Journal of the American Statistical Association

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

A Versatile Estimation Procedure Without Estimating the Nonignorable Missingness Mechanism

Jiwei Zhao & Yanyuan Ma

To cite this article: Jiwei Zhao & Yanyuan Ma (2021): A Versatile Estimation Procedure Without Estimating the Nonignorable Missingness Mechanism, Journal of the American Statistical Association, DOI: 10.1080/01621459.2021.1893176

To link to this article: https://doi.org/10.1080/01621459.2021.1893176

JOURNAL OF THE AMERICAN STATISTICAL ASSOCIATION
2021, VOL. 00, NO. 0, 1–15: Theory and Methods
https://doi.org/10.1080/01621459.2021.1893176

A Versatile Estimation Procedure Without Estimating the Nonignorable Missingness Mechanism

Jiwei Zhao[a] and Yanyuan Ma[b]

[a]Department of Biostatistics and Medical Informatics, University of Wisconsin-Madison, Madison, WI; [b]Department of Statistics, Pennsylvania State University, University Park, PA

ABSTRACT

We consider the estimation problem in a regression setting where the outcome variable is subject to nonignorable missingness and identifiability is ensured by the shadow variable approach. We propose a versatile estimation procedure where modeling of missingness mechanism is completely bypassed. We show that our estimator is easy to implement and we derive the asymptotic theory of the proposed estimator. We also investigate some alternative estimators under different scenarios. Comprehensive simulation studies are conducted to demonstrate the finite sample performance of the method. We apply the estimator to a children's mental health study to illustrate its usefulness.

ARTICLE HISTORY
Received February 2019
Accepted February 2021

KEYWORDS
Asymptotic normality; Identifiability; Missingness mechanism; Nonignorable missing data; Semiparametric theory; Shadow variable

1. Introduction

In statistical data analysis, the issue of missing values is a rule rather than an exception. There are often many missing data in biomedical and health related studies, social sciences, and survey sampling. How to appropriately address missingness is fascinating but challenging, and has drawn much attention to statisticians in the past several decades.

In the missing data literature, the missingness mechanism is a key concept and a fundamental and useful taxonomy to distinguish different problems. The missingness is named ignorable if it depends on the observed data only; otherwise, it is named nonignorable. Rich literatures exist on handling ignorable missing data (Rubin 1987; Robins, Rotnitzky, and Zhao 1994; Schafer 1997; Little and Rubin 2002; Tsiatis 2006; Kim and Shao 2013; Molenberghs et al. 2014). However, in many practical situations, it is highly likely that the missingness actually also depends on the missed variables themselves hence is nonignorable. Research for nonignorable missing data is not yet as complete due to its difficulties. Simply applying existing methods for ignorable missing data to nonignorable ones may lead to biased parameter estimation, incorrect standard errors and, as a consequence, incorrect statistical inference and conclusions.

One notorious issue for analyzing nonignorable missing data is the model identifiability. Here, identifiability means that any two different sets of parameters produce two different models. To achieve identifiability, it is well known in the literature (Robins and Ritov 1997) that assumptions are needed for either the data generation process, or the mechanism model, or both. In a multivariate analysis setting, Tang et al. (2003) established a model identifiability condition when maximum pseudo likelihood estimation was used, where they primarily assumed that the missingness only depends on the missed variable. d'Haultfoeuille (2010) adopted some similar assumption and proposed to solve the model identifiability issue nonparametrically. By identifying an appropriate shadow variable (Kott 2014), much more flexible nonignorable missingness mechanism can be adopted, such as Shao and Zhao (2013), Wang, Shao, and Kim (2014), Zhao and Shao (2015), etc. Although these earlier literature named the variable *nonresponse instrument*, to avoid confusion with the literature on instrumental variables for missing data (Newey and Powell 2003; Tchetgen Tchetgen and Wirth 2017; Sun et al. 2018), in the current article, we follow the work of Miao and Tchetgen Tchetgen (2016), Zhao and Ma (2018), and Miao et al. (2019) and term it *shadow variable*. Shadow variable is prevalent in survey sampling designs and is available in many empirical studies (Kott 2014). Since the nonignorable missingness mechanism is difficult to verify, the flexibility in modeling the mechanism is highly appreciated in applications, and the shadow variable approach provides one such possibility. With a suitable shadow variable, the exact model identifiability conditions still need to be investigated on a case-by-case basis. More details of the use of the shadow variable strategy in this article, the specific identifiability conditions, and the validity in applications are presented in Section 2.

Provided that the model is identifiable, the other controversial part in analyzing nonignorable missing data is on modeling the missingness mechanism. Because of its dependence on the unobserved data, it is nearly impossible to verify the mechanism model in practice except for a few special scenarios (d'Haultfoeuille 2010). In the literature, there are many parametric modeling attempts for the mechanism (Ibrahim and Lipsitz 1996; Rotnitzky and Robins 1997; Qin, Leung, and Shao 2002; Chang and Kott 2008; Wang, Shao, and Kim 2014; Morikawa and Kim 2016), but parametric mechanism model is generally

CONTACT Jiwei Zhao *jiwei.zhao@wisc.edu* Department of Biostatistics and Medical Informatics, University of Wisconsin-Madison, Madison, WI 53726.
Supplementary materials for this article are available online. Please go to *www.tandfonline.com/r/JASA*.

considered to be restrictive. Kim and Yu (2011) and Shao and Wang (2016) extended the parametric mechanism to a semiparametric framework which contains a more flexible nonparametric component. However, these semiparametric mechanism models are also confined to a special structure and can still be misspecified.

Due to the difficulty in modeling the missingness mechanism, in this article we completely avoid this practice. We propose a versatile estimation procedure which does not require modeling or estimating the missingness mechanism. The key idea of our proposal is to view the mechanism as a nuisance parameter in a semiparametric model, and to project away its effect via semiparametric treatment (Bickel et al. 1993; Tsiatis 2006). In the estimator we construct in this work, only a working model for the mechanism is needed in the implementation, and the working model does not have to contain the true mechanism.

Our procedure requires estimating integrals depending on the probability density function (pdf) or probability mass function (pmf) of the covariate variable. Because covariates are fully observed, this is a complete-data problem and many statistical methods exist in the literatures. We propose to estimate the integral through empirical expectation if the integral can be viewed as a marginal expectation, and to estimate it through nonparametric regression technique, such as kernels, if the integral can be converted into a conditional expectation. Our procedure is more robust compared to parametric estimation of the pdf/pmf, and is simpler to implement compared to nonparametric estimation. It is also worthwhile to mention that it is technically challenging to establish the asymptotic theory of the proposed estimator, which requires extensive use of bilinear operators in combination with semiparametric treatments.

Compared to the current literature, this work has the following distinctive features hence makes novel contributions. First, under the semiparametric modeling framework, we directly work on the likelihood function and clearly pinpoint the conditions under which the model is identifiable. Previous works usually study some modified versions of the likelihood, such as pseudo likelihood (Tang et al. 2003; Zhao and Shao 2015; Zhao and Ma 2018) or conditional likelihood (Zhao 2017, 2018; Zhao and Shao 2017), and avoid treating the nonparametric component by incorporating various modeling assumptions. Our techniques to studying the semiparametric likelihood here are very different. Second, we rigorously characterize the complete geometric structure of our semiparametric model, which not only produces the versatile estimation approach, but also portrays the potential to create other types of estimators for the parameter of interest. Third, despite the unknown nonignorable missingness mechanism, our proposed approach completely overcomes the hurdle of either modeling or estimating it. Our approach literally encompasses a class of estimators, and it is practically convenient to use in the sense that any arbitrary working model of the mechanism will always generate an asymptotically consistent estimator. Fourth, it is not a standard exercise to develop the asymptotic theory for the proposed estimator in this article. We extensively use the semiparametric technique as well as the properties of bilinear operators.

