Identifiability and Estimation under Missing Not at Random Mechanisms

Abstract

Conducting valid statistical analyses is challenging in the presence of missing-not-at-random (MNAR) data, where the missingness mechanism is dependent on the missing values themselves even conditioned on the observed data. Here, we consider a MNAR model that generalizes several prior popular MNAR models in two ways: first, it is less restrictive in terms of statistical independence assumptions imposed on the underlying joint data distribution, and second, it allows for all variables in the observed sample to have missing values. This MNAR model corresponds to a so-called *criss-cross* structure considered in the literature on graphical models of missing data that prevents nonparametric identification of the entire missing data model. Nonetheless, part of the complete-data distribution remains nonparametrically identifiable. By exploiting this fact and considering a rich class of exponential family distributions, we establish sufficient conditions for identification of the complete-data distribution as well as the entire missingness mechanism. We then propose methods for testing the independence restrictions encoded in such models using odds ratio as our parameter of interest. We adopt two semiparametric approaches for estimating the odds ratio parameter and establish the corresponding asymptotic theories: one involves maximizing a conditional likelihood with order statistics and the other uses estimating equations. The utility of our methods is illustrated via simulation studies.

1 INTRODUCTION

Conducting valid statistical analyses is challenging in the presence of missing data as the observed data may not be representative of the population of interest. According to the terminology of Rubin [1976], a missingness mechanism is called missing-at-random (MAR) if it only depends on the observed data values, and it is called missing-not-at-random (MNAR) if it is dependent on the missing values themselves even conditioned on the observed data. Under a MAR model, identification of a target parameter as a function of the observed data is a relatively straightforward task, and estimation strategies are well-studied, ranging from likelihood-based methods such as expectation maximization [Dempster et al., 1977, Little and Rubin, 2002], to multiple imputation [Rubin, 1987], inverse probability weighting [Robins et al., 1994, Li et al., 2013], and semiparametric methods closely related to the estimation of causal parameters [Robins et al., 1995, Tsiatis, 2006]. On the other hand, MNAR mechanisms are substantially more complicated and under-studied, yet they are construed as the most prevalent form of missingness mechanisms in practice.

In the presence of MNAR mechanisms, it is generally not possible to express the underlying *complete-data* distribution as a function of the *observed data* distribution without imposing additional assumptions. A lack of identification result implies that there exist at least two models that differ in their respective complete-data distribution but share the same observed data distribution. A well-known example of a non-identified MNAR mechanism is the non-ignorable non-response model in survey sampling, where the response variable directly causes its own missingness, often referred to as a *self-censoring* missingness mechanism. Other MNAR models include scenarios where missingness of a variable depends on other variables that themselves could be missing.

Common approaches for making progress in non-identified MNAR models include imposing, often untestable, (semi)parametric assumptions that yield identification [Wu and Carroll, 1988, Little and Rubin, 2002, Zhao and Shao, 2015]. For instance, in order to deal with the self-censoring mechanism involving a univariate response variable, several authors have considered the presence of a fully observed variable along with certain assumptions to identify

Submitted to the 39th Conference on Uncertainty in Artificial Intelligence (UAI 2023). **To be used for reviewing only**.

and estimate distributional quantities involving the response variable – e.g., Wang et al. [2014] considers a *shadow variable* that is not determinant of the underlying missingness, and Sun et al. [2018] considers an *instrumental variable* that is dependent with the missingness indicator of the response variable but independent of the response variable itself (marginally or conditioned on other fully observed variables). Other approaches include conducting sensitivity analysis [Rotnitzky et al., 1998, Scharfstein and Irizarry, 2003, Scharfstein et al., 2021] or obtaining nonparametric bounds for parameters of interest [Horowitz and Manski, 2000]. A recent line of work considers missing data models with a collection of independence restrictions among variables and corresponding missingness indicators that can be represented by directed acyclic graphs (DAGs); see Mohan and Pearl [2021], Nabi et al. [2022] for detailed reviews.

In this work, we consider a MNAR model that corresponds to a graphical characterization, the *criss-cross* structure discussed in Nabi and Bhattacharya [2022], where missingness of the response variable depends on the missingness of covariates and vice versa. This kind of missingness is common in cross-sectional and survey studies. Unlike most prior work, all variables in our model can be subject to missingness, i.e., our results do not rely on the presence of fully observed variables. Furthermore, the MNAR model under study generalizes several prior popular missing data models, including the permutation model [Robins, 1997], the block-conditional MAR model [Zhou et al., 2010], and the block-parallel model [Mohan et al., 2013], making it less restrictive in terms of statistical independence assumptions imposed on the underlying joint data distribution.

The criss-cross MNAR structure prevents nonparametric identification of the entire missing data model. We show, however, part of the complete-data distribution remains nonparametrically identifiable. We consider a quantitative measure, based on the rank of a *Jacobian matrix*, to examine the amount of information in the identifiable part that would be sufficient for recovering the entire complete-data law, a.k.a. the *target law*, as a function of only partially observed data. We explore these sufficient conditions extensively in the rich class of exponential family distributions. We further extend these results to higher dimensional parameter spaces and explore identifiability conditions for the entire missingness selection model, studied under *full law* identification. Aside from identification arguments, we explore procedures for testing independence relations among variables that are themselves missing in terms of an *odds ratio* parameterization of the complete-data law, as well as other model assumptions. We propose semiparametric estimating equations and conditional likelihoods based on order statistics to compute parameters that can be used for model selection purposes. Asymptotic properties of these two approaches are studied. We show empirically that the estimating equation approach is more efficient compared to the conditional likelihood approach while the latter is more robust to misspecifications of the missingness selection model.

The paper is organized as follows. We describe our notation and a brief overview of missing data DAGs in Section 2, and formally define the MNAR model under study in Section 3. We first consider univariate settings and discuss our (non)parametric identification and semiparametric estimation results in Sections 4 and 5, respectively, followed by generalizations to multidimensional covariate spaces in Section 6. The simulation results are provided in Section 7, followed by conclusions in Section 8. All proofs are deferred to supplementary materials.

