 93.6 and 96.4.

Table 8: Lengths of the 95% CI (Theoretical Data).

	Number of Variables								
	2	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
CD	0.157	0.184	0.144	0.148	0.236	0.102	0.184	0.151	0.180
LD	0.189	0.259	0.226	0.235	0.384	0.213	0.358	0.339	0.390
EMB	0.178	0.209	0.196	0.200	0.301	0.160	0.275	0.229	0.281
DA1	0.176	0.207	0.187	0.192	0.293	0.145	0.256	0.208	0.253
DA2	0.177	0.208	0.194	0.198	0.298	0.158	0.271	0.223	0.274
FCS1	0.178	0.209	0.237	0.211	0.324	0.248	0.306	0.223	0.299
FCS2	0.178	0.209	0.197	0.201	0.302	0.161	0.275	0.228	0.281
D-SI	0.143	0.174	0.133	0.149	0.244	0.103	0.205	0.150	0.188
S-SI	0.157	0.184	0.161	0.155	0.238	0.145	0.188	0.149	0.186

Table 9: Computational Time (Theoretical Data).

	Number of Variables								
	2	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
EMB	0.46	0.53	0.53	0.59	0.71	**0.78**	**0.97**	**1.27**	**1.69**
DA2	**0.10**	**0.16**	**0.29**	**0.42**	**0.55**	1.09	1.39	2.22	3.63
FCS2	2.47	5.98	14.48	21.33	25.40	54.71	59.14	85.69	133.17

Note: Reported values are the time in seconds to perform multiple imputation, which is averaged over 1,000 simulation runs. The fastest results are in boldface.

8.3.2 Realistic Case

This section presents the results of the Monte Carlo simulation for the realistic case, where the correlation matrix and the regression coefficients are based on the real data (CIA 2016; Freedom House 2016). The results in this section reinforce the findings in Section 8.3.1.

Table 10 shows the Bias and RMSE values for the regression coefficient β_1. The overall conclusions are similar to **Table 6**. When between-imputation iterations are ignored, there are no discernible effects on bias and efficiency in EMB and DA, but FCS may occasionally suffer from small bias.

Table 11 gives the coverage probability of the 95% CI for β_1. The overall conclusions are similar to **Table 7**, except that DA1 is confidence invalid even when $p = 3$. This implies that we cannot ignore between-imputation iterations in MCMC-based approaches even when the number of variables is small. On the other hand, EMB is confidence valid and we can safely ignore between-imputation iterations in EMB. Again, this is an important finding in the current study.

Table 10: Bias and RMSE (Realistic Data).

		Number of Variables								
		2	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
CD	Bias	0.003	0.002	0.002	0.002	0.001	0.002	0.000	0.002	0.002
	RMSE	0.074	0.086	0.068	0.067	0.066	0.065	0.070	0.069	0.075
LD	Bias	**0.034**	**0.047**	**0.037**	**0.054**	**0.082**	**0.099**	**0.083**	**0.072**	**0.085**
	RMSE	0.095	0.128	0.104	0.118	0.141	0.154	0.157	0.159	0.188
EMB	Bias	0.001	0.002	0.002	0.005	0.001	0.000	0.000	0.002	0.006
	RMSE	0.084	0.113	0.091	0.090	0.089	0.092	0.102	0.099	0.110
DA1	Bias	0.006	0.001	0.003	0.003	0.001	0.001	0.001	0.001	0.002
	RMSE	0.084	0.112	0.090	0.089	0.087	0.091	0.100	0.096	0.105
DA2	Bias	0.009	0.000	0.002	0.004	0.002	0.004	0.000	0.001	0.001
	RMSE	0.084	0.111	0.089	0.088	0.086	0.090	0.098	0.094	0.102
FCS1	Bias	0.007	**0.013**	0.006	0.005	0.002	0.008	0.006	**0.012**	0.000
	RMSE	0.084	0.106	0.081	0.081	0.080	0.081	0.086	0.083	0.088
FCS2	Bias	0.007	0.001	0.002	0.002	0.003	0.005	0.002	0.003	0.005
	RMSE	0.084	0.112	0.088	0.088	0.086	0.090	0.097	0.093	0.100
D-SI	Bias	**0.188**	**0.075**	**0.011**	**0.035**	**0.037**	**0.047**	**0.023**	**0.034**	**0.059**
	RMSE	0.207	0.163	0.115	0.118	0.118	0.123	0.130	0.127	0.151
S-SI	Bias	0.005	**0.014**	0.007	0.006	0.002	0.006	0.005	0.009	0.006
	RMSE	0.089	0.116	0.096	0.095	0.091	0.094	0.100	0.102	0.105

