Statistica Sinica **34** (2024), 1649-1673
doi:https://doi.org/10.5705/ss.202021.0404

# A UNIFIED INFERENCE FRAMEWORK FOR MULTIPLE IMPUTATION USING MARTINGALES

Qian Guan and Shu Yang*

*North Carolina State University*

*Abstract:* Multiple imputation is widely used to handle missing data. Although Rubin's combining rule is simple, it is not clear whether the standard multiple imputation inference is consistent when coupled with the commonly used full-sample estimators. Here, we establish a unified martingale representation of multiple imputation for a wide class of asymptotically linear full-sample estimators. This representation invokes the wild bootstrap inference to provide a consistent variance estimation under a correct specification of the imputation models. As a motivating application, we use the proposed method to estimate the average causal effect (ACE) with partially observed confounders in a causal inference. Our framework applies to asymptotically linear ACE estimators, including the regression imputation, weighting, and matching estimators. Lastly, we extend the proposed method to include scenarios in which both the outcome and the confounders are subject to missingness, and when the data are missing not at random.

*Key words and phrases:* Causality, congeniality, influence function, martingale representation, weighted bootstrap.

## 1. Introduction

Missing data are ubiquitous in practice. A widely used approach to handle incomplete/missing data is multiple imputation (MI). The National Research Council recommends MI as one of its preferred approaches to addressing missing data (National Research Council (2010)). The idea of MI is to fill the missing values multiple times by sampling from the posterior predictive distribution of the missing values, given the observed values. Then, we can apply full-sample analyses straightforwardly to the imputed data sets. These multiple results are summarized by an easy-to-implement combining rule for inference (Rubin (1987)). MI can provide valid frequentist inferences in various applications (e.g., Clogg et al. (1991)). However, some authors have found that Rubin's variance estimator is not always consistent (e.g., Fay (1992), Kott (1995), Fay (1996), Binder and Sun (1996), Wang and Robins (1998), Robins and Wang (2000), Nielsen (2003) and Kim et al. (2006)). To ensure the validity of Rubin's variance estimation, imputations must be proper (Rubin (1987)). A sufficient condition for proper imputation is the congeniality condition of Meng (1994), imposed on both the

*Corresponding author.

imputation model and the subsequent full-sample analysis. However, even with a correctly specified imputation model, Yang and Kim (2016) show that MI is not necessarily congenial for the method of moments estimation, and so common statistical procedures may be incompatible with MI. Nevertheless, given the popularity of MI in practice, it is important to develop a valid inference procedure for using MI in statistical inference.

As a motivating application, we focus on causal inference with partially observed confounders. Causal inference is a central goal in many disciplines, such as medicine, econometrics, and the political and social sciences. When all confounders that influence both the treatment and the outcome are observed, the average causal effect (ACE) of the treatment is identifiable (Imbens and Rubin (2015)). Many ACE estimators have been proposed to adjust for confounders, including regression imputation (Hahn (1998); Heckman, Ichimura and Todd (1997)), (augmented) propensity score weighting (Horvitz and Thompson (1952); Rosenbaum and Rubin (1983); Robins, Rotnitzky and Zhao (1994); Bang and Robins (2005); Cao, Tsiatis and Davidian (2009)), and matching (Rosenbaum (1989); Stuart (2010); Abadie and Imbens (2016)). Others use MI for causal inference with partially observed confounders, for example, Qu and Lipkovich (2009), Crowe, Lipkovich and Wang (2010), Mitra and Reiter (2011), and Seaman and White (2014). Although many full-sample estimators are available for estimating the ACE, few works have examined the validity of Rubin's variance estimator when using these estimators for causal inference.

We establish a novel martingale representation of MI for a general class of asymptotically linear full-sample estimators under a correct specification of the imputation models. Our key insight is that the MI estimator is intrinsically created in a sequential manner. First, the posterior samples of the parameters are drawn from the posterior distribution, which is asymptotically equivalent to the sampling distribution of the maximum likelihood estimator (MLE) based on the Bernstein–von Mises theorem (van der Vaart (2000, Chap. 10)). Second, we draw the posterior predictive samples of the missing data, conditioned on the observed data. This conceptualization leads to an asymptotically linear expression of the MI estimator in terms of a sequence of random variables that have conditional mean zero, given the sigma algebra generated from the preceding variables (i.e., a martingale representation). The martingale representation invokes the wild/weighted bootstrap procedure (Wu (1986); Liu (1988)) to provide a valid variance estimation and inference, regardless of which full-sample estimator is adopted in the MI.

We show the asymptotic validity of our proposed bootstrap inference method for the MI estimator using the martingale central limit theory (Hall and Heyde (1980)) and the asymptotic property of the weighted sampling of martingale difference arrays (Pauly (2011)). Although the validity of the proposed method is based on the asymptotic results as the sample size goes to infinity, the simulation

results demonstrate that it performs well for finite samples. We also compare the proposed method with the improper MI approaches proposed by Wang and Robins (1998) and Robins and Wang (2000). An improper MI uses a Monte Carlo imputation to compute the MLE and, therefore, requires that the imputation size $m$ be large in order to reduce the Monte Carlo error. In contrast, our proposed method allows the imputation size $m$ to be fixed at a small value. This property is appealing for releasing multiply imputed data sets for public usage. Moreover, an improper MI deals only with regular estimators, but not with nonregular estimators, such as the matching estimators. The proposed method can be applied to a wide range of the ACE estimators adopted in MI, including the outcome regression, weighting, and matching estimators. Indeed, the results of our simulation studies indicate that Rubin's variance estimator overestimates the variance for the inverse probability weighting (IPW) and matching estimators, because these two estimators are not self-efficient (Meng (1994); Xie and Meng (2017)), whereas the proposed variance estimation procedure is consistent for all types of estimators.

Importantly, our framework can easily accommodate scenarios in which both the outcome and the confounders have missing values, and when the missing data are missing not at random (MNAR). In the former case, we need only add the imputation step for the missing outcomes. In the latter case, we need to modify the imputation model by further considering the missing data probability model in the data likelihood function. Our research is likely to bridge the advantages of MI and its wide applications in causal inference and missing data analyses.

The rest of the paper is organized as follows. Section 2 introduces general asymptotically linear estimators and common estimators in causal inference. Section 3 describes the general MI used to fill in missing values that facilitate full-sample estimators. Section 4 presents the martingale representation for the MI estimators and the wild bootstrap inference procedure, and establishes its validity. Section 5 extends the proposed method to scenarios with other causal estimands, those in which both the outcome and the confounders have missing values, and those in which the confounders are MNAR. In Section 6, we evaluate the finite-sample performance of the proposed method using simulation studies. In Section 7, we apply the proposed wild bootstrap inference method to data from a U.S. National Health and Nutrition Examination Survey. Section 8 concludes the paper.

## 2. Background

### 2.1. General setup

We introduce a general setup and illustrate it using common estimators of the ACE in causal inference. Suppose we observe $n$ independent and identically distributed (i.i.d.) samples $\mathbf{L} = \{L_i : i = 1, \ldots, n\}$ governed by the distribution

$\mathbb{P}(L)$. We are interested in an inference about the target parameter, a functional of the observed data distribution, $\tau = \tau(\mathbb{P})$, for example, the mean of the distribution $\mathbb{P}$. For simplicity of presentation, we assume $\tau$ to be a one-dimensional parameter. An extension to a multi-dimensional parameter is feasible at the cost of heavier notation. Let $\hat{\tau}_n$ denote a generic estimator of $\tau$. We focus on the class of asymptotically linear estimators. This class of estimators includes the common regular and asymptotically linear (RAL) estimators, which can be expressed by

$$\hat{\tau}_n - \tau = \frac{1}{n}\sum_{i=1}^{n} \psi(L_i) + o_{\mathbb{P}}(n^{-1/2}), \tag{2.1}$$

where $\{\psi(L_i) : i = 1, \ldots, n\}$ are i.i.d., with $\mathbb{E}\{\psi(L_i)\} = 0$ and $\mathbb{E}\{\psi(L_i)^2\} < \infty$. The random variable $\psi(L_i)$ is called the influence function of $\hat{\tau}_n$, and captures the first-order asymptotic behavior of $\hat{\tau}_n$ (Bickel et al. (1993)). For the regularity conditions, see, for example, Newey (1990). For a given estimator, upon identifying its influence function, we can characterize the asymptotic distribution and construct corresponding confidence intervals (CIs) for the target parameter. The class of estimators also includes possibly nonregular asymptotically linear estimators, which can be expressed by

$$\hat{\tau}_n - \tau = \frac{1}{n}\sum_{i=1}^{n} \psi_i(\mathbf{L}) + o_{\mathbb{P}}(n^{-1/2}), \tag{2.2}$$

where the individual component $\psi_i(\mathbf{L})$ may depend on the full sample, and therefore is not i.i.d., but satisfies $\mathbb{E}\{\psi_i(\mathbf{L})\} = 0$ and $\mathbb{E}\{\psi_i(\mathbf{L})^2\} < \infty$. The matching estimator is an example, as we illustrate later. For simplicity, we also call $\psi_i(\mathbf{L})$ the influence function of $\hat{\tau}_n$.

### 2.2. Motivating application: Estimating the ACE

We explain the general framework by applying it to estimate the ACE. Let $X$ be a vector of $p$-dimensional covariates, $A \in \{0, 1\}$ be a binary treatment, with zero and one being the labels for the control and the active treatments, respectively, and $Y$ be the outcome of interest. Suppose we observe $n$ i.i.d. samples $\mathbf{L} = \{L_i = (A_i, X_i, Y_i) : i = 1, \ldots, n\}$.

Following Neyman (1923) and Rubin (1974), we use the potential outcomes framework to formulate the causal parameter of interest. Under the stable unit treatment value assumption (Rubin (1980)), for each level of treatment $a$, there exists a potential outcome $Y(a)$, representing the outcome had the unit, possibly contrary to the fact, been given treatment $a$. We make the causal consistency assumption that links the observed outcome to the potential outcomes; that is, the observed outcome $Y$ is the potential outcome $Y(A)$ under the actual treatment. We focus on estimating the ACE $\tau = \mathbb{E}\{Y(1) - Y(0)\}$. Our methodology also applies to the broader class of causal estimands in Li, Morgan and Zaslavsky

(2018); we discuss the extension to other causal estimands in Section 5.1. For simplicity of exposition, denote

$$\mu_a(X) = \mathbb{E}\{Y(a) \mid X\} \text{ and } e(X) = \mathbb{P}(A = 1 \mid X),$$

where $\mu_a(X)$ is an outcome mean function for $a = 0, 1$, and $e(X)$ is the propensity score.

It is well known that under the common assumptions in the causal inference literature, including the treatment ignorability and overlap assumptions (Assumptions 1 and 2 in the Supplementary Material), the ACE can be identified using various estimators, including the outcome regression, augmented inverse probability weighting (AIPW), IPW and matching estimators; see Imbens (2004) and Rosenbaum (2002) for surveys of these estimators. These common estimators are asymptotically linear and belong to the class of estimators in our general setup. We review these estimators below and identify their influence functions in the Supplementary Material.

