Journal of the American Statistical Association

September 2022 ▸ Volume 117 ▸ Number 539

Taylor & Francis
Taylor & Francis Group

Journal of the American Statistical Association

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/uasa20

Semiparametric Inference for Nonmonotone Missing-Not-at-Random Data: The No Self-Censoring Model

Daniel Malinsky, Ilya Shpitser & Eric J. Tchetgen Tchetgen

JOURNAL OF THE AMERICAN STATISTICAL ASSOCIATION
2022, VOL. 117, NO. 539, 1415–1423: Theory and Methods
https://doi.org/10.1080/01621459.2020.1862669

Semiparametric Inference for Nonmonotone Missing-Not-at-Random Data: The No Self-Censoring Model

Daniel Malinsky[a], Ilya Shpitser[b], and Eric J. Tchetgen Tchetgen[c]

[a]Department of Biostatistics, Columbia University, New York, NY; [b]Department of Computer Science, Johns Hopkins University, Baltimore, MD; [c]Department of Statistics, The Wharton School of the University of Pennsylvania, Philadelphia, PA

ABSTRACT
We study the identification and estimation of statistical functionals of multivariate data missing nonmonotonically and not-at-random, taking a semiparametric approach. Specifically, we assume that the missingness mechanism satisfies what has been previously called "no self-censoring" or "itemwise conditionally independent nonresponse," which roughly corresponds to the assumption that no partially observed variable directly determines its own missingness status. We show that this assumption, combined with an odds ratio parameterization of the joint density, enables identification of functionals of interest, and we establish the semiparametric efficiency bound for the nonparametric model satisfying this assumption. We propose a practical augmented inverse probability weighted estimator, and in the setting with a (possibly high-dimensional) always-observed subset of covariates, our proposed estimator enjoys a certain double-robustness property. We explore the performance of our estimator with simulation experiments and on a previously studied dataset of HIV-positive mothers in Botswana. Supplementary materials for this article are available online.

ARTICLE HISTORY
Received September 2019
Accepted December 2020

KEYWORDS
Double-robustness; Identification; Missing data; Missingness-not-at-random

1. Introduction

Missing data are a pervasive feature of observations in almost every area of scientific study. Techniques for accounting for missing data in settings where the missingness depends only on observed data ("missingness-at-random" or MAR) are well-developed (Robins, Rotnitzky, and Zhao 1994; Tsiatis 2006; Little and Rubin 2014). The situation is substantially more difficult, however, in multivariate settings when the probability of missingness may depend on unobserved parts of the data ("missingness-not-at-random" or MNAR) and when the patterns of missingness are nonmonotone (see Robins 1997; Rotnitzky and Robins 1997; Scharfstein, Rotnitzky, and Robins 1999). Nonmonotone missingness may occur, for example, when there are complex patterns of drop-out/re-entry in a longitudinal study or when, as is often the case in practice, exposure, covariate, and outcome variables may each be subject to missing values with no specific pattern across the sample. Robins and Gill (1997) and Vansteelandt, Rotnitzky, and Robins (2007) argued that missingness-not-at-random should be expected in nonmonotone longitudinal settings if, for example, a research subject's decision to re-enter depends in part on the evolution of attributes that would have been recorded in missed visits. MNAR is also common in settings where social stigma makes nonresponse to some research questions (e.g., about HIV status, sexual activity, or drug use) dependent on other imperfectly observed or censored questions (Marra et al. 2017).

Recent work on nonparametric or semiparametric inference in nonmonotone MNAR settings has proceeded by positing some set of restrictions on the missingness mechanism sufficient for identifying a functional or parameter of interest (Rotnitzky, Robins, and Scharfstein 1998; Robins, Rotnitzky, and Scharfstein 2000; Vansteelandt, Rotnitzky, and Robins 2007; Zhou, Little, and Kalbfleisch 2010; Li et al. 2013; Shpitser 2016; Sadinle and Reiter 2017; Sun et al. 2018; Tchetgen Tchetgen, Wang, and Sun 2018). We adopt the identifying assumption introduced in Shpitser (2016) and Sadinle and Reiter (2017)—called "no self-censoring" in the former and "itemwise conditionally independent nonresponse" in the latter—which allows for both missingness-not-at-random and nonmonotonicity. Specifically, we assume only that each measured but sometimes missing variable is conditionally independent of its missingness indicator given all other variables (which may also be missing) and all other missingness indicators. Mechanistically interpreted, this means that no variable is a direct cause of its own missingness status. Our parameter of interest is any measurable function of the full data distribution (e.g., a marginal mean, correlation, or regression parameter), which is identified from the observed data under this assumption.

Shpitser (2016) introduced a pseudolikelihood-based inverse probability weighted (IPW) estimator for the "no self-censoring" model. His approach is consistent but (as is common for IPW estimators) not efficient and relies on a specific parameterization of the missingness mechanism. Sadinle and Reiter (2017) introduced a modeling strategy that

CONTACT Daniel Malinsky dsm2128@cumc.columbia.edu Department of Biostatistics, Columbia University, New York, NY 10027.
Supplementary materials for this article are available online. Please go to www.tandfonline.com/r/JASA.

requires specifying joint distributions or multivariate kernel density estimation, which can be prohibitively challenging in practice and lacks desirable inferential guarantees such as $\sqrt{n}$-consistency and asymptotic normality for estimating a particular parameter of interest. In contrast to these approaches, we present a semiparametric analysis yielding influence function (IF) based estimators which have a number of desirable properties (Newey 1990; Bickel et al. 1993; Van der Vaart 2000; Tsiatis 2006). Furthermore, our approach exploits an odds ratio parameterization of joint densities due to Chen (2007, 2010), thereby enabling convenient and congenial specification of the various components of the likelihood. Finally, the estimator we propose benefits from a certain appealing double-robustness property, which can mitigate the threat of model misspecification in the setting where a possibly high-dimensional set of always-observed covariates is also available.

We begin by introducing our central assumption, which we show implies nonparametric identification of both the probability of each missingness pattern and our parameter of interest. Next we use semiparametric theory to derive an observed data influence function of a pathwise differentiable functional on a nonparametric full data model. Because the no self-censoring model is nonparametric saturated (as defined by Robins (1997)), this influence function is unique and efficient. We then consider the case where there is available an additional vector of always-observed covariates and demonstrate that our proposed estimator is doubly robust. We explore the performance of our estimator in simulated data, and conclude with an application to a cohort study of HIV-positive mothers in Botswana.

2. The Model

Suppose the underlying data-generating process yields iid samples of (R, L) with full data vector $L = (L_1, \dots, L_K)'$ and missingness indicators $R = (R_1, \dots, R_K)'$. R takes values in $\{0,1\}^K$ where 1 corresponds to "observed" and 0 corresponds to "missing." That is, $R_i = 1$ if L_i is observed and $R_i = 0$ otherwise. L may be continuous or discrete. In slight abuse of notation, we use equations such as $R = r$ and $R = 1$ as shorthand for $(R_1, \dots, R_K) = (r_1, \dots, r_K)$ and $(R_1, \dots, R_K) = (1, \dots, 1)^K$ (an identity vector of length K). Also for any vector A we use $A_{-i} = (A_1, \dots, A_{i-1}, A_{i+1}, \dots, A_K)'$ to denote the vector A with ith entry removed. Let $L_{(r)}$ be the subvector of the elements of L that are observed when $R = r$. The observed data is comprised of iid realizations of the vector $(R, L_{(R)})$. We use $p(\cdot)$ to denote a distribution or density function.

