Published in final edited form as:

Biometrics. 2011 September ; 67(3): 799–809. doi:10.1111/j.1541-0420.2010.01538.x.

Multiple Imputation for Missing Values Through Conditional Semiparametric Odds Ratio Models

Hua Yun Chen^{1,*}, Hui Xie^{1,**}, and Yi Qian^{2,***}

¹Division of Epidemiology & Biostatistics, School of Public Health, University of Illinois at Chicago, Chicago, IL 60612

²Department of Marketing, Kellogg School of Management, Northwestern University, Evanston, IL 60208

Summary

Multiple imputation is a practically useful approach to handling incompletely observed data in statistical analysis. Parameter estimation and inference based on imputed full data have been made easy by Rubin's rule for result combination. However, creating proper imputation that accommodates flexible models for statistical analysis in practice can be very challenging. We propose an imputation framework that uses conditional semiparametric odds ratio models to impute the missing values. The proposed imputation framework is more flexible and robust than the imputation approach based on the normal model. It is a compatible framework in comparison to the approach based on fully conditionally specified models. The proposed algorithms for multiple imputation through the Monte Carlo Markov Chain sampling approach can be straightforwardly carried out. Simulation studies demonstrate that the proposed approach performs better than existing, commonly used imputation approaches. The proposed approach is applied to imputing missing values in bone fracture data.

Keywords

Acceptance-rejection sampling; Dirichlet process prior; Gibbs sampler; Hybrid MCMC; Molecular dynamics algorithm; Nonparametric Bayesian inference; Rejection control

1. Introduction

Missing data occur in almost all studies involving human subjects. It is well known that analysis based on complete cases can be biased and the efficiency in estimation can be substantially reduced. Many statistical methods have been developed in recent decades (Little and Rubin, 2002; Ibrahim et al., 2005) to address the problems of potential bias and loss of efficiency in parameter estimation and inference due to missing data. Utilization of those methods in practice is largely affected by the difficulties in implementing them in software packages due to the methods' complexity and computational intensiveness. Multiple imputation (Rubin, 1987, 1996; Scheuren, 2005) stands out as an approach that can address both the complexity and computational intensiveness through the Monte Carlo simulation approach to Bayesian inference.

* hychen@uic.edu . ** huixie@uic.edu . *** yiqian@northwestern.edu .

Supplementary Materials Web Appendices A and B, referenced in Sections 3 and 4 respectively, are available under the Paper Information link at the Biometrics website <http://www.biometrics.tibs.org>.

Some standard statistical software packages, such as SAS and STATA, along with a number of stand-alone packages, such as MICE and IVEware, have implemented multiple imputation in various forms (Schafer, 1999; van Buuren et al., 1999; Raghunathan et al., 2001). See Horton and Lipsitz (2001) for a review of the software and Harel and Zhou (2007) for a summary. These software packages have made the application of the multiple imputation approach substantially easy in practice. However, the modeling approaches implemented in current software packages have serious limitations. One such limitation of certain software packages (Kenward and Carpenter, 2007; Schafer, 1997) is that the imputation models are inflexible in modeling a mixture of discrete and continuous data. For example, a joint normal model may be assumed for the data despite the discrete nature of some variables, and the normal model can have difficulties incorporating interaction and higher-order terms. This can result in bias in estimation based on the imputed data (van Buuren, 2007; Yu et al., 2007). Alternative methods, such as the fully conditionally modeling approach, or the sequential imputation approach (van Buuren et al., 1999; Raghunathan et al., 2001; Gelman and Raghunathan, 2001), potentially suffer from incompatibility in model specification and, as a result, theory to support the use of those methods is hard to obtain. See Kenward and Carpenter (2007) for a comparison of the relative merits of different imputation methods and for a discussion of the key issues yet to be addressed with the imputation methods currently implemented in software packages. To address the issues with the currently implemented imputation approaches, we propose a novel imputation framework based on conditional semiparametric odds ratio models (Chen, 2003, 2004, 2007). By using the conditional semiparametric odds ratio modeling framework, we can simultaneously address both the issue of the inflexibility of the joint normal model and the issue of the possible incompatibility of the fully conditional models. The semiparametric odds ratio model can be naturally obtained from an odds ratio decomposition of the conditional density (Chen, 2003). It generalizes generalized linear models and is very flexible in modeling the mixture of discrete and continuous data. Chen (2004) uses it for modeling covariate distribution in the missing covariate problems. When the number of missing groups, *i.e.*, groups of variables observed or missing together, is large and each missing group can take many distinct values, computation of the maximum likelihood estimator can be prohibitively expensive because of the combinatory nature of the algorithm. The imputation approach based on Markov Chain Monte Carlo allows us to handle the missing data problem with a much higher dimension.

To apply the approach to imputing missing values, we need first to study Bayesian inference under the semiparametric odds ratio model. Since the Bayesian approach is used as a convenient analytical framework with a large sample size, we choose weak priors to minimize their impact on the ultimate inference. One difficulty to overcome is the specification of the prior distribution for the nonparametric parameter. Although many prior distributions for the nonparametric parameters (Antoniak, 1974; Escobar, 1994; Walker et al., 1999; Müller and Quintana, 2004) can be used, the Dirichlet process prior makes the computation relatively simple in our case when an approximation is applied. We propose to use the Dirichlet process prior (Ferguson, 1973, 1974; Blackwell and MacQueen, 1973; Escobar, 1994) for the nonparametric parameters in the model. Another difficulty to overcome in Bayesian inference under the semiparametric model is finding relatively efficient sampling methods to simulate the posterior distribution (MacEachern et al., 1999). We propose to use the Gibbs sampler with a hybrid Monte Carlo step (Liu, 2001) to sample from the parameter distribution associated with the consecutive conditional models given the full data and the rest of the parameters. Once the algorithm for simulating the posterior distribution is established, imputing missing values based on the conditional semiparametric odds ratio model becomes very easy because the distributions to sample from are all discrete with probability masses on a finite number of known points, and the probability masses can be easily computed. In fact, the imputed values for each missing data group under the

proposed framework can be viewed as weighted draws from the observed values of that missing group. Similar to the hot deck imputation approach (Little and Rubin, 2002), restrictions on the values a variable can take, such as bounds or semi-continuity, which can be a major problem in other imputation approaches, are no longer a problem using the proposed framework. The proposed approach imputes the missing values of a sampling unit by the weighted draws from the combinations of the observed values from different missing groups. In this respect, it may be considered as a special hot deck imputation. However, in contrast to the hot deck imputation approach, which is often *ad hoc*, the weights in our approach are computed based on the semiparametric model and the imputation is proper in Rubin's sense.

The remainder of this article is organized as follows. In Section 2, we first briefly introduce the setting of the missing data problem along with the notation used. The conditional semiparametric odds ratio models for multivariate data are then introduced. In Section 3, Bayesian inference under the conditional semiparametric odds ratio model is first studied in detail and algorithms for drawing Monte Carlo samples from the posterior distributions are proposed. Building on these developments, an algorithm for imputing missing values in incompletely observed data is then proposed. Simulations are conducted in Section 4 to evaluate the performance of the proposed imputation approach. The approach is applied to imputing missing values in bone fracture data in Section 5. The article concludes with a discussion of issues that may arise when applying the proposed approach in practice, and further research needed to resolve those issues.

2. Models for the missing data problem

2.1 The missing data problem

Consider the general missing data problem where the full data from a sampling unit are denoted by Y . Suppose that the variables in Y are divided into t groups as $Y_1, \dots, Y_t$, where Y_j is a vector of $d_j \geq 1$ dimension for $j = 1, \dots, t$. We assume that all variables in Y_j are either observed or missing together, $j = 1, \dots, t$. Each Y_j is called a missing group in this paper. As will be seen in the next section, one advantage of dealing with a missing group rather than an individual variable is that the amount of modeling can be reduced and the imputation can be carried out relatively easily. Let $R = (R_1, \dots, R_t)$ denote the missing data indicator for Y . $R_j = 1$ if Y_j is observed, and $R_j = 0$ otherwise. Let $R_j(Y_j)$ denote the observed data on Y_j such that $R_j(Y_j) = Y_j$ if $R_j = 1$, and $R_j(Y_j) = (-\infty, +\infty)^{d_j}$ if $R_j = 0$, for $j = 1, \dots, t$. Further, let $R(Y) = \{R_1(Y_1), \dots, R_t(Y_t)\}$. Given the model parameters, the observed data from a sampling unit are $\{R, R(Y)\}$. The whole sample consists of n i.i.d. replicates of $\{R, R(Y)\}$.

