Journal of Survey Statistics and Methodology (2024) **12**, 183–210

JOINT IMPUTATION OF GENERAL DATA

MICHAEL W. ROBBINS *

High-dimensional complex survey data of general structures (e.g., containing continuous, binary, categorical, and ordinal variables), such as the US Department of Defense's Health-Related Behaviors Survey (HRBS), often confound procedures designed to impute any missing survey data. Imputation by fully conditional specification (FCS) is often considered the state of the art for such datasets due to its generality and flexibility. However, FCS procedures contain a theoretical flaw that is exposed by HRBS data—HRBS imputations created with FCS are shown to diverge across iterations of Markov Chain Monte Carlo. Imputation by joint modeling lacks this flaw; however, current joint modeling procedures are neither general nor flexible enough to handle HRBS data. As such, we introduce an algorithm that efficiently and flexibly applies multiple imputation by joint modeling in data of general structures. This procedure draws imputations from a latent joint multivariate normal model that underpins the generally structured data and models the latent data via a sequence of conditional linear models, the predictors of which can be specified by the user. We perform rigorous evaluations of HRBS imputations created with the new algorithm and show that they are convergent and of high quality. Lastly, simulations verify that the proposed method performs well compared to existing algorithms including FCS.

KEY WORDS: Fully conditional specification; Joint modeling; Markov Chain Monte Carlo; Missing data; Multiple imputation.

Michael W. Robbins is a Senior Statistician with the RAND Corporation, Pittsburgh, PA 15213, USA.

This work was partially funded by the US Department of Defense and from grant R21AG058123 from the National Institutes of Health. The views expressed are those of the author and do not necessarily represent the views of RAND, the Department of Defense, or the National Institutes of Health.

*Address correspondence to Michael W. Robbins, RAND Corporation, Pittsburgh, PA 15213, USA; E-mail: mrobbins@rand.org.

https://doi.org/10.1093/jssam/smad034 Advance access publication 12 September 2023

Statement of Significance

This article illustrates a novel, flexible, and computationally efficient procedure for imputation of missing values in high-dimensional datasets of a general structure via joint modeling. The method outperforms existing state-of-the-art techniques in a general simulation study and is shown to produce high-quality imputations in a data application wherein procedures that use fully conditional specification (i.e., mice) yield divergent imputations.

1. INTRODUCTION

Missing data present one of the classical problems of statistical analyses. Imputation, in which missing values are replaced with plausible entries according to some sort of statistical model, is a highly popular approach for addressing missing data as it yields completed datasets that can be analyzed with traditional techniques. Modern approaches to imputation have tended to settle within a Bayesian paradigm wherein imputations are sampled at random from a posterior predictive distribution; this begets the multiple imputation framework in which estimators of uncertainty can be adjusted for imputation error through the creation of several imputed datasets. Most commonly used imputation procedures generate imputations iteratively via Markov Chain Monte Carlo (MCMC) in hopes that after a burn-in period of iterations, the imputations will represent draws from the posterior distribution of the missing data given the observed data. Reviews of missing data, imputation, and multiple imputation are numerous—examples include Rubin (1987, 1996), Schafer (1999), Carpenter and Kenward (2012), and Little and Rubin (2020).

This article is motivated by the problem of imputation in the US Department of Defense’s 2018 Health-Related Behaviors Survey (HRBS) of service members, which is designed to assess health behaviors that have the potential to impact readiness and to assess the overall well-being of the US Armed Forces. HRBS data are high-dimensional (approximately 33,000 respondents across with nearly 300 variables) and have a difficult-to-model data structure (e.g., the survey includes complicated skip logic and many items are sparsely distributed binary variables). As demonstrated later, we find that no existing imputation algorithm is well suited to handle the complexities of HRBS data.

The current state of the art for missing data problems in large-scale surveys akin to HRBS is often considered imputation by fully conditional specification (FCS, Raghunathan et al. 2001; Van Buuren et al. 2006; Van Buuren and

Groothuis-Oudshoorn 2011; White et al. 2011), also known as chained equations, wherein each variable is imputed from a conditional model that potentially includes all other variables. This process naturally lends itself to imputation of variables of general structure (e.g., binary, unordered categorical, ordinal); furthermore, transformation (e.g., Robbins and White 2011; Robbins 2014; Lee and Carlin 2017) or predictive mean matching (Little 1988) can be applied to preserve continuous marginal distributions that are nonstandard. Conditional modeling and imputation may be performed with random forests (Doove et al. 2014; Shah et al. 2014) or regression trees (Burgette and Reiter 2010; Doove et al. 2014) within FCS procedures. Since the conditional models can be, in theory, incompatible with one another, FCS does not necessarily sample imputations from a valid joint distribution, and as such, the imputations are not guaranteed to converge across iterations of MCMC. In spite of its theoretical flaws, FCS is thought to perform well in practice (Lee and Carlin 2010; White et al. 2011; Van Buuren 2018) and is widely used and available across a host of software (e.g., Raghunathan et al. 2002; Van Buuren and Groothuis-Oudshoorn 2011; Su et al. 2011; Honaker et al. 2011). However, application of FCS methods to HRBS induces problems. Imputation of sparse binary variables by logistic regression yields clearly erroneous marginal distributions, whereas divergence across iterations of MCMC is observed when predictive mean matching is applied.

Imputation algorithms that sample from valid joint distributions have been developed (e.g. Schafer 2017; Gondara and Wang 2018; Hoff 2018; Yoon et al. 2018; Zhao and Schafer 2018; Quartagno and Carpenter 2020; Erler et al. 2021; Grund et al. 2021). However, these procedures tend to be incompatible or highly inefficient with data of general structures or high dimensions. For example, these procedures often lack the flexibility to impose selected conditional dependencies within imputation modeling, which renders them computationally infeasible with HRBS data.

Here, we introduce a new procedure that borrows from earlier ideas (Carpenter and Kenward 2012; Robbins et al. 2013) and addresses the theoretical and empirical issues encountered with FCS. This algorithm imposes a latent multivariate normal process to facilitate the imputation of continuous, binary, unordered categorical, and ordinal (i.e., ordered categorical) variables. To ensure theoretical validity, procedure draws imputations from a joint model while building that model from a sequence of linear conditional models. Modeling in such a fashion enables flexibility in the selection of conditional relationships that permitted between variables. The sweep operator (Goodnight 1979) optimizes the computational performance of the algorithm. When the method is applied to HRBS, diagnostics of both marginal and multivariate distributions indicate strong performance with convergence observed over MCMC iterations. The new procedure also has the potential to be dramatically more computationally efficient than FCS with high-dimensional data.

2. THE HEALTH-RELATED BEHAVIORS SURVEY DATA

The HRBS, which has been administered in some form for over 30 years, has been described as the US Department of Defense's "flagship survey for understanding the health, health-related behaviors, and well-being of service members… the HRBS asks questions about health-related issues that can affect force readiness or the ability to meet the demands of military life" (https://www.rand.org/nsrd/projects/hrbs.html). The 2018 version of the survey was administered to service members in the US Armed Forces of all ranks and pay grades (excluding generals and admirals), components (including active duty, reserve, and National Guard but excluding those currently deployed), and branches. The sample was stratified by component, pay grade, branch, and gender, and disproportionate sampling across strata was used to account for response rates that were expected to vary across strata and increase counts in less prevalent strata. In particular, strata involving Coast Guard and Marine Corps, as well as women, were oversampled, whereas strata involving the Air Force were sampled at smaller rates. The target population for the 2018 HRBS included 2,170,000 service members, 400,000 of which were sampled, yielding a total of 33,641 respondents (for a response rate of 8.4 percent). Survey weights that account for sample design and nonresponse were developed. See Meadows et al. (2020a, 2020b) for details.

Missingness in HRBS data occurs through two primary means: (1) drop out, which occurs when an individual stops midway through and fails to return to complete the survey, and (2) refusal, which occurs when an individual fails to respond to a specific item on the survey but does respond to some subsequent items. The bulk of the missingness in the 2018 HRBS (approximately 94 percent) was due to drop out, which leads to a nearly (though not entirely) monotonic missingness pattern. Missingness rates in the data range from less than 0.1 percent for items appearing early in the survey to seven percent for items that occur later. Approximately 89 percent of cases are complete.

The survey data contain 265 items split across 14 separate modules, many of which regard sensitive topics such as drug and alcohol use, sexual behavior, and gambling addiction. Most of the items are binary (e.g., yes/no), and many have sparse distributions (e.g., a very low prevalence of "yes" responses). Note that several of the survey items are subject to skip logic in that they are only asked of respondents who provide specific answers to other questions. Specifically, the survey instrument contains both parent (e.g., Did you deploy in the past year?) and child questions (e.g., For how long did you deploy?), where child questions are asked only of those who provided certain response to the parent questions. As such, HRBS mandates an imputation algorithm that has the flexibility to select conditional dependence structures for each variable. That is, there is no basis upon which a relationship between a parent question and a child question can be estimated (since child questions are only observed for individuals who provide a specific answer to the parent question).

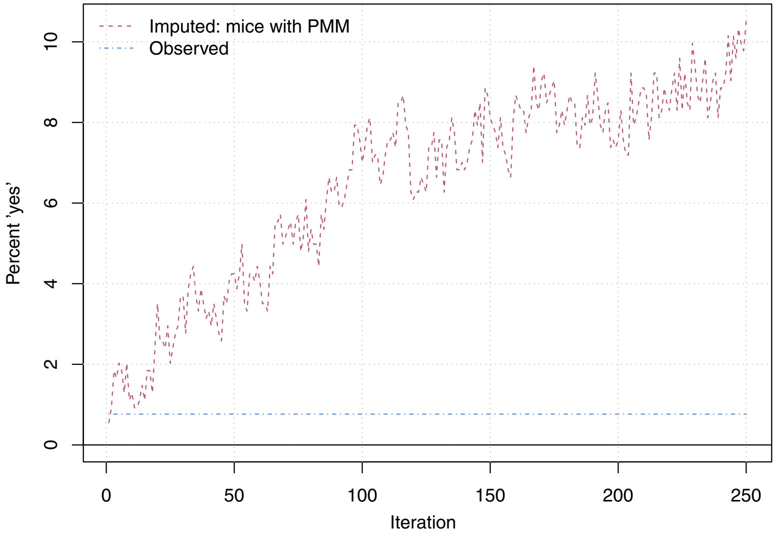

Figure 1. Trace Plot Across Iterations of MCMC of the Mean (in Percent) of Imputations Created using `mice` with PMM for HRBS Variable Q40A: "In the Past 12 Months, Have You Used Marijuana or Hasish?"