It is well known that the full data distribution $p(L)$ is not identified from observed data distribution $p(R, L_{(R)})$ without a restriction on the missingness mechanism. We assume the following condition holds:

Assumption 1 (No self-censoring).

$$R_i \perp\!\!\!\perp L_i \mid R_{-i}, L_{-i} \tag{1}$$

for all $i = 1, \dots, K$.

We also make the following standard assumption:

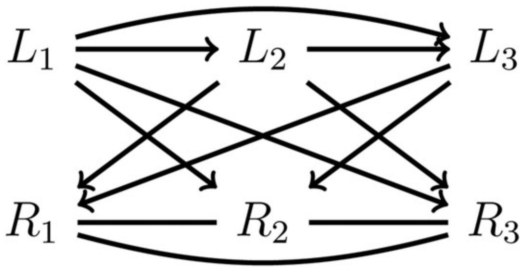

Figure 1. A chain graph representation of the no self-censoring independence model, for $K = 3$ variables.

Assumption 2 (Positivity).

$$p(R = 1|L) > \sigma > 0 \tag{2}$$

w.p.1 for some constant σ.

Shpitser (2016) and Sadinle and Reiter (2017) discussed the no self-censoring assumption (1) extensively. We only note a few features here which make the model interesting. First, the assumption does not place any restriction on the observed data distribution, that is, the model is nonparametric saturated (see Shpitser 2016; Sadinle and Reiter 2017, 2018). Second, if any of the K independence assumptions in (1) are false, the joint distribution is no longer nonparametrically identified (Mohan, Pearl, and Tian 2013). Thus, the model defined by (1) is an appealing starting point for analyzing MNAR data, either as a substantive model or in the course of sensitivity analysis.

Several authors have used graphical models to represent missingness models and establish identification results (Mohan, Pearl, and Tian 2013; Shpitser, Mohan, and Pearl 2015; Tian 2015; Mohan and Pearl 2018; Bhattacharya et al. 2019). So far, these have focused mostly on missingness models that can be represented by directed acyclic graphs. The independence model defined by (1) can be summarized graphically by a chain graph (Lauritzen 1996). A chain graph $\mathcal{G} = (V, E)$ with vertices V and edges E is a mixed graph that may contain both directed and undirected edges. We associate the vertices V in a graph $\mathcal{G}$ with the random variables $(R_1, \dots, R_K, L_1, \dots, L_K)$. Specifically, the chain graph representation of the no self-censoring model contains undirected edges between all pairs of R vertices, directed edges from each L_i to each R_j such that $j \neq i$, and directed edges among all L vertices. None of the undirected edges among R vertices may be directed without changing the model, however any of the directed edges among L vertices may be reversed without changing the model. An example is shown for $K = 3$ variables in Figure 1.

The absence of any edge between each L_i and the corresponding R_i corresponds to the independence assumption (1).[1] This follows from the pairwise Markov property for chain graphs. A vertex V_i is called a descendent of V_j if there is a directed path from V_j to V_i in $\mathcal{G}$. Let $\text{nd}_i(\mathcal{G})$ denote the set of nondescendents of vertex V_i in $\mathcal{G}$. The pairwise Markov property states that for nonadjacent vertices V_i, V_j such that $V_j \in \text{nd}_i(\mathcal{G})$, $V_i \perp\!\!\!\perp V_j \mid \text{nd}_i(\mathcal{G}) \setminus V_j$. Taking any R_i, L_i in Figure 1 to be V_i, V_j we see this corresponds exactly to the no self-censoring assumption

[1] "Self-censoring" edges $L_i \to R_i$ are also called "self-masking" in Mohan, Thoemmes, and Pearl (2018) and Tu et al. (2019).

(1). Since there are no other absent edges (nonadjacencies) in $\mathcal{G}$, the model imposes no additional independence restrictions. The MNAR model corresponding to dropping undirected edges between missingness indicators, discussed in, for example, Mohan, Pearl, and Tian (2013) and Mohan and Pearl (2018), is a submodel of ours. Mohan, Thoemmes, and Pearl (2018) used graphical models to study identifiability when (1) is violated but additional parametric assumptions hold. Though we do not emphasize graphical concepts in our results below, the graphical perspective can provide significant insight into nonparametric identifiability. In particular, a version of our first identification result in the next section can be derived by inspecting the Markov factorization for a chain graph, which we describe in the Appendix included as supplementary material to this article.

Sadinle and Reiter (2017) showed that the full data law is nonparametrically identified under (1). However, their identification (and estimation) approach requires modeling the full-data joint density. In contrast, we will establish a novel but complementary identification result and variationally independent parameterization for the probability of missingness $p(R|L)$, which makes inverse probability weighting immediately feasible and obviates the need to model the joint density $p(L)$.

3. Identification

To fix ideas, suppose that our target parameter of interest is $\beta = \mathbb{E}[B] \equiv \mathbb{E}[b(L)]$ for some known function b of the full data. Later, we consider more general parametric or semiparametric full data models. To express this parameter as a function of the observed data, we make use of an odds ratio (OR) parameterization of the joint density discussed in, for example, Chen (2007, 2010). Chen shows how to factorize an arbitrary joint density into a product of variationally independent components, namely a combination of univariate conditionals and conditional odds ratios. We make use of this factorization to show, first, that the probability of missingness $p(R|L)$ is identified for every missingness pattern. Then, the odds ratio function $\mathrm{OR}(R, L) \equiv \mathrm{OR}(R, L; r_0 = 1, l_0 = 0) = \frac{p(R|L)}{p(R=1|L)} \frac{p(R=1|L=0)}{p(R|L=0)}$ is also identified.[2] This enables identification of β. All proofs not in the main body are deferred to the Appendix.

The missingness probability $p(R|L)$ can be expressed using the odds ratio parameterization (Chen 2010, eq. (3)):

$$p(R|L) = \frac{\prod_{i=2}^{K} \mathrm{OR}(R_i, (R_1, \ldots, R_{i-1})|(R_{i+1}, \ldots, R_K) = 1, L) \prod_{i=1}^{K} p(R_i|R_{-i} = 1, L)}{\sum_r \prod_{i=2}^{K} \mathrm{OR}(r_i, (r_1, \ldots, r_{i-1})|(R_{i+1}, \ldots, R_K) = 1, L) \prod_{i=1}^{K} p(r_i|R_{-i} = 1, L)}, \quad (3)$$

where $\mathrm{OR}(R_i, (R_1, \ldots, R_{i-1})|(R_{i+1}, \ldots, R_K) = 1, L) = \frac{p(R_i|(R_{i+1},\ldots,R_K)=1,R_1,\ldots,R_{i-1},L)p(R_i=1|R_{-i}=1,L)}{p(R_i|R_{-i}=1,L)p(R_i=1|(R_{i+1},\ldots,R_K)=1,R_1,\ldots,R_{i-1},L)}$. Under the no self-censoring assumption (1), each term in this ratio can be written as a function of the observed data.

Theorem 1. $p(R|L)$ is nonparametrically identified under (1).