We also would like to point out that, our framework is based on a correctly specified regression model, hence is most appropriate for studying the association between outcome variable and covariates. If the interest is different, for example, if the interest is in studying the expectation of the outcome, other modeling approaches need to be considered. For example, Miao et al. (2019) studied identification and inference under a general pattern mixture parameterization, so their approach does not require a correctly specified parametric regression model. In addition, Miao et al. (2019) proposed a set of identification conditions in their framework that only involve the observed data and hence are testable empirically.

The rest of the article is as follows. In Section 2, we clarify notations and assumptions, describe the shadow variable strategy and lay down the model identification conditions. In Section 3, we study the situation where the whole covariate vector serves as the shadow variable. We derive the efficient score, propose our estimator and establish the asymptotic theory. The parallel results under the more general situation where part of the covariate serves as the shadow variable is established in Section 4. A few alternative estimators under different scenarios are investigated in Section 5. In Section 6, we conduct comprehensive simulation studies to demonstrate the finite sample performance of our proposed methods under various situations. In Section 7, we analyze a data concerning a children's mental health study. The article is concluded with a discussion in Section 8.

2. Notations, Assumptions, and Identifiability

Consider the regression model $f_{Y|\mathbf{X}}(y, \mathbf{x}; \boldsymbol{\beta})$, the conditional probability density/mass function of Y given $\mathbf{X}$, where $\boldsymbol{\beta}$ is a p-dimensional unknown parameter to be estimated. Regression model of this type has been used to study the association between an outcome variable Y and a set of covariates $\mathbf{X}$ in the literatures. Indeed, when data are subject to missingness, the model $f_{Y|\mathbf{X}}(y, \mathbf{x}; \boldsymbol{\beta})$ continues to be used by scientific investigators in various disciplines. Therefore, in the current article, the estimation and inference of $\boldsymbol{\beta}$ is our primary scientific interest, and our major statistical interest is to understand how the missingness will affect the estimation and inference of $\boldsymbol{\beta}$.

Throughout the article, the covariate $\mathbf{X}$ is fully observed and let the pdf/pmf of $\mathbf{X}$ be $f_{\mathbf{X}}(\cdot)$. The outcome variable Y is subject to missingness. Let the binary variable R be the missingness indicator, with $R = 1$ for an observed Y and $R = 0$ for a missing Y. Write the missingness mechanism as $\text{pr}(R = 1 \mid Y, \mathbf{X})$. We observe N independent and identically distributed realizations of $(R, RY, \mathbf{X})$, written as $\{(r_i, r_i y_i, \mathbf{x}_i)\}, i = 1, \ldots, N$. Without loss of generality, we assume that the first n subjects are completely observed, that is, $r_i = 1$ for $i = 1, \ldots, n$, while the remaining $N - n$ subjects have $r_i = 0$ for $i = n + 1, \ldots, N$.

We adopt the shadow variable framework, that is, we assume that the covariate $\mathbf{X}$ can be decomposed as $\mathbf{X} = (\mathbf{U}^{\mathrm{T}}, \mathbf{Z}^{\mathrm{T}})^{\mathrm{T}}$, such that

$$f_{Y|\mathbf{X}}(y, \mathbf{x}) \neq f_{Y|\mathbf{X}}(y, \mathbf{u}) \text{ and}$$
$$\text{pr}(R = 1 \mid y, \mathbf{x}) = \text{pr}(R = 1 \mid y, \mathbf{u}) = \pi(y, \mathbf{u}). \quad (1)$$

The variable $\mathbf{Z}$ is termed the shadow variable. This implies that part of the covariate, $\mathbf{Z}$, is independent of the missingness indicator R conditional on the outcome Y and the other part

of the covariate $\mathbf{U}$. Consequently, while $\mathbf{Z}$ appears in the model $f_{Y|\mathbf{X}}(y,\mathbf{x})$, it does not in the model $\text{pr}(R=1\mid Y,\mathbf{X})$, hence is shadowed out. Note that $\mathbf{Z}$ can be $\mathbf{X}$, hence the whole covariate $\mathbf{X}$ itself is the shadow variable, but $\mathbf{Z}$ cannot be empty, which degenerates to the no shadow variable situation. The shadow variable assumption is popularly used in the survey sampling literature (Kott 2014) and the nonignorable missing data literature (Shao and Zhao 2013; Wang, Shao, and Kim 2014; Zhao and Shao 2015; Miao and Tchetgen Tchetgen 2016; Zhao and Ma 2018; Miao et al. 2019). It is often convenient to use in applications. For example, in a children's mental health study (Ibrahim, Lipsitz, and Horton 2001), the teacher's assessment of the psychopathology status of the student, which suffered from nonignorable missing data, was specified as the outcome variable. Besides the teacher's assessment, the dataset also included a separate parent's report about the psychopathology status of the child. As indicated in Ibrahim, Lipsitz, and Horton (2001), the teacher's response rate may be related to her assessment of the student but is unlikely to be related to a separate parent's report after conditioning on the teacher's assessment and all other fully observed covariates; moreover, the parent's report is likely highly correlated with that of the teacher (Miao et al. 2019). In this case, the parent's report constitutes a valid shadow variable.

As we discussed in Section 1, even when a shadow variable is present, the model identifiability conditions still need to be investigated on a case-by-case basis. In our situation, without any extra conditions on $f_{Y|\mathbf{X}}(y,\mathbf{x};\boldsymbol{\beta})$, the joint distribution of $(R,Y,\mathbf{X})$ depends on the unknown components $\boldsymbol{\beta}$, $\pi(y,\mathbf{u})$ and $f_{\mathbf{X}}(\mathbf{x})$, and may still be unidentifiable. The following example illustrates this point.

Example 1. Let Y be a discrete variable with three possible values $0,1,2$. Let $X=Z$, which is also a discrete variable with two possible values $0,1$. The missingness mechanism is $\text{pr}(R=1\mid y)=\pi(y)$. Let $f_{Y|\mathbf{X}}(y,x;\boldsymbol{\beta})$ be the saturated model. Thus, the model $f_{Y|\mathbf{X}}(y,x;\boldsymbol{\beta})$ contains four parameters, $\pi(y)$ contains three parameters, and $\text{pr}(X=1)$ contains one parameter. We have a total of eight free parameters in the joint distribution of (R,RY,X).

We will show that it is impossible to identify the eight parameters. Consider the data generation process

$$\begin{aligned}
&\text{pr}(Y=2\mid X=0)=\frac{6}{7}(1+\alpha),\ \ \text{pr}(Y=1\mid X=0)=\alpha,\\
&\text{pr}(Y=2\mid X=1)=\frac{3}{7}(1+\alpha),\ \ \text{pr}(Y=1\mid X=1)=7\alpha,\\
&\pi(y)=\frac{7}{32}\{(1-3y+2y^2)-(13-32y+12y^2)\alpha\}^{-1},\\
&\text{pr}(X=1)=\frac{1}{2},
\end{aligned}$$

where α is a constant in the range $(1/32, 25/(13\times 32))$. Now

$$\pi(y)f_{Y|\mathbf{X}}(y,x;\boldsymbol{\beta})=\frac{1}{32}1^{I(y=2,x=1)}2^{I(y=2,x=0)}7^{I(y=1,x=1)}1^{I(y=1,x=0)}4^{I(y=0,x=1)}1^{I(y=0,x=0)}$$

is free of α, hence the likelihood function is free of α. Therefore, this model is unidentifiable.

Example 1 indicates that, to achieve the model identifiability, one needs to impose further conditions. We show that the completeness condition below will guarantee the identifiability.

