Journal of the American Statistical Association

ISSN: 0162-1459 (Print) 1537-274X (Online) Journal homepage: https://www.tandfonline.com/loi/uasa20

On Inverse Probability Weighting for Nonmonotone Missing at Random Data

BaoLuo Sun & Eric J. Tchetgen Tchetgen

To cite this article: BaoLuo Sun & Eric J. Tchetgen Tchetgen (2018) On Inverse Probability Weighting for Nonmonotone Missing at Random Data, Journal of the American Statistical Association, 113:521, 369-379, DOI: 10.1080/01621459.2016.1256814

To link to this article: https://doi.org/10.1080/01621459.2016.1256814

JOURNAL OF THE AMERICAN STATISTICAL ASSOCIATION
2018, VOL. 113, NO. 521, 369–379, Theory and Methods
https://doi.org/10.1080/01621459.2016.1256814

On Inverse Probability Weighting for Nonmonotone Missing at Random Data

BaoLuo Sun[a] and Eric J. Tchetgen Tchetgen[a,b]

[a]Department of Biostatistics, Harvard School of Public Health, Boston, MA; [b]Department of Epidemiology, Harvard School of Public Health, Boston, MA

ABSTRACT

The development of coherent missing data models to account for nonmonotone missing at random (MAR) data by inverse probability weighting (IPW) remains to date largely unresolved. As a consequence, IPW has essentially been restricted for use only in monotone MAR settings. We propose a class of models for nonmonotone missing data mechanisms that spans the MAR model, while allowing the underlying full data law to remain unrestricted. For parametric specifications within the proposed class, we introduce an unconstrained maximum likelihood estimator for estimating the missing data probabilities which is easily implemented using existing software. To circumvent potential convergence issues with this procedure, we also introduce a constrained Bayesian approach to estimate the missing data process which is guaranteed to yield inferences that respect all model restrictions. The efficiency of standard IPW estimation is improved by incorporating information from incomplete cases through an augmented estimating equation which is optimal within a large class of estimating equations. We investigate the finite-sample properties of the proposed estimators in extensive simulations and illustrate the new methodology in an application evaluating key correlates of preterm delivery for infants born to HIV-infected mothers in Botswana, Africa. Supplementary materials for this article are available online.

ARTICLE HISTORY
Received August 2015
Accepted October 2016

KEYWORDS
Augmented IPW; Bayes; Nonmonotone missing at random data

1. Introduction

Missing data are a major complication that occurs frequently in empirical research. Nonresponse in sample surveys, dropout or noncompliance in clinical trials and data excision by error or to protect confidentiality are but a few examples of ways in which full data are unavailable and our ability to make accurate inferences may be compromised. Missingness could also be introduced into a study by design, for example, multi-stage sampling plans to reduce the cost associated with measurements for all subjects. In many practical situations, the missing data pattern is nonmonotone, that is, there is no nested pattern of missingness such that observing variable X_k implies that variable X_j is also observed, for any $j < k$. Nonmonotone missing data patterns may occur, for instance, when individuals who dropped out of a longitudinal study reenter at later time points or in a cross-sectional regression analysis in which the outcome and covariates may be missing in patterns that are arbitrary across persons. The missing data process is said to be missing-completely-at-random (MCAR) if it is independent of both observed and unobserved variables in the full data, and missing-at-random (MAR) if, conditional on the observed variables, the process is independent of the unobserved ones (Rubin 1976; Little and Rubin 2002). A missing data process which is neither MCAR nor MAR is said to be missing-not-at-random (MNAR).

While complete case (CC) analysis is the easiest to implement and often used in practice, the method is generally known to produce biased estimates when the missingness mechanism is not MCAR (Little and Rubin 2002), although in regression settings, a CC analysis remains unbiased provided the missingness process does not depend on the outcome given observed covariates included in the regression model (Little and Rubin 2002; Little and Zhang 2011). Other commonly used procedures include last-observation-carried-forward analysis most commonly used in longitudinal studies and other single imputation techniques. However, such ad-hoc approaches typically provide valid inferences only under restrictive and often unrealistic conditions (Molenberghs et al. 2004; Siddiqui and Ali 1998; Little and Rubin 2002). More principled methods to appropriately account for missing data include parametric likelihood or Bayesian inference (Little and Rubin 2002; Horton and Laird 1999; Ibrahim and Chen 2000; Ibrahim, Chen, and Lipsitz 2002; Ibrahim et al. 2005) and parametric multiple imputation (MI) inference (Rubin 1977; Schafer 1999) which is widely used through its incorporation into mainstream statistical software (Horton and Lipsitz 2001).

Inverse probability weighting (IPW) (Horvitz and Thompson 1952; Little and Rubin 2002; Robins, Rotnitzky, and Zhao 1994; van der Laan and Robins 2003; Tsiatis 2006; Li et al. 2013; Seaman and White 2013) is another method to reduce selection bias from missing data or unequal sampling fractions. IPW estimation does not require specification of the full-data likelihood, but the missingness mechanism needs to be modeled. The development of coherent models and practical estimation procedures for the response mechanism of nonmonotone missing data is challenging, even under the assumption that the data are MAR. To the best of our knowledge, and

CONTACT Eric J. Tchetgen Tchetgen *etchetge@hsph.harvard.edu* Departments of Biostatistics & Epidemiology, Harvard School of Public Health, 655 Huntington Avenue, Boston, MA 02115.

Supplementary materials for this article are available online. Please go to *www.tandfonline.com/r/JASA*.

as discussed in the seminal missing data book by Tsiatis (2006), there currently is not available, a general approach to model an arbitrary nonmonotone missing data generating process strictly imposing MAR only. This represents an important gap in the missing data literature, which has essentially restricted the use of IPW estimation to monotone missing data settings.

There has been some debate in literature about the plausibility of the MAR assumption with nonmonotone missing data, and it has been argued that MNAR may be a more natural mechanism under such settings (Robins and Gill 1997; Little and Rubin 2002). Methods based on nonmonotone MNAR generally require and can be sensitive to additional parametric assumptions for the full data and missingness mechanism (Troxel, Lipsitz, and Harrington 1998; Ibrahim, Chen, and Lipsitz 2001), or for just the missingness mechanism (Rotnitzky, Robins, and Scharfstein 1998). An analysis assuming MAR may be preferable to one assuming MCAR even if the missingness mechanism is strictly MNAR (Little and Rubin 2002; Molenberghs et al. 2014), and in some empirical settings yield more accurate predictions of the missing values than those based on MNAR for nonmonotone missing data (Rubin, Stern, and Vehovar 1995). The analytic simplifications with methods based on MAR for the often nuisance missingness mechanism benefit the main focus of inquiry (Schafer and Graham 2002; Schafer 2003). In addition, estimation under the MAR assumption provides a principled framework for anchoring inference in the presence of incomplete data (Molenberghs et al. 2014). Such inference can and should subsequently be supplemented with sensitivity analyses to assess the extent to which a violation of MAR might lead to bias (Robins, Rotnitzky, and Scharfstein 1999).

In this article, we propose a class of models for arbitrary nonmonotone MAR data patterns. To estimate the missingness mechanism required for IPW estimation, we present two approaches: unconstrained maximum likelihood estimation (UMLE) and constrained Bayesian estimation (CBE). The first approach is easily implemented in standard software, say using existing procedures in SAS or R. However, despite this appealing feature, as we illustrate in the simulation studies, UMLE has a major drawback, in that the estimator may not be defined in finite samples, even if all regression models are correctly specified. This problematic feature of the approach is mainly due to certain natural restrictions of the model. In addition to UMLE, we introduce a CBE approach (Gelfand, Smith, and Lee 1992) which largely resolves any convergence difficulty and is easily implemented in standard Bayesian software packages. As IPW may be inefficient in practice, we improve its asymptotic efficiency by recovering available information from incomplete cases through implementing an augmented IPW (AIPW) estimator which is optimal within a very large class of AIPW estimators. The approach, which combines the proposed estimators of the nonmonotone missing data process with ideas originating from the seminal work by Robins, Rotnitzky, and Zhao (1994) and further developed by van der Laan and Robins (2003) and Tsiatis (2006), holds appeal in the fact that it leverages available information from incomplete cases without having to specify a model of the full data distribution. We present a simulation study to investigate the finite-sample properties of both constrained and unconstrained inferences in the context of logistic regression with nonmonotone missing outcome and covariates, followed by an analysis of preterm delivery on a cohort of women in Botswana to illustrate an application of the methods.

2. Notation and Assumptions

Let $L = (L_1, \ldots, L_K)'$ be a random K-vector representing the complete data. Let R be the scalar random variable encoding the different missing data patterns. For missing data pattern $R = m$, where $1 \leq m \leq 2^K$, we only observe $L_{(m)} \subseteq L$. For each of n individuals, we observe an independently and identically distributed realization of $(R_i, L_{(R)i})$, $i = 1, 2, \ldots, n$, and we suppress the subject index i when not essential. We reserve $R = 1$ to denote complete cases. Let $\mathbb{P}_n$ denote the empirical measure $\mathbb{P}_n f(O) = n^{-1} \sum_i f(O_i)$.

