Identification and Semiparametric Efficiency Theory of Nonignorable Missing Data With a Shadow Variable

WANG MIAO, Department of Probability and Statistics, Peking University, China
LAN LIU, School of Statistics, University of Minnesota at Twin Cites , USA
YILIN LI, Department of Probability and Statistics, Peking University, China
ERIC J. TCHETGEN TCHETGEN, Department of Statistics, The Wharton School of the University of Pennsylvania, USA
ZHI GENG, School of Mathematics and Statistics, Beijing Technology and Business University, China

We consider identification and estimation with an outcome missing not at random (MNAR). We study an identification strategy based on a so-called *shadow variable*. A shadow variable is assumed to be correlated with the outcome, but independent of the missingness process conditional on the outcome and fully observed covariates. We describe a general condition for nonparametric identification of the full data law under MNAR using a valid shadow variable. Our condition is satisfied by many commonly-used models; moreover, it is imposed on the complete cases, and therefore has testable implications with observed data only. We characterize the semiparametric efficiency bound for the class of regular and asymptotically linear estimators, and derive a closed form for the efficient influence function. We describe a doubly robust and locally efficient estimation method and evaluate its performance on both simulation data and a real data example about home pricing.

Additional Key Words and Phrases: Doubly robust estimation, efficient influence function, identification, missing not at random, shadow variable

1 INTRODUCTION

Methods for missing data have received much attention in statistics and related areas. Suppose we are interested in an outcome prone to missingness. Data are said to be missing at random (MAR) if the missingness only depends on the observed data; otherwise, data are said to be missing not at random (MNAR). Consider inference about a full data functional with an outcome prone to missing values. The underlying full data law is identified under MAR and methods to make inference abound, to name a few, likelihood methods based on the expectation–maximization (EM) [Dempster et al. 1977], multiple imputation [Rubin 1987; Schenker and Welsh 1988], inverse probability weighting [Horvitz and Thompson 1952], and doubly robust methods [Bang and Robins 2005; Tsiatis 2006; Van der Laan and Robins 2003]. Among them, the doubly robust approach is in principle most robust, because it requires correct specification of either the full data law or of the missingness process, but not necessarily both, while likelihood or imputation methods require correct specification of the full data law, and likewise inverse probability weighting has to rely on correct specification of the missingness process. Because doubly robust methods effectively double one's chances to reduce bias due to model misspecification, such methods have grown in popularity in recent years for estimation with missing data and other forms of coarsening data [Tsiatis 2006; Van der Laan and Robins 2003]. Missing data have also been intensively studied in a range of modern fields, for example, in matrix completion [Jin et al. 2022; Mao et al. 2019], high-dimensional regression [Loh and Wainwright 2012], sparse principal component analysis [Zhu et al. 2022], classification [Tony Cai and Zhang 2019].

However, it is possible that MNAR occurs as missingness may depend on the missing values even after conditioning on the observed data. Compared to MAR, MNAR is much more challenging. As recently noted by Miao et al. [2016] and Wang et al. [2014], even fully parametric models are often non-identifiable under MNAR, that is, the parameters are not uniquely determined in spite

of infinite samples. Previous authors have studied the problem of identification under MNAR. Among them, Heckman [1979]'s outcome–selection model rests on a pair of parametric models for the outcome and the missingness process. Little [1993, 1994] introduced a pattern-mixture parametrization for incomplete data, which specifies the distribution of the outcome for each missing data pattern separately. Little studied identification of pattern-mixture models by imposing restrictions on unknown parameters across different missing data patterns, for example, setting the missing data distribution equal to that of the observed data. Fay [1986] and Ma et al. [2003] used graphical models for the missing data mechanism and studied identification for categorical variables. Rotnitzky et al. [1998] and Robins et al. [2000] developed sensitivity analysis methods given a completely known association between the outcome and the missingness process. Das et al. [2003], Tchetgen Tchetgen and Wirth [2017], Sun et al. [2018], and Liu et al. [2020] proposed identification conditions for nonparametric and semiparametric regression models with the help of an instrumental variable, which affects the missingness process but not the outcome.

Identification under MNAR is sometimes possible, if a fully observed correlate of the outcome is known to be independent of the missingness process, after conditioning on fully observed covariates and the outcome itself. Such a correlate, which we refer to as a *shadow variable*, is available in many empirical studies such as in survey sampling designs [Kott 2014]. Even with a shadow variable, identification often requires additional conditions. In the context of outcome-selection parametrization, D'Haultfœuille [2010] established identification of an additive regression model with a nonparametric propensity score model and proposed nonparametric estimation methods; Wang et al. [2014] studied identification with a parametric propensity score model and proposed inverse probability weighted estimation; Zhao and Shao [2015] studied identification of a parametric outcome model with a nonparametric propensity score model and developed pseudo-likelihood estimation methods; Miao and Tchetgen Tchetgen [2016] discussed identification of location-scale models and proposed doubly robust estimation. However, their various identification conditions involve the missing values and prior knowledge about the data generating mechanism, and therefore cannot be justified with observed data.

For estimation, several methods initially developed for MAR have recently been extended to handling MNAR data under suitable conditions, such as likelihood-based estimation [Greenlees et al. 1982; Tang et al. 2014], inverse probability weighting [Scharfstein et al. 1999], and regression based estimation [Fang et al. 2018; Vansteelandt et al. 2007]. In contrast, doubly robust estimation for MNAR data is not well developed. For some exceptions, see for instance Scharfstein and Irizarry [2003] and Vansteelandt et al. [2007] who proposed doubly robust estimators by assuming a completely known selection bias, i.e., the association between the outcome of interest and the missingness process. However, this approach may only be useful from the perspective of sensitivity analysis and its utility may be limited in most practical settings by overwhelming uncertainty about the unidentified selection bias. Miao and Tchetgen Tchetgen [2016] used a shadow variable to estimate the selection bias and proposed a suite of doubly robust estimators under more stringent identifying conditions, which were inspired by an unpublished initial draft of the current paper; however, both papers failed to develop the semiparametric theory for such estimators and to formally characterize their efficiency bound. Morikawa and Kim [2021] developed semiparametric efficiency theory when missing mechanism is parametric and correctly specified, so their estimator does not have the doubly robustness property. Zhao and Ma [2021] proposed another semiparametric efficiency estimation procedure where modeling of missingness mechanism is completely bypassed.

In this paper, we establish a novel identification condition and semiparametric inference under a general pattern mixture parametrization with a shadow variable. Given a shadow variable, we show that the full data distribution is nonparametrically identified under certain completeness condition in Section 3. In contrast to previous approaches that impose restrictions either on the full data law

or on the missing data distribution for the purpose of identification, our identifying condition only involves the observed data, and thus can be justified empirically. As a result, given a valid shadow variable, identification can be assessed with the observed data. In Section 4, we develop general semiparametric efficiency theory for MNAR data with a shadow variable, by characterizing the set of influence functions of any pathwise differentiable nonparametric functional of interest and the corresponding semiparametric efficiency bound. We derive a closed form for the efficient influence function and offer a one-step construction of the efficient estimator given a $n^{1/2}$-consistent and doubly robust initial estimator. In Section 5, we study the performance of a variety of estimators via both a series of simulations and a Home Pricing example. We conclude and discuss in Section 6, and relegate proofs to the Appendix.

2 PRELIMINARY

Throughout the paper, we let Y denote the outcome prone to missing values, R the missingness indicator with $R = 1$ if Y is observed and $R = 0$ otherwise, and X a vector of fully observed covariates. We use lower-case letters for realized values of the corresponding variables, for example, y for a value of the outcome variable Y. We use f to denote a probability density or mass function. Vectors are assumed to be column vectors unless explicitly transposed. Suppose one has also fully observed a variable Z that satisfies the following assumption of a shadow variable.

Assumption 1. $Z \perp\!\!\!\perp R \mid (X, Y)$ *and* $Z \not\perp\!\!\!\perp Y \mid (R = 1, X)$.

Assumption 1 formalizes the idea that the missingness process may depend on (X, Y), but not on the shadow variable Z after conditioning on (X, Y). Therefore, Assumption 1 allows for missingness not at random. Assumption 1 is analogous to the "nonresponse instrument" assumption previously made by D'Haultfœuille [2010]; Wang et al. [2014], and Zhao and Shao [2015], although we do not use such terminology to avoid confusion with literature on instrumental variables for missing data [Newey and Powell 2003; Sun et al. 2018; Tchetgen Tchetgen and Wirth 2017]. Figure 1 presents graphical model examples that illustrate the assumption. The second part of Assumption 1 in principle can be tested with the observed data. But the first part involves missing values of Y, however interestingly, it is sometimes refutable as pointed out by D'Haultfœuille [2010], that is, it can be rejected with observed data if the solution of a certain integral equation does not exist. Nonetheless, Assumption 1 may be reasonable in many empirical applications. For example, in a study of mental health of children in Connecticut [Zahner et al. 1992], researchers aimed at evaluating the prevalence of students with abnormal psychopathological status based on their teacher's assessment prone to missingness. As indicated by Ibrahim et al. [2001], a separate parental assessment constitutes a valid shadow variable in this study. Several other examples are described by Zhao and Ma [2018]; Zhao and Shao [2015] and Wang et al. [2014].

The full data contain n independent and identically distributed samples of (X, Y, Z), but in the observed data the values of Y are missing for $R = 0$. The observed data distribution is captured by $p(Y, R = 1 \mid X, Z)$, $p(R = 0 \mid X, Z)$ and $p(X, Z)$, which are functionals of the joint distribution $p(X, Y, Z, R)$. However, given the observed data distribution, the joint distribution may not be uniquely determined even with infinite samples, which is known as the identification problem in missing data analysis. Considering a joint distribution model $p(X, Y, Z, R; \theta)$ indexed by a possibly infinite dimensional parameter θ, it is said to be identifiable if and only if θ is uniquely determined by the observed data distribution $p(Y, R = 1 \mid X, Z)$, $p(R = 0 \mid X, Z)$ and $p(X, Z)$. Because $p(R, X, Z)$ is identified without extra assumptions, we focus on identification of $p(Y \mid R, X, Z)$.

Assumption 1 is key to identification of $p(Y \mid R, X, Z)$. Otherwise, if Z may affect the missingness after conditioning on (X, Y), then even fully parametric models may not be identified [Miao et al. 2016; Wang et al. 2014]. Without the shadow variable, only certain bounds can be obtained. In the

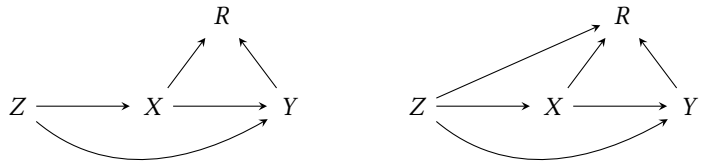

Fig. 1. Two diagram examples describing the relationship between the shadow variable Z, missingness indicator R, outcome Y, and covariates X: Assumption 1 holds in the graph on the left, but not in the one on the right.

next section, we will elaborate how one could use a shadow variable to improve identification of MNAR data, and discuss extra conditions that are required for identification.

3 A NOVEL IDENTIFICATION CONDITION

We factorize $p(Y, R \mid X, Z)$ as

$$p(Y, R \mid X, Z) = p(Y \mid R, X, Z)p(R \mid X, Z),$$

with $p(Y \mid R, X, Z)$ encoding the outcome distribution for different data patterns: $R = 1$ for the observed data and $R = 0$ for the missing data. Although $p(Y \mid R = 1, X, Z)$ can be obtained from complete cases, the missing data distribution $p(Y \mid R = 0, X, Z)$ is not directly available from the observed data under MNAR.