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

The completeness condition is widely used for model identifiability across various disciplines; to name a few, Newey and Powell (2003), d'Haultfoeuille (2010), Hu and Shiu (2018), and Miao et al. (2019). It is satisfied for many commonly used models of $f_{Y|\mathbf{X}}(y,\mathbf{x})$. For example, in exponential families, where $f_{Y|\mathbf{X}}(y,\mathbf{x})=s(y,\mathbf{u})t(\mathbf{x})\exp\{\boldsymbol{\mu}(\mathbf{x})^{\mathrm{T}}\boldsymbol{\tau}(y,\mathbf{u})\}$, with $s(y,\mathbf{u})>0$, $t(\mathbf{x})>0$, $\boldsymbol{\tau}(y,\mathbf{u})$ is one-to-one in y, and the support of $\boldsymbol{\mu}(\mathbf{x})$ is an open set, then the completeness condition holds. This was documented in classic textbook such as Lehmann and Romano (2006, Theorem 4.3.1). Therefore, commonly seen regression models, such as the linear regression for continuous Y and the logistic regression for binary Y, satisfy the completeness condition. For the situation where both y and $\mathbf{z}$ are discrete with finite support $\{y_1,\ldots,y_s\}$ and $\{\mathbf{z}_1,\ldots,\mathbf{z}_t\}$, Newey and Powell (2003) noted that, the completeness condition implicitly requires that $t\geq s$, that is, the shadow variable has a no smaller support than the variable with missing values. This also explains why the model in Example 1 is unidentifiable. We state the identifiability result in Lemma 1, with its proof in Appendix A.1.

Lemma 1. Under the shadow variable assumption (1) and Condition 1, all unknown components $\boldsymbol{\beta}$, $\pi(y,\mathbf{u})$ and $f_{\mathbf{X}}(\mathbf{x})$ are identifiable.

In lieu of Condition 1, other conditions can be adopted to achieve identifiability and are very specific to the particular model. For example, Tang et al. (2003) considered a special case of (1) and aimed at estimating $\boldsymbol{\beta}$ only. To achieve the identification of $\boldsymbol{\beta}$ in a linear regression with normal errors, they required the cardinality of the support of the shadow variable to be at least three. Zhao and Shao (2015) considered a generalized linear model $f_{Y|\mathbf{X}}(y,\mathbf{x};\boldsymbol{\beta})$ and devised different identifiability conditions depending on whether the dispersion parameter is known or not. The identification problem was solved fully nonparametrically in Miao et al. (2019); furthermore, their conditions only involved the observed data distribution and therefore was testable empirically.

Given an identifiable model, we proceed to consider the estimation problem. If we adopt a likelihood approach, even though our sole interest is in estimating $\boldsymbol{\beta}$, we cannot avoid the estimation of both $\pi(y,\mathbf{u})$ and $f_{\mathbf{X}}(\mathbf{x})$. While the estimation of $f_{\mathbf{X}}(\mathbf{x})$ is a standard problem since there is no missing data in the variable $\mathbf{X}$, the estimation of $\pi(y,\mathbf{u})$ is challenging. Due to the missingness in Y, the $\pi(\cdot)$ model is usually unverifiable and can be easily misspecified in practice.

Aware of this difficulty, we propose to estimate $\boldsymbol{\beta}$ while avoiding modeling or estimating the missingness mechanism. Instead, we only need to posit a working model for $\pi(y,\mathbf{u})$, which could be misspecified. We show that, using an arbitrary working model $\pi^*(y,\mathbf{u})$, our estimator of $\boldsymbol{\beta}$ is always consistent and asymptotically normal, hence our procedure is robust to mechanism misspecification.

For ease of illustration, also with its own importance, in Section 3, we first analyze a special case of (1) where the whole covariate serves as the shadow variable, that is, $\mathbf{X} = \mathbf{Z}$, such that

$$f_{Y|\mathbf{X}}(y, \mathbf{x}) \neq f_Y(y), \text{ and}$$
$$\text{pr}(R = 1 \mid y, \mathbf{x}) = \text{pr}(R = 1 \mid y) = \pi(y). \quad (2)$$

Analysis under the general assumption (1), which turns out to be statistically very different and mathematically more challenging, is conducted in Section 4.

3. Proposed Estimator Under Special Assumption (2)

3.1. Estimation Procedure

Under (2), the joint pdf of $(\mathbf{X}, RY, R)$ is

$$f_{\mathbf{X}}(\mathbf{x})\{f_{Y|\mathbf{X}}(y, \mathbf{x}; \boldsymbol{\beta})\pi(y)\}^r \left\{1 - \int f_{Y|\mathbf{X}}(t, \mathbf{x}; \boldsymbol{\beta})\pi(t)d\mu(t)\right\}^{1-r}.$$

Because $\boldsymbol{\beta}$ is the parameter of interest while $f_{\mathbf{X}}(\mathbf{x})$ and $\pi(y)$ are nuisance, we take a semiparametric approach and derive the nuisance tangent space, its orthogonal complement and the efficient score with respect to $\boldsymbol{\beta}$. In Appendix A.2, we first derive that the nuisance tangent space $\Lambda = \Lambda_{f_{\mathbf{X}}} \oplus \Lambda_\pi$, where

$$\Lambda_{f_{\mathbf{X}}} = [\mathbf{a}(\mathbf{x}) : E\{\mathbf{a}(\mathbf{X})\} = \mathbf{0}],$$
$$\Lambda_\pi = \left[r\mathbf{b}(y) - (1-r)\frac{\int f_{Y|\mathbf{X}}(t, \mathbf{x}; \boldsymbol{\beta})\mathbf{b}(t)\pi(t)d\mu(t)}{\int f_{Y|\mathbf{X}}(t, \mathbf{x}; \boldsymbol{\beta})\{1 - \pi(t)\}d\mu(t)} : \text{for all } \mathbf{b}(y)\right],$$

where $\oplus$ stands for the addition of two spaces that are orthogonal to each other. We also derive the orthogonal complement of Λ to be

$$\Lambda^\perp = \left[\mathbf{a}(\mathbf{x}, r, ry) : E\{\mathbf{a}(\mathbf{x}, R, RY) \mid \mathbf{x}\} = \mathbf{0}, E\left\{\mathbf{a}(\mathbf{X}, 1, y) - \mathbf{a}(\mathbf{X}, 0, 0) \mid y\right\} = \mathbf{0}\right].$$

The form of $\Lambda^\perp$ permits many possibilities for constructing estimating equations for $\boldsymbol{\beta}$. Among all elements in $\Lambda^\perp$, the most interesting one is the efficient score, defined as the orthogonal projection of the score vector $\mathbf{S}_{\boldsymbol{\beta}}$ onto $\Lambda^\perp$, where

$$\mathbf{S}_{\boldsymbol{\beta}}(\mathbf{x}, r, ry, \boldsymbol{\beta}) = r\frac{\mathbf{f}_{\boldsymbol{\beta}}(y, \mathbf{x}; \boldsymbol{\beta})}{f_{Y|\mathbf{X}}(y, \mathbf{x}; \boldsymbol{\beta})} - (1-r)\frac{\int \mathbf{f}_{\boldsymbol{\beta}}(t, \mathbf{x}; \boldsymbol{\beta})\pi(t)d\mu(t)}{\int f_{Y|\mathbf{X}}(t, \mathbf{x}; \boldsymbol{\beta})\{1 - \pi(t)\}d\mu(t)}. \quad (3)$$