We assume that the missing data process is MAR (Rubin 1976; Robins, Rotnitzky, and Zhao 1994). IPW methodology essentially requires for unbiasedness that MAR holds for all persons in the population and so, more specifically, we shall assume everywhere MAR (Seaman et al. 2013), also sometimes called missing-always-at-random (Mealli and Rubin 2015) such that $\forall i, \gamma$,

$$\Pr\{R_i = m | l_i; \gamma\} = \Pr\{R_i = m | l_i^*; \gamma\}, \quad \forall m, l_i, l_i^* \quad \text{such that } l_{(m)i} = l^*_{(m)i} \tag{1}$$

where (l_i, l_i^*) represents a pair of possible values of L_i, so that the conditional probability of having missing data pattern m, which we denote by $\pi_m(l_{(m)})$, depends only on the observed variables for that pattern. The finite or infinite-dimensional parameter indexing the missing data mechanism is denoted by γ. Throughout, we also make the positivity assumption that $\forall i$,

$$\pi_1(l_i) > \sigma > 0 \quad \forall l_i \quad \text{in the support of } L_i, \tag{2}$$

for a fixed positive constant σ. That is, the probability of being a complete case is bounded away from zero with probability 1. Assumption (2) is necessary for identification of the full data law and smooth functionals of the latter (Robins, Rotnitzky, and Zhao 1994), and ensures finite asymptotic variance of the IPW and AIPW estimators.

A key implication of assumptions (1) and (2) is that the missing data process is nonparametrically identified. We note that for likelihood-based methods the weaker assumption of realized MAR (Seaman et al. 2013) already implies that if separate parameters index the missing data mechanism and the full data distribution, efficient estimation of the parameters of the missing data process can be obtained by maximizing its partial likelihood, ignoring the part of the likelihood corresponding to the full data.

3. Estimation of Missing Data Mechanism

Although the missingness mechanism is in principle nonparametrically identified under assumptions (1) and (2), in practice estimation typically entails specifying parametric models as dictated by the curse of dimensionality, since L is typically of moderate to high dimension (Robins and Ritov 1997). To motivate our discussion of nonmonotone missing data models,

we briefly review strategies for modeling some common missing data structures. In the simple case of two missing data patterns, that is, $R = 1, 2$, the probability of being a complete case is $1 - \pi_2(L_{(2)})$ and the parameters γ of a model $\pi_2(L_{(2)}; \gamma)$ can be estimated by maximizing the likelihood function

$$\prod_i \left\{1 - \pi_2\left(L_{(2)}\gamma\right)\right\}^{\mathbb{1}(R_i=1)} \left\{\pi_2\left(L_{(2)}\gamma\right)\right\}^{1-\mathbb{1}(R_i=1)}.$$

The two-missing-data-pattern scenario arises in familiar settings such as in regression analysis with incomplete data only on the outcome for a subset of the sample.

When $M > 2$ the missing data is said to be monotone if for some ordering of the variables in L, the kth variable is observed only if the $k-1$th variable was observed, and therefore one can sort the missing data patterns in such a way that $L_{(m+1)} \subset L_{(m)}$ for $m = 1, \ldots, M-1$. Some of the earliest works in this area include weighting methods to adjust for non-response in panel studies with monotone missing data patterns (Little and David 1983). In general, any monotone response mechanism can be modeled using a discrete hazard function (Robins, Rotnitzky, and Zhao 1994; Tsiatis 2006) by defining

$$\lambda_m(L_{(m)}) = \begin{cases} \Pr(R = m | R \leq m, L), & m \neq 1. \\ 1, & m = 1. \end{cases}$$

The discrete hazard $\lambda_m(\cdot)$ is a function of $L_{(m)}$ only since

$$\frac{\Pr(R = m|L)}{\Pr(R \leq m|L)} = \frac{\pi_m(L_{(m)})}{1 - \sum_{j>m} \pi_j(L_{(j)})}$$

and $L_{(j)} \subset L_{(m)}$ for all $j > m$ by the monotone missing data structure. Defining

$$K_m(L_{(m)}) = \Pr(R < m|L) = \prod_{j \geq m} \left\{1 - \lambda_j(L_{(j)})\right\}, \quad m \neq 1,$$

the conditional probability for each missing data pattern is

$$\pi_m(L_{(m)}) = \begin{cases} K_{m+1}(L_{(m+1)})\lambda_m(L_{(m)}), & m < M. \\ \lambda_m(L_{(m)}), & m = M. \end{cases}$$

and in particular the complete case probability is

$$\pi_1(L) = K_2\left(L_{(2)}\right) = \Pr(R < 2|L) = \prod_{j \geq 2} \left\{1 - \lambda_j\left(L_{(j)}\right)\right\}$$

To estimate the hazard functions $\lambda_m(L_{(m)})$, in practice we may run a series of logistic regressions of the indicator variable $\mathbb{1}(R = m)$ on $L_{(m)}$ among individuals with $R \leq m$, $m = 2, \ldots, M$. Alternatively, one may pool information by allowing $\lambda_m(L_{(m)})$ to share parameters across m.

3.1. The Failure of Standard Polytomous Regression

For nonmonotone missing data patterns, the nesting of patterns $L_{(m+1)} \subset L_{(m)}$ is no longer available, and building coherent models for the conditional probabilities of the various missing data patterns is challenging even under assumptions (1) and (2) (Robins, Rotnitzky, and Zhao 1994; Robins and Gill 1997; Tsiatis 2006). A straightforward approach to model $\pi_m(L_{(m)})$ using standard polytomous regression for the multinomial missing data process will often have the unintended consequence of imposing more restrictive conditions than what MAR assumption (1) strictly entails (Robins and Gill 1997). We illustrate this using an example from Robins and Gill (1997), which we adapt to a general bivariate pattern (Little and Rubin 2002, pp. 18–19). Suppose the full data are bivariate $L = (L_1, L_2)$ and one encodes the missing data patterns as follows: $R = 1$ if L is observed; $R = 2$ if one only observes $L_{(2)} = L_1$; $R = 3$ if one only observes $L_{(3)} = L_2$; and $R = 4$ if neither variable is observed. In general, the MAR assumption (1) for this scenario is $\forall \gamma$,

$$\Pr\{R = m|L; \gamma\} = \Pr\{R = m|L_{(m)}; \gamma\}, \quad m = 1, 2, 3, 4.$$

A standard polytomous logistic regression for R corresponds to

$$\Pr\{R = m|L; \gamma\} = \frac{\exp(\gamma_{0m} + \gamma_{1m}L_1 + \gamma_{2m}L_2)}{1 + \sum_{k=2}^{4} \exp(\gamma_{0k} + \gamma_{1k}L_1 + \gamma_{2k}L_2)}, \quad m = 2, 3, 4. \tag{3}$$

By the MAR assumption, since for $R = 4$ neither variable is observed, the probability $\Pr\{R = 4|L\}$ depends on neither L_1 nor L_2 so that $\gamma_{1j} = \gamma_{2j} = 0$ for $j = 2, 3, 4$. Therefore assuming model (3) under MAR implies MCAR. In general, it can be shown using a similar argument that the missing data pattern probabilities modeled using polytomous logistic regression can at most depend on the intersection of the sets of observed variables $L_{(m)}$, $m = 2, 3, \ldots, M$ (i.e., the set of fully observed variables), which is strictly stronger than the MAR assumption (1). This suggests that standard polytomous regression is ill-suited as modeling strategy for nonmonotone missing data process under MAR.

As a remedy, Robins and Gill proposed a large class of models for the missing data mechanism, which they call the randomized monotone missingness (RMM) processes, that are guaranteed to be MAR for a nonmonotone missing data mechanism without necessarily being MCAR (Robins and Gill 1997). This class of models does not span the space of all MAR models and therefore it is indeed possible to test whether the proposed class of models includes the true missing data mechanism. However, estimation of the missing data mechanism within this class is complex and computationally demanding, even for small to moderate sample sizes and number of different missing data patterns, and no software is currently available to implement the approach, which has limited its widespread adoption. In this article, we take a different direction and propose a class of models for nonmonotone missing data that spans the entire MAR model (with the class of RMM processes being a possible submodel) and therefore, with enough data such that nonparametric models can be used reliably, in principle one would not be able to reject MAR based on the observed data.