We use the odds ratio function to encode the deviation between the observed and missing data distributions:

$$\mathrm{OR}(X, Y, Z) = \frac{p(Y \mid R = 0, X, Z)p(Y = 0 \mid R = 1, X, Z)}{p(Y \mid R = 1, X, Z)p(Y = 0 \mid R = 0, X, Z)}. \tag{1}$$

Here, we use $Y = 0$ as a reference value, although any other value within the support of Y may be chosen by the analyst. The odds ratio function generalizes the approach of Little [1993, 1994] that imposes a known relationship between the data patterns. For instance, $\mathrm{OR}(X, Y, Z) = 1$ corresponds to identical data patterns $p(Y \mid R = 0, X, Z) = p(Y \mid R = 1, X, Z)$ or missingness at random. In the following, we establish the key role of the odds ratio function in nonignorable missing data analysis and propose to identify it with a shadow variable. We also make the following standard assumption:

Assumption 2. $0 < p(R = 1|X, Y, Z) < 1$ *with probability 1.*

Assumption 2 implies that $\mathrm{OR}(X, Y, Z) > 0$ and $E\{\mathrm{OR}(X, Y, Z) \mid R = 1, X, Z\} < +\infty$ hold with probability 1. Following the convention of expressing a joint density in terms of the odds ratio function and two baseline distributions [Chen 2003, 2004, 2007; Kim and Yu 2011; Osius 2004], we have the following results in the presence of a shadow variable.

Proposition 1. *Given Assumption 1 and 2, we have that for all* (X, Y, Z)

$$\mathrm{OR}(X, Y, Z) = \mathrm{OR}(X, Y) = \frac{p(R = 0 \mid X, Y)p(R = 1 \mid X, Y = 0)}{p(R = 0 \mid X, Y = 0)p(R = 1 \mid X, Y)}, \tag{2}$$

$$p(Y, R \mid X, Z) = c(X, Z)p(R \mid X, Y = 0)p(Y \mid R = 1, X, Z)\{\mathrm{OR}(X, Y)\}^{1-R}, \tag{3}$$

$$c(X, Z) = \frac{p(R = 1 \mid X)}{p(R = 1 \mid X, Y = 0)} \frac{p(Z \mid R = 1, X)}{p(Z \mid X)},$$

$$p(R = 1 \mid X, Y = 0) = \frac{E\{\mathrm{OR}(X, Y) \mid R = 1, X\}}{p(R = 0 \mid X)/p(R = 1 \mid X) + E\{\mathrm{OR}(X, Y) \mid R = 1, X\}}. \tag{4}$$

These results are straightforward to verify by applying the shadow variable Assumption 1. Identity (2) indicates that the odds ratio function also captures the impact of the outcome itself on the propensity score $p(R = 1 \mid X, Y)$, and is thus a measure of the selection bias, i.e., the degree to which the missingness departs from MAR. Under the shadow variable setting, the odds ratio function only depends on X and Y, which we therefore denote by $\mathrm{OR}(X, Y)$. A special case of the odds ratio function is the exponential tilting parameter of Scharfstein and Irizarry [2003] and Kim and Yu [2011], who assume a logistic propensity score model. However, they require that the exponential tilting parameter is known a priori or available from a follow-up study of nonrespondents. But here in principle, we allow for a nonparametric propensity score model with an unknown odds ratio function, and we aim to identify it using a shadow variable.

Identity (3) reveals a factorization of $p(Y, R \mid X, Z)$ that is determined by the odds ratio function $\mathrm{OR}(X, Y)$, the complete-case outcome distribution $p(Y \mid R = 1, X, Z)$, and the propensity score evaluated at the reference level $Y = 0$; we refer to the latter two as the baseline outcome distribution and the baseline propensity score, respectively. Because $p(Y \mid R = 1, X, Z)$ can be uniquely determined from complete cases, from (3)–(4), identification of $p(Y, R \mid X, Z)$ rests on $\mathrm{OR}(X, Y)$. This is further illustrated with the following results, which are implied from (3), and we omit the proof.

Proposition 2. *Given Assumption 1 and 2, we have that*

$$E\{\widetilde{\mathrm{OR}}(X, Y) \mid R = 1, X, Z\} = \frac{p(Z \mid R = 0, X)}{p(Z \mid R = 1, X)}, \tag{5}$$

where $\widetilde{\mathrm{OR}}(X, Y) = \mathrm{OR}(X, Y)/E\{\mathrm{OR}(X, Y) \mid R = 1, X\}$,

$$p(R = 1 \mid X, Y) = \frac{p(R = 1 \mid X, Y = 0)}{p(R = 1 \mid X, Y = 0) + \mathrm{OR}(X, Y)p(R = 0 \mid X, Y = 0)},$$

$$p(Y \mid R = 0, X, Z) = \frac{\mathrm{OR}(X, Y)p(Y \mid R = 1, X, Z)}{E\{\mathrm{OR}(X, Y) \mid R = 1, X, Z\}}, \tag{5}$$

These identities reveal the central role of the odds ratio function in identification task: (2) shows how $p(R = 1 \mid X, Y)$, known as the propensity score, depends on the outcome through the odds ratio function; (2) shows that under the shadow variable assumption, the missing data distribution and thus the full data distribution can be recovered by integrating the odds ratio function with the complete-case distribution.

Identify (5) offers an essential equation for identification of $\mathrm{OR}(X, Y)$. With $p(Z \mid R = 0, X)$, $p(Z \mid R = 1, X)$ and $p(Y \mid R = 1, X, Z)$ obtained from the observed data, (5) is a Fredholm integral equation of the first kind, with $\widetilde{\mathrm{OR}}(X, Y)$ to be solved for. Because $\mathrm{OR}(X, Y) = \widetilde{\mathrm{OR}}(X, Y)/\widetilde{\mathrm{OR}}(X, Y = 0)$, identification of $\mathrm{OR}(X, Y)$ is implied by uniqueness of the solution to (5), which is guaranteed by a completeness condition.

Condition 1 (Completeness of $p(Y \mid R = 1, X, Z)$ in Z)**.** *For all square-integrable function* $h(X, Y)$*,* $E\{h(X, Y) \mid R = 1, X, Z\} = 0$ *almost surely if and only if* $h(X, Y) = 0$ *almost surely.*

The completeness condition is widely used in identification problems, such as in the instrumental variable identification [D'Haultfœuille 2011; Newey and Powell 2003]. The completeness condition we propose here only involves the observed data, which is advantageous in that in principle, it can be justified without extra model assumptions on the missing data distribution. We will return to the completeness condition later in this section after the following main identification result.

Theorem 1. *Under Assumptions 1, 2 and Condition 1, equation* (5) *has a unique solution, and thus the odds ratio function* $\mathrm{OR}(X, Y)$ *is identified. Therefore, the joint distribution* $p(X, Y, Z, R)$ *is identified.*

Theorem 1 shows how we achieve identification using a shadow variable: Assumption 1 results in equation (5) for the odds ratio function, and Condition 1 guarantees uniqueness of its solution. After identifying the odds ratio function, one can recover $p(Y \mid R = 0, X, Z)$ from (2) and then identify $p(Y, R \mid X, Z)$ and its functionals. In contrast to previous identification results derived under the outcome–selection factorization, we provide an alternative strategy to achieve identification for nonignorable missing data via the pattern-mixture factorization. The result characterizes the largest class of nonparametric models that are identifiable. The shadow variable is key to identification of the odds ratio function, without which, nonparametric identification is impossible because (5) is no longer available, and one has to resort to stringent parametric models such as Heckman's (1979) selection model or normal mixture models [Miao et al. 2016].

Our approach has the advantage that the identification Condition 1 can be justified with observed data. Although previous authors have described several identification conditions for the shadow variable setting, however, their various conditions are imposed either on the propensity score $p(R = 1 \mid X, Y)$, the full data distribution $p(Y \mid X, Z)$, or on both. Thus, their conditions involve missing values and cannot be justified empirically. For example, Wang et al. [2014] required monotonicity in the outcome of the propensity score and the full data likelihood ratio; Zhao and Shao [2015] considered a generalized linear model for the full data distribution; D'Haultfœuille [2010] required a completeness condition on the full data distribution. In contrast, our identification strategy only rests on completeness of the observed data distribution $p(Y \mid R = 1, X, Z)$, which does not involve missing values. As a result, under the shadow variable setting, identification or lack thereof can be assessed with only the observed data, a fact previously thought to be impossible.

Given a shadow variable Z, the completeness Condition 1 guarantees nonparametric identification of the odds ratio function. Completeness has been studied in various identification problems. Commonly-used parametric and semiparametric models such as exponential families and location-scale families satisfy the completeness condition. For a review and examples of completeness, see Newey and Powell [2003], D'Haultfœuille [2011], Hu and Shiu [2018] and the references therein. These previous results can be used as a basis to study completeness. Condition 1 implicitly requires that Z has a larger support than Y; for instance, if Y is categorical, then Z needs to have at least as many levels as Y. However, if the odds ratio function belongs to a parametric/semiparametric model class, the completeness condition can be weakened. We further illustrate the completeness condition with three examples.

Example 1 (Binary case)**.** *Consider binary Y and Z, then a saturated model for the odds ratio function can be parametrized as* $\mathrm{OR}(Y) = 1 + \gamma Y, \gamma > -1$*, and* (5) *implies that*

$$\frac{1 + \gamma E(Y \mid R = 1, Z = 1)}{1 + \gamma E(Y \mid R = 1)} = \frac{p(Z = 1 \mid R = 0)}{p(Z = 1 \mid R = 1)}.$$

If $Z \not\perp Y \mid R = 1$, then $p(Y \mid R = 1, Z)$ satisfies the completeness condition, and γ is identified by

$$\gamma = \frac{p(Z = 1 \mid R = 0) - p(Z = 1 \mid R = 1)}{p(Z = 1 \mid R = 1)E(Y \mid R = 1, Z = 1) - p(Z = 1 \mid R = 0)E(Y \mid R = 1)},$$

which is consistent with the result of Ma et al. [2003].

Example 2 (Exponential families)**.** *For continuous Y and Z, if*

$$p(Y \mid R = 1, X, Z) = s(X, Y)t(X, Z) \exp\{\mu(X, Z)^{\mathrm{T}} \tau(X, Y)\},$$

with $t(X, Z) > 0$, $s(X, Y) \geq 0$, $\tau(X, Y)$ one-to-one in Y, and the support of $\mu(X, Z)$ contains an open set, then completeness condition holds for $p(Y \mid R = 1, X, Z)$, as noted by Newey and Powell [2003].

Example 3 (Parametric odds ratio function). *Consider the case with binary Z and $Y \sim \text{Uniform}(0,1)$. The completeness Condition 1 is obviously not met, and thus $\text{OR}(Y)$ is not identifiable in nonparametric models. However, if the odds ratio function belongs to a parametric model $\text{OR}(Y;\gamma) = 1+\gamma Y, \gamma > -1$, then γ is identified as long as $Y \not\perp\!\!\!\perp Z \mid R=1$, which is testable.*

4 SEMIPARAMETRIC EFFICIENCY THEORY

In this section, we establish the semiparametric efficiency theory including estimation and inference about a pathwise differentiable functional of the full data law with the outcome MNAR by leveraging a shadow variable.