Note: Biased results are in boldface, i.e., Bias > 0.010.

Table 12 shows the CI lengths. The overall conclusions are similar to **Table 8**. One difference is that the CI length by FCS1 is slightly short.

Table 13 displays the computational time required to generate multiple imputations. The overall conclusions are similar to **Table 9**.

Table 11: Coverage of the 95% CI (Realistic Data).

	Number of Variables								
	2	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
CD	94.6	95.3	95.8	94.7	95.2	96.4	94.6	95.3	94.8
LD	**92.2**	**91.6**	**92.8**	**91.5**	**86.8**	**85.0**	**89.8**	**90.0**	**90.8**
EMB	94.3	94.1	94.7	93.9	96.1	94.2	94.0	94.4	94.7
DA1	94.1	**92.2**	94.4	**93.4**	95.7	**92.2**	**93.1**	**92.9**	**93.1**
DA2	94.0	94.0	94.8	94.4	95.9	94.5	93.8	95.0	95.0
FCS1	94.6	94.7	96.3	**96.7**	**97.0**	**97.0**	**96.7**	**96.9**	**97.7**
FCS2	94.7	93.8	95.5	95.7	96.4	94.3	94.8	95.2	96.1
D-SI	**32.7**	**74.5**	**79.2**	**77.6**	**77.7**	**74.1**	**75.3**	**75.1**	**68.8**
S-SI	**87.9**	**83.2**	**82.3**	**82.5**	**84.2**	**82.1**	**81.0**	**80.3**	**81.2**

Note: Confidence invalid results are in boldface, i.e., outside of 93.6 and 96.4.

Table 12: Lengths of the 95% CI (Realistic Data).

	Number of Variables								
	2	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
CD	0.279	0.334	0.268	0.266	0.267	0.261	0.278	0.274	0.289
LD	0.333	0.441	0.389	0.412	0.436	0.457	0.516	0.543	0.631
EMB	0.314	0.429	0.364	0.356	0.362	0.359	0.397	0.396	0.432
DA1	0.313	0.414	0.348	0.342	0.343	0.337	0.370	0.364	0.390
DA2	0.315	0.423	0.356	0.351	0.353	0.351	0.383	0.380	0.410
FCS1	0.315	0.416	0.353	0.348	0.350	0.350	0.382	0.380	0.406
FCS2	0.316	0.429	0.359	0.355	0.358	0.352	0.389	0.386	0.413
D-SI	0.288	0.380	0.292	0.289	0.291	0.278	0.302	0.294	0.315
S-SI	0.281	0.325	0.262	0.257	0.259	0.255	0.269	0.267	0.277

Table 13: Computational Time (Realistic Data).

	Number of Variables								
	2	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
EMB	0.14	0.15	0.16	0.20	0.23	0.28	0.36	**0.44**	**0.53**
DA2	**0.04**	**0.05**	**0.06**	**0.10**	**0.15**	**0.22**	**0.33**	0.47	0.67
FCS2	1.05	2.55	4.22	8.92	12.02	15.59	20.82	26.78	35.95

Note: Reported values are the time in seconds to perform multiple imputation, which is averaged over 1,000 simulation runs. The fastest results are in boldface.