2.1 HRBS and Existing Imputation Algorithms

Although we provide further details in section 6, application of existing imputation algorithms to HRBS data induces problems that are briefly overviewed here. Procedures that apply joint modeling and are designed for general data, such as `jomo` (Quartagno and Carpenter 2020), `sbgcop` (Hoff 2018), and HCMM (Murray and Reiter 2016) crash and therefore fail to produce imputations, likely due to their inability to selectively model conditional relationships. Similarly, `mice` (Van Buuren and Groothuis-Oudshoorn 2011), which utilizes FCS, when applied in conjunction with random forests or regression trees also fails to produce imputations. When `mice` is applied with logistic and polytmous regression for binary and categorical variables, imputations are created that are clearly erroneous. Furthermore, `mice` when applied with predictive mean matching (PMM, Little 1988) produces imputations that diverge across iterations of MCMC. An example of this divergence is illustrated via a trace plot seen in figure 1 for a representative item.

It is clear that HRBS requires a new imputation algorithm that is general and flexible enough to handle its complexities. The new procedure is outlined in section 4, although relevant background for imputation procedures in general is first provided in section 3.

3. AN IMPUTATION PRIMER

Relevant imputation methods are founded on the concept of data augmentation (DA, Tanner and Wong 1987). DA is designed for cases where the desired objective of sampling from a posterior distribution $P(\theta|y)$ is difficult, but for some latent variable z, sampling from $P(z|y, \theta)$ and $P(\theta|y, z)$ is simple, where $P(\cdot)$ is general notation for a probabilistic density. As such, DA involves iteratively sampling from $P(z|y, \theta)$ and $P(\theta|y, z)$ to yield valid draws from $P(\theta, z|y)$. In missing data models, it is common to let y represent the observed data in the DA formulation, z represent the missing data, and θ model parameters. As such, imputation via DA involves iteratively alternating between an imputation step (or I Step), which involves sampling updated imputations from the density of the missing data given the observed data and the parameters sampled from the previous iteration, and a parameter step (P Step) wherein one samples parameters from the density of the parameters given the observed data and the imputations sampled from the preceding I Step.

To illustrate the DA process with more formal notation, let $\boldsymbol{\chi}_{\text{obs}}$ denote the observed data and $\boldsymbol{\chi}_{\text{mis}}$ denote the missing data, while $\boldsymbol{\chi} = \{\boldsymbol{\chi}_{\text{obs}}, \boldsymbol{\chi}_{\text{mis}}\}$ gives the complete data. Furthermore, $\boldsymbol{\Theta}$ is a set of model parameters that govern the distribution of $\boldsymbol{\chi}$. The objective is to sample imputations from $P(\boldsymbol{\chi}_{\text{mis}}|\boldsymbol{\chi}_{\text{obs}}, \boldsymbol{\Theta})$. Letting $\boldsymbol{\chi}_{\text{mis}}^{(t)}$ and $\boldsymbol{\Theta}^{(t)}$ represent samples of $\boldsymbol{\chi}_{\text{mis}}$ and $\boldsymbol{\Theta}$ drawn at the t^{th} iteration, these are updated within the $(t+1)^{th}$ iteration via:

I Step: Draw $\boldsymbol{\chi}_{\text{mis}}^{(t+1)}$ from $P(\boldsymbol{\chi}_{\text{mis}}|\boldsymbol{\chi}_{\text{obs}}, \boldsymbol{\Theta}^{(t)})$.
P Step: Draw $\boldsymbol{\Theta}^{(t+1)}$ from $P(\boldsymbol{\Theta}|\boldsymbol{\chi}_{\text{obs}}, \boldsymbol{\chi}_{\text{mis}}^{(t+1)})$.

As $t \to \infty$, convergence is observed in that $\{\boldsymbol{\chi}_{\text{mis}}^{(t)}, \boldsymbol{\Theta}^{(t)}\}$ represents a random draw from $P(\boldsymbol{\chi}_{\text{mis}}, \boldsymbol{\Theta}|\boldsymbol{\chi}_{\text{obs}})$. Validity of estimators derived from the imputed data is contingent upon the missing at random assumption (in the nomenclature of Little and Rubin 2020).

Gibbs sampling (Geman and Geman 1984) is used to update imputations within the I Step. Letting $\boldsymbol{\chi} = \{\boldsymbol{X}_1, \ldots, \boldsymbol{X}_p\}$, within the $(t+1)^{th}$ iteration, we sequentially update $\boldsymbol{X}_j^{(t)}$ for each j by replacing values that were originally missing (in $\boldsymbol{X}_j$) with draws from

$$P\Big(\boldsymbol{X}_j|\boldsymbol{X}_1^{(t+1)}, \ldots, \boldsymbol{X}_{j-1}^{(t+1)}, \boldsymbol{X}_{j+1}^{(t)}, \ldots, \boldsymbol{X}_p^{(t)}, \boldsymbol{\Theta}^{(t)}\Big),$$

which serves to create $\boldsymbol{X}_j^{(t+1)}$. In the event that $\boldsymbol{\chi}$ follows a Gaussian distribution, multivariate normal theory can be used to form of each of the above conditional models given a mean vector and covariance matrix extracted from $\boldsymbol{\Theta}^{(t)}$. However, joint modeling in this manner for more general data, which may contain binary, unordered categorical, or ordinal variables, is more complicated. Elaborating, one can construct a joint model via a sequence conditional model using

$$P(\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_p | \boldsymbol{\Theta}) = \prod_{j=1}^{p} P(\boldsymbol{X}_j | \boldsymbol{X}_1, \ldots, \boldsymbol{X}_{j-1}, \boldsymbol{\theta}_j^*),$$

where $\boldsymbol{\Theta} = \{\boldsymbol{\theta}_1^*, \ldots, \boldsymbol{\theta}_p^*\}$. Given the specific marginal structure of each $\boldsymbol{X}_j$, models for

$$P(\boldsymbol{X}_j | \boldsymbol{X}_1, \ldots, \boldsymbol{X}_{j-1}) \tag{1}$$

and $\boldsymbol{\theta}_j^*$ may be easily determined for $j = 1, \ldots, p$, which yields a valid joint density. Nonetheless, sampling from

$$P(\boldsymbol{X}_j | \boldsymbol{X}_1, \ldots, \boldsymbol{X}_{j-1}, \boldsymbol{X}_{j+1}, \ldots, \boldsymbol{X}_p) \tag{2}$$

for $j = 1, \ldots, p$ in a manner that is congenial with the resulting joint density, as is required for Gibbs sampling, often presents an intractable (or computationally infeasible) problem for high-dimensional data of a general structure.

FCS circumvents the above problem by modeling each conditional expression of the form in (2) instead of addressing the joint distribution. As such, in lieu of a P Step, FCS samples model parameters for each conditional model within each phase of the Gibbs sampling. That is, for each $j = 1, \ldots, p$, imputations for $\boldsymbol{X}_j$ at the $(t+1)^{th}$ iteration are determined via

$$\boldsymbol{\theta}_j^{(t+1)} \sim P\Big(\boldsymbol{\theta}_j | \boldsymbol{X}_1^{(t+1)}, \ldots, \boldsymbol{X}_{j-1}^{(t+1)}, \boldsymbol{X}_j^{(t)}, \ldots, \boldsymbol{X}_p^{(t)}\Big),$$

$$\boldsymbol{X}_j^{(t+1)} \sim P\Big(\boldsymbol{X}_j | \boldsymbol{X}_1^{(t+1)}, \ldots, \boldsymbol{X}_{j-1}^{(t+1)}, \boldsymbol{X}_{j+1}^{(t)}, \ldots, \boldsymbol{X}_p^{(t)}, \boldsymbol{\theta}_j^{(t+1)}\Big),$$

where $\boldsymbol{\theta}_j$ indicates model parameters for the density seen in (2). Since the sequence conditional expressions given by (2) may define an incoherent joint distribution when modeled separately, there is no guarantee that $\{\boldsymbol{\chi}_{\text{mis}}^{(t)}, \boldsymbol{\Theta}^{(t)}\}$ will converge to $P(\boldsymbol{\chi}_{\text{mis}}, \boldsymbol{\Theta} | \boldsymbol{\chi}_{\text{obs}})$ across iterations with FCS; in fact, divergence is possible. Most references that discuss convergence in FCS methods (e.g., White et al. 2011; Van Buuren 2018) recommend the use of a small number of iterations of MCMC (usually as low as five, which is the default in several algorithms), perhaps to hedge against the possibility of divergence.

Researchers have noted performance issues with FCS when applied in high-dimensional datasets (e.g., Loh et al. 2019); nonetheless, it has observed prevalent usage when applied in a large-scale survey (e.g., Schenker et al. 2006).

4. A JOINT IMPUTATION ALGORITHM FOR DATA OF GENERAL STRUCTURES

Here, we introduce a novel imputation method which is designed to accomplish the following:

(1) Sample imputations from a coherent joint distribution;
(2) Have the flexibility to impute variables of a variety of structures (e.g., continuous, binary, unordered categorical, ordinal);
(3) Afford the user the ability to determine which conditional relationships are permitted within the imputation model;
(4) Be computationally feasible and efficient in high-dimensional datasets.

In light of the above, the procedure is referred to as General Efficient Regression-Based Imputation with Latent (GERBIL) processes from here out.

In the development of the GERBIL method, we revisit the data augmentation framework, but instead of assuming that the latent process z (as described at the beginning of section 3) represents only the missing data whereas the other process y is the observed data, we assume that there is a latent data system that underpins all data values (observed or missing) and that the collected data instead represent available knowledge regarding this system in that some variables may be fully or partially observed.