3.2. Proposed Nonmonotone Missing Data Model

Our approach involves modeling the conditional probability for each missing data pattern separately as

$$\Pr\{R = m|L\} = \pi_m(L_{(m)}), \quad m = 2, \ldots, M. \tag{4}$$

The probability of observing complete data is

$$\Pr\{R = 1|L\} = \pi_1(L) = 1 - \sum_{m=2}^{M} \pi_m(L_{(m)}), \tag{5}$$

which depends on the union set of observed variables $\bigcup_{m=2}^{M} L_{(m)}$. To ground ideas, consider as a parametric submodel of (4) the series of simple logistic models

$$\pi_m\left(L_{(m)};\gamma_m\right)=\left\{1+\exp\left[-\gamma_m\left(1,L_{(m)}\right)^T\right]\right\}^{-1},\quad m=2,\ldots,M,$$
$$\pi_1(L;\gamma)=1-\sum_{m=2}^{M}\left\{1+\exp\left[-\gamma_m\left(1,L_{(m)}\right)^T\right]\right\}^{-1},\quad \gamma=(\gamma_2,\ldots,\gamma_M). \tag{6}$$

By assumption (2), model (6) must satisfy the constraint

$$1-\sum_{m=2}^{M}\pi_m\left(L_{(m)};\gamma_m\right)>\sigma \quad \text{with probability 1.} \tag{7}$$

Consider the UMLE estimator of γ, defined as the value which maximizes the unconstrained log-likelihood function corresponding to missing data model (6).

$$\sum_{i=1}^{N}\left\{\left[\sum_{m=2}^{M}\mathbb{1}(R_i=m)\log\pi_m\left(L_{(m)i};\gamma_m\right)\right]+\mathbb{1}(R_i=1)\log\times\left[1-\sum_{k=2}^{M}\pi_k\left(L_{(k)i};\gamma_k\right)\right]\right\} \tag{8}$$

with corresponding score equation

$$\mathbb{P}_n\left\{\left[\frac{\mathbb{1}(R=1)}{\pi_1\left(L_{(1)}\right)}-\frac{\mathbb{1}(R=m)}{\pi_m\left(L_{(m)}\right)}\right]\pi_m(1-\pi_m)\left(1,L_{(m)}\right)^T\right\}=0 \tag{9}$$

for the parameters γ_m for missing data pattern m, where γ_m and $(1, L_{(m)})^T$ have the same dimension.

It may be in practice that maximizing (8) fails to converge. This could happen if there is at least one individual for whom the empirical version of constraint (7) is not satisfied in the process of finding the maximum, in which case the fitted complete case probability may be near zero or possibly negative, a real possibility especially at small or moderate sample sizes. Thus, we have referred to (8) as an unconstrained log-likelihood function, as it does not naturally impose constraint (7).

Note that even if the missingness mechanism were known, constraint (7) which depends on $\bigcup_{m=2}^{M} L_{(m)}$ can only be observed for complete case individuals. In fact, only complete cases need to satisfy the constraint to ensure that the UMLE can be computed in practice. Thus, one could in principle attempt to maximize the observed data log-likelihood (8) together with the observable constraints

$$\mathbb{1}(R_i=1)\sum_{k=2}^{M}\pi_k\left(L_{(k)i};\gamma_k\right)<1-\sigma^* \quad \text{for } i=1,2,\ldots,N, \tag{10}$$

where σ^* is a user-specified small positive constant. Still, this is potentially computationally prohibitive, since there are as many constraints as complete case observations.

Instead, in addition to UMLE, we develop a constrained Bayesian estimation approach where samples are drawn from the unconstrained posterior conditional distribution for γ and only those draws that fall into the constrained parameter space (10) are retained (Gelfand, Smith, and Lee 1992). An additional appeal of this approach is that the posterior credible intervals of γ are guaranteed to satisfy constraint (10), which is useful if one wishes to perform hypothesis testing to identify significant predictors in the missing data regression models. Constrained Bayesian estimation has been used previously in several other settings, for instance to estimate risk ratio and relative excess risk regressions (Chu and Cole 2010, 2011); however, to the best of our knowledge, it has not been used in the current context. To implement the approach, we specify a diffuse prior distribution $g(\gamma)$ for $\gamma=(\gamma_2,\ldots,\gamma_M)$ under model (6) and incorporate constraint (10) in the posterior distribution of γ. Under the constrained Bayesian model, the posterior distribution of γ is proportional to

$$f(\gamma|\text{data})\propto f(\text{data}|\gamma)g(\gamma)=\prod_{i=1}^{N}\left\{\prod_{m=2}^{M}\left\{\pi_m\left(L_{(m)i};\gamma_m\right)\right\}^{\mathbb{1}(R_i=m)}\times\Omega(\gamma,L_i)^{\mathbb{1}(R_i=1)}\right\}g(\gamma) \tag{11}$$

where

$$\Omega(\gamma,L_i)=\left\{\left[1-\sum_{k=2}^{M}\pi_k\left(L_{(k)i};\gamma_k\right)\right]\times\mathbb{1}\left[\sum_{k=2}^{M}\pi_k\left(L_{(k)i};\gamma_k\right)<1-\sigma^*\right]\right\}.$$

We define the CBE estimator of γ as the posterior mode (or mean) from distribution (11).

We note that in practice there may be some missing data patterns that are sparsely observed. In such cases, a simple approach entails combining across patterns with small event probabilities and estimating the missingness process under an additional assumption that the probability of any pattern within the combined set only depends on the intersection set of variables observed for all patterns in the combined set. Although the suggested approach to handle sparse patterns may introduce some bias, we do not anticipate the magnitude of this bias to be significant provided the combined set of patterns remains relatively rare compared to other more prominent missing data patterns. If the combination of sparse patterns gives rise to a monotone missing data pattern in the overall dataset, then the standard approach of modeling variationally independent discrete hazards described earlier may be used. The probabilities of the missing data patterns are not variationally independent because of the nesting of patterns, that is, probability of pattern m depends on hazards from m to M while that of pattern $m+1$ depends on hazards $m+1$ to M. The proposed approach subsumes monotone nonresponse patterns as a special case. However, some care is needed to ensure that the parameterization of models for each pattern respects their natural nesting in this setting. Nonetheless it will lead to a complicated estimation procedure without any apparent benefit in bias reduction or efficiency. Therefore in practice, existing discrete hazard function models should be used to construct weights with monotone missing data patterns.

4. IPW Inference

Suppose we observe n iid realizations of the vector L, and we wish to make inferences about the parameter β_0 which is the unique solution of the full data population estimating equation

$$E\{M(L;\beta_0)\} = 0 \tag{12}$$

where expectation is taken over the distribution of the complete data L. Note that we do not require a model for the distribution of the full data L; in fact, estimation is possible under certain weak regularity conditions (van der Vaart 1998) as long as full data unbiased estimating functions exist. In the presence of missing data, the estimating function in (12) may only be evaluated for complete cases, which may be a highly selective subsample even under MAR. This motivates the use of IPW estimating functions of complete cases to form the following population estimating equation

$$E\left\{\frac{\mathbb{1}(R=1)}{\pi_1(L)}M(L;\beta_0)\right\} = 0. \tag{13}$$

The unbiasedness of the above estimating equation holds by straightforward iterated expectations. We note that the IPW estimator $\hat{\beta}_{\text{ipw}}$ which solves empirical versions of (13) is inefficient especially when the fraction of complete cases is small, since incomplete cases are discarded except in that they may be included in the estimation of the weights $\pi_1(L;\hat{\gamma})$. In the next section, we will describe a strategy to recover information from incomplete cases by augmenting estimating function (13) to gain efficiency.

The IPW estimating equations framework encompasses a great variety of settings under which investigators may wish to account for nonmonotone missing data. This includes IPW of the full data score equation, where the score function is such an unbiased estimating function, given a model $f(L|\beta)$ for the law of the full data, in which case (13) reduces to

$$E\left\{\frac{\mathbb{1}(R=1)}{\pi_1(L)}\frac{\partial \log f(L|\beta)}{\partial\beta}\bigg|_{\beta_0}\right\} = 0. \tag{14}$$

Note that Equation (14) does not necessarily correspond to the observed data score equation and will therefore generally not achieve the efficiency bound for the model. Estimation can also be extended to classes of semiparametric models which specify only certain marginal relationships in L and in which scientific interest focuses on some low dimensional functional $\beta = \beta(F_L)$ of the distribution F_L of the full data L. For instance, in many health related applications it is common to specify a model $g(X,\beta)$ for the conditional mean of the outcome response Y given a set of covariates $X = (X_1, X_2, \ldots, X_P)^T$. Here $L = (Y, X)$ and either the outcome or any covariate may be missing. Then, the parameter of interest can be identified by the population IPW estimating equation

$$E\left\{\frac{\mathbb{1}(R=1)}{\pi_1(L)}\left[Y - g(X,\beta_0)\right]h(X)\right\} = 0,$$

where $h(X)$ is a user-specified function of X of the same dimension as β_0. Regression parameters in semiparametric models for right censored failure time data can likewise be identified by similar IPW population estimating equations, for example, Cox proportional hazards regression and Aalen's additive hazards regression. Analogous estimating equations are also available for longitudinal and clustered data. In all cases, empirical estimating equations are obtained by replacing population expectations with their empirical counterparts, and $\pi_1(L)$ with a consistent estimator.