2 PRELIMINARIES

Let Z be a vector of random variables with finite support and probability density $p(Z)$. Given a finite sample, variables in Z, indexed here by k, may have missing instances. Let R be the corresponding vector of binary missingness indicators where $R_k = 1$ if Z_k is observed and $R_k = 0$ if Z_k is missing. We only observe a coarsened version of Z in our sample, which we denote by Z^*. Each $Z^*_k \in Z$ is deterministically defined as follows: $Z^*_k = Z_k$ if $R_k = 1$ and $Z^* =$ "?" if $R_k = 0$. Z has a counterfactual connotation as it corresponds to variables "had they been fully observed" or "had R been set to one" (no missingness) – see Bhattacharya et al. [2019]. We use lowercase z to denote the observed realization of Z.

Following the literature on graphical models of missing data, it is descriptive to use directed acyclic graphs (DAGs) to encode assumptions in a given missing data model. A DAG $\mathcal{G}(V)$ is a set of vertices V connected by directed edges such that there are no directed cycles. The statistical model of a DAG $\mathcal{G}(V)$ is a set of distributions that factorize as $p(V) = \prod_{V_i \in V} p(V_i \mid \text{pa}_{\mathcal{G}}(V_i))$, where $\text{pa}_{\mathcal{G}}(V_i)$ denotes parents (direct causes) of V_i in $\mathcal{G}(V)$; when the vertex set is clear from the context, $\mathcal{G}(V)$ is abbreviated as $\mathcal{G}$. Using the conventions in Mohan et al. [2013], Bhattacharya et al. [2019], a missing data DAG (or mDAG for short) is defined over the set of vertices that correspond to variables in $V = \{Z, R, Z^*\}$. In addition to acyclicity, a mDAG restricts the presence of certain edges: each $Z^*_k \in Z^*$ has only two parents (Z_k and R_k), Z^*_k does not have any outgoing edges and variables in R cannot point to variables in Z. As an example, Fig. 1 illustrates the self-censoring mechanism in (a), the shadow variable setup in (b), and the instrumental variable approach in (c). Here, Y is the non-response variable, and X, W are fully observed variables. Deterministic edges are drawn in gray in all mDAGs.

A missing data model associated with a mDAG $\mathcal{G}$ is the set of distributions $p(Z, R, Z^*)$ that factorize as

$$\prod_{V_i \in Z} p(V_i \mid \text{pa}_{\mathcal{G}}(V_i)) \times \prod_{R_k \in R} p(R_k \mid \text{pa}_{\mathcal{G}}(R_k)). \quad (1)$$

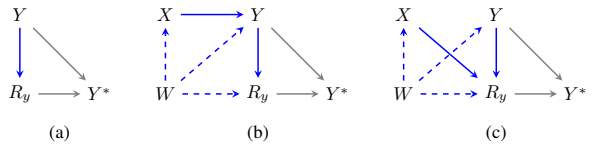

Figure 1: (a) Self-censoring MNAR mechanism; (b) Shadow variable setup considered in Wang et al. [2014]; (c) Instrumental variable setup considered in Sun et al. [2018]. A dashed edge implies potential dependence between the endpoint variables.

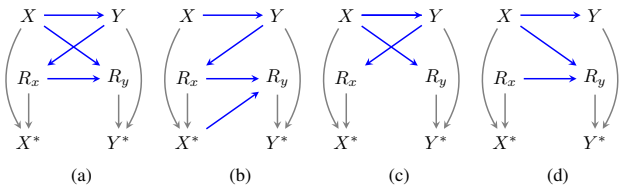

Figure 2: (a) Criss-cross MNAR model; (b) Permutation model [Robins, 1997]; (c) Block-parallel model [Mohan et al., 2013]; (d) Block-conditional MAR model [Zhou et al., 2010].

We exclude the factors $p(Z_k^* \mid Z_k, R_k)$ which are deterministically defined.Similar to a DAG, a mDAG encodes a set of ordinary conditional independence restrictions which can be easily read via Markov properties and d-separation rules: given disjoint subsets of vertices A, B, C, the DAG global Markov property states that if $A \perp_{\text{d-sep}} B \mid C$ in $\mathcal{G}(V)$, then $A \perp B \mid C$ in $p(V)$ [Pearl, 2009]. We refer to $p(Z)$ as the *target law*, $p(R \mid Z)$ as the *missingness mechanism*, and $p(R, Z^*)$ as the *observed data law*. The product of target law and missingness mechanism, i.e., $p(Z, R)$, is referred to as the *full law*. Note that in addition to partially missing variables, we may also have variables that are fully observed. However, in this work, we allow for the possibility of having all variables be partially missing in our model.

Aside from the mDAG factorization, an *odds ratio* parameterization of the full law (or parts of it) can be useful in handling missing data models as it is illustrated by our methods in later sections; for more use of such parameterization see Nabi et al. [2020], Malinsky et al. [2021]. Given disjoint sets of variables A, B, C and reference values $A = a_0, B = b_0$, the odds ratio parameterization of $p(A = a, B = b \mid C)$, given by Chen [2007], is as follows:

$$\frac{1}{Z(C)} \times p(a \mid b_0, C) \times p(b \mid a_0, C) \times \text{OR}(a, b \mid C), \quad (2)$$

where $\text{OR}(A = a, B = b \mid C)$ is defined as

$$\frac{p(A = a \mid B = b, C)}{p(A = a_0 \mid B = b, C)} \times \frac{p(A = a_0 \mid B = b_0, C)}{p(A = a \mid B = b_0, C)},$$

and $Z(C) = \sum_{A,B} p(A|B = b_0, C) \times p(B|A = a_0, C) \times \text{OR}(A, B \mid C)$ is the normalizing term.

3 THE MNAR MISSING DATA MODEL

We partition Z into two disjoint sets X and Y, where the missingness of X and Y depend on each other as follows:

$$(i)\ R_x \perp X \mid Y \qquad (ii)\ R_y \perp Y \mid X, R_x \qquad (3)$$

The above set of assumptions can be represented via the mDAG shown in Fig. 2(a), which corresponds to the so-called *criss-cross* structure discussed in Nabi and Bhattacharya [2022]. This missing data model is a supermodel of several popular models in the literature such as the *permutation model* [Robins, 1997] shown in Fig. 2(b), *block-parallel model* [Mohan et al., 2013] shown in Fig. 2(c), and *block-conditional MAR model* [Zhou et al., 2010] shown in Fig. 2(d). For instance, the permutation model implies the following set of independence restrictions: *(i)* $R_x \perp X \mid Y$ and *(ii)* $R_y \perp Y, X \mid X^*, R_x$. The independence restriction in *(ii)* implies $R_y \perp Y \mid X, R_x = 1$ and $R_y \perp Y, X \mid R_x = 0$. These assumptions are a superset of the assumptions made in the criss-cross model, as defined in (3). For more detailed comparisons across the aforementioned models, see Nabi et al. [2022].