4.1 Defining the Latent Process

As in section 3, let $\chi = \{X_1, \ldots, X_p\}$ denote that collected data (which may contain missing values). We assume that each variable in χ has either a continuous, categorical, binary, or ordinal distribution. Extensions involving semi-continuous data and right-censored data are discussed in section 7. For simplicity, we assume that binary variables take on value 0 or 1, and we assume that if X_j is unordered categorical or ordinal with $k_j > 2$ possible values, then $X_j \in \{1, \ldots, k_j\}$. We reformat the data so that if X_j is unordered categorical, it is represented by $k_j - 1$ nested binary variables. However, missingness is imposed in a nested binary variable for cases where the categorical variable was observed to fall into a category antecedent to the one corresponding to the that binary variable. To elaborate, a categorical variable X_j is reformatted into variables $X^*_{j'}, \ldots, X^*_{j'+k_j-2}$ for some index j' as follows:

$$X^*_{j'+\ell-1} = \begin{cases} ?, & \text{if } X_j < \ell \text{ or } X_j = ?, \\ 1, & \text{if } X_j = \ell, \\ 0, & \text{if } X_j > \ell, \end{cases} \tag{3}$$

for $1 \leqslant \ell \leqslant k_j - 1$ where "?" indicates a missing value. All ?s in $X^*_{j'}, \ldots, X^*_{j'+k_j-2}$ are imputed. Let $\chi^* = \{X^*_1, \ldots, X^*_q\}$ denote the (expanded) reformatted data, where $q \geqslant p$ and where $\boldsymbol{\chi}^*$ contains only continuous, binary, and ordinal variables. Note that variables that are not unordered categorical are copied over from $\boldsymbol{\chi}$ to $\boldsymbol{\chi}^*$. For an unordered categorical variable $\boldsymbol{X}_j$, we suggest ordering the categories from least to most prevalent when creating the nested variables; this will minimize the number of missing values that are artificially imposed.

The formulation in (3) represents a nested version of the manner in which semicontinuous (i.e., mixed discrete/continuous) data are frequently handled in imputation algorithms (Robbins et al. 2013). Specifically, the categorical variable is first broken down into two variables: (1) a binary variable that indicates whether or not the original variable falls into the first category and (2) a categorical variable that is set as the value of the original variable but is missing when the original variable falls into the first category. Next, this second (categorical) variable is dissected in a similar manner—this yields a second binary variable that is unity when the original variable fell into the second category, missing when it fell into the first, and zero otherwise, along with a third variable that is unordered categorical and contains missing values for cases where the original categorical variable fell into one of the first two categories. This process is repeated until all categories are embodied by nested binary variables. The advantage of this process is that it allows the nested variables to be (conditionally) independent of one another and is easily reversed following imputation.

Borrowing from the idea of probit modeling, akin to how it has been previously applied in imputation settings (Carpenter and Kenward 2012), we assume that a multivariate Gaussian distribution underpins $\boldsymbol{\chi}^*$. Specifically, $\boldsymbol{\psi} = \{\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_q\}$ indicates the underlying latent process. We assume that $\boldsymbol{\psi} \sim N_q(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ for a mean vector $\boldsymbol{\mu}$ and variance matrix $\boldsymbol{\Sigma}$. The process of observed data $\boldsymbol{\chi}^*$ is generated from the latent process $\boldsymbol{\psi}$ as follows:

If $\boldsymbol{X}_j^*$ is continuous,

$$\boldsymbol{X}_j^* = F_j^{-1}(\Phi(\boldsymbol{Z}_j)), \tag{4}$$

where $F_j(\cdot)$ is the marginal cumulative distribution function (CDF) of $\boldsymbol{X}_j^*$, in that $F_j(x) = \Pr(\boldsymbol{X}_j \leqslant x)$ where $\Pr(A)$ gives the probability of event A, and where $\Phi(\cdot)$ denotes the CDF of a standard normal random variable. Of course, prior to imputation, the observed data should be transformed to have a standard normal distribution via the inverse transformation $\boldsymbol{Z}_j = \Phi^{-1}(F_j(\boldsymbol{X}_j^*))$. Transformations of this type may be performed with a parametric density (e.g., Robbins and White 2011; Robbins et al. 2013) or in a nonparametric manner with a kernel or empirical distribution (Robbins 2014). This formulation serves to link the continuous data via a Gaussian copula (Nelsen 2009).

If $\boldsymbol{X}_j^*$ is binary, a probit-type model is imposed:

$$\boldsymbol{X}_j^* = \begin{cases} 0, & \text{if } \boldsymbol{Z}_j < 0, \\ 1, & \text{if } \boldsymbol{Z}_j \geqslant 0. \end{cases}$$

Lastly, if $\boldsymbol{X}_j^*$ is ordinal where $\boldsymbol{X}_j^* \in \{1, 2, \ldots, k_j\}$,

$$\boldsymbol{X}_j^* = i \quad \text{if } \tau_{j,i-1} < \boldsymbol{Z}_j \leqslant \tau_{j,i},$$

for $i \in \{1, \ldots, k_j\}$, where $\tau_{j,i} = \Phi^{-1}(\Pr(\boldsymbol{X}_j^* \leqslant i))$ for $i \in \{1, \ldots, k_j - 1\}$ and where we set $\tau_{j,0} = -\infty$ and $\tau_{j,k_j} = \infty$.

Note that the latent multivariate normal process can be modeled conditionally upon a set of fully observed predictors; these variables can obey any distribution and need not be underpinned by a normal density. For simplicity, we do not condition on such variables here.

4.2 Imputation of the Latent Process

The P Step of GERBIL builds upon ideas presented in Robbins et al. (2013), which addressed missingness in continuous variables. The objective of the P Step is to determine values of the mean vector $\boldsymbol{\mu}$ and variance matrix $\boldsymbol{\Sigma}$ of the latent multivariate Gaussian process (Robbins et al. 2013); however, these quantities are modeled indirectly. Specifically, we build a joint model for $\boldsymbol{\psi}$ by stating linear forms for conditional models seen in (1) in that $\boldsymbol{Z}_j$ is allowed to depend on variables that precede it in sequence but not those that antecede it. That is, we assume

$$\boldsymbol{Z}_j = \boldsymbol{V}_j \boldsymbol{\beta}_j + \sigma_j \boldsymbol{\varepsilon}_j, \tag{5}$$

for $j = 1, \ldots, q$, where $\boldsymbol{V}_j$ denotes an $n \times \kappa_j$ predictor matrix of which the columns are some subset of the columns of $\{1; \boldsymbol{Z}_1; \ldots; \boldsymbol{Z}_{j-1}\}$, with 1 indicating a vector of ones, and where $\boldsymbol{\beta}_j$ denotes a length-κ_j vector of regression coefficients—the flexibility to selectively reduce the size of the predictor set for each conditional model is crucial in our setting as referenced previously. Also, $\boldsymbol{\varepsilon}_j$ is mean-zero Gaussian noise, and σ_j is a positive scalar. This model imposes that $P(\boldsymbol{Z}_j | \boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_{j-1}) = P(\boldsymbol{Z}_j | \boldsymbol{V}_j)$, in that the conditional distribution is independent of terms excluded from $\boldsymbol{V}_j$. Note that the predictor matrix $\boldsymbol{V}_j$ for a $\boldsymbol{Z}_j$ that corresponds to a nested binary variable within an unordered categorical variable should exclude any other nested variables from that same categorical variable. Likewise, in accordance with the skip logic seen within HRBS data, the predictor list for child questions should exclude the respective parent question.

Using a noninformative prior for $\boldsymbol{\Theta} = \{\boldsymbol{\beta}_1, \sigma_1, \ldots, \boldsymbol{\beta}_q, \sigma_q\}$ in that $P(\boldsymbol{\Theta}) \propto \prod_{j=1}^{q} 1/\sigma_j^2$, the posterior distributions of $\boldsymbol{\beta}_j$ and σ_j^2 (given fully observed $\boldsymbol{\psi}$) are derived as follows. If $\boldsymbol{X}_j^*$ is binary, we fix $\sigma_j^2 = 1$, which is in accordance with traditional probit modeling. Otherwise,

$$\sigma_j^2 | \boldsymbol{\psi} \sim \text{Inv-}\chi^2(n - \kappa_j, s_j^2), \tag{6}$$

where letting the superscript T indicate a matrix transpose, $s_j^2 = (\boldsymbol{Z}_j - \boldsymbol{V}_j\hat{\boldsymbol{\beta}}_j)^T (\boldsymbol{Z}_j - \boldsymbol{V}_j\hat{\boldsymbol{\beta}}_j)/(n - \kappa_j)$ with $\hat{\boldsymbol{\beta}}_j = (\boldsymbol{V}_j^T\boldsymbol{V}_j)^{-1}\boldsymbol{V}_j^T\boldsymbol{Z}_j$ and with Inv-$\chi^2(\cdot, \cdot)$ denoting an inverse chi-square distribution. Likewise,

$$\boldsymbol{\beta}_j | \sigma_j^2, \boldsymbol{\psi} \sim \mathrm{N}_{\kappa_j}(\hat{\boldsymbol{\beta}}_j, \sigma_j^2(\boldsymbol{V}_j^T\boldsymbol{V}_j)^{-1}). \tag{7}$$

Given imputed values of the latent process, $\boldsymbol{\psi}^{(t)} = \{\boldsymbol{Z}_1^{(t)}; \ldots; \boldsymbol{Z}_q^{(t)}\}$ at the t^{th} iteration, the P Step involves sampling $\boldsymbol{\beta}_j^{(t)}$ and $\sigma_j^{(t)}$ from $P(\boldsymbol{\beta}_j, \sigma_j | \boldsymbol{Z}_1^{(t)}, \ldots, \boldsymbol{Z}_{j-1}^{(t)})$ for $j = 1, \ldots, q$ in accordance with (6), when needed, and (7) above.