Here $\mathbf{f}_{\boldsymbol{\beta}}(y, \mathbf{x}; \boldsymbol{\beta}) \equiv \partial f_{Y|\mathbf{X}}(y, \mathbf{x}; \boldsymbol{\beta})/\partial\boldsymbol{\beta}$ is a vector of the same dimension as $\boldsymbol{\beta}$. In Appendix A.3, we show that

$$\mathbf{S}_{\text{eff}}(\mathbf{x}, r, ry) = \frac{r\mathbf{f}_{\boldsymbol{\beta}}(y, \mathbf{x}; \boldsymbol{\beta})}{f_{Y|\mathbf{X}}(y, \mathbf{x}; \boldsymbol{\beta})} - \frac{(1-r)\int \mathbf{f}_{\boldsymbol{\beta}}(t, \mathbf{x}; \boldsymbol{\beta})\pi(t)d\mu(t)}{\int f_{Y|\mathbf{X}}(t, \mathbf{x}; \boldsymbol{\beta})\{1 - \pi(t)\}d\mu(t)} - r\mathbf{b}(y) + \frac{(1-r)\int \mathbf{b}(t)f_{Y|\mathbf{X}}(t, \mathbf{x}; \boldsymbol{\beta})\pi(t)d\mu(t)}{\int f_{Y|\mathbf{X}}(t, \mathbf{x}; \boldsymbol{\beta})\{1 - \pi(t)\}d\mu(t)},$$

where $\mathbf{b}(y)$ is the solution to the integral equation

$$\int \left\{\mathbf{f}_{\boldsymbol{\beta}}(y, \mathbf{x}; \boldsymbol{\beta}) + \frac{\int \mathbf{f}_{\boldsymbol{\beta}}(t, \mathbf{x}; \boldsymbol{\beta})\pi(t)d\mu(t)}{1 - \int f_{Y|\mathbf{X}}(t, \mathbf{x}; \boldsymbol{\beta})\pi(t)d\mu(t)}f_{Y|\mathbf{X}}(y, \mathbf{x}; \boldsymbol{\beta})\right\} \quad (4)$$
$$\times f_{\mathbf{X}}(\mathbf{x})d\mu(\mathbf{x})$$
$$= \int \left\{\mathbf{b}(y)f_{Y|\mathbf{X}}(y, \mathbf{x}; \boldsymbol{\beta}) + \frac{\int \mathbf{b}(t)f_{Y|\mathbf{X}}(t, \mathbf{x}; \boldsymbol{\beta})\pi(t)d\mu(t)}{1 - \int f_{Y|\mathbf{X}}(t, \mathbf{x}; \boldsymbol{\beta})\pi(t)d\mu(t)}f_{Y|\mathbf{X}}(y, \mathbf{x}; \boldsymbol{\beta})\right\}f_{\mathbf{X}}(\mathbf{x})d\mu(\mathbf{x}).$$

Note that because $\mathbf{S}_{\text{eff}}(\mathbf{x}, 1, y) = \mathbf{f}_{\boldsymbol{\beta}}(y, \mathbf{x}; \boldsymbol{\beta})/f_{Y|\mathbf{X}}(y, \mathbf{x}; \boldsymbol{\beta}) - \mathbf{b}(y)$ and the efficient score exists and is unique, hence the solution $\mathbf{b}(y)$ to (4) exists and is unique.

Despite of the results above, the efficient score $\mathbf{S}_{\text{eff}}$ is not readily implementable because it contains the unknown quantities $f_{\mathbf{X}}(\mathbf{x})$ and $\pi(y)$. As we have pointed out, we aim to avoid estimating or even modeling $\pi(y)$. Thus, we propose to adopt a working model of the mechanism, denoted $\pi^*(y)$. We show in Appendix A.4 that in the construction of $\mathbf{S}_{\text{eff}}(\mathbf{x}, r, ry)$, we can adopt $\pi^*(y)$ and the resulting "working model based efficient score" $\mathbf{S}^*_{\text{eff}}(\mathbf{x}, r, ry)$ still has mean zero. On the other hand, the integrations in (4) can be viewed as expectations with respect to the covariate $\mathbf{X}$. Because $\mathbf{X}$ is fully observed, we recommend to approximate the expectations using their corresponding empirical versions. Combining these two aspects, we propose the following flexible estimation procedure.

Algorithm 1. Algorithm under special assumption (2)

Step 1. Posit a working model for $\pi(y)$, denoted as $\pi^*(y)$.

Step 2. Obtain $\widehat{\mathbf{b}}^*(y, \boldsymbol{\beta})$ by solving the integral equation

$$\frac{1}{N}\sum_{i=1}^N \left\{\mathbf{f}_{\boldsymbol{\beta}}(y, \mathbf{x}_i; \boldsymbol{\beta}) + \frac{\int \mathbf{f}_{\boldsymbol{\beta}}(t, \mathbf{x}_i; \boldsymbol{\beta})\pi^*(t)d\mu(t)}{1 - \int f_{Y|\mathbf{X}}(t, \mathbf{x}_i; \boldsymbol{\beta})\pi^*(t)d\mu(t)}f_{Y|\mathbf{X}}(y, \mathbf{x}_i; \boldsymbol{\beta})\right\} \quad (5)$$
$$= \frac{1}{N}\sum_{i=1}^N \left\{\mathbf{b}(y)f_{Y|\mathbf{X}}(y, \mathbf{x}_i; \boldsymbol{\beta}) + \frac{\int \mathbf{b}(t)f_{Y|\mathbf{X}}(t, \mathbf{x}_i; \boldsymbol{\beta})\pi^*(t)d\mu(t)}{1 - \int f_{Y|\mathbf{X}}(t, \mathbf{x}_i; \boldsymbol{\beta})\pi^*(t)d\mu(t)}f_{Y|\mathbf{X}}(y, \mathbf{x}_i, \boldsymbol{\beta})\right\}.$$

Step 3. Insert $\widehat{\mathbf{b}}^*(y, \boldsymbol{\beta})$ into the efficient score expression to obtain

$$\mathbf{S}^*_{\text{eff}}\{\mathbf{x}, r, ry, \boldsymbol{\beta}, \widehat{\mathbf{b}}^*(\cdot, \boldsymbol{\beta})\} = \frac{r\mathbf{f}_{\boldsymbol{\beta}}(y, \mathbf{x}; \boldsymbol{\beta})}{f_{Y|\mathbf{X}}(y, \mathbf{x}; \boldsymbol{\beta})} - \frac{(1-r)\int \mathbf{f}_{\boldsymbol{\beta}}(t, \mathbf{x}; \boldsymbol{\beta})\pi^*(t)d\mu(t)}{\int f_{Y|\mathbf{X}}(t, \mathbf{x}; \boldsymbol{\beta})\{1 - \pi^*(t)\}d\mu(t)} - r\widehat{\mathbf{b}}^*(y, \boldsymbol{\beta}) + \frac{(1-r)\int \widehat{\mathbf{b}}^*(t, \boldsymbol{\beta})f_{Y|\mathbf{X}}(t, \mathbf{x}; \boldsymbol{\beta})\pi^*(t)d\mu(t)}{\int f_{Y|\mathbf{X}}(t, \mathbf{x}; \boldsymbol{\beta})\{1 - \pi^*(t)\}d\mu(t)}.$$

Step 4. Solve the estimating equation $\sum_{i=1}^N \mathbf{S}^*_{\text{eff}}\{\mathbf{x}_i, r_i, r_i y_i, \boldsymbol{\beta}, \widehat{\mathbf{b}}^*(\cdot, \boldsymbol{\beta})\} = \mathbf{0}$ to obtain the estimator $\widehat{\boldsymbol{\beta}}$.

In Step 2 of Algorithm 1, (5) is a Type II Fredholm integral equation and has a unique solution, hence we obtain $\widehat{\mathbf{b}}^*(y, \boldsymbol{\beta})$

using the method proposed in Atkinson (1976). We point out that if we choose $\pi^*(y) = 0$ regardless of the fact that $\pi(y)$ is not a constant, then Algorithm 1 is much simplified and it will reduce to the pseudo likelihood estimator of Tang et al. (2003).