Rubin (1976) and Little and Rubin (2002) classify missing data into three categories based on the missing data mechanism. They are missing completely at random (MCAR) when $P(R|Y)$ does not depend on Y , missing at random (MAR) when $P(R|Y)$ is a function of $R(Y)$ only, and nonignorable missing (NI) or not missing at random (NMAR) when $P(R|Y)$ depends on the missing values. In this paper, we present a general methodology that is applicable to both ignorable and nonignorable missing data problems.

For missing data with arbitrary missing patterns, a key component for carrying out imputation is to specify a model for the full data. Let the density of Y under a product of Lebesgue measures and/or count measures be decomposed into consecutive conditional densities as

$$g(y_t, \dots, y_1) = \prod_{j=1}^t g_j(y_j | y_{j-1}, \dots, y_1).$$

The semiparametric odds ratio models we propose for imputation are built on this decomposition, which we describe in detail in the following section.

2.2 The semiparametric odds ratio model for the full data

Let $(y_{j0}, \dots, y_{j10})$ be a fixed point in the sample. For a given conditional density $g_j(y_j|y_{j-1}, \dots, y_{j10})$, define the odds ratio function relative to $(y_{j0}, \dots, y_{j10})$ as

$$\eta_j\{y_j; (y_{j-1}, \dots, y_{j10})|y_{j0}, \dots, y_{j10}\} = \frac{g_j(y_j|y_{j-1}, \dots, y_{j10})g_j(y_{j0}|y_{(j-1)0}, \dots, y_{j10})}{g_j(y_j|y_{(j-1)0}, \dots, y_{j10})g_j(y_{j0}|y_{j-1}, \dots, y_{j10})}.$$

For notational simplicity, we use $\eta_j\{y_j; (y_{j-1}, \dots, y_{j10})\}$ to denote $\eta_j\{y_j; (y_{j-1}, \dots, y_{j10})|y_{j0}, \dots, y_{j10}\}$ in the remainder of the paper. Chen (2003, 2004, 2007) shows that the conditional density can be rewritten as

$$g_j\{y_j|y_{j-1}, \dots, y_{j10}\} = \frac{\eta_j\{y_j; (y_{j-1}, \dots, y_{j10})\}g_j(y_{j0}|y_{(j-1)0}, \dots, y_{j10})}{\int \eta_j\{y_j; (y_{j-1}, \dots, y_{j10})\}g_j(y_j|y_{(j-1)0}, \dots, y_{j10})dy_j}.$$

In general, the choice of the fixed point can be arbitrary. In practice, certain choices, such as the center of the data points, may make computation numerically more stable.

One of the advantages of the odds ratio representation of a conditional density is that when there is no additional constraint imposed on the density, the odds ratio function and the conditional density at a fixed point, $g_j(y_j|y_{(j-1)0}, \dots, y_{j10})$, are variation independent. This means that we can separately model the odds ratio function and the conditional density at a fixed condition without creating any incompatibility in the density expression (Gelman and Speed, 1993, 1999). For many parametric models, the odds ratio representation of the conditional density is a reparametrization of the model parameters. This point can be illustrated with the following examples.

Consider, for example, the case of two variables with (Y_1, Y_2) following a bivariate normal distribution with marginal means m_1 and m_2 , variances τ_1^2 and τ_2^2 , and correlation coefficient ρ . Then $g_1(y)$ is normal with mean $\mu_1 = m_1$ and variance $\sigma_1^2 = \tau_1^2$, and

$$g_2(y_2|y_1) = \frac{1}{\sqrt{2\pi}\sigma_2} \exp\left[-\frac{1}{2\sigma_2^2}\{y_2 - \mu_2(y_1)\}^2\right],$$

where $\mu_2(y_1) = \beta_0 + \beta_1 y_1$, $\beta_0 = m_2 - \beta_1 m_1$, $\beta_1 = \rho\tau_2/\tau_1$, and $\sigma_2^2 = (1 - \rho^2)\tau_2^2$. The odds ratio function for the conditional normal density is

$$\eta_2(y_2; y_1) = \exp\left\{\frac{\beta_1}{\sigma_2^2}(y_2 - y_{20})(y_1 - y_{10})\right\}$$

and the conditional density at a fixed condition is

$$g_2(y_2|y_{10}) = \frac{1}{\sqrt{2\pi}\sigma_2} \exp\left\{-\frac{1}{2\sigma_2^2}(y_2 - \mu_{20})^2\right\}.$$

The parameters in the odds ratio function and the conditional density at a fixed point are, respectively, β_1/σ_2^2 and $(\sigma_2^2, \beta_0 + \beta_1 y_{10})$. When $y_{10} = 0$, it is obvious to see that the two sets of parameters are variation independent. This is also generally true even if $y_{10} \neq 0$.

For a general conditional normal density,

$$g_j(y_j|y_{j-1}, \dots, y_1) = \frac{1}{\sqrt{2\pi}\sigma_j} \exp\left\{-\frac{1}{2\sigma_j^2}(y_j - \mu_j)^2\right\},$$

where μ_j and σ_j may depend on $(y_{j-1}, \dots, y_1)$. The odds ratio function for the conditional normal density is

$$\eta_j\{y_j; (y_{j-1}, \dots, y_1)\} = \exp\left[\left(y_j - y_{j0}\right)\left\{\left(\mu_j\sigma_j^{-2} - \mu_{j0}\sigma_{j0}^{-2}\right) - \frac{y_j + y_{j0}}{2}(\sigma_j^{-2} - \sigma_{j0}^{-2})\right\}\right]$$

and the conditional density at a fixed condition is

$$g_j(y_j|y_{(j-1)0}, \dots, y_{10}) = \frac{1}{\sqrt{2\pi}\sigma_{j0}} \exp\left\{-\frac{1}{2\sigma_{j0}^2}(y_j - \mu_{j0})^2\right\}.$$

Under both the homogeneous variance assumption and the linear regression model assumption, *i.e.*, σ_j^2 does not depend on $(y_{j-1}, \dots, y_1)$ and $\mu_j = \beta_0 + \beta_1 y_1 + \dots + \beta_{j-1} y_{j-1}$, the odds ratio function reduces to

$$\eta_j\{y_j; (y_{j-1}, \dots, y_1)\} = \exp\left\{\sum_{k=1}^{j-1} \frac{\beta_k}{\sigma^2} (y_j - y_{j0})(y_k - y_{k0})\right\}.$$

The parameters in the odds ratio function and the conditional density at a fixed point are, respectively, $(\beta_k/\sigma^2, k = 1, \dots, j-1)$ and $(\sigma^2, \beta_0 + \beta_1 y_{10} + \dots + \beta_{j-1} y_{(j-1)0})$. The two sets of parameters are also variation independent. Similar reparametrization can be obtained for the generalized linear models, where

$$g_j(y_j|y_{j-1}, \dots, y_1) = \exp\left[\frac{1}{\phi_j} \left\{\xi_j y_j - b(\xi_j)\right\} + c(y_j, \phi_j)\right]$$

and ξ_j and ϕ_j may depend on $(y_{j-1}, \dots, y_1)$. Note that when ϕ_j does not depend on $(y_{j-1}, \dots, y_1)$, the odds ratio function for this density becomes

$$\eta_j\{y_j; (y_{j-1}, \dots, y_1)\} = \exp\left\{(y_j - y_{j0})(\xi_j - \xi_{j0})/\phi\right\}$$

and the conditional density at a fixed condition is

$$g_j(y_i|y_{(j-1)0}, \dots, y_{10}) = \exp \left[\frac{1}{\phi} \{ \xi_{j0} y_j - b(\xi_{j0}) \} + c(y_i, \phi_j) \right].$$

When the canonical link function is used in the generalized linear model, $\xi_j = \beta_0 + \beta_1 y_1 + \dots + \beta_{j-1} y_{j-1}$. As a result, the parameters in the odds ratio function are $(\beta_k/\phi, k = 1, \dots, j-1)$ and the parameters in the conditional density at a fixed point are $(\phi, \beta_0 + \beta_1 y_{10} + \dots + \beta_{j-1} y_{(j-1)0})$. The two sets of parameters are also variation independent.

