A Multiplicative Instrumental Variable Model for Data Missing Not-at-Random

Yunshu Zhang, Chan Park, Jiewen Liu, Yonghoon Lee, Mengxin Yu,
James M. Robins, Eric J. Tchetgen Tchetgen

September 29, 2025

Abstract

Instrumental variable (IV) methods offer a valuable approach to account for outcome data missing not-at-random. A valid missing data instrument is a measured factor which (i) predicts the nonresponse process and (ii) is independent of the outcome in the underlying population. For point identification, all existing IV methods for missing data including the celebrated Heckman selection model, a priori restrict the extent of selection bias on the outcome scale, therefore potentially understating uncertainty due to missing data. In this work, we introduce an IV framework which allows the degree of selection bias on the outcome scale to remain completely unrestricted. The new approach instead relies for identification on (iii) a key multiplicative selection model, which posits that the instrument and any hidden common correlate of selection and the outcome, do not interact on the multiplicative scale. Interestingly, we establish that any regular statistical functional of the missing outcome is nonparametrically identified under (i)-(iii) via a *single-arm Wald ratio estimand* reminiscent of the standard Wald ratio estimand in causal inference. For estimation and inference, we characterize the influence function for any functional defined on a nonparametric model for the observed data, which we leverage to develop semiparametric multiply robust IV estimators. Several extensions of the methods are also considered, including the important practical setting of polytomous and continuous instruments. Simulation studies illustrate the favorable finite sample performance of proposed methods, which we further showcase in an HIV study nested within a household health survey study we conducted in Mochudi, Botswana, in which interviewer characteristics are used as instruments to correct for selection bias due to dependent nonresponse in the HIV component of the survey study.

1 Introduction

Missing data is a common and critical issue that impacts most empirical studies conducted in the health and social sciences. Failure to address selection bias induced by missing data can lead to biased and inconsistent results. Rubin (1976) was an early proponent for formal statistical methods for addressing missing data, including introducing language for classifying missing data mechanisms into three broad categories: missing completely at random (MCAR), missing at random (MAR), and missing not at random (MNAR). In a setting where the sampled data consists of an outcome variable of primary interest which is potentially missing, and a set of fully observed baseline covariates, MCAR assumes that the missingness mechanism is completely independent of both the outcome and covariates, in which case a standard complete-case approach to handle missing data, which restricts analyses, e.g. a regression model of the outcome on covariates, to the subset of observations with fully observed data is guaranteed to be unbiased, although potentially inefficient. In contrast, under MAR, one assumes that conditional on fully observed covariates, the missingness mechanism is independent of the outcome variable. Under this assumption, most well-defined statistical functionals of the the full data are identifiable from the observed data using methods such as complete-case inverse probability weighting, direct maximum likelihood inference, multiple imputation and augmented inverse probability weighting (Seaman et al., 2012; Rubin, 2018; Van der Laan et al., 2011; Robins et al., 1994; Tsiatis, 2006) , all of which are viable options for obtaining valid inference, while formally accounting for selection bias. Lastly, missing data are said to be MNAR, if say, an unmeasured factor induces missingness and is also associated with the outcome, rendering the two dependent. In such setting, identification can be

significantly more challenging, and the aforementioned methods designed for MCAR or MAR are generally invalid due to the latent dependence between the missingness mechanism and the unobserved outcome. Unfortunately, whether missing data mechanisms are MAR or MNAR is untestable from the observed data alone without a different strong assumption. As a result, there has been growing interest over the recent years in the development of statistical methods that can accommodate MNAR mechanisms.

To identify a target population functional or parameter of interest under MNAR, some existing statistical methods have relied on strong parametric model restrictions, which can be sensitive to even slight specification error (Diggle and Kenward, 1994; Wu and Carroll, 1988; Roy, 2003; Miao et al., 2016). Alternatively, sensitivity analysis has been proposed as a framework to assess the potential impact of MNAR mechanisms on inference in nonparametric and semiparametric models (Rotnitzky and Robins, 1997; Robins et al., 2000). A different strand of work has considered nonparametric identification and semiparametric inference leveraging a so-called shadow variable, apriori known to be independent of the missingness mechanism conditional on the potential unobserved outcome variable and measured covariates (Miao et al., 2024; Kott, 2014; Miao and Tchetgen Tchetgen, 2016; Li et al., 2023). While, partly inspired by methods originally proposed by Heckman (1979, 1997) and further developed by Das et al. (2003); Newey et al. (1990) , Tchetgen Tchetgen and Wirth (2017) and Sun et al. (2018) considered an instrumental variable (IV) approach to account for selection bias for an outcome missing not at random. An IV in a missing outcome data setting is a fully observed variable, which is apriori known to be independent of the outcome in the underlying population, and is a strong correlate of the missingness process, conditional on observed covariates. Though Heckman's selection model relies on crucial parametric models specification to identify a population outcome regression function in view, Sun et al. (2018) established necessary and sufficient conditions for identifying the full data joint distribution within this IV framework; and Tchetgen Tchetgen and Wirth (2017) considered a sufficient condition for IV identification of a regression function under a homogoneity condition which restricts the degree of selection bias on the scale of the outcome model.

In this paper, we propose a new multiplicative instrumental variable (MIV) model for selection bias due to outcome missing not at random, which provides identification of any full data functional of the outcome, while avoiding the homogeneity restrictions of previous methods on the outcome model scale. Our key identifying assumption is that the IV and any unobserved common correlate of both the missingness mechanism and the outcome, impact the selection probability via a product form, reflecting independent mechanisms of actions. For example, in a study estimating HIV seroprevalence among adults in Mochudi, Botswana, data were collected through a household survey (Novitsky et al., 2015). Unfortunately, a substantial fraction of enumerated households had one or more participants who did not complete the HIV testing component of the survey visit. The concern is that, failure to test for HIV may be associated with unobserved health-seeking behaviors and other risk factors for HIV infection, in which case missing data on HIV status is most likely not at random. In this setting, because interviewers are assigned at random, an *interviewer's characteristics*, such as years of experience as a survey field worker, years of education and other personal characteristics are strong predictors of whether or not an interviewee agrees to test for HIV, that are exogenous and therefore independent of an *interviewee's unmeasured health-seeking behavior*; therefore interviewer characteristics are strong candidate instruments. Furthermore, it is unlikely that the mechanism by which an interviewer's characteristics might lead to nonresponse interacts with one by which an interviewee's individual risk factors for HIV infection might lead to nonresponse. Such *independent mechanisms of action* is faithfully represented by the proposed multiplicative selection instrumental variable model.