3.2. Theoretical Property

To theoretically analyze $\widehat{\boldsymbol{\beta}}$, the technical difficulties mainly stem from quantifying the difference between the solutions of the integral equations (4) and (5). To proceed, we first introduce some notation. We define

$$u_1(y) = E\{f_{Y|\mathbf{X}}(y, \mathbf{X}_i; \boldsymbol{\beta})\} = \int f_{Y|\mathbf{X}}(y, \mathbf{x}; \boldsymbol{\beta}) f_{\mathbf{X}}(\mathbf{x}) d\mu(\mathbf{x}),$$

$$u_2(t, y) = E\left\{\frac{f_{Y|\mathbf{X}}(y, \mathbf{X}_i; \boldsymbol{\beta}) f_{Y|\mathbf{X}}(t, \mathbf{X}_i; \boldsymbol{\beta})}{1 - \int f_{Y|\mathbf{X}}(t, \mathbf{X}_i; \boldsymbol{\beta}) \pi^*(t) d\mu(t)}\right\} \pi^*(t),$$

$$\mathbf{v}(y) = E\Bigg\{\mathbf{f}_{\boldsymbol{\beta}}(y, \mathbf{X}_i; \boldsymbol{\beta}) + \frac{\int \mathbf{f}_{\boldsymbol{\beta}}(t, \mathbf{X}_i; \boldsymbol{\beta}) \pi^*(t) d\mu(t)}{1 - \int f_{Y|\mathbf{X}}(t, \mathbf{X}_i; \boldsymbol{\beta}) \pi^*(t) d\mu(t)} f_{Y|\mathbf{X}}(y, \mathbf{X}_i; \boldsymbol{\beta})\Bigg\},$$

and the linear operation $\mathcal{A}(\cdot, y)$ on $\mathbf{b}(\cdot)$ as

$$\mathcal{A}(\mathbf{b})(y) \equiv \mathbf{b}(y) u_1(y) + \int \mathbf{b}(t) u_2(t, y) d\mu(t).$$

Similarly, let

$$u_{1i}(y) = f_{Y|\mathbf{X}}(y, \mathbf{x}_i; \boldsymbol{\beta}),$$

$$u_{2i}(t, y) = \frac{f_{Y|\mathbf{X}}(y, \mathbf{x}_i; \boldsymbol{\beta}) f_{Y|\mathbf{X}}(t, \mathbf{x}_i; \boldsymbol{\beta})}{1 - \int f_{Y|\mathbf{X}}(t, \mathbf{x}_i; \boldsymbol{\beta}) \pi^*(t) d\mu(t)} \pi^*(t),$$

$$\mathbf{v}_i(y) = \mathbf{f}_{\boldsymbol{\beta}}(y, \mathbf{x}_i; \boldsymbol{\beta}) + \frac{\int \mathbf{f}_{\boldsymbol{\beta}}(t, \mathbf{x}_i; \boldsymbol{\beta}) \pi^*(t) d\mu(t)}{1 - \int f_{Y|\mathbf{X}}(t, \mathbf{x}_i; \boldsymbol{\beta}) \pi^*(t) d\mu(t)} f_{Y|\mathbf{X}}(y, \mathbf{x}_i; \boldsymbol{\beta}).$$

Note that $u_{1i}, u_{2i}, \mathbf{v}_i$ depend on the ith observation only through $\mathbf{x}_i$. Also define

$$\widehat{u}_1(y) = N^{-1} \sum_{i=1}^{N} u_{1i}(y), \quad \widehat{u}_2(t, y) = N^{-1} \sum_{i=1}^{N} u_{2i}(t, y),$$

$$\widehat{\mathbf{v}}(y) = N^{-1} \sum_{i=1}^{N} \mathbf{v}_i(y).$$

Similar to $\mathcal{A}$, we define the linear operator

$$\mathcal{A}_i(\mathbf{b})(y) \equiv \mathbf{b}(y) u_{1i}(y) + \int \mathbf{b}(t) u_{2i}(t, y) d\mu(t),$$

$$\widehat{\mathcal{A}}(\mathbf{b})(y) \equiv N^{-1} \sum_{i=1}^{N} \mathcal{A}_i(\mathbf{b})(y).$$

We also introduce some regularity conditions.

(A1) $0 \le \pi^*(t) < 1 - \delta$ for all t, where $0 < \delta < 1$ is a constant.

(A2) The true parameter value of $\boldsymbol{\beta}$ belongs to a bounded domain. The support sets of $f_{\mathbf{X}}(\mathbf{x}), f_Y(y), \pi(y)$ are compact.

(A3) The functions $u_{1i}(y)$, $u_{2i}(t, y)$ are bounded and have bounded derivatives with respect to y and t on their support. The score function $\mathbf{S}_{\boldsymbol{\beta}}(\mathbf{x}, y; \boldsymbol{\beta}) \equiv \mathbf{f}_{\boldsymbol{\beta}}(y, \mathbf{x}; \boldsymbol{\beta}) / f_{Y|\mathbf{X}}(y, \mathbf{x}; \boldsymbol{\beta})$ is bounded, hence its orthogonal projection $\mathbf{b}^*(y)$ is also bounded.

Under these regularity conditions, the following result, with its proof in Appendix A.5, guarantees that $\|\mathcal{A}(\mathbf{b})\|_\infty$ is bounded above and below by a constant times $\|\mathbf{b}\|_\infty$, hence $\mathcal{A}$ is sufficiently well behaved.

Lemma 2. Under the regularity conditions (A1)–(A3), there exist constants $0 < c_1 < c_2 < \infty$ such that $c_1 \|\mathbf{b}\|_\infty \le \|\mathcal{A}(\mathbf{b})\|_\infty \le c_2 \|\mathbf{b}\|_\infty$.

Further, we have the following result, with its proof given in Appendix A.6, concerning the asymptotic distribution of $\widehat{\boldsymbol{\beta}}$.

Theorem 1. For any choice of $\pi^*(y)$, under conditions (A1)–(A3), $\widehat{\boldsymbol{\beta}}$ satisfies

$$\sqrt{N}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \to N\{\mathbf{0}, \mathbf{A}^{-1}\mathbf{B}(\mathbf{A}^{-1})^{\mathrm{T}}\},$$

in distribution when $N \to \infty$, where

$$\mathbf{A} = E\left[\frac{d\mathbf{S}^*_{\text{eff}}\{\mathbf{X}_i, R_i, R_i Y_i, \boldsymbol{\beta}, \mathbf{b}^*(\cdot, \boldsymbol{\beta})\}}{d\boldsymbol{\beta}^{\mathrm{T}}}\right],$$

$$\mathbf{B} = \text{var}[\mathbf{S}^*_{\text{eff}}\{\mathbf{X}_i, R_i, R_i Y_i, \boldsymbol{\beta}, \mathbf{b}^*(\cdot, \boldsymbol{\beta})\} - \mathbf{h}(\mathbf{X}_i)],$$

$$\mathbf{h}(\mathbf{x}_i) = \int \Big[\{\pi(y) - 1\}\{\mathbf{v}_i(y) - \mathcal{A}_i(\mathbf{b}^*)(y)\} + \mathcal{A}^{-1}\{\mathbf{v}_i - \mathcal{A}_i(\mathbf{b}^*)\}(y) u_1(y)\Big] d\mu(y).$$

Here

$$\frac{d\mathbf{S}^*_{\text{eff}}\{\mathbf{X}_i, R_i, R_i Y_i, \boldsymbol{\beta}, \mathbf{b}^*(\cdot, \boldsymbol{\beta})\}}{d\boldsymbol{\beta}^{\mathrm{T}}} \equiv \frac{\partial \mathbf{S}^*_{\text{eff}}\{\mathbf{X}_i, R_i, R_i Y_i, \boldsymbol{\beta}, \mathbf{b}^*(\cdot, \boldsymbol{\beta})\}}{\partial \boldsymbol{\beta}^{\mathrm{T}}} + \frac{\partial \mathbf{S}^*_{\text{eff}}\{\mathbf{X}_i, R_i, R_i Y_i, \boldsymbol{\beta}, \mathbf{b}^*(\cdot, \boldsymbol{\beta})\}}{\partial \mathbf{b}^{*\mathrm{T}}} \frac{\partial \mathbf{b}^*(\cdot, \boldsymbol{\beta})}{\partial \boldsymbol{\beta}^{\mathrm{T}}}.$$