To fix ideas, let $\pi_1(L;\hat{\gamma}) = 1 - \sum_{m=2}^{M}\{1 + \exp[-\hat{\gamma}_m (1, L_{(m)})^T]\}^{-1}$ where $\hat{\gamma} = (\hat{\gamma}_2, \ldots, \hat{\gamma}_M)$ is either the UMLE (assuming it can be computed) or CBE estimate. Then, an estimate for the parameter of interest β_0 is given by the solution $\hat{\beta}_{\text{ipw}}$ to the inverse probability weighted estimating equation

$$\mathbb{P}_n\left\{\frac{\mathbb{1}(R=1)}{\pi_1(L;\hat{\gamma})}M(L;\beta)\right\} = 0. \tag{15}$$

Subject to standard regularity conditions and assuming that the missing data model given in (6) is correctly specified, we show in the supplementary material that $\hat{\beta}_{\text{ipw}}$ is consistent and asymptotically normal

$$\sqrt{n}(\hat{\beta}_{\text{ipw}} - \beta_0) \xrightarrow{d} N\big(0, E\{\nabla_\beta \Gamma(\beta_0,\gamma_0)\}^{-1}\text{var} \times [\Gamma(\beta_0,\gamma_0) - W(\beta_0,\gamma_0)]\, E\{\nabla_\beta \Gamma(\beta_0,\gamma_0)\}^{-1^T}\big) \tag{16}$$

where $\Gamma(\beta,\gamma) = \{\mathbb{1}(R=1)/\pi_1(L;\gamma)\}M(L;\beta)$, S_{γ_0} is the score function (9) for the missing data mechanism evaluated at the truth and

$$W(\beta_0,\gamma_0) = E\left[\Gamma(\beta_0,\gamma_0)S_{\gamma_0}^T\right] E\left[S_{\gamma_0}S_{\gamma_0}^T\right]^{-1} S_{\gamma_0}.$$

The asymptotic variance in (16) can be consistently estimated by replacing the terms under expectation with empirical averages evaluated at $(\hat{\beta}_{\text{ipw}}, \hat{\gamma})$

$$\widehat{E}\{\nabla_\beta \Gamma(\hat{\beta},\hat{\gamma})\}^{-1}\widehat{\text{var}}[\Gamma(\hat{\beta},\hat{\gamma}) - \widehat{W}(\hat{\beta},\hat{\gamma})]\widehat{E}\{\nabla_\beta \Gamma(\hat{\beta},\hat{\gamma})\}^{-1^T}. \tag{17}$$

Although the posterior mode (or mean) is asymptotically efficient by the Bernstein–von Mises Theorem (van der Vaart 1998), in finite sample the CBE estimate may not necessarily correspond to the solution of the score function (9). For inference under the constrained Bayesian approach, we therefore apply a finite-sample correction to the variance estimate

$$\widehat{E}\{\nabla_\beta \Gamma(\hat{\beta},\hat{\gamma})\}^{-1}\widehat{\text{var}}[\Gamma(\hat{\beta},\hat{\gamma}) - \widehat{W}(\hat{\beta},\hat{\gamma}) + \widehat{E}\{W(\hat{\beta},\hat{\gamma})\}] \times \widehat{E}\{\nabla_\beta \Gamma(\hat{\beta},\hat{\gamma})\}^{-1^T} \tag{18}$$

so that the term in $\widehat{\text{var}}[\cdot]$ has mean zero empirically. The correction term $\widehat{E}\{W(\hat{\beta},\hat{\gamma})\}$ is expected to vanish as sample size increases. A conservative, albeit more easily implementable, estimate of the asymptotic variance in (16) is obtained by the standard sandwich variance formula (Robins, Rotnitzky, and Zhao 1994)

$$\widehat{E}\{\nabla_\beta \Gamma(\hat{\beta},\hat{\gamma})\}^{-1}\widehat{\text{var}}[\Gamma(\hat{\beta},\hat{\gamma})]\widehat{E}\{\nabla_\beta \Gamma(\hat{\beta},\hat{\gamma})\}^{-1^T}. \tag{19}$$

4.1. Improved IPW Estimator via Augmentation

The efficiency of the IPW estimator introduced in the previous section, which only makes direct use of complete cases, can be improved by incorporating information from individuals with missing data via augmentation of the IPW estimating equation

(Robins, Rotnitzky, and Zhao 1994; van der Laan and Robins 2003; Tsiatis 2006). The approach is based on a result by Robins, Rotnitzky, and Zhao (1994) who show that under assumptions (1) and (2), all regular and asymptotically linear (RAL) estimators based on observed data, of a functional β_0, can be shown to be asymptotically equivalent to an estimator solving

$$\mathbb{P}_n\left\{\frac{\mathbb{1}(R=1)}{\pi_1(L)}U(L;\beta)+A\left(R,L_{(R)}\right)\right\}=0. \tag{20}$$

$U(L;\beta)$ is an element of $\mathbb{U}^F$, the set of all full data estimating equations of β_0, and $A(R,L_{(R)})$ is an element of the space $\mathbb{A}$ spanned by all scores of the missing data mechanism which are of the form

$$\left\{\sum_{r\neq 1}\left[\frac{\mathbb{1}(R=1)}{\pi_1(L)}-\frac{\mathbb{1}(R=r)}{\pi_r\left(L_{(r)}\right)}\right]t_r\left(L_{(r)}\right)\right\},$$

where $t_r(L_{(r)})$ is an arbitrary q-dimensional function of the observed data $L_{(r)}$ corresponding to missing data pattern $R=r$ (Robins, Rotnitzky, and Zhao 1994). The class of estimating equations obtained by varying $U(L)$ over $\mathbb{U}^F$ and $A(R,L_{(R)})$ over $\mathbb{A}$ is referred to as augmented estimating equations, since it entails augmenting a standard IPW estimating equation by an arbitrary score function of the missingness process (Robins, Rotnitzky, and Zhao 1994; Tsiatis 2006). In principle, one can therefore construct an efficient estimator by identifying the optimal full data estimating function $U_{\text{opt}}\in\mathbb{U}^F$ paired with the optimal choice of augmentation $A_{\text{opt}}\in\mathbb{A}$ to use in Equation (20). Unfortunately the optimal index leading to a semiparametric efficient estimator is generally not available in the closed form and often computationally prohibitive in most problems of interest. Instead, we take a more practical approach to improve efficiency by using a restricted class of estimators (Tsiatis 2006).

We illustrate the approach using an example with two levels of missingness. Suppose the full data $L_i=(Y_i,\boldsymbol{X}_i)$ is independent and identically distributed for $i=1,2,\ldots,n$, where Y is the binary response variable and $\boldsymbol{X}=(X_1,X_2)^T$ are two univariate covariates. For a subsample of individuals, only (Y,X_1) was observed. Let the missing data indicator be $R_i=1$ if the ith individual is a complete case and $R_i=2$ if we only observe $L_{(2)i}=(Y_i,X_{1i})$. Suppose we assume the substantive model to be

$$\Pr(Y=1|\boldsymbol{X})=\left[1+\exp(\boldsymbol{\beta}^T\boldsymbol{X})\right]^{-1}=\mu(\boldsymbol{X},\boldsymbol{\beta}),$$

and we are interested in estimating $\boldsymbol{\beta}=(\beta_0,\beta_1,\beta_2)^T$, then the class of all augmented IPW estimators (AIPW) will be any estimator that solves

$$\mathbb{P}_n\left\{\frac{\mathbb{1}(R=1)}{\pi_1(L)}\boldsymbol{h}_{3\times1}(\boldsymbol{X},\boldsymbol{\beta})[Y-\mu(\boldsymbol{X},\boldsymbol{\beta})]+\left[\frac{\mathbb{1}(R=1)}{\pi_1(L)}-\frac{\mathbb{1}(R=2)}{\pi_2\left(L_{(2)}\right)}\right]\boldsymbol{f}_{3\times1}(Y,X_1)\right\}=0.$$

The functions $\boldsymbol{h}_{3\times1}(\boldsymbol{X},\boldsymbol{\beta})$ and $\boldsymbol{f}_{3\times1}(Y,X_1)$ are any arbitrary functions of $\boldsymbol{X}$ and (Y,X_1) respectively, where the subscripts denote their dimensions. The optimal AIPW estimator in terms of asymptotic variance corresponds to a specific choice which we denote as $\boldsymbol{h}^{\text{opt}}_{3\times1}(\boldsymbol{X},\boldsymbol{\beta})$ and $\boldsymbol{f}^{\text{opt}}_{3\times1}(Y,X_1)$. The optimal choice $(\boldsymbol{h}^{\text{opt}}_{3\times1},\boldsymbol{f}^{\text{opt}}_{3\times1})$ is only available in the closed form in special simple settings, and typically require solving complicated integral equations for each observation (Robins, Rotnitzky, and Zhao 1994; Tsiatis 2006). This will generally be the case for nonmonotone nonresponse, and therefore we consider a more practical approach, which we introduce here in the simple case with two levels of missingness, in the interest of simplifying the presentation. The supplement includes a detailed description of the approach for general nonmonotone patterns.