Our main contributions beyond nonparametric identification of any given smooth functional of the outcome distribution in the target population, including the outcome mean, include the characterization of the influence function for the functional of interest under a semiparametric full data model. In addition, we propose a multiply robust estimator for the population outcome mean functional with high-order asymptotic bias, and as a result, of any functional that can be expressed as the solution to a moment equation, under the proposed multiplicative selection IV model. Our work is closely related to and builds upon recent advances by Liu et al. (2025); Lee et al. (2025), where an analogous MIV model is proposed for making inferences about a causal effect. Crucially, we extend their results from causal inference to missing data problems, whereby, we generalize their MIV framework from the case of a single binary IV, to more general IV settings, where the instrument may be polytomous, continuous, or multivariate. As we demonstrate, these generalizations are not trivial extensions of prior results and require new technical developments, particularly with respect to the semiparametric estimation and efficiency theory, therefore providing several new results which are of

independent interest both for missing data and causal IV settings.

The rest of the paper is organized as follows. Section 2 formally introduces notation for the paper and defines the multiplicative selection model. Section 3 presents the identification and estimation results for the population mean under missingness, starting with the binary IV case. This section largely restates the results of Liu et al. (2025) and Lee et al. (2025), reformulated in missing-data notation. Section 4 contains our main contributions, extending the identification and estimation results to the setting of general IVs. We derive the influence function of a smooth functional of the outcome, and we establish the asymptotic properties of the corresponding semiparametric estimator. Section 5 demonstrates the performance of our proposed estimators through simulation studies. Section 6 applies our method to the HIV seroprevalence study in Mochudi, Botswana. Section 7 concludes with a discussion of potential future extensions.

2 Multiplicative Selection Model

2.1 Notations and Assumptions

Suppose one observes n independent and identically distributed (i.i.d) samples on $O = (X, RY, R, Z)$, such that Y is not observed if $R = 0$. Let X denote observed covariates, Y is the outcome of interest, and Z is the IV. Y can be either continuous, binary, or polytomous, and X is a vector with possibly all three types of variables. Suppose that one aims to estimate a functional $\psi_0 = \psi(F)$ of the full data distribution F of (X, Y), that uniquely solves the full data moment equation

$$0 = \mathbb{E}\left[h\left(Y;\psi_0\right)\right] = \mathbb{P}\left(R=1\right)\mathbb{E}\left[h\left(Y;\psi_0\right)|R=1\right] + \mathbb{P}\left(R=0\right)\mathbb{E}\left[h\left(Y;\psi_0\right)|R=0\right],$$

where h is a pre-specified function and ψ_0 is the true value of some parameter of interest. The following are two standard examples.

- If $h\left(Y;\psi_0\right) = Y - \psi_0$, then $\psi_0 = \mathbb{E}\left[Y\right]$ is the population mean;
- If $h\left(Y;\psi_0\right) = \mathbb{1}\left\{Y \geq \psi_0\right\} - q$, where $\mathbb{1}\left\{\cdot\right\}$ is the indicator function, then, supposing that Y is continuous $\psi_0 = F_Y^{-1}(q)$ is the q-th quantile of the population.

Suppose further that there exist latent factors U, which induce dependence between Y and the missingness indicator R, such that the missingness mechanism is as a result MNAR. Because Y is observed for subjects with $R = 1$, the key challenge is to estimate the unobservable functional $\beta := \mathbb{E}\left[h\left(Y;\psi_0\right)|R=0\right]$. To identify β, we assume the multiplicative IV model with dependence structure in the population illustrated in the following the Directed Acyclic Graph (DAG) in Figure 1, which encodes the following missing data IV core conditions.

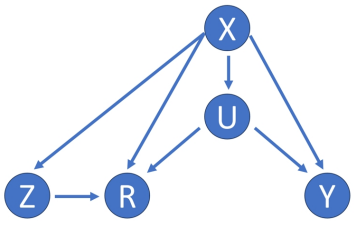

Figure 1: DAG of the multiplicative selection model with unmeasured confounders U included.

Assumption 1 (weak ignorability)**.** *Y independent of Z, R given (U, X).*

Assumption 2 (conditional independence)**.** *U independent of Z given X.*

Assumption 3 (multiplicative selection model and instrumental relevance)**.**
$\mathbb{P}\left(R=0|Z,U,X\right) = \exp\left\{\alpha_z\left(Z,X\right) + \alpha_u\left(U,X\right)\right\}$*, where $\alpha_z(\cdot,\cdot)$ and $\alpha_u(\cdot,\cdot)$ are arbitrary functions satisfying $\alpha_z\left(z,x\right) \neq \alpha_z\left(z',x\right)$ for all x and $z \neq z'$ where the natural constraint $\mathbb{P}(R=0|Z,U,X) \in [0,1)$ holds almost surely.*

Assumption 1 is quite a mild condition as U is not required to be observed. Assumption 2 indicates that the source of variation induced by the instrument must essentially be exogenous relative to the outcome process, a critical design consideration in selecting an appropriate instrument. Assumption 3 formalizes the multiplicative selection model condition and IV relevance. An immediate implication of 3 is that $\tilde{\pi}(z,u,x) \neq \tilde{\pi}(z',u,x)$ for all u,x and $z \neq z'$, where $\tilde{\pi}(Z,U,X) = \mathbb{P}(R=1|Z,U,X)$; a standard IV relevance condition.

We use the following notations throughout the paper:

$$\begin{aligned}
&\pi_z(X) = \mathbb{P}(R=1|Z=z,X), && \pi(X) = \mathbb{P}(R=1|X),\\
&\pi_0 = \mathbb{P}(R=0), && \rho_z(X) = \mathbb{P}(Z=z|X),\\
&\mu_z(X) = \mathbb{E}\{Rh(Y;\psi_0)|Z=z,X\}, && \delta^R(X) = \pi_1(X) - \pi_0(X),\\
&\delta^Y(X) = \mu_1(X) - \mu_0(X), && \delta(X) = \delta^Y(X)/\delta^R(X).
\end{aligned}$$

Note that $\delta(X)$ is a single-arm version of the Wald ratio estimand, a standard target of inference in instrumental variable causal inference literature, tailored here to address missing data.

To simplify the exposition, we first consider the binary IV setting in Section 3 and subsequently extend our results to the general IV setting in Section 4.

3 Identification and Estimation for Binary IV

In this section, we present the identification and estimation results for the population mean when Z is binary. These results essentially restate those of Liu et al. (2025) and Lee et al. (2025) in missing-data notation; we refer the reader to their papers for detailed proofs.