To provide some concrete intuition for this result, we illustrate the case where $K = 3$.

$$\begin{aligned}
&p(R_1, R_2, R_3|L)\\
&= p(R_1|R_{-1} = 1, L)p(R_2|R_{-2} = 1, L)p(R_3|R_{-3} = 1, L)\\
&\quad \times \mathrm{OR}(R_1, R_2|R_3 = 1, L)\mathrm{OR}(R_2, R_3|R_1 = 1, L)\\
&\mathrm{OR}(R_1, R_3|R_2, L)/C(L)\\
&= p(R_1|R_{-1} = 1, L)p(R_2|R_{-2} = 1, L)p(R_3|R_{-3} = 1, L)\\
&\quad \times \mathrm{OR}(R_1, R_2|R_3 = 1, L)\mathrm{OR}(R_2, R_3|R_1 = 1, L)\\
&\mathrm{OR}(R_1, R_3|R_2 = 1, L)\Gamma(R_1, R_2, R_3|L)/C(L)\\
&= p(R_1|R_{-1} = 1, L_{-1})p(R_2|R_{-2} = 1, L_{-2})p(R_3|R_{-3} = 1, L_{-3})\\
&\quad \times \mathrm{OR}(R_1, R_2|R_3 = 1, L_3)\mathrm{OR}(R_2, R_3|R_1 = 1, L_1)\\
&\mathrm{OR}(R_1, R_3|R_2 = 1, L_2)\Gamma(R_1, R_2, R_3)/C(L),
\end{aligned}$$

where

$$\begin{aligned}
C(L) &= \sum_{r_1,r_2,r_3} p(r_1|R_{-1} = 1, L)p(r_2|R_{-2} = 1, L)p(r_3|R_{-3} = 1, L)\\
&\quad \times \mathrm{OR}(r_1, r_2|R_3 = 1, L)\mathrm{OR}(r_2, r_3|R_1 = 1, L)\mathrm{OR}(r_1, r_3|r_2, L)\\
&= \sum_{r_1,r_2,r_3} p(r_1|R_{-1} = 1, L_{-1})p(r_2|R_{-2} = 1, L_{-2})p(r_3|R_{-3} = 1, L_{-3})\\
&\quad \times \mathrm{OR}(r_1, r_2|R_3 = 1, L_3)\mathrm{OR}(r_2, r_3|R_1 = 1, L_1)\\
&\quad \mathrm{OR}(r_1, r_3|R_2 = 1, L_2)\Gamma(r_1, r_2, r_3).
\end{aligned}$$

The first equality follows from (3) using the following basic identity implied by the definition of the odds ratio: $\mathrm{OR}(R_3, (R_1, R_2)|L) = \mathrm{OR}(R_1, R_3|R_2, L)\mathrm{OR}(R_2, R_3|R_1 = 1, L)$. The second equality follows since $\mathrm{OR}(R_1, R_3|R_2, L)$ can be written equivalently as $\mathrm{OR}(R_1, R_3|R_2 = 1, L)\Gamma(R_1, R_2, R_3|L)$ where $\Gamma(R_1, R_2, R_3|L) = \frac{\mathrm{OR}(R_1,R_3|R_2,L)}{\mathrm{OR}(R_1,R_3|R_2=1,L)}$, producing an expression in terms of pairwise odds ratios and a 3-way interaction on the odds ratio scale. The final equality follows by assumption (1). Each pairwise odds ratio

$$\begin{aligned}
&\mathrm{OR}(R_i, R_j|R_k = 1, L)\\
&= \frac{p(R_i|R_k = 1, R_j, L)p(R_i = 1|(R_j, R_k) = 1, L)}{p(R_i|(R_j, R_k) = 1, L)p(R_i = 1|R_k = 1, R_j, L)}\\
&= \frac{p(R_i|R_k = 1, R_j, L_{-i})p(R_i = 1|(R_j, R_k) = 1, L_{-i})}{p(R_i|(R_j, R_k) = 1, L_{-i})p(R_i = 1|R_k = 1, R_j, L_{-i})}
\end{aligned}$$

and by symmetry

$$\begin{aligned}
&\mathrm{OR}(R_i, R_j|R_k = 1, L)\\
&= \frac{p(R_j|R_k = 1, R_i, L)p(R_j = 1|(R_i, R_k) = 1, L)}{p(R_j|(R_i, R_k) = 1, L)p(R_j = 1|R_k = 1, R_i, L)}\\
&= \frac{p(R_j|R_k = 1, R_i, L_{-j})p(R_j = 1|(R_i, R_k) = 1, L_{-j})}{p(R_j|(R_i, R_k) = 1, L_{-j})p(R_j = 1|R_k = 1, R_i, L_{-j})}
\end{aligned}$$

therefore $\mathrm{OR}(R_i, R_j|R_k = 1, L) = \mathrm{OR}(R_i, R_j|R_k = 1, L_k)$ a function of only L_k. A similar symmetry argument implies that $\Gamma(R_1, R_2, R_3|L)$ is independent of L. Finally $p(R_i|R_{-i} = 1, L) = p(R_i|R_{-i} = 1, L_{-i})$ which establishes that the numerator (and therefore also the normalizing constant) is function of only observed data. We discuss a more abstract derivation which makes use of the chain graph Markov factorization in the Appendix.

As an immediate corollary to Theorem 1, we have that the odds ratio $\mathrm{OR}(R, L)$ is identified for any K.

[2] We define the odds ratio with fixed reference values $r_0 = 1$ and $l_0 = 0$ throughout and suppress the reference value in our notation for brevity.

Corollary 1. Under assumptions (1) and (2),

$$\mathrm{OR}(R,L) = \exp\{(1-R_1)\delta h_1(L_{-1}) + \cdots + (1-R_K)\delta h_K(L_{-K})\}, \tag{4}$$

where

$$\delta h_i(L_{-i}) = \log\left(\frac{p(R_i=0|R_{-i}=1,L_{-i})}{p(R_i=1|R_{-i}=1,L_{-i})}\right) - \log\left(\frac{p(R_i=0|R_{-i}=1,L_{-i}=0)}{p(R_i=1|R_{-i}=1,L_{-i}=0)}\right) \quad \text{for } i=1,\ldots,K.$$

Finally, we have the following result:

Corollary 2. Under assumptions (1) and (2), β is identified by

$$\beta = \mathbb{E}[B] = \mathbb{E}\left\{\sum_r \frac{\mathbb{E}[\mathrm{OR}(r,L)b(L)|R=1,L_{(r)}]}{\mathbb{E}[\mathrm{OR}(r,L)|R=1,L_{(r)}]}\mathbb{I}(R=r)\right\}. \tag{5}$$

Proof. By the bivariate odds ratio factorization (Chen 2007, eq. (1)), we have that for any measurable function b:

$$\mathbb{E}[B|R=r,L_{(r)}] = \frac{\mathbb{E}[\mathrm{OR}(r,L)b(L)|R=1,L_{(r)}]}{\mathbb{E}[\mathrm{OR}(r,L)|R=1,L_{(r)}]}.$$

Therefore,

$$\begin{aligned}\mathbb{E}[B] &= \sum_{l_{(r)}}\sum_r \frac{\mathbb{E}[\mathrm{OR}(r,L)b(L)|R=1,L_{(r)}]}{\mathbb{E}[\mathrm{OR}(r,L)|R=1,L_{(r)}]}p(r,l_{(r)}) \\ &= \mathbb{E}\left\{\sum_r \frac{\mathbb{E}[\mathrm{OR}(r,L)b(L)|R=1,L_{(r)}]}{\mathbb{E}[\mathrm{OR}(r,L)|R=1,L_{(r)}]}\mathbb{I}(R=r)\right\}.\end{aligned}$$

(Note that one may replace sums over values of L with integrals as appropriate for continuous components of L.) □

Thus, β can be expressed as a functional of the observed data distribution $p(R, L_{(R)})$ under assumptions (1) and (2), provided all terms in the odds ratio factorization of $p(R|L)$ are defined. A necessary condition for the latter is that all patterns of the form $R_{-i} = 1$ for each i (all covariates but one are observed) have support in the data.