The proposed approach entails approximating $\boldsymbol{h}^{\text{opt}}_{3\times1}(\boldsymbol{X},\boldsymbol{\beta})$ and $\boldsymbol{f}^{\text{opt}}_{3\times1}(Y,X_1)$ with a linear combination of basis functions. For instance, the choice of basis functions $\boldsymbol{J}^h_{6\times1}(\boldsymbol{X})=\{1,X_1,X_2,X_1^2,X_2^2,X_1X_2\}^T$ and $\boldsymbol{J}^f_{6\times1}(Y,X_1)=\{1,Y,X_1,Y^2,X_1^2,YX_1\}^T$ allows for quadratic relationships in (X_1,X_2) and (Y,X_1) respectively. The approximations to $(\boldsymbol{h}^{\text{opt}}_{3\times1},\boldsymbol{f}^{\text{opt}}_{3\times1})$ are $\boldsymbol{h}^*_{3\times1}=A_{3\times6}\boldsymbol{J}^h_{6\times1}$ and $\boldsymbol{f}^*_{3\times1}=B_{3\times6}\boldsymbol{J}^f_{6\times1}$, respectively, where $A_{3\times6}$ and $B_{3\times6}$ are arbitrary constant matrices. We can then consider the class of augmented estimators

$$\mathbb{P}_n\left\{\frac{\mathbb{1}(R=1)}{\pi_1(L)}\boldsymbol{h}^*_{3\times1}(\boldsymbol{X})[Y-\mu(\boldsymbol{X},\boldsymbol{\beta})]+\left[\frac{\mathbb{1}(R=1)}{\pi_1(L)}-\frac{\mathbb{1}(R=2)}{\pi_2\left(L_{(2)}\right)}\right]\boldsymbol{f}^*_{3\times1}(Y,X_1)\right\}=0. \tag{A1}$$

It is possible to estimate the unique constant matrices $A_{3\times6}$ and $B_{3\times6}$ in the class of estimators (A1) which give the optimal efficiency in the class. This estimator is guaranteed to be more efficient asymptotically compared to the simple IPW estimator typically used by analysts which solves

$$\mathbb{P}_n\left\{\frac{\mathbb{1}(R=1)}{\pi_1(L)}(1,X_1,X_2)^T\{Y-\mu(\boldsymbol{X},\boldsymbol{\beta})\}\right\}=0.$$

An appeal of the proposed approach of approximating the optimal functions with linear combinations of basis functions is that it does not require specification of the full data law beyond the substantive model of interest as well as assumptions (1) and (2) to estimate the weights $\pi_1(L;\hat{\gamma})$.

5. Simulation

In this section, we report a simulation study to investigate the finite-sample properties of the proposed estimators. Independent and identically distributed $(Y,\boldsymbol{X})$ is generated where $\boldsymbol{X}=(X_1,X_2,X_3)$ follow the truncated normal distributions $X_1\sim\text{N}(\mu=1,\sigma=0.5)$, $X_2\sim\text{N}(\mu=X_1+X_1^2,\sigma=0.5)$ and $X_3\sim\text{N}(\mu=X_2+0.8X_1X_2,\sigma=0.5)$ on the support $\boldsymbol{X}\in[0,2]^3$. The binary outcome variable Y is then generated with the substantive model

$$\text{logit}\Pr(Y=1|\boldsymbol{X})=\beta_0+\beta_1X_1+\beta_2X_2+\beta_3X_3. \tag{21}$$

The generated full data are then induced with missing values following missing data model (6) with three patterns $L_{(1)}=(Y,\boldsymbol{X})$, $L_{(2)}=(Y,X_1)$, and $L_{(3)}=(X_2,X_3)$ generated under the nonmonotone MAR mechanism

$$\begin{aligned}\text{logit}\Pr\{R=2|Y,\boldsymbol{X}\}&=\{\gamma_{20}+\gamma_{21}Y+\gamma_{22}X_1\}\\ \text{logit}\Pr\{R=3|Y,\boldsymbol{X}\}&=\{\gamma_{30}+\gamma_{31}X_2+\gamma_{32}X_3\},\end{aligned} \tag{22}$$

Table 1. Absolute value of empirical bias (|Bias|), empirical root mean squared error (RMSE), average width of confidence interval (AWCI) and empirical coverage rate of Wald-type 95% confidence interval (% Cov) in estimation of $\boldsymbol{\beta}$ in the substantive model with MAR mechanism† which depends on both outcome and covariates (scenario 1), based on 1000 simulation replicates of sample sizes $n = 1000, 2000$. The true value of $\boldsymbol{\beta}$ is $(\beta_0, \beta_1, \beta_2, \beta_3) = (-2.5, 0.7, 0.8, 1.0)$.

		β_0				β_1				β_2				β_3			
n	Method	|Bias|	RMSE	AWCI	% Cov	|Bias|	RMSE	AWCI	% Cov	|Bias|	RMSE	AWCI	% Cov	|Bias|	RMSE	AWCI	% Cov
1000	Full MLE	35	282	1052	94.1	6	301	1184	95.2	6	293	1131	95.0	14	278	1048	93.8
	CC	233	490	1587	88.0	252	632	2312	93.8	152	520	2024	95.6	111	466	1752	94.2
	IPW	63	426	1574	95.0	66	608	2336	93.0	0	525	2099	95.3	21	463	1789	95.6
	AIPW	107	429	1419	92.8	54	552	1908	91.4	15	532	1962	93.6	50	474	1706	94.3
	NORM	422	651	1410	73.9	383	579	1444	77.9	52	523	1576	86.7	447	696	1475	69.3
	MICE	68	426	1539	92.2	684	893	2087	69.8	430	731	2176	82.9	1	468	1752	91.4
2000	Full MLE	12	188	738	95.4	4	211	835	94.5	9	201	797	95.4	19	192	739	94.7
	CC	248	384	1109	84.1	227	488	1629	91.6	148	397	1416	93.4	117	348	1225	93.1
	IPW	34	292	1096	94.0	7	446	1667	94.7	6	379	1473	94.7	25	332	1254	94.2
	AIPW	50	273	984	93.7	10	379	1360	92.8	2	380	1389	93.4	41	334	1199	93.0
	NORM	393	520	988	63.5	461	558	1012	56.1	83	378	1090	84.1	447	588	1026	54.9
	MICE	47	289	1071	93.6	681	792	1455	51.8	452	623	1515	71.0	8	343	1244	91.7

All entries are original values multiplied by 1000, except for the coverage rate expressed as %.
†True value of γ in the missing data model is $(\gamma_{20}, \gamma_{21}, \gamma_{22}, \gamma_{30}, \gamma_{31}, \gamma_{32}) = (-0.8, -1.8, 0.2, -1.2, 0.3, 0.3)$.

which depends on both the covariates and the outcome (scenario 1), with the probability of being a complete case $\Pr\{R = 1|Y, \boldsymbol{X}\} = 1 - \Pr\{R = 2|Y, \boldsymbol{X}\} - \Pr\{R = 3|Y, \boldsymbol{X}\}$. This mechanism may be reasonable when, for example, certain combinations of variables in a survey data increase the risks of personal identification and are withheld from the analysts due to confidentiality concerns (Molenberghs et al. 2014). The missing data process is generated from a multinomial distribution with the probabilities in (22) and only the corresponding observed data for the sampled pattern contributes to estimation. We perform 1000 replicates each with sample sizes $n = 1000, 2000$. Each simulation replicate has approximately 30% –40% of complete cases. The simulation results from the MAR missing data mechanism (22) are summarized in Table 1. The supplementary Appendix describes simulation settings for a MAR missing data mechanism which is independent of the outcome Y given covariates (scenario 2), as well as a MNAR missing data mechanism (scenario 3), which has been argued as a more natural mechanism with nonmonotone missing data (Robins and Gill 1997; Little and Rubin 2002), to investigate the properties of the proposed estimators under additional scenarios. The results for these two scenarios are summarized in Tables 2 and 3, respectively.

For IPW and AIPW estimators, the missing data models are specified as (22). The choice of basis functions for AIPW estimation includes linear and quadratic terms, and detailed description of its implementation is included in the supplementary materials. The parameters γ of the missing data models are estimated using CBE to construct the weights, and the substantive model is correctly specified as (21). We obtain the CBE estimator of γ as the posterior mean of distribution (11) with independent diffuse priors $\gamma \sim N(0, 10^2)$ and $\sigma^* = 10^{-8}$. Adaptive Gibbs sampling (Gilks, Best, and Tan 1995) was implemented through "BRugs," the R interface to the OpenBUGS MCMC software (Lunn et al. 2009). More details on the implementation as well as the sample OpenBUGS code for estimation of missing data model are included in the supplementary materials.

Fully parametric estimation approaches include the general location model with binary Y (Schafer 1997). For comparison to the proposed methods, two likelihood-based MI methods are included in the simulation: full predictive distribution sampling assuming multivariate normality (NORM) (Schafer 1997) and multivariate imputation by chained equations (MICE) (van Buuren and Oudshoorn 2000; van Buuren and Groothuis-Oudshoorn 2011) based on the variables $(Y, \boldsymbol{X})$ in the substantive and missing data models. For NORM, the imputed Y value

Table 2. Absolute value of empirical bias (|Bias|), empirical root mean squared error (RMSE), average width of confidence interval (AWCI) and empirical coverage rate of Wald-type 95% confidence interval (% Cov) in estimation of $\boldsymbol{\beta}$ in the substantive model with MAR mechanism† which depends on the covariates only (scenario 2), based on 1000 simulation replicates of sample sizes $n = 1000, 2000$. The true value of $\boldsymbol{\beta}$ is $(\beta_0, \beta_1, \beta_2, \beta_3) = (-2.5, 0.7, 0.8, 1.0)$. All entries are original values multiplied by 1000, except for the coverage rate expressed as %.