3.1 Identification of the Functional

Following Liu et al. (2025) and Lee et al. (2025), the missing population mean β can be identified via the single-arm Wald ratio functional $\delta(X)$:

$$\beta = \mathbb{E}_{X|R=0}[\delta(X)|R=0].$$

Based on this identification result, one may in principle proceed to construct a plug-in estimator for β

$$\hat{\beta}_{ID} = n_0^{-1} \sum_{R_i=0} \hat{\delta}(X_i),$$

where n_0 is the number of subjects with missing outcomes and $\hat{\delta}$ is an estimator of δ. However, even if $\hat{\delta}$ is consistent, using say flexible machine learning, the plug-in bias will in general be of first order with convergence rate substantially slower than $\sqrt{n}$ (Robins et al., 2017). To address this limitation of the plug-in estimator, we develop a de-biased estimator which is guaranteed to have second-order bias, based on the influence function of β, using modern semiparametric theory (Robins et al., 2017; Chernozhukov et al., 2018).

3.2 Influence Function of the Functional

Following Liu et al. (2025) and Lee et al. (2025), the influence function (IF) of β under a nonparametric model which places no restriction on the observed data distribution has the explicit form:

$$\phi(\beta,O,P) = \frac{2Z-1}{\rho_Z(X)} \cdot \frac{1-\pi(X)}{\pi_0\delta^R(X)} [Rh(Y;\psi_0) - R\delta(X) - \mu_0(X) + \pi_0(X)\delta(X)] + \frac{1-R}{\pi_0}[\delta(X) - \beta]. \quad (1)$$

Based on the derived IF, following Liu et al. (2025) and Lee et al. (2025), one may construct the semiparametric estimator for the mean functional among incomplete cases β.

$$\hat{\beta}_{IF} = \frac{1}{n}\sum_{i=1}^{n} \frac{2Z_i-1}{\hat{\rho}_{Z_i}(X_i)} \cdot \frac{1-\hat{\pi}(X)}{\hat{\pi}_0\hat{\delta}^R(X_i)} \left[Rh(Y_i) - R\hat{\delta}(X_i) - \hat{\mu}_0(X_i) + \hat{\pi}_0(X_i)\hat{\delta}(X_i)\right] + \frac{1}{n}\sum_{i=1}^{n} \frac{1-R_i}{\hat{\pi}_0} \cdot \hat{\delta}(X_i)$$

where $\hat{\pi}_z, \hat{\pi}(X), \hat{\pi}_0, \hat{\rho}_z(X), \hat{\mu_z}(X), \hat{\delta}^R(X), \hat{\delta}^Y(X), \hat{\delta}(X)$ are estimators of π_z, $\pi(X)$, π_0, $\rho_z(X)$, $\mu_z(X)$, $\delta^R(X)$, $\delta^Y(X)$, $\delta(X)$. Note that estimating the Wald-type ratio $\delta(X)$ can be unstable due to its ratio structure. To mitigate this issue, one may employ the Forster-Warmuth estimator for counterfactual regression problems of Yang et al. (2023). This estimator potentially attains the minimax rate of estimation of such a regression function under weaker conditions than a standard substitution-based estimator, see Liu et al. (2025) and Lee et al. (2025) for details.

Proceeding as in Liu et al. (2025) and Lee et al. (2025), one can show that

$$\begin{aligned}\sqrt{n}\left(\hat{\beta}-\beta\right) = &\underbrace{\sqrt{n}\int \phi(\beta, O, P)\, dP_n(O)}_{\text{CLT}} + \underbrace{\sqrt{n}R\left(\hat{P}, P\right)}_{\text{Second-order remainder}} \\ &+ \underbrace{\sqrt{n}\int \left\{\phi\left(\beta, O, \hat{P}\right) - \phi(\beta, O, P)\right\} d\left\{P_n(O) - P(O)\right\}}_{\text{Empirical process}},\end{aligned} \tag{2}$$

where P stands for the true distribution, P_n is the empirical distribution from the observed data, $\hat{P}$ denotes the estimated distribution, and $\phi(\beta, \cdot, \cdot)$ is the influence function corresponding to $\hat{\beta}$. The first term takes the form of a standard sample average and is asymptotically normal by the Central Limit Theorem (CLT). The third term is an empirical process term, which is $o_P(1)$ under certain Donsker conditions. However, this term can still be negligible through the use of cross-fitting and double machine learning techniques, even when the Donsker conditions are not satisfied, as we demonstrate later in Section 4.3. The key component is the remainder term $R(\hat{P}, P)$. The use of the IF guarantees that the corresponding estimator has a second-order remainder, which can be written as a sum of products of the estimation errors of nuisance functions. The following equation provides the explicit form of this remainder term, also characterized in Liu et al. (2025) and Lee et al. (2025).

$$\begin{aligned}R\left(\hat{P}, P\right) = &\mathbb{E}_P\left[\frac{1}{\pi_0}\left\{\hat{\delta}(X) - \delta(X)\right\}\left\{\hat{\pi}(X) - \pi(X)\right\}\right] + o_P\left(n^{-1/2}\right) \\ &+ \mathbb{E}_P\left[\frac{1-\hat{\pi}(X)}{\pi_0 \hat{\delta}^R(X)}\left\{\hat{\delta}(X) - \delta(X)\right\}\left\{\hat{\delta}^R(X) - \delta^R(X)\right\}\right] \\ &- \mathbb{E}_P\left[\frac{\delta^R(X)\left\{1-\hat{\pi}(X)\right\}}{\pi_0 \hat{\delta}^R(X)\hat{\rho}_1(X)}\left\{\hat{\delta}(X) - \delta(X)\right\}\left\{\hat{\rho}_0(X) - \rho_0(X)\right\}\right] \\ &- \mathbb{E}_P\left[\frac{1-\hat{\pi}(X)}{\pi_0 \hat{\delta}^R(X)\hat{\rho}_1(X)\hat{\rho}_0(X)}\left\{\hat{\mu}_0(X) - \mu_0(X)\right\}\left\{\hat{\rho}_0(X) - \rho_0(X)\right\}\right] \\ &+ \mathbb{E}_P\left[\frac{\hat{\delta}(X)\left\{1-\hat{\pi}(X)\right\}}{\pi_0 \hat{\delta}^R(X)\hat{\rho}_1(X)\hat{\rho}_0(X)}\left\{\hat{\pi}_0(X) - \pi_0(X)\right\}\left\{\hat{\rho}_0(X) - \rho_0(X)\right\}\right]\end{aligned} \tag{3}$$

Therefore, if the nuisance function estimators satisfy $R\left(\hat{P}, P\right) = o_P\left(n^{-1/2}\right)$, the IF-based estimator $\hat{\beta}_{IF}$ from sample splitting is asymptotically normal:

$$\sqrt{n}(\hat{\beta}_{IF} - \beta) \xrightarrow{D} N(0, \sigma^2),$$

where $\sigma^2 = \mathbb{V}\left[\phi(\beta, O, P)\right]$.