Remark 1. One can easily verify that $E\{\mathbf{v}_i - \mathcal{A}_i(\mathbf{b}^*)\}(y) = \mathbf{0}$ and $E\mathcal{A}^{-1}\{\mathbf{v}_i - \mathcal{A}_i(\mathbf{b}^*)\}(y) = \mathbf{0}$, hence $\mathbf{h}(\mathbf{x}_i) \in \Lambda_{f_{\mathbf{X}}}$. Thus, if fortunately the working model $\pi^*(y)$ is chosen as the true mechanism $\pi(y)$, then $E(\mathbf{S}_{\text{eff}}\mathbf{h}^{\mathrm{T}}) = 0$ because $\mathbf{S}_{\text{eff}} \in \Lambda^{\perp}$. This means $\mathbf{B}$ in Theorem 1 can be further simplified to $\mathbf{B} = \text{var}[\mathbf{S}_{\text{eff}}\{\mathbf{X}_i, R_i, R_i Y_i, \boldsymbol{\beta}, \mathbf{b}(\cdot, \boldsymbol{\beta})\}] + \text{var}\{\mathbf{h}(\mathbf{X}_i)\}$ under this situation. Thus, we can view $h(\mathbf{x}_i)$ as the additional term to account for the cost from empirically approximating the integrals in (4).

4. Proposed Estimator Under General Assumption (1)

Under the general model (1), the joint pdf of $(\mathbf{X}, RY, R)$ is

$$f_{\mathbf{X}}(\mathbf{u}, \mathbf{z})\{f_{Y|\mathbf{X}}(y, \mathbf{u}, \mathbf{z}; \boldsymbol{\beta})\pi(y, \mathbf{u})\}^r \times \left\{1 - \int f_{Y|\mathbf{X}}(t, \mathbf{u}, \mathbf{z}; \boldsymbol{\beta})\pi(t, \mathbf{u}) d\mu(t)\right\}^{1-r},$$

where $\boldsymbol{\beta}$ is still the parameter of interest, and the functions $f_{\mathbf{X}}(\cdot)$ and $\pi(\cdot)$ are the nuisance parameters. Because the mechanism model $\pi(y, \mathbf{u})$ now also depends on partial covariate $\mathbf{u}$, the situation is much different from that considered in Section 3. We in fact show in Appendices A.7 and A.8 that the nuisance

tangent space orthogonal complement in this case is

$$\Lambda^{\perp} = \big\{\mathbf{a}(\mathbf{u},\mathbf{z},r,ry) : E\{\mathbf{a}(\mathbf{u},\mathbf{z},R,RY) \mid \mathbf{u},\mathbf{z}\} = \mathbf{0}, \\ E\big[\{\mathbf{a}(\mathbf{u},\mathbf{Z},1,y) - \mathbf{a}(\mathbf{u},\mathbf{Z},0,0)\} \mid y,\mathbf{u}\big] = \mathbf{0}\big\},$$

and the efficient score for parameter $\boldsymbol{\beta}$ is

$$\begin{aligned}&\mathbf{S}_{\text{eff}}(\mathbf{u},\mathbf{z},r,ry)\\ &= \frac{r\mathbf{f}_{\boldsymbol{\beta}}(y,\mathbf{u},\mathbf{z};\boldsymbol{\beta})}{f_{Y|\mathbf{X}}(y,\mathbf{u},\mathbf{z};\boldsymbol{\beta})} - \frac{(1-r)\int \mathbf{f}_{\boldsymbol{\beta}}(t,\mathbf{u},\mathbf{z};\boldsymbol{\beta})\pi(t,\mathbf{u})d\mu(t)}{\int f_{Y|\mathbf{X}}(t,\mathbf{u},\mathbf{z};\boldsymbol{\beta})\{1-\pi(t,\mathbf{u})\}d\mu(t)}\\ &\quad -r\mathbf{b}(y,\mathbf{u}) + \frac{(1-r)\int \mathbf{b}(t,\mathbf{u})f_{Y|\mathbf{X}}(t,\mathbf{u},\mathbf{z};\boldsymbol{\beta})\pi(t,\mathbf{u})d\mu(t)}{\int f_{Y|\mathbf{X}}(t,\mathbf{u},\mathbf{z};\boldsymbol{\beta})\{1-\pi(t,\mathbf{u})\}d\mu(t)},\end{aligned}$$

where $\mathbf{f}_{\boldsymbol{\beta}}(y,\mathbf{u},\mathbf{z};\boldsymbol{\beta}) \equiv \partial f_{Y|\mathbf{X}}(y,\mathbf{u},\mathbf{z};\boldsymbol{\beta})/\partial\boldsymbol{\beta}$, and $\mathbf{b}(y,\mathbf{u})$ satisfies the integral equation

$$\begin{aligned}&\int\Bigg[\frac{\mathbf{f}_{\boldsymbol{\beta}}(y,\mathbf{u},\mathbf{z};\boldsymbol{\beta})}{f_{Y|\mathbf{X}}(y,\mathbf{u},\mathbf{z};\boldsymbol{\beta})} - \mathbf{b}(y,\mathbf{u})\\ &+\frac{\int \mathbf{f}_{\boldsymbol{\beta}}(t,\mathbf{u},\mathbf{z};\boldsymbol{\beta})\pi(t,\mathbf{u})d\mu(t)}{\int f_{Y|\mathbf{X}}(t,\mathbf{u},\mathbf{z};\boldsymbol{\beta})\{1-\pi(t,\mathbf{u})\}d\mu(t)}\\ &-\frac{\int \mathbf{b}(t,\mathbf{u})f_{Y|\mathbf{X}}(t,\mathbf{u},\mathbf{z};\boldsymbol{\beta})\pi(t,\mathbf{u})d\mu(t)}{\int f_{Y|\mathbf{X}}(t,\mathbf{u},\mathbf{z};\boldsymbol{\beta})\{1-\pi(t,\mathbf{u})\}d\mu(t)}\Bigg]\\ &\times f_{\mathbf{Z}|\mathbf{U}}(\mathbf{z},\mathbf{u})f_{Y|\mathbf{X}}(y,\mathbf{u},\mathbf{z};\boldsymbol{\beta})d\mu(\mathbf{z}) = \mathbf{0}. \qquad (6)\end{aligned}$$

Similar to the complete shadow case in Section 3, the solution $\mathbf{b}(y,\mathbf{u})$ exists and is unique. Note that we used the decomposition $f_{\mathbf{X}}(\mathbf{u},\mathbf{z}) = f_{\mathbf{Z}|\mathbf{U}}(\mathbf{z},\mathbf{u})f_{\mathbf{U}}(\mathbf{u})$ above. We can see that the space $\Lambda^{\perp}$ has slightly different form from its counterpart in Section 3, caused by the additional inclusion of $\mathbf{U}$ in the mechanism model. Nevertheless, in Appendix A.9 we verify that a misspecified $\pi(y,\mathbf{u})$ model, $\pi^*(y,\mathbf{u})$, can be employed in the construction of $\mathbf{S}_{\text{eff}}(\mathbf{u},\mathbf{z},r,ry)$ and the mean zero property of the efficient score will still retain.