The common estimators require correct specifications of different parts of the observed data distribution, including the outcome model and propensity score.

**Assumption 1 (Outcome model).** *The parametric model $\mu_a(X; \beta_a)$ is a correct specification for $\mu_a(X)$, for $a = 0, 1$; i.e., $\mu_a(X) = \mu_a(X; \beta_a^*)$, where $\beta_a^*$ is the true model parameter.*

**Assumption 2 (Propensity score model).** *The parametric model $e(X; \alpha)$ is a correct specification for $e(X)$; i.e., $e(X) = e(X; \alpha^*)$, where $\alpha^*$ is the true model parameter.*

**Example 1.** *The outcome regression estimator is $\hat{\tau}_{n,\text{reg}} = n^{-1} \sum_{i=1}^n \tau_{\text{reg},i}$, where*

$$\tau_{\text{reg},i} = \mu_1(X_i; \hat{\beta}_1) - \mu_0(X_i; \hat{\beta}_0). \tag{2.3}$$

**Example 2.** *The IPW estimator is $\hat{\tau}_{n,\text{IPW}} = n^{-1} \sum_{i=1}^n \tau_{\text{IPW},i}$, where*

$$\tau_{\text{IPW},i} = \frac{A_i Y_i}{e(X_i; \hat{\alpha})} - \frac{(1 - A_i) Y_i}{1 - e(X_i; \hat{\alpha})}. \tag{2.4}$$

**Example 3.** *The AIPW estimator is $\hat{\tau}_{n,\text{AIPW}} = n^{-1} \sum_{i=1}^n \tau_{\text{AIPW},i}$, where*

$$\begin{aligned} \tau_{\text{AIPW},i} = {} & \frac{A_i Y_i}{e(X_i; \hat{\alpha})} + \left\{1 - \frac{A_i}{e(X_i; \hat{\alpha})}\right\} \mu_1(X_i; \hat{\beta}_1) \\ & - \frac{(1 - A_i) Y_i}{1 - e(X_i; \hat{\alpha})} - \left\{1 - \frac{1 - A_i}{1 - e(X_i; \hat{\alpha})}\right\} \mu_0(X_i; \hat{\beta}_0). \end{aligned} \tag{2.5}$$

**Example 4 (Matching).** *For unit $i$, denote the imputed potential outcomes as*

$$\hat{Y}_i(1) = \begin{cases} M^{-1}\sum_{j\in\mathcal{J}_X(i)} Y_j & \text{if } A_i = 0, \\ Y_i & \text{if } A_i = 1, \end{cases} \quad \hat{Y}_i(0) = \begin{cases} Y_i & \text{if } A_i = 0, \\ M^{-1}\sum_{j\in\mathcal{J}_X(i)} Y_j & \text{if } A_i = 1. \end{cases}$$

*The matching estimator of $\tau$ is*

$$\hat{\tau}^{(0)}_{n,\text{mat}} = \frac{1}{n}\sum_{i=1}^{n}\{\hat{Y}_i(1) - \hat{Y}_i(0)\} = \frac{1}{n}\sum_{i=1}^{n}(2A_i - 1)\left(Y_i - M^{-1}\sum_{l\in\mathcal{J}_X(i)} Y_l\right). \quad (2.6)$$

*where $M$ ($M \geq 1$) is the number of matches and $\mathcal{J}_X(i)$ is the index set of the nearest $M$ neighbors for unit $i$ in its opposite treatment group based on the matching variable $X$.*

The above estimators are asymptotically linear with the influence functions given in the Supplementary Material.

## 3. MI to Deal with Missing Values

### 3.1. General MI

Continuing with the general setup in Section 2.1, we now consider the case where $L$ is $q$-dimensional and $L = (L_{[1]}, \ldots, L_{[q]})$ contains missing values. Let $R = (R_{[1]}, \ldots, R_{[q]})$ be the vector of missing indicators, such that $R_{[j]} = 1$ if the $j$th component $L_{[j]}$ is observed, and zero if it is missing. In addition, let $1_q$ denote the $q$-vector of ones. We write $L = (L_R, L_{\overline{R}})$, where $L_R$ and $L_{\overline{R}}$ represent the observed and missing parts of $L$, respectively. This notation depends on the missingness pattern; for example, if $R_{[1]} = 1$ and $R_{[j]} = 0$, for $j = 2, \ldots, q$, then $L_R = L_{[1]}$ and $L_{\overline{R}} = (L_{[2]}, \ldots, L_{[q]})$. With missing values in $L$, the full-sample estimator $\hat{\tau}_n$ is not feasible to calculate.

To facilitate applying a full-sample estimator, MI creates multiple complete data sets by filling in missing values. Assume unit $i$ has the complete data $Z_i = (L_i, R_i)$ and the observed data $Z_{\text{obs},i} = (L_{R_i,i}, R_i)$. Denote $\mathbf{Z} = (Z_1, \ldots, Z_n)$ and $\mathbf{Z}_{\text{obs}} = (Z_{\text{obs},1}, \ldots, Z_{\text{obs},n})$. Assume that the observed data likelihood is $f(\mathbf{Z}_{\text{obs}}; \theta)$, with the true parameter value $\theta_0$. The MI procedure proceeds as follows.

**Step MI-1.** Create $m$ complete data sets by filling in missing values using imputed values generated from the posterior predictive distribution. Specifically, to create the $j$th imputed data set, first generate $\theta^{*(j)}$ from the posterior distribution $p(\theta \mid \mathbf{Z}_{\text{obs}})$, and then generate $L^{*(j)}_{\overline{R}_i,i}$ from $f(L_{\overline{R}_i,i} \mid Z_{\text{obs},i}; \theta^{*(j)})$ for each missing $L_{\overline{R}_i,i}$.

**Step MI-2.** Apply a full-sample estimator of $\tau$ to each imputed data set. Let $\hat{\tau}^{(j)}$ be the estimator applied to the $j$th imputed data set, and $\hat{V}^{(j)}$ be the full-sample variance estimator for $\hat{\tau}^{(j)}$.

**Step MI-3.** Use Rubin's combining rule to summarize the results from the multiple imputed data sets. The MI estimator of $\tau$ is $\hat{\tau}_{\mathrm{MI}} = m^{-1}\sum_{j=1}^{m}\hat{\tau}^{(j)}$, and Rubin's variance estimator is

$$\hat{V}_{\mathrm{MI}}(\hat{\tau}_{\mathrm{MI}}) = W_m + (1+m^{-1})B_m, \tag{3.1}$$

where $W_m = m^{-1}\sum_{j=1}^{m}\hat{V}^{(j)}$ and $B_m = (m-1)^{-1}\sum_{j=1}^{m}(\hat{\tau}^{(j)} - \hat{\tau}_{\mathrm{MI}})^2$.

**Remark 1.** In Step MI-1, the full/observed data likelihood has to be specified and fitted for MI, which can be challenging in the presence of several, if not many variables. In practice, we suggest specifying the full data likelihood as a product of a sequence of conditional models of one variable, given the preceding variables, allowing model flexibility for each variable (e.g., the error distribution matches the variable type, which is logistic for a binary variable). Furthermore, we can perform a model diagnosis after the imputation to assess the goodness-of-fit; see the real-data application in Section 7 for an example.

### 3.2. CI in the presence of confounders missing at random

We explain our method by estimating the ACE, assuming the confounders are missing at random (MAR), in the sense of Rubin (1976). Extensions to settings with missing outcomes and different missingness mechanisms are provided in Section 5. We now consider the case in which values are missing from $X = (X_{[1]}, \ldots, X_{[p]})$, a $p$-dimensional vector. Accordingly, let $R_X = (R_{[1]}, \ldots, R_{[p]})$ be the vector of missing indicators, such that $R_{[j]} = 1$ if the $j$th component $X_{[j]}$ is observed, and zero if it is missing. We write $X = (X_{R_X}, X_{\overline{R}_X})$, where $X_{R_X}$ and $X_{\overline{R}_X}$ represent the observed and missing parts of $X$, respectively. With values missing from $X$, the aforementioned full-sample estimators (2.3)–(2.6) are not feasible to calculate. Estimating of the ACE requires further assumptions. Following most of the empirical literature, we impose the MAR assumption.

**Assumption 3 (Missingness at random).** *We have* $X_{\overline{R}_X} \perp\!\!\!\perp R_X \mid Z_{\mathrm{obs}}$.

Assumption 3 holds if the observed data capture all the information related to missingness. Under Assumption 3, $f(A_i, X_i, Y_i, R_{Xi}; \theta) = f(A_i, X_{R_{Xi},i}, Y_i, R_{Xi}; \theta)$ $f(X_{\overline{R}_{Xi},i} | A_i, X_{R_{Xi},i}, Y_i, R_{Xi} = 1_p; \theta)$ is identifiable, which justifies the likelihood-based or Bayesian inference. Moreover, by the Bayes rule, the posterior distribution of the missing data can be expressed as

$$\begin{aligned}
&f(X_{\overline{R}_{Xi},i} \mid A_i, X_{R_{Xi},i}, Y_i, R_{Xi}; \theta^{*(j)}) \propto f(A_i, X_{\overline{R}_{Xi},i}, X_{R_{Xi},i}, Y_i, R_{Xi}; \theta^{*(j)}) \\
&= f(R_{Xi} \mid Y_i, X_{R_{Xi},i}, X_{\overline{R}_{Xi},i}, A_i; \theta^{*(j)}) f(Y_i, X_{R_{Xi},i}, X_{\overline{R}_{Xi},i}, A_i; \theta^{*(j)}) \\
&\propto f(Y_i, X_{R_{Xi},i}, X_{\overline{R}_{Xi},i}, A_i; \theta^{*(j)}) \\
&\propto f(Y_i \mid X_{R_{Xi},i}, X_{\overline{R}_{Xi},i}, A_i; \theta^{*(j)}) f(A_i \mid X_{R_{Xi},i}, X_{\overline{R}_{Xi},i}; \theta^{*(j)}) \\
&\quad f(X_{\overline{R}_{Xi},i} \mid X_{R_{Xi},i}; \theta^{*(j)}),
\end{aligned} \tag{3.2}$$

Table 1. Simulation results of the full-sample point estimators and MI point estimators based on 5,000 simulated data sets

| Method $\hat{\tau}_n$ | $\mathbb{V}(\hat{\tau}_n)$ $(\times 10^4)$ | $\mathbb{V}(\hat{\tau}_{\mathrm{MI}})$ $(\times 10^4)$ | $\mathbb{V}(\hat{\tau}_{\mathrm{MI}} - \hat{\tau}_n)$ $(\times 10^4)$ | $\mathrm{cov}(\hat{\tau}_{\mathrm{MI}} - \hat{\tau}_n, \hat{\tau}_n)$ $(\times 10^4)$ |
|---|---|---|---|---|
| Regression | 24 | 35 | 11 | 0 |
| IPW | 62 | 66 | 22 | -9 |
| AIPW | 25 | 36 | 12 | 0 |
| matching | 30 | 38 | 15 | -4 |

where (3.2) follows because $f(R_{Xi} \mid Y_i, X_{R_{Xi},i}, X_{\overline{R}_{Xi},i}, A_i; \theta^{*(j)}) = f(R_{Xi} \mid Y_i, X_{R_{Xi},i}, A_i; \theta^{*(j)})$, by Assumption 3. The MI procedure uses the imputation model for $X_{\overline{R}_{Xi},i}$, which does not depend on the missingness pattern probability for $R_{Xi}$.