The importance of the criss-cross graphical characterization is that in the presence of such structure, the target law is not nonparametrically identifiable as a function of the observed data distribution [Nabi and Bhattacharya, 2022], similar to the presence of self-censoring structure shown in Fig. 1(a). See Bhattacharya et al. [2019] for sufficient conditions under which the target law is nonparametrically identifiable and Nabi et al. [2020] for necessary and sufficient conditions under which the full law is nonparametrically identifiable, in a given mDAG.

4 IDENTIFICATION ARGUMENTS

4.1 NONPARAMETRIC IDENTIFICATION

Bhattacharya et al. [2019] proved that the conditional density of $p(R_y \mid R_x = 0, X)$ is not nonparametrically identifiable in the criss-cross model. This directly implies that the full law is not nonparametrically identified as a function of the observed data law. Nabi and Bhattacharya [2022] further proved that the target law is not identified either by providing a counterexample using binary variables for X and Y. We verify the lack of nonparametric identification of the target law in Appendix A, using continuous variables following normal distributions.

The conditional distribution $p(X \mid Y)$ is, however, nonparametrically identified. This is because using the independence

assumptions in display (3) and Bayes rule, we can write:

$$p(X \mid Y) = p(X \mid Y, R_x = 1) = \frac{p(X, Y, R_x = 1)}{\int p(x, Y, R_x = 1)dx},$$

where the marginal distribution $p(X, Y, R_x = 1)$ equals:

$$\frac{p(X, Y, R_x = 1, R_y = 1)}{p(R_y = 1 \mid R_x = 1, X, Y)} = \frac{p(X, Y, R_x = 1, R_y = 1)}{p(R_y = 1 \mid R_x = 1, X)},$$

and thus it is identified. The probabilistic operation of taking the full law and dividing it by the conditional density of $p(R_y \mid \text{pa}_{\mathcal{G}}(R_y))$ (evaluated at $R = 1$) corresponds to an intervention on R_y that sets it to one. This provides an intuitive inverse probability weighting estimation strategy for parameters involving the conditional density of X given Y. See Section 5.2 for a discussion on estimation and Nabi et al. [2022] for more details on the interventional view to identification in graphical models of missing data.

We take advantage of the nonparametric identification of $p(X \mid Y)$ in two ways: one is by combining this knowledge with consideration of a class of exponential family distributions to provide sufficient conditions for the identification of target and full laws (Section 4.2), and the other is by exploiting the knowledge in $p(X \mid Y)$ to estimate the odds ratio between X and Y as a method of an independence test, using either a conditional likelihood approach (Section 5.1) or a generalized estimating equation (GEE) approach (Section 5.2).

4.2 PARAMETRIC IDENTIFICATION

We first consider identification of the target law $p(X, Y)$ when X is assumed to be univariate. We generalize our identification results to multivariate X in Section 6.

4.2.1 Target law identification

Assume $p(X)$ and $p(Y \mid X)$ belong to the exponential family distribution. That is,

$$p(x) \sim \exp\left\{\frac{x\eta_x - b_x(\eta_x)}{\Phi_x} + c_x(x;\ \Phi_x)\right\} \tag{4}$$
$$p(y \mid x) \sim \exp\left\{\frac{y\eta - b(\eta)}{\Phi} + c(y;\ \Phi)\right\}, g(\mu(\eta)) = \alpha + \beta x,$$

where b, c, b_x, c_x are known functions, $\Phi, \Phi_x > 0$ are dispersion parameters that may be known or unknown, and g is a known one-to-one, third-order continuously differentiable link function. Let $\mu(\eta) := \mathbb{E}[Y|X]$ and $\mu_x(\eta_x) := \mathbb{E}[X]$. From the exponential family theory, we know that $b'(\eta) = \mu(\eta)$ and $b'_x(\eta_x) = \mu_x$. If $\mu = g^{-1}$, then g is called the canonical link function and is denoted by g_c. We outline sufficient conditions for identifying the parameter vector $\theta = (\alpha, \beta, \Phi, \eta_x, \Phi_x)$ in the following theorem.

Theorem 1. *Assume the model in display (4) and X takes $k+1$ distinct values $x_0, x_1, \cdots, x_k$. Let $\varphi = [g \circ \mu]^{-1}$, $\zeta = b([g \circ \mu]^{-1})$. Define the following equations:*

$$\phi_i(\theta) = \{\varphi(\alpha + x_i\beta) - \varphi(\alpha + x_0\beta)\}/\Phi$$
$$\zeta_i(\theta) = \frac{-\zeta(\alpha + x_1\beta) + \zeta(\alpha + x_0\beta)}{\Phi} + \frac{\eta_x(x_1 - x_0)}{\Phi_x} + c(x_1;\ \Phi_x) - c(x_0;\ \Phi_x).$$

Define the Jacobian matrix $J = \partial(\Phi,\ Z)/\partial\theta$, where $\Phi = \{\phi_1, \ldots, \phi_k\}$ and $Z = \{\zeta_1, \ldots, \zeta_k\}$. Under regularity conditions (detailed in Appendix B.1), the target law $p(X, Y)$ is identifiable if

$$(i)\ k \geq dim(\theta), \quad (ii)\ \textit{Jacobian matrix } J \textit{ has full rank.}$$

See Appendix B.1 for a proof. To provide an insight into Theorem 1, we emphasize the following observation: for any two distinct points of X, say x_1, and x_0, we have

$$\frac{p(x_1 \mid y)}{p(x_0 \mid y)} = \frac{p(y \mid x_1)}{p(y \mid x_0)} \times \frac{p(x_1)}{p(x_0)}. \tag{5}$$

The left-hand side of equation (5) is identified, therefore as we vary the choice of distinct points of X, we are getting a series of equations that connect the identified conditional distribution $p(X \mid Y)$ to the target law. The rank of the Jacobian matrix J provides a quantitative measure for the amount of information about the target law that is reflected in the conditional distribution $p(X \mid Y)$. When J is full rank, we are able to obtain a unique solution of the target law, as a function of observed data law, by solving a system of equations. In the case of J being rank deficient, we observe that removing some columns of J can lead J to be full rank. Removing columns from J has the interpretation of assuming the corresponding parameters to be known, which yields sufficient conditions for identification claims. A similar argument is made by Zhao and Shao [2015] in the non-ignorable non-response model (a.k.a. self-censoring) where X is assumed to be fully observed and the parametric marginal density of X is known.