We next calculate $\boldsymbol{\mu}^{(t)}$ and $\boldsymbol{\Sigma}^{(t)}$, the mean vector and covariance matrix of the process $\boldsymbol{\psi}$ at the t^{th} iteration, from the parameter set $\{\boldsymbol{\beta}_1^{(t)}, \sigma_1^{(t)}, \ldots, \boldsymbol{\beta}_q^{(t)}, \sigma_q^{(t)}\}$; section A.1 of the supplementary data online provides illustration of such calculations.

The I Step for the $(t+1)^{th}$ of GERBIL involves sampling $\boldsymbol{\psi}^{(t+1)}$ from $P(\boldsymbol{\psi} | \boldsymbol{\chi}_{\text{obs}}^*, \boldsymbol{\mu}^{(t)}, \boldsymbol{\Sigma}^{(t)})$, where $\boldsymbol{\chi}_{\text{obs}}^*$ includes the fully and partial observed information regarding $\boldsymbol{\psi}$ from $\boldsymbol{\chi}^*$. Since $\boldsymbol{\chi}^*$ is uniquely determined from $\boldsymbol{\psi}$, we do not need to recalculate $\boldsymbol{\chi}^*$ at each iteration to align with the data augmentation framework. First, we use $\boldsymbol{\mu}^{(t)}$ and $\boldsymbol{\Sigma}^{(t)}$ to find the parameters that define $P(\boldsymbol{Z}_j | \boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_{j-1}, \boldsymbol{Z}_{j+1}, \ldots, \boldsymbol{Z}_p)$ for each $j = 1, \ldots, q$, which is Gaussian since $\boldsymbol{\psi}$ multivariate normal. We execute Gibbs sampling from this distribution. For each $j \in \{1, \ldots, q\}$, let

$$\mu_{j|\cdot}^{(t+1)} = E\Big[\boldsymbol{Z}_j | \boldsymbol{Z}_1^{(t+1)}, \ldots, \boldsymbol{Z}_{j-1}^{(t+1)}, \boldsymbol{Z}_{j+1}^{(t)}, \ldots, \boldsymbol{Z}_p^{(t)}, \boldsymbol{\mu}^{(t)}, \boldsymbol{\Sigma}^{(t)}\Big],$$

$$\sigma_{j|\cdot}^{(t+1)} = \mathrm{Var}\Big(\boldsymbol{Z}_j | \boldsymbol{Z}_1^{(t+1)}, \ldots, \boldsymbol{Z}_{j-1}^{(t+1)}, \boldsymbol{Z}_{j+1}^{(t)}, \ldots, \boldsymbol{Z}_p^{(t)}, \boldsymbol{\mu}^{(t)}, \boldsymbol{\Sigma}^{(t)}\Big).$$

Multivariate normal theory is used to determine $\mu_{j|\cdot}^{(t+1)}$ and $\sigma_{j|\cdot}^{(t+1)}$. Details are provided in section A.1 of the supplementary data online.

If $\boldsymbol{X}_j^*$ is continuous:

- For cases where $\boldsymbol{X}_j^*$ is observed, set $\boldsymbol{Z}_j^{(t+1)} = \boldsymbol{Z}_j$;
- For cases where $\boldsymbol{X}_j^*$ is missing, sample $\boldsymbol{Z}_j^{(t+1)}$ from $\mathrm{N}(\mu_{j|\cdot}^{(t+1)}, \sigma_{j|\cdot}^{(t+1)})$.

Note that if $\boldsymbol{X}_j^*$ is binary or ordinal, only partial information is known regarding $\boldsymbol{Z}_j$, even for cases where $\boldsymbol{X}_j^*$ is observed. This information is incorporated in the sampling scheme for binary $\boldsymbol{X}_j^*$ as follows:

- For cases where $\boldsymbol{X}_j^*$ is missing, sample $\boldsymbol{Z}_j^{(t+1)}$ from $\mathrm{N}(\mu_{j|\cdot}^{(t+1)}, \sigma_{j|\cdot}^{(t+1)})$;
- For cases with $\boldsymbol{X}_j^* = 0$, draw $\boldsymbol{Z}_j^{(t+1)}$ from $\mathrm{trN}(\mu_{j|\cdot}^{(t+1)}, \sigma_{j|\cdot}^{(t+1)}, -\infty, 0)$; and
- For cases with $\boldsymbol{X}_j^* = 1$, draw $\boldsymbol{Z}_j^{(t+1)}$ from $\mathrm{trN}(\mu_{j|\cdot}^{(t+1)}, \sigma_{j|\cdot}^{(t+1)}, 0, \infty)$.

In the above, $\mathrm{trN}(\mu, \sigma^2, a, b)$ is a truncated normal distribution with mean μ, variance σ^2, and bounds of a and b. That is, $X \sim \mathrm{trN}(\mu, \sigma^2, a, b)$ implies $X \equiv (Z|a < Z < b)$ with $Z \sim N(\mu, \sigma^2)$. To find $\boldsymbol{Z}_j^{(t+1)}$ if $\boldsymbol{X}_j^*$ is ordinal with k_j categories:

- For cases where $\boldsymbol{X}_j^*$ is missing, sample $\boldsymbol{Z}_j^{(t+1)}$ from $\mathrm{N}(\mu_{j|\cdot}^{(t+1)}, \sigma_{j|\cdot}^{(t+1)})$;
- For cases with $\boldsymbol{X}_j^* = i$ where $1 \leqslant i \leqslant k_j$, draw $\boldsymbol{X}_j^{(t+1)}$ from $\mathrm{trN}(\mu_{j|\cdot}^{(t+1)}, \sigma_{j|\cdot}^{(t+1)}, \tau_{j,i-1}, \tau_{j,i})$.

Herein, we again set $\tau_{j,0} = -\infty$ and $\tau_{j,k_j} = \infty$.

To initialize the MCMC procedure, we find that setting $\mu_{j|\cdot}^{(0)} = 0$ and $\sigma_{j|\cdot}^{(0)} = 1$ and sampling $\boldsymbol{\psi}^{(0)} = \{\boldsymbol{Z}_1^{(0)}; \ldots; \boldsymbol{Z}_q^{(0)}\}$ according to the rules above performs sufficiently well. Of course, more rigorous options could be implemented.

4.3 Derivation of Final Imputations

After a burn-in period of b iterations, the MCMC procedure is stopped, and $\boldsymbol{\psi}^{(b)} = \{\boldsymbol{Z}_1^{(b)}, \ldots, \boldsymbol{Z}_q^{(b)}\}$ indicates the final imputed version of the latent data. The final imputations for the (reformatted) recorded dataset are denoted $\tilde{\boldsymbol{\chi}}^* = \{\tilde{\boldsymbol{X}}_1^*, \ldots, \tilde{\boldsymbol{X}}_q^*\}'$ and are derived from $\boldsymbol{\psi}^{(b)}$ as follows.

If $\boldsymbol{X}_j^*$ is continuous, $\tilde{\boldsymbol{X}}_j^* = F_j^{-1}(\Phi(\boldsymbol{Z}_j^{(b)}))$; see Robbins et al. (2013) and Robbins (2014) for specifics regarding transformation and untransformation of marginal distributions. If $\boldsymbol{X}_j^*$ is binary,

$$\tilde{\boldsymbol{X}}_j^* = \begin{cases} 0, & \text{if } \boldsymbol{Z}_j^{(b)} < 0, \\ 1, & \text{if } \boldsymbol{Z}_j^{(b)} \geqslant 0, \end{cases}$$

and if $\boldsymbol{X}_j^*$ is ordinal with k_j categories,

$$\tilde{\boldsymbol{X}}_j^* = i \quad \text{if } \tau_{j,i-1} < \boldsymbol{Z}_j^{(b)} \leqslant \tau_{j,i}.$$

for $i \in \{1, \ldots, k_j\}$.

The nesting structure described in (3), in which case an unordered categorical variable $\boldsymbol{X}_j \subseteq \boldsymbol{\chi}$ from the original dataset has been represented by $\{\boldsymbol{X}^*_{j'}, \ldots, \boldsymbol{X}^*_{j'+k_j-2}\} \subseteq \boldsymbol{\chi}^*$ for some j' in the expanded dataset, is then reversed. This creates the final imputed dataset $\tilde{\boldsymbol{\chi}} = \{\tilde{\boldsymbol{X}}_1, \ldots, \tilde{\boldsymbol{X}}_p\}$, which is accomplished after setting

$$\tilde{\boldsymbol{X}}_j = \begin{cases} 1, & \text{if } \boldsymbol{X}^*_{j'} = 1, \\ 2, & \text{if } \boldsymbol{X}^*_{j'+1} = 1 \text{ and } \boldsymbol{X}^*_{j'} = 0, \\ \vdots & \\ k_j - 1, & \text{if } \boldsymbol{X}^*_{j'+k_j-2} = 1 \text{ and } \boldsymbol{X}^*_i = 0 \text{ for each } i \in \{j', \ldots, j' + k_j - 3\}, \\ k_j, & \text{if } \boldsymbol{X}^*_i = 0 \text{ for each } i \in \{j', \ldots, j' + k_j - 2\}, \end{cases}$$

for all categorical $\boldsymbol{X}_j$ and setting other variables contained in $\boldsymbol{\chi}$ equal to their corresponding imputed version in $\tilde{\boldsymbol{\chi}}^*$.

To apply multiple imputation (Rubin 1987, 1996), the entire process illustrated above is repeated independently m times to procedure m separately imputed datasets. Well known combining rules are used to pool the datasets and adjust estimators for imputation error.

Note that the marginal transformations that are applied to continuous variables in (4) assume that $F_j(x) = \Pr(\boldsymbol{X}_j \leqslant x)$ is known for each relevant j and likewise that $\tau_{j,i} = \Phi^{-1}\{\Pr(\boldsymbol{X}^*_j \leqslant i)\}$ is assumed known for each ordinal $\boldsymbol{X}_j$. In practice, these quantities are estimated which may induce bias into the transformations in missingness mechanisms that are not missing completely at random (borrowing the terminology of Little and Rubin 2020). However, earlier studies involving continuous data (Robbins et al. 2013; Robbins 2014) find no evidence of substantial bias stemming from transformations. Note also that the copula framework applied to continuous variables requires that following the marginal transformations, the transformed variables obey a multivariate normal distribution (i.e., relationships between variables are linear). The aforementioned studies have also shown that in practice, bivariate relationships are often more linear following such transformations than before.