### 3.3. Issue of standard inference with MI

The variance of the MI estimator can be decomposed to

$$\mathbb{V}(\hat{\tau}_{\mathrm{MI}}) = \mathbb{V}(\hat{\tau}_n) + \mathbb{V}(\hat{\tau}_{\mathrm{MI}} - \hat{\tau}_n) + 2\mathrm{cov}(\hat{\tau}_{\mathrm{MI}} - \hat{\tau}_n, \hat{\tau}_n).$$

In Rubin's variance estimator (3.1), $W_m$ estimates the within-imputation variance $\mathbb{V}(\hat{\tau}_n)$, and $(1 + m^{-1})B_m$ estimates the between-imputation variance $\mathbb{V}(\hat{\tau}_{\mathrm{MI}} - \hat{\tau}_n)$. However, it ignores the covariance between $\hat{\tau}_{\mathrm{MI}} - \hat{\tau}_n$ and $\hat{\tau}_n$. Rubin's variance estimator is asymptotically unbiased only under the congeniality condition (Meng (1994)), that is, $\mathrm{cov}(\hat{\tau}_{\mathrm{MI}} - \hat{\tau}_n, \hat{\tau}_n) = o(1)$. Therefore, Rubin's variance estimator using a different full-sample estimator $\hat{\tau}_n$ may be inconsistent.

For illustration, we conduct a numerical experiment to assess the congeniality condition for the outcome regression, IPW, AIPW, and matching estimators of the ACE. The data-generating mechanism is described in scenario (a) in Section 6. For each simulated data set, we compute the full-sample point estimators $\hat{\tau}_n$, assuming the confounders are fully observed and the MI point estimators $\hat{\tau}_{\mathrm{MI}}$. Table 1 presents the simulations results of the variances of the full-sample point estimators and the MI point estimators, and the covariance between $\hat{\tau}_{\mathrm{MI}} - \hat{\tau}_n$ and $\hat{\tau}_n$. The covariance is significantly negative for the IPW and matching estimators. Rubin's variance estimator overestimates the variances of these estimators and, as a result, MI is not congenial for them. Thus, the congeniality condition required for MI can be quite restrictive for general ACE estimation.

## 4. A Martingale Representation of the MI Estimators of Causal Effects

### 4.1. A novel martingale representation

Based on the unified linear form of the full-sample estimator, as in (2.1) or (2.2), we express the MI estimator in general form as

$$\hat{\tau}_{\mathrm{MI}} - \tau = \frac{1}{m}\sum_{j=1}^{m}(\hat{\tau}^{(j)} - \tau) = \frac{1}{nm}\sum_{i=1}^{n}\sum_{j=1}^{m}\psi(L_i^{*(j)}) + o_{\mathbb{P}}(n^{-1/2}), \tag{4.1}$$

where $L_i^{*(j)} = (L_{R_i,i}, L_{\overline{R}_i,i}^{*(j)})$ and $o_{\mathbb{P}}(n^{-1/2})$ is from (2.1), or

$$\hat{\tau}_{\mathrm{MI}} - \tau = \frac{1}{m}\sum_{j=1}^{m}(\hat{\tau}^{(j)} - \tau) = \frac{1}{nm}\sum_{i=1}^{n}\sum_{j=1}^{m}\psi_i(\mathbf{L}^{*(j)}) + o_{\mathbb{P}}(n^{-1/2}), \tag{4.2}$$

where $\mathbf{L}^{*(j)} = (L_1^{*(j)}, \ldots, L_n^{*(j)})$ and $o_{\mathbb{P}}(n^{-1/2})$ is from (2.2). In the following, we explain our framework using (4.1); the same exposition applies to (4.2) by replacing $\psi(L_i)$ with $\psi_i(\mathbf{L})$ and $L_i^{*(j)}$ with $\mathbf{L}^{*(j)}$.

To express (4.1) further, it is important to understand the properties of the posterior distribution and the imputed values $L_i^{*(j)}$. Using the Bernstein–von Mises theorem (van der Vaart (2000); Chap. 10), under the regularity conditions described in Assumption 4, conditioned on the observed data, the posterior distribution $p(\theta \mid \mathbf{Z}_{\mathrm{obs}})$ converges to a normal distribution with mean $\hat{\theta}$ and variance $n^{-1}\mathcal{I}_{\mathrm{obs}}^{-1}$ almost surely, where $\hat{\theta}$ is the MLE of $\theta_0$, and $\mathcal{I}_{\mathrm{obs}}^{-1}$ is the inverse of the Fisher information matrix. Let $S(\theta; L, R)$ be the score function of $\theta$. In the presence of missing data, define the mean score function $\bar{S}(\theta_0; Z_{\mathrm{obs},i}) = \mathbb{E}\{S(\theta_0; L_i, R_i) \mid Z_{\mathrm{obs},i}, \theta_0\}$.

The MLE $\hat{\theta}$ can be viewed as the solution to the mean score equation $\sum_{i=1}^{n}\bar{S}(\theta; Z_{\mathrm{obs},i}) = 0$. Under the regularity conditions described in Assumption 4, we can then express $\hat{\theta} - \theta_0 = n^{-1}\mathcal{I}_{\mathrm{obs}}^{-1}\sum_{i=1}^{n}\bar{S}(\theta_0; Z_{\mathrm{obs},i}) + o_{\mathbb{P}}(n^{-1/2})$. It is insightful to write (4.1) as

$$\begin{aligned}\hat{\tau}_{\mathrm{MI}} - \tau = {} & \frac{1}{nm}\sum_{i=1}^{n}\sum_{j=1}^{m}\Big[\psi(L_i^{*(j)}) - \mathbb{E}\{\psi(L_i) \mid \mathbf{Z}_{\mathrm{obs}}, \hat{\theta}\}\Big] \\ & + \frac{1}{nm}\sum_{i=1}^{n}\sum_{j=1}^{m}\mathbb{E}\{\psi(L_i) \mid \mathbf{Z}_{\mathrm{obs}}, \hat{\theta}\} + o_{\mathbb{P}}(n^{-1/2}),\end{aligned} \tag{4.3}$$

where we recall $\mathbf{Z}_{\mathrm{obs}} = (Z_{\mathrm{obs},1}, \ldots, Z_{\mathrm{obs},n})$. Now, by a Taylor expansion of $\mathbb{E}\{\psi(L_i) \mid \mathbf{Z}_{\mathrm{obs}}, \hat{\theta}\}$ around the true value $\theta_0$,

$$\hat{\tau}_{\mathrm{MI}} - \tau = \frac{1}{nm}\sum_{i=1}^{n}\sum_{j=1}^{m}\Big[\psi(L_i^{*(j)}) - \mathbb{E}\{\psi(L_i) \mid \mathbf{Z}_{\mathrm{obs}}, \hat{\theta}\}\Big] \tag{4.4}$$

$$+\frac{1}{nm}\sum_{i=1}^{n}\sum_{j=1}^{m}[\mathbb{E}\{\psi(L_i)\mid \mathbf{Z}_{\text{obs}},\theta_0\}+\Gamma\mathcal{I}_{\text{obs}}^{-1}\bar{S}(\theta_0;Z_{\text{obs},i})]+o_{\mathbb{P}}(n^{-1/2}),$$

where $\Gamma=\mathbb{E}\big[\mathbb{E}\{\psi(L_i)S(\theta_0;L_i,R_i)\mid \mathbf{Z}_{\text{obs}},\theta_0\}-\mathbb{E}\{\psi(L_i)\mid \mathbf{Z}_{\text{obs}},\theta_0\}\bar{S}(\theta_0;Z_{\text{obs},i})\big]^{\mathrm{T}}$.

Based on (4.4), we can write

$$n^{1/2}(\hat{\tau}_{\text{MI}}-\tau)=\sum_{k=1}^{n+nm}\xi_{n,k}+o_{\mathbb{P}}(n^{-1/2}), \tag{4.5}$$

where

$$\xi_{n,k}=\begin{cases}\dfrac{1}{n^{1/2}}\left[\mathbb{E}\{\psi(L_i)\mid \mathbf{Z}_{\text{obs}},\theta_0\}+\Gamma\mathcal{I}_{\text{obs}}^{-1}\bar{S}(\theta_0;Z_{\text{obs},i})\right], & \text{if } k=i,\\ \dfrac{1}{n^{1/2}m}\big[\psi(L_i^{*(j)})-\mathbb{E}\{\psi(L_i)\mid \mathbf{Z}_{\text{obs}},\hat{\theta}\}\big], & \text{if } k=n+(i-1)m+j,\end{cases}$$

where $i=1,\dots,n$ and $j=1,\dots,m$. For the decomposition in (4.5), the first $n$ terms of $\xi_{n,k}$ contribute to the variability of $\hat{\tau}_{\text{MI}}$, because of the unknown parameters, and the remaining $nm$ terms of $\xi_{n,k}$ contribute to the variability of $\hat{\tau}_{\text{MI}}$, because of the imputations given the parameter values, reflecting the sequential MI procedure.