In the following, we model the odds ratio function parametrically, which we denote by $\eta_j\{y_j; (y_{j-1}, \dots, y_1), \theta_j\}$, and we model $g_j(y_j|y_{(j-1)0}, \dots, y_{10})$ nonparametrically, which we denote by $f_j(y_j)$. For notational convenience, we assume that $\eta_1(y_1) = 1$. The joint model under this framework becomes

$$g(y_i, \dots, y_1 | \theta_2, \dots, \theta_i; f_1, \dots, f_i) = \prod_{j=1}^i \frac{\eta_j\{y_j; (y_{j-1}, \dots, y_1), \theta_j\} f_j(y_j)}{\int \eta_j\{y_j; (y_{j-1}, \dots, y_1), \theta_j\} f_j(y_j) dy_j}. \quad (1)$$

Many different parametric models can be used for the odds ratio function. The most convenient model is perhaps the bilinear form for the logarithm of the odds ratio function. That is,

$$\log \eta_j\{y_j; (y_{j-1}, \dots, y_1)\} = \sum_{k=1}^{j-1} \theta_{jk} (y_j - y_{j0}) (y_k - y_{k0}).$$

As evident in the preceding examples, all generalized linear models with a canonical link function have this form of odds ratio function. In general, high-order terms may be introduced into the model as

$$\log \eta_j\{y_j; (y_{j-1}, \dots, y_1), \theta\} = \sum_{k=1}^{j-1} \sum_{m_k=1}^{M_k} \sum_{l=1}^L \theta_{jklm_k} (y_j - y_{j0})^l (y_k - y_{k0})^{m_k}, \quad (2)$$

which reduces to the bilinear form when $M_k = L = 1$ for $k = 1, \dots, j-1$. Alternatively, the bilinear form of the log-odds ratio may hold after some transformations, such as,

$$\log \eta_j\{y_j; (y_{j-1}, \dots, y_1), \theta\} = \sum_{k=1}^{j-1} \theta_{jk} \{h_j(y_j) - h_j(y_{j0})\} \{h_k(y_k) - h_k(y_{k0})\},$$

where $h_k, k = 1, \dots, j$, are known monotone functions. More general situations with an unknown form of $h_k, k = 1, \dots, j$, may be considered. But we restrict ourselves to (2) with known L and M in this article. Note that if we allow L and M to be estimated, (2) can approximate any log-odds ratio function smoothly enough. Finally, note that when part or all of $y_k, k = 1, \dots, j$, are vectors, let $y_k = (y_{k1}, \dots, y_{1d_k})$. Model (2) can be rewritten as

$$\log \eta_j\{y_j; (y_{j-1}, \dots, y_1), \theta\} = \sum_{k=1}^{j-1} \sum_{|m_k|=1}^{M_k} \sum_{l=1}^L \theta_{jklm_k} \prod_{u=1}^{d_j} (y_{ju} - y_{ju0})^{l_u} \prod_{v=1}^{d_k} (y_{kv} - y_{kv0})^{m_{kv}}, \quad (3)$$

where $l = (l_1, \dots, l_d)$, $m_k = (m_{k1}, \dots, m_{kd_k})$, and $|k| = \sum_u k_u$ and $|l| = \sum_v l_v$. We assume from now on that odds ratio functions are specified up to an unknown parameter θ , where $\theta = (\theta_2, \dots, \theta_t)$ and θ_j is the parameter for η_j , $j = 2, \dots, t$.

3. Imputation under the odds ratio modeling framework

3.1 Bayesian inference with fully observed data

Assume in this subsection that data are fully observed as $(Y_1^i, \dots, Y_t^i)$, $i = 1, \dots, n$. The likelihood under the semiparametric odds ratio model (1) is

$$\prod_{i=1}^n \prod_{j=1}^t \frac{\eta_j \{Y_j^i; (Y_{j-1}^i, \dots, Y_1^i), \theta_j\} f_j(Y_j^i)}{\int \eta_j \{y_j; (Y_{j-1}^i, \dots, Y_1^i), \theta_j\} f_j(y_j) dy_j}.$$

Since (θ_j, f_j) , $j = 1, \dots, t$, are, respectively, parameters from different conditional distributions, we assume that they are distinct parameters. As a result, independent priors are assumed for the parameters with different j . Further, for any given j , the priors for (θ_j, f_j) may be reasonably assumed independent because θ_j resembles a location parameter and f_j resembles a scale parameter. More specifically, we assume the prior distribution for θ_j has the density $\psi_j(\theta_j)$. For convenience, we set $\psi_1(\theta_1) \equiv 1$. The prior distribution for f_j is

assumed to be a Dirichlet process as $\mathcal{D}_j(c_j F_j)$, where $c_j > 0$ and F_j is a probability

distribution on $R_j^{d_j}$ for $j = 1, \dots, t$. In practice, we choose c_j and F_j , and the hyperparameter in ψ_j such that they yield relatively noninformative priors. See the Web Appendix for more details on the choice of these parameters. Under the prior specification, the joint distribution for the data and the parameters is

$$\prod_{j=1}^t \left[\left\{ \prod_{i=1}^n \frac{\eta_j \{Y_j^i; (Y_{j-1}^i, \dots, Y_1^i), \theta_j\} f_j(Y_j^i)}{\int \eta_j \{y_j; (Y_{j-1}^i, \dots, Y_1^i), \theta_j\} f_j(y_j) dy_j} \right\} \mathcal{D}_j(c_j F_j) \psi_j(\theta_j) \right]. \quad (4)$$

The posteriors for $f_1, (\theta_j, f_j)$, $j = 2, \dots, t$, are the product of

$$\begin{aligned} \mathcal{P}(f_1 | Y^i, i=1, \dots, n) &\propto \left\{ \prod_{i=1}^n p_1(Y_1^i | f_1) \right\} \mathcal{D}_1(c_1 F_1) \propto \mathcal{D}_1\left(c_1 F_1 + \sum_{i=1}^n \delta_{Y_1^i}\right), \\ \mathcal{P}(\theta_j, f_j | Y^i, i=1, \dots, n) &\propto \left\{ \prod_{i=1}^n p_j(Y_j^i | Y_{j-1}^i, \dots, Y_1^i, \theta_j, f_j) \right\} \mathcal{D}_j(f_j) \psi_j(\theta_j) \\ &\propto \left[\prod_{i=1}^n \frac{\eta_j \{Y_j^i; (Y_{j-1}^i, \dots, Y_1^i), \theta_j\}}{\int \eta_j \{y_j; (Y_{j-1}^i, \dots, Y_1^i), \theta_j\} f_j(y_j) dy_j} \right] \psi_j(\theta_j) \mathcal{D}_j\left(c_j F_j + \sum_{i=1}^n \delta_{Y_j^i}\right), \end{aligned}$$

where $j = 2, \dots, t$, $Y^i = (Y_1^i, \dots, Y_t^i)$, and δ_{Y_j} denotes the point measure at Y_j . Note that the posteriors for (θ_j, f_j) for different j can be computed independently of each other given that the data are fully observed. We concentrate on a given j in the remainder of this section.

The posterior for θ_j involves intractable integrations. The Gibbs sampler can be used to simulate the posterior. Note that direct draws from the conditional distributions of θ_j given f_j and the data, and of f_j given θ_j and the data, are difficult if not impossible to obtain. To overcome these difficulties, we first modify the Dirichlet processes based on the following heuristic arguments. If the observed data points were the only points at which potential data could be observed, it would be natural to assume a Dirichlet process prior with the mean

distribution having probability mass only on the observed data points. In this case, the Dirichlet process prior becomes the Dirichlet distribution, which describes the distribution of the probability mass on the observed data points. By analogy to the empirical distribution approximating the true continuous distribution, it appears reasonable to assume that the prior can be approximated by such a Dirichlet distribution even when Y_j is a continuous variable.

This is equivalent to replacing $\mathcal{D}_j(c_j F_j + n F_{nj})$ with $\mathcal{D}\{(c_j + n) F_{nj}\}$, where $F_{nj} = \frac{1}{n} \sum_i \delta_{y_j^i}$. Note that

$$c_j F_j + n F_{nj} = (c_j + n) \left\{ F_{nj} + \frac{c_j}{n + c_j} (F_j - F_{nj}) \right\}.$$

The second term is of lower order in n compared to the first term on the right-hand side of the preceding equation. This suggests that the replacement is approximately correct for a large n . This simplification reduces the complication in sampling from the posterior of f_j . After the modification of the priors, we can apply the Gibbs sampler to sample from the posterior distributions. Note that the conditional distributions to be sampled in the Gibbs sampler do not have a closed-form expression and cannot be directly sampled. There are many ways to circumvent this problem. One approach is first to use one-dimensional Laplace approximations to the conditional distributions, and then sample from the reference distribution based on the Laplace approximation and apply rejection control to yield a Markov chain sample (Tierney, 1994). Note that this approach requires us to perform maximizations in each step, which can be computationally costly. An alternative approach based on the hybrid Monte Carlo sample (Duan et al., 1987; Liu, 2001) is attractive because the parameters in each $f_j, j = 2, \dots, t$, are likely highly correlated. The hybrid Monte Carlo approach first generates a candidate sample based on one of the deterministic algorithms used to approximate molecular dynamics in physics. A rejection control step is then added to determine the move of the chain. See Liu (2001) for a detailed account of this approach. Let $\lambda_j = \log f_j$. The Gibbs sampler for sampling $(\lambda_j, \theta_j), j = 1, \dots, t$, iteratively can be described as follows.