Following Liu et al. (2025) and Lee et al. (2025), the IF-based estimator has the multiple robustness property as another benefit. The IF $\phi(\beta, O, P)$ is unbiased under the union of three models, i.e., if one of the following holds:

1. $\delta(X)$ is correct, and at least one of $\mu_0(X), \pi_0(X)$ and $\mu_1(X), \pi_1(X)$ is correct;
2. $\delta^R(X), \rho_Z(X), \pi(X)$ are correct;
3. $\delta(X)$ and $\rho_Z(X)$ are correct.

4 Extension to General Instrumental Variables

In this section, we present new results that extend the multiplicative IV model for binary IV to a more general setting which accommodates more general forms of instruments, including polytomous, continuous, and multi-dimensional IVs, which are particularly common in missing data IV settings. Building on the approach from the previous section, we begin by presenting the identification formula for β and then derive its IF. The IF derived for the binary IV setting can be shown to be recovered as a special case of the general formulation. We further construct the corresponding IF-based estimator and derive its remainder term, which we show is of second-order. The asymptotic properties of the estimator and its multiple robustness are also established based on the analysis of the remainder term.

For the general scenario, the following notations are used throughout this section:

$$\mu(X) = \mathbb{E}_{R,Y|X}\{Rh(Y;\psi_0)\,|X\}, \qquad \delta^Y(Z,X) = \mu_Z(X) - \mu(X),$$
$$\delta^R(Z,X) = \pi_Z(X) - \pi(X), \qquad \delta(Z,X) = \delta^Y(Z,X)/\delta^R(Z,X).$$

Note that the functions $\delta^Y(Z,X)$, $\delta^R(Z,X)$, and $\delta(Z,X)$ are defined as contrasts between quantities for a given level Z and the corresponding marginal mean, rather than as differences between the two levels in the binary IV case. Consequently, they depend jointly on Z and X, rather than on X alone. While these general quantities cannot be directly reduced to the binary case, they are closely related. For example, in the proof of Corollary 1 in the supplementary materials, we show that

$$\delta^R(Z=1,X) = \delta^R(X)\,\mathbb{P}(Z=0 \mid X).$$

4.1 Identification of the Functional

The following theorem establishes that *the generalized Wald-type functional* $\delta(Z,X)$ identifies the target parameter β.

Theorem 1. *The incomplete-case population mean parameter β can be identified by*

$$\beta = \mathbb{E}_{Z,X|R=0}\left[\delta(Z,X)\,|R=0\right].$$

Analogous to Section 3.1, a plug-in estimator based on the identification formula can be constructed as $\hat{\beta}_{\mathrm{ID}} = n_0^{-1}\sum_{R_i=0}\hat{\delta}(Z_i,X_i)$. However, analogous to the binary IV case, its bias will generally be of first order, motivating the following developments involving the corresponding IF which is guaranteed to have second order bias.

4.2 Influence Function of the Functional

The following result gives the IF for β under the semiparametric model.

Theorem 2. *In the scenario of general IV Z, the influence function of the missing population mean β is*

$$\phi(\beta,O,P) = \{g(Z,X) - g(X)\}\left[Rh(Y) - \mu(Z,X) - \delta(Z,X)\{R - \pi(Z,X)\}\right] + \frac{1-R}{\pi_0}\{\delta(Z,X) - \beta\},$$

where

$$g(Z,X) = \frac{1-\pi(Z,X)}{\pi_0\delta^R(Z,X)}, \text{ and } g(X) = \mathbb{E}_{Z|X}\left[g(Z,X)\,|X\right].$$

It is worth noting that the IF in the previous theorem generalizes the result of binary IV, i.e., (1) is a special case of Theorem 2 when Z is binary.

Corollary 1. *The IF in Theorem 2 is reduced to* (1) *when Z is binary.*

Theorem 2 provides the IF for the missing population mean β. Given this theorem, it is straightforward to derive the IF for the mean of the whole population, as stated in the following corollary.

Corollary 2. *In the scenario of general IV, the functional*

$$\mathbb{E}\left[h\left(Y;\psi_0\right)\right]=\mathbb{P}\left(R=1\right)\alpha+\pi_0\beta$$

as the population mean has IF given by

$$\begin{aligned}&\pi_0\left\{g\left(Z,X\right)-g\left(X\right)\right\}\left[Rh\left(Y\right)-\mu\left(Z,X\right)-\delta\left(Z,X\right)\left\{R-\pi\left(Z,X\right)\right\}\right]\\&+\alpha\left\{R-\mathbb{P}\left(R=1\right)\right\}+\beta\left(1-R-\pi_0\right)+R\left\{h\left(Y;\psi_0\right)-\alpha\right\}+\left(1-R\right)\left\{\delta\left(Z,X\right)-\beta\right\}.\end{aligned}$$

Using IF, each functional can be semiparametrically estimated. For example, we estimate the missing population mean β by

$$\hat{\beta}_{IF}=\frac{1}{n}\sum_{i=1}^{n}\left\{\hat{g}\left(Z_i,X_i\right)-\hat{g}\left(X_i\right)\right\}\left[R_ih\left(Y_i\right)-\hat{\mu}\left(Z_i,X_i\right)-\hat{\delta}\left(Z_i,X_i\right)\left\{R_i-\hat{\pi}\left(Z_i,X_i\right)\right\}\right]+\frac{1-R_i}{\hat{\pi}_0}\cdot\hat{\delta}\left(Z_i,X_i\right),$$

where $\hat{g}\left(Z_i,X_i\right),\hat{g}\left(X_i\right),\hat{\mu}\left(Z_i,X_i\right),\hat{\delta}\left(Z_i,X_i\right),\hat{\pi}\left(Z_i,X_i\right)$ are estimators of $g\left(Z_i,X_i\right),g\left(X_i\right),\mu\left(Z_i,X_i\right),\delta\left(Z_i,X_i\right),\pi\left(Z_i,X_i\right)$.

Next, we present another main contribution of our paper. The following theorem derives the explicit form of the remainder term for the IF-based estimator in the general IV setting and shows that it is of second order.