We highlight that our identification framework is highly generalizable. As the dimensionality of the distribution increases, the core of the theorem remains unchanged. We delve into the generalization of Theorem 1 thoroughly in Section 6. In addition, the method proposed is not restricted to the exponential family distributions, while focusing on this family results in clean and concise identification characterizations. We will further demonstrate in Section 4.2.2 that the full law identification is easier to establish within the exponential family.

In Appendix C, we show the utilization of Theorem 1 in establishing sufficient conditions for target law identification in widely used exponential family distributions, including normal, Bernoulli, exponential, and Poisson distributions

with either canonical or inverse links. The second condition in Theorem 1, namely that the Jacobian matrix must be of full rank, has different implications on what specific knowledge is required for θ in advance. For instance, under normal distributions with an inverse link discussed in Appendix C.2 or exponential distributions discussed in Appendix C.7, the target law is identified without any further restrictions on the parameter vector θ. While in certain other distributions, the full-rank requirement of the Jacobian matrix implies that part of θ must be known apriori. For instance, in bivariate normal distributions with a canonical link discussed in Appendix C.1, it is essential for identification arguments that at least the marginal mean of either X or Y is known. We emphasize that Theorem 1 only provides sufficient, not necessary, identification conditions. This means that stronger-than-needed characterizations might be established.

4.2.2 Full law identification

Under the conditions of Theorem 1, we can use the joint factorization of the full law in the criss-cross model to show that the conditional density of R_x given Y, a.k.a. the propensity score of R_x, is identified: $p(X, Y, R_x = 1, R_y = r_y) = p(X, Y) \times p(R_x = 1 \mid Y) \times p(R_y = r_y \mid X, R_x = 1)$, for $r_y = 0, 1$. To fully identify the full law, we need to show whether the full law evaluated at $R_x = 0$, i.e., $p(X, Y, R_x = 0, R_y = r_y)$, is identified or not, or equivalently whether or not the propensity score of R_y evaluated at $R_x = 0$, i.e., $p(R_y = 1 \mid R_x = 0, X)$, is identified. The question of full law identification translates into the nonexistence of any two distinct propensity scores for R_y, e.g., $p_1(R_y \mid X, R_x) \neq p_2(R_y \mid X, R_x)$, such that $\int [\, p_1(R_y = 1 \mid R_x = 0, x) - p_2(R_y = 1 \mid R_x = 0, x) \,] \; p(x \mid Y) \; dx = 0$. Let $h(X) = p_1(R_y = 1 \mid R_x = 0, X) - p_2(R_y = 1 \mid R_x = 0, X)$. This condition then implies that if $\mathbb{E}[h(X) \mid Y] = 0$, then it must be the case that $h(X) = 0$ for the full law to be identified. This relates to the *completeness* condition described below.

Condition 1. *For any function $h(X)$ with finite mean, $\mathbb{E}\{h(X) \mid Y\} = 0$ implies $h(X) = 0$ almost surely.*

With the completeness condition introduced, we can establish identification of the full law as follows.

Lemma 1. *Given the conditions in Theorem 1 and Condition 1, the full law $p(X, Y, R_x, R_y)$ is identified.*

See Appendix B.3 for a proof. Identification under the completeness condition is widely seen among previous works [Newey and Powell, 2003, Miao et al., 2015, Zhao and Ma, 2022]. As a special case, full law identification can be established from the completeness property of the exponential family distributions. More specifically, Condition 1 is guaranteed to hold if $p(X \mid Y)$ takes the following form:

$$p(X \mid Y) = s(X)\, t(Y) \exp\left[\mu(Y)^T \tau(X)\right],$$

where $s(X) > 0$, $\tau(X)$ is one-to-one in X, and the support of $\mu(Y)$ is an open set.

We show that the specific examples discussed in Appendices C.1, C.3, C.4, C.5, and C.6 all have $p(X \mid Y)$ lie in the exponential family, therefore the full law is guaranteed to be identified (under conditions outlined in Theorem 1). In examples discussed in Appendices C.2 and C.7, $p(X \mid Y)$ falls out of the exponential family, therefore the full law may or may not be identified.

5 ESTIMATION AND INFERENCE

Our primary target of inference is the odds ratio between X and Y, denoted by $\mathrm{OR}(X, Y)$ and defined in (2). Since the conditional density $p(X \mid Y)$ is nonparametrically identified, this odds ratio is also nonparametrically identified. In order to estimate this parameter, we establish two semiparametric methods outlined below. Hereafter, we use n to denote the size of the completely observed samples and N the size of all samples.

5.1 CONDITIONAL LIKELIHOOD WITH ORDER STATISTICS

As our first approach to estimate $\mathrm{OR}(X, Y)$, we adopt the conditional likelihood approach based on order statistics, motivated by the fact that $p(X \mid Y, R_x = 1, R_y = 1)$ equals

$$\frac{p(R_x = 1, R_y = 1 \mid Y, X)}{\int p(R_x = 1, R_y = 1 \mid Y, x) p(x \mid Y) dx}\, p(X \mid Y),$$

where $p(R_x = 1, R_y = 1 \mid Y, X) = p(R_y = 1 \mid R_x = 1, X) p(R_x = 1 \mid Y)$ is a multiplier of a function of Y-only and a function of X-only, and $\int p(R_x = 1, R_y = 1 \mid Y, X) p(X \mid Y) dX$ is a function of Y-only. Consider the following conditional likelihood $p(x_1, \ldots, x_n \mid r_{x_1} = r_{y_1} = 1, \ldots, r_{x_n} = r_{y_n} = 1, y_1, \ldots, y_n, \tilde{X})$ which equals