- (1) Fit an independence model to the data, *i.e.*, set $\eta \equiv 1$ (or $\theta = 0$) and f_j equal to the empirical marginal probability mass function for each j .
- (2) Given the data, sample (θ_j, λ_j) using the hybrid Monte Carlo approach.
- (3) Repeat step 2 for $j = 1, \dots, t$.
- (4) Repeat steps 2 and 3 until convergence.

Technical details on the hybrid Monte Carlo approach in step 2 are provided in the Web Appendix A.

3.2 Imputing missing values

For incompletely observed data, let $\pi(r, y_1, \dots, y_t, \gamma)$ denote the missing data probability for missing pattern r with parameter γ , which is distinct from the parameters in the full data model. Given the full data, the missing data indicators, $R^i, i = 1, \dots, n$, are distributed as

$$\prod_{i=1}^n \prod_r \left\{ \pi(r, Y_1^i, \dots, Y_t^i, \gamma) \right\}^{1_{\{R_i=r\}}}.$$

Assume that γ has a prior with density $\xi(\gamma)$, which is independent of the prior for the parameters in the full data model. The joint distribution of (R^i, Y^i) , $i = 1, \dots, n$, and parameters $(\gamma, \theta_2, \dots, \theta_t, f_1, \dots, f_t)$ is proportional to

$$\left\{ \prod_{i=1}^n \pi(R_i, Y_1^i, \dots, Y_t^i, \gamma) \right\} \xi(\gamma) \left(\prod_{j=1}^t \left[\prod_{i=1}^n \frac{\eta_j \{Y_j^i; (Y_{j-1}^i, \dots, Y_1^i), \theta_j\} f_j(Y_j^i)}{\int \eta_j \{y_j; (Y_{j-1}^i, \dots, Y_1^i), \theta_j\} f_j(y_j) dy_j} \right] \mathcal{D}_j(c_j F_j) \psi_j(\theta_j) \right). \quad (5)$$

With the augmented data (R^i, Y^i) , $i = 1, \dots, n$, it is easy to see that the posterior for $(\gamma, \theta_2, \dots, \theta_t, f_1, \dots, f_t)$ is the product of the posterior for $(\theta_2, \dots, \theta_t, f_1, \dots, f_t)$ given in the previous section and the posterior for γ ,

$$P(\gamma | R^i, Y^i, i=1, \dots, n) \propto \prod_{i=1}^n \prod_r \left\{ \pi(r, Y_1^i, \dots, Y_t^i, \gamma) \right\}^{\mathbb{1}_{\{R^i=r\}}} \xi(\gamma). \quad (6)$$

Assume that samples from this posterior can be obtained. Given the augmented data, samples from the posterior distribution can be obtained for all parameters.

For incompletely observed data, the Gibbs sampler discussed in the previous section can be applied when data are augmented by missing values. As a result, we need to impute the missing values in each Gibbs sampler cycle once the parameters have been drawn. Suppose that $(\gamma, \theta_2, \dots, \theta_t, \lambda_1, \dots, \lambda_t)$ is drawn by applying one round of the algorithm from the

previous section and the method for drawing γ . The missing data, $\bar{R}^i(Y^i)$, can then be drawn from the conditional distribution of the missing data given the observed data, $R^i(Y^i)$. This conditional distribution may involve evaluating summations with a huge number of terms. To avoid this problem, we choose to impute the missing values component-wise. Suppose that $R_{-j}^i = 0$ and $Y_{-j}^i = (Y_k^i, k \neq j)$ is known. The distribution of Y_j^i given R^i, Y_{-j}^i and the parameters has a density proportional to

$$\left[\pi(R_i, Y_1^i, \dots, Y_t^i, \gamma) \prod_{l=j}^t \frac{\eta_l \{Y_l^i; (Y_{l-1}^i, \dots, Y_1^i), \theta_l\}}{\int \eta_l \{y_l; (Y_{l-1}^i, \dots, Y_1^i), \theta_l\} f_l(y_l) dy_l} \right] f_j(Y_j^i). \quad (7)$$

Since f_j is a discrete distribution with probability masses on a finite number of points, sampling Y_j from the preceding distribution is very simple because it is also a discrete distribution with probability masses w_{ijf_j} , $j = 1, \dots, K_j$. The factor, w_{ij} , can be straightforwardly computed when we recognize that the integral in the denominator of the conditional density is a summation.

In summary, the algorithm for carrying out imputation based on the Markov Chain Monte Carlo sampling approach can be described as follows.

- (1) Initially, an independence model is fit to the incompletely observed data. The missing values are then imputed using step 4, which, in this case, is approximately equivalent to random draws from the observed values for each variable.
- (2) Draw γ from the distribution with a density proportional to (6).
- (3) Carry out one step of the hybrid Monte Carlo sampling algorithm for (λ_j, θ_j) , $j = 1, \dots, t$.

- (4) For each $i = 1, \dots, n$, and missing group $Y_j, j = 1, \dots, t$, if Y_j^i is missing, impute Y_j^i from the conditional distribution of Y_j^i given $(Y_{-j}^i, R_i, \gamma, \theta_j, f_j)$, which is the discrete distribution in (7).
- (5) Repeat steps 2–4 until convergence.

At convergence, the generated values are taken as imputed values for missing data. Multiple imputed datasets can be obtained through additional iterations. To make the imputed datasets less dependent on each other, different imputed datasets are obtained with a relatively large number of iterations in between. Alternatively, parallel chains may be run simultaneously to obtain different imputed datasets.

The above derivation works for both ignorable and nonignorable missing data problems. Under the missing at random (Rubin, 1976; Little and Rubin, 2002) assumption, the missing data probability does not depend on the missing Y values. As a result, the missing data probability can be excluded from the imputation step. Because the priors for γ and for the parameters in the full data model are independent, the posteriors for γ and for the other parameters are also independent. This means that, under the MAR assumption, we do not need to model the missing data probability for the imputation and step 2 in the above algorithm can be removed.

4. Simulation study

In this section, we evaluate the performance of the proposed method through simulation studies. Datasets with two binary variables, three normal variables, and one Poisson variable, are simulated. Conditional models are used in the simulation of the joint distribution of the six variables. Specifically, Y_1 is a standard normal random variable. Given Y_1 , Y_2 is binary with

$$\text{logit}\{P(Y_2=1|Y_1)\}=\beta_{20}+\beta_{21}Y_1.$$

Given Y_1 and Y_2 , Y_3 is normally distributed with mean

$$\mu_1(Y_1, Y_2)=\beta_{30}+\beta_{31}Y_1+\beta_{32}Y_2+\beta_{33}Y_1 * Y_2,$$

and unit variance. Given $Y_j, j = 1, 2, 3$, Y_4 is Poisson distributed with rate parameter

$$\log\lambda(Y_1, Y_2, Y_3)=\beta_{40}+\beta_{41}Y_1+\beta_{42}Y_2+\beta_{43}Y_3.$$

Given $Y_1, \dots, Y_4$, Y_5 is normally distributed with mean

$$\mu_2(Y_j, j=1, \dots, 4)=\beta_{50}+\beta_{51}Y_1+\beta_{52}Y_2+\beta_{53}Y_3+\beta_{54}Y_4+\beta_{55}Y_3 * Y_4,$$

and unit variance. Given $Y_j, j = 1, \dots, 5$, Y_6 is binary with

$$\text{logit}\{P(Y_6=1|Y_j, j=1, \dots, 5)\}=\beta_{60}+\beta_{61}Y_1+\beta_{62}Y_2+\beta_{63}Y_3+\beta_{64}Y_4+\beta_{65}Y_5+\beta_{66}Y_2*Y_4.$$

The full data are simulated from two sets of models. In the first set, no interaction terms are involved in the model, which corresponds to setting $(\beta_{33}, \beta_{55}, \beta_{66}) = (0, 0, 0)$. In the second set, substantial interactions are included with $(\beta_{33}, \beta_{55}, \beta_{66}) = (1.0, 1.0, -1.0)$.

Although the proposed approach can in principle handle nonignorable missing data, finding the parameter estimator can be very difficult when some parameters in the model are nearly unidentifiable, which is not uncommon in such problems. This can result in very slow mixing or non-convergence in the Monte Carlo simulation. To avoid the problem, we only simulate two types of missing data mechanisms. In the first type, we simulate MCAR missing data, with each variable being subject to 10% randomly selected missing values. Overall, only about 53% of the sample has complete observations for all six variables. In the second type, we simulate MAR missing data, with the first three variables being completely observed and each of the other three variables being subject to missing values. The missing data probability depends on the first three variables through linear logistic regression models of the form

$$\text{logit} \{P(R_k=1|Y_j, j=1, \dots, 6)\} = \alpha_{k0} + \alpha_{k1}Y_1 + \alpha_{k2}Y_2 + \alpha_{k3}Y_3, \quad k=4, 5, 6,$$

where $\alpha_{k0} = 2$ and $\alpha_{k1} = \alpha_{k2} = \alpha_{k3} = 0.5$. On average, about 48% of the observations have complete data for all variables.