Theorem 3. *In the scenario of general IV, the IF-based estimator for the missing population mean $\hat{\beta}_{IF}$ has the second-order remainder term:*

$$\begin{aligned}R\left(\hat{P},P\right)=&\mathbb{E}_P\left[\frac{1}{\pi_0}\left\{\hat{\delta}\left(X\right)-\delta\left(X\right)\right\}\left\{\hat{\pi}\left(Z,X\right)-\pi\left(Z,X\right)\right\}\right]\\&+\mathbb{E}_P\left[\hat{g}\left(Z,X\right)\left\{\hat{\delta}\left(X\right)-\delta\left(X\right)\right\}\left\{\hat{\delta}^R\left(Z,X\right)-\delta^R\left(Z,X\right)\right\}\right]\\&+\mathbb{E}_P\left[\delta^R\left(Z,X\right)\left\{\hat{\delta}\left(X\right)-\delta\left(X\right)\right\}\left[\mathbb{E}_{\hat{P}}\left\{\hat{g}\left(Z,X\right)|X\right\}-\mathbb{E}_P\left\{\hat{g}\left(Z,X\right)|X\right\}\right]\right]\\&+\mathbb{E}_P\left[\left\{\hat{\mu}\left(X\right)-\mu\left(X\right)\right\}\left[\mathbb{E}_{\hat{P}}\left\{\hat{g}\left(Z,X\right)|X\right\}-\mathbb{E}_P\left\{\hat{g}\left(Z,X\right)|X\right\}\right]\right]\\&-\mathbb{E}_P\left[\hat{\delta}\left(X\right)\left\{\hat{\pi}\left(X\right)-\pi\left(X\right)\right\}\left[\mathbb{E}_{\hat{P}}\left\{\hat{g}\left(Z,X\right)|X\right\}-\mathbb{E}_P\left\{\hat{g}\left(Z,X\right)|X\right\}\right]\right].\end{aligned}$$

The remainder term for the general IV case has a structure analogous to that of the binary IV case. Specifically, the quantity $\mathbb{E}_{\hat{P}}\left\{\hat{g}\left(Z,X\right)\mid X\right\}-\mathbb{E}_P\left\{\hat{g}\left(Z,X\right)\mid X\right\}$ in (3) corresponds to $\hat{\rho}_0(X)-\rho_0(X)$ in Theorem 3, both of which capture the error in estimating the conditional density of Z given X.

The main advantage of having a second-order bias is that it requires much weaker conditions for the estimator's asymptotic normality. Specifically, to ensure that the remainder term is negligible, we need $R(\hat{P},P)=o_P(n^{-1/2})$. For estimators not derived from the IF, such as the identification-based estimator, the remainder term is typically of the same order as the bias of the nuisance function estimators. Consequently, these nuisance functions must be estimated at a rate faster than $\sqrt{n}$, which generally requires the use of parametric models. However, parametric models may be misspecified in practice, potentially introducing bias.

In contrast, nonparametric methods offer greater flexibility but usually cannot achieve $\sqrt{n}$-rate convergence. Fortunately, since the remainder term is second-order, it is sufficient for the product of the nuisance estimation rates to be faster than $n^{-1/2}$, for instance, if each nuisance function is estimated at a rate faster than $n^{-1/4}$. As a result, both the remainder and empirical process terms in the decomposition (2) become negligible, and only the sample average term remains, which is asymptotically normal by the CLT. The asymptotic properties of the IF-based estimator are formalized in the following corollary.

Corollary 3. *In the scenario of general IV, if the nuisance function estimators satisfy $R\left(\hat{P},P\right)=o_P\left(n^{-1/2}\right)$, the IF-based estimator $\hat{\beta}_{IF}$ applying sample splitting in Section 4.3 is asymptotically normal:*

$$\sqrt{n}(\hat{\beta}_{IF}-\beta)\xrightarrow{D}N(0,\sigma^2),$$

where $\sigma^2=\mathbb{V}\left[\phi\left(\beta,O,P\right)\right]$

Additionally, the corollary provides a consistent variance estimator for the IF-based estimator The variance can be estimated using the influence function, which enables valid inference procedures, including hypothesis testing and the construction of confidence intervals.

$$\hat{\mathbb{V}}\left(\hat{\beta}_{IF}\right) = n^{-2}\sum_{i=1}^{n}\phi\left(\beta, O_i, \hat{P}\right). \tag{4}$$

Another benefit of the IF-based estimator is its multiple robustness, which also arises from the higher-order bias structure. When parametric models are used to estimate the nuisance functions, this property ensures that the estimator remains consistent even if some of the models are misspecified. The following corollary presents several scenarios under which the estimator is valid.

Corollary 4. *In the scenario of general Z, $\phi(\beta, O, P)$ is unbiased under the union of three models, that is if one of the following holds:*

1. *$\delta(X), \mu(X), \pi(X)$ are correct;*
2. *$\pi(Z, X), \rho_Z(X)$ are correct;*
3. *$\delta(X)$ and $f(Z|X)$ are correct.*

4.3 Cross-Fitting and Double Machine Learning

To derive estimators for the target functional, it is essential to estimate nuisance functions such as $\mu(Z, X)$ and $\pi(Z, X)$. Parametric methods are commonly used for this purpose; however, if some of the parametric models are misspecified, the resulting estimators will be inconsistent, as discussed in the previous section. To improve robustness and accuracy, nonparametric methods, including modern machine learning algorithms, have been applied to estimate these nuisance functions. Despite their flexibility, such methods may introduce regularization bias and overfitting, which can prevent the estimators from achieving the standard convergence rate. In particular, the empirical process term in (2) may fail to be $o_P(1)$, causing the variance estimator in (4) to be invalid.

To address this issue, cross-fitting and double machine learning techniques have been developed to improve the robustness and accuracy of causal inference, particularly in econometrics and machine learning (Chernozhukov et al., 2018). Cross-fitting involves partitioning the data into multiple folds and using disjoint subsets for training and evaluation. This reduces overfitting and ensures that estimates are not overly dependent on a particular data subset. Double machine learning extends this idea by using machine learning algorithms to estimate nuisance functions in a two-step procedure: first, models are trained on the estimation fold to estimate the nuisance functions; second, these estimates are used to evaluate the influence function on the evaluation fold. This separation induces independence between the nuisance function estimation and the evaluation of the influence function, which helps ensure that the empirical process term in (2) becomes negligible. Algorithm 1 outlines the detailed steps for implementing the cross-fitted version of the IF-based estimator $\hat{\beta}_{\mathrm{IF}}$.

5 Simulation Study

We conduct simulation studies to evaluate the performance of our proposed estimator. The simulation consists of two parts: the first considers a one-dimensional binary IV, while the second extends the setting to a general IV scenario with two binary IVs.