4. Semiparametric Theory

Suppose that the full data law $p(L;\eta)$ is indexed by an infinite-dimensional parameter η. Of interest is a finite-dimensional parameter $\beta = \beta(\eta)$. Further, we assume the conditional model $p(R|L;\gamma)$ is indexed by an infinite-dimensional parameter γ. We are interested in deriving the efficient influence function for β in the nonparametric no self-censoring model, that is, the model satisfying (1) and (2) but otherwise unrestricted. We denote this model by $\mathcal{M}_{\text{nsc}}$. Among all regular and asymptotically linear (RAL) estimators of β in $\mathcal{M}_{\text{nsc}}$, the estimator based on the efficient influence function (that is, solving the efficient IF estimating equation with all nuisance models estimated nonparametrically) achieves, under sufficient regularity conditions, the minimum asymptotic variance and is said to achieve the semiparametric efficiency bound (Bickel et al. 1993).

In what follows we define $\pi_j(L) \equiv p(R = r_j|L)$ for missingness patterns $j = 1,\ldots,J$. We reserve J for the complete-case pattern, that is, $R = r_J = 1$. Before presenting the main result for $\mathcal{M}_{\text{nsc}}$, we present the efficient influence function for β in a different nonparametric model, not necessarily satisfying the no self-censoring assumption. This influence function is easier to derive and will later on suggest an estimator that is both easier to implement and exhibits an interesting double-robustness property. Let the conditional model $p(R|L)$ be parameterized as $\log\frac{\pi_j(L)}{\pi_J(L)} = h_j(L;\gamma)$. For the next result, we assume that the log odds ratio is a known function of L. Specifically, we assume for the moment that $h_j(L;\gamma) = h_{1,j}(L=0;\gamma) + h_{2,j}(L)$ and denote by $\mathcal{M}_{\text{odds}}$ the nonparametric model where $h_{2,j}$ is known for $j = 1,\ldots,J$. We use $\phi_{\text{full}}(\beta)$ to denote the full data influence function for β. (E.g., if $\beta = \mathbb{E}[b(L)]$ then $\phi_{\text{full}}(\beta) = b(L) - \beta$.)

Lemma 1. In $\mathcal{M}_{\text{odds}}$, the efficient influence function for β is

$$\begin{aligned}\phi_{\text{odds}}(\beta) =& \frac{\mathbb{I}(R=1)}{\pi_J(L)}\phi_{\text{full}}(\beta) \\ &+ \sum_{j=1}^{J-1}\mathbb{I}(R=r_j)\mathbb{E}[\phi_{\text{full}}(\beta)|R=r_j,L_{(r_j)}] \\ &- \frac{\mathbb{I}(R=1)}{\pi_J(L)}\sum_{j=1}^{J-1}\pi_j(L)\mathbb{E}[\phi_{\text{full}}(\beta)|R=r_j,L_{(r_j)}].\end{aligned}$$

This result is a version of Theorem 4.1 in Robins, Rotnitzky, and Scharfstein (2000), though we provide a simple and self-contained proof in the Appendix using our notation. Next, we return to the model $\mathcal{M}_{\text{nsc}}$, satisfying the no self-censoring assumption and where the odds ratio is not known a priori. Deriving the influence function for β involves two steps: first noticing that the odds ratio is point identified in $\mathcal{M}_{\text{nsc}}$ by the results in the previous section, and second "adjusting" the above IF for nonparametric estimation of the odds ratio, subject to the no self-censoring restriction.

Theorem 2. In $\mathcal{M}_{\text{nsc}}$, the efficient influence function for β is $\phi_{\text{nsc}}(\beta) = -\mathbb{E}\left[\frac{\partial}{\partial\beta}\phi_{\text{odds}}(\beta)\right]^{-1} \times \left(\phi_{\text{odds}}(\beta) + \phi_{\text{adj}}(\beta)\right)$, with $\phi_{\text{odds}}(\beta)$ from Lemma 1 and

$$\begin{aligned}\phi_{\text{adj}}(\beta) = -\sum_{i=1}^{K}&\mathbb{E}[(1-R_i)|L_{-i}] \\ &\frac{\mathbb{I}(R_{-i}=1)}{p(R_{-i}=1|L_{-i})}\left(\frac{R_i}{p(R_i=1|R_{-i},L_{-i})}-1\right) \\ &\times\frac{p(R_i=1|R_{-i},L_{-i})}{p(R_i=0|R_{-i},L_{-i})}\mathbb{E}[\Delta(R,L)|R_i=0,L_{-i}]\end{aligned}$$

with $\Delta(R,L) \equiv \phi_{\text{full}}(\beta) - \mathbb{E}[\phi_{\text{full}}(\beta)|R,L_{(R)}]$.

Therefore, the semiparametric efficiency bound in $\mathcal{M}_{\text{nsc}}$ is given by the variance of ϕ_{nsc}.

5. Double-Robustness in Settings With Always-Observed Covariates

In some settings, there may be available an additional set of always-observed covariates X. For example, X may consist of baseline measurements (with no missing values) in a longitudinal study with complex patterns of missingness at follow-up times $1,\ldots,K$. We may assume that our fundamental identifying assumptions on the missingness mechanism (1) and (2) hold conditional on X, that is,

$$R_i \perp\!\!\!\perp L_i | R_{-i}, L_{-i}, X \tag{6}$$

for $i = 1, \ldots, K$ and

$$p(R = 1|L, X) > \sigma' > 0 \tag{7}$$

w.p.1 for some constant σ'. Versions of Lemma 1 and Theorem 2 hold under these assumptions, where X is added to the conditioning set in all appropriate places. In particular, the conditional odds ratio $\mathrm{OR}(R, L|X) = \exp\{(1-R_1)\delta h_1(L_{-1}, X) + \cdots + (1 - R_K)\delta h_K(L_{-K}, X)\}$ where $\delta h_i(L_{-i}, X) = \log(\frac{p(R_i=0|R_{-i}=1,L_{-i},X)}{p(R_i=1|R_{-i}=1,L_{-i},X)}) - \log(\frac{p(R_i=0|R_{-i}=1,L_{-i}=0,X)}{p(R_i=1|R_{-i}=1,L_{-i}=0,X)})$, $\pi_j(L, X) = p(R = r_j|L, X)$, and likewise for other terms which were previously considered only functions of L.