We discuss the mean properties of $\xi_{n,k}$ in order to create suitable $\sigma$-fields in the martingale presentation. For $k=i$, where $i=1,\dots,n$, we have

$$\begin{aligned}\mathbb{E}(\xi_{n,k})&=\frac{1}{n^{1/2}}\mathbb{E}\left[\mathbb{E}\{\psi(L_i)\mid \mathbf{Z}_{\text{obs}},\theta_0\}+\Gamma\mathcal{I}_{\text{obs}}^{-1}\bar{S}(\theta_0;Z_{\text{obs},i})\right]\\&=\frac{1}{n^{1/2}}\mathbb{E}\{\psi(L_i)\}+\frac{1}{n^{1/2}}\Gamma\mathcal{I}_{\text{obs}}^{-1}\mathbb{E}\{\bar{S}(\theta_0;Z_{\text{obs},i})\}=0,\end{aligned} \tag{4.6}$$

where $\mathbb{E}\{\psi(L_i)\}=0$ and $\mathbb{E}\{\bar{S}(\theta_0;Z_{\text{obs},i})\}=0$ are from the mean zero property of the influence function and the mean score function. For $k=n+(i-1)m+j$, where $i=1,\dots,n$ and $j=1,\dots,m$, we have

$$\begin{aligned}\mathbb{E}(\xi_{n,k}\mid \mathbf{Z}_{\text{obs}})&=\frac{1}{n^{1/2}m}\mathbb{E}\left[\psi(L_i^{*(j)})-\mathbb{E}\{\psi(L_i)\mid \mathbf{Z}_{\text{obs}},\hat{\theta}\}\mid \mathbf{Z}_{\text{obs}}\right]\\&=\frac{1}{n^{1/2}m}\left[\mathbb{E}\{\psi(L_i^{*(j)})\mid \mathbf{Z}_{\text{obs}}\}-\mathbb{E}\{\psi(L_i)\mid \mathbf{Z}_{\text{obs}},\hat{\theta}\}\right]=0,\end{aligned} \tag{4.7}$$

where the last equality follows because, given $\mathbf{Z}_{\text{obs}}$, the posterior predictive distribution of $L_i^{*(j)}$ follows the distribution $f(L_i\mid \mathbf{Z}_{\text{obs}};\hat{\theta})$, by the Bernstein–von Mises theorem (van der Vaart (2000, Chap. 10)). Consider the $\sigma$-fields $\mathcal{F}_{n,k}=\sigma\{\mathbb{N}\}$ if $k=i$, with $\mathbb{N}$ being the null set, and $\mathcal{F}_{n,k}=\sigma\{\mathbf{Z}_{\text{obs}}\}$ if $k=n+(i-1)m+j$, where $i=1,\dots,n$ and $j=1,\dots,m$. Therefore, by (4.6) and (4.7),

$$\left\{\sum_{i=1}^{k}\xi_{n,i},\mathcal{F}_{n,k},1\le k\le n(1+m)\right\}\text{ is a martingale for each } n\ge 1.$$

Equation (4.4) is a martingale representation of the MI estimator by expressing the MI estimator in terms of a series of random variables that have mean zero, conditional on the sigma algebra generated from the preceding variables. This martingale representation is used to construct the bootstap replicate for the variance estimation.

### 4.2. Wild bootstrap for the MI estimator

Invoked by the martingale representation, we propose a wild bootstrap procedure (Wu (1986); Liu (1988)), which provides a valid variance estimation and inference of the linear statistic for martingale difference arrays, based on the martingale central limit theory, for estimating the variance of $\hat{\tau}_{\mathrm{MI}}$.

**Step 1.** Sample $u_k$, for $k = 1, \ldots, n + nm$, to satisfy that $\mathbb{E}(u_k \mid \mathbf{Z}_{\mathrm{obs}}) = 0$, $\mathbb{E}(u_k^2 \mid \mathbf{Z}_{\mathrm{obs}}) = 1$, and $\mathbb{E}(u_k^4 \mid \mathbf{Z}_{\mathrm{obs}}) < \infty$.

**Step 2.** Compute the bootstrap replicate as $T^* = n^{-1/2} \sum_{k=1}^{n+nm} \hat{\xi}_{n,k} u_k$, where

$$\hat{\xi}_{n,k} = \begin{cases} \dfrac{1}{n^{1/2}} \left[ \mathbb{E}\{\psi(L_i) \mid \mathbf{Z}_{\mathrm{obs}}, \hat{\theta}\} + \hat{\Gamma} \hat{\mathcal{I}}_{\mathrm{obs}}^{-1} \bar{S}(\hat{\theta}; Z_{\mathrm{obs},i}) \right], & \text{if } k = i, \\ \dfrac{1}{n^{1/2} m} \left[ \psi(L_i^{*(j)}) - \mathbb{E}\{\psi(L_i) \mid \mathbf{Z}_{\mathrm{obs}}, \hat{\theta}\} \right], & \text{if } k = n + (i-1)m + j, \end{cases}$$

where $i = 1, \ldots, n$ and $j = 1, \ldots, m$.

**Step 3.** Repeat Step 1–Step 2 $B$ times, and estimate the variance of $\hat{\tau}_{\mathrm{MI}}$ using the sample variance of the $B$ copies of $T^*$.

**Remark 2.** There are many choices for generating $u_k$, such as the standard normal distribution, Mammen's two-point distribution (Mammen (1993)),

$$u_k = \begin{cases} \dfrac{1 - 5^{1/2}}{2}, & \text{with probability } \dfrac{1 + 5^{-1/2}}{2}, \\ \dfrac{5^{1/2} + 1}{2}, & \text{with probability } \dfrac{1 - 5^{-1/2}}{2}, \end{cases}$$

a simpler distribution with probability 0.5 of being 1 and probability 0.5 of being $-1$, or the Poisson distribution with parameter one re-centered at zero (Beyersmann, Termini and Pauly (2013)). Our simulation study shows that the wild bootstrap procedure is not sensitive to the choice of the sampling distribution of $u_k$. In particular, one can also use nonparametric bootstrap weights; that is, let $u_k = (nm + n)^{-1/2}(W_k - \overline{W})$, where $\{W_k : k = 1, \ldots, n(m + 1)\}$ follows a multinomial distribution with $n(m + 1)$ draws on $n(m + 1)$ cells with equal probability, and $\overline{W} = (nm + n)^{-1} \sum_{k=1}^{n(m+1)} W_k$.

Several authors have used a nonparametric bootstrap to estimate the variance of MI estimators. Schomaker and Heumann (2018) combined MI with a bootstrap to perform an inference for the quantity of interest. However, their discussions

are restricted to the MLEs of the model parameters and require a bootstrap on top of MI, which is computationally intensive. Moreover, in the causal inference literature, in the absence of missing data, Abadie and Imbens (2008) show that a nonparametric bootstrap cannot provide a consistent variance estimation for the matching estimators of the ACE, owing to the nonsmooth nature of the matching procedure. Note that the proposed wild bootstrap procedure with the nonparametric bootstrap weights differs from the naive bootstrap. The martingale representation and wild bootstrap procedure work for asymptotically linear ACE estimators, including the matching estimator.

**Remark 3.** In Step 2, we approximate $\xi_{n,k}$, which involves the MLE $\hat{\theta}$, the estimated observed Fisher information, and the conditional expectations taken with respect to the distribution of the missing values, given the observed values. These estimators are readily available from the posterior draws, or can be approximated by using Monte Carlo integration based on the imputed values. For example, we approximate $\mathbb{E}\{\psi(L_i) \mid \mathbf{Z}_{\text{obs}}, \hat{\theta}\}$ by $M^{-1}\sum_{j=1}^{M}\psi(L_i^{*(j)})$. Thus, the computation is not as intimidating as it appears, although it is heavier than Rubin's combining rule. However, as shown in Theorem 1, the proposed inference procedure is valid, whereas Rubin's method may not be.

We show the asymptotic validity of the above bootstrap inference method in the following theorem, with regularity assumptions.

**Assumption 4.** *Suppose the standard conditions hold for the MLE $\hat{\theta}$ to be $n^{1/2}$-consistent for $\theta_0$:*

1. *$Z_{\text{obs},1}, \ldots, Z_{\text{obs},n}$ are i.i.d. and follow $f(z \mid \theta)$;*
2. *$\theta$ is identifiable; that is, if $\theta \neq \theta'$, then $f(z \mid \theta) \neq f(z \mid \theta')$;*
3. *the density $f(z \mid \theta)$ has a common support (not depending on $\theta$);*
4. *the parameter space contains an open set, of which the true parameter $\theta_0$ is an interior point.;*
5. *for every $z$ in the support, $f(z \mid \theta)$ is three times differentiable with respect to $\theta$, the third derivative is continuous in $\theta$, and $\int \partial^3 \log f(z \mid \theta)/\partial\theta^3 \mathrm{d}z < \infty$;*
6. *for any $\theta_0$ in the parameter space, there exists a positive number $c$ and a function $M(z)$, such that $|\partial^3 \log f(z \mid \theta)/\partial\theta^3| \leq M(z)$, for all $z$ in the support, $\theta_0 - c < \theta < \theta_0 + c$, with $\mathbb{E}_{\theta_0}\{M(Z)\} < \infty$.*

Define $\bar{\psi}(\theta; Z_{\text{obs},i}) = \mathbb{E}\{\psi(L_i) \mid Z_{\text{obs},i}, \theta\}$.

**Assumption 5.** *$\bar{\psi}(\theta; \mathbf{Z}_{\text{obs}})$, $\mathbb{V}\{\psi(L_i) \mid \mathbf{Z}_{\text{obs}}, \theta\}$, $\bar{S}(\theta; Z_{\text{obs},i})$, and $\mathbb{V}\{S(\theta; L_i, R_i) \mid Z_{\text{obs},i}, \theta\}$ are continuous functions of $\theta$.*

**Assumption 6.** $\mathbb{E}[\{\bar{\psi}(\theta; \mathbf{Z}_{\text{obs}})\}^4] < \infty$ *and* $\mathbb{E}[\{\bar{S}(\theta; Z_{\text{obs},i})\}^4] < \infty$, *for* $\theta$ *in a neighborhood of* $\theta_0$.

**Assumption 7.** $\{\bar{\psi}(\theta; \mathbf{Z}_{\text{obs}} - \bar{\psi}(\theta_0; \mathbf{Z}_{\text{obs}})\}^2$ *and* $\{\bar{S}(\theta; Z_{\text{obs},i}) - \bar{S}(\theta_0; Z_{\text{obs},i})\}^2$ *belong to a Donsker class.*

Assumption 4 is the standard assumption in the literature to guarantee the consistency of the MLE (van der Vaart (2000)). Assumption 5 is imposed to guarantee sufficient smoothness on the conditional mean and variance functions for the influence function and the score function. It holds for the general estimands, such as themean-type estimands, and the commonly used class of parametric models, such as the exponential family. For Assumption 6, the moment conditions are used to invoke the central limit theory, and typically hold for the general estimands and parametric models, coupled with the bounded moment conditions for $L$. In practice, $L$ often has a bounded support, and thus the bounded moment conditions are reasonable. Assumption 7 ensures the convergence of the empirical process to its limiting version (Kennedy (2016)). Interested readers can consult Kennedy (2016) for details and examples of the Donsker class.

**Theorem 1.** *Suppose that Assumptions* 1, 2 *and* 4–7 *hold. Suppose that* $f(L_{\overline{R}_i,i} \mid Z_{\text{obs},i}; \theta)$ *is specified correctly. Then, for MI that adopts the full-sample estimator that satisfies* (2.1) *or* (2.2), *we have*

$$\sup_r \left| \mathbb{P}(n^{1/2} T^* \le r \mid \mathbf{Z}_{\text{obs}}) - \mathbb{P}\{n^{1/2}(\hat{\tau}_{\text{MI}} - \tau) \le r\} \right| \xrightarrow{\mathbb{P}} 0,$$

*as* $n \to \infty$.

We provide the proof of Theorem 1 in the Supplementary Material, which draws on the martingale central limit theory (Hall and Heyde (1980)) and the asymptotic property of the weighted sampling of martingale difference arrays (Pauly (2011)). Theorem 1 indicates that the distribution of the wild bootstrap statistic consistently estimates the distribution of the MI estimator.

Theorem 1 requires that the imputation model $f(L_{\overline{R}_i,i} \mid Z_{\text{obs},i}; \theta)$ be specified correctly (the congeniality condition of Meng (1994)). This requirement is needed, not only for the consistency of the MI variance estimator, but also for the consistency of the MI point estimator. The following corollaries hereafter clarify the required imputation models in various scenarios.