5.1 Binary Instrumental Variable

In the first part, we design a data-generating process involving a single binary IV. Two-dimensional covariates $X = (X_1, X_2)$ are independently drawn from a uniform distribution $U(0, 1)$, and an unobserved confounder U is independently drawn from a normal distribution $N(4, 0.5^2)$. The instrumental variable Z is generated from a logistic model conditional on X. The missingness indicator R follows the proposed multiplicative IV

Algorithm 1 Cross-Fitting to Derive $\hat{\beta}_{IF}$

Input: Dataset $\{(Y_i, X_i, Z_i, R_i)\}_{i=1}^n$, number of folds K.
Initialization: Partition the data into K folds.
for each fold $k \in \{1, \ldots, K\}$ **do**
Define training set $\mathcal{I}_k^c$ (all data except the k-th fold) and validation set $\mathcal{I}_k$ (the k-th fold).
Estimate the nuisance functions using the training set $\mathcal{I}_k^c$: $\hat{g}(Z_i, X_i)$, $\hat{g}(X_i)$, $\hat{\mu}(Z_i, X_i)$, $\hat{\delta}(Z_i, X_i)$, $\hat{\pi}(Z_i, X_i)$, $\hat{\pi}_0$.
for each $i \in \mathcal{I}_k$ **do**
Calculate the uncentered influence function component:

$$\tilde{\phi}\left(\beta, O_i, \hat{P}\right) = \left\{\hat{g}\left(Z_i, X_i\right) - \hat{g}\left(X_i\right)\right\}\left[R_i h\left(Y_i\right) - \hat{\mu}\left(Z_i, X_i\right) - \hat{\delta}\left(Z_i, X_i\right)\left\{R_i - \hat{\pi}\left(Z_i, X_i\right)\right\}\right] + \frac{1 - R_i}{\hat{\pi}_0} \cdot \hat{\delta}\left(Z_i, X_i\right)$$

end for
end for
Aggregate the results:
Calculate the mean of the influence function components over all data points:

$$\hat{\beta}_{IF} = \frac{1}{n} \sum_{I=1}^{n} \tilde{\phi}\left(\beta, O_i, \hat{P}\right)$$

Output: The cross-fitted estimator $\hat{\beta}_{IF}$.

model. The outcome Y is drawn from a normal distribution, with its expectation defined as a linear function of X modulated by a tilting term that depends on U. The detailed data-generating mechanism is provided below.

$$P\left(Z = 1|X\right) = \frac{\exp\left(-1 + X_1 + X_2\right)}{1 + \exp\left(-1 + X_1 + X_2\right)},$$
$$P\left(R = 0|Z, U, X\right) = \exp\left\{\left(-X_1 - X_2 - \frac{U}{4}\right) + Z\left(X_1 + X_2 + 1\right)\right\},$$
$$Y \sim N\left(\left(X_1 + X_2\right) \cdot \exp\left(\frac{U}{6}\right), 0.5^2\right).$$

We focus on estimating the mean of the missing outcome β, whose true value is 1.8. The two estimators introduced in the paper—the identification-based estimator $\hat{\beta}_{\text{ID}}$ and the influence-function-based estimator $\hat{\beta}_{\text{IF}}$—are used to estimate β. For the IF-based estimators, we apply five-fold cross-fitting to reduce regularization bias and use a median adjustment over 11 repetitions to mitigate sensitivity to particular random splits. Variance of the IF-based estimators can be estimated via (4), and the 95% confidence interval can be derived correspondingly.

We estimate the IV model using a generalized linear model with a logit link, incorporating all the covariates X. To estimate all remaining nuisance functions—including the missingness and outcome models—we employ the Super Learner algorithm, using both covariates and instrumental variables as inputs. The ensemble library consists of a variety of base learners: generalized linear models, multivariate adaptive regression splines, and multi-layer perceptrons. The sample sizes are set to 500, 750, and 1000, respectively, and each experiment is replicated 300 times. The degrees of freedom for the series of basis functions are set to 4, 5, and 6, respectively, depending on the sample size.

The results are shown in Figure 2 and Table 1, which display boxplots and summarize the bias, variance, and mean squared error (MSE) of various estimators. Coverage rates for the 95% confidence interval are reported only for the IF-based estimators, as a consistent variance estimator may not exist for the ID-based estimator. The recommended IF-based estimator generally exhibits smaller bias, variance, and MSE

compared to the ID-based estimator, consistent with its theoretical asymptotic efficiency. The coverage rates of the IF-based estimators remain close to the nominal 95%, supporting the consistency of the variance estimator derived from the influence function.

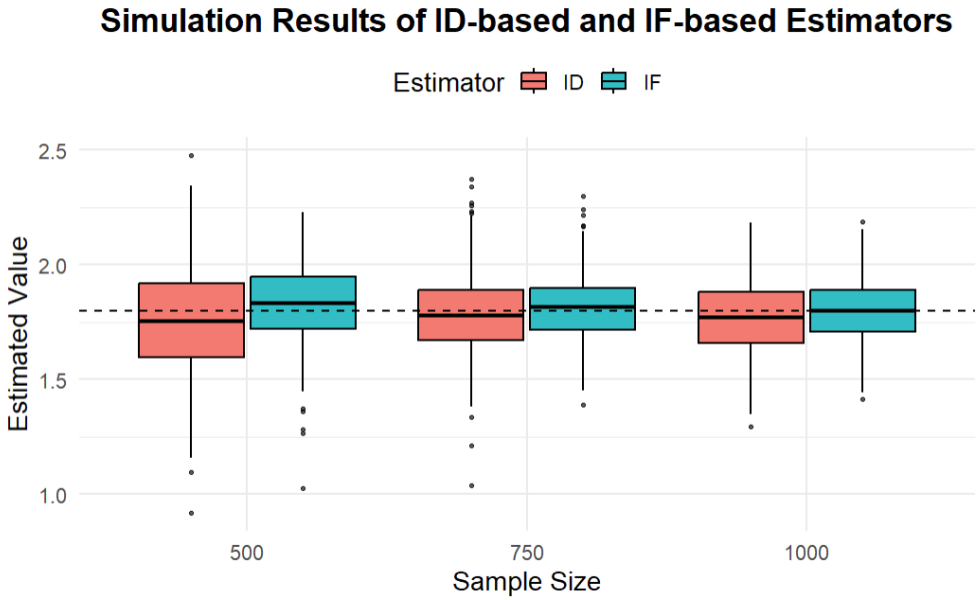

Figure 2: Boxplots of Estimators for the Binary Instrumental Variable Scenario Across Varying Sample Sizes. Dotted line indicates the true value.

Table 1: Simulation Results for the Binary IV Scenario: Bias, Variance, and MSE of Estimators Across Different Sample Sizes. Coverage rates of IF-based estimators are also included.