9 Conclusions

This article assessed the relative performance of the three multiple imputation algorithms (DA, FCS, and EMB). In both theoretical and realistic settings (**Table 7** and **Table 11**), if between-imputation iterations were ignored, the MCMC algorithms (DA and FCS) did not attain confidence validity. The nominal 95% CIs by DA and FCS without iterations were different from 95% coverage beyond the margin of error in 1,000 simulation runs. This is because the CI lengths by DA without iterations were generally too short, and the CI lengths by FCS are generally too long (**Table 8** and **Table 12**). Based on Schafer (1997: 139), this can be explained by choices for starting values. DA uses EM as a single starting value for *M* chains that understates missing data uncertainty (Schafer 2016: 22) while FCS uses random draws as *M* over-dispersed starting values that overstates missing data uncertainty (van Buuren and Groothuis-Oudshoorn 2011: 6). Without iterations, imputed values depend on the choice of starting values.

DA and FCS can be both confidence valid under the large number of iterations; however, the assessment of convergence in MCMC is notoriously difficult. Furthermore, the convergence properties of FCS are currently under debate due to possible incompatibility (Li, Yu, and Rubin 2012; Zhu and Raghunathan 2015). On the other hand, the current study found that EMB was confidence valid regardless of the situations. Therefore, EMB is a confidence proper imputation algorithm without iterations, which allows us to avoid a painful decision-making process of how to judge the convergence to generate confidence proper multiple imputations. This finding is useful in the missing data literature. For example, while ratio imputation is often used in official statistics (Takahashi, Iwasaki, and Tsubaki 2017), multiple ratio imputation does not exist in the literature. The EMB algorithm was applied to ratio imputation to create multiple ratio imputation (Takahashi 2017a; Takahashi 2017b).

No simulation studies can include all the patterns of relevant data (Kropko et al. 2014: 511). Therefore, the current study focused on two types of data, (1) theoretical and (2) realistic. Although the author believes that the two data generation processes cover data types relevant to many social research situations, the results in any simulation studies must be read with caution (Hardt, Herke, and Leonhart 2014: 11). Future research should delve into other data types, such as small-*n* data, large-*p* data, categorical data, and non-normal data, to name a few.

Additional File

The additional file for this article can be found as follows:

- **Data for Tables 1 and 2.** Political and Economic Data from CIA (2016) and Freedom House (2016). DOI: https://doi.org/10.5334/dsj-2017-037.s1

Acknowledgements

The author wishes to thank Dr. Manabu Iwasaki (Seikei University), Dr. Michiko Watanabe (Keio University), and Dr. Takayuki Abe (Keio University) for the helpful comments. The author also wishes to thank the two anonymous reviewers for their comments that improved the quality of the article. Note that part of this article in its very early version was presented at the 59th World Statistics Congress of the International Statistical Institute (Takahashi and Ito 2013).

Competing Interests

The author has no competing interests to declare.

References

Abe, T and **Iwasaki, M** 2007 Evaluation of statistical methods for analysis of small-sample longitudinal clinical trials with dropouts. *Journal of the Japanese Society of Computational Statistics*, 20(1): 1–18. DOI: https://doi.org/10.5183/jjscs1988.20.1

Acemoglu, D, Johnson, S and **Robinson, J A** 2005 Institutions as the fundamental cause of long-run growth, Aghion, P and Durlauf, S (eds.). In: *Handbook of Economic Growth*. North Holland: Elsevier.

Allison, P D 2002 *Missing Data*. Thousand Oaks, CA: Sage Publications. DOI: https://doi.org/10.4135/9781412985079

Baraldi, A N and **Enders, C K** 2010 An introduction to modern missing data analyses. *Journal of School Psychology*, 48(1): 5–37. DOI: https://doi.org/10.1016/j.jsp.2009.10.001

Barnard, J and **Rubin, D B** 1999 Small-sample degrees of freedom with multiple imputation. *Biometrika*, 86(4): 948–955. DOI: https://doi.org/10.1093/biomet/86.4.948

Barro, R J 1997 *Determinants of Economic Growth: A Cross-Country Empirical Study*. Cambridge, MA: MIT Press.