$$\frac{\prod_{i=1}^n p(x_i \mid y_i)}{\sum_{\text{permutation of } x} \prod_{i=1}^n p\left(x_{(i)} \mid y_i\right)},$$

where $\tilde{X}$ denotes the order statistics $(x_{(1)}, \ldots, x_{(n)})$ and the permutation is over all possible permutations of $\{1, \ldots, n\}$. By exploiting the information available in this conditional likelihood, it is possible to estimate some parameters, such as the odds ratio, in the model of $p(X \mid Y)$. The nice feature of applying this conditional likelihood is that for each subject i, the corresponding terms $p(R_x = 1, R_y = 1 \mid Y, X)$ and $\int p(R_x = 1, R_y = 1 \mid Y, x) p(x \mid Y) dx$ are all canceled out during the above derivations; therefore, this conditional likelihood approach is robust to the model misspecification of the propensity scores, i.e., neither $p(R_y = 1 \mid R_x = 1, X)$ nor $p(R_x = 1 \mid Y)$ need to be correctly specified in order to have a consistent estimation of the odds ratio.

Since the above conditional likelihood has the computation complexity of order $n!$, in reality, we approximate the conditional likelihood with the following pairwise pseudo-likelihood

$$\prod_{i<k} \frac{p\left(x_i \mid y_i\right) p\left(x_k \mid y_k\right)}{p\left(x_i \mid y_i\right) p\left(x_k \mid y_k\right)+p\left(x_i \mid y_k\right) p\left(x_k \mid y_i\right)} = \prod_{i<k} \frac{1}{1+Q\left(x_i, y_i ; x_k, y_k\right)},$$

where $Q\left(x_i, y_i ; x_k, y_k\right)$ is the inverse of OR and equals

$$\left\{p\left(x_i \mid y_k\right) p\left(x_k \mid y_i\right)\right\} / \left\{p\left(x_i \mid y_i\right) p\left(x_k \mid y_k\right)\right\}.$$

Therefore, by analyzing the completely observed subjects from the biased sample $p(X \mid Y, R_x = 1, R_y = 1)$, we are able to estimate the odds ratio OR between X and Y. This conditional likelihood approach was first proposed in [Kalbfleisch, 1978] for hypothesis testing and then was used in a variety of statistical problems including both parameter estimation [Liang and Qin, 2000] and variable selection [Zhao et al., 2018]; see [Chen, 2021] for a more comprehensive exposition.

To illustrate the above pairwise pseudo-likelihood, we first consider a special case that $X \mid Y \sim \mathbb{N}(\alpha + \beta Y, \sigma^2)$, then

$$OR = \exp\left(\frac{\beta}{\sigma^2}(x_i - x_k)(y_i - y_k)\right) = \exp\left[\frac{\beta}{\sigma^2}\left(w_j v_j\right)\right],$$

where $w_j = -\operatorname{sign}\left(y_i - y_k\right)$ and $v_j = (x_i - x_k)\left|y_i - y_k\right|$, $j = 1, \ldots, n(n-1)/2$ corresponds to each pair of $(i, k), i, k = 1, \ldots, n$. Hence, the logarithm of the above pairwise pseudo-likelihood can be written as

$$-\sum_j \log\left\{1 + \exp\left[\frac{\beta}{\sigma^2}\left(w_j v_j\right)\right]\right\}.$$

Thus, one can obtain the estimate of the parameter $\frac{\beta}{\sigma^2}$, denoted as θ hereafter, by performing the logistic regression with response u_k and covariate v_k without the intercept term, where

$$u_k = \begin{cases} 1 & \text{if } y_i - y_k > 0 \\ 0 & \text{if } y_i - y_k < 0. \end{cases}$$

Denote $\tilde{\theta}$ the parameter estimate. Our result below demonstrates the asymptotic normality of $\tilde{\theta}$.

Theorem 2. *Denote $Q(x_i, y_i; x_k, y_k; \theta) = Q_{ik}(\theta)$ and $\zeta_{ik}(\theta) = \partial \log\{1 + Q_{ik}(\theta)\}/\partial\theta$. Assume that $\mathbb{E}\|\zeta_{12}(\theta)\|^2 < \infty$ for any θ in the parameter space. Then,*

$$\sqrt{N}(\tilde{\theta} - \theta_0) \xrightarrow{d} \mathbb{N}(0, A^{-1}BA^{-1}),$$

where $A = \mathbb{E}\left\{R_{x_1} R_{y_1} R_{x_2} R_{y_2} \partial\zeta_{12}(\theta_0)/\partial\theta\right\}$ and $B = 4\mathbb{E}\left\{R_{x_1} R_{y_1} R_{x_2} R_{y_2} R_{x_3} R_{y_3} \zeta_{12}(\theta_0)\zeta_{13}(\theta_0)\right\}$.

See Appendix D.1 for a proof. The aforementioned pairwise pseudo-likelihood is favorable under a large sample size given its computational efficiency. However, the pairwise pseudo-likelihood estimator is generally inefficient. To improve efficiency, groupwise pseudo-likelihood can be adopted. Instead of picking two observations at a time, groupwise pseudo-likelihood uses more than two observations as a group. For example, with a group size of three, we will have

$$L \propto \prod_{i<j<k} \frac{p(x_i \mid y_i)\, p(x_j \mid y_j)\, p(x_k \mid y_k)}{\sum_{P:\text{ permutation of } (i,j,k)} p(x_{P(i)} \mid y_i)\, p(x_{P(j)} \mid y_j)\, p(x_{P(k)} \mid y_k)}.$$

Increased group size gives better efficiency with the cost of computational time. The final choice of group size should base on the consideration of computational time and statistical efficiency. Computational techniques with adaptive Monte Carlo approximation and Metropolis algorithm for directly maximizing the conditional likelihood are also well established and can be found in Chapter 4 of Chen [2021].