4.1 Brief introduction to the semiparametric efficiency theory

We shortly review the semiparametric efficiency theory, and refer to the monographs Bickel et al. [1993] and Tsiatis [2006] for detailed exposition of the theory. Consider a class of parametric models $p(X;\theta)$, with score function S_θ. For any differentiable parameter $\psi(\theta)$, the well-known Cramér-Rao lower bound for the covariance of unbiased estimator is

$$V_\psi = \nabla_\theta \psi(\theta) \left\{E(S_\theta S_\theta^{\mathrm{T}})\right\}^{-1} \nabla_\theta \psi(\theta)^{\mathrm{T}}.$$

Most reasonable estimators in a range of statistical inference problems are regular and asymptotically linear (RAL). Asymptotic linearity means

$$\hat{\psi} - \psi = n^{-1} \sum_{i=1}^{n} \phi\,(X_i) + o_p(n^{-1/2}),$$

where $\phi(X)$ is the *influence function* for ψ with $E\{\phi\,(X)\} = 0$ and $E\{\phi(X)\phi^{\mathrm{T}}(X)\} < \infty$ non-singular, and regularity refers to $n^{1/2}(\hat{\psi}-\psi)$ converging to a limiting distribution that does not depend on the local data generating process, which rules out superefficient estimators such as the Hodges' estimator. According to Tsiatis [2006], most common estimators are RAL, and the covariance of RAL estimators is lower bounded by V_ψ. When a finite-dimensional parameter ψ is of interest rather than the entire data distribution, one could adopt a range of flexible nonparametric models containing infinite-dimensional nuisance parameters instead of fully parametric models for nuisance parameters. And such models are called *semiparametric models* [Bickel et al. 1993]. To overcome the challenges caused by the infinite-dimensional parameters, a semiparametric model can be approximated by a sequence of parametric submodels which contain the true data generating process. Therefore, the semi-parametric efficiency bound, which is the lower bound of asymptotic variance for all RAL estimators, is defined by the supremum of the Cramér-Rao bounds for all parametric submodels. The corresponding influence function and estimator attaining the semiparametric efficiency bound are called the efficient influence function and efficient estimator, respectively.

4.2 The space of all influence functions

Let ψ be a full data functional defined as the solution to a given estimation equation $E\{U(X,Y,Z;\psi)\} = 0$; for instance, the outcome mean $\psi = E(Y)$ corresponds to $U(X,Y,Z;\psi) = Y-\psi$ and a p-dimensional regression coefficients ψ of $E(Y \mid X,Z;\psi)$ corresponds to $U(X,Y,Z;\psi) = h(X,Z)\{Y - E(Y \mid X,Z;\psi)\}$, where $h(\cdot)$ is a p-dimensional linearly independent function. For missing data problems, solving for ψ requires evaluation of $E\{U(X,Y,Z;\psi) \mid R,X,Z\}$ for both $R=0$ and 1. Although $E\{U(X,Y,Z;\psi) \mid R=0,X,Z\}$ cannot be evaluated directly from the observed data, it can be derived from the complete-case distribution $p(Y \mid R=1,X,Z)$ and the odds ratio function $\text{OR}(X,Y)$ according to (2). As a result, our inferential framework assumes a correctly specified odds ratio model $\text{OR}(X,Y;\gamma)$.

Let $p(Y, R \mid X, Z; \theta)$ denote a semiparametric or nonparametric model for the joint distribution of (Y, R) conditional on (X, Z), indexed by a possibly infinite-dimensional parameter θ. The parameter θ consists of two variationally independent components: $\theta = (\gamma, \eta)$, γ for the odds ratio model $\text{OR}(X, Y; \gamma)$ and η for the baseline regression and the baseline propensity score. We let $\text{NIF}(\psi, \theta)$ denote the full data influence function for ψ in the nonparametric model of $p(Y, R \mid X, Z)$ with $E\{\nabla_\psi \text{NIF}\} = -1$, for example, $\text{NIF}(\psi, \theta) = Y - \psi$ for $\psi = E(Y)$. For notational simplicity, we use $w = w(X, Y) = 1/p(R = 1 \mid X, Y)$ to denote the inverse probability weight. Let $\mathcal{H}^{(X,Z)}$ denote a generic Hilbert space consisting of all measurable vector functions $h(X, Z)$ of (X, Z) with finite variance equipped with the covariance inner product. Although semiparametric efficiency is well studied under MAR, it is more challenging for MNAR data. We consider the following model which allows for uncertainty of the odds ratio function with a shadow variable:

(i) the shadow variable Assumption 1, 2 and the completeness Condition 1 hold; and the odds ratio function follows a parametric model, i.e., $\text{OR}(X, Y, Z) = \text{OR}(X, Y; \gamma)$ with an unknown and finite dimensional parameter γ.

In (i), the baseline regression and the baseline propensity score remain nonparametric, and thus (i) in fact contains a large class of semiparametric models for the joint distribution. In previous work, Robins et al. [2000]; Rotnitzky and Robins [1997], and Vansteelandt et al. [2007] have studied semiparametric efficiency for MNAR data a–ssuming that the odds ratio $\text{OR}(X, Y, Z)$ is completely known. This model class does not impose the shadow variable assumption as the odds ratio and the baseline propensity score may depend on Z. However, this model is not entirely of interest because the exact odds ratio function is seldom known in practice. Model (i) is different from the semiparametric models of Zhao and Ma [2021] who requires a fully parametric model for $p(Y \mid X, Z)$ and leaves the propensity score $p(R = 1 \mid X, Y)$ nonparametric; it is also different from the model of Morikawa and Kim [2021] who considers a fully parametric propensity score model that in fact specifies parametric forms for both the odds ratio function $\text{OR}(X, Y)$ and the baseline propensity score $p(R = 1 \mid X, Y = 0)$.

To derive the set of influence functions for all regular and asymptotically linear (RAL) estimators of ψ assuming (i), we first consider the following submodel:

(i*) the shadow variable Assumption 1, 2 and the completeness Condition 1 hold; and the odds ratio function is completely known, i.e., $\text{OR}(X, Y, Z)$ equals a given function $\text{OR}(X, Y)$ for all (X, Y, Z);

Note that model (i) is a generalization of (i*) by allowing for unknown selection bias. We denote

$$\text{IF}_0(\psi, \theta) = wR \cdot \text{NIF}(\psi, \theta) + (1 - wR)E\{\text{NIF}(\psi, \theta) \mid R = 0, X\},$$

and for arbitrary $h \in \mathcal{H}^{(X,Z)}$, we denote

$$T(h; \theta) = (1 - wR)\{h - E(h \mid R = 0, X)\},$$
$$\text{IF}_1(h; \psi, \theta) = \text{IF}_0(\psi, \theta) + T(h; \theta).$$

One can verify that $\text{IF}_0(\psi, \theta)$ is in fact an observed data influence function for ψ under model (i*), i.e., when γ is known. In the Appendix, we show that the orthogonal complement to the observed data tangent space under (i*), denoted by $\mathcal{T}^\perp$, is

$$\mathcal{T}^\perp = \{T(h; \theta) : h \in \mathcal{H}^{(X,Z)}\};$$

and the space of all observed data influence functions for ψ under (i*) is

$$\{\text{IF}_1(h; \psi, \theta) : h \in \mathcal{H}^{(X,Z)}\}.$$

However, results derived under (i*) do not account for the uncertainty about the unknown odds ratio model. Under model (i) allowing for a parametric odds ratio model with unknown parameters, we have the following results.

Theorem 2. *Under model* (i) *and the regularity conditions described by Bickel et al. [1993], we have that*

(a) *the observed data score function of γ is*

$$S_\gamma = \{p(R=1 \mid X,Z) - R\}E\{\nabla_\gamma \log \mathrm{OR}(X,Y;\gamma) \mid R=0,X,Z\};$$

and the set of influence functions for all RAL estimators of γ is

$$\{\mathrm{IF}_\gamma(g;\theta) = [E\{T(g;\theta)S_\gamma^{\mathrm{T}}\}]^{-1} \cdot T(g;\theta) : T(g;\theta) \in \mathcal{T}^\perp\};$$

(b) *the set of influence functions for all RAL estimators of ψ is*

$$\left\{ \mathrm{IF}_2(g,h;\psi,\theta) = \mathrm{IF}_1(h;\psi,\theta) + E\{\nabla_\gamma \mathrm{IF}_1(h;\psi,\theta)\} \cdot \mathrm{IF}_\gamma(g;\theta) : g,h \in \mathcal{H}^{(X,Z)} \right\}.$$

Theorem 2 shows the impact of the odds ratio model on the influence functions of ψ. As a special case, when the odds ratio function is completely known as in (i) or (i*), we have $\mathrm{IF}_2(g,h;\psi,\theta) = \mathrm{IF}_1(h;\psi,\theta)$; if further the missingness is at random, i.e., $\mathrm{OR}(X,Y) = 1$ for all (X,Y), then $\mathrm{IF}_1(h;\psi,\theta)$ reduces to an influence function under MAR.

4.3 The efficient influence function

We let $\Pi(\cdot \mid \mathcal{T}^\perp)$ denote the orthogonal projection onto $\mathcal{T}^\perp$, the orthogonal complement to the observed data tangent space in model (i*). The following result gives the efficient influence function.

Theorem 3. *Under model* (i)*, we have that*

(a) *the efficient influence function for γ is*

$$\mathrm{EIF}_\gamma(\theta) = \{E(S_\gamma^{\mathrm{eff}}(S_\gamma^{\mathrm{eff}})^{\mathrm{T}})\}^{-1}S_\gamma^{\mathrm{eff}},$$

where $S_\gamma^{\mathrm{eff}} = \Pi(S_\gamma \mid \mathcal{T}^\perp)$ is the efficient score of γ;

(b) *the efficient influence function for ψ is*

$$\mathrm{EIF}_\psi(\psi,\theta) = \mathrm{IF}_1^{\mathrm{eff}}(\psi,\theta) + E\{\nabla_\gamma \mathrm{IF}_1^{\mathrm{eff}}(\psi,\theta)\} \cdot \mathrm{EIF}_\gamma(\theta),$$

with

$$\mathrm{IF}_1^{\mathrm{eff}}(\psi,\theta) = \mathrm{IF}_0(\psi,\theta) - \Pi\{\mathrm{IF}_0(\psi,\theta) \mid \mathcal{T}^\perp\}.$$

As shown in (b), $\mathrm{IF}_1^{\mathrm{eff}}(\psi,\theta)$ is in fact the efficient influence function of ψ in model (i*) where the odds ratio parameter γ is known; by taking account of the impact of estimating γ, which is captured by $E\{\nabla_\gamma \mathrm{IF}_1^{\mathrm{eff}}(\psi,\theta)\} \cdot \mathrm{EIF}_\gamma(\theta)$, we obtain the efficient influence function of ψ in model (i).

The efficient influence function involves the projection $\Pi(\cdot \mid \mathcal{T}^\perp)$, which is in general complicated. Nonetheless, we show that this is available in closed form as summarized below.

Theorem 4. *Under model* (i)*, any function of the observed data can be written as $m(RY,R,X,Z) = (1-R)m_0(X,Z) + R \cdot m_1(X,Y,Z)$, and we have that*

$$\Pi(m \mid \mathcal{T}^\perp) = (1-wR)\left\{K - \frac{Q \cdot E(K \mid R=0,X)}{E(Q \mid R=0,X)}\right\},$$

with

$$\begin{aligned} Q &= 1/E\{w \mid R=0,X,Z\}, \\ K &= Q \cdot E(m_0 - m_1 \mid R=0,X,Z). \end{aligned}$$

For illustration, in the Corollary 1 we derive the efficient influence function when both Y and Z are binary.

Corollary 1. *Consider binary Y and Z, then under model* (i)*, we have that*

$$S_\gamma^{\text{eff}} = (1 - wR)\{Z - E(Z \mid R = 0, X)\}\frac{(G_1 - G_0)\nabla_\gamma \log \text{OR}(X, Y = 1; \gamma)}{E(w \mid R = 0, X)},$$

with $G_z = E(Y \mid R = 0, X, Z = z)$ for $z = 0, 1$, and that

$$\Pi(\text{IF}_0 \mid \mathcal{T}^\perp) = (1 - wR)\{Z - E(Z \mid R = 0, X)\}\frac{H_1 - H_0}{E(w \mid R = 0, X)},$$

with $H_z = E[w\{E(\text{NIF} \mid R = 0, X) - \text{NIF}\} \mid R = 0, X, Z = z]$ for $z = 0, 1$.