Bodner, T E 2008 What improves with increased missing data imputations? *Structural Equation Modeling*, 15: 651–675. DOI: https://doi.org/10.1080/10705510802339072

Carpenter, J R and **Kenward, M G** 2013 *Multiple Imputation and its Application*. Chichester, West Sussex: A John Wiley & Sons Publication.

Carsey, T M and **Harden, J J** 2014 *Monte Carlo Simulation and Resampling Methods for Social Science*. Thousand Oaks, CA: Sage Publications. DOI: https://doi.org/10.4135/9781483319605

Central Intelligence Agency 2016 *The World Factbook*. Available at: https://www.cia.gov/library/publications/the-world-factbook/index.html [Last accessed November 27, 2016].

Cheema, J R 2014 Some general guidelines for choosing missing data handling methods in educational research. *Journal of Modern Applied Statistical Methods*, 13(2): 53–75. DOI: https://doi.org/10.22237/jmasm/1414814520

Cranmer, S J and **Gill, J** 2013 We have to be discrete about this: A non-parametric imputation technique for missing categorical data. *British Journal of Political Science*, 43(2): 425–449. DOI: https://doi.org/10.1017/S0007123412000312

Deng, Y, Chang, C, Ido, M S and **Long, Q** 2016 Multiple imputation for general missing data patterns in the presence of high-dimensional data. *Scientific Reports*, 6(21689): 1–10. DOI: https://doi.org/10.1038/srep21689

de Waal, T, Pannekoek, J and **Scholtus, S** 2011 *Handbook of Statistical Data Editing and Imputation*. Hoboken, NJ: John Wiley & Sons. DOI: https://doi.org/10.1002/9780470904848

Do, B C and **Batzoglou, S** 2008 What is the expectation maximization algorithm? *Nature Biotechnology*, 26(8): 897–899. DOI: https://doi.org/10.1038/nbt1406

Donders, A R T, van der Heijden, G J M G, Stijnen, T and **Moons, K G M** 2006 Review: A gentle introduction to imputation of missing values. *Journal of Clinical Epidemiology*, 59: 1087–1091. DOI: https://doi.org/10.1016/j.jclinepi.2006.01.014

Enders, C K 2010 *Applied Missing Data Analysis*. New York, NY: The Guilford Press.

Feng, Y 2003 *Democracy, Governance, and Economic Performance: Theory and Evidence*. Cambridge, MA: The MIT Press.

Freedom House 2016 *Freedom in the World 2016*. Available at: https://freedomhouse.org/report/freedom-world/freedom-world-2016 [Last accessed November 30, 2016].

Gill, J 2008 *Bayesian Methods: A Social and Behavioral Sciences Approach*, Second Edition. London: Chapman & Hall/CRC.

Graham, J W 2009 Missing data analysis: Making it work in the real world. *Annual Review of Psychology*, 60: 549–576. DOI: https://doi.org/10.1146/annurev.psych.58.110405.085530

Graham, J W, Olchowski, A E and **Gilreath, T D** 2007 How many imputations are really needed? Some practical clarifications of multiple imputation theory. *Prevention Science*, 8(3): 206–213. DOI: https://doi.org/10.1007/s11121-007-0070-9

Gujarati, D N 2003 *Basic Econometrics*, Fourth Edition. Boston, MA: McGraw-Hill.