In practice, the missing data are usually first imputed by a model compatible with the model for analysis. The actual model fit to the imputed data may be a little different from the model used for imputation. In the simulation, we try to emulate this situation to see how the estimator from the imputed data behave. The analysis of the simulated data is divided into two steps: the imputation step and the analysis step. In the imputation step, a fully Bayesian conditional semiparametric odds ratio model with the correct conditional distribution is first fit to the simulated incomplete data using the Gibbs sampler approach. At convergence, twenty imputed complete datasets are generated. In the analysis step, the imputed datasets are analyzed using the respective parametric models given earlier. Rubin's rule for combination is then applied to the estimates from the multiply imputed datasets. Note that a semiparametric odds ratio model compatible with the parametrically specified model for data analysis almost always exists.

To reduce the amount of computation in the imputation step, the observed data are rounded to one decimal place. The observed data, however, are not rounded in the analysis step. We use both the method by Geweke (1992) and the time series plots for convergence diagnosis on the Markov chains. The diagnosis suggests that the Markov chains converge to the stationary distributions within 1000 burn-in iterations for all test runs. The autocorrelation plots from these test runs also show that the lag-50 draws are effectively uncorrelated. To be cautious, we use 2000 burn-in cycles and 150 iterations between imputations in the simulation studies. The simulation results are obtained based on 500 replicates of a sample size of 400 for the full data. Two analyses are performed on the imputed datasets. One uses 20 imputed datasets and the other uses five. Because of the similarity of the results, we only present the results based on five imputed datasets.

In addition to the estimates based on the multiply imputed datasets, we also compute the estimates based on the full data (FD) and on the complete cases (CC), respectively, for the purpose of comparison. Two major multiple imputation methods currently implemented in the software are included in the comparison with our multiple imputation method. The first method is imputation based on the joint normal (JN) model, or its variant, termed the conditional Gaussian model (CGM). The CGM uses the location-scale model for a mixture

of discrete and continuous outcomes. In the location-scale model, the log-linear model is used to model the categorical variables, and a conditional joint normal model is used to model the remaining continuous outcomes. We use the function *impCgm* in the missing data library of Splus 8.0 (Insightful Corp.), which implements the method. When using the method for our simulated datasets, Y_2 and Y_6 are categorical variables and the other variables are considered as continuous variables. Note that Y_4 is a count variable. To conform to the data type, we post-process the imputed datasets. The post-processing rounds the imputed Y_4 values to the nearest integer and all negative integer values are then changed to zero. The second method is multiple imputation using fully conditional specifications. We use the R package MICE for this method. Given all the other variables at current values, the following default imputation methods are used to sequentially impute each variable: predictive mean matching for numeric data and logistic regression imputation for binary data. In CGM, interaction terms $Y_1 * Y_2$, $Y_2 * Y_4$ and $Y_3 * Y_4$ are combined with variables Y_1 to Y_6 to form a new data matrix, in which the missing values are multiply imputed in the imputation step.

Tables 1 and 2 show the results for the simulated data without interaction terms in the models. Results in Table 1 show that, when data are missing completely at random, all three imputation approaches perform well. The amount of bias is comparable to that of the complete-case analysis. The variance is generally smaller for the imputation approaches, suggesting that more power is gained by using the imputation approaches. The performance of the three imputation approaches is similar, with the joint normal approach performing slightly better in terms of the variances of the estimators. Table 2 shows that, when data are missing at random, the complete-case analysis is subject to sizable bias. The imputation approach based on the joint normal model has sizable bias for estimators of some of the parameters. In the tables, such values under the imputation approaches are highlighted in boldface. In contrast, both the MICE and our proposed approach perform well both in terms of bias and the variability of the imputation estimators.

Tables 3 and 4 show the results for the simulated data with interactions. Because of the relatively poor performance of the imputation based on the joint normal model, we replace it with an imputation based on conditional Gaussian models, an improved version of the joint normal imputation approach. When data are missing completely at random, both the CGM and MICE imputation approaches can perform poorly in terms of bias. Our approach performs well in terms of bias and variability compared to the complete-case analysis and the other two imputation approaches. When data are missing at random, our approach continues to perform well, while estimators from CGM and MICE imputation approaches can have substantial bias. The poor performance of the CGM approach in this case suggests that conditional log-linear normal models cannot approximate the simulated model well. On the other hand, the worse performance of the MICE in this setting may reflect that the conditional models used for the imputation are in conflict with each other.

In summary, the simulations demonstrate that imputations based on the JN model (or CGM) or the MICE can perform well in models without interactions. But they can perform poorly in accommodating interactions in the models. The simulation results further suggest that incorrect imputation models can induce substantial bias, and flexible and robust models for imputation are needed to correct such bias.

We also simulate sample sizes of 200 and 800. Since these results are mostly similar to the case with a sample size of 400, they are included in the Web Appendix B. To investigate the influence of the order of the variables in the imputation model on the results, we also simulate cases of imputation models with a different order than the analysis model. The results suggest that the order has relatively little impact on the results when no interactions

are involved. When interaction terms are involved, using the reverse order of conditioning to impute missing data can lead to large bias. The simulation results are included in the Web Appendix B. This issue on variable order may be addressed by including high-order terms in the odds ratio model because, theoretically, the odds ratio model can approximate any model as long as sufficiently high-order terms are included. In practice, however, this may not be easy to do. We recommend using the same order of conditioning for imputing missing values and for analyzing the imputed full data when significant interaction effects are present.

5. Application to the bone fracture data

The bone fracture data were collected in a matched case-control study of hip fracture in male veterans. Cases are matched with controls in race and age. Potential risk factors such as body mass index, smoke status, whether having dementia, etc. are evaluated for their effects on hip fracture. The data were analyzed in Chen (2004) using the maximum likelihood approach with a semiparametric odds ratio model for the covariates. Theoretically, the maximum likelihood of Chen (2004) is efficient. However, one problem encountered in applying the maximum likelihood approach is that the computation of the integrals involved in the incomplete data likelihood evaluation, even though reduced to summations, can have so many terms that it is impractical to evaluate them in a limited amount of time. As a compromise, different values of the continuous variables are grouped to make the computation feasible. The Monte Carlo sampling approach can usually handle computation problems with a higher dimension than the direct numeric evaluation. As a result, imputation approaches are more attractive for such problems with a large number of covariates. For the bone fracture data, we are able to carry out the imputation without rounding any of the continuous variables. It also enables us to answer such questions as whether interactions are significant or not.

In analyzing the data, consecutive conditional odds ratio models are used in the imputation model. The order of the conditioning in the imputation model is age, race, etoh, smoke, dementia, antiseiz, levoT4, antichol, bmi, log(hgb), albumin, fracture. This means that we model *age*, *race* conditional on *age*, *etoh* conditional on *race* and *age*, and so on. As with the simulations, we run 2000 Gibbs sampling cycles before generating imputed datasets for the final analysis. The interval between the imputed datasets is set to 150 Gibbs sampling cycles and a total of 20 imputed datasets are generated. In order to speed the convergence, the starting parameters are estimated by fitting one imputed dataset obtained from MICE. This helps reduce the burn-in period relative to starting from parameter values estimated under the independence assumption. We also consider the other two MI methods: MICE and CGM as used in the simulation studies. Compared with these two methods, our proposed method takes a longer computation time to convergence.

The results on the analyses of the imputed datasets are listed in Table 5 along with the results from complete-case analysis and the estimates from the other two MI methods. All multiple imputation methods show that the *LevoT4* is insignificant in contrast to the result from the complete-case analysis. The imputation estimates based on five imputed datasets are mostly close to those based on 20 imputed datasets. However, the estimates for some of the variables, such as *LevoT4*, have a relatively large change in magnitude, which may suggest that more than five imputed datasets are needed for this example (Harel, 2007; Graham et al., 2007). There are large numerical differences between some parameter estimates among the three multiple imputation methods. In particular, the signs of the estimates for log(hgb), albumin, and albumin*log(hgb) are opposite for MICE when compared to those from CGM and our proposed MI method. Table 5 also reports the results, excluding the three interaction terms. All multiple imputation methods show that the *LevoT4*

is insignificant in contrast to the result from the complete-case analysis. The results from all three MI methods are more comparable to each other for this analysis.