It follows immediately that

$$\begin{aligned}\phi_{\text{odds}}(\beta) = {} & \frac{\mathbb{I}(R = 1)}{\pi_J(L, X)}\phi_{\text{full}}(\beta) \\ & + \sum_{j=1}^{J-1} \mathbb{I}(R = r_j)\mathbb{E}[\phi_{\text{full}}(\beta)|R = r_j, L_{(r_j)}, X] \\ & - \frac{\mathbb{I}(R = 1)}{\pi_J(L, X)} \sum_{j=1}^{J-1} \pi_j(L, X)\mathbb{E}[\phi_{\text{full}}(\beta)|R = r_j, L_{(r_j)}, X]\end{aligned}$$

and replacing assumptions (1) and (2) with (6) and (7), $\phi_{\text{nsc}}(\beta) = -\mathbb{E}\left[\frac{\partial}{\partial\beta}\phi_{\text{odds}}(\beta)\right]^{-1} \times \left(\phi_{\text{odds}}(\beta) + \phi_{\text{adj}}(\beta)\right)$ with $\phi_{\text{odds}}(\beta)$ as above and

$$\begin{aligned}\phi_{\text{adj}}(\beta) = -\sum_{i=1}^{K} & \mathbb{E}[(1 - R_i)|L_{-i}, X] \\ & \frac{\mathbb{I}(R_{-i} = 1)}{p(R_{-i} = 1|L_{-i}, X)}\left(\frac{R_i}{p(R_i = 1|R_{-i}, L_{-i}, X)} - 1\right) \\ & \times \frac{p(R_i = 1|R_{-i}, L_{-i}, X)}{p(R_i = 0|R_{-i}, L_{-i}, X)}\mathbb{E}[\Delta(R, L, X)|R_i = 0, L_{-i}, X],\end{aligned}$$

where $\Delta(R, L, X) \equiv \phi_{\text{full}}(\beta) - \mathbb{E}[\phi_{\text{full}}(\beta)|R, L_{(R)}, X]$.

Interestingly, in the setting with always-observed covariates X we have an estimator that is doubly robust. Specifically, the estimating function $\phi_{\text{odds}}(\beta)$ above is mean-zero at the true value of β in the union model where the odds ratio is correctly specified and for each pattern either the pattern probability $\pi_j(L, X)$ or pattern mixture regression model $\mathbb{E}[B|R = r_j, L_{(r_j)}, X]$ is correctly specified, but possibly not both. Let $\mathcal{M}_{\text{OR}}$ denote the model where $\mathrm{OR}(R, L|X)$ is correctly specified, $\mathcal{M}_{\pi,j}$ denote the model $p(R = r_j|L = 0, X; \psi_j)$ parameterized by ψ_j and $\mathcal{M}_{\text{PM},j}$ denote the model $\mathbb{E}[B|R = r_j, L_{(r_j)} = 0, X; \mu_j]$ parameterized by μ_j. Define the union model $\mathcal{M}_{\text{union}} = \cap_j \mathcal{M}_j$, where $\mathcal{M}_j = (\mathcal{M}_{\pi,j} \cup \mathcal{M}_{\text{PM},j}) \cap \mathcal{M}_{\text{OR}}$. We use $\psi_0 = (\psi_{1,0}, \ldots, \psi_{J,0})$ and $\mu_0 = (\mu_{1,0}, \ldots, \mu_{J,0})$ to denote the true parameter vectors.

Theorem 3. Let $\phi_{\text{odds}}(\beta, \mathrm{OR}, \mu, \psi)$ as defined above. When $(\psi_0, \mu_0) \in \mathcal{M}_{\text{union}}$, $\mathbb{E}[\phi_{\text{odds}}(\beta, \mathrm{OR}, \mu_0, \psi_0)] = 0$.

Note that the double-robustness property obtained in Theorem 3 requires that the odds ratio $\mathrm{OR}(R, L|X)$ is a known function of L and X. Since the odds ratio appears in both the pattern mixture regressions and the pattern probabilities $\pi_j(L, X)$, double-robustness in this setting does not protect against *arbitrary* misspecification of either the regression models or pattern probabilities: only components of these models *variationally independent of the odds ratio* may be misspecified without necessarily sacrificing unbiasedness of the estimating equation. Specifically, this implies that at most one of $\pi_j(L = 0, X)$ or the regression function $\mathbb{E}[\phi_{\text{full}}|R = r_j, L_{(r_j)} = 0, X] = \frac{\mathbb{E}[\mathrm{OR}(r_j,L|X)\phi_{\text{full}}|R=1,L_{(r_j)}=0,X]}{\mathbb{E}[\mathrm{OR}(r_j,L|X)|R=1,L_{(r_j)}=0,X]}$ for each pattern may be misspecified. Furthermore, the quantifier over patterns means that for some patterns one may correctly specify only the pattern probability and for other patterns one may only specify the pattern mixture regression without sacrificing unbiasedness.

It is instructive to contrast our double-robustness result with another recently proposed doubly robust estimator for a MNAR model, the "discrete choice model" (DCM) estimator in Tchetgen Tchetgen, Wang, and Sun (2018). The model for the missingness mechanism introduced in that paper is motivated by some behavioral assumptions underlying observed patterns of nonresponse in the data. In that model (which is neither properly a superset nor subset of our model $\mathcal{M}_{\text{nsc}}$), Tchetgen Tchetgen, Wang, and Sun (2018) proposed an estimator which is doubly robust in the sense of requiring that either the pattern mixture regression or missingness mechanism is correctly specified. In our case we also require that the odds ratio is correctly specified for each pattern. However, in the special setting where the data only contains complete cases and every "leave-one-out" pattern (i.e., patterns where L_i is missing but L_{-i} is observed, for each i), then the no self-censoring and DCM models coincide. That is, if every missingness pattern besides the complete case and "leave-one-out" patterns have zero probability, then assumption (1) and the DCM independence assumption place exactly the same restriction on the missingness mechanism. Tchetgen Tchetgen, Wang, and Sun (2018) expressed the DCM assumption as $L_{(-r)}|R = r, L_{(r)} \sim L_{(-r)}|R = 1, L_{(r)}$ for all $r \neq 1$ where $L_{(-r)}$ denotes the unobserved subvector of L when $R = r$ (see also Little 1993). With only complete cases and "leave-one-out" patterns this is simplifies to $L_i|R_i = 0, R_{-i} = 1, L_{-i} \sim L_i|R = 1, L_{-i}$. In the same setting, the no self-censoring assumption $L_i|R_i = 0, R_{-i}, L_{-i} \sim L_i|R_i = 1, R_{-i}, L_{-i}$ amounts to $L_i|R_i = 0, R_{-i} = 1, L_{-i} \sim L_i|R = 1, L_{-i}$. Therefore here the models coincide and thus have the same influence function. Moreover, in this restricted pattern setting the odds ratio function $\mathrm{OR}(R, L)$ is only involved in the missingness mechanism but not in the pattern mixture regression models, so an estimator based on ϕ_{odds} recovers the same double-robustness property in Tchetgen Tchetgen, Wang, and Sun (2018).

In the next section, we propose an estimator for β based on parametric specification of each component of the estimating equation. When we use a doubly robust estimator for the odds ratio components, the resulting estimator for β will be doubly robust in the sense just described.

6. The Proposed Estimator

In applications, it is common to specify parametric models for the odds ratio as well as the pattern probabilities $\pi_j(L, X)$ and the pattern mixture regressions, particularly if L has more than two continuous components. A convenient choice may be to assume a logistic model for R_i in terms of two components, one of which appears only in the odds ratio and both of which appear in the pattern probabilities $\pi_j(L, X)$. For example,

$$\begin{aligned} \text{logit} p(R_i = 1 | R_{-i} = 1, L_{-i}, X; \psi_i) \\ = \delta h_i(L_{-i}, X; \psi_{i,LX}) + h_{i,X}(X; \psi_{i,X}), \end{aligned} \tag{8}$$

where $\psi_i = (\psi_{i,LX}, \psi_{i,X})'$. Note that $\delta h_i(L_{-i}, X; \psi_{i,LX})$ is a component of the odds ratio and so $\delta h_i(L_{-i}, X; \psi_{i,LX}) = 0$ when $L_{-i} = 0$. The second component $h_{i,X}(X; \psi_{i,X})$ is a function of X which is needed for the $\pi_j(L, X)$ but does not appear in the odds ratio. The parameterized odds ratio is then $\text{OR}(\cdot; \psi_{LX})$ where $\psi_{LX} \equiv (\psi_{1,LX}, \ldots, \psi_{K,LX})'$ and we write $\widehat{\text{OR}}(\cdot) \equiv \text{OR}(\cdot; \hat{\psi}_{LX})$ for the estimated odds ratio. Also let $\psi_X \equiv (\psi_{1,X}, \ldots, \psi_{K,X})'$, $\psi \equiv (\psi_X, \psi_{LX})'$, and denote each estimated pattern probability by $\pi_j(L, X; \hat{\psi})$. We use $\mu \equiv (\mu_1, \mu_2)'$ to parameterize the pattern mixture regression functions.