The manner in which we handle categorical variables is, to our knowledge, novel. Alternative approaches proposed by other authors do not impose missingness in nested variables (Allison 2002; Honaker et al. 2011; Carpenter and Kenward 2012)—imputed values of the categorical variable are then set as the category that observes the highest value among the imputed nested variables. However, rigorous evaluations of this approach are scarce, as noted by Carpenter and Kenward (2012). In contrast, our proposed approach performs well in simulations (see section 5).

Note that sweep operator (Beaton 1964; Goodnight 1979) is used to dramatically improve the computational efficiency of the GERBIL algorithm in both the P Step and I Step. Further details on the use of this operation are provided in section A.2 of the supplementary data online.

4.4 Advantages over Existing Methods

GERBIL applies joint modeling which avoids the theoretical issues encountered with FCS and guarantees that GERBIL imputations will converge across iterations of MCMC. That is, the use of joint modeling gives GERBIL a strong advantage over all implementations of FCS (e.g., `mice`, `mi`, IVEware) regardless of the conditional model used for imputation. Furthermore, strategic use of the sweep operator in GERBIL ensures that it may be more computationally efficient than existing FCS software. In addition, most current implementations of imputation by joint modeling (e.g., Schafer 2017; Zhao and Schafer 2018) do not facilitate general data structures.

The R package `jomo` (Carpenter and Kenward 2012; Quartagno and Carpenter 2020), which uses a latent Gaussian process to underpin noncontinuous variables, is perhaps most closely aligned with GERBIL in terms of utility, but GERBIL has a number of operational advantages over `jomo`. Specifically, `jomo` does not build the joint model from a sequence of conditional models as seen in (5) but instead directly estimates the covariance matrix. Estimation of a covariance matrix that is subject to restrictions (e.g., the diagonal elements that correspond to binary variables must be set to 1) is difficult in practice as the result may not be positive semidefinite. `jomo` addresses this issue by using a guess-and-check Metropolis-Hastings algorithm, and further applies a guess-and-check method in lieu of sampling from a truncated normal distribution. These issues lead to infeasibility of the algorithm when applied to high-dimensional, complex data such as the HRBS data studied in section 6. Lastly, `jomo` does not let its user specify dependencies (which is crucial for HRBS and similar data)—collinearities in the data may render estimation of the covariance matrix infeasible.

The GERBIL procedure provides a more natural method by which covariance matrices of the latent process can be estimated. By setting the conditional error variance of the models for binary variables to be one (instead of attempting to restrict diagonal elements of a covariance matrix to be one), we ensure that the resulting covariance matrix will be positive semidefinite and can be estimated using appropriate Bayesian techniques. Furthermore, variables can be dropped from specific conditional models in (5) while maintaining a positive semidefinite covariance matrix, enabling the user to avoid multicollinearities and impose desired conditional dependence structures.

Hoff (2018) introduces a rank-based approach to estimating parameters of a copula model that underpins general data. This method may be easily extended to impute missing data and is implemented for such purpose in the R package `sbgcop` (Hoff 2018). This approach is theoretically similar to that of `jomo` and as such contains some of the same drawbacks (e.g., lack of flexibility regarding which dependencies are enabled, which may lead to its use being infeasible in high-dimensional data from complex surveys such as HRBS). Unlike `jomo`, however, `sbgcop` circumvents the need to restrict variances that correspond to binary variables to unity through sampling latent data via a correlation matrix (although covariances are indeed estimated through the Gibbs sampling process). Additionally, `sbgcop` does not directly enable the imputation of unordered categorical variables. Murray and Reiter (2016) introduce a joint modeling procedure for mixed data based on Dirichlet process mixtures (this method is implemented within the R package `MixedDataImpute`), and the R packages `mdmb` (Grund et al. 2021) and `JointAI` (Erler et al. 2021) develop imputations via a joint model for general data under the guise of a specific analysis model of interest. However, these procedures have the same practical drawbacks as `jomo` and `sbgcop` and lack computational feasibility in data of high dimensions. Procedures based upon deep learning have been developed (Gondara and Wang 2018; Yoon et al. 2018), but these have been shown to underperform in comparison to FCS methods (Wang et al. 2022).

5. SIMULATIONS

In this section, we perform a simulation study to evaluate the effectiveness of GERBIL and compare its performance to that of several existing procedures. It is not computationally feasible to perform simulations with data that mimic HRBS due to the dimensionality and complexity of those data. We instead perform simulations using smaller synthetic datasets in hopes of showing GERBIL performs comparably to existing procedures that fail when applied to HRBS. The synthetic data are not designed to favor any particular method(s) but are instead designed to be general and applicable for all methods. Since the performance of an imputation algorithm is dependent upon compatibility between the model used for imputation and those used for analysis (e.g., Robbins and White 2014; Grund et al. 2021), we will consider a wide variety of models for analysis (including those which are more complicated than models used for the analysis of HRBS data) in order the more fully assess the capabilities of the GERBIL algorithm.

First, we generate a dataset that contains six variables with differing marginal structures, loosely outlined as follows:

- X_1—Unordered categorical;
- X_2—Continuous (fully observed);
- X_3—Continuous;

- X_4—Binary (generated from a probit-type model);
- X_5—Ordinal (generated from a probit-type model); and
- X_6—Binary (generated from a logistic model).

Elaborating, X_1 is generated from a multinomial distribution with 4 categories. A latent process that underpins $X_2 \ldots, \ldots X_5$ is generated from a multivariate normal distribution with while conditioning on X_1. Further, X_6 is generated from a logistic model conditional on $X_1 \ldots, \ldots X_5$. Nonnegligible associations exist between all variables. We generate $n = 2,\ 000$ observations of each variable.

Missingness is stochastically imposed in the synthetic data using the following three mechanisms. In each case, around a third of the observations are missing (excluding X_2).

(1) MCAR: Missingness probabilities are independent of any other data characteristics.
(2) MAR: Missingness probabilities depend upon only the fully observed variable X_2.
(3) NMAR: Missingness probabilities in variable X_j depend upon only X_j for $j \in \{1, 3, \ldots, 6\}$.

These mechanisms are designed in line with the nomenclature of Little and Rubin (2020). Further details on the data generating and missingness mechanisms are provided in section A.5 of the supplementary data online. Note that missingness rates in each variable (with the exception of X_2) are approximately 33 percent under each mechanism.

Next, the missing values are imputed using six distinct methods, three of which utilize FCS, whereas the others implement joint modeling. Specifically, comparisons to FCS are performed using the implementation available in the R package `mice` (Van Buuren and Groothuis-Oudshoorn 2011). Within `mice`, one can assign different methods of imputation to each variable, with Gaussian imputation available for continuous variables, logistic regression for binary variables, and polytomous regression for categorical variables. `mice` also implements predictive mean matching (PMM, Little 1988), which uses a nearest neighbor-type approach based on a predictive model and is often applied to handle continuous variables that may have non-Gaussian marginal distributions, as well as classification tress and random forests. These techniques can also be applied to binary, unordered categorical, and ordinal variables within `mice`. We also compare against the R packages `jomo` and `sbgcop`, both of which employ joint modeling (as described in section 4.4). In summary, the various methods used for imputation in the simulations are:

(1) sbgcop: The `sbgcop` package is used for imputation (1.1 seconds per 100 iterations). The categorical variable is handled in accordance with (3).

(2) jomo: The `jomo` package is used for imputation (2 seconds per 100 iterations when applied within this simulation setting).
(3) HCMM: The hierarchically coupled mixture model procedure from the R package `MixedDataImpute` is used for imputation (4.1 seconds per 100 iterations).
(4) Logistic: `mice` is used with logistic regression for binary variables, polytomous regression for categorical variables, ordered logistic regression for ordinal variables, and Gaussian imputation for continuous variables (20 seconds per 100 iterations).
(5) PMM: `mice` is used with PMM for all variables (3 seconds per 100 iterations).
(6) CART: `mice` with imputation by classification trees is performed for all variables (1 minute per 100 iterations).
(7) GERBIL: General Efficient Regression-Based Imputation with Latent processes as proposed in section 4 (3.8 seconds per 100 iterations). An empirical distribution transformation (Robbins 2014) is applied to continuous variables.

We also considered `mice` with random forests (eight minutes per 100 iterations), but due to its computational burden, it was excluded from the larger simulation study. Abbreviated simulations show it does not perform as well as the other `mice` methods. The computing times listed are performed on a Windows machine with a 2.8 GHz processor and 32.0 GB of RAM. Due to its use of the SWEEP operator, GERBIL will improve in computational efficiency in comparison to the `mice` methods as the dimensionality of the data increases. Comparisons to `JointAI` (1.5 minutes per 100 iterations) may also be informative (in particular, if a single analysis model is of interest) but are excluded here for brevity—however, in simulations not shown, `JointAI` performs comparably to the other existing joint modeling procedures.

We use 15 iterations of MCMC for the `mice` methods, 60 iterations for GERBIL, `jomo`, and HCMM, and 120 iterations for `sbgcop`; more iterations of the non-`mice` methods are used because of their relative computational efficiency and because `mice` is shown to converge somewhat quicker in the setting of these simulations. All possible inter-variable dependencies are enabled for the `mice` methods and GERBIL. To adjust for imputation error, we use multiple imputation (Rubin 1987, 1996) with $m = 40$ independently imputed datasets for each method. This selection of m is in line with the recommendations of Graham et al. (2007).

We use $N = 5{,}000$ replications for this simulation study—that is, the above process of simulating and imputing data is repeated independently 5,000 times. The following parameters are tracked in each replication for each method:

- Means and the variance-covariance matrix of the dataset $\{\boldsymbol{X}_{1,1}, \ldots, \boldsymbol{X}_{1,4}, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_6\}$ where the $\boldsymbol{X}_{1,k}$ for $k \in \{1, \ldots, 4\}$ are categorical indicators underpinning $\boldsymbol{X}_1$ (although the mean and variance of $\boldsymbol{X}_2$ are excluded). There are 8 mean parameters calculated with 8 variances and 36 covariances.