6. Discussion

We propose a conditional semiparametric odds ratio modeling framework for imputing missing values in incompletely observed data. The modeling framework is more flexible than existing approaches. Unlike the joint normal model, high-order terms can be naturally incorporated into the semiparametric odds ratio model. The modeling framework can also accommodate different shapes of distributions and it is a consistent framework in comparison to some *ad hoc* approaches, such as the fully conditional specification approach or the sequential imputation approach. Our modeling framework is particularly useful for imputing missing values, where the imputed data are subsequently analyzed by different models because the imputation modeling framework is more general than the generalized linear models that are frequently used in practice. The proposed imputation models are consistent with many models commonly used for analysis of imputed datasets. Since models used for data analysis can be richer in fine features than semiparametric odds ratio models, computational complexity can be a big hurdle to overcome when methods for handling incompletely observed data, such as the maximum likelihood through EM, are used directly. We find that our imputation framework balances the conflicting goals well.

Some issues remain and further work is needed to address them. The use of the Dirichlet distribution to replace the Dirichlet process prior substantially reduces the computation burden. Simulation results suggest that the performance is good. However, a direct draw of random measures from a Dirichlet process may be possible. When using conditional semiparametric odds ratio models for imputation, the imputation model is likely larger than models for data analysis. Although this is not the focus of this article, it would be of substantial interest to determine its impact on the results obtained by using Rubin's combination rule. When the number of variables in the dataset is large, the number of parameters in the log-linear odds ratio model can increase rapidly. This can also happen with moderately high dimensional data when nonlinear terms in the odds ratio models are considered. It is, therefore, useful to build a variable selection mechanism into the imputation framework. Many newly developed variable selection procedures can be adopted. These issues will be formally addressed in future research. The program used for simulation and data analysis in the paper can be downloaded from <http://www.uic.edu/~hychen/Research/Methods.html>.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

We would like to thank the Editor, the Association Editor, and two referees for their thoughtful comments, which greatly helped us in revising this paper.

References

- Antoniak C. Mixture of Dirichlet processes with applications to Bayesian nonpara-metric problems. *Annals of Statistics*. 1974; 2:1152–1174.
- Blackwell D, MacQueen JB. Ferguson distribution via Pólya urn skemes. *Annals of Statistics*. 1973; 1:353–355.
- Chen HY. A note on prospective analysis of outcome-dependent samples. *Journal of the Royal Statistical Society*. 2003:575–584.Ser. B

- Chen HY. Nonparametric and Semiparametric models for missing covariates in parametric regressions. *Journal of the American Statistical Association*. 2004; 99:1176–1189.
- Chen HY. A semiparametric odds ratio model for measuring association. *Biometrics*. 2007; 63:413–421. [PubMed: 17688494]
- Duane S, Kennedy AD, Pendleton BJ, Roweth D. Hybrid Monte Carlo. *Physics Letters B*. 1987; 195(2):216–222.
- Escobar MD. Estimating normal means with a Dirichlet process prior. *Journal of the American Statistics Association*. 1994; 89:268–277.
- Ferguson TS. A Bayesian analysis of some nonparametric problems. *Annals of Statistics*. 1973; 1:209–230.
- Ferguson TS. Prior distribution on space of probability measures. *Annals of Statistics*. 1974; 2:615–629.
- Gelman A, Raghunathan TE. Discussion on “conditionally specified distributions: an introduction”. *Statistical Science*. 2001; 15:268–269.
- Gelman A, Speed TP. Characterizing a joint probability distribution by conditionals. *Journal of the Royal Statistical Society*. 1993; 55:185–188.Ser. B
- Gelman A, Speed TP. Corrigendum: Characterizing a joint probability distribution by conditionals. *Journal of the Royal Statistical Society*. 1999; 61:483.Ser. B
- Geweke, J. Evaluating the accuracy of sampling-based approaches to calculating posterior moments. In: Bernardo, JM.; Berger, JO.; Dawid, AP.; Smith, AFM., editors. *Bayesian Statistics 4*. Clarendon Press; Oxford: 1992.
- Graham JW, Olchowski AE, Gilreath TD. How many imputations are really needed? Some practical clarifications of multiple imputation theory. *Preventive Science*. 2007; 8:206–213.
- Harel O. Inferences on missing information under multiple imputation and two-stage multiple imputation. *Statistical Methodology*. 2007; 4:75–89.
- Harel O, Zhou XH. Multiple imputation: Review of theory, implementation and software. *Statistics in Medicine*. 2007; 26:3057–3077. [PubMed: 17256804]
- Horton NJ, Lipsitz SR. Multiple imputation in practice: Comparison of software packages for regression models with missing variables. *American Statistician*. 2001; 55:244–254.
- Ibrahim JG, Chen MH, Lipsitz SR, Herring AH. Missing-data methods for generalized linear models: A comparative review. *Journal of the American Statistical Association*. 2005; 100:332–346.
- Kenward MG, Carpenter J. Multiple imputation: current perspectives. *Statistical Methods in Medical Research*. 2007; 16:199–218. [PubMed: 17621468]
- Little, RJA.; Rubin, DB. *Statistical Analysis with Missing Values*. Wiley; New York: 2002.
- Liu, JS. *Monte Carlo Strategies in scientific computing*. Springer; New York: 2001.
- MacEachern SN, Clyde M, Liu JS. Sequential importance sampling for nonparametric Bayes models: the next generation. *Canadian Journal of Statistics*. 1999; 27:251–267.
- Müller P, Quintana FA. Nonparametric Bayesian Data Analysis. *Statistical Science*. 2004; 19:95–110.
- Raghunathan TE, Lepkowski JM, van Hoewyk J, Solenberger P. A multivariate technique for multiply imputing missing values using a sequence of regression models. *Survey Methodology*. 2001; 27:85–95.
- Rubin DB. Inference and missing data. *Biometrika*. 1976; 63:581–92.
- Rubin, DB. *Multiple Imputation for Nonresponse in Surveys*. John Wiley; New York: 1987.
- Rubin DB. Multiple imputation after 18 years. *Journal of the American Statistical Association*. 1996; 91:473–489.
- Schafer, JL. *Analysis of Incomplete Multivariate Data*. Chapman and Hall / CRC Press; London: 1997.
- Schafer JL. Multiple imputation: a primer. *Statistical Methods in Medical Research*. 1999; 8:3–15. [PubMed: 10347857]
- Scheuren F. Multiple imputation: How it began and continues. *The American Statistician*. 2005; 59:315–319.
- Tierney T. Markov chains for exploring posterior distributions (with discussion). *Annals of Statistics*. 1994; 22:1701–1762.

- van Buuren S, Boshuizen HC, Knook DL. Multiple imputation of missing blood pressure covariates in survival analysis. *Statistics in Medicine*. 1999; 18:681–694. [PubMed: 10204197]
- van Buuren S. Multiple imputation of discrete and continuous data by fully conditional specification. *Statistical Methods in Medical Research*. 2007; 16:219–242. [PubMed: 17621469]
- Walker SG, Damien P, Laud PW, Smith AFM. Bayesian nonparametric inference for random distribution and related functions (with discussion). *Journal of the Royal Statistical Society*. 1999; 61:485–527.Ser. B
- Yu LM, Burton A, Rivero-Arias O. Evaluation of Software for multiple imputation of semi-continuous data. *Statistical Methods for Medical Research*. 2007; 16:243–258.

Table 1

Simulation results for the MCAR data without interaction.