Hardt, J, Herke, M and **Leonhart, R** 2012 Auxiliary variables in multiple imputation in regression with missing X: A warning against including too many in small sample research. *BMC Medical Research Methodology,* 12(184): 1–13. DOI: https://doi.org/10.1186/1471-2288-12-184

Honaker, J and **King, G** 2010 What to do about missing values in time series cross-section data. *American Journal of Political Science*, 54(2): 561–581. DOI: https://doi.org/10.1111/j.1540-5907.2010.00447.x

Honaker, J, King, G and **Blackwell, M** 2011 Amelia II: A program for missing data. *Journal of Statistical Software*, 45(7): 1–47. DOI: https://doi.org/10.18637/jss.v045.i07

Honaker, J, King, G and **Blackwell, M** 2016 *Package 'Amelia'*. Available at: http://cran.r-project.org/web/packages/Amelia/Amelia.pdf [Last accessed November 30, 2016].

Horowitz, J L 2001 The bootstrap, Heckman, J J and Leamer, E (eds.). In: *Handbook of Econometrics*, 5. North Holland: Elsevier. DOI: https://doi.org/10.1016/s1573-4412(01)05005-x

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*, 61(1): 79–90. DOI: https://doi.org/10.1198/000313007X172556

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. DOI: https://doi.org/10.1198/000313007X172556

Hughes, R A, Sterne, J A C and **Tilling, K** 2016 Comparison of imputation variance estimators. *Statistical Methods in Medical Research*, 25(6): 2541–2557. DOI: https://doi.org/10.1177/0962280214526216

King, G, Honaker, J, Joseph, A and **Scheve, K** 2001 Analyzing incomplete political science data: An alternative algorithm for multiple imputation. *American Political Science Review*, 95(1): 49–69.

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(4): 497–519. DOI: https://doi.org/10.1093/pan/mpu007

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(5): 624–632. DOI: https://doi.org/10.1093/aje/kwp425

Lee, K J and **Carlin, J B** 2012 Recovery of information from multiple imputation: A simulation study. *Emerging Themes in Epidemiology*, 9(3): 1–10. DOI: https://doi.org/10.1186/1742-7622-9-3

Leite, W and **Beretvas, S** 2010 The performance of multiple imputation for Likert-type items with missing data. *Journal of Modern Applied Statistical Methods*, 9(1): 64–74. DOI: https://doi.org/10.22237/jmasm/1272686820

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*, 11(14): 1–35.

Little, R J A and **Rubin, D B** 2002 *Statistical Analysis with Missing Data*, Second Edition. Hoboken, NJ: John Wiley & Sons. DOI: https://doi.org/10.1002/9781119013563

McNeish, D 2017 Missing data methods for arbitrary missingness with small samples. *Journal of Applied Statistics*, 44(1): 24–39. DOI: https://doi.org/10.1080/02664763.2016.1158246

Mooney, C Z 1997 *Monte Carlo Simulation*. Thousand Oaks, CA: Sage Publications. DOI: https://doi.org/10.4135/9781412985116

Raghunathan, T 2016 *Missing Data Analysis in Practice*. Boca Raton, FL: CRC Press.

Rubin, D B 1987 *Multiple Imputation for Nonresponse in Surveys*. New York, NY: John Wiley & Sons. DOI: https://doi.org/10.1002/9780470316696

Schafer, J L 1997 *Analysis of Incomplete Multivariate Data*. Boca Raton, FL: Chapman & Hall/CRC. DOI: https://doi.org/10.1201/9781439821862

Schafer, J L 2016 *Package 'norm2'*. Available at: https://cran.r-project.org/web/packages/norm2/norm2.pdf [Last accessed November 30, 2016].