		β_0				β_1				β_2				β_3			
n	Method	|Bias|	RMSE	AWCI	% Cov	|Bias|	RMSE	AWCI	% Cov	|Bias|	RMSE	AWCI	% Cov	|Bias|	RMSE	AWCI	% Cov
1000	Full MLE	4	267	1046	95.4	1	299	1181	94.0	8	278	1129	95.7	3	257	1047	96.5
	CC	48	476	1858	95.1	30	668	2618	95.0	14	591	2298	94.5	34	516	2009	95.7
	IPW	51	491	1915	95.5	35	719	2722	94.2	17	618	2385	94.9	40	541	2078	95.1
	AIPW	136	471	1632	94.5	47	623	2137	92.4	19	656	2280	92.6	83	570	2008	93.4
	NORM	73	394	1334	89.8	125	492	1438	86.8	48	560	1536	83.1	40	480	1419	85.8
	MICE	60	400	1674	95.5	1105	1183	1834	32.5	800	990	2246	67.4	5	503	1974	93.1
2000	Full MLE	10	190	739	95.3	8	213	833	95.1	6	205	797	94.4	4	190	737	94.6
	CC	38	326	1297	95.3	6	466	1826	95.6	3	406	1597	95.4	19	364	1395	95.1
	IPW	42	338	1337	95.7	17	503	1916	94.4	3	422	1659	94.9	19	370	1442	94.9
	AIPW	68	310	1130	94.4	3	410	1502	93.5	3	427	1599	94.1	39	380	1397	94.7
	NORM	64	290	918	87.2	149	369	1112	86.3	28	394	1111	83.4	49	353	996	83.4
	MICE	84	283	1181	94.2	1086	1129	1289	10.5	773	876	1592	50.8	33	360	1385	93.6

†True value of γ in the missing data model is $(\gamma_{20}, \gamma_{21}, \gamma_{30}, \gamma_{31}, \gamma_{32}) = (-0.8, 0.2, -1.2, 0.3, 0.3)$.

Table 3. Absolute value of empirical bias (|Bias|), empirical root mean squared error (RMSE), average width of confidence interval (AWCI) and empirical coverage rate of Wald-type 95% confidence interval (% Cov) in estimation of $\boldsymbol{\beta}$ in the substantive model with MNAR mechanism† (scenario 3), based on 1000 simulation replicates of sample sizes $n = 1000, 2000$. The true value of $\boldsymbol{\beta}$ is $(\beta_0, \beta_1, \beta_2, \beta_3) = (-2.5, 0.7, 0.8, 1.0)$. All entries are original values multiplied by 1000, except for the coverage rate expressed as %.

		β_0				β_1				β_2				β_3			
n	Method	|Bias|	RMSE	AWCI	% Cov	|Bias|	RMSE	AWCI	% Cov	|Bias|	RMSE	AWCI	% Cov	|Bias|	RMSE	AWCI	% Cov
1000	Full MLE	26	277	1046	94.2	2	306	1184	94.7	11	291	1132	94.7	8	276	1048	93.8
	CC	590	706	1481	61.7	564	792	2121	82.6	104	482	1852	95.1	86	426	1609	93.7
	IPW	114	395	1432	92.4	333	642	2072	90.6	64	477	1862	95.9	47	425	1611	93.4
	AIPW	108	394	1328	89.8	271	611	1844	87.5	61	502	1805	94.0	14	443	1567	92.3
	NORM	266	538	1369	82.2	376	572	1505	80.3	154	506	1624	88.2	504	724	1467	66.3
	MICE	33	402	1444	91.2	208	593	1965	87.4	17	554	1962	89.7	107	477	1676	91.4
2000	Full MLE	9	187	738	95.0	4	213	834	95.3	5	205	798	95.2	0	189	738	93.7
	CC	611	668	1039	36.3	573	691	1490	66.5	60	334	1299	95.0	71	294	1130	94.1
	IPW	139	296	1004	90.5	332	492	1465	87.3	19	334	1307	94.8	31	289	1130	95.5
	AIPW	174	304	927	85.7	227	414	1304	89.0	28	337	1275	94.1	10	289	1103	95.2
	NORM	223	394	975	80.1	303	427	1072	77.7	87	357	1149	89.2	493	608	1024	51.7
	MICE	60	289	1014	90.6	235	454	1384	85.3	31	387	1384	89.8	112	339	1182	90.4

†True value of γ in the missing data model is $(\gamma_{20}, \gamma_{21}, \gamma_{22}, \gamma_{23}, \gamma_{24}, \gamma_{30}, \gamma_{31}, \gamma_{32}, \gamma_{33}, \gamma_{34}) = (-1.0, -1.5, 0.3, -0.1, -0.2, -1.4, 0.4, 0.4, -0.8, 0.1)$.

is dichotomized to the binary $\mathbb{1}(Y > 0)$. The imputation model for outcome Y in MICE is the correctly specified substantive model (21), while the continuous variables $\boldsymbol{X}$ are imputed via predictive mean matching (Rubin 1986). The results for both MI methods are based on five imputed datasets in each simulation replicate. Finally, we also implement unweighted complete case (CC) regression to evaluate the magnitude of selection bias and carry out MLE based on the full data (Full MLE) to assess the extent of efficiency loss due to missing data.

Under the MAR missing data mechanism (scenario 1), the CC estimator has substantial bias irrespective of sample size since the missing data model involves both the outcome and covariates, with 95% confidence intervals which undercover. The IPW and AIPW estimators, which place no further restrictions on the full data law beyond the substantive model of interest, are generally the least biased among the approaches in the simulation. Among the estimators which account for missing data, the NORM estimator is generally the most efficient, although it also has substantial bias in estimation of all parameters with noticeable undercoverage of 95% confidence intervals, since it places strong assumptions on the full data law by specifying multivariate normality. Compared to the NORM estimator, the MICE estimator is less biased for estimation of (β_0, β_3), but also less efficient. The imputation model for the binary outcome is correctly specified as the substantive model, and predictive mean matching is less vulnerable to misspecification than explicit models for the distribution of the missing values conditional on the observed ones (Andridge and Little 2010).

The efficiency of the IPW and CC estimator is similar, although in instances where the CC estimator is biased such differences of efficiency are not meaningful. At $n = 1000$ the NORM/MICE estimators sometimes show smaller RMSE compared to the IPW/AIPW estimators due to the former being more efficient, albeit more biased. Increasing sample size does not substantially mitigate this bias, and therefore at $n = 2000$ the IPW estimator shows smaller RMSE than the NORM/MICE estimators, and the AIPW estimator generally has the smallest RMSE among the methods used to account for missing data in this simulation. The empirical relative efficiency comparing AIPW to IPW is ranges from 0.7 to 0.9 based on the squared ratio of their average confidence interval widths. By incorporating incomplete cases in the estimation, the AIPW estimator in this simulation is more efficient than the MICE estimator (but still less efficient than the NORM estimator). For the current data-generating mechanism, the estimated variances of the IPW/AIPW estimators are generally biased downward compared to empirical variances in finite samples, leading to undercoverage of the 95% confidence intervals, but show improvement at the larger sample size $n = 2000$.

Under scenario 2, the NORM and MICE estimators are similarly biased with 95% confidence intervals which undercover, with the former being the most efficient among the estimators adjusting for missing data. The efficiency of AIPW and MICE estimators are similar. The CC estimator has low bias, since in this case the missing data mechanism depends on only the covariates in the regression model (Little and Rubin 2002; Little and Zhang 2011; White and Carlin 2010). Finally, under the MNAR mechanism in scenario 3, all four estimators IPW, AIPW, NORM, and MICE are biased. However, the bias of IPW/AIPW estimators is smaller than that of the CC estimator, as the missing data mechanism depends at least in part on some of the observed variables in each missing data pattern. Therefore, assuming MAR in accounting for missing data is able to mitigate some but not all selection bias.

6. Application

The empirical application concerns a study of the association between maternal exposure to highly active antiretroviral therapy (HAART) during pregnancy and birth outcomes among HIV-infected women in Botswana. A detailed description of the study cohort has been presented elsewhere (Chen et al. 2012). The entire study cohort consists of 33148 obstetrical records abstracted from 6 sites in Botswana for 24 months. Our current analysis focuses on the subset of women who were known to be HIV positive ($n = 9711$). The birth outcome of interest is preterm delivery, defined as delivery < 37 weeks gestation. 6.7% of the outcomes are unobserved. The exposure of interest is whether a woman continued HAART from before pregnancy or not (68.9% missing), and the convariate includes whether $CD4^+$ cell count is less than 200 μL (53.4% missing) (Table 4). Our goal is to correlate these factors with preterm delivery.

Table 4. Tabulation of nonmonotone missing data patterns as a percentage of total data ($n = 9711$). Missing variables are indicated by 0.