4.4 The efficient estimation

Theorems 3–4 provide a theoretical efficiency bound for all regular and asymptotically linear estimators of ψ in model (i), and offer a closed form for the efficient influence function. Consider the union model $M_1 \cup M_2$ that assumes either (M_1) $p(Y, Z \mid R = 1, X; \beta)$ and $\text{OR}(X, Y; \gamma)$ are correctly specified, or (M_2) $p(R = 1 \mid Y = 0, X; \alpha)$ and $\text{OR}(X, Y; \gamma)$ are correctly specified. For nuisance parameter η, we refer to Wang et al. [2014] for an inverse probability weighted estimator $\hat{\alpha}$ and the Appendix for a regression based estimator $\hat{\beta}$. General results of Robins and Rotnitzky [2001] imply that in the aforementioned union model $M_1 \cup M_2$, EIF_ψ and EIF_γ are also the efficient influence functions for ψ and γ, respectively. Let $\hat{E}$ denote the empirical mean. It follows that $\hat{\gamma}^{\text{eff}}$, the solution to

$$\hat{E}\{\text{EIF}_\gamma(\hat{\gamma}^{\text{eff}}, \hat{\alpha}, \hat{\beta})\} = 0$$

and $\hat{\psi}^{\text{eff}}$ the solution to

$$\hat{E}\{\text{EIF}_\psi(\hat{\psi}^{\text{eff}}, \hat{\alpha}, \hat{\beta}, \hat{\gamma}^{\text{eff}})\} = 0$$

with $\hat{\alpha}, \hat{\beta}$ estimates of the nuisance parameters, are locally semiparametric efficient in model (i) at the intersection submodel $M_1 \cap M_2$; that is, $\hat{\gamma}_{\text{eff}}$ and $\hat{\psi}_{\text{eff}}$ attain the semiparametric efficiency bound for model (i) when both baseline models happen to hold.

Following from the general theory for estimating equations, the proposed estimators are also asymptotically normal under regularity conditions described by Newey and McFadden [1994], which we do not replicate. The variance estimate for the efficient estimator can be obtained by the sandwich estimator. Based on normal approximations, confidence intervals can be constructed.

Under the union model, the efficient estimator can also be obtained based on an initial doubly robust $n^{1/2}$-consistent estimator $(\hat{\psi}_{\text{dr}}, \hat{\gamma}_{\text{dr}})$ by a one-step construction following Bickel et al. [1993],

$$\hat{\gamma}_{\text{eff}} = \hat{\gamma}_{\text{dr}} + \hat{E}\{\text{EIF}_\gamma(\hat{\alpha}, \hat{\beta}, \hat{\gamma}_{\text{dr}})\},$$

$$\hat{\psi}_{\text{eff}} = \hat{\psi}_{\text{dr}} + \hat{E}\{\text{EIF}(\hat{\psi}_{\text{dr}}, \hat{\alpha}, \hat{\beta}, \hat{\gamma}_{\text{dr}})\}.$$

The above approach requires a doubly robust estimation of γ, ψ, which has been proposed by Miao and Tchetgen Tchetgen [2016] under the shadow variable setting.

5 NUMERICAL EXAMPLES

5.1 Simulations

We study the finite sample performance of the proposed methods on estimation of the outcome mean $\psi = E(Y)$ via simulations. We generate a covariate $X \sim N(0, 1)$, and then generate (Y, Z, R) with a

normal model for the baseline outcome distribution, a logistic model for the baseline propensity score, and odds ratio function. We consider the following four types of generated datasets:

$$\begin{aligned}
M_1 :&\mathrm{OR}(X,Y)=\exp(0.5Y), \quad \text{logit } p(R=1\mid Y=0,X)=0.5+0.4X,\\
&Y\mid R=1,X,Z\sim N(X+Z,1), \quad Z\mid R=1,X\sim N(-0.4X^2,1);\\
M_2 :&\mathrm{OR}(X,Y)=\exp(0.5Y), \quad \text{logit } p(R=1\mid Y=0,X)=X+1.5X^2,\\
&Y\mid R=1,X,Z\sim N(X+Z,1), \quad Z\mid R=1,X\sim N(-0.4X^2,1);\\
M_3 :&\mathrm{OR}(X,Y)=\exp(0.5Y), \quad \text{logit } p(R=1\mid Y=0,X)=0.5+0.4X,\\
&Y\mid R=1,X,Z\sim N(X-0.3X^2+Z,1), \quad Z\mid R=1,X\sim N(X-0.4X^2,1);\\
M_4 :&\mathrm{OR}(X,Y)=\exp(-0.3Y), \quad \text{logit } p(R=1\mid Y=0,X)=0.5+0.4X+X^2,\\
&Y\mid R=1,X,Z\sim N(X+0.3X^2+Z,1), \quad Z\mid R=1,X\sim N(X-0.4X^2,1).
\end{aligned}$$

For these settings, the missing data proportions are around 35%. We generate data from the four combinations of the baseline models, but employ a simpler model for estimation:

$$\mathrm{OR}(X,Y)=\exp(-\gamma Y), \quad \text{logit } p(R=1\mid X,Y=0)=\alpha_0+\alpha_1 X,$$

$$Y\mid R=1,X,Z\sim N(\beta_{10}+\beta_{11}X+\beta_{12}Z,\sigma_1^2), \quad Z\mid R=1,X\sim N(\beta_{20}+\beta_{21}X^2,\sigma_2^2).$$

Then M_3 has a correct baseline propensity score model but an incorrect baseline outcome model, which we denoted as TF in the following. And the other three situations are similarly defined. We apply the efficient estimator (EFF) and compare it with a doubly robust but not efficient estimator from Miao and Tchetgen Tchetgen [2016], a regression based estimator, an inverse probability weighted estimator from Wang et al. [2014] (IPW) as well as a naive estimator (MAR) assuming MAR obtained by linear regression on complete cases. We simulate 1000 replicates under 500 and 1000 sample sizes for each combination and summarize the results with boxplots.

Figure 2 presents the results for the outcome mean, and Figure 3 for the odds ratio parameter. Table 1 reports coverage probability of the 0.95 confidence interval estimated with the method in the Appendix. In FT of Figure 2, the baseline propensity score is incorrect but the baseline outcome model is correct. As a result, the inverse probability weighted estimator is biased with coverage probability below the nominal level. In TF, the baseline propensity score is correct but the baseline outcome model is incorrect, thus the regression based estimator is biased. In TT, both models are correct, the efficient estimator performs best among all estimators in terms of variance. In FF, neither of the two models is correct and all four methods are biased. In summary, bias of the efficient estimator is small whenever at least one model is correctly specified. We also observe that as expected, the naive estimator assuming MAR is biased in all cases. The performance of the estimators for the odds ratio parameter is analogous. As a conclusion, we recommend the efficient approach for inference about the mean and odds ratio parameter as well as to evaluate the magnitude of selection bias.

5.2 A home pricing example

We apply the proposed method to a home pricing dataset extracted from the China Family Panel Studies. The dataset was collected from 3126 households in China. The outcome of interest is log of current home price (in 10^4 RMB yuan), of which 596 (21.8%) values are missing, because the house owner does not respond in the survey, nor is the price available from the real estate market. Completely available covariates include log of construction price, province, urban (1 for urban household, 0 rural), travel time to the nearest business center, house building area, family size, house story height, log of family income, and refurbish status.

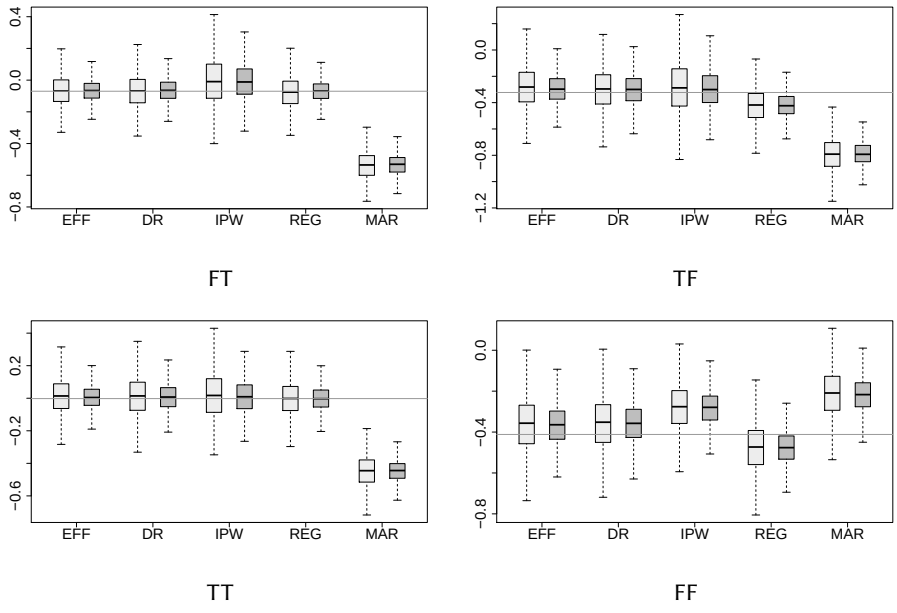

Fig. 2. Boxplots of estimators of the outcome mean. White boxes are for sample size 500, and gray ones for 1000. The horizontal line marks the true value of the parameter.

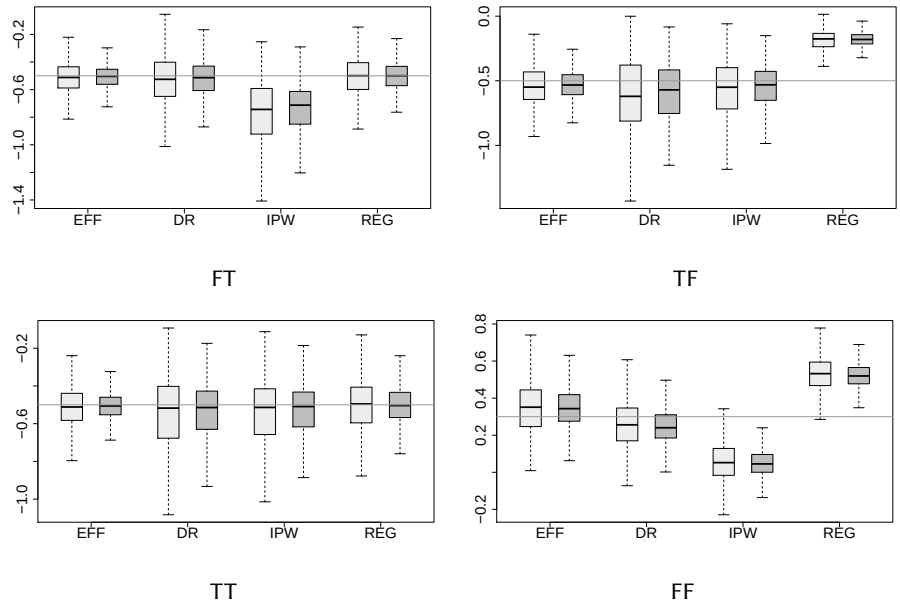

Fig. 3. Boxplots of estimators of the odds ratio parameter. White boxes are for sample size 500, and gray ones for 1000. The horizontal line marks the true value of the parameter.

Table 1. Coverage probability of 0.95 confidence interval

	Outcome mean (ψ)				Odds ratio parameter (γ)			
	EFF	DR	IPW	REG	EFF	DR	IPW	REG
FT	0.962	0.955	0.918	0.957	0.959	0.951	0.907	0.943
	0.954	0.959	0.917	0.952	0.941	0.941	0.737	0.942
TF	0.965	0.965	0.925	0.89	0.964	0.936	0.951	0.034
	0.976	0.974	0.953	0.817	0.958	0.949	0.947	0.002
TT	0.963	0.937	0.924	0.937	0.964	0.944	0.941	0.925
	0.952	0.941	0.935	0.944	0.965	0.948	0.946	0.955
FF	0.936	0.931	0.901	0.867	0.957	0.911	0.724	0.032
	0.911	0.891	0.843	0.672	0.925	0.772	0.457	0.003

Note: Confidence intervals are obtained with the sandwich variance estimator. The result of each situation includes two rows, of which the first stands for sample size 500, and the second for 1000.