5.2 GENERALIZED ESTIMATING EQUATIONS

In the estimation approach presented in Section 5.1, we need to specify the conditional density function $p(X \mid Y)$ either fully parametrically or semiparametrically. Alternatively, the model $p(X \mid Y)$ can be semiparametrically specified. For instance, assuming $\mathbb{E}(X \mid Y) = h(Y; \theta)$ with $h(\cdot)$ a known function and θ the unknown parameter of interest, we have the following estimating equation

$$\mathbb{E}\Big[\frac{R_x \times R_y}{\pi(X)} \times f(Y) \times (X - E(X \mid Y))\Big] = 0,$$

for any arbitrary function $f(Y)$. Hereafter, we denote $\pi(X) = p(R_y = 1 \mid R_x = 1, X)$ and $w(Y) = p(R_x = 1 \mid Y)$. Note that the model $\pi(X)$ does not involve any missing data, so any off-the-shelf statistical method can be applied to model $\pi(X)$. To better illustrate our proposed method, we do not particularly discuss the method for estimating $\pi(X)$ here.

Thus, the estimator of the parameter θ, denoted as $\hat{\theta}$, can be obtained by solving the following empirical version of the estimating equation

$$\frac{1}{N}\sum_{i=1}^{N} \frac{R_{x_i} \times R_{y_i}}{\pi(x_i)} \times f(y_i) \times (x_i - h(y_i; \theta)) = 0.$$

In the following, we develop the asymptotic normality of the estimator $\hat{\theta}$. In particular, we also identify the optimal choice of $f(y)$, $f_{opt}(y)$, such that it achieves the best possible estimation efficiency among all choices of arbitrary function $f(y)$. For simplicity, we denote $\Psi(X, Y, R_x, R_y; \theta) = \frac{R_x \times R_y}{\pi(X)} \times f(Y) \times (X - h(Y; \theta))$.

Theorem 3. *Assume that* $\mathbb{E}\|\Psi(X, Y, R_x, R_y; \theta)\|^2 < \infty$ *for any* θ *in the parameter space. Then,*

(a) For any function $f(Y)$*, we have*

$$\sqrt{N}(\hat{\theta} - \theta_0) \xrightarrow{d} \mathbb{N}(0, C^{-1}D(C^{-1})^T),$$

where $D = \mathbb{E}\left\{\frac{R_x R_y}{\pi(X)^2}(X - h(Y;\theta))^2 f(Y) f(Y)^T\right\}$, $C = \mathbb{E}\left\{\frac{R_x R_y}{\pi(X)} a(Y) f(Y)^T\right\}$, *and* $a(Y) = \frac{\partial h(Y;\theta)}{\partial \theta}\Big|_{\theta=\theta_0}$.

(b) The optimal choice of $f(Y)$ *is*

$$f_{opt}(Y) = \left[\mathbb{E}\left\{\frac{(X - h(Y;\theta))^2}{\pi(X)} \mid Y\right\}\right]^{-1} a(Y).$$

See Appendix D.2 for a proof.

5.3 ALTERNATIVE ESTIMATION TARGETS

In addition to the associational relation between X and Y, one might be interested in testing additional model assumptions, e.g., whether the missingness of X is indeed influenced by Y or not. This can be easily set up by rewriting the propensity score of R_x using a parameterization that encodes the odds ratio between R_x and Y as $p(R_x = 1 \mid y) = \{1 + \exp(\lambda + \eta(y))\}^{-1}$ where $\eta(y) := \log(\text{OR}(R_x = 0, y))$ and $\lambda = \log[p(R_x = 0 \mid y_0)/p(R_x = 1 \mid y_0)]$. Under the conditions of Theorem 1, $\eta(y)$ would be identified. Exploring detailed estimation strategies are left to future work.

It is worth pointing out that under the conditions of Theorem 1 and Condition 1, one can simply estimate the entire parameter vector of the full law, assuming the parametric forms of the propensity scores in the missingness mechanism are known. More flexible estimation approaches are possible if one is willing to make additional modeling assumptions. For instance, in addition to independence restrictions in display (3), we may assume $p(R_y = 1 \mid R_x, X)$ is not a function of X when $R_x = 0$. This reduces down the crisscross model to the permutation MNAR model proposed by Robins [1997], where the full law is nonparametrically identified and the model is nonparametrically saturated, i.e., it imposes no restriction on the observed data law. In this case, we can proceed with nonparametric influence function based estimation, as discussed in Appendix E.

6 MULTIDIMENSIONAL X

We now discuss how our identification arguments can be easily generalized to higher dimensional vector spaces. For a reasonable representation of sampling distributions, we extend Theorem 1 to instances where X follows either a multivariate normal or a multinomial distribution. The corresponding identification theories under these two scenarios are provided in Appendix B.2; generalization to other sampling distributions can be carried out in a similar fashion.

As two special cases, we consider X to follow a multivariate normal or a multinomial distribution while $Y \mid X$ follows a normal distribution under the canonical link. We assume that the first condition in Theorem 1 is satisfied by having sufficient observations.

Example 1. *(X is multivariate normal and $Y \mid X$ is normal under canonical link) Suppose*

$$X \sim \mathbb{N}_d(\mu, \Sigma), \qquad Y \mid X \sim \mathbb{N}(\alpha + X^T\beta, \Phi).$$

Assume the nuisance parameter Σ *is known. The unknown vector of parameters is* $\theta = (\alpha, \beta, \Phi, \mu)$*. A sufficient condition for identification of the target law* $p(X, Y)$ *is for the intercept* α *to be known. According to Lemma 1, the full law is also identified.*

Example 2. *(X is multinomial and $Y \mid X$ is normal under canonical link) Suppose*

$$X \sim \text{Multinomial}_d(n, p), \quad Y \mid X \sim \mathbb{N}(\alpha + X^T\beta, \Phi),$$

where $p = (p_1, \ldots, p_d)$ *is the vector of event probabilities, and* n *is the number of trials. We can write* $p(x) = \exp[x^T\eta + c(x)]$ *where* $\eta = (\log p_1, \ldots, \log p_d)$, $c(x) = \log \frac{n!}{x_1! \ldots x_d!}$*. Assume* n *is known. The unknown vector of parameters is* $\theta = (\alpha, \beta, \Phi, \eta)$*. A sufficient condition for identification of the target law* $p(X, Y)$ *is for the intercept* α *to be known, or knowing at least one element of* η*. According to Lemma 1, the full law is also identified.*