- Estimated regression coefficients, and standard errors of those coefficients, for all fully specified conditional models of the form $P(\boldsymbol{X}_j|\boldsymbol{X}_1 \ldots, \boldsymbol{X}_{j-1}, \boldsymbol{X}_{j+1}, \boldsymbol{X}_6)$ for $j \in \{1, \ldots, 6\}$. For continuous and ordinal variables, we fit a basic linear model. For binary variables, we fit a logistic regression, and for the categorical variable, we fit a multinomial log-linear model via the nnet package in R (Venables and Ripley 2002). There are 58 regression parameters tabulated with 58 standard errors on those parameters.

We calculate root mean square error (rMSE) for all parameters and coverage rates for a subset of parameters.

We let $\hat{\theta}^{[r]}(x)$ denote the value of a parameter θ estimated at the r^{th} replication for imputation method x ($\hat{\theta}^{[r]}(x)$ is calculated as the average of separate estimates of θ produced for each of the multiply imputed datasets). For method x, we calculate the rMSE in the estimate of θ as follows:

$$\text{rMSE}_\theta(x) = \sqrt{\frac{1}{N}\sum_{r=1}^{N}[\hat{\theta}^{[r]}(x) - \theta]^2}.$$

The rMSE is calculated for all parameters listed above.

We compare the rMSE of GERBIL to the rMSE of each of the competing methods. Table 1 shows the portion of the 168 parameters for which GERBIL yields the better (i.e., smaller) rMSE for each method under each missingness mechanism. We see that in all cases, GERBIL performs better for a majority of the parameters. The exact rMSE seen for each of the six methods under all missingness mechanisms is reported in the tables seen in section A.7 of the supplementary data online.

We next consider the accuracy of the interval estimates produced using multiple imputation for each of the methods. That is, if θ is the mean of a variable or a regression coefficient, we use Rubin's combining rules (Rubin 1987) across the multiply imputed datasets to approximate the variance of $\hat{\theta}^{[r]}(x)$, which we denote $T^{[r]}(x)$ at the r^{th} replication. Then, for these parameters, we calculate the coverage of a $(1-\alpha)$ percent confidence interval around θ as $N^{-1}\sum_{r=1}^{N} C_\theta^{[r]}(x)$ where

$$C_\theta^{[r]}(x) = \begin{cases} 1, & \text{if } \theta \in \{\hat{\theta}^{[r]}(x) \pm t_{1-\alpha/2, d^{[r]}}\sqrt{T^{[r]}(x)}\}, \\ 0, & \text{otherwise,} \end{cases}$$

and where $t_{\alpha,\nu}$ is the $100\alpha^{th}$ percentile of a t distribution with ν degrees of freedom (where the degrees of freedom at the r^{th} replication, $d^{[r]}$, are calculated from the within- and between-imputation variances).

Table 1. The Portion of the 168 Parameters for Which the rMSE for the Respective Method in the Respective Missingness Mechanism Is Greater Than the rMSE Yielded by GERBIL

	sbgcop	jomo	HCMM	Logistic	PMM	CART
MCAR	0.643	0.613	0.601	0.595	0.601	0.738
MAR	0.690	0.661	0.649	0.589	0.655	0.762
NMAR	0.589	0.595	0.625	0.619	0.619	0.708

Box plots across the 66 parameters for which the coverage rates were calculated are shown in figure 2 for each method and missingness mechanism. The estimated rates approximate the coverage of a 95 percent confidence interval for the parameters. NMAR results are excluded from the figure since all methods provide poor coverage under NMAR missingness and those results do not further inform the comparative performance of the methods.

Figure 2 shows that GERBIL systematically provides estimated coverage that is close to 95 percent. The HCMM, Logistic, and PMM methods perform reasonably well; however, the other methods fail to yield reliable coverage. Exact rates of coverage are reported in tables provided in section A.7 of the supplementary data online.

In summary, GERBIL met our aspiration of performing no worse than the procedures the existing procedures within our simulation study. In fact, GERBIL was shown to outperform those methods in several regards.

6. IMPUTATION OF HRBS DATA

This section details imputation of HRBS data, including imputations created with GERBIL and comparisons to imputations created using existing methods when feasible.

The skip logic structure to HRBS data, as discussed in section 2, is addressed as follows for imputation. Regardless of the imputation method used, all child questions are imputed for all cases, including cases for which the respective question was legitimately skipped. A postimputation editing process determined which imputed values should be overwritten as legitimate skips (as is needed for cases in which the parent question was missing). That is, if a respondent had an imputed value on a parent question that indicated no deployment in the prior year, a nonzero imputed value of the child question of how long the deployment lasted was retained and “cleaned” later so that the parent–child questions were consistent. In contrast, if the imputed value of the parent question indicated that the respondent did not deploy, the child question was marked as a skip.

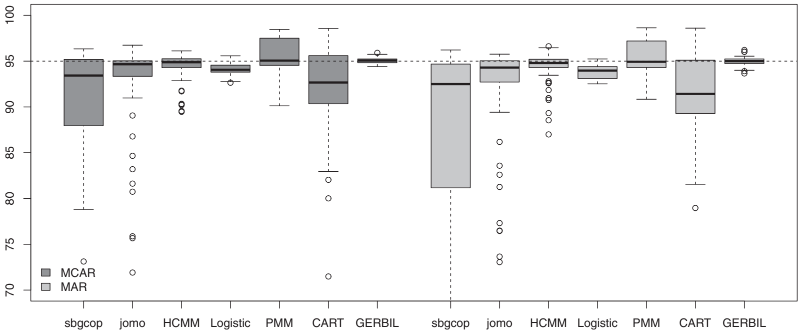

Figure 2. Boxplots of the Simulated Coverage Rates for the 95 Percent Confidence Intervals of 66 Separate Parameters under Various Methods and Missingness Mechanisms.

When using a method that builds conditional models (and thereby has the flexibility to select predictors for those models), including GERBIL and `mice`, the imputation model for a child question is not allowed to include a parent question (and vice versa) to avoid collinearity issues. Similarly, each conditional model is reduced as needed for sparsely distributed variables so that the number of predictors is never greater than the number of affirmative observed responses.

As noted in section 2, the `jomo`, `sbgcop`, and HCMM algorithms were unable to produce imputations across the full HRBS dataset. This is likely a consequence of those methods being unable to allow its user to specify dependence structures; that is, they cannot account for collinearity issues created when estimating covariances between parent and child questions. Furthermore, there are concerns about the computational efficiency of those methods when applied to data of the dimensionality of HRBS. We also applied `mice` to HRBS data with several different approaches for estimating the conditional models. Note that we were unable to produce imputations of HRBS data using `mice` with imputation by classification and regression trees, linear discriminant analysis, and random forests—this is likely a consequence of computational issues induced by the dimensionality and/or sparsity of the data.

In all, we were able to generate imputations for the full HRBS dataset using three methods: logistic (`mice` with logistic and polytomous regression for binary and categorical variables and PMM for continuous variables), PMM (`mice` with PMM for all variables), and GERBIL. All methods implement conditional models that align with the requirements noted above; the `mice` methods do so with expressions in the form of (2), whereas GERBIL builds models based on (1). For all methods, the variables are indexed in the order in which they appear on the survey. In accordance with existing literature (e.g., Schenker et al. 2006), survey weights are included as a covariate in the imputation model. After adding binary indicators for parent questions and including

14 fully observed supplementary (i.e., nonsurvey) variables, the imputation model contains 284 variables, of which 27 are modeled as continuous, 186 as binary, 52 as ordinal, and 19 as unordered categorical. Consequentially, the expanded dataset used to model the latent process has 346 variables.

GERBIL is more computationally efficient when applied to HRBS data than `mice`. When run on a Windows machine with a 2.8-GHz processor and 32.0 GB of RAM, one iteration of GERBIL takes two minutes with HRBS data, whereas one iteration of `mice` with logistic regression takes one hour. Although recent versions of `mice` have improved the computational efficiency of PMM, GERBIL is also approximately five times faster than `mice` with PMM when applied to HRBS data. As such, we ran no more than 250 iterations for both of the `mice` methods—this proved sufficient for illustrating issues with the performance of `mice`. Furthermore, we ran GERBIL with up to 5,000 iterations. Due to the computational intensity of all procedures, most of our diagnostics involve one imputed dataset; that is, we do not consider multiple imputation, although multiple chains are involved in the calculation of convergence diagnostics for GERBIL presented shortly.

To assess the quality of the imputations, we use a variety of metrics, beginning with marginal diagnostics. Let $\hat{\mu}_{\text{obs},j}$ denote the mean of observed values of variable j and $\hat{\mu}_{\text{imp},j}$ denote the mean of the imputed values of variable j. For each variable and for a variety of imputation methods, we calculate

$$\delta_j = \frac{|\hat{\mu}_{\text{mis},j} - \hat{\mu}_{\text{obs},j}|}{\hat{\mu}_{\text{obs},j}}. \tag{8}$$

Box plots of the values of $\log(\delta_j + 1)$ across all variables for each imputation method and for varying burn-in periods are shown in figure 3. The log transformation is used to reduce the effect of extreme values of δ_j on the visualization.

Although δ_j is an imperfect measure ($\hat{\mu}_{\text{mis},j}$ may be rightfully different from $\hat{\mu}_{\text{obs},j}$), the figure indicates issues with both versions of imputations created using `mice`. Specifically, when `mice` with logistic modeling is applied, the means of the imputed values clearly diverge from their respective means found using only observed data and should thus be considered erroneous. Sparse binary variables are most problematic in this regard. In that vein, issues with logistic regression in sparse data have been identified previously (Devika et al. 2016). For instance, the "one in ten rule" (Harrell et al. 1996; Peduzzi et al. 1996) is clearly violated in this application, as are relaxations of it (e.g., Vittinghoff and McCulloch 2007). In fact, imputation literature suggests the use of PMM in place of logistic regression in sparse data (Van Buuren 2018). However, issues are observed when PMM is used as well. Specifically, some degree of divergence is present. GERBIL, however, offers stable performance when up to 5,000 iterations are considered.