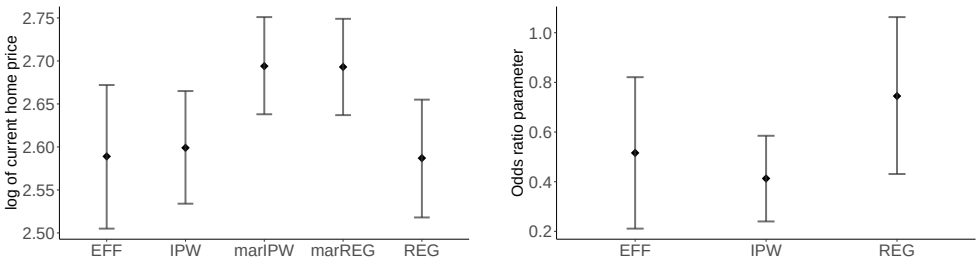

Fig. 4. Point estimates and 95% confidence intervals of log of current home price and the odds ratio parameter γ in Home Pricing example. *marREG* and *marIPW* respectively stand for standard regressi on estimation and inverse probability weighted estimation under MAR.

The construction price of a house is related to the current price, however, we expect that it is independent of nonresponse conditional on the current price and fully observed covariates. Therefore, we use log of construction price as a shadow variable Z. Let X denote the vector of all other covariates including the intercept, we assume the following models,

$$\mathrm{OR}(X,Y)=\exp(-\gamma Y), \qquad \text{logit}\, p(R=1\mid X,Y=0)=X^{\mathrm{T}}\alpha,$$
$$Y\mid R=1,X,Z\sim N\left\{(X^{\mathrm{T}},Z)\beta_1,\sigma_1^2\right\}, \quad Z\mid R=1,X\sim N(X^{\mathrm{T}}\beta_2,\sigma_2^2).$$

We summarize estimates of the outcome mean and the odds ratio model in Figure 4.

Estimates for the odds ratio parameter produced by the proposed methods depart significantly from zero, providing empirical evidence of selection bias due to missingness and showing potential bias of standard estimation methods that assume MAR. The proposed method results in slightly lower estimates of home price on the log scale than those obtained by standard methods assuming MAR; however, the deviation is more notable on the original scale and amount to significant bias equal to 1.46×10^4 RMB yuan.

6 DISCUSSION

We have developed a general semiparametric framework for identification and inference about any functional of the full data law in the presence of nonignorable missing outcome data with the aid of a shadow variable. Under certain completeness condition, we describe the largest class of nonparametric models that are identifiable by the approach. Our approach reveals the central role of the odds ratio function and the shadow variable in identification of full data distribution. The identification conditions we propose only involve the observed data, and thus can be justified empirically. Our identification results establish the basis for statistical inference in both this paper and a early published companion paper [Miao and Tchetgen Tchetgen 2016], which builds directly on a prior draft of the current manuscript. When the shadow variable Assumption 1 does not hold, the odds ratio function is in general not identified, and one can conduct sensitivity analysis to check how results would change according to the impact of the shadow variable. We refer to Robins et al. [2000] for details for sensitivity analysis. The proposed identification, estimation, and semiparametric efficiency theory readily extends to missing covariate problems considered by Miao and Tchetgen Tchetgen [2018] and Yang et al. [2019], who employ a shadow variable identifying condition, however do not provide a framework for semiparametric inference. The proposed methods can also be extended to longitudinal data analysis, which is often subject to dropout or missing data. Their potential use for such complicated settings will be studied elsewhere.

APPENDIX

Proof of Theorem 1. Under the shadow variable Assumption 1, from Proposition 1 we have

$$E\{\widetilde{\mathrm{OR}}(X,Y) \mid R=1,X,Z\} = \frac{p(Z \mid R=0,X)}{p(Z \mid R=1,X)}, \tag{A.1}$$

$$\widetilde{\mathrm{OR}}(X,Y) = \frac{\mathrm{OR}(X,Y)}{E\{\mathrm{OR}(X,Y) \mid R=1,X\}}.$$

Based on these two equalities, we prove identification of $\mathrm{OR}(X,Y)$ under Assumption 1. Because $p(Y \mid R=1,X,Z)$ and $p(Z \mid R=1,X)$ can be obtained from the observed data, for any candidate of $\mathrm{OR}(X,Y)$, $E\{\widetilde{\mathrm{OR}}(X,Y) \mid R=1,X,Z\}$ can be computed from the observed data. Suppose $\mathrm{OR}^*(X,Y)$ is the truth and $\mathrm{OR}'(X,Y)$ is a candidate that

$$E\{\widetilde{\mathrm{OR}}'(X,Y) \mid R=1,X,Z\} = \frac{p(Z \mid R=0,X)}{p(Z \mid R=1,X)}.$$

We have

$$E\{\widetilde{\mathrm{OR}}'(X,Y) - \widetilde{\mathrm{OR}}^*(X,Y) \mid R=1,X,Z\} = 0,$$

which together with Condition 1 implies that $\widetilde{\mathrm{OR}}'(X,Y) = \widetilde{\mathrm{OR}}^*(X,Y)$. Therefore, (A.1) must have a unique solution, that is, $\widetilde{\mathrm{OR}}(X,Y)$ is identified and hence $\mathrm{OR}(X,Y)$ is identified by $\mathrm{OR}(X,Y) = \widetilde{\mathrm{OR}}(X,Y)/\widetilde{\mathrm{OR}}(X,Y=0)$. □

We need the following lemma to prove Theorem 2. We derive the ortho-tangent space $\mathcal{T}^{\perp}$ under Assumption 1 with a completely known odds ratio function, i.e.,

(i*) the shadow variable Assumption 1 and the completeness Condition 1 hold; and the odds ratio function is completely known, i.e., $\mathrm{OR}(X,Y,Z) = \mathrm{OR}(X,Y)$ for a given function $\mathrm{OR}(X,Y)$ for all (X,Y,Z).

Note that in model (i*), the full data model $p(L)$ is nonparametric.

Lemma A.1. *Under model* (i*)*, the ortho-complement to the observed data tangent space is*

$$\mathcal{T}^{\perp} = \left\{ \; T(h;\theta) : h = h(X,Z) \in \mathcal{H}^{(X,Z)} \; \right\}, \tag{A.2}$$

with

$$T(h;\theta) = \{1 - wR\}\{h - E(h \mid R = 0, X)\}.$$

Before we prove Lemma A.1, we review some preliminary results. Let $L = (X, Y, Z)$ and $O = (X, RY, Z, R)$ denote the full data and observed data, respectively. Suppose we are interested in a full data functional ψ solving $E\{U(X, Y, Z; \psi)\} = 0$. We let $\mathcal{T}_1^{\mathrm{F}}$ and $\mathcal{T}_2^{\mathrm{F}}$ denote the full data tangent space of $p(L)$ and the propensity score model $\pi(L) = p(R = 1 \mid L)$, respectively. The full data tangent space for $p(L, R)$ is $\mathcal{T}^{\mathrm{F}} = \mathcal{T}_1^{\mathrm{F}} \oplus \mathcal{T}_2^{\mathrm{F}}$, where $\oplus$ denotes direct summation of spaces. Let $\mathcal{T}_1$, $\mathcal{T}_2$, and $\mathcal{T}$ denote the observed data tangent space of $p(L)$, $\pi(L)$, and $p(L, R)$, respectively.

Rotnitzky and Robins [1997] and Robins et al. [2000] showed that $\mathcal{T} = \overline{\mathcal{T}_1 + \mathcal{T}_2}$, with $\overline{\mathcal{S}}$ denoting the close linear span of the set $\mathcal{S}$. Let $\mathcal{H}^{(L,R)}$, $\mathcal{H}^{(O)}$, and $\mathcal{H}^{(X,Z)}$ denote the respective Hilbert spaces of all measurable functions of (L, R), O, and (X, Z) with finite variance equipped with covariance inner product. Robins et al. [2000] established that

$$\mathcal{T}_2^{\perp} = \{b(O) : b(O) \in (\mathcal{T}_2^{\mathrm{F}})^{\perp}\}. \tag{A.3}$$

and Rotnitzky and Robins [1997] derived the following lemma.

Lemma A.2 (Rotnitzky and Robins [1997])**.** *Assume that the full data model $p(L)$ is nonparametric, and that the propensity score follows an arbitrary semiparameteric model $p(R \mid L; \alpha)$ indexed by a possibly infinite dimensional parameter α. Then*

$$\mathcal{T}_1^{\perp} = \{T_1(a) = (1 - R)a - wR \cdot E\{(1 - R)a \mid L\} : a = a(O) \in \mathcal{H}^{(O)}\},$$

with $w = 1/\pi(L)$ the inverse probability weight.

Proof of Lemma A.1. In model (i*), we wish to characterize the ortho-complement to the tangent space in the observed data model, that is,

$$\mathcal{T}^{\perp} = \mathcal{T}_1^{\perp} \cap \mathcal{T}_2^{\perp}.$$

Under the shadow variable Assumption 1, we have that $\pi(L) = \pi(X, Y)$, and thus the inverse probability weight $w = 1/\pi(L)$ does not depend on Z. Applying Lemma A.2, the ortho-complement of the observed data tangent space for $p(L)$ is

$$\mathcal{T}_1^{\perp} = \{T_1(a) = (1 - R)a - wR \cdot E\{(1 - R)a \mid L\} : a = a(O) \in \mathcal{H}^{(O)}\}.$$

Additionally, under model (i*), we can verify that

$$\mathcal{T}_2^{\mathrm{F}} = \{T_2(g) = \{R - \pi\}g \text{ for all } g = g(X) \in \mathcal{H}^{(X,Z)}\}.$$

Note that any element in $\mathcal{T}_1^{\perp}$ is a function of the observed data, from (A.3), we have that

$$\mathcal{T}^{\perp} = \mathcal{T}_1^{\perp} \cap \mathcal{T}_2^{\perp} = \mathcal{T}_1^{\perp} \cap (\mathcal{T}_2^{\mathrm{F}})^{\perp} = \{T_1 \in \mathcal{T}_1^{\perp} : E(T_1 T_2) = 0 : T_2 \in \mathcal{T}_2^{\mathrm{F}}\}.$$

Thus, the elements of $\mathcal{T}^{\perp}$ can be expressed as

$$T(a) = (1 - R)a - wR \cdot E\{(1 - R)a \mid L\}, \tag{A.4}$$

with $a = a(O) \in \mathcal{H}^{(O)}$ such that

$$\begin{aligned}
& 0 = E[T(a)T_2(g)] : T_2(g) \in \mathcal{T}_2^{\mathrm{F}}, \\
\Leftrightarrow \quad & 0 = E[\{(1-R)a - wR \cdot E((1-R)a \mid L)\}\{R-\pi\}g] \text{ for all } g = g(X) \in \mathcal{H}^{(X,Z)}, \\
\Leftrightarrow \quad & 0 = E[\{(1-R)a - wR \cdot E((1-R)a \mid L)\}\{R-\pi\} \mid X], \\
\Leftrightarrow \quad & 0 = E\{(1-R)a \mid X\}, \\
& \text{upon writing } a(O) = a_1(L)R + a_2(X,Z)(1-R), \\
\Leftrightarrow \quad & 0 = E\{(1-R)(a_1R + a_2(1-R)) \mid X\}, \\
\Leftrightarrow \quad & 0 = E\{(1-R)a_2 \mid X\}, \\
\Leftrightarrow \quad & 0 = E\{a_2 \mid R=0, X\}.
\end{aligned}$$

Hence, $a_2(X,Z)$ can be written as $a_2(X,Z) = h - E(h \mid R=0,X)$ for an arbitrary function $h = h(X,Z) \in \mathcal{H}^{(X,Z)}$, and (A.4) can be written as $T(h) = \{1-wR\}\{h - E(h \mid R=0,X)\}$. As a result, we have that

$$\mathcal{T}^{\perp} = \{T(h) = \{1-wR\}\{h - E(h \mid R=0,X)\} : h = h(Z,X) \in \mathcal{H}^{(X,Z)}\}.$$

□

Let $\mathrm{NIF}(\psi,\theta)$ denote the full data influence function of ψ in the nonparametric model $p(X,Y,Z;\theta)$. One can verify that

$$\mathrm{IF}_0(\psi,\theta) = wR \cdot \mathrm{NIF}(\psi,\theta) + (1-wR)E\{\mathrm{NIF}(\psi,\theta) \mid R=0,X\},$$

is an observed data influence function for ψ in model (i*), then according to Newey [1994] we have the set of all observed data influence functions under (i*), which is $\mathrm{IF}_0(\psi,\theta) + \mathcal{T}^{\perp}$.