Pattern (R)	Preterm Delivery	Low CD4$^+$	Cont. HAART	%
1	1	1	1	10.5
2	0	1	1	0.7
3	1	0	1	18.3
4	0	0	1	1.6
5	1	1	0	33.9
6	0	1	0	1.5
7	1	0	0	30.6
8	0	0	0	2.9

Complete cases are given in the first pattern $R = 1$

Table 5. Analysis for outcome preterm delivery with estimated odds ratios from logistic regression. Wald 95% confidence intervals for IPW /AIPW estimators are based on estimated asymptotic variances. The standard errors for MICE and NORM are estimated from $M = 10$ imputed samples. Asterisk denote significance at 0.05 α-level.

Method	Low CD4$^+$	Cont. HAART
CC	0.78 (0.54, 1.14)	1.14 (0.83, 1.57)
IPW	0.97 (0.64, 1.45)	1.32 (0.99, 1.77)
AIPW	0.88 (0.65, 1.19)	1.31 (1.03, 1.66)*
MICE	1.15 (0.90, 1.47)	1.20 (0.94, 1.52)
NORM	1.15 (0.96, 1.39)	1.25 (1.07, 1.46)*

Complete cases consist of 10.5% of the data. We applied the proposed IPW and AIPW estimators in logistic regression as well as performed CC analysis. We also provide results for MICE and imputation assuming multivariate normality (NORM) for comparison (Table 5). MICE specifies a univariate imputation model for each of the incomplete variables preterm delivery, low CD4$^+$, and continued HAART treatment. The binary variables are imputed using logistic regressions, to provide a total of $M = 10$ imputed datasets before pooling the results in the final analysis. The imputed values for missing variables L in NORM are dichotomized to the binary values $\mathbb{1}(L > 0)$.

The IPW estimator of the logistic regression for the preterm delivery uses to estimate the weight a missing data model of the form given by (6), which includes the main effects of observed variables $L_{(m)}$ for each missing data pattern $m = 2, \ldots, 8$. Given the fairly large sample size ($n = 9711$), the results for IPW are similar using UMLE and CBE to estimate the missing data process, consistent with findings from both the simulation study and asymptotic theory. Hence, only results for CBE are presented for the IPW estimator in Table 5. The results of CBE for the missing data model parameters γ are shown in Table 6, and suggest that assuming MAR and a correctly specified missing data model, the variables preterm delivery, low CD4$^+$ count, and continued HAART all influence the missing data process, as shown by the exclusion of zero from the 95% credible intervals of their respective parameters γ for at least one missing data pattern. In particular, the dependence of the missing data process on the outcome variable preterm delivery in missing data patterns $R = 3, 5, 7$ suggests that unweighted CC estimates should differ from adjusted estimates.

The IPW and AIPW estimated odds ratio for the preterm delivery associated with the exposure of interest, continued HAART treatment from before pregnancy, increased by approximately 16%, respectively, compared to CC estimates. This association becomes significant at 0.05 α-level with AIPW estimation. The odds ratio point estimates by MICE and NORM for the association between preterm delivery and exposure of interest increased by 5% and 10% respectively compared to CC estimates. The latter tends to produce smaller standard errors, in agreement with theory and simulation study, and the association is significant at 0.05 α-level. The association between preterm delivery and low CD4$^+$ count is not significant at 0.05 α-level for all the estimation methods employed in this analysis, although the odds ratio point estimate for IPW/AIPW and MICE/NORM increased by approximately 10–20% and 47%, respectively, compared to the CC estimate. The observed range of relative efficiencies of AIPW compared to IPW, AIPW compared to MICE and AIPW compared to NORM are 0.5–0.7, 1.0–1.4, and 2.4–2.5, respectively.

Differences between MICE/NORM and IPW/AIPW estimates may reflect differences of modeling assumptions, since the former relies on model assumptions about full data univariate conditional or multivariate laws while the latter relies on a model for the missing data mechanism. In the current application, neither the conditional distribution of covariates in the full data nor the missing data model is of primary scientific interest. Although model compatibility of the conditional laws specified in MICE may be an issue (White, Royston, and Wood 2011; van Buuren 2007), simulation studies suggest that this may not be a serious problem in certain practical settings (van Buuren et al. 2006). In general, more efficient estimators can be obtained by specifying a full data model, and the NORM estimates indeed have the smallest standard errors among the methods being compared. However, in this particular application, the proposed AIPW estimator produces standard errors which are comparable to those of MICE, while at the same time entirely avoiding the need to model the full data law. This is in agreement with simulation study results which show similar efficiency between the AIPW and MICE estimators.

7. Discussion

We have proposed a simple yet general class of missing data models for nonmonotone MAR mechanisms which makes no assumption about the full data distribution. Our models are

Table 6. Posterior medians with 95% credible intervals from constrained Bayesian estimation (CBE) of missing data model parameters γ for each missing data pattern $R = m, m = 2, 3, \ldots, 8$. Asterisk denotes exclusion of zero from the credible interval.

R	intercept	Preterm delivery	Low CD4$^+$ Count	Cont. HAART
2	− 4.90(−5.43,−4.43)*		0.89(0.24, 1.47)*	− 0.44(−1.00, 0.10)
3	− 2.43(−2.55,−2.31)*	0.20(0.09, 0.32)*		1.08(0.96, 1.19)*
4	− 4.77(−5.16,−4.37)*			0.80(0.36, 1.25)*
5	− 0.51(−0.56,−0.46)*	− 0.48(−0.58, −0.38)*	− 0.47(−0.59, −0.36)*	
6	− 4.26(−4.44,−4.08)*		0.39(−0.11, 0.81)	
7	− 0.90(−0.96,−0.85)*	0.34(0.25, 0.44)*		
8	− 3.51(−3.63,−3.39)*			

explicit in their dependence on only the observed variables, and the proposed IPW estimator can easily be implemented using existing software. The article makes two important contributions, first we describe a simple UMLE approach to estimate the missing data mechanism that is straightforward to implement although that may suffer from convergence issues in small samples. Our second contribution offers a remedy to failure of UMLE by introducing a constrained Bayesian estimator which circumvents any potential convergence difficulty encountered with UMLE. Another contribution shows that AIPW can achieve substantial gains in efficiency over simple IPW estimators by recovering information from incomplete cases, while avoiding having to model the full data distribution. Assuming correct specification of the model for nonresponse, the proposed IPW/AIPW estimators corrects the bias of CC analysis and may be used whenever one has available a full data estimating equation and the nonmonotone MAR missing data mechanism potentially depends also on the outcome. The constrained Bayesian estimator is guaranteed to produce valid probability weights for subsequent estimation of a full data regression or other functionals of interest. In addition, constrained Bayesian estimation of the missing data model parameters is able to elucidate important variables that influence the missingness process by studying the properties of the Monte Carlo approximations to their posterior distributions (e.g., posterior medians and 95% credible intervals, as illustrated in the application). Constrained Bayesian estimation under a parametric model for the missing data process also allows for sensitivity analysis under a unified framework to explore the possibility that the process is MNAR, which is part of future work.

Finally, Robins and Gill have argued that the class of RMM models represents the most general plausible physical mechanism for generating nonmonotone missing data (Robins and Gill 1997). Therefore, they have effectively argued that any model within our class that is not RMM may be difficult to motivate scientifically. We emphasize that the perspective we have presented is completely agnostic as to whether a particular submodel of MAR may be more scientifically meaningful than another; in fact, RMM, like any other submodel of MAR, can be accommodated by the proposed approach, but would require placing additional constraints while sampling from the posterior, to ensure that one remains within the submodel. This will necessarily result in a more complicated fitting procedure, with little apparent benefit for bias reduction or efficiency gain. This is because, as well established in the missing data literature, it is generally advisable for efficiency considerations in IPW estimation under MAR, that one estimates the probability of a complete case using as richly parameterized a regression as empirically feasible (Robins, Rotnitzky, and Zhao 1994). This implies that even if RMM is correctly specified, one would generally benefit from including correlates of the full data estimating equation into a model for the missing data mechanism, even if such variables do not necessarily correlate with the missing data process. On the other hand, care should be taken when variables predictive of nonresponse, but weakly correlate with the full data estimating equation, are included in the missing data model, since their inclusion tends to reduce precision of the full data parameter estimates with little benefit in terms of bias (Little and Vartivarian 2005). We believe such efficiency considerations trump any concern for scientific interpretation of the model for the missing data process, particularly since after all, the missing data process is technically a nuisance parameter not of primary scientific interest.

Supplementary Materials

Additional results: Supplement provides results described in the article and detailed description of implementing the AIPW estimator, sample OpenBUGS code for the estimation of weights as well as additional results in the simulation study. (pdf).