To further investigate convergence issues and to diagnose whether PMM imputations are erroneous, we focus on four representative binary variables:

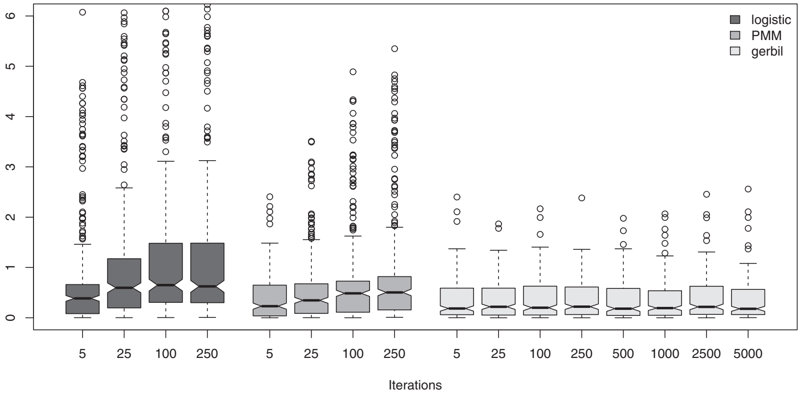

Figure 3. Box Plots of $\log(\delta_j + 1)$from (8) across All Variables for Various Imputation Method and Burn-In Periods (Iterations). The logistic and PMM methods use `mice`**.**

(1) Q22B—"In the past 12 months… I rode in a car or other vehicle driven by someone who had too much to drink" (0.56 percent of cases are imputed; of observed cases, 3.1 percent respond "yes").

(2) Q40A—"In the past 12 months, have you used marijuana or hashish?" (1.6 percent of cases are imputed; of observed cases, 0.76 percent respond "yes").

(3) Q43B—"In the past 12 months have you used… prescription sedatives, tranquilizers, muscle relaxers, or barbiturates?" (1.7 percent of cases are imputed; of observed cases, 6.9 percent respond "yes").

(4) Q97—"In the past 12 months, have you ever had to lie to people important to you about how much you gambled?" (6.4 percent of cases are imputed; of observed cases, 0.45 percent respond "yes").

Trace plots of the mean of `mice` with PMM and GERBIL imputations for these variables across the first 250 iterations of MCMC are shown in figure 4. The logistic method is omitted from the figure as those imputations are clearly erroneous and diverge quickly. Furthermore, the figure shows evidence of divergence across iterations of PMM. Discrepancies between the observed and imputed marginal distributions are not evidence in themselves that imputations are erroneous, and it remains possible that convergence will occur with more iterations. However, the values reported in the figure enter the realm of implausibility. For instance, with Q97 after 250 iterations of PMM, we see that imputed values constitute 6.4 percent of the total cases but contribute 48.0 percent of the "yes" responses (130 imputed yeses versus 141 observed yeses).

In contrast, GERBIL imputations appear to be stable across iterations of MCMC. The mean of the GERBIL imputations slightly exceeds that of the observed data; however, this is reasonable given that the items in question

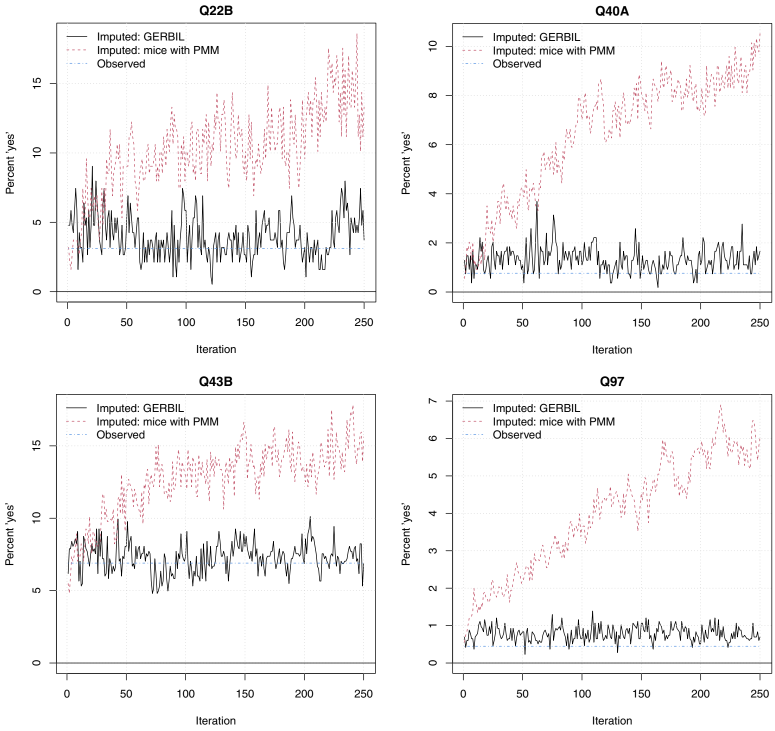

Figure 4. Trace Plots across Iterations of MCMC of the Mean (in Percent) of Imputations Created with Two Methods for Four Binary HRBS Variables.

pertain to sensitive topics and therefore those that would answer “yes” may be more likely to refuse or drop out. To diagnose whether convergence has occurred for GERBIL imputations, we calculate the $\hat{R}$ statistic of Gelman and Rubin (1992). This statistic is calculated across ten independently generated chains of imputed data with 250 iterations of MCMC each; in line with the guidance of Gelman and Rubin (1992), calculations involve only the latter half of iterations. The values of the $\hat{R}$ are 1.020, 1.005, 1.010, and 1.018 for the four variables described above. Most imputations procedures consider $\hat{R} < 1.1$ to be indicative of convergence (Su et al. 2011; Gelman et al. 2013). The $\hat{R}$ statistic was not calculated for `mice` due to its computational burden.

We have indications that the observed divergence of the PMM procedure is not a consequence of the sparsity of the data but is instead related to FCS sampling from an incoherent joint distribution. To elaborate, we ran the PMM algorithm while enabling dependencies within the conditional models that are in line with the sequential structure seen in (1) as opposed to the fully conditional models of (2). Results (omitted for brevity) indicate that divergence is no longer observed in this case. See Meadows et al. (2020a, 2020b) for further details.

Multivariate properties of the HRBS imputations as found using the various methods are shown in section A.3 of the supplementary data online. Therein, further issues with the mice imputations are illustrated, and the GERBIL imputations are shown to be of high quality. We also consider posterior predictive *p*-values (He et al. 2010; Burgette and Reiter 2010) as a diagnostic tool. Results for these evaluations are shown in section A.4 of the supplementary data online.

7. DISCUSSION

To summarize, HRBS data were shown to present a unique and significant challenge for imputation algorithms. However, the proposed GERBIL method was demonstrated to be able to accomplish what existing methods could not: it created high-quality imputations of the HRBS data, and as such, this important dataset can be analyzed without fear of the imputations biasing the findings.

GERBIL is able to accomplish what was previously infeasible by satisfying the objectives stated at the beginning of section 4. Specifically, imputations are sampled from a coherent joint distribution in that the data augmentation framework of Tanner and Wong (1987) is obeyed, thereby ensuring MCMC convergence. Unlike existing procedures that accomplish that objective, GERBIL is also easily applied in large datasets. To elaborate, existing general imputation procedures (e.g., Van Buuren and Groothuis-Oudshoorn 2011; Su et al. 2011; Quartagno and Carpenter 2020) are usually computationally onerous or simply inoperable when applied to high-dimensional data. However, the proposed GERBIL method was upward of 30 times faster than `mice` when applied in the HRBS data example provided here, and `jomo` procedure was not able to process HRBS data. The need for efficient imputation algorithms with high-dimensional data is amplified by the fact that big data are becoming increasingly prevalent and that studies have shown the need for exhaustive variable selection when building imputation models (Robbins and White 2014).

Multilevel modeling is an important tool that has been well-studied within the imputation framework. Approaches for FCS are found within the `mice` package. Furthermore, approaches for multilevel models with continuous data have been developed (Schafer and Yucel 2002; Yucel 2008; Goldstein et al. 2009) and extended for use with mixed data within the `jomo` and `JointAI` (Erler et al. 2021) packages, for example. Since a latent multivariate normal model underpins both the machinery used by both GERBIL and `jomo`, the methodology used to apply `jomo` with multilevel models could be extended for use within the GERBIL procedure.

Several authors have pointed out that bias can result in estimators involving complex survey weights when multiple imputation is used (e.g., Kott 1995;

Kim et al. 2006). Several authors have studied this issue, with Seaman et al. (2012) and others suggesting that for best results, one should use the weights, all covariates used to produce the weights and for final analysis, and their interactions should be used in the imputation model. Quartagno et al. (2020) extend these concepts by stratifying the data on the basis of the weight and use multilevel modeling for imputation which, as noted above, could be incorporated within GERBIL. For large datasets with missingness scattered throughout all or most variables (such as the HRBS), it is impractical to include all possible interactions within imputation models.

Note there is potential that (when one is selective with regard to the predictors used within the conditional models) the ordering of the variables may affect the imputations with GERBIL. As the variable ordering used in our data example was natural due to the nearly monotonic nature of the missingness, we did not explore this issue here and leave it for further work.

Additional points of discussion are seen in section A.6 of the supplemental material online.

Supplementary Materials

Supplementary materials are available online at academic.oup.com/jssam.

REFERENCES

Allison, P. D. (2002), *Missing Data. Sage University Paper Series on Quantitative Applications in the Social Sciences*. Thousand Oaks, CA: Sage Publications.

Beaton, A. E. (1964), "The Use of Special Matrix Operators in Statistical Calculus," Technical Report, Princeton, NJ: Educational Testing Service, RB-64-51, i-222.

Burgette, L. F., and Reiter, J. P. (2010), "Multiple Imputation for Missing Data via Sequential Regression Trees," *American Journal of Epidemiology*, 172, 1070–1076.