**Corollary 1.** *For the scenario with confounders MAR, the assumption that the imputation model* $f(L_{\overline{R}_i,i} \mid Z_{\text{obs},i}; \theta)$ *is specified correctly in Theorem* 1 *implies that the outcome distribution* $f(Y_i \mid X_i, A_i; \theta)$, *the propensity score model* $f(A_i \mid X_i; \theta)$, *and the confounder distribution* $f(X_{\overline{R}_{Xi},i} \mid X_{R_{Xi},i}; \theta)$ *should be specified correctly.*

## 5. Extensions

### 5.1. Different causal estimands

Our inference framework extends to a wide class of causal estimands, as long as the estimand admits an asymptotically linear full-sample estimator, as in (2.1). For example, we can consider the ACEs over a subset of the population (Crump et al. (2006); Li, Morgan and Zaslavsky (2018)), including the ACE on the treated. We can also consider nonlinear causal estimands. For example, for a binary outcome, the log of the causal risk ratio is

$$\log \mathrm{CRR} = \log \frac{\mathbb{P}\{Y(1)=1\}}{\mathbb{P}\{Y(0)=1\}} = \log \frac{\mathbb{E}\{Y(1)\}}{\mathbb{E}\{Y(0)\}},$$

and the log of the causal odds ratio is

$$\log \mathrm{COR} = \log \frac{\mathbb{P}\{Y(1)=1\}/\mathbb{P}\{Y(1)=0\}}{\mathbb{P}\{Y(0)=1\}/\mathbb{P}\{Y(0)=0\}} = \log \frac{\mathbb{E}\{Y(1)\}/[1-\mathbb{E}\{Y(1)\}]}{\mathbb{E}\{Y(0)\}/[1-\mathbb{E}\{Y(0)\}]}.$$

The key insight is that under Assumptions 1 and 2, we can estimate $\mathbb{E}\{Y(a)\}$ using commonly used estimators, denoted by $\hat{\mathbb{E}}\{Y(a)\}$, for $a = 0, 1$. We can then obtain an estimator for $\log \mathrm{CRR}$ as $\log[\hat{\mathbb{E}}\{Y(1)\}/\hat{\mathbb{E}}\{Y(0)\}]$. By the Taylor expansion, we can linearize these estimators and establish a similar linear form as (2.1), which serves as the basis for constructing the weighted bootstrap inference.

### 5.2. Missingness not at random

If Assumption 3 fails, the missing pattern also depends on the missing values themselves, even after controlling for the observed data, a scenario known as MNAR. In our motivating example discussed in Section 7, the family poverty ratio is likely to be MNAR because subjects with higher income may be less likely to disclose their income information (Davern et al. (2005)). In general, MNAR occurs frequently for sensitive questions related to, for example, alcohol consumption, income, and so on.

Causal inference with data MNAR is more challenging because the full data distribution, and therefore the ACE, are not identifiable, in general. To use MI in causal inference with confounders MNAR, we require identification conditions that ensure that the full data distribution is identifiable. For example, Wang, Shao and Kim (2014) introduce a nonresponse instrument as a sufficient condition for the identifiability of the observed likelihood. Miao, Ding and Geng (2016) investigate the identifiability of normal and normal mixture models with nonignorable missing data. Yang, Wang and Ding (2019) propose an outcome-independence missingness mechanism, under which, the missing-data mechanism is independent of the outcome, given the treatment and confounders, and establish general identification conditions.

Our proposed method can easily extend to the scenario in which the confounders are MNAR when additional assumptions are made for the identifiability of the full data distribution. After the identification check, we need only modify the posterior predictive distribution of $X^{(j)}_{\overline{R}_i,i}$. For example, following Yang, Wang and Ding (2019), we assume that the missingness pattern $R$ is independent of the outcome, given the treatment and confounders.

**Assumption 8 (Outcome-independent missingness).** *We have* $Y \perp\!\!\!\perp R_X \mid (A, X_{R_X}, X_{\overline{R}_X})$.

Under the regularity conditions in Yang, Wang and Ding (2019), $f(A, X, Y, R_X)$ is identifiable (Yang, Wang and Ding (2019)). Then, in Step MI-1, the posterior distribution of $X^{(j)}_{\overline{R}_{Xi},i}$ can be decomposed to

$$\begin{aligned}
&f(X_{\overline{R}_{Xi},i} \mid A_i, X_{R_{Xi},i}, Y_i, R_{Xi}; \theta^{*(j)}) \\
&\propto f(Y_i \mid X_{R_{Xi},i}, X_{\overline{R}_{Xi},i}, A_i; \theta^{*(j)}) \times f(R_{Xi} \mid X_{R_{Xi},i}, X_{\overline{R}_{Xi},i}, A_i; \theta^{*(j)}) \\
&\quad f(A_i \mid X_{R_{Xi},i}, X_{\overline{R}_{Xi},i}; \theta^{*(j)}) f(X_{\overline{R}_{Xi},i} \mid X_{R_{Xi},i}; \theta^{*(j)}).
\end{aligned}$$

After imputation, the wild bootstrap steps remain the same.

**Corollary 2.** *For the scenario with confounders MNAR, the assumption that the imputation model* $f(L_{\overline{R}_i,i} \mid Z_{\text{obs},i}; \theta)$ *is specified correctly in Theorem* 1 *implies that the outcome distribution* $f(Y_i \mid X_i, A_i; \theta)$*, propensity score model* $f(A_i \mid X_i; \theta)$*, confounder distribution* $f(X_{\overline{R}_{Xi},i} \mid X_{R_{Xi},i}; \theta)$*, and missingness model* $f(R_{Xi} \mid X_i, A_i; \theta)$ *should be specified correctly.*

### 5.3. Partially observed outcome and confounders

In some cases, both the outcome and the confounders are subject to missingness. Our framework can easily accommodate this scenario by adding an outcome imputation step to the MI procedure.

We now introduce another missingness indicator $R_Y$ for $Y$; that is, $R_Y = 1$ if $Y$ is observed, and $R_Y = 0$ otherwise. In Step MI-1, we first generate $\theta^{*(j)}$ from the posterior distribution $p(\theta \mid \mathbf{Z}_{\text{obs}})$. Then, for unit $i$ with $R_Y = 1$, generate $X^{*(j)}_{\overline{R}_{Xi},i}$ from $f(X_{\overline{R}_{Xi},i}, \mid A_i, X_{R_{Xi},i}, Y_i, R_i, R_{Yi} = 1; \theta^{*(j)})$; for unit $i$ with $R_Y = 0$, generate $X^{*(j)}_{\overline{R}_{Xi},i}$ and $Y_i^{*(j)}$ from $f(X_{\overline{R}_{Xi},i}, Y_i \mid A_i, X_{R_{Xi},i}, R_{Xi}, R_{Yi} = 0; \theta^{*(j)})$ to create the $j$th imputed data set. Then, the MI estimator can be written in a general form with both imputed outcome and confounders as

$$\hat{\tau}_{\text{MI}} - \tau = \frac{1}{nm} \sum_{i=1}^{n} \sum_{j=1}^{m} \psi(A_i, X_i^{*(j)}, Y_i^{*(j)}) + o_{\mathbb{P}}(1).$$

Accordingly, the martingale difference arrays in the wild bootstrap procedure can be written as

$$\hat{\xi}_{n,k} = \begin{cases} \dfrac{1}{n^{1/2}}[\mathbb{E}\{\psi(A_i, X_i, Y_i) \mid \mathbf{Z}_{\text{obs}}, \hat{\theta}\} + \hat{\Gamma}\hat{\mathcal{I}}_{\text{obs}}^{-1}\bar{S}(\hat{\theta}; Z_{\text{obs},i})], & \text{if } k = i, \\ \dfrac{1}{n^{1/2}m}[\psi(A_i, X_i^{*(j)}, Y_i^{*(j)}) - \mathbb{E}\{\psi(A_i, X_i, Y_i) \mid \mathbf{Z}_{\text{obs}}, \hat{\theta}\}], & \text{if } k = n + (i-1)m + j, \end{cases}$$

where $i = 1, \ldots, n$ and $j = 1, \ldots, m$. The other steps in the MI and wild bootstrap procedures remain the same, as when only the confounders have missing values.

**Corollary 3.** *For the scenario where both the outcome and the confounders are subject to missingness, the assumption that the imputation model* $f(L_{\overline{R}_i,i} \mid Z_{\text{obs},i}; \theta)$ *is specified correctly in Theorem* 1 *implies Corollary* 1 *under MAR and Corollary* 2 *under MNAR.*

## 6. Simulation Study

We conduct simulation studies to evaluate the finite-sample performance of the proposed inference when MI adopts different full-sample estimators, including the outcome regression, IPW, AIPW, and matching estimators.

For each sample, the confounders $X = (X_{[1]}, X_{[2]})$ are sampled from a multivariate normal distribution with mean $(0, 0)$, variance $(1, 1)$, and correlation coefficient 0.2. The potential outcomes follow $Y(0) = 2 + 3X_{[1]} + 2X_{[2]} + \epsilon(0)$ and $Y(1) = 1 + 2X_{[1]} + X_{[2]} + \epsilon(1)$, where $\epsilon(0) \sim \mathcal{N}(0, \sigma_0^2)$ and $\epsilon(1) \sim \mathcal{N}(0, \sigma_1^2)$, with $\sigma_0 = \sigma_1 = 1$, and $\epsilon(0)$ and $\epsilon(1)$ are independent. Thus the true value of the ACE is $\tau = -1$. We generate the treatment indicator $A$ from Bernoulli$\{\pi_A(X)\}$ and $\pi_A(X) = P(A = 1 \mid X) = \Phi(-0.2 + 0.3X_{[1]} + 0.4X_{[2]})$, where $\Phi(\cdot)$ is the cumulative density function for the standard normal distribution. In the sample, we assume $A$ and $X_{[1]}$ are fully observed, but $X_{[2]}$ and $Y$ can be partially observed, with missing indicators $R_{[2]}$ and $R_Y$, respectively. We consider four scenarios:

(a) $X_{[2]}$ is MAR; that is, its missingness depends only on the observed data. Let $R_{[2]} \sim$ Bernoulli$\{\pi_{R1}(A, X_{[1]}, Y)\}$, where $\pi_{R1}(A, X_{[1]}, Y) = \Phi(-0.1 + 0.1A + 0.5X_{[1]} + 0.2Y)$, with the missingness rate being about 45%. Moreover, the inference procedure assumes a correct missingness mechanism;

(b) $X_{[2]}$ is MNAR; that is, its missingness depends on unobserved data. Let $R_{[2]} \sim$ Bernoulli$\{\pi_{R2}(A, X_{[1]}, X_{[2]})\}$, where $\pi_{R2}(A, X_{[1]}, X_{[2]}) = \Phi(0.2 + 1X_{[2]})$, with the missingness rate being about 45%. Moreover, the inference procedure assumes a correct missingness mechanism;

(c) $X_{[2]}$ is MNAR as in scenario (b), but the inference procedure assumes an incorrect missingness mechanism;

(d) both $X_{[2]}$ and $Y$ are MNAR, with the missingness indicators $R_{[2]}$ and $R_Y$, respectively. Let $R_{[2]} \sim$ Bernoulli$\{\pi_R(X_{[2]})\}$, where $\pi_R(X_{[2]}) = \Phi(0.8 + 1X_{[2]})$, with the missingness rate being about 30%. Let $R_Y \sim$

Bernoulli$\{\pi_Y(A,X)\}$, where $\pi_Y(A,X) = \Phi(1 + 0.2A + 0.5X_{[1]} + 0.5X_{[2]})$, with the missingness rate being about 20%.