In an effort to construct an estimator similar in spirit to $\widehat{\boldsymbol{\beta}}$ in Section 3, we realize that we would have to handle the $\mathbf{U}$ part and the $\mathbf{Z}$ part of the covariates differently because they play different roles. In fact, while we could be totally "empirical" with respect to $\mathbf{Z}$, we would have to remain "nonparametric" with respect to $\mathbf{U}$. Specifically, recognizing that the left hand side of (6) is a conditional expectation, we approximate the integral equation (6) by

$$\begin{aligned}&\frac{1}{N}\sum_{i=1}^{N}\Bigg[\frac{\mathbf{f}_{\boldsymbol{\beta}}(y,\mathbf{u},\mathbf{z}_i;\boldsymbol{\beta})}{f_{Y|\mathbf{X}}(y,\mathbf{u},\mathbf{z}_i;\boldsymbol{\beta})} - \mathbf{b}(y,\mathbf{u})\\ &+\frac{\int \mathbf{f}_{\boldsymbol{\beta}}(t,\mathbf{u},\mathbf{z}_i;\boldsymbol{\beta})\pi^*(t,\mathbf{u})d\mu(t)}{\int f_{Y|\mathbf{X}}(t,\mathbf{u},\mathbf{z}_i;\boldsymbol{\beta})\{1-\pi^*(t,\mathbf{u})\}d\mu(t)}\\ &-\frac{\int \mathbf{b}(t,\mathbf{u})f_{Y|\mathbf{X}}(t,\mathbf{u},\mathbf{z}_i;\boldsymbol{\beta})\pi^*(t,\mathbf{u})d\mu(t)}{\int f_{Y|\mathbf{X}}(t,\mathbf{u},\mathbf{z}_i;\boldsymbol{\beta})\{1-\pi^*(t,\mathbf{u})\}d\mu(t)}\Bigg]\\ &\times f_{Y|\mathbf{X}}(y,\mathbf{u},\mathbf{z}_i;\boldsymbol{\beta})K_h(\mathbf{u}_i-\mathbf{u}) = \mathbf{0}, \qquad (7)\end{aligned}$$

utilizing the nonparametric regression technique, where $K_h(\cdot) = K(\cdot/h)/h$, $K(\cdot)$ is a kernel function and h is a bandwidth, with their conditions detailed later. Once $\widehat{\mathbf{b}}^*(t,\mathbf{u})$ is obtained from solving (7), we can then proceed to construct the estimating equation and obtain the estimator. For completeness, we write out the algorithm.

Algorithm 2. Algorithm under general assumption (1)

Step 1. Posit a working model for $\pi(y,\mathbf{u})$, denoted as $\pi^*(y,\mathbf{u})$.
Step 2. Obtain $\widehat{\mathbf{b}}^*(y,\mathbf{u},\boldsymbol{\beta})$ by solving the integral equation (7).
Step 3. Insert $\widehat{\mathbf{b}}^*(y,\mathbf{u},\boldsymbol{\beta})$ into the efficient score expression to obtain

$$\begin{aligned}&\mathbf{S}^*_{\text{eff}}\{\mathbf{u},\mathbf{z},r,ry,\boldsymbol{\beta},\widehat{\mathbf{b}}^*(\cdot,\mathbf{u},\boldsymbol{\beta})\}\\ &= \frac{r\mathbf{f}_{\boldsymbol{\beta}}(y,\mathbf{u},\mathbf{z};\boldsymbol{\beta})}{f_{Y|\mathbf{X}}(y,\mathbf{u},\mathbf{z};\boldsymbol{\beta})} - \frac{(1-r)\int \mathbf{f}_{\boldsymbol{\beta}}(t,\mathbf{u},\mathbf{z};\boldsymbol{\beta})\pi^*(t,\mathbf{u})d\mu(t)}{\int f_{Y|\mathbf{X}}(t,\mathbf{u},\mathbf{z};\boldsymbol{\beta})\{1-\pi^*(t,\mathbf{u})\}d\mu(t)}\\ &\quad -r\widehat{\mathbf{b}}^*(y,\mathbf{u},\boldsymbol{\beta})\\ &\quad +\frac{(1-r)\int \widehat{\mathbf{b}}^*(t,\mathbf{u},\boldsymbol{\beta})f_{Y|\mathbf{X}}(t,\mathbf{u},\mathbf{z};\boldsymbol{\beta})\pi^*(t,\mathbf{u})d\mu(t)}{\int f_{Y|\mathbf{X}}(t,\mathbf{u},\mathbf{z};\boldsymbol{\beta})\{1-\pi^*(t,\mathbf{u})\}d\mu(t)}.\end{aligned}$$

Step 4. Solve the estimating equation $\sum_{i=1}^{N}\mathbf{S}^*_{\text{eff}}\{\mathbf{u}_i,\mathbf{z}_i,r_i,r_iy_i,\boldsymbol{\beta},\widehat{\mathbf{b}}^*(\cdot,\mathbf{u}_i,\boldsymbol{\beta})\} = \mathbf{0}$ to obtain the estimator for $\boldsymbol{\beta}$. We still denote as $\widehat{\boldsymbol{\beta}}$.

Like (5), (7) is also a Type II Fredholm integral equation and has a unique solution, so it can also be solved by the method proposed in Atkinson (1976). Note that only $\widehat{\mathbf{b}}^*(y,\mathbf{u}_i,\boldsymbol{\beta})$'s are needed in Algorithm 2, instead of the generic function $\widehat{\mathbf{b}}^*(y,\mathbf{u},\boldsymbol{\beta})$. If we happen to select $\pi^*(y,\mathbf{u}) = 0$, even though we are aware that $\pi(y,\mathbf{u})$ is not a constant, the computation will be much simplified and the estimator degenerates to that in Zhao and Shao (2015).

To study the theoretical property of $\widehat{\boldsymbol{\beta}}$, the technical difficulties mainly stem from quantifying the difference between the solutions of the integral equations (6) and (7). To proceed, we first introduce some notation. We define

$$\begin{aligned}u_1(y,\mathbf{u}) &= f_{\mathbf{U}}(\mathbf{u})E\left\{f_{Y|\mathbf{X}}(y,\mathbf{u},\mathbf{Z}_i;\boldsymbol{\beta}) \mid \mathbf{U}=\mathbf{u}\right\}\\ &= f_{\mathbf{U}}(\mathbf{u})\int f_{Y|\mathbf{X}}(y,\mathbf{u},\mathbf{z};\boldsymbol{\beta})f_{\mathbf{Z}|\mathbf{U}}(\mathbf{z},\mathbf{u})d\mu(\mathbf{z}),\\ u_2(t,y,\mathbf{u}) &= f_{\mathbf{U}}(\mathbf{u})E\\ &\quad\left\{\frac{f_{Y|\mathbf{X}}(y,\mathbf{u},\mathbf{Z}_i;\boldsymbol{\beta})f_{Y|\mathbf{X}}(t,\mathbf{u},\mathbf{Z}_i;\boldsymbol{\beta})}{1-\int f_{Y|\mathbf{X}}(t,\mathbf{u},\mathbf{Z}_i;\boldsymbol{\beta})\pi^*(t,\mathbf{u})d\mu(t)} \mid \mathbf{U}=\mathbf{u}\right\}\\ &\quad\times\pi^*(t,\mathbf{u}),\\ \mathbf{v}(y,\mathbf{u}) &= f_{\mathbf{U}}(\mathbf{u})E\Bigg\{\mathbf{f}_{\boldsymbol{\beta}}(y,\mathbf{u},\mathbf{Z}_i;\boldsymbol{\beta})\\ &\quad+\frac{\int \mathbf{f}_{\boldsymbol{\beta}}(t,\mathbf{u},\mathbf{Z}_i;\boldsymbol{\beta})\pi^*(t,\mathbf{u})d\mu(t)}{1-\int f_{Y|\mathbf{X}}(t,\mathbf{u},\mathbf{Z}_i;\boldsymbol{\beta})\pi^*(t,\mathbf{u})d\mu(t)}\\ &\quad\times f_{Y|\mathbf{X}}(y,\mathbf{u},\mathbf{Z}_i;\boldsymbol{\beta}) \mid \mathbf{U}=\mathbf{u}\Bigg\},\end{aligned}$$

and the linear operation $\mathcal{A}(\cdot,y,\mathbf{u})$ on $\mathbf{b}(\cdot)$ as

$$\mathcal{A}(\mathbf{b})(y,\mathbf{u}) \equiv \mathbf{b}(y,\mathbf{u})u_1(y,\mathbf{u}) + \int \mathbf{b}(t,\mathbf{u})u_2(t,y,\mathbf{u})d\mu(t).$$