Carpenter, J., and Kenward, M. (2012), *Multiple Imputation and Its Application*, Hoboken, NJ: John Wiley & Sons.

Devika, S., Jeyaseelan, L., and Sebastian, G. (2016), "Analysis of Sparse Data in Logistic Regression in Medical Research: A Newer Approach," *Journal of Postgraduate Medicine*, 62, 26–31.

Doove, L. L., Van Buuren, S., and Dusseldorp, E. (2014), "Recursive Partitioning for Missing Data Imputation in the Presence of Interaction Effects," *Computational Statistics & Data Analysis*, 72, 92–104.

Erler, N. S., Rizopoulos, D., and Lesaffre, E. M. E. H. (2021), "Jointai: Joint Analysis and Imputation of Incomplete Data in r," *Journal of Statistical Software*, 100, 1–56.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., and Rubin, D. B. (2013), *Bayesian Data Analysis*,. CRC press.

Gelman, A. E., and Rubin, D. B. (1992), "Inference from Iterative Simulation Using Multiple Sequences (with Discussion)," *Statistical Science*, 7, 457–472.

Geman, D., and Geman, S. (1984), "Stochastic Relaxation, Gibbs Distributions, and the Bayesian Reconstruction of Images," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6, 721–741.

Goldstein, H., Carpenter, J., Kenward, M. G., and Levin, K. A. (2009), "Multilevel Models with Multivariate Mixed Response Types," *Statistical Modelling*, 9, 173–197.

Gondara, L., and Wang, K. (2018), "Mida: Multiple Imputation Using Denoising Autoencoders," in Pacific-Asia Conference on Knowledge Discovery and Data Mining, pp. 260–272, Cham, Switzerland: Springer.

Goodnight, J. H. (1979), "A Tutorial on the SWEEP Operator," *The American Statistician*, 33, 149–158.

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, 206–213.

Grund, S., Lüdtke, O., and Robitzsch, A. (2021), "Multiple Imputation of Missing Data in Multilevel Models with the r Package Mdmb: A Flexible Sequential Modeling Approach," *Behavior Research Methods*, 53, 2631–2649.

Harrell, F. E., Jr, Lee, K. L., and Mark, D. B. (1996), "Multivariable Prognostic Models: Issues in Developing Models, Evaluating Assumptions and Adequacy, and Measuring and Reducing Errors," *Statistics in Medicine*, 15, 361–387.

He, Y., Zaslavsky, A., Landrum, M., Harrington, D., and Catalano, P. (2010), "Multiple Imputation in a Large-Scale Complex Survey: A Practical Guide," *Statistical Methods in Medical Research*, 19, 653–670.

Hoff, P. (2018), "Sbgcop: Semiparametric Bayesian Gaussian Copula Estimation and Imputation," R Package Version 0.980.

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

Kim, J. K., Brick, J. M., Fuller, W. A., and Kalton, G. (2006), "On the Bias of the Multiple-Imputation Variance Estimator in Survey Sampling," *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 68, 509–521.

Kott, P. S. (1995), "A Paradox of Multiple Imputation." Technical Report, National Agricultural Statistics Service, Fairfax, VA. Presented at the Joint Statistical Meetings, August 1995, Orlando, FL.

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.

———. (2017), "Multiple Imputation in the Presence of Non-Normal Data," *Statistics in Medicine*, 36, 606–617.

Little, R. J., and Rubin, D. B. (2020), *Statistical Analysis with Missing Data* (3rd ed.), Hoboken, NJ: John Wiley & Sons.

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

Loh, W. Y., Eltinge, J., Cho, M. J., and Li, Y. (2019), "Classification and Regression Trees and Forests for Incomplete Data from Sample Surveys," *Statistica Sinica*, 29, 431–453.

Meadows, S. O., Engel, C. C., Collins, R. L., Beckman, R. L., Breslau, J., Bloom, E. L., Dunbar, M. S., Gilbert, M. L., Grant, D. M., Hawes-Dawson, J., Holliday, S. B., MacCarthy, S., Pedersen, E. R., Robbins, M. W., Rose, A. J., Ryan, J., Schell, T. L., and Simon, M. (2020a), *2018 Department of Defense Health Related Behaviors Survey (HRBS): Results for the Active Component*, Santa Monica, CA: RAND Corporation.

———. (2020b), *2018 Department of Defense Health Related Behaviors Survey (HRBS): Results for the Reserve Component*, Santa Monica, CA: RAND Corporation.

Murray, J. S., and Reiter, J. P. (2016), "Multiple Imputation of Missing Categorical and Continuous Values via Bayesian Mixture Models with Local Dependence," *Journal of the American Statistical Association*, 111, 1466–1479.

Nelsen, R. B. (2009), *An Introduction to Copulas* (2nd ed.), New York, NY: Springer.

Peduzzi, P., Concato, J., Kemper, E., Holford, T. R., and Feinstein, A. R. (1996), "A Simulation Study of the Number of Events per Variable in Logistic Regression Analysis," *Journal of Clinical Epidemiology*, 49, 1373–1379.

Quartagno, M., and Carpenter, J. (2020), "jomo: A package for Multilevel Joint Modelling Multiple Imputation." Available at https://CRAN.R-project.org/package=jomo.

Quartagno, M., Carpenter, J., and Goldstein, H. (2020), "Multiple Imputation with Survey Weights: A Multilevel Approach," *Journal of Survey Statistics and Methodology*, 8, 965–989.

Raghunathan, T., Lepkowski, J., Van Hoewyk, J., and Solenberger, P. (2001), "A Multivariate Technique for Multiply Imputing Missing Values Using a Sequence of Regression Models," *Survey Methodology*, 27, 85–95.

Raghunathan, T. E., Solenberger, P. W., and Van Hoewyk, J. (2002), *IVEware: Imputation and Variance Estimation Software*, Ann Arbor, MI: Survey Methodology Program, Survey Research Center, Institute for Social Research, University of Michigan.

Robbins, M. W., and White, T. K. (2011), "Farm Commodity Payments and Imputation in the Agricultural Resource Management Survey," *American Journal of Agricultural Economics*, 93, 606–612.

———. (2014), "Direct Payments, Cash Rents, Land Values, and the Effects of Imputation in U.S. Farm-Level Data," *Agricultural and Resource Economics Review*, 43, 451–470.

Robbins, M. W. (2014), "The Utility of Nonparametric Transformations for Imputation of Survey Data," *Journal of Official Statistics*, 30, 675–700.

Robbins, M. W., Ghosh, S. K., and Habiger, J. D. (2013), "Imputation in High-Dimensional Economic Data as Applied to the Agricultural Resource Management Survey," *Journal of the American Statistical Association*, 108, 81–95.

Rubin, D. B. (1987), *Multiple Imputation for Nonresponse in Surveys*, New York, NY: John Wiley & Sons.

———. (1996), "Multiple Imputation after 18+ Years," *Journal of the American Statistical Association*, 91, 473–489.

Schafer, J. L. (1999), "Multiple Imputation: A Primer," *Statistical Methods in Medical Research*, 8, 3–15.

———. (2017), "Mix: Estimation/Multiple Imputation for Mixed Categorical and Continuous Data," R Package Version 1.0-10.

Schafer, J. L., and Yucel, R. M. (2002), "Computational Strategies for Multivariate Linear Mixed-Effects Models with Missing Values," *Journal of Computational and Graphical Statistics*, 11, 437–457.

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, 924–933.

Seaman, S. R., White, I. R., Copas, A. J., and Li, L. (2012), "Combining Multiple Imputation and Inverse-Probability Weighting," *Biometrics*, 68, 129–137.

Shah, A. D., Bartlett, J. W., Carpenter, J., Nicholas, O., and Hemingway, H. (2014), "Comparison of Random Forest and Parametric Imputation Models for Imputing Missing Data Using Mice: A Caliber Study," *American Journal of Epidemiology*, 179, 764–774.

Su, Y.-S., Yajima, M., Gelman, A. E., and Hill, J. (2011), "Multiple Imputation with Diagnostics (mi) in r: Opening Windows into the Black Box," *Journal of Statistical Software*, 45, 1–31.

Tanner, M. A., and Wong, W. H. (1987), "The Calculation of Posterior Distributions by Data Augmentation (with Discussion)," *Journal of the American Statistical Association*, 82, 528–540.

Van Buuren, S. (2018), *Flexible Imputation of Missing Data*, Boca Raton, FL: CRC Press.

Van Buuren, S., Brand, J. P. L., Groothuis-Oudshoorn, C. G. M., and Rubin, D. B. (2006), "Fully Conditional Specification in Multivariate Imputation," *Journal of Statistical Computation and Simulation*, 76, 1049–1064.

Van Buuren, S., and Groothuis-Oudshoorn, K. (2011), "Mice: Multivariate Imputation by Chained Equations in r," *Journal of Statistical Software*, 45, 1–68.

Venables, W. N., and Ripley, B. D. (2002), *Modern Applied Statistics with S* (4th ed.), New York: Springer.

Vittinghoff, E., and McCulloch, C. E. (2007), "Relaxing the Rule of Ten Events per Variable in Logistic and Cox Regression," *American Journal of Epidemiology*, 165, 710–718.

Wang, Z., Akande, O., Poulos, J., and Li, F. (2022), “Are Deep Learning Models Superior for Missing Data Imputation in Surveys? Evidence from an Empirical Comparison,” *Survey Methodology*, 48, 375–399.

White, I. R., Royston, P., and Wood, A. M. (2011), “Multiple Imputation Using Chained Equations: Issues and Guidance for Practice,” *Statistics in Medicine*, 30, 377–399.

Yoon, J., Jordon, J., and Schaar, M. (2018), “Gain: Missing Data Imputation using Generative Adversarial Nets,” in International Conference on Machine Learning, pp. 5689–5698, Stockholm, Sweden: PMLR 80.

Yucel, R. M. (2008), “Multiple Imputation Inference for Multivariate Multilevel Continuous Data with Ignorable Non-Response,” *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 366, 2389–2403.

Zhao, J. H., and Schafer, J. L. (2018), “pan: Multiple Imputation for Multivariate Panel or Clustered Data.” R Package Version 1.6.