Parameter	FD			CC			JN			MICE			Imp		
	Bias	SE	CR	Bias	SE	CR	Bias	SE (RSE)	CR	Bias	SE (RSE)	CR	Bias	SE (RSE)	CR
$\beta_{21} = 2.0$	0.01	0.20	97	0.04	0.29	96	-0.03	0.22(0.22)	96	0.01	0.22(0.22)	96	0.01	0.22 (0.22)	96
$\beta_{31} = 1.0$	0.00	0.06	94	0.00	0.09	94	-0.04	0.07(0.07)	94	-0.01	0.07(0.07)	95	0.00	0.07 (0.07)	95
$\beta_{32} = -3.0$	0.01	0.13	94	0.01	0.17	95	-0.08	0.13(0.14)	94	0.02	0.14(0.14)	94	0.02	0.14 (0.13)	94
$\beta_{41} = 0.0$	0.00	0.06	96	0.01	0.08	96	0.02	0.07(0.07)	94	0.01	0.07(0.07)	95	0.00	0.07 (0.07)	96
$\beta_{42} = 1.0$	0.00	0.15	96	-0.01	0.21	95	-0.08	0.16(0.17)	92	-0.01	0.17(0.17)	95	-0.01	0.17 (0.17)	94
$\beta_{43} = 0.0$	0.00	0.04	96	-0.00	0.05	96	-0.01	0.04(0.04)	94	-0.00	0.04(0.04)	94	-0.00	0.04 (0.04)	94
$\beta_{51} = 0.0$	0.00	0.08	95	0.00	0.12	94	-0.02	0.09(0.10)	95	0.00	0.10(0.10)	94	0.00	0.10 (0.10)	95
$\beta_{52} = -1.0$	-0.01	0.22	95	0.00	0.30	94	0.06	0.23(0.25)	96	-0.01	0.26(0.25)	93	-0.01	0.26 (0.25)	94
$\beta_{53} = 1.0$	0.00	0.05	94	0.00	0.07	96	0.01	0.06(0.06)	95	0.00	0.06(0.06)	93	0.00	0.06 (0.06)	94
$\beta_{54} = 0.0$	-0.00	0.04	96	-0.00	0.05	96	-0.00	0.04(0.04)	95	-0.00	0.04(0.04)	96	-0.00	0.04 (0.04)	96
$\beta_{61} = -1.0$	-0.04	0.25	93	-0.06	0.33	95	0.02	0.23(0.25)	97	-0.05	0.28(0.26)	94	-0.05	0.28 (0.26)	93
$\beta_{62} = 0.0$	-0.01	0.56	94	0.01	0.80	94	-0.08	0.55(0.61)	96	0.01	0.62(0.61)	96	0.01	0.62 (0.61)	97
$\beta_{63} = 0.0$	0.04	0.19	94	0.06	0.26	94	0.03	0.21(0.21)	95	0.05	0.21(0.21)	93	0.05	0.21 (0.21)	94
$\beta_{64} = 0.0$	-0.01	0.10	96	-0.00	0.15	98	-0.01	0.11(0.12)	97	-0.01	0.11(0.12)	97	-0.01	0.12 (0.12)	96
$\beta_{65} = 2.0$	-0.03	0.13	93	-0.02	0.19	95	-0.06	0.15(0.15)	91	-0.02	0.15 (0.15)	94	-0.02	0.15 (0.15)	94

Note: "JN" denotes MI assuming joint normal distribution and is fitted using the function impGauss in the missing data library of Splus 8.0 (Insightful). After the imputation, the imputed values are post-processed to conform to the data type as follows: for binary variables, the imputed value is converted to the closer value of one or zero; for count variables, the imputed value is rounded off to the closest integer, and negative integer values are then changed to zero. "MICE" is multiple imputation using the Chained Equations. The R package MICE 1.16 is used. The default imputation method is used to impute each univariate, given all the rest. Predictive mean matching is used for numeric data, logistic regression imputation for binary data, and polytomous regression imputation for categorical data. "Imp" is the multiple imputation method using our proposed method. Bias: estimated-truth, SE: standard error estimate from simulation, RSE: Average of Rubin's standard error estimate, CR: 95% confidence interval coverage rate (in percentage).

Table 2

Simulation results for the MAR data without interaction.

Parameter	FD			CC			JN			MICE			Imp		
	Bias	SE	CR	Bias	SE	CR	Bias	SE (RSE)	CR	Bias	SE (RSE)	CR	Bias	SE (RSE)	CR
$\beta_{21} = 2.0$	0.04	0.21	95	0.40	0.33	85	0.04	0.21(0.21)	95	0.04	0.21(0.21)	95	0.04	0.21 (0.21)	95
$\beta_{31} = 1.0$	-0.00	0.06	96	-0.16	0.09	58	0.00	0.06(0.06)	94	-0.00	0.06(0.06)	96	-0.00	0.06 (0.06)	96
$\beta_{32} = -3.0$	0.01	0.12	95	0.17	0.16	84	-0.00	0.13(0.13)	93	0.01	0.12(0.13)	95	0.01	0.12 (0.13)	95
$\beta_{41} = 0.0$	-0.00	0.06	93	-0.00	0.08	95	-0.00	0.07(0.07)	95	-0.00	0.06(0.07)	96	-0.00	0.07 (0.07)	95
$\beta_{42} = 1.0$	-0.00	0.15	95	0.00	0.21	95	-0.04	0.15(0.17)	95	0.00	0.17(0.17)	94	0.00	0.18 (0.17)	95
$\beta_{43} = 0.0$	-0.00	0.04	96	-0.00	0.05	95	0.00	0.04(0.04)	96	-0.00	0.04(0.04)	96	-0.00	0.04 (0.04)	96
$\beta_{51} = 0.0$	-0.00	0.08	95	-0.00	0.12	95	0.00	0.10(0.10)	94	-0.00	0.10(0.09)	95	-0.00	0.09 (0.09)	95
$\beta_{52} = -1.0$	0.01	0.22	94	0.01	0.29	96	-0.01	0.24(0.25)	95	0.01	0.25(0.24)	94	0.01	0.25 (0.24)	94
$\beta_{53} = 1.0$	-0.00	0.05	94	0.00	0.07	96	-0.00	0.06(0.06)	95	-0.00	0.06(0.06)	93	-0.00	0.06 (0.06)	96
$\beta_{54} = 0.0$	-0.00	0.04	95	-0.00	0.05	93	-0.00	0.05(0.05)	94	-0.00	0.05(0.05)	95	-0.00	0.05 (0.05)	94
$\beta_{61} = -1.0$	-0.00	0.23	96	-0.02	0.35	95	0.09	0.22(0.26)	96	-0.01	0.27 (0.26)	95	-0.02	0.27(0.26)	95
$\beta_{62} = 0.0$	-0.05	0.55	94	-0.07	0.81	94	-0.19	0.55(0.61)	96	-0.06	0.66 (0.63)	96	-0.05	0.66 (0.63)	95
$\beta_{63} = 0.0$	0.03	0.18	95	0.04	0.27	94	0.04	0.19(0.21)	96	0.03	0.23(0.22)	93	0.03	0.23 (0.22)	93
$\beta_{64} = 0.0$	-0.01	0.11	95	-0.01	0.16	94	-0.01	0.12(0.13)	97	-0.01	0.13(0.13)	96	-0.01	0.15 (0.14)	94
$\beta_{65} = 2.0$	-0.03	0.13	94	-0.02	0.20	94	-0.10	0.15(0.16)	91	-0.03	0.17 (0.17)	95	-0.02	0.18(0.17)	94

See Table 1 for definitions of abbreviations.

Table 3

Simulation results on the MCAR data with interactions.

Parameter	FD			CC			CGM			MICE			Imp		
	Bias	SE	CR	Bias	SE	CR	Bias	SE (RSE)	CR	Bias	SE (RSE)	CR	Bias	SE (RSE)	CR
$\beta_{21} = 1.0$	0.01	0.13	96	0.02	0.19	96	0.02	0.14(0.14)	96	-0.13	0.12(0.16)	93	0.01	0.14(0.14)	96
$\beta_{31} = 1.0$	0.00	0.08	94	0.00	0.11	94	0.01	0.08(0.08)	95	-0.06	0.09(0.12)	97	0.00	0.08(0.08)	96
$\beta_{32} = -3.0$	0.00	0.11	94	-0.01	0.15	95	0.01	0.11(0.12)	98	0.37	0.13(0.23)	70	0.02	0.12(0.12)	95
$\beta_{33} = 1.0$	0.00	0.11	94	-0.01	0.15	95	-0.06	0.12(0.12)	92	-0.18	0.13(0.20)	95	-0.01	0.12(0.12)	95
$\beta_{41} = 0.0$	0.00	0.07	96	0.01	0.10	94	-0.01	0.08(0.08)	96	0.15	0.07(0.09)	65	0.01	0.08(0.08)	95
$\beta_{42} = 1.0$	0.00	0.14	95	0.00	0.20	96	0.00	0.15(0.16)	95	-0.32	0.12(0.20)	67	0.00	0.15(0.15)	95
$\beta_{43} = 0.0$	-0.00	0.03	96	-0.00	0.05	95	0.00	0.04(0.04)	95	-0.07	0.03(0.05)	67	-0.00	0.04(0.04)	94
$\beta_{51} = 0.0$	0.00	0.09	96	0.00	0.13	96	0.10	0.13(0.19)	98	-0.09	0.15(0.23)	99	0.02	0.12(0.14)	97
$\beta_{52} = -1.0$	-0.01	0.20	95	-0.01	0.27	94	-0.10	0.30(0.41)	99	0.09	0.29(0.48)	99	-0.05	0.27(0.32)	98
$\beta_{53} = 1.0$	0.00	0.06	96	0.00	0.09	96	0.16	0.13(0.14)	87	0.33	0.17(0.20)	70	-0.02	0.09(0.10)	97
$\beta_{54} = 0.0$	-0.00	0.05	94	0.00	0.07	93	-0.18	0.14(0.11)	69	-0.11	0.13(0.15)	94	-0.01	0.07(0.07)	96
$\beta_{55} = 1.0$	0.00	0.02	95	0.00	0.02	93	-0.12	0.07(0.05)	36	-0.14	0.07(0.07)	59	0.00	0.03(0.03)	93
$\beta_{61} = -1.0$	-0.02	0.24	95	-0.06	0.33	94	-0.08	0.29(0.27)	94	-0.22	0.24(0.26)	91	-0.03	0.27(0.27)	96
$\beta_{62} = 0.0$	-0.00	0.59	95	0.02	0.78	95	0.02	0.67(0.63)	94	0.07	0.52(0.63)	98	0.01	0.66(0.63)	94
$\beta_{63} = 0.0$	-0.02	0.18	96	-0.04	0.25	95	-0.02	0.21(0.20)	95	0.03	0.18(0.19)	97	-0.02	0.20(0.20)	95
$\beta_{64} = 1.0$	0.05	0.27	95	0.11	0.37	94	0.03	0.27(0.28)	96	-0.29	0.23(0.33)	86	0.06	0.31(0.29)	95
$\beta_{65} = 0.0$	0.00	0.04	94	0.01	0.06	95	0.00	0.05(0.05)	96	0.02	0.04(0.05)	94	0.00	0.05(0.05)	96
$\beta_{66} = -1.0$	-0.05	0.30	94	-0.13	0.41	92	-0.01	0.31(0.32)	95	0.35	0.24(0.37)	86	-0.07	0.34(0.33)	96