7 SIMULATIONS

We now examine the finite sample behavior of our proposed estimation strategies, namely (i) non-optimal GEE, (ii) optimal GEE, and (iii) conditional likelihood with order statistics. We conduct simulation studies of (X, Y) following bivariate normal distribution

$$\begin{pmatrix} Y \\ X \end{pmatrix} \sim \mathbb{N}\left[\begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}\right],$$

with $\mu_1 = 2$, $\mu_2 = 0.4$, $\sigma_1 = 1$, $\sigma_2 = 3$, $\rho = 0.3$. The missingness mechanism is set as follows:

$$p(R_x = 1 \mid Y) = \text{expit}(-0.5 + Y),$$
$$p(R_y = 1 \mid X, R_x) = \text{expit}(2 - R_x + 0.7X).$$

Under this setup, approximately 5% of observations have both X and Y missing, 16% of observations have X missing and Y observed, 25% of observations have X observed and Y missing and 54% of observations have both X and Y observed. Under the above setup, we have

$$X \mid Y \sim \mathbb{N}(\alpha + \beta Y, \sigma^2) = N(-1.4 + 0.9Y, 8.19)$$
$$\text{OR} = \exp\left\{\frac{\beta}{\sigma^2}(x_i - x_k)(y_i - y_k)\right\}.$$

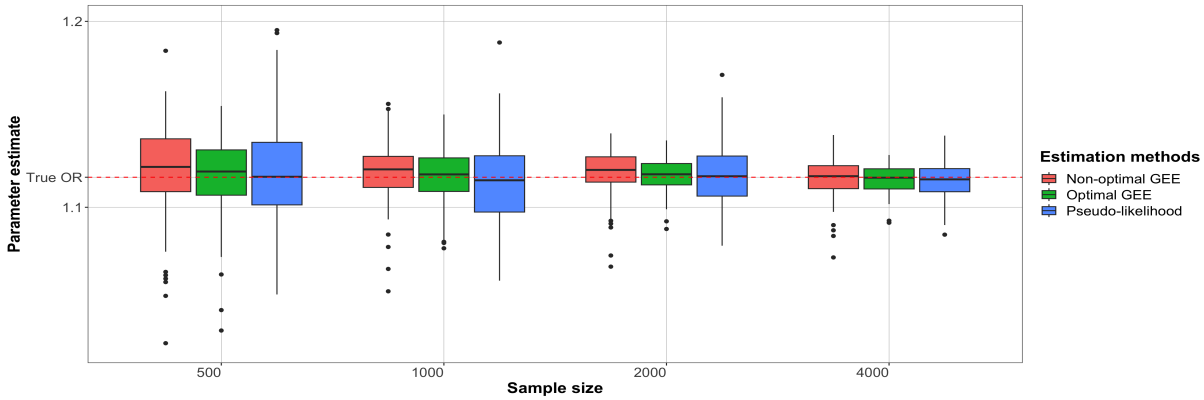

Figure 3: OR estimation with varying sample size.

Table 1: Parameter estimates with varying sample size.

		Non-optimal GEE		Optimal GEE	
N	Statistics	α	β	α	β
500	bias	0.1411	-0.0335	0.0613	-0.0028
	MSE	0.0199	0.0011	0.0038	0.0000
	SD	0.7980	0.3346	0.7830	0.3037
1000	bias	0.1079	-0.0281	0.1010	-0.0248
	MSE	0.0116	0.0008	0.0102	0.0006
	SD	0.6586	0.2601	0.6142	0.2467
2000	bias	-0.0820	0.0348	-0.0332	0.0140
	MSE	0.0067	0.0012	0.0011	0.0002
	SD	0.7864	0.3081	0.7043	0.2722
4000	bias	-0.0213	0.0088	-0.0242	0.0097
	MSE	0.0005	0.0001	0.0006	0.0001
	SD	0.5989	0.2249	0.4927	0.1795

Assuming the nuisance parameters σ_1, σ_2 are known, we aim at estimating α and β with non-optimal and optimal GEE approaches. We further estimate the odds ratio when $(x_i - x_k)(y_i - y_k) = 1$ using all three aforementioned methods. For non-optimal GEE, we choose $f(Y) = (1, Y)$. Note that for the optimal GEE, $f_{opt}(Y)$ might be a function of α, β. In such scenarios, to construct $\hat{f}_{opt}(Y)$, we utilize the estimated values $\hat{\alpha}$ and $\hat{\beta}$, obtained as medians over 100 simulation runs from the non-optimal GEE. All code necessary to reproduce our simulations is included with this submission.

We evaluate the performance of our three proposed estimators based on three main criteria: (i) finite sample behavior as sample size increases, (ii) bias behavior as a result of model misspecification for $p(R_y = 1 \mid X, R_x = 1)$, and (iii) efficiency behavior as a result of varying the correlation between X and Y. For each case, we conduct 100 simulation runs. The empirical comparisons for the second and third criteria are deferred to Appendix F due to page limits.

Figure 3 illustrates how the odds ratio estimation varies across a range of sample sizes from 500 to 4000. In order to ensure a fair comparison across the three methods, we assume that the intercept α of $\mathbb{E}(X \mid Y)$ is known for both non-optimal and optimal GEEs. The results demonstrate that all three methods yield unbiased estimates with reduced estimation uncertainty as the sample size increases. The conditional likelihood estimators are less efficient followed by non-optimal GEE, especially when the sample size is small. Overall, all three methods provide comparable OR estimates with small bias, mean-squared error (MSE), and standard deviation (SD) when the sample size is large.

Apart from OR estimation, the GEE approach is also capable of estimating the intercept α. Table 1 compares the performance of the two GEEs for estimating α and β, in terms of bias, MSE, and SD. As expected, the results show that the optimal GEE method outperforms the non-optimal GEE method in terms of smaller SD, regardless of the sample size. Additionally, for small sample sizes, the optimal GEE exhibits smaller bias and MSE than the non-optimal GEE. For additional simulations, see Appendix F.

8 CONCLUSIONS

In this paper, we considered a MNAR model which, like the self-censoring missingness mechanism, is an impediment to nonparametric identification of the complete-data distribution. We provided sufficient identification assumptions for both target and full laws by examining the rich class of exponential family distributions. We provided different semiparametric estimation strategies for computing parameters of the underlying joint distribution that can be used for pairwise independence tests and model selection purposes. An interesting avenue for future work is the exploration of a doubly-robust estimation theory that would enable the use of more flexible machine learning and statistical models in computing various model parameters.