We generate 5,000 Monte Carlo samples with size $n = 3000$ for each scenario. In MI, the missing-data mechanism is specified according to the above scenarios and other components of the distribution are specified correctly. We use noninformative priors for the parameters. Suppose that the prior distribution for each coefficient in the outcome model, the propensity score model, and missing indicator model is $\mathcal{N}(0, 100)$; the prior distribution for the variance parameters $\sigma_0$ and $\sigma_1$ in the outcome regression model is Gamma$(0.01, 0.01)$; the prior distribution for the mean of $X$ is $(0, 0)$; the prior distribution for the variance covariance matrix of X is $I_2$, where $I_2$ is the two-dimensional identity matrix. Further information on the priors and posterior sampling are provided in the Supplementary Material. We consider three sizes of MI with $m = 5, 10,$ or 100. To generate the posterior samples of the missing values $X_{\overline{R}}^{*(j)}$, we use Gibbs sampling with 5,000 iterations, discard the first 2,000 burn-in samples, and randomly choose $m$ posterior samples from the remaining 3,000 draws. For each imputed data set, we calculate the full-sample point estimators and variance estimators of the ACE using an outcome regression, IPW, AIPW, and matching, and then use Rubin's method to obtain the corresponding MI estimators $\hat{\tau}_{\mathrm{MI}}$ and Rubin's variance estimators $\hat{V}_{\mathrm{MI}}$. For the matching estimator, we set the number of matches as $M = 1$.

We compare the standard MI inference and the proposed bootstrap inference. For the standard MI inference, the $100(1-\alpha)\%$ CIs are calculated as $(\hat{\tau}_{\mathrm{MI}} - t_{\nu,1-\alpha/2}\hat{V}_{\mathrm{MI}}^{1/2}, \hat{\tau}_{\mathrm{MI}} + t_{\nu,1-\alpha/2}\hat{V}_{\mathrm{MI}}^{1/2})$, where $t_{\nu,1-\alpha/2}$ is the $100(1-\alpha/2)\%$ quantile of the $t$ distribution with degrees of freedom $\nu = (m-1)\lambda^{-2}$, with $\lambda = (1+m^{-1})B_m/\{W_m + (1+m^{-1}B_m)\}$. For the proposed bootstrap procedure, we use $B = 1000$, generate the weights $\mu_k$ from Mammen's two-point distribution, as suggested in Remark 2, and calculate the variance estimate $\hat{V}_{\mathrm{BS}}$. The corresponding $100(1-\alpha)\%$ CI is estimated using two methods: (i) a quantile-based CI $(\hat{\tau}_{\mathrm{MI}} - q^*_{1-\alpha/2}, \hat{\tau}_{\mathrm{MI}} - q^*_{\alpha/2})$, where $q^*_{1-\alpha/2}$ and $q^*_{\alpha/2}$ are the $(1-\alpha/2)$th and $(\alpha/2)$th quantiles, respectively, of $T^*$; and (ii) the Wald-type CI $(\hat{\tau}_{\mathrm{MI}} - z_{1-\alpha/2}\hat{V}_{\mathrm{BS}}^{1/2}, \hat{\tau}_{\mathrm{MI}} + z_{1-\alpha/2}\hat{V}_{\mathrm{BS}}^{1/2})$, where $z_{1-\alpha/2}$ is the $(1-\alpha/2)$th quantile of the standard normal distribution.

We assess the performance in terms of the relative bias of the variance estimator and the coverage rate of CIs. The relative bias of the variance estimators is calculated as $\{\mathbb{E}(\hat{V}_{\mathrm{MI}}) - \mathbb{V}(\hat{\tau}_{\mathrm{MI}})\}/\mathbb{V}(\hat{\tau}_{\mathrm{MI}}) \times 100\%$ and $\{\mathbb{E}(\hat{V}_{\mathrm{BS}}) - \mathbb{V}(\hat{\tau}_{\mathrm{MI}})\}/\mathbb{V}(\hat{\tau}_{\mathrm{MI}}) \times 100\%$. The coverage rate of the $100(1-\alpha)\%$ CI is estimated by the percentage of Monte Carlo samples for which the CIs contain the true value.

Tables 2–5 present the simulation results for the four scenarios. When the imputation model is specified correctly, as in scenarios (a), (b), and (d), the MI point estimator has small biases for all full-sample estimators. In addition, as

Table 2. Simulation results: point estimate (Monte Carlo mean of point estimates), true variance (Monte Carlo variance of point estimates), relative bias of the variance estimator, and the coverage and mean width of the interval estimate using Rubin's method and the proposed wild bootstrap method under scenario (a), with missingness at random.

| Method $\hat{\tau}_n$ | $m$ | Point est ($\times 10$) | True var ($\times 10^4$) | Relative Bias (%) | | Coverage (%) for 95% CI | | | Mean width ($\times 10^2$) for 95% CI | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | Rubin | BS | Rubin | BS | | Rubin | BS | |
| | | | | | | | Quantile | Wald | | Quantile | Wald |
| | 5 | -10.0 | 35.8 | -2.1 | 1.9 | 94.3 | 94.9 | 95.4 | 23.9 | 23.6 | 24.1 |
| Regression | 10 | -10.0 | 34.9 | -1.9 | 3.7 | 94.6 | 95.3 | 95.8 | 23.1 | 23.6 | 24.0 |
| | 100 | -10.0 | 33.8 | -1.4 | 5.6 | 94.8 | 95.6 | 95.9 | 22.6 | 23.4 | 23.9 |
| | 5 | -10.0 | 68.0 | **25.8** | **-0.3** | 96.0 | 93.9 | 94.7 | 35.6 | 31.1 | 31.9 |
| IPW | 10 | -10.0 | 66.3 | **27.4** | **0.3** | 96.3 | 94.2 | 94.6 | 34.9 | 30.8 | 31.6 |
| | 100 | -10.0 | 64.4 | **29.7** | **1.2** | 96.3 | 94.2 | 94.7 | 34.4 | 30.4 | 31.3 |
| | 5 | -10.0 | 36.6 | 3.0 | -3.9 | 95.2 | 94.4 | 94.9 | 24.8 | 23.2 | 23.7 |
| AIPW | 10 | -10.0 | 35.7 | 3.0 | -2.7 | 94.9 | 94.5 | 95.0 | 24.0 | 23.1 | 23.5 |
| | 100 | -10.0 | 34.6 | 3.7 | -1.1 | 95.3 | 94.7 | 95.3 | 23.5 | 22.9 | 23.4 |
| | 5 | -10.0 | 39.1 | **18.2** | **-4.5** | 96.5 | 94.4 | 95.0 | 27.5 | 23.9 | 24.4 |
| Matching | 10 | -10.0 | 37.8 | **18.7** | **-3.5** | 96.5 | 94.5 | 95.1 | 26.6 | 23.7 | 24.2 |
| | 100 | -10.0 | 36.4 | **20.1** | **-2.1** | 96.9 | 94.4 | 95.0 | 26.0 | 23.4 | 23.9 |

Table 3. Simulation results under scenario (b), with missingness not at random.

| Method $\hat{\tau}_n$ | $m$ | Point est ($\times 10$) | True var ($\times 10^4$) | Relative Bias (%) | | Coverage (%) for 95% CI | | | Mean width ($\times 10^2$) for 95% CI | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | Rubin | BS | Rubin | BS | | Rubin | BS | |
| | | | | | | | Quantile | Wald | | Quantile | Wald |
| | 5 | -10.0 | 34.5 | -0.5 | 2.8 | 94.6 | 95.2 | 95.7 | 23.6 | 23.3 | 23.8 |
| Regression | 10 | -10.0 | 33.6 | 0.9 | 4.4 | 94.8 | 95.4 | 95.7 | 22.9 | 23.2 | 23.7 |
| | 100 | -10.0 | 32.9 | -0.1 | 5.6 | 94.8 | 95.5 | 96.0 | 22.5 | 23.1 | 23.6 |
| | 5 | -10.0 | 67.5 | **28.0** | **0.3** | 96.4 | 94.5 | 94.8 | 35.7 | 30.9 | 31.7 |
| IPW | 10 | -10.0 | 65.6 | **30.6** | **1.3** | 96.7 | 94.6 | 95.0 | 35.0 | 30.6 | 31.4 |
| | 100 | -10.0 | 64.2 | **29.8** | **1.4** | 96.7 | 94.7 | 95.0 | 34.5 | 30.4 | 31.2 |
| | 5 | -10.0 | 35.5 | 5.0 | -2.3 | 95.2 | 94.8 | 95.2 | 24.6 | 23.1 | 23.5 |
| AIPW | 10 | -10.0 | 34.5 | 5.6 | -0.7 | 95.5 | 94.9 | 95.5 | 23.9 | 22.9 | 23.4 |
| | 100 | -10.0 | 33.6 | 5.7 | -0.5 | 95.5 | 95.1 | 95.4 | 23.4 | 22.8 | 23.2 |
| | 5 | -10.0 | 38.0 | **21.0** | **-3.5** | 96.9 | 94.8 | 95.4 | 27.5 | 23.7 | 24.2 |
| Matching | 10 | -10.0 | 36.7 | **21.8** | **-2.1** | 96.9 | 95.0 | 95.5 | 26.5 | 23.5 | 24.0 |
| | 100 | -10.0 | 35.6 | **22.4** | **-1.1** | 97.0 | 94.9 | 95.3 | 25.9 | 23.2 | 23.7 |

$m$ increases, the variance of the MI point estimator becomes smaller, suggesting that using more imputations can improve the efficiency of the MI estimator. Across different choices of m, the relative bias of the proposed variance estimator stays small, and the accuracy of the estimator is less sensitive to the choice of m. Rubin's variance estimator is unbiased for the outcome regression estimator and the AIPW estimator; however, it overestimates the variances of the IPW

Table 4. Simulation results under scenario (c) when the true missing mechanism is MNAR but MAR is assumed.