Corollary 2. *In model* (i*)*, the set of influence functions for all RAL estimators of* ψ *is* $\mathrm{IF}_0(\psi,\theta) + \mathcal{T}^{\perp}$*, i.e.,*

$$\left\{ \ \mathrm{IF}_1(h;\psi,\theta) = \mathrm{IF}_0(\psi,\theta) + T(h;\theta) : h = h(X,Z) \in \mathcal{H}^{(X,Z)}. \ \right\}$$

Proof of Theorem 2. We prove that the results hold within all parametric submodels of the semiparametric model, and then the results hold for the semiparametric model by aggregating all submodels. Consider submodel $p(Y,R \mid X,Z;\theta_t)$ indexed by t, i.e., a path in the semiparametric model (i), with $\theta_t = (\gamma_t,\eta_t)$ and θ_0 equal to the true value of θ. We let S_t denote the observed data score function in the submodel; we use $\Pi(\cdot \mid \mathcal{T}^{\perp})$ to denote the projection onto $\mathcal{T}^{\perp}$.

(a) We first derive the observed data score function S_γ. The full data likelihood $pr(Y,R \mid X,Z;\gamma)$ can be written as

$$\frac{p(R \mid X, Y=0)p(Y \mid R=1,X,Z)\mathrm{OR}(X,Y;\gamma)^{1-R}}{\int p(R \mid X, Y=0)p(Y \mid R=1,X,Z)\mathrm{OR}(X,Y;\gamma)^{1-R}dRdY},$$

and the observed data likelihood is

$$\{p(Y,R=1 \mid X,Z;\gamma)\}^R\{p(R=0 \mid X,Z;\gamma)\}^{1-R};$$

then the full data score function of γ is

$$S_\gamma^{\mathrm{F}} = (1-R)\nabla_\gamma \log \mathrm{OR}(X,Y;\gamma) - E\{(1-R)\nabla_\gamma \log \mathrm{OR}(X,Y;\gamma) \mid X,Z\},$$

and the observed data score function of γ is

$$S_\gamma = R \cdot S_\gamma^{\mathrm{F}} + (1-R)E\{S_\gamma^{\mathrm{F}} \mid R=0,X,Z\}.$$

After some algebra, we can verify that

$$S_\gamma = \{p(R=1 \mid X,Z) - R\}E\{\nabla_\gamma \log \mathrm{OR}(X,Y;\gamma) \mid R=0,X,Z\}.$$

Next, following from the fact that the orthogonal complement to the nuisance tangent space under model (i) is exactly the space $\mathcal{T}^\perp$, and therefore from Tsiatis [2006, Theorem 4.2], the space of influence functions for all RAL estimator for γ is

$$\{\mathrm{IF}_\gamma(g;\theta) = [E\{T(g;\theta)S_\gamma^{\mathrm{T}}\}]^{-1}T(g;\theta) : T(g;\theta) \in \mathcal{T}^\perp\}. \tag{A.5}$$

(b) For any t and $h = h(X,Z)$, we let ψ_t denote the solution to

$$E_t\{\mathrm{IF}_1(h;\psi_t,\theta_t)\} = 0,$$

where E_t denotes expectation with respect to $p(Y,R \mid X,Z;\theta_t)$. Therefore, we have that

$$\begin{aligned} 0 &= \nabla_t E_t\{\mathrm{IF}_1(h;\psi_t,\theta_t)\} \\ &= E\{\mathrm{IF}_1(h;\psi,\theta)S_t\} + E\{\nabla_t \mathrm{IF}_1(h;\psi_t,\theta_t)\} \\ &= E\{\mathrm{IF}_1(h;\psi,\theta)S_t\} + E\{\nabla_\psi \mathrm{IF}_1(h;\psi,\theta)\}\nabla_t\psi_t \\ &\quad + E\{\nabla_\gamma \mathrm{IF}_1(h;\psi,\theta)\}\nabla_t\gamma_t + E\{\nabla_\eta \mathrm{IF}_1(h;\psi,\theta)\}\nabla_t\eta_t. \end{aligned} \tag{A.6}$$

In order to derive the form of influence functions for ψ under model (i), we prove that $E\{\nabla_\eta \mathrm{IF}_1(h;\psi,\theta)\} = 0$ by separately showing that $E\{\nabla_\eta \mathrm{IF}_0(h,\psi,\theta)\} = 0$ and that $E\{\nabla_\eta T(h;\theta)\} = 0$ for all $h = h(X,Z)$. Let η_i denote the ith component of η and η_{-i} the others. A similar argument to Miao and Tchetgen Tchetgen [2016, Lemma A1] indicates double robustness of $\mathrm{IF}_0(h;\psi,\theta)$ against misspecification of the baseline model parameters η, that is, for all $\eta_i + \delta_i$ in an open neighborhood of η_i, $E\{\mathrm{IF}_0(h;\psi,\gamma,\eta_i+\delta_i,\eta_{-i})\} = 0$. We thus have

$$\begin{aligned} &E\{\nabla_{\eta_i}\mathrm{IF}_0(h;\psi,\theta)\} \\ &= E_\theta\left\{\lim_{\delta_i\to 0}\frac{\mathrm{IF}_0(h;\psi,\gamma,\eta_i+\delta_i,\eta_{-i}) - \mathrm{IF}_0(h;\psi,\gamma,\eta_i,\eta_{-i})}{\delta_i}\right\} \\ &= \lim_{\delta_i\to 0} E\left\{\frac{\mathrm{IF}_0(h;\psi,\gamma,\eta_i+\delta_i,\eta_{-i}) - \mathrm{IF}_0(h;\psi,\gamma,\eta_i,\eta_{-i})}{\delta_i}\right\} = 0. \end{aligned}$$

Therefore, we have $E\{\nabla_\eta \mathrm{IF}_0(h;\psi,\theta)\} = 0$.
Given γ, Lemma A.1 implies that $E\{T(h;\theta)S_\eta\} = 0$ for any $T(h;\theta) \in \mathcal{T}^\perp$. Thus, $E\{\nabla_\eta T(h;\theta)\} = -E\{T(h;\theta)S_\eta\} = 0$, and as a result,

$$E\{\nabla_\eta \mathrm{IF}_1(h;\psi,\theta)\} = 0. \tag{A.7}$$

In addition, because for any h, $\mathrm{IF}_1(h;\psi,\theta)$ is an influence function for ψ when γ is known, we have that

$$E\{\nabla_\psi \mathrm{IF}_1(h;\psi,\theta)\} = -1. \tag{A.8}$$

Newey [1994] shows that for any influence function IF_γ of γ,

$$\nabla_t\gamma_t = E(\mathrm{IF}_\gamma S_t). \tag{A.9}$$

From (A.6)–(A.9), we have

$$\nabla_t\psi_t = E[\{\mathrm{IF}_1(h;\psi,\theta) + E(\nabla_\gamma \mathrm{IF}_1(h;\psi,\theta)) \cdot \mathrm{IF}_\gamma(g;\theta)\}S_t],$$

which implies from Newey [1994] that for any h and $g \in \mathcal{H}^{(X,Z)}$,

$$\mathrm{IF}_2(h,g;\psi,\theta) = \mathrm{IF}_1(h;\psi,\theta) + E\{\nabla_\gamma \mathrm{IF}_1(h;\psi,\theta)\} \cdot \mathrm{IF}_\gamma(g;\theta) \tag{A.10}$$

is an influence function for ψ in model (i).

In fact, (A.10) represents all influence functions for ψ in model (i) as we demonstrate below. Given any $h_0(X,Z), g_0(X,Z)$, Newey [1994] implies that the following linear variety is the set of all influence functions for ψ assuming (i),

$$\mathrm{IF}_2(h_0, g_0; \psi, \theta) + \text{ ortho-complement to the tangent space assuming (i).}$$

Moreover, the ortho-complement to the tangent space under model (i) can be represented as $\{T(h;\theta) \in \mathcal{T}^{\perp} : E\{T(h;\theta)\cdot S_\gamma\} = 0\}$, which is equivalent to

$$\{T(h;\theta) \in \mathcal{T}^{\perp} : E\{\nabla_\gamma T(h;\theta)\} = 0\},$$

by noting that $E\{\nabla_\gamma T(h;\theta)\} = -E\{T(h;\theta)\cdot S_\gamma\}$. Therefore, the space of all influence functions for ψ assuming (i) is

$$\{\mathrm{IF}_2(h_0, g_0; \psi, \theta) + T(h;\theta) : T(h;\theta) \in \mathcal{T}^{\perp}, E\{\nabla_\gamma T(h;\theta)\} = 0\},$$

that is,

$$\begin{aligned}
& \mathrm{IF}_2(h_0, g_0; \psi, \theta) + T(h;\theta) \\
= \;& \mathrm{IF}_1(h_0; \psi, \theta) + E\{\nabla_\gamma \mathrm{IF}_1(h_0; \psi, \theta)\}\cdot \mathrm{IF}_\gamma(g_0;\theta) + T(h;\theta) \\
= \;& \mathrm{IF}_1(h_0 + h; \psi, \theta) + E\{\nabla_\gamma \mathrm{IF}_1(h_0; \psi, \theta)\}\cdot \mathrm{IF}_\gamma(g_0;\theta) \\
= \;& \mathrm{IF}_1(h_0 + h; \psi, \theta) + E\{\nabla_\gamma \mathrm{IF}_1(h_0 + h; \psi, \theta)\}\cdot \mathrm{IF}_\gamma(g_0;\theta) \\
= \;& \mathrm{IF}_2(h_0 + h, g_0; \psi, \theta).
\end{aligned}$$

As a result, any influence function for ψ assuming (i) can be represented in the form of (A.10).

□

Proof of Theorem 3. (a) This is implied from the result of Tsiatis [2006, Theorem 4.2] that $\mathrm{EIF}_\gamma(\theta) = \{E(S_\gamma^{\mathrm{eff}}(S_\gamma^{\mathrm{eff}})^{\mathrm{T}})\}^{-1} S_\gamma^{\mathrm{eff}}$, with $S_\gamma^{\mathrm{eff}} = \Pi(S_\gamma \mid \mathcal{T}^{\perp})$.