We propose a straightforward augmented IPW (AIPW) estimator for β, where the augmentation term incorporates information from all the missingness patterns. Denote by $\hat{\beta}_{\text{AIPW}}$ the solution to

$$\begin{aligned} \mathbb{P}_n \Bigg(& \frac{\mathbb{I}(R=1)}{\pi_J(L, X; \hat{\psi})} \phi_{\text{full}}(\beta) + \sum_{j=1}^{J-1} \mathbb{I}(R = r_j) \\ & \frac{\mathbb{E}[\widehat{\text{OR}}(r_j, L|X) \phi_{\text{full}}(\beta) | R = 1, L_{(r_j)}, X; \hat{\mu}_1]}{\mathbb{E}[\widehat{\text{OR}}(r_j, L|X) | R = 1, L_{(r_j)}, X; \hat{\mu}_2]} \\ & - \frac{\mathbb{I}(R=1)}{\pi_J(L, X; \hat{\psi})} \sum_{j=1}^{J-1} \pi_j(L, X; \hat{\psi}) \\ & \frac{\mathbb{E}[\widehat{\text{OR}}(r_j, L|X) \phi_{\text{full}}(\beta) | R = 1, L_{(r_j)}, X; \hat{\mu}_1]}{\mathbb{E}[\widehat{\text{OR}}(r_j, L|X) | R = 1, L_{(r_j)}, X; \hat{\mu}_2]} \Bigg) = 0, \end{aligned}$$

where $\mathbb{P}_n$ denotes the sample average. This (empirically) solves the estimating equation $\mathbb{E}[\phi_{\text{odds}}(\beta; \hat{\psi}_X, \hat{\mu}, \widehat{\text{OR}})] = 0$ with estimators for all nuisance functions plugged-in. To achieve double-robustness, we must use an estimator of the odds ratio which is consistent in the union model. One such estimator, proposed in Tchetgen Tchetgen, Robins, and Rotnitzky (2010) and Tan (2019) for a full data semiparametric problem, can readily be adapted to the missing data setting. The doubly robust estimator is $\widehat{\text{OR}}_{dr} \equiv \text{OR}(\cdot; \hat{\psi}_{LX})$ where each $\hat{\psi}_{i,LX}$ solves the estimating equation $\mathbb{P}_n(r(\psi_{i,LX}; \hat{\psi}_{i,X}, \hat{\mu})) = 0$ and

$$\begin{aligned} r(\psi_{i,LX}; \hat{\psi}_{i,X}, \hat{\mu}) = & \Big(R_i - \text{expit}(h_{i,X}(X; \hat{\psi}_{i,X})) \Big) \\ & \left(\frac{\partial \delta h_i(L_{-i}, X; \psi_{i,LX})}{\partial \psi_{i,LX}} - \mathbb{E}\left[\frac{\partial \delta h_i(L_{-i}, X; \psi_{i,LX})}{\partial \psi_{i,LX}} | X; \hat{\mu} \right] \right) \\ & \times \exp\left\{ -\delta h_i(L_{-i}, X; \psi_{i,LX}) \right\} \mathbb{I}(R_{-i} = 1). \end{aligned}$$

The resulting estimator $\hat{\psi}_{LX}$ is consistent if either $p(R_i = 1 | R_{-i} = 1, L_{-i} = 0, X)$ or $\mathbb{E}\left[\frac{\partial \delta h_i(L_{-i}, X; \psi_{i,LX})}{\partial \psi_{i,LX}} | R = 1, X \right]$ for each $i = 1, \ldots, K$ is correctly specified. In practice, a common choice of functional form for $\delta h_i(L_{-i}, X; \psi_{i,LX})$ is linear in L_{-i} (as is assumed in Tan (2019)) so the derivatives reduce to L_{-i}.

Let $V(\cdot)$ be the vector of stacked estimating equations for parameters $\hat{\beta}_{\text{AIPW}}$, $\hat{\psi}_X$, $\hat{\mu}$, and $\widehat{\text{OR}}_{dr}$ using the doubly robust odds ratio estimator above. Let $\Omega = (\beta, \psi_X, \mu, \text{OR})'$ be the combined set of parameters. Ultimately our procedure solves the estimating equation $\mathbb{P}_n \left[V(\widehat{\Omega}) \right] = 0$.

Theorem 4. In the union model $\mathcal{M}_{\text{union}}$, $\widehat{\Omega}$ is consistent and asymptotically normal with influence function $\mathbb{E}\left[\frac{\partial V(\Omega)}{\partial \Omega} \right]^{-1} V(\Omega)$.

Thus, the proposed estimator has the benefit of being simple to implement and doubly robust with respect to the always-observed covariates. We note that implementing an estimator based on the nonparametric IF ϕ_{nsc} would be asymptotically more efficient when all parametric models are correctly specified, but also considerably more complicated to specify correctly because of the $p(R_{-i} = 1 | L_{-i}, X)$ terms; each of these would require correctly specifying a joint distribution and then marginalizing by integrating or summing L_i (that is, calculating $p(L_{-i}) = \int_{l_i} p(L) dL_i$). As far as we are able to determine, although locally semiparametric efficient in $\mathcal{M}_{\text{nsc}}$, an estimator based on ϕ_{nsc} fails to be doubly robust as the entire missing data mechanism must be correctly specified. So, in the interests of ease-of-implementation and double-robustness, we propose the simpler $\hat{\beta}_{\text{AIPW}}$ and explore its performance with simulations in the next section.

7. Simulation Study

In our simulations, we specifically consider the case where $K = 3$ and $\beta \equiv \mathbb{E}[L_3]$, that is, we are simply interested in the marginal mean of the "outcome" variable, L_3. First we examine the setting where all variables are sometimes missing (X is empty) and then the setting with an always-observed vector $X = (X_1, X_2)$. In both cases, we sample (R, L, X) from a conditional Gaussian chain graph model satisfying the no self-censoring assumption; that is, missingness patterns $R = r$ are sampled according to a multinomial distribution and $L, X | R = r$ is normal $N(\mu_0(r), \Sigma_0)$. With this parametric data-generating process, imposing the no self-censoring assumption amounts to setting certain mean parameters (interaction terms) to zero; see Højsgaard, Edwards, and Lauritzen (2012, pp. 119–120). The precise parameter values chosen for the simulation study are detailed in the Appendix. We also carried out versions of the following simulations with binary data; those results are deferred to the Appendix.