| Method $\hat{\tau}_n$ | $m$ | Point est ($\times 10$) | True var ($\times 10^4$) | Relative Bias (%) | | Coverage (%) for 95% CI | | | Mean width ($\times 10^2$) for 95% CI | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | Rubin | BS | Rubin | BS | | Rubin | BS | |
| | | | | | | | Quantile | Wald | | Quantile | Wald |
| Regression | 5 | -11.5 | 34.6 | 1.7 | 10.9 | 27.2 | 29.1 | 30.2 | 23.7 | 24.3 | 24.8 |
| | 10 | -11.5 | 33.8 | 1.8 | 12.3 | 25.6 | 24.1 | 24.6 | 23.2 | 24.1 | 24.6 |
| | 100 | -11.5 | 33.2 | 1.4 | 13.0 | 23.9 | 27.9 | 28.9 | 22.8 | 24.0 | 24.5 |
| IPW | 5 | -12.0 | 130.1 | **31.5** | **1.1** | 66.1 | 54.5 | 53.7 | 46.3 | 39.1 | 40.5 |
| | 10 | -12.0 | 127.8 | **31.3** | **-1.4** | 64.9 | 53.1 | 51.9 | 45.6 | 38.6 | 40.0 |
| | 100 | -12.0 | 126.4 | **33.3** | **-1.8** | 64.7 | 52.1 | 50.9 | 45.4 | 38.2 | 39.6 |
| AIPW | 5 | -11.5 | 36.3 | 6.0 | -0.7 | 31.0 | 27.5 | 28.6 | 24.7 | 23.5 | 24.0 |
| | 10 | -11.5 | 35.5 | 5.8 | 0.2 | 29.0 | 26.5 | 27.8 | 24.1 | 23.3 | 23.8 |
| | 100 | -11.5 | 34.9 | 5.5 | 0.5 | 27.6 | 26.3 | 27.4 | 23.8 | 23.2 | 23.7 |
| Matching | 5 | -11.6 | 38.7 | **26.2** | **-1.3** | 40.9 | 29.4 | 30.8 | 28.1 | 24.2 | 24.7 |
| | 10 | -11.6 | 37.5 | **26.6** | **-0.5** | 38.4 | 27.8 | 29.1 | 27.3 | 23.9 | 24.4 |
| | 100 | -11.6 | 36.6 | **26.7** | **-0.2** | 36.5 | 27.2 | 28.6 | 26.7 | 23.6 | 24.1 |

estimator and the matching estimator, for example, by as much as 29.7% and 20.1% in scenario (a). Because of the variance overestimation, the coverage rate of Rubin's method exceeds the nominal level for the IPW and matching estimators, all exceeding 96%, and some reaching 97.3%. In contrast, our proposed wild bootstrap procedure for variance estimation is unbiased for all four ACE estimators, and therefore the coverage rate of the CIs based on this method is close to the nominal level. Moreover, the proposed method is not sensitive to the number of imputations $m$ or the choice of a quantile-based or Wald-type CI. However, in scenario (c), when the true missing-data mechanism is MNAR, but the inference procedure assumes MAR, the MI point estimator has large biases and all the CIs have poor coverage rates; see Table 4.

Other methods have been developed for MI inference. For example, Xie and Meng (2017) propose a doubling variance approach for a more conservative variance estimation when Rubin's method underestimates the variance. However, it further overestimates the variance of the MI estimators in our simulation settings, such that the performance is even worse than that of Rubin's method. Meng and Rubin (1992) and Chan and Meng (2022) propose likelihood ratio-based procedures for multiply imputed data inference. However, these procedures are not easily implemented for the variance and CI construction for the treatment effect estimation.

## 7. An Application

We apply our method to a data set from the 2015–2016 U.S. National Health and Nutrition Examination Survey to estimate the ACE of education on general

Table 5. Simulation results under scenario (d), where both the outcome and the confounders are missing and MNAR is assumed.

| Method $\hat{\tau}_n$ | $m$ | Point est ($\times 10$) | True var ($\times 10^4$) | Relative Bias (%) | | Coverage (%) for 95% CI | | | Mean width ($\times 10^2$) for 95% CI | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | Rubin | BS | Rubin | BS | | Rubin | BS | |
| | | | | | | | Quantile | Wald | | Quantile | Wald |
| Regression | 5 | -10.0 | 35.6 | -2.4 | -1.5 | 94.6 | 94.7 | 95.2 | 23.7 | 23.2 | 23.7 |
| | 10 | -10.0 | 34.3 | -0.9 | 0.7 | 94.9 | 95.0 | 95.7 | 23.1 | 23.0 | 23.5 |
| | 100 | -10.0 | 33.4 | -0.5 | 2.2 | 95.0 | 95.3 | 95.7 | 22.6 | 22.9 | 23.4 |
| IPW | 5 | -10.0 | 68.5 | **28.6** | **-2.7** | 96.3 | 94.2 | 94.7 | 36.6 | 30.8 | 31.6 |
| | 10 | -10.0 | 65.9 | **32.7** | **-0.8** | 96.7 | 94.5 | 94.9 | 35.6 | 30.4 | 31.3 |
| | 100 | -10.0 | 64.0 | **34.3** | **-0.2** | 97.3 | 94.5 | 95.1 | 35.2 | 30.1 | 30.9 |
| AIPW | 5 | -10.0 | 36.5 | 7.3 | -3.9 | 95.5 | 94.4 | 94.9 | 25.4 | 23.2 | 23.7 |
| | 10 | -10.0 | 34.9 | 9.7 | -1.3 | 96.1 | 94.6 | 95.4 | 24.5 | 23.0 | 23.5 |
| | 100 | -10.0 | 33.8 | 10.2 | 0.1 | 96.1 | 94.9 | 95.3 | 23.9 | 22.8 | 23.3 |
| Matching | 5 | -10.0 | 39.5 | **18.5** | **-4.7** | 96.6 | 94.1 | 94.6 | 27.8 | 24.0 | 24.5 |
| | 10 | -10.0 | 37.7 | **21.4** | **-2.6** | 97.1 | 94.5 | 95.0 | 26.8 | 23.7 | 24.2 |
| | 100 | -10.0 | 36.5 | **22.1** | **-1.5** | 97.2 | 94.8 | 95.6 | 26.2 | 23.5 | 24.0 |

health satisfaction. The general health satisfaction outcome ($Y$) is fully observed, with a lower value indicating better satisfaction. A sample of 4,845 individuals is divided into two groups: one (76%) with at least high school education, denoted as $A = 1$, and the other (24%) with an education level lower than high school, denoted as $A = 0$. The covariates $X$ consist of four categorical variables, namely, age, race, gender, and marital status, and one continuous variable, namely, the family poverty ratio, which is truncated at zero and five. About 10% of the family poverty ratio values are missing. The other four covariates are fully observed.

The general health satisfaction outcome ($Y$) is an ordinal variable, with distinct values $1, 2, 3, 4, 5$. We introduce a latent continuous variable $Y^*$ to link the ordinal outcome to the continuous space with support $(-\infty, +\infty)$:

$$Y = \begin{cases} 1 & \text{if } Y^* < 1, \\ [Y^*] & \text{if } 1 \le Y^* \le 5, \\ 5 & \text{if } Y^* > 5, \end{cases}$$

where $[\cdot]$ represents rounding to the nearest integer. Because the family poverty ratio $X_{[1]}$ is a continuous variable truncated at zero and five, we introduce another latent variable $X^*_{[1]}$ to link the recorded truncated family poverty ratio values to the full continuous space $(-\infty, +\infty)$:

$$X_{[1]} = \begin{cases} 0 & \text{if } X^*_{[1]} < 0, \\ X^*_{[1]} & \text{if } 0 \le X^*_{[1]} \le 5, \\ 5 & \text{if } X^*_{[1]} > 5. \end{cases}$$

Table 6. Result for the ACE of education on general health satisfaction: point estimates, variance of the point estimators, and 95% CI estimated using Rubin's method and proposed wild bootstrap method.

| Method | Point est | Rubin | | BS | |
|---|---|---|---|---|---|
| | | Var est ($\times 10^4$) | 95% CI | Var est ($\times 10^4$) | 95% CI Wald |
| Regression | -0.36 | 19 | (-0.45,-0.27) | 19 | (-0.45,-0.27) |
| IPW | -0.25 | 65 | (-0.41,-0.10) | 54 | (-0.40,-0.11) |
| AIPW | -0.27 | 32 | (-0.38,-0.16) | 31 | (-0.38,-0.16) |
| Matching | -0.25 | 40 | (-0.37,-0.12) | 28 | (-0.35,-0.14) |

Accordingly, let $X^*$ include the latent family poverty ratio variable $X^*_{[1]}$ and the other four variables. To facilitate imputation and estimation, we assume the latent outcome $Y^*$ follows a linear regression model, that is, $Y^*(a) = X^{*\mathrm{T}}\beta_a + \epsilon(a)$, where $\epsilon(a) \sim \mathcal{N}(0, \sigma_a^2)$, for $a = 0, 1$. The treatment indicator follows Bernoulli$\{\pi_A(X^*)\}$, with $\pi_A(X^*) = \Phi(X^{*\mathrm{T}}\alpha)$. The missing indicator follows Bernoulli$\{\pi_R(X^*, A)\}$, with $\pi_R(X^*, A) = \Phi\{(X^*, A)^{\mathrm{T}}\gamma\}$, under which the missingness of the family poverty ratio probably depends on the missing values themselves, but not the outcome variable (i.e., Assumption 8). In addition, we assume the latent family poverty ratio follows a linear regression model with the other covariates, that is, $X^*_{\overline{R}_X} = X_{R_X}\eta + \epsilon_X$, where $X^*_{\overline{R}_X} = X^*_{[1]}$ represents the latent family poverty ratio, and $X_R$ represents the other four covariates, $\epsilon_X \sim \mathcal{N}(0, \sigma_X^2)$. We conduct model diagnoses in the Supplementary Material and the diagnosis plots show that the proposed model fits the data well. Given the outcome model and the covariate model, the missing values of the family poverty ratio can be imputed by $f(X^*_{\overline{R}_X} \mid A, X_{R_X}, Y, R_X; \theta^{*(j)}) \propto f(Y^* \mid X^*, A; \theta^{*(j)}) f(R_X \mid X^*, A; \theta^{*(j)}) f(A \mid X^*; \theta^{*(j)}) f(X^*_{\overline{R}_X} \mid X_{R_X}; \theta^{*(j)})$ given each posterior sample of the parameters $\theta^{*(j)}$. Further details about the priors and the posterior sampling are provided in the Supplementary Material.

For each imputed data set, we consider the full-sample point estimators of the ACE using an outcome regression, IPW, AIPW, and matching based on the propensity score to reduce the dimensionality of the matching variable (Abadie and Imbens (2016)). We compare Rubin's variance estimator and the proposed wild bootstrap variance estimator. Table 6 shows that education has a significantly positive effect on general health satisfaction. The variances for the IPW estimator and matching estimator estimated using Rubin's method are larger than those estimated using the wild bootstrap method, whereas the two methods give similar results for the regression estimator and the AIPW estimator. This suggests that Rubin's method works well for the regression estimator and the AIPW estimator, but might overestimate the variances of the IPW and matching estimators, which is consistent with our observations in the simulation studies.