(b) To derive the efficient influence function for ψ, we choose g and h such that $\mathrm{IF}_2(g, h; \psi, \theta)$ falls in the observed data tangent space under model (i). Because $\Pi(\mathrm{IF}_0 \mid \mathcal{T}^{\perp}) \in \mathcal{T}^{\perp}$, there exists $h^{\mathrm{eff}}(X,Z)$ such that $T(h^{\mathrm{eff}}) = -\Pi(\mathrm{IF}_0 \mid \mathcal{T}^{\perp})$, and we let $\mathrm{IF}_1^{\mathrm{eff}} = \mathrm{IF}_0 + T(h^{\mathrm{eff}}) = \Pi(\mathrm{IF}_0 \mid \mathcal{T})$. We further choose $g^{\mathrm{eff}}(X,Z)$ such that $\mathrm{EIF}_\gamma = T(g^{\mathrm{eff}})$ is the efficient influence function for γ. Then we have that

$$\begin{aligned}
\mathrm{EIF}_\psi &= \mathrm{IF}_1^{\mathrm{eff}}(\psi, \theta) + E\{\nabla_\gamma \mathrm{IF}_1^{\mathrm{eff}}(\psi, \theta)\}\cdot \mathrm{EIF}_\gamma \\
&= \Pi(\mathrm{IF}_0 \mid \mathcal{T}) + E\{\nabla_\gamma \Pi(\mathrm{IF}_0 \mid \mathcal{T})\}\cdot T(g^{\mathrm{eff}}).
\end{aligned}$$

Note that $\mathcal{T}$ is the observed data tangent space assuming (i*), and it is contained in the observed data tangent space assuming (i). Hence, $T(g^{\mathrm{eff}})$ and $\Pi(\mathrm{IF}_0 \mid \mathcal{T})$ belong to the latter space and so does EIF_ψ. Therefore, EIF_ψ is the efficient influence function for ψ.

□

Proof of Theorem 4. Consider the space $\mathcal{T}^{\perp} = \{T(h) : h = h(X,Z) \in \mathcal{H}^{(X,Z)}\}$, with

$$\begin{aligned}
T(h) &= \{1 - wR\}\{h - E[h \mid R = 0, X]\} \qquad (A.11)\\
&= \{(1-R) - R(w-1)\}\{h - E(h \mid R = 0, X)\}.
\end{aligned}$$

We show how to project onto the space $\mathcal{T}^{\perp}$, that is, we wish to find $T(h^*) = \Pi(m \mid \mathcal{T}^{\perp})$ for any function $m = m(RY, R, X, Z)$ of the observed data. First note that for any m, there exist a function m_0 of (X,Z) and m_1 of (X,Y,Z), such that $m(RY,R,X,Z) = (1-R)m_0(X,Z) + Rm_1(X,Y,Z)$. We therefore wish to find $h^* = h^*(X,Z)$ that solves

$$E[\{m - T(h^*)\}T(h)] = 0 \text{ for all } h = h(X,Z) \in \mathcal{H}^{X,Z}. \qquad (A.12)$$

For any $h = h(X, Z)$, letting $\Delta(h) = h - E(h \mid X, R = 0)$, we have that

$$\begin{aligned} 0 &= E[\{m - T(h^*)\}T(h)] \\ &= E\left[\begin{array}{c} \{(1-R)m_0 + Rm_1 - ((1-R) - R(w-1))\Delta(h^*)\} \\ \cdot\{(1-R) - R(w-1)\}\Delta(h) \end{array}\right] \\ &= E\left\{\begin{array}{c} (1-R)m_0\Delta(h) - m_1R(w-1)\cdot\Delta(h) - (1-R)\Delta(h)\Delta(h^*) \\ -R(w-1)^2\Delta(h^*)\Delta(h) \end{array}\right\}. \end{aligned}$$

By $R(w-1) = 1 - R - (1 - wR)$ and $E[\{wR - 1\}g(X, Y, X)] = 0$, we have

$$\begin{aligned} 0 &= E\left\{\begin{array}{c} (1-R)m_0\Delta(h) - (1-R)m_1\Delta(h) - (1-R)\Delta(h^*)\Delta(h) \\ -(1-R)(w-1)\cdot\Delta(h^*)\Delta(h) \end{array}\right\} \\ &= E[\{m_0 - m_1 - w\Delta(h^*)\}\cdot\{(1-R)\Delta(h)\}] \\ &= E\left[E\{m_0 - m_1 - w\Delta(h^*) \mid R = 0, X, Z\}\cdot\{(1-R)\Delta(h)\}\right] \\ &= E\left[\Delta(E\{m_0 - m_1 - w\Delta(h^*) \mid R = 0, X, Z\})\cdot\{(1-R)\Delta(h)\}\right], \end{aligned}$$

and by letting $h = E\{m_0 - m_1 - w\Delta(h^*) \mid R = 0, X, Z\}$, we conclude that

$$\begin{aligned} 0 &= \Delta(E\{m_0 - m_1 - w\Delta(h^*) \mid R = 0, X, Z\}) \\ &= \Delta(E(m_0 - m_1 \mid R = 0, X, Z)) - \Delta(h^*)E(w \mid R = 0, X, Z) \\ &\quad + E\{\Delta(h^*)E(w \mid R = 0, X, Z) \mid R = 0, X\}. \end{aligned}$$

Letting

$$\begin{aligned} Q &= Q(X, Z) = 1/E\{w(X, Y) \mid R = 0, X, Z\}, \\ K &= K(X, Z) = Q\cdot E(m_0 - m_1 \mid R = 0, X, Z), \end{aligned}$$

then the above equation can be written as

$$\begin{aligned} 0 &= \Delta(K/Q) - \Delta(h^*)/Q + E\{\Delta(h^*)/Q \mid R = 0, X\} \\ \Leftrightarrow 0 &= Q\Delta(K/Q) - \Delta(h^*) + Q\cdot E\{\Delta(h^*)/Q \mid R = 0, X\} \\ \Rightarrow 0 &= E\{Q\Delta(K/Q) \mid R = 0, X\} + E(Q \mid R = 0, X)\cdot E\{\Delta(h^*)/Q \mid R = 0, X\}. \end{aligned}$$

This implies that

$$E\{\Delta(h^*)/Q \mid R = 0, X\} = -\frac{E\{Q\Delta(K/Q) \mid R = 0, X\}}{E(Q \mid R = 0, X)},$$

and thus

$$\begin{aligned} \Delta(h^*) &= Q\Delta(K/Q) + Q\cdot E\{\Delta(h^*)/Q \mid R = 0, X\} \\ &= Q\Delta(K/Q) - \frac{Q\cdot E\{Q\Delta(K/Q) \mid R = 0, X\}}{E(Q \mid R = 0, X)}. \\ &= K - \frac{Q\cdot E(K \mid R = 0, X)}{E(Q \mid R = 0, X)}. \end{aligned}$$

As a result, the projection of any function $m = (1-R)m_0(X, Z) + Rm_1(X, Y, Z)$ of the observed data onto the space $\mathcal{T}^\perp$ is

$$\begin{aligned} \Pi(m \mid \mathcal{T}^\perp) &= T(h^*) = (1 - wR)\Delta(h^*), \\ &= (1 - wR)\left\{K - \frac{Q\cdot E(K \mid R = 0, X)}{E(Q \mid R = 0, X)}\right\}, \end{aligned} \tag{A.13}$$

completing the proof. □

Proof of Corollary 1. In case Z and Y are binary, we start by deriving a simplified form of the projection for any function $m(RY, R, X, Z) = (1-R)m_0(X, Z) + Rm_1(X, Y, Z)$ of the observed data:

$$\Pi(m \mid \mathcal{T}^{\perp}) = (1 - wR)\left\{K - \frac{Q \cdot E(K \mid R = 0, X)}{E(Q \mid R = 0, X)}\right\},$$

and then we generalize the results to obtain the efficient influence function. Because

$$\begin{aligned} Q &= 1/E(w \mid R = 0, X, Z) \\ K &= Q \cdot E(m_0 - m_1 \mid R = 0, X, Z) = E\{Q(m_0 - m_1) \mid R = 0, X, Z\}, \end{aligned}$$

we have that

$$\begin{aligned} & K - \frac{Q \cdot E(K \mid R = 0, X)}{E(Q \mid R = 0, X)} \\ = \; & E\{Q(m_0 - m_1) \mid R = 0, X, Z\} - \frac{Q \cdot E\{Q(m_0 - m_1) \mid R = 0, X\}}{E(Q \mid R = 0, X)} \end{aligned}$$

Let $Q_z = Q(X, Z = z), E_z = E(m_0 - m_1 \mid R = 0, X, Z = z)$, and note that Z is binary, then we can write

$$\begin{aligned} E\{Q(m_0 - m_1) \mid R = 0, X, Z\} &= (Q_1E_1 - Q_0E_0)Z + Q_0E_0 \\ E\{Q(m_0 - m_1) \mid R = 0, X\} &= (Q_1E_1 - Q_0E_0)E(Z \mid R = 0, X) + Q_0E_0 \\ Q &= (Q_1 - Q_0)Z + Q_0 \\ E(Q \mid R = 0, X) &= (Q_1 - Q_0)E(Z \mid R = 0, X) + Q_0, \end{aligned}$$

and thus,

$$\begin{aligned} & K - \frac{Q \cdot E(K \mid R = 0, X)}{E(Q \mid R = 0, X)} \\ &= \frac{Q_1Q_0(E_1 - E_0)Z + Q_1Q_0(E_0 - E_1)E(Z \mid R = 0, X)}{(Q_1 - Q_0)E(Z \mid R = 0, X) + Q_0} \\ &= \{Z - E(Z \mid R = 0, X)\}\frac{E_1 - E_0}{(1/Q_0 - 1/Q_1)E(Z \mid R = 0, X) + 1/Q_1} \\ &= \{Z - E(Z \mid R = 0, X)\}\frac{E_1 - E_0}{E\{w(X, Y) \mid R = 0, X\}}. \end{aligned}$$

Therefore, we have

$$\Pi(m \mid \mathcal{T}^{\perp}) = (1 - wR)\{Z - E(Z \mid R = 0, X)\}\frac{E_1 - E_0}{E\{w(X, Y) \mid R = 0, X\}}. \tag{A.14}$$

We next apply this result to derive $S_\gamma^{\text{eff}} = \Pi(S_\gamma \mid \mathcal{T}^{\perp})$ and $\Pi(\text{IF}_0 \mid \mathcal{T}^{\perp})$. Note that $\log \text{OR}(X, Y = 0; \gamma) = 0$ and that for binary Y we have $\nabla_\gamma \log \text{OR}(X, Y; \gamma) = \nabla_\gamma \log \text{OR}(X, Y = 1; \gamma) \cdot Y$, then the observed data score function of γ is

$$\begin{aligned} S_\gamma &= \{p(R = 1 \mid X, Z) - R\}E\{\nabla_\gamma \log \text{OR}(X, Y; \gamma) \mid R = 0, X, Z\} \\ &= \{p(R = 1 \mid X, Z) - R\}E(\nabla_\gamma \log \text{OR}(X, Y = 1; \gamma) \cdot Y \mid R = 0, X, Z). \end{aligned}$$

As a result, we obtain S_γ^{eff} from (A.14) by letting

$$\begin{aligned} m_0 &= p(R = 1 \mid X, Z)\nabla_\gamma \log \text{OR}(X, Y = 1; \gamma) \cdot E(Y \mid R = 0, X, Z) \\ m_1 &= \{p(R = 1 \mid X, Z) - 1\}\nabla_\gamma \log \text{OR}(X, Y = 1; \gamma) \cdot E(Y \mid R = 0, X, Z), \\ E_z &= \nabla_\gamma \log \text{OR}(X, Y = 1; \gamma)E(Y \mid R = 0, X, Z = z), \quad z = 0, 1. \end{aligned}$$