	$n = 500$		$n = 750$		$n = 1000$	
	$\hat{\beta}_{ID}$	$\hat{\beta}_{IF}$	$\hat{\beta}_{ID}$	$\hat{\beta}_{IF}$	$\hat{\beta}_{ID}$	$\hat{\beta}_{IF}$
Bias	-0.049	0.030	-0.017	0.020	-0.029	-0.00064
Var	0.058	0.029	0.036	0.020	0.030	0.018
MSE	0.061	0.029	0.037	0.021	0.031	0.018
Coverage Rate (%)	/	97.3	/	97.0	/	96.7

5.2 General Instrumental Variable

In the second part of our study, we consider a more general setting by introducing an additional binary IV into the data-generating process. The two-dimensional covariates $X = (X_1, X_2)$ and the unmeasured confounder U are independently drawn from the standard uniform distribution $U(0, 1)$. The instrumental variables $Z = (Z_1, Z_2)$ are both generated from logistic distributions. The observation indicator R is still determined by a multiplicative IV model, now influenced by both Z_1 and Z_2. The outcome model for Y remains unchanged. The full data-generating formulas are provided below.

$$P(Z_1 = 1|X) = \frac{\exp\{(-1 + X_1 + X_2)/4\}}{1 + \exp\{(-1 + X_1 + X_2)/4\}}$$

$$P(Z_2 = 1|X) = \frac{\exp\{(X_1 - X_2)/4\}}{1 + \exp\{(X_1 - X_2)/4\}}$$

$$P(R = 0|Z, U, X) = \exp\left[\frac{1}{4}\{8 + X_1 - X_2 - U + Z_1(-1 - X_1 - X_2) + Z_2(8 + X_1 - X_2)\}\right]$$

The target estimand is again the mean of the missing outcome β, with the true value now set to 1.07. The two estimators $\hat{\beta}_{\text{ID}}$ and $\hat{\beta}_{\text{IF}}$ are implemented in the same manner, using the same basis functions in the Super Learner algorithm and identical tuning parameters. Since there are $2 \times 2 = 4$ possible levels of the two binary IVs, additional nuisance functions are required for prediction. The sample size is increased to 10,000.

Figure 3 and Table 2 present the simulation results for the general IV scenario. The IF-based estimator exhibits slightly smaller bias but slightly larger variance. Compared to the binary IV scenario, the differences among the two estimators are less pronounced. One possible explanation for the limited advantage of the IF-based estimators over the ID-based estimator is that the general IV setting involves more levels, making the numerator $\delta^R(X, Z)$ more likely to approach zero. This increases numerical instability may also lead to an overestimation of the variance by the IF-based method. A potential remedy is to remove extreme outliers in the IF values across the dataset.

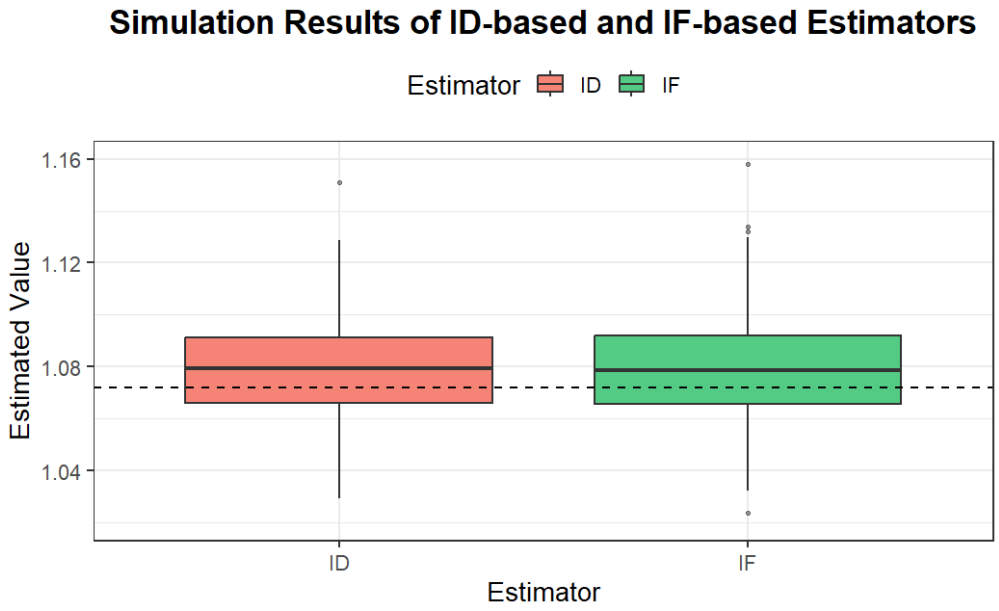

Figure 3: Boxplots of Estimators in the General Instrumental Variable Scenario. The dashed line indicates the true value.

Table 2: Bias, Variance, MSE, and Coverage Rate of Estimators in the General Instrumental Variable Scenario.

	$\hat{\beta}_{ID}$	$\hat{\beta}_{IF}$
Bias	0.00718	0.00688
Variance	0.000350	0.000403
MSE	0.000400	0.000449
Coverage Rate (%)	/	96.7

6 Real Data Application

To illustrate the proposed IV method, we examine data from a household survey conducted in Mochudi, Botswana, designed to estimate HIV seroprevalence among adults while addressing selective nonresponse in HIV testing. The survey included 4,997 adults aged 16 to 64, of whom 4,045 (81%) provided complete HIV test results. Among the remaining 952 individuals with missing HIV status ($R = 0$), 111 (2%) consented to

testing but lacked final results, and 841 (17%) explicitly refused to participate in the HIV testing component. Such nonparticipation, even among those successfully contacted, presents a likely source of selection bias.

The dataset contains fully observed covariates, including participant gender, age, and the number of nights spent away from home in the past month, which is an established risk factor for HIV. We consider interviewer characteristics as candidate instruments: gender, age, and years of experience. These attributes are expected to influence an individual's likelihood of completing HIV testing, but are unlikely to directly affect HIV status, especially given that interviewer assignments were randomized prior to the survey.

Potential unmeasured confounders, such as the health-seeking behavior of the participants, may jointly affect both their HIV status and their likelihood of responding to the test. Since these confounders pertain to participants rather than interviewers, it is reasonable to assume that they do not interact with interviewer-level instruments in the response mechanism. This supports the assumption of a separable and independent response model, making the multiplicative IV framework appropriate for this setting.

We estimate HIV seroprevalence using our proposed estimator based on the influence function. Since the instrumental variable is a three-dimensional vector, we apply the generalized IV approach described in Section 4. To facilitate analysis, the two continuous instruments—age and years of experience—are discretized into categorical variables based on their quartiles.

We estimate the IV model using a generalized linear model with a logit link, incorporating the relevant covariates. To estimate the remaining nuisance functions from the missingness and outcome models, we employ the Super Learner algorithm, using both covariates and instrumental variables as inputs. The ensemble library includes a range of basis learners: generalized linear models, multivariate adaptive regression splines, generalized additive models, polynomial splines, random forests, single-layer perceptrons, and multi-layer perceptrons. To mitigate overfitting and reduce bias from machine learning-based estimation, we apply five-fold cross-fitting throughout the procedure.