7.1. Setting 1

In Figure 2, we compare the performance of our AIPW estimator against a popular multiple imputation method for missing data: multivariate imputation by chained equations or MICE (van Buuren and Groothuis-Oudshoorn 2010). Though MICE uses a series of flexible models to impute missing values based on covariate information, the consistency of this imputation procedure depends on the assumption that missing data mechanism is MAR. Here we have generated data that are MNAR, satisfying assumption (1), and therefore, as expected, MICE

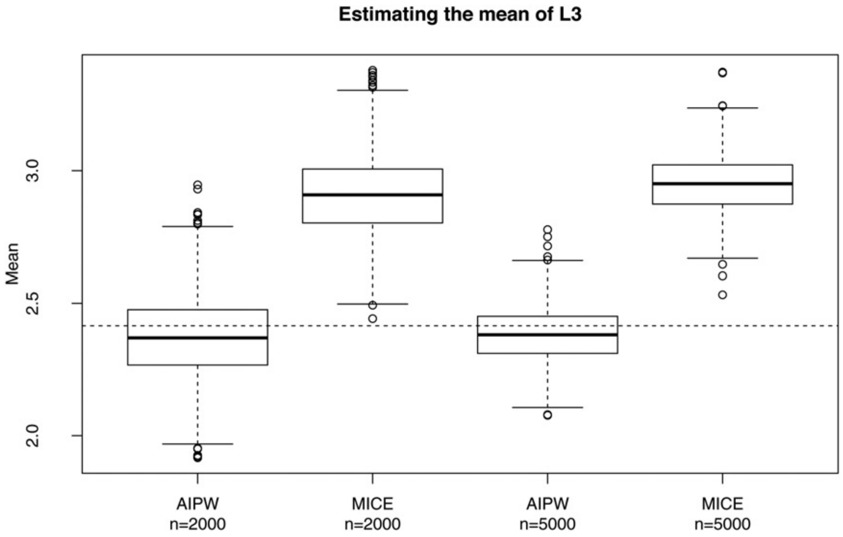

Figure 2. Estimates of $\mathbb{E}[L_3]$ in the first simulation setting. Boxplots are calculated from 1000 trials at sample sizes $n = 2000$ and $n = 5000$. The horizontal dashed line indicates the true value, which is 2.415593. MICE estimates are clearly biased upward. AIPW estimates, as expected, concentrate around the true value.

Table 1. Simulation results illustrating the double-robustness property.

Setting	Bias	Percent bias	MSE	Var
Outcome Reg. Misspec.	−0.16	0.98	0.14	0.12
Missingness Prob. Misspec.	−0.16	0.96	0.039	0.013
Both Misspec.	0.96	5.75	7.68	6.77
Both Correct	−0.068	0.41	0.020	0.016

NOTE: The true value of $\mathbb{E}[L_3]$ is 16.71512 and sample size is $n = 5000$. Bias, percent bias, mean squared error, and variance are calculated over 1000 trials.

performs poorly despite the quite simple parametric model for the full data. With all nuisance models correctly specified, the proposed AIPW estimator is seen to be unbiased.

7.2. Setting 2

In the second setting, with always observed vector $X = (X_1, X_2)$, we explored the double-robustness property. From the conditional Gaussian form of the data-generating process, we know that the outcome regressions are correctly specified as log-linear functions of $L_{(r)}$ and X. Likewise, the logit probability of each missingness indicator R_i in Equation (8) is a linear function of L_{-i} and X. To illustrate our double-robustness result, these nuisance models were misspecified by replacing (X_1, X_2) with $(\log(\frac{1}{X_1} + \frac{1}{X_2}), \sqrt{X_1 X_2})$ in all outcome regressions and missingness probabilities. We used the doubly robust estimator $\widehat{\text{OR}}_{dr}$ for the odds ratio. Results for 1000 trials with $n = 5000$ samples are shown in Table 1. The AIPW estimator is seen to be unbiased when either the outcome regressions or missingness probabilities are misspecified, but not both.

8. An Application to HIV Data

We applied the proposed AIPW estimator to data from an observational study of HIV-positive mothers in Botswana. Specifically, we are interested in the relationship between continuing highly active antiretroviral therapy (HAART) during pregnancy and adverse birth outcomes, such as preterm delivery. The full dataset, abstracted from 6 sites in Botswana, is described in detail in Chen et al. (2012). Following Tchetgen Tchetgen, Wang, and Sun (2018), Sun and Tchetgen Tchetgen (2018), and Shpitser (2016), we focus on HIV-positive women ($n = 9711$) and 3 variables: HAART exposure during pregnancy, preterm delivery, and an indicator of low CD4$^+$ count (<200 μL). The question of interest is whether HAART continuation (68.9% missing) is associated with premature delivery (6.7% missing), satisfied by CD4$^+$ count (53.4% missing). The analysis is complicated by a small number of complete cases (10.5%) and nonmonotone patterns of missingness.

By way of estimating the association of interest and as further illustration of the proposed approach, we first obtain an estimate of the joint distribution of the variables of interest. In Table 2, we compare the joint distribution as estimated by our proposed AIPW estimator with complete-case analysis as well as MICE. Substantial differences between our AIPW procedure and complete case analysis are evident, for example, in the probability of observing HAART continuation, no low CD4$^+$ count, and no preterm delivery (third column). Differences with MICE are less pronounced. From the estimated joint distribution, we can obtain an estimate of the odds ratios between HAART and preterm delivery at both levels of CD4$^+$ count: these are 2.72 (95% CI: 1.26, 5.53) and 1.24 (95% CI: 0.91, 1.73) for low CD4$^+$ count and moderate or high CD4$^+$ count, respectively. Ninety-five percent confidence intervals in parentheses are computed by bootstrap (percentile) over 1000 subsamples. Using the same procedure, IPW (i.e., our estimator with pattern mixture regression models set to zero) produces estimates 2.51 (95% CI: 1.16, 4.95) and 1.15 (95% CI: 0.83, 1.63), respectively, MICE produces estimates 1.60 (95% CI: 1.09, 2.90) and 1.52 (95% CI: 0.98, 1.70), while compete-case analysis produces 2.16 (95% CI: 0.98, 4.41) and 1.00 (95% CI: 0.71, 1.42), respectively. It is interesting to compare our AIPW estimates to the results of Tchetgen Tchetgen, Wang, and Sun (2018), who analyzed the same data with the aforementioned discrete choice model estimator. Their

Table 2. Estimated joint distribution of HAART continuation during pregnancy (H), low CD4$^+$ count (C), and preterm delivery (P) in HIV-infected women in Botswana: comparing complete case analysis and MICE with estimation by AIPW.

H,C,P	1,1,1	1,1,0	1,0,0	1,0,1	0,1,1	0,1,0	0,0,0	0,0,1
AIPW	0.0122	0.0268	0.5206	0.1722	0.0091	0.0542	0.1618	0.0430
MICE	0.0164	0.0355	0.5156	0.1705	0.0140	0.0483	0.1641	0.0357
Complete cases	0.0137	0.0342	0.3320	0.0979	0.0333	0.1802	0.2380	0.0705

analysis is based on a different MNAR assumption than the one considered here, as discussed in Section 5. They reported an estimated odds ratio association of 1.158 (95% CI: 0.869, 1.560) under a main effect-only logistic regression of preterm delivery on HAART continuation and CD4$^+$ count. Similarly, Shpitser (2016) also reported an estimated odds ratio association of 1.032 (95% CI: 0.670, 1.394) using his pseudolikelihood-based IPW estimator under the no self-censoring assumption. Neither analysis by Tchetgen Tchetgen et al. and Shpitser was able to detect a significant association between preterm delivery and HAART continuation conditional on CD4$^+$ count; in contrast, the proposed AIPW estimator detected a significant association for mothers with low CD4$^+$ count. This underscores how the assumed missingness model may have quite important implications for the substantive scientific conclusions or policy recommendations supported by the data.