We further derive $\Pi(\mathrm{IF}_0 \mid \mathcal{T}^{\perp})$. Because

$$\mathrm{IF}_0 \quad = \quad wR \cdot \mathrm{NIF} + (1 - wR)E(\mathrm{NIF} \mid R = 0, X),$$

according to (A.14) we obtain $\Pi(\mathrm{IF}_0 \mid \mathcal{T}^{\perp})$ by letting

$$\begin{aligned} m_0 &= E(\mathrm{NIF} \mid R = 0, X), \\ m_1 &= w \cdot \mathrm{NIF} + (1 - w)E(\mathrm{NIF} \mid R = 0, X), \\ E_z &= E[w\{E(\mathrm{NIF} \mid R = 0, X) - \mathrm{NIF}\} \mid R = 0, X, Z = z], \quad z = 0, 1, \end{aligned}$$

completing the proof. □

Details for regression based estimation

We specify working models both for the baseline regression $p(Y, Z \mid R = 1, X; \beta)$ and the odds ratio function $\mathrm{OR}(X, Y; \gamma)$. We use $S(X, Y, Z; \beta) = \partial \log\{p(Y, Z \mid R = 1, X; \beta)\}/\partial\beta$ to denote the complete-case score function of β. Letting $\tilde{E}$ denote the expectation with respect to the working model we specify, $\hat{E}$ the empirical mean, and $h(X, Z)$ a user-specified vector function which depends on Z, we solve

$$\hat{E}\{R \cdot S(X, Y, Z; \hat{\beta})\} = 0, \tag{A.15}$$

$$\hat{E}[(1 - R)\{h(X, Z) - \tilde{E}(h(X, Z) \mid R = 0, X; \hat{\beta}, \hat{\gamma}_{\mathrm{reg}})\}] = 0, \tag{A.16}$$

$$\hat{E}[(1 - R)\tilde{E}\{U(\hat{\psi}_{\mathrm{reg}}) \mid R = 0, X, Z; \hat{\beta}, \hat{\gamma}_{\mathrm{reg}}\} + R \cdot U(\hat{\psi}_{\mathrm{reg}})] = 0. \tag{A.17}$$

to obtain $\hat{\beta}$ and the regression based estimator $(\hat{\gamma}_{\mathrm{reg}}, \hat{\psi}_{\mathrm{reg}})$. Equation (A.15) results in a complete-case estimator of β, and (A.16)–(A.17) lead to regression based estimators of γ and ψ, respectively. The conditional expectation $\tilde{E}$ in (A.16)–(A.17) are evaluated under the conditional density $p(Y, Z \mid R = 0, X, \hat{\beta}, \hat{\gamma}_{\mathrm{reg}})$, which is determined by working models $p(Y, Z \mid R = 1, X; \hat{\beta})$ and $\mathrm{OR}(X, Y; \hat{\gamma}_{\mathrm{reg}})$ as in (6)–(7).

ACKNOWLEDGMENT

We are grateful for valuable comments from the editor and two anonymous reviewers. This work was partially supported by National Key R&D Program of China (2020YFE0204200, 2022YFA1008100), National Natural Science Foundation of China (12071015, 12292980), NSF grant DMS-1916013, and NIH grant (R01AI27271, R01CA222147, R01AG065276, R01GM139926, R01HL155417-01).

REFERENCES

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

Peter J Bickel, Chris A J Klaassen, YA'Acov Ritov, and Jon A Wellner. 1993. *Efficient and Adaptive Estimation for Semiparametric Models.* Johns Hopkins University Press, Baltimore.

Hua Yun Chen. 2003. A note on the prospective analysis of outcome-dependent samples. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 65 (2003), 575–584.

Hua Yun Chen. 2004. Nonparametric and semiparametric models for missing covariates in parametric regression. *J. Amer. Statist. Assoc.* 99 (2004), 1176–1189.

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

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

Arthur P Dempster, Nan M Laird, and Donald B Rubin. 1977. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)* 39 (1977), 1–38.

Xavier D'Haultfœuille. 2010. A new instrumental method for dealing with endogenous selection. *Journal of Econometrics* 154 (2010), 1–15.

Xavier D'Haultfœuille. 2011. On the completeness condition in nonparametric instrumental problems. *Econometric Theory* 27 (2011), 460–471.

Fang Fang, Jiwei Zhao, and Jun Shao. 2018. Imputation-based adjusted score equations in generalized linear models with nonignorable missing covariate values. *Statistica Sinica* 28 (2018), 1677–1701.

Robert E Fay. 1986. Causal models for patterns of nonresponse. *J. Amer. Statist. Assoc.* 81 (1986), 354–365.

John S Greenlees, William S Reece, and Kimberly D Zieschang. 1982. Imputation of missing values when the probability of response depends on the variable being imputed. *J. Amer. Statist. Assoc.* 77 (1982), 251–261.

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

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

Yingyao Hu and Ji-Liang Shiu. 2018. Nonparametric identification using instrumental variables: Sufficient conditions for completeness. *Econometric Theory* 34 (2018), 659–693.

Joseph G Ibrahim, Stuart R Lipsitz, and Nick Horton. 2001. Using auxiliary data for parameter estimation with non-ignorably missing outcomes. *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 50 (2001), 361–373.

Huaqing Jin, Yanyuan Ma, and Fei Jiang. 2022. Matrix Completion with Covariate Information and Informative Missingness. *Journal of Machine Learning Research* 23, 180 (2022), 1–62.

Jae Kwang Kim and Cindy Long Yu. 2011. A semiparametric estimation of mean functionals with nonignorable missing data. *J. Amer. Statist. Assoc.* 106 (2011), 157–165.

Phillip S. Kott. 2014. Calibration Weighting When Model and Calibration Variables Can Differ. In *Contributions to Sampling Statistics*, Fulvia Mecatti, Luigi Pier Conti, and Giovanna Maria Ranalli (Eds.). Springer, Cham, 1–18.

Roderick JA Little. 1993. Pattern-mixture models for multivariate incomplete data. *J. Amer. Statist. Assoc.* 88 (1993), 125–134.

Roderick JA Little. 1994. A class of pattern-mixture models for normal incomplete data. *Biometrika* 81 (1994), 471–483.

Lan Liu, Wang Miao, Baoluo Sun, James Robins, and Eric Tchetgen Tchetgen. 2020. Identification and inference for marginal average treatment effect on the treated with an instrumental variable. *Statistica Sinica* 30, 3 (July 2020), 1517–1541.

Po-Ling Loh and Martin J. Wainwright. 2012. High-dimensional regression with noisy and missing data: Provable guarantees with non-convexity. *The Annals of Statistics* 40, 3 (2012), 1637–1664.

Wen Qing Ma, Zhi Geng, and Yong Hua Hu. 2003. Identification of graphical models for nonignorable nonresponse of binary outcomes in longitudinal studies. *Journal of Multivariate Analysis* 87 (2003), 24–45.

Xiaojun Mao, Song Xi Chen, and Raymond KW Wong. 2019. Matrix completion with covariate information. *J. Amer. Statist. Assoc.* 114, 525 (2019), 198–210.

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

Wang Miao and Eric Tchetgen Tchetgen. 2016. On varieties of doubly robust estimators under missingness not at random with a shadow variable. *Biometrika* 103 (2016), 475–482.

Wang Miao and Eric Tchetgen Tchetgen. 2018. Identification and inference with nonignorable missing covariate data. *Statistica Sinica* 28 (2018), 2049–2067.

Kosuke Morikawa and Jae Kwang Kim. 2021. Semiparametric optimal estimation with nonignorable nonresponse data. *The Annals of Statistics* 49, 5 (2021), 2991–3014.

Whitney K Newey. 1994. The asymptotic variance of semiparametric estimators. *Econometrica* 62 (1994), 1349–1382.

Whitney K Newey and Daniel McFadden. 1994. Large sample estimation and hypothesis testing. In *Handbook of Econometrics*, R. F. Engle and D. L. McFadden (Eds.). Vol. 4. Elsevier, Amsterdam, 2111–2245.

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

Gerhard Osius. 2004. The association between two random elements: A complete characterization and odds ratio models. *Metrika* 60 (2004), 261–277.

James Robins and Andrea Rotnitzky. 2001. Comment on the Bickel and Kwon article, 'Inference for semiparametric models: Some questions and an answer'. *Statistica Sinica* 11 (01 2001), 920–936.

James M. Robins, Andrea Rotnitzky, and Daniel O. Scharfstein. 2000. Sensitivity Analysis for Selection bias and unmeasured Confounding in missing Data and Causal inference models. In *Statistical Models in Epidemiology, the Environment, and Clinical Trials*, M. Elizabeth Halloran and Donald Berry (Eds.). Springer New York, New York, 1–94.

A Rotnitzky and JM Robins. 1997. Analysis of semi-parametric regression models with non-ignorable non-response. *Statistics in Medicine* 16 (1997), 81–102.

Andrea Rotnitzky, James Robins, and Daniel O Scharfstein. 1998. Semiparametric regression for repeated outcomes with nonignorable nonresponse. *J. Amer. Statist. Assoc.* 93 (1998), 1321–1339.

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

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

Daniel O Scharfstein, Andrea Rotnitzky, and James Robins. 1999. Adjusting for nonignorable drop-out using semiparametric nonresponse models. *J. Amer. Statist. Assoc.* 94 (1999), 1096–1120.

Nathaniel Schenker and Alan H Welsh. 1988. Asymptotic results for multiple imputation. *The Annals of Statistics* 16, 4 (1988), 1550–1566.

BaoLuo Sun, Lan Liu, Wang Miao, Kathleen Wirth, James Robins, and Eric Tchetgen Tchetgen. 2018. Semiparametric Estimation With Data Missing Not at Random Using an Instrumental Variable. *Statistica Sinica* 28 (2018), 1965–1983.

Niansheng Tang, Puying Zhao, and Hongtu Zhu. 2014. Empirical likelihood for estimating equations with nonignorably missing data. *Statistica Sinica* 24 (2014), 723–747.

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

T Tony Cai and Linjun Zhang. 2019. High dimensional linear discriminant analysis: optimality, adaptive algorithm and missing data. *Journal of the Royal Statistical Society Series B: Statistical Methodology* 81, 4 (2019), 675–705.

Anastasios Tsiatis. 2006. *Semiparametric Theory and Missing Data.* Springer, New York.

Mark J Van der Laan and James Robins. 2003. *Unified Methods for Censored Longitudinal Data and Causality.* Springer, New York.

Stijn Vansteelandt, Andrea Rotnitzky, and James Robins. 2007. Estimation of regression models for the mean of repeated outcomes under nonignorable nonmonotone nonresponse. *Biometrika* 94 (2007), 841–860.

Sheng Wang, Jun Shao, and Jae Kwang Kim. 2014. An instrumental variable approach for identification and estimation with nonignorable nonresponse. *Statistica Sinica* 24 (2014), 1097–1116.

S. Yang, L. Wang, and P. Ding. 2019. Causal inference with confounders missing not at random. *Biometrika* 106, 4 (09 2019), 875–888.

Gwendolyn EP Zahner, Walter Pawelkiewicz, John J DeFrancesco, and Jean Adnopoz. 1992. Children's mental health service needs and utilization patterns in an urban community: an epidemiological assessment. *Journal of the American Academy of Child & Adolescent Psychiatry* 31 (1992), 951–960.

Jiwei Zhao and Yanyuan Ma. 2018. Optimal pseudolikelihood estimation in the analysis of multivariate missing data with nonignorable nonresponse. *Biometrika* 105 (2018), 479–486.

Jiwei Zhao and Yanyuan Ma. 2021. A Versatile Estimation Procedure without Estimating the Nonignorable Missingness Mechanism. *J. Amer. Statist. Assoc.* 0, ja (2021), 1–44.

Jiwei Zhao and Jun Shao. 2015. Semiparametric pseudo-likelihoods in generalized linear models with nonignorable missing data. *J. Amer. Statist. Assoc.* 110 (2015), 1577–1590.

Ziwei Zhu, Tengyao Wang, and Richard J Samworth. 2022. High-dimensional principal component analysis with heterogeneous missingness. *Journal of the Royal Statistical Society. Series B: Statistical Methodology* 84, 5 (2022), 2000–2031.