Our estimate of HIV seroprevalence among nonrespondents is 38.4% (95% CI: 33.1%–43.7%), substantially higher than the observed prevalence among respondents, which is 21.4% (95% CI: 20.2%–22.7%). This discrepancy highlights a significant degree of selection bias when nonrespondents are ignored. The estimated HIV seroprevalence for the overall adult population in Mochudi is 24.7% (95% CI: 18.6%–30.8%), which aligns closely with findings from previous studies, such as the 25.8% estimate reported by Sun et al. (2018).

7 Discussion

In this paper, we proposed a novel approach to handle nonignorable missing data by introducing a multiplicative selection model that incorporates instrumental variables and allows for the presence of unmeasured confounders. By leveraging semiparametric theory, we derived the influence function and constructed an estimator with desirable properties such as multiple robustness, and interpretability. Our framework generalizes existing work by extending from binary to general instrumental variables, including multi-level, continuous, and multi-dimensional cases. Simulation studies and a real-world application to an HIV seroprevalence survey in Botswana demonstrate the practical utility of our method.

It is important to note that the IF we derived for the general IV setting is inefficient, unlike in the binary IV case. The inefficiency arises because the target parameter is over-identified when the IV is polytomous, continuous, or multi-dimensional. Consequently, multiple identification formulas lead to multiple candidate IFs, and the one we obtained is only a particular instance that may not coincide with the efficient influence function (EIF). Obtaining the EIF requires projecting the IF onto the tangent space of the multiplicative IV model. In the supplementary material, we characterized this tangent space under the imposed restrictions. However, this derivation is highly non-trivial and structurally complex, which makes the projection extremely difficult and the resulting efficient estimator impractical for use. Nevertheless, although our estimator does not attain the semiparametric efficiency bound, it still provides the important advantage of reducing higher-order asymptotic bias.

Future research directions include extending the proposed model to longitudinal and developing tools for sensitivity analysis under possible violations of the model assumptions. Although the proposed estimator has desirable asymptotic properties, its finite-sample performance may vary. In particular, unstable outliers can occur when the denominators in the ratio structure of the Wald-type statistics are close to zero, which is especially common in the general IV setting where ratios must be evaluated for each IV level. Studying

the finite-sample behavior of the estimator and identifying potential remedies to improve its stability would be an interesting direction for future work.

References

Victor Chernozhukov, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins. Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, 21:1–68, 2018.

Mitali Das, Whitney K Newey, and Francis Vella. Nonparametric estimation of sample selection models. *The Review of Economic Studies*, 70:33–58, 2003.

Peter Diggle and Michael G Kenward. Informative drop-out in longitudinal data analysis. *Journal of the Royal Statistical Society Series C: Applied Statistics*, 43(1):49–73, 1994.

James Heckman. Instrumental variables: A study of implicit behavioral assumptions used in making program evaluations. *J Hum Resour*, 32:441–462, 1997.

James J Heckman. Sample selection bias as a specification error. *Econometrica*, 47:153–161, 1979.

Phillip S Kott. Calibration weighting when model and calibration variables can differ. In L. P. Conti & G. M. Ranalli F. Mecatti, editor, *Contributions to Sampling Statistics*, pages 1–18. Cham: Springer, 2014.

Yonghoon Lee, Mengxin Yu, Jiewen Liu, Chan Park, Yunshu Zhang, James M. Robins, and Eric J. Tchetgen Tchetgen. Inference on nonlinear counterfactual functionals under a multiplicative iv model, 2025. URL `https://arxiv.org/abs/2507.15612`.

Wei Li, Wang Miao, and Eric Tchetgen Tchetgen. Non-parametric inference about mean functionals of non-ignorable non-response data without identifying the joint distribution. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 85(3):913–935, 2023.

Jiewen Liu, Chan Park, Yonghoon Lee, Yunshu Zhang, Mengxin Yu, James M. Robins, and Eric J. Tchetgen Tchetgen. The multiplicative instrumental variable model, 2025. URL `https://arxiv.org/abs/2507.09302`.

Wang Miao and E. J. Tchetgen Tchetgen. On varieties of doubly robust estimators under missingness not at random with a shadow variable. *Biometrika*, 2:475–482, 2016.

Wang Miao, Peng Ding, and Zhi Geng. Identifiability of normal and normal mixture models with nonignorable missing data. *J Am Stat Assoc*, 111:1673–1683, 2016.

Wang Miao, Lan Liu, Yilin Li, Eric J Tchetgen Tchetgen, and Zhi Geng. Identification and semiparametric efficiency theory of nonignorable missing data with a shadow variable. *ACM/JMS Journal of Data Science*, 1(2):1–23, 2024.

Whitney K Newey, James L Powell, and James R Walker. Semiparametric estimation of selection models: some empirical results. *The american economic review*, 80(2):324–328, 1990.

Vlad Novitsky, Denise Kühnert, Sikhulile Moyo, Erik Widenfelt, Lillian Okui, and Max Essex. Phylodynamic analysis of hiv sub-epidemics in mochudi, botswana. *Epidemics*, 13:44–55, 2015.

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 Daniel O Scharfstein. Sensitivity analysis for selection bias and unmeasured confounding in missing data and causal inference models. In *Statistical Models in Epidemiology, the Environment, and Clinical Trials*, pages 1–94. Springer, New York: Springer, 2000.

James M. Robins, Lingling Li, Eric Tchetgen Tchetgen, Rajarshi Mukherjee, and Aad W. van der Vaart. Minimax estimation of a functional on a structured high-dimensional model. *Annals of Statistics*, 45(5): 1951–1987, 2017. doi: 10.1214/16-AOS1515.

Andrea Rotnitzky and James Robins. Analysis of semi-parametric regression models with non-ignorable non-response. *Statistics in medicine*, 16(1):81–102, 1997.

Jason Roy. Modeling longitudinal data with nonignorable dropouts using a latent dropout class model. *Biometrics*, 59(4):829–836, 2003.

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

Donald B Rubin. Multiple imputation. In *Flexible imputation of missing data, second edition*, pages 29–62. Chapman and Hall/CRC, 2018.

Shaun R Seaman, Ian R White, Andrew J Copas, and Leah Li. Combining multiple imputation and inverse-probability weighting. *Biometrics*, 68(1):129–137, 2012.

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

Eric J Tchetgen Tchetgen and Kathleen E Wirth. A general instrumental variable framework for regression analysis with outcome missing not at random. *Biometrics*, 73:1123–1131, 2017.

Anastasios A Tsiatis. *Semiparametric theory and missing data*. Springer, 2006.

Mark J Van der Laan, Sherri Rose, et al. *Targeted learning: causal inference for observational and experimental data*, volume 4. Springer, 2011.

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.

Yachong Yang, Arun Kumar Kuchibhotla, and Eric Tchetgen Tchetgen. Forster-warmuth counterfactual regression: A unified learning approach. *arXiv preprint arXiv:2307.16798*, 2023.