References

Andridge, R. R., and Little, R. J. A. (2010), "A Review of Hot Deck Imputation for Survey Non-Response," *International Statistical Review*, 78, 40–64. [376]

Chen, J. Y., Ribaudo, H. J., Souda, S., Parekh, N., Ogwu, A., Lockman, S., Powis, K., Dryden-Peterson, S., Creek, T., Jimbo, W., Madidimalo, T., Makhema, J., Essex, M., and Shapiro, R. L. (2012), "Highly Active Antiretroviral Therapy and Adverse Birth Outcomes Among Hiv-Infected Women in Botswana," *The Journal of Infectious Diseases*, 206, 1695–1705. [376]

Chu, H., and Cole, S. R. (2010), "Estimation of Risk Ratios in Cohort Studies with Common Outcomes: A Bayesian Approach," *Epidemiology*, 21, 855–862. [372]

Chu, H., and Cole, S. R. (2011), "Estimating the Relative Excess Risk due to Interaction: A Bayesian Approach," *Epidemiology*, 22, 242–248. [372]

Gelfand, A. E., Smith, A. F. M., and Lee, T.-M. (1992), "Bayesian Analysis of Constrained Parameter and Truncated Data Problems Using Gibbs Sampling," *Journal of the American Statistical Association*, 87, 523–532. [370,372]

Gilks, W., Best, N., and Tan, K. (1995), "Adaptive Rejection Metropolis Sampling Within Gibbs Sampling," *Applied Statistics*, 44, 455–472. [375]

Horton, N. J., and Laird, N. M. (1999), "Maximum Likelihood Analysis of Generalized Linear Models With Missing Covariates," *Statistical Methods in Medical Research* 8, 37–50. [369]

Horton, N. J., and Lipsitz, S. R. (2001), "Multiple Imputation in Practice: Comparison of Software Packages for Regression Models With Missing Variables," *The American Statistician*, 55, 244–254. [369]

Horvitz, D., and Thompson, D. (1952), "A Generalization of Sampling Without Replacement From a Finite Universe," *Journal of the American Statistical Association*, 47, 663–685. [369]

Ibrahim, J. G., and Chen, M.-H. (2000), "Power Prior Distributions for Regression Models," *Statistical Science*, 15, 46–60. [369]

Ibrahim, J. G., Chen, M.-H., and Lipsitz, S. R. (2001), "Missing Responses in Generalised Linear Mixed Models When the Missing Data Mechanism is Nonignorable," *Biometrika*, 88, 551–564. [370]

Ibrahim, J. G., Chen, M.-H., and Lipsitz, S. R. (2002), "Bayesian Methods for Generalized Linear Models With Covariates Missing at Random," *Canadian Journal of Statistics*, 30, 55–78. [369]

Ibrahim, J. G., Chen, M.-H., Lipsitz, S. R., and Herring, A. H. (2005), "Missing-Data Methods for Generalized Linear Models: A Comparative Review," *Journal of the American Statistical Association*, 100, 332–346. [369]

Li, L., Shen, C., Li, X., and Robins, J. M. (2013), "On Weighting Approaches for Missing Data," *Statistical Methods in Medical Research*, 22, 14–30. [369]

Little, R., and David, M. (1983), "Weighting Adjustments for Non-Response in Panel Surveys," *Bureau of the Census Technical Report*. [371]

Little, R. J. A., and Vartivarian, S. (2005), "Does Weighting for Nonresponse Increase the Variance of Survey Means?," *Survey Methodology*, 31, 147–177. [378]

Little, R. J., and Rubin, D. B. (2002), *Statistical Analysis with Missing Data*, New York: Wiley. [369,370,371,375,376]

Little, R. J., and Zhang, N. (2011), "Subsample Ignorable Likelihood for Regression Analysis With Missing Data," *Journal of the Royal Statistical Society*, Series C, 60, 591–605. [369,376]

Lunn, D., Spiegelhalter, D., Thomas, A., and Best, N. (2009), "The Bugs Project: Evolution, Critique and Future Directions," *Statistics in Medicine*, 28, 3049–3067. [375]

Mealli, F., and Rubin, D. B. (2015), "Clarifying Missing at Random and Related Definitions, and Implications when Coupled with Exchangeability," *Biometrika*, 102, 995–1000. [370]

Molenberghs, G., Fitzmaurice, G., Kenward, M., Tsiatis, A., and Verbeke, G. (2014), *Handbook of Missing Data Methodology*, Handbooks of Modern Statistical Methods, Boca Raton, FL: CRC Press. [370,375]

Molenberghs, G., Thijs, H., Jansen, I., and Beunckens, C. (2004), "Analyzing Incomplete Longitudinal Clinical Trial Data," *Biostatistics*, 5, 445–464. [369]

Robins, J. M., and Gill, R. D. (1997), "Non-Response Models for the Analysis of Non-Monotone Ignorable Missing Data," *Statistics in Medicine*, 16, 39–56. [370,371,375,378]

Robins, J. M., and Ritov, Y. (1997), "Toward a Curse of Dimensionality Appropriate (coda) Asymptotic Theory for Semi-Parametric Models," *Statistics in Medicine*, 16, 285–319. [370]

Robins, J. M., Rotnitzky, A., and Scharfstein, D. O. (1999), "Sensitivity Analysis for Selection Bias and Unmeasured Confounding in Missing Data and Causal Inference Models," in Statistical Models in Epidemiology: The Environment and Clinical Trials, eds. M. E. Halloran and D. Berry, New York: Springer-Verlag. [370]

Robins, J. M., Rotnitzky, A., and Zhao, L. P. (1994), "Estimation of Regression Coefficients When Some Regressors are Not Always Observed," *Journal of the American Statistical Association*, 89, 846–866. [369,370,371,373,374,378]

Rotnitzky, A., Robins, J. M., and Scharfstein, D. O. (1998), "Semiparametric Regression for Repeated Outcomes With Nonignorable Nonresponse," *Journal of the American Statistical Association*, 93, 1321–1339. [370]

Rubin, D. B. (1976), "Inference and Missing Data," *Biometrika*, 63, 581–592. [369,370]

——— (1977), "Formalizing Subjective Notions About the Effect of Nonrespondents in Sample Surveys," *Journal of the American Statistical Association*, 72, 538–543. [369]

——— (1986), "Statistical Matching using File Concatenation With Adjusted Weights and Multiple Imputations," *Journal of Business and Economic Statistics*, 4, 87–94. [376]

Rubin, D. B., Stern, H. S., and Vehovar, V. (1995), "Handling "Don't Know" Survey Responses: The Case of the Slovenian Olebiscite," *Journal of the American Statistical Association* 90, 822–828. [370]

Schafer, J. (1997), *Analysis of Incomplete Multivariate Data*, Boca Raton, FL: Chapman and Hall. [375]

——— (1999), "Multiple Imputation: A Primer," *Statistical Methods in Medical Research* 8, 3–15. [369]

——— (2003), "Multiple Imputation in Multivariate Problems When the Imputation and Analysis Models Differ," *Statistica Neerlandica*, 57, 19–35. [370]

Schafer, J. L., and Graham, J. W. (2002), "Missing Data: Our View of the State of the Art," *Psychological Methods*, 7, 147–177. [370]

Seaman, S., Galati, J., Jackson, D., and Carlin, J. (2013), "What is Meant by Missing at Random?" *Statistical Science* 28, 257–268. [370]

Seaman, S. R., and White, I. R. (2013), "Review of Inverse Probability Weighting for Dealing With Missing Data," *Statistical Methods in Medical Research* 22, 278–295. [369]

Siddiqui, O., and Ali, M. W. (1998), "A Comparison of the Random-Effects Pattern Mixture Model With Last-Observation-Carried-Forward (locf) Analysis in Longitudinal Clinical Trials With Dropouts," *Journal of Biopharmaceutical Statistics*, 8, 545–563. [369]

Troxel, A. B., Lipsitz, S. R., and Harrington, D. P. (1998), "Marginal Models for the Analysis of Longitudinal Measurements With Nonignorable Non-Monotone Missing Data," *Biometrika* 85, 661–672. [370]

Tsiatis, A. (2006), *Semiparametric Theory and Missing Data*, New York: Springer. [369,370,371,374]

van Buuren, S. (2007), "Multiple Imputation of Discrete and Continuous Data by Fully Conditional Specification," *Statistical Methods in Medical Research*, 16, 219–242. [377]

van Buuren, S., Brand, J., Oudshoorn, C., and Rubin, D. (2006), "Fully Conditional Specification in Multivariate Imputation," *Journal of Statistical Computation and Simulation*, 76, 1049–1064. [377]

van Buuren, S., and Groothuis-Oudshoorn, K. (2011), "mice: Multivariate Imputation by Chained Equations in r," *Journal of Statistical Software* 45, 1–67. [375]

van Buuren, S., and Oudshoorn, C. (2000), "Multivariate Imputation by Chained Equations: Mice v1.0 users Manual," *Leiden: TNO Prevention and Health*. [375]

van der Laan, M. J., and Robins, J. M. (2003), *Unified Methods for Censored Longitudinal Data and Causality*, New York: Springer. [369,370,374]

van der Vaart, A. (1998), *Asymptotic Statistics*, Cambridge, UK: Cambridge University Press. [373]

White, I. R., and Carlin, J. B. (2010), "Bias and Efficiency of Multiple Imputation Compared With Complete-Case Analysis for Missing Covariate Values," *Statistics in Medicine* 29, 2920–2931. [376]

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. [377]