9. Discussion

We have introduced a practical and straightforward-to-implement AIPW estimator for functions of data missing-not-at-random, under the "no self-censoring" or "itemwise conditionally independent nonresponse" assumption, which places no restrictions on the observed data. Our estimator improves on the efficiency and flexibility of previously proposed estimators (Shpitser 2016; Sadinle and Reiter 2017) and when a subset of covariates are always observed, enjoys a certain double-robustness property (provided the odds ratio function encoding the association between R and L is correctly specified). We demonstrated in simulations that the estimator is an attractive alternative to popular multiple imputation procedures when the missing data mechanism is MNAR. Our analysis of HIV data from Botswana demonstrates that the proposed estimator can potentially make an important practical difference in applied problems with acute missingness.

Supplementary Materials

Supplementary materials available online include additional simulation details, additional simulation results, and proofs of theoretical results not proved in the body of the paper.

Funding

The authors gratefully acknowledge funding from the National Institutes of Health, grant R01 AI127271-01 A1.

References

Bhattacharya, R., Nabi, R., Shpitser, I., and Robins, J. M. (2019), "Identification in Missing Data Models Represented by Directed Acyclic Graphs," in *Proceedings of the 35th Conference on Uncertainty in Artificial Intelligence*. [1416]

Bickel, P. J., Klaassen, C. A. J., Ritov, Y., and Wellner, J. A. (1993), *Efficient and Adaptive Estimation for Semiparametric Models*, Baltimore, MD: Johns Hopkins University Press. [1416,1418]

Chen, H. Y. (2007), "A Semiparametric Odds Ratio Model for Measuring Association," *Biometrics*, 63, 413–421. [1416,1417,1418]

——— (2010), "Compatibility of Conditionally Specified Models," *Statistics & Probability Letters*, 80, 670–677. [1416,1417]

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. [1421]

Højsgaard, S., Edwards, D., and Lauritzen, S. (2012), *Graphical Models With R*, Boston, MA: Springer. [1420]

Lauritzen, S. L. (1996), *Graphical Models*, Oxford: Clarendon Press. [1416]

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. [1415]

Little, R. J. A. (1993), "Pattern-Mixture Models for Multivariate Incomplete Data," *Journal of the American Statistical Association*, 88, 125–134. [1419]

Little, R. J. A., and Rubin, D. B. (2014), *Statistical Analysis With Missing Data* (2nd ed.), New York: Wiley. [1415]

Marra, G., Radice, R., Bärnighausen, T., Wood, S. N., and McGovern, M. E. (2017), "A Simultaneous Equation Approach to Estimating HIV Prevalence With Nonignorable Missing Responses," *Journal of the American Statistical Association*, 112, 484–496. [1415]

Mohan, K., and Pearl, J. (2018), "Graphical Models for Processing Missing Data," arXiv no. 1801.03583. [1416,1417]

Mohan, K., Pearl, J., and Tian, J. (2013), "Graphical Models for Inference With Missing Data," in *Advances in Neural Information Processing Systems*, pp. 1277–1285. [1416,1417]

Mohan, K., Thoemmes, F., and Pearl, J. (2018), "Estimation With Incomplete Data: The Linear Case," in *Proceedings of the 27th International Joint Conference on Artificial Intelligence*, pp. 5082–5088. [1416,1417]

Newey, W. K. (1990), "Semiparametric Efficiency Bounds," *Journal of Applied Econometrics*, 5, 99–135. [1416]

Robins, J. M. (1997), "Non-Response Models for the Analysis of Non-Monotone Non-Ignorable Missing Data," *Statistics in Medicine*, 16, 21–37. [1415,1416]

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. [1415]

Robins, J. M., Rotnitzky, A., and Scharfstein, D. O. (2000), "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, pp. 1–94. [1415,1418]

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. [1415]

Rotnitzky, A., and Robins, J. (1997), "Analysis of Semi-Parametric Regression Models With Non-Ignorable Non-Response," *Statistics in Medicine*, 16, 81–102. [1415]

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. [1415]

Sadinle, M., and Reiter, J. P. (2017), "Itemwise Conditionally Independent Nonresponse Modelling for Incomplete Multivariate Data," *Biometrika*, 104, 207–220. [1415,1416,1417,1422]

——— (2018), "Sequential Identification of Nonignorable Missing Data Mechanisms," *Statistica Sinica*, 28, 1741–1759. [1416]

Scharfstein, D. O., Rotnitzky, A., and Robins, J. M. (1999), "Adjusting for Nonignorable Drop-Out Using Semiparametric Nonresponse Models," *Journal of the American Statistical Association*, 94, 1096–1120. [1415]

Shpitser, I. (2016), "Consistent Estimation of Functions of Data Missing Non-Monotonically and Not at Random," in *Advances in Neural Information Processing Systems*, pp. 3144–3152. [1415,1416,1421,1422]

Shpitser, I., Mohan, K., and Pearl, J. (2015), "Missing Data as a Causal and Probabilistic Problem," in *Proceedings of the 31st Conference on Uncertainty in Artificial Intelligence*. [1416]

Sun, B. L., Liu, L., Miao, W., Wirth, K., Robins, J., and Tchetgen Tchetgen, E. J. (2018), "Semiparametric Estimation With Data Missing Not at Random Using an Instrumental Variable," *Statistica Sinica*, 28, 1965–1983. [1415]

Sun, B. L., and Tchetgen Tchetgen, E. J. (2018), "On Inverse Probability Weighting for Nonmonotone Missing at Random Data," *Journal of the American Statistical Association*, 113, 369–379. [1421]

Tan, Z. (2019), "On Doubly Robust Estimation for Logistic Partially Linear Models," arXiv no. 1901.09138. [1420]

Tchetgen Tchetgen, E. J., Robins, J. M., and Rotnitzky, A. (2010), "On Doubly Robust Estimation in a Semiparametric Odds Ratio Model," *Biometrika*, 97, 171–180. [1420]

Tchetgen Tchetgen, E. J., Wang, L., and Sun, B. L. (2018), "Discrete Choice Models for Nonmonotone Nonignorable Missing Data: Identification and Inference," *Statistica Sinica*, 28, 2069–2088. [1415,1419,1421]

Tian, J. (2015), "Missing at Random in Graphical Models," in *Proceedings of the 18th International Conference on Artificial Intelligence and Statistics*. [1416]

Tsiatis, A. (2006), *Semiparametric Theory and Missing Data*, New York: Springer. [1415,1416]

Tu, R., Zhang, C., Ackermann, P., Mohan, K., Hedvig Kjellström, and Zhang, K. (2019), "Causal Discovery in the Presence of Missing Data," in *The 22nd International Conference on Artificial Intelligence and Statistics*, pp. 1762–1770. [1416]

van Buuren, S., and Groothuis-Oudshoorn, K. (2010), "mice: Multivariate Imputation by Chained Equations in R," *Journal of Statistical Software*, 45, 1–68. [1420]

Van der Vaart, A. W. (2000), *Asymptotic Statistics*, Cambridge: Cambridge University Press. [1416]

Vansteelandt, S., Rotnitzky, A., and Robins, J. (2007), "Estimation of Regression Models for the Mean of Repeated Outcomes Under Nonignorable Nonmonotone Nonresponse," *Biometrika*, 94, 841–860. [1415]

Zhou, Y., Little, R. J. A., and Kalbfleisch, J. D. (2010), "Block-Conditional Missing at Random Models for Missing Data," *Statistical Science*, 25, 517–532. [1415]