## 8. Conclusion

We have established a unified inference framework for MI using a martingale and a wild bootstrap inference for consistent variance estimation. Our framework allows a wide class of asymptotically linear full-sample estimators. We demonstrate its utility in estimating the ACE with missing values. The simulation results indicate good finite-sample performance of the proposed method when MI adopts different full-sample estimators, including the outcome regression, IPW, AIPW, and matching estimators. Our framework works well when the missing mechanism is either MAR or MNAR.

Our framework can also be extended in the following directions. First, MI originated for survey data, which often contain design weights (or sample weights) to account for sample selection. If the sampling weights are noninformative, the sample data follow the population model, and therefore the imputation can ignore the sampling weights. However, if the sampling weights are informative, then the sample data distribution differs from that of the population model, in which case, the imputation must consider the sampling weights. The full Bayesian imputation is difficult (if not impossible) to implement in this case. To mitigate this problem, Kim and Yang (2017) and Wang, Kim and Yang (2018) propose an approximate Bayesian computation techniques that can be used for MI in complex sampling. It would be interesting to extend the martingale representation to this setting in future work. Second, in the current work, we assume that the imputer's model and the analyst's model are the same and are specified correctly. Xie and Meng (2017) argue that the uncongeniality of the imputer's model and the analyst's model is the rule, but not an exception. Their findings suggest that even when both models are specified correctly, if the imputation model is more saturated than the analysis model, then the standard MI inference may be invalid. In future work, we will extend our framework to this setting for consistent inference allowing uncongeniality.

## Supplementary Material

The online Supplementary Material contains common ACE estimators and their influence functions, proofs, the priors and MCMC details for the simulation study and application, and the model diagnosis in the application. The R code to implement the proposed method is available at `https://github.com/qianguan/miATE`.

## Acknowledgments

Yang's research was partially supported by the NSF DMS 1811245, NSF SES 2242776, NIH 1R01AG066883, and 1R01ES031651.

## References


Abadie, A. and Imbens, G. W. (2008). On the failure of the bootstrap for matching estimators. *Econometrica* **76**, 1537–1557.

Abadie, A. and Imbens, G. W. (2016). Matching on the estimated propensity score. *Econometrica* **84**, 781–807.

Bang, H. and Robins, J. M. (2005). Doubly robust estimation in missing data and causal inference models. *Biometrics* **61**, 962–973.

Beyersmann, J., Termini, S. D. and Pauly, M. (2013). Weak convergence of the wild bootstrap for the Aalen–Johansen estimator of the cumulative incidence function of a competing risk. *Scandinavian Journal of Statistics* **40**, 387–402.

Bickel, P. J., Klaassen, C., Ritov, Y. and Wellner, J. (1993). *Efficient and Adaptive Inference in Semiparametric Models*. Johns Hopkins University Press, Baltimore.

Binder, D. A. and Sun, W. (1996). Frequency valid multiple imputation for surveys with a complex design. In *Proceedings of the Survey Research Methods Section, ASA (1996)*, 281–286. American Statistical Association, Alexandria.

Cao, W., Tsiatis, A. A. and Davidian, M. (2009). Improving efficiency and robustness of the doubly robust estimator for a population mean with incomplete data. *Biometrika* **96**, 723–734.

Chan, K. W. and Meng, X.-L. (2022). Multiple improvements of multiple imputation likelihood ratio tests. *Statist. Sinica* **32**, 1489–1514.

Clogg, C. C., Rubin, D. B., Schenker, N., Schultz, B. and Weidman, L. (1991). Multiple imputation of industry and occupation codes in census public-use samples using Bayesian logistic regression. *J. Amer. Statist. Assoc.* **86**, 68–78.

Crowe, B. J., Lipkovich, I. A. and Wang, O. (2010). Comparison of several imputation methods for missing baseline data in propensity scores analysis of binary outcome. *Pharmaceut. Statist.* **9**, 269–279.

Crump, R., Hotz, V. J., Imbens, G. and Mitnik, O. (2006). Moving the goalposts: Addressing limited overlap in the estimation of average treatment effects by changing the estimand. Technical report. National Bureau of Economic Research, Cambridge.

Davern, M., Rodin, H., Beebe, T. J. and Call, K. T. (2005). The effect of income question design in health surveys on family income, poverty and eligibility estimates. *Health Services Research* **40**, 1534–1552.

Fay, R. E. (1992). When are inferences from multiple imputation valid. In *Proceedings of the Survey Research Methods Section, ASA (1992)*, 227–32.

Fay, R. E. (1996). Alternative paradigms for the analysis of imputed survey data. *J. Amer. Statist. Assoc.* **91**, 490–498.

Hahn, J. (1998). On the role of the propensity score in efficient semiparametric estimation of average treatment effects. *Econometrica* **66**, 315–331.

Hall, P. and Heyde, C. (1980). *Martingale Limit Theory and Its Application*. Academic Press, Boston.

Heckman, J. J., Ichimura, H. and Todd, P. E. (1997). Matching as an econometric evaluation estimator: Evidence from evaluating a job training programme. *Rev. Econ. Stud.* **64**, 605–654.

Horvitz, D. G. and Thompson, D. J. (1952). A generalization of sampling without replacement from a finite universe. *J. Amer. Statist. Assoc.* **47**, 663–685.

Imbens, G. W. (2004). Nonparametric estimation of average treatment effects under exogeneity: A review. *Rev. Econ. Stat.* **86**, 4–29.

Imbens, G. W. and Rubin, D. B. (2015). *Causal Inference in Statistics, Social, and Biomedical Sciences*. Cambridge University Press, Cambridge.

Kennedy, E. H. (2016). Semiparametric theory and empirical processes in causal inference. In *Statistical Causal Inferences and Their Applications in Public Health Research*, 141–167. Springer.

Kim, J. K., Brick, J., Fuller, W. A. and Kalton, G. (2006). On the bias of the multiple-imputation variance estimator in survey sampling. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **68**, 509–521.

Kim, J. K. and Yang, S. (2017). A note on multiple imputation under complex sampling. *Biometrika* **104**, 221–228.

Kott, P. (1995). A paradox of multiple imputation. In *Proceedings of the Survey Research Methods Section, ASA (1995)*, 384–389.

Li, F., Morgan, K. L. and Zaslavsky, A. M. (2018). Balancing covariates via propensity score weighting. *J. Amer. Statist. Assoc.* **113**, 390–400.

Liu, R. Y. (1988). Bootstrap procedures under some non-iid models. *Ann. Statist.* **16**, 1696–1708.

Mammen, E. (1993). Bootstrap and wild bootstrap for high dimensional linear models. *Ann. Statist.* **21**, 255–285.

Meng, X.-L. (1994). Multiple-imputation inferences with uncongenial sources of input. *Statist. Sci.* **9**, 538–558.

Meng, X.-L. and Rubin, D. B. (1992). Performing likelihood ratio tests with multiply-imputed data sets. *Biometrika* **79**, 103–111.

Miao, W., Ding, P. and Geng, Z. (2016). Identifiability of normal and normal mixture models with nonignorable missing data. *J. Amer. Statist. Assoc.* **111**, 1673–1683.

Mitra, R. and Reiter, J. P. (2011). Estimating propensity scores with missing covariate data using general location mixture models. *Stat. Med.* **30**, 627–641.

National Research Council (2010). *Panel on Handling Missing Data in Clinical Trials. The Prevention and Treatment of Missing Data in Clinical Trials*. National Academies Press, Washington (DC).

Newey, W. K. (1990). Semiparametric efficiency bounds. *Journal of Applied Econometrics* **5**, 99–135.

Neyman, J. (1923). Sur les applications de la thar des probabilities aux experiences Agaricales: Essay de principle (English translation of excerpts by D. Dabrowska and T. Speed). *Statist. Sci.* **5**, 465–472.

Nielsen, S. F. (2003). Proper and improper multiple imputation. *Int. Stat. Rev.* **71**, 593–607.

Pauly, M. (2011). Weighted resampling of martingale difference arrays with applications. *Electron. J. Stat.* **5**, 41–52.

Qu, Y. and Lipkovich, I. (2009). Propensity score estimation with missing values using a multiple imputation missingness pattern (MIMP) approach. *Stat. Med.* **28**, 1402–1414.

Robins, J. M., Rotnitzky, A. and Zhao, L. P. (1994). Estimation of regression coefficients when some regressors are not always observed. *J. Amer. Statist. Assoc.* **89**, 846–866.

Robins, J. M. and Wang, N. (2000). Inference for imputation estimators. *Biometrika* **87**, 113–124.

Rosenbaum, P. R. (1989). Optimal matching for observational studies. *J. Amer. Statist. Assoc.* **84**, 1024–1032.

Rosenbaum, P. R. (2002). *Observational Studies*. 2nd Edition. Springer, New York.

Rosenbaum, P. R. and Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika* **70**, 41–55.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized

studies. *J. Educ. Psychol.* **66**, 688–701.

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

Rubin, D. B. (1980). Comment on "Randomization analysis of experimental data: The Fisher randomization test" by D. Basu. *J. Amer. Statist. Assoc.* **75**, 591–593.

Rubin, D. B. (1987). *Multiple Imputation for Nonresponse in Surveys*. Wiley, New York.

Schomaker, M. and Heumann, C. (2018). Bootstrap inference when using multiple imputation. *Stat. Med.* **37**, 2252–2266.

Seaman, S. and White, I. (2014). Inverse probability weighting with missing predictors of treatment assignment or missingness. *Comm. Statist. Theory Methods* **43**, 3499–3515.

Stuart, E. A. (2010). Matching methods for causal inference: A review and a look forward. *Statist. Sci.* **25**, 1–21.

van der Vaart, A. W. (2000). *Asymptotic Statistics*. Cambridge University Press, Cambridge.

Wang, N. and Robins, J. M. (1998). Large-sample theory for parametric multiple imputation procedures. *Biometrika* **85**, 935–948.

Wang, S., Shao, J. and Kim, J. K. (2014). An instrumental variable approach for identification and estimation with nonignorable nonresponse. *Statist. Sinica* **24**, 1097–1116.

Wang, Z., Kim, J. and Yang, S. (2018). Approximate Bayesian inference under informative sampling. *Biometrika* **105**, 91–102.

Wu, C.-F. J. (1986). Jackknife, bootstrap and other resampling methods in regression analysis. *Ann. Statist.* **14**, 1261–1295.

Xie, X. and Meng, X.-L. (2017). Dissecting multiple imputation from a multi-phase inference perspective: What happens when God's, imputer's and analyst's models are uncongenial? *Statist. Sinica* **27**, 1485–1545.

Yang, S. and Kim, J. K. (2016). A note on multiple imputation for method of moments estimation. *Biometrika* **103**, 244–251.

Yang, S., Wang, L. and Ding, P. (2019). Causal inference with confounders missing not at random. *Biometrika* **106**, 875–888.

Qian Guan

Department of Statistics, North Carolina State University, Raleigh, NC 27607, USA.

E-mail: guanqian913@gmail.com

Shu Yang

Department of Statistics, North Carolina State University, Raleigh, NC 27607, USA.

E-mail: syang24@ncsu.edu

(Received November 2021; accepted November 2022)