Note: "CGM" stands for "Conditional Gaussian Model", which implements the location-scale model for a mixture of discrete and continuous outcomes. In the approach, the log-linear model is used to model the categorical variables, and a conditional joint normal outcome is then used to model the remaining continuous outcomes. In the simulation study, Y2 and Y6 are categorical variables and the rest are continuous variables. The method imputes Y4, which is a count variable, as a normal outcome. The imputed values of Y4 are rounded off to the closest integers, and negative integer values are then changed to zero. The function *impCgm* in the missing data library of Splus 8.0 (Insightful Corp.) is used. "MICE" is multiple imputation using the Chained Equations. The R package MICE 1.16 is used. The default imputation method is used to impute each univariate, given all the rest. Predictive mean matching is used for numeric data, logistic regression imputation for binary data, and polytomous regression imputation for categorical data. Interaction terms Y1 * Y2, Y2 * Y4 and Y3 * Y4 are used for imputing all terms.

Table 4

Simulation results for the MAR data with interactions.

Parameter	FD			CC			CGM			MICE			Imp		
	Bias	SE	CR	Bias	SE	CR	Bias	SE(RSE)	CR	Bias	SE(RSE)	CR	Bias	SE(RSE)	CR
$\beta_{21} = 1.0$	0.02	0.13	95	0.80	0.35	34	0.02	0.14(0.13)	95	0.02	0.13(0.13)	95	0.02	0.13(0.13)	95
$\beta_{31} = 1.0$	0.00	0.08	94	-0.16	0.12	73	0.00	0.08(0.08)	96	0.00	0.08(0.08)	94	0.00	0.08(0.08)	94
$\beta_{32} = -3.0$	0.01	0.11	96	0.15	0.17	84	0.01	0.11(0.11)	94	0.01	0.11(0.11)	96	0.01	0.11(0.11)	96
$\beta_{33} = 1.0$	0.00	0.11	94	-0.03	0.17	95	0.00	0.11(0.11)	96	0.00	0.12(0.11)	94	0.00	0.12(0.11)	94
$\beta_{41} = 0.0$	0.00	0.07	94	0.01	0.10	96	-0.00	0.08(0.08)	95	0.02	0.09(0.09)	94	0.00	0.08(0.08)	96
$\beta_{42} = 1.0$	0.00	0.14	97	0.00	0.18	94	-0.02	0.15(0.15)	95	-0.08	0.16(0.16)	93	0.00	0.14(0.15)	95
$\beta_{43} = 0.0$	-0.00	0.03	94	-0.00	0.05	96	-0.00	0.04(0.04)	96	0.04	0.04(0.04)	86	0.01	0.04(0.04)	94
$\beta_{51} = 0.0$	-0.00	0.09	93	-0.00	0.13	94	0.00	0.15(0.23)	99	-0.08	0.22(0.33)	99	0.00	0.12(0.14)	98
$\beta_{52} = -1.0$	0.01	0.20	93	0.01	0.26	93	-0.23	0.31(0.47)	97	-1.07	0.53(0.68)	72	-0.01	0.26(0.29)	97
$\beta_{53} = 1.0$	0.00	0.06	95	0.00	0.09	96	0.45	0.14(0.19)	26	1.13	0.32(0.35)	5	0.03	0.11(0.12)	96
$\beta_{54} = 0.0$	-0.00	0.05	95	-0.00	0.06	94	-0.06	0.10(0.11)	95	0.15	0.13(0.18)	90	-0.01	0.07(0.08)	96
$\beta_{55} = 1.0$	0.00	0.02	94	0.00	0.03	95	-0.23	0.06(0.07)	2	-0.44	0.13(0.18)	15	-0.02	0.07(0.05)	90
$\beta_{61} = -1.0$	-0.01	0.23	95	-0.03	0.32	95	-0.15	0.27(0.26)	91	-0.01	0.27(0.27)	94	-0.02	0.27(0.27)	94
$\beta_{62} = 0.0$	-0.00	0.55	96	-0.01	0.73	96	0.11	0.65(0.63)	94	-0.11	0.56(0.61)	97	-0.00	0.63(0.63)	95
$\beta_{63} = 0.0$	-0.00	0.17	94	-0.00	0.26	96	-0.04	0.21(0.21)	94	-0.00	0.20(0.21)	97	0.00	0.22(0.21)	95
$\beta_{64} = 1.0$	0.04	0.25	94	0.05	0.32	97	0.04	0.29(0.30)	97	0.02	0.30(0.30)	96	0.05	0.30(0.30)	96
$\beta_{65} = 0.0$	-0.00	0.04	95	-0.00	0.07	95	-0.01	0.06(0.05)	93	-0.00	0.05(0.05)	98	-0.00	0.06(0.05)	95
$\beta_{66} = -1.0$	-0.05	0.28	95	-0.05	0.37	97	-0.06	0.33(0.34)	96	0.01	0.30(0.33)	97	-0.05	0.34(0.34)	96

See Table 3 for definitions of abbreviations.

Table 5

Analysis of the imputed bone fracture data.

Variable	Method			
	CC	MICE	CGM	IMPA
Without interactions				
Etoh	1.39(0.39)	1.23(0.31)	1.18(0.30)	1.24(0.31)
Smoke	0.93(0.40)	0.62(0.30)	0.51(0.29)	0.67(0.32)
Dementia	2.51(0.72)	1.61(0.47)	1.54(0.45)	1.56(0.46)
Antiseiz	3.31(1.06)	2.51(0.64)	2.44(0.60)	2.56(0.62)
LevoT4	2.01(1.02)	0.92(0.64)	0.88(0.55)	0.97(0.62)
AntiChol	-1.92(0.77)	-1.49(0.59)	-0.91(0.48)	-1.62(0.55)
Albumin	-0.91(0.35)	-1.03(0.28)	-1.01(0.26)	-1.01(0.30)
BMI	-0.10(0.04)	-0.10(0.03)	-0.11(0.03)	-0.10(0.03)
log(HGB)	-2.60(1.20)	-3.39(0.93)	-3.18(0.88)	-3.20(0.96)
With interactions				
Etoh	1.41(0.40)	1.13(0.29)	1.15(0.30)	1.27(0.31)
Smoke	-9.21(5.69)	-5.32(4.34)	-3.05(4.52)	-2.97(4.54)
Dementia	2.80(0.79)	1.69(0.47)	1.54(0.47)	1.60(0.48)
Antiseiz	4.12(1.29)	2.45(0.62)	2.51(0.63)	2.67(0.66)
LevoT4	3.15(1.34)	0.41(0.65)	1.03(0.63)	1.00(0.66)
AntiChol	5.08(4.15)	-0.72(1.99)	-1.26(2.34)	-2.87(2.29)
Albumin	5.90(4.04)	-3.07(3.40)	2.53(2.97)	2.80(3.02)
BMI	-0.12(0.04)	-0.12(0.03)	-0.11(0.03)	-0.11(0.03)
log(HGB)	4.60(5.99)	-7.56(4.80)	1.02(4.35)	1.46(4.43)
smoke*loghgb	4.05(2.28)	2.40(1.74)	1.40(1.79)	1.82(1.80)
AntiChol*albumin	-2.36(1.40)	0.02(0.55)	0.07(0.62)	0.36(0.65)
Albumin*loghgb	-2.67(1.67)	0.95(1.35)	-1.43(1.19)	-1.58(1.22)

CC: Complete-case analysis; MICE: multiple imputation using the Chained Equations; CGM: multiple imputation using the conditional Gaussian model; IMPA: imputation estimator based on 20 imputed datasets; IMPB: imputation estimate based on 5 imputed datasets. Standard error estimates are in parentheses.