Schafer, J L and **Graham, J W** 2002 Missing data: Our view of the state of the art. *Psychological Methods*, 7(2): 147–177. DOI: https://doi.org/10.1037//1082-989X.7.2.147

Schafer, J L and **Olsen, M K** 1998 Multiple imputation for multivariate missing-data problems: A data analyst's perspective. *Multivariate Behavioral Research*, 33: 545–571. DOI: https://doi.org/10.1207/s15327906mbr3304_5

Schenker, N, Raghunathan, T E, Chiu, 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(475): 924–933. DOI: https://doi.org/10.1198/016214505000001375

Scheuren, F 2005 Multiple imputation: How it began and continues. *The American Statistician*, 59(4): 315–319. DOI: https://doi.org/10.1198/000313005X74016

Shara, N, Yassin, S A, Valaitis, E, Wang, H, Howard, B V, Wang, W, Lee, E T and **Umans, J G** 2015 Randomly and non-randomly missing renal function data in the strong heart study: A comparison of imputation methods. *PLOS ONE*, 10(9): 1–11. DOI: https://doi.org/10.1371/journal.pone.0138923

Stuart, E A, Azur, M, Frangakis, C and **Leaf, P** 2009 Multiple imputation with large data sets: A case study of the children's mental health initiative. *American Journal of Epidemiology*, 169(9): 1133–1139. DOI: https://doi.org/10.1093/aje/kwp026

Takahashi, M 2017a Multiple ratio imputation by the EMB algorithm: Theory and simulation. *Journal of Modern Applied Statistical Methods*, 16(1): 630–656. DOI: https://doi.org/10.22237/jmasm/1493598840

Takahashi, M 2017b Implementing multiple ratio imputation by the EMB algorithm (R). *Journal of Modern Applied Statistical Methods*, 16(1): 657–673. DOI: https://doi.org/10.22237/jmasm/1493598900

Takahashi, M and **Ito, T** 2013 "Multiple imputation of missing values in economic surveys: Comparison of competing algorithms," *Proceedings of The 59th World Statistics Congress of the International Statistical Institute (ISI)*, 3240–3245. Hong Kong, China.

Takahashi, M and **Ito, T** 2014 Comparison of competing algorithms of multiple imputation: Analysis using large-scale economic data. *Research Memoir of Official Statistics*, (71): 39–82.

Takahashi, M, Iwasaki, M and **Tsubaki, H** 2017 Imputing the mean of a heteroskedastic log-normal missing variable: A unified approach to ratio imputation. *Statistical Journal of the IAOS*, 33(3), in press. DOI: https://doi.org/10.3233/SJI-160306

van Buuren, S 2012 *Flexible Imputation of Missing Data*. Boca Raton, FL: Chapman & Hall/CRC. DOI: https://doi.org/10.1201/b11826

van Buuren, S and **Groothuis-Oudshoorn, K** 2011 mice: multivariate imputation by chained equations in R. *Journal of Statistical Software*, 45(3): 1–67. DOI: https://doi.org/10.18637/jss.v045.i03

van Buuren, S, Groothuis-Oudshoorn, K, Roxitzsch, A, Vink, G, Doove, L and **Jolani, S** 2015 *Package 'mice'*. Available at: https://cran.r-project.org/web/packages/mice/mice.pdf [Last accessed November 30, 2016].

von Hippel, P T 2016 New confidence intervals and bias comparisons show that maximum likelihood can beat multiple imputation in small samples. *Structural Equation Modeling*, 23(3): 422–437. DOI: https://doi.org/10.1080/10705511.2015.1047931

Zhu, J and **Raghunathan, T E** 2015 Convergence properties of a sequential regression multiple imputation algorithm. *Journal of the American Statistical Association*, 110(511): 1112–1124. DOI: https://doi.org/10.1080/01621459.2014.948117

How to cite this article: Takahashi, M 2017 Statistical Inference in Missing Data by MCMC and Non-MCMC Multiple Imputation Algorithms: Assessing the Effects of Between-Imputation Iterations. *Data Science Journal,* 16: 37, pp. 1–17, DOI: https://doi.org/10.5334/dsj-2017-037

Submitted: 30 November 2016 **Accepted:** 23 June 2017 **Published:** 28 July 2017

]u[*Data Science Journal* is a peer-reviewed open access journal published by Ubiquity Press.

OPEN ACCESS