References

Rohit Bhattacharya, Razieh Nabi, Ilya Shpitser, and James Robins. Identification in missing data models represented by directed acyclic graphs. In *Proceedings of the Thirty Fifth Conference on Uncertainty in Artificial Intelligence (UAI-35th)*. AUAI Press, 2019.

Hua Yun Chen. A semiparametric odds ratio model for measuring association. *Biometrics*, 63:413–421, 2007.

Hua Yun Chen. *Semiparametric Odds Ratio Model and Its Applications*. Chapman and Hall/CRC, 2021.

A.P. Dempster, N.M. Laird, and D.B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B*, 39:1–38, 1977.

Joel L Horowitz and Charles F Manski. Nonparametric analysis of randomized experiments with missing covariate and outcome data. *Journal of the American statistical Association*, 95(449):77–84, 2000.

John D Kalbfleisch. Likelihood methods and nonparametric tests. *Journal of the American Statistical Association*, 73 (361):167–170, 1978.

Lingling Li, Changyu Shen, Xiaochun Li, and James M Robins. On weighting approaches for missing data. *Statistical methods in medical research*, 22(1):14–30, 2013.

Kung-Yee Liang and Jing Qin. Regression analysis under non-standard situations: a pairwise pseudolikelihood approach. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 62(4):773–786, 2000.

Roderick JA Little and Donald B Rubin. *Statistical Analysis with Missing Data*. Wiley Series in Probability and Statistics. Wiley, 2002. ISBN 9780471183860.

Daniel Malinsky, Ilya Shpitser, and Eric J Tchetgen Tchetgen. Semiparametric inference for nonmonotone missing-not-at-random data: the no self-censoring model. *Journal of the American Statistical Association*, pages 1–9, 2021.

Wang Miao, Lan Liu, Eric Tchetgen Tchetgen, and Zhi Geng. Identification, doubly robust estimation, and semiparametric efficiency theory of nonignorable missing data with a shadow variable. *arXiv preprint arXiv:1509.02556*, 2015.

Karthika Mohan and Judea Pearl. Graphical models for processing missing data. *Journal of the American Statistical Association*, pages 1–16, 2021.

Karthika Mohan, Judea Pearl, and Jin Tian. Graphical models for inference with missing data. In *Advances in Neural Information Processing Systems 26*, pages 1277–1285. Curran Associates, Inc., 2013.

Razieh Nabi and Rohit Bhattacharya. On testability and goodness of fit tests in missing data models. *arXiv preprint arXiv:2203.00132*, 2022.

Razieh Nabi, Rohit Bhattacharya, and Ilya Shpitser. Full law identification in graphical models of missing data: Completeness results. In *Proceedings of the Twenty Seventh International Conference on Machine Learning (ICML-20)*, 2020.

Razieh Nabi, Rohit Bhattacharya, Ilya Shpitser, and James Robins. Causal and counterfactual views of missing data models. *arXiv preprint arXiv:2210.05558*, 2022.

Whitney K Newey and James L Powell. Instrumental variable estimation of nonparametric models. *Econometrica*, 71(5):1565–1578, 2003.

Judea Pearl. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, 2 edition, 2009. ISBN 978-0521895606.

James M. Robins. Non-response models for the analysis of non-monotone non-ignorable missing data. *Statistics in Medicine*, 16:21–37, 1997.

James M Robins, Andrea Rotnitzky, and Lue Ping Zhao. Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association*, 89(427):846–866, 1994.

James M Robins, Andrea Rotnitzky, and Lue Ping Zhao. Analysis of semiparametric regression models for repeated outcomes in the presence of missing data. *Journal of the American Statistical Association*, 90(429):106–121, 1995.

Andrea Rotnitzky, James M Robins, and Daniel O Scharfstein. Semiparametric regression for repeated outcomes with nonignorable nonresponse. *Journal of the American Statistical Association*, 93(444):1321–1339, 1998.

Donald B Rubin. Inference and missing data. *Biometrika*, 63(3):581–592, 1976.

Donald B. Rubin. *Multiple Imputation for Nonresponse in Surveys*. New York: Wiley & Sons, 1987.

Daniel O Scharfstein and Rafael A Irizarry. Generalized additive selection models for the analysis of studies with potentially nonignorable missing outcome data. *Biometrics*, 59(3):601–613, 2003.

Daniel O Scharfstein, Razieh Nabi, Edward H Kennedy, Ming-Yueh Huang, Matteo Bonvini, and Marcela Smid. Semiparametric sensitivity analysis: Unmeasured confounding in observational studies. *arXiv preprint arXiv:2104.08300*, 2021.

BaoLuo Sun, Lan Liu, Wang Miao, Kathleen Wirth, James Robins, and Eric J Tchetgen Tchetgen. Semiparametric estimation with data missing not at random using an instrumental variable. *Statistica Sinica*, 28(4):1965, 2018.

Anastasios Tsiatis. *Semiparametric Theory and Missing Data*. Springer-Verlag New York, 1st edition edition, 2006.

Sheng Wang, Jun Shao, and Jae Kwang Kim. An instrumental variable approach for identification and estimation with nonignorable nonresponse. *Statistica Sinica*, pages 1097–1116, 2014.

Margaret C Wu and Raymond J Carroll. Estimation and comparison of changes in the presence of informative right censoring by modeling the censoring process. *Biometrics*, pages 175–188, 1988.

Jiwei Zhao and Yanyuan Ma. A versatile estimation procedure without estimating the nonignorable missingness mechanism. *Journal of the American Statistical Association*, 117(540):1916–1930, 2022.

Jiwei Zhao and Jun Shao. Semiparametric pseudolikelihoods in generalized linear models with nonignorable missing data. *Journal of the American Statistical Association*, 110(512):1577–1590, 2015.

Jiwei Zhao, Yang Yang, and Yang Ning. Penalized pairwise pseudo likelihood for variable selection with nonignorable missing data. *Statistica Sinica*, 28(4):2125–2148, 2018.

Yan Zhou, Roderick J. A. Little, and Kalbfleisch John D. Block-conditional missing at random models for missing data. *Statistical Science*, 25(4):517–532, 2010.