Similarly, let

$$\begin{aligned}u_{1i}(y,\mathbf{u}) &= K_h(\mathbf{u}_i-\mathbf{u})f_{Y|\mathbf{X}}(y,\mathbf{u},\mathbf{z}_i;\boldsymbol{\beta}),\\ u_{2i}(t,y,\mathbf{u}) &= K_h(\mathbf{u}_i-\mathbf{u})\frac{f_{Y|\mathbf{X}}(y,\mathbf{u},\mathbf{z}_i;\boldsymbol{\beta})f_{Y|\mathbf{X}}(t,\mathbf{u},\mathbf{z}_i;\boldsymbol{\beta})}{1-\int f_{Y|\mathbf{X}}(t,\mathbf{u},\mathbf{z}_i;\boldsymbol{\beta})\pi^*(t,\mathbf{u})d\mu(t)}\\ &\quad\times\pi^*(t,\mathbf{u}),\end{aligned}$$

$$\mathbf{v}_i(y,\mathbf{u}) = K_h(\mathbf{u}_i-\mathbf{u})\Big\{\mathbf{f}_{\boldsymbol\beta}(y,\mathbf{u},\mathbf{z}_i;\boldsymbol\beta) + \frac{\int \mathbf{f}_{\boldsymbol\beta}(t,\mathbf{u},\mathbf{z}_i;\boldsymbol\beta)\pi^*(t,\mathbf{u})d\mu(t)}{1-\int f_{Y|\mathbf{X}}(t,\mathbf{u},\mathbf{z}_i;\boldsymbol\beta)\pi^*(t,\mathbf{u})d\mu(t)} \times f_{Y|\mathbf{X}}(y,\mathbf{u},\mathbf{z}_i;\boldsymbol\beta)\Big\}.$$

Note that $u_{1i}, u_{2i}, \mathbf{v}_i$ depend on the ith observation only through $\mathbf{x}_i$. Also define

$$\widehat{u}_1(y,\mathbf{u}) = N^{-1}\sum_{i=1}^N u_{1i}(y,\mathbf{u}),$$

$$\widehat{u}_2(t,y,\mathbf{u}) = N^{-1}\sum_{i=1}^N u_{2i}(t,y,\mathbf{u}), \quad \widehat{\mathbf{v}}(y,\mathbf{u}) = N^{-1}\sum_{i=1}^N \mathbf{v}_i(y,\mathbf{u}).$$

Similar to $\mathcal{A}$, we define the linear operator

$$\mathcal{A}_i(\mathbf{b})(y,\mathbf{u}) \equiv \mathbf{b}(y,\mathbf{u})u_{1i}(y,\mathbf{u}) + \int \mathbf{b}(t,\mathbf{u})u_{2i}(t,y,\mathbf{u})d\mu(t),$$

$$\widehat{\mathcal{A}}(\mathbf{b})(y,\mathbf{u}) \equiv N^{-1}\sum_{i=1}^N \mathcal{A}_i(\mathbf{b})(y,\mathbf{u}).$$

We need the following conditions on the true functions, kernel function and the bandwidth.

(B1) $0 \le \pi^*(t,\mathbf{u}) < 1-\delta$ for all $(t,\mathbf{u})$, where $0 < \delta < 1$ is a constant.

(B2) The true parameter value of $\boldsymbol\beta$ belongs to a bounded domain. The support sets of $f_{\mathbf{Z}|\mathbf{U}}(\mathbf{z},\mathbf{u}), f_Y(y), \pi(y,\mathbf{u})$ are compact.

(B3) The functions $u_{1i}(y,\mathbf{u}), u_{2i}(t,y,\mathbf{u})$ are bounded and have bounded derivatives with respect to y and t on their support. The score function $\mathbf{S}_{\boldsymbol\beta}(\mathbf{u},\mathbf{z},y;\boldsymbol\beta) \equiv \mathbf{f}_{\boldsymbol\beta}(y,\mathbf{u},\mathbf{z};\boldsymbol\beta)/f_{Y|\mathbf{X}}(y,\mathbf{u},\mathbf{z};\boldsymbol\beta)$ is bounded, hence its orthogonal projection $\mathbf{b}^*(y,\mathbf{u})$ is also bounded.

(B4) The univariate kernel function $K(\cdot)$ is bounded and symmetric, has a bounded derivative and compact support $[-1,1]$, and satisfies $\int K(u)du = 1$, $\mu_m = \int u^m K(u)du \neq 0$, $\int u^r K(u)du = 0$ for $r = 1,\ldots,m-1$. $K_h(u) = K(u/h)/h$. The d-dimensional kernel function is a product of d univariate kernel functions, that is, $K(\mathbf{u}) = \prod_{j=1}^d K(u_j)$, and $K_h(\mathbf{u}) = \prod_{j=1}^d K_h(u_j) = h^{-d}\prod_{j=1}^d K(u_j/h)$ for $\mathbf{u} = (u_1,\ldots,u_d)^{\mathrm{T}}$ and bandwidth h. Here d is the dimension of $\mathbf{u}$.

(B5) The bandwidth h satisfies $h \to 0$, $Nh^{2d} \to \infty$ and $Nh^{2m} \to 0$.

Under these regularity conditions, we have the following lemma to ensure that $\|\mathcal{A}(\mathbf{b})\|_\infty$ is bounded by $\|\mathbf{b}\|_\infty$. Its proof is in Appendix A.10,

Lemma 3. Under the regularity conditions (B1)–(B3), there exist constants $0 < c_1 < c_2 < \infty$ such that $c_1\|\mathbf{b}\|_\infty \le \|\mathcal{A}(\mathbf{b})\|_\infty \le c_2\|\mathbf{b}\|_\infty$.

The theoretical property of $\widehat{\boldsymbol\beta}$ is summarized in Theorem 2, with its proof in Appendix A.11.

Theorem 2. For any choice of $\pi^*(y,\mathbf{u})$, under conditions (B1)–(B5), $\widehat{\boldsymbol\beta}$ satisfies

$$\sqrt{N}(\widehat{\boldsymbol\beta}-\boldsymbol\beta) \to N\{\mathbf{0}, \mathbf{A}^{-1}\mathbf{B}(\mathbf{A}^{-1})^{\mathrm{T}}\},$$

in distribution when $N \to \infty$, where

$$\mathbf{A} = E\left[\frac{d\mathbf{S}^*_{\text{eff}}\{\mathbf{X}_i,R_i,R_iY_i,\boldsymbol\beta,\mathbf{b}^*(\cdot,\boldsymbol\beta)\}}{d\boldsymbol\beta^{\mathrm{T}}}\right],$$

$$\mathbf{B} = \text{var}[\mathbf{S}^*_{\text{eff}}\{\mathbf{X}_i,R_i,R_iY_i,\boldsymbol\beta,\mathbf{b}^*(\cdot,\boldsymbol\beta)\} - \mathbf{h}(\mathbf{X}_i)],$$

$$\mathbf{h}(\mathbf{x}_i) = \int \big[\{\pi(y,\mathbf{u})-1\}\{\mathbf{v}_i(y,\mathbf{u}) - \mathcal{A}_i(\mathbf{b}^*)(y,\mathbf{u})\} + \mathcal{A}^{-1}\{\mathbf{v}_i - \mathcal{A}_i(\mathbf{b}^*)\}(y,\mathbf{u})u_1(y,\mathbf{u})\big]d\mu(y,\mathbf{u}).$$

Here

$$\frac{d\mathbf{S}^*_{\text{eff}}\{\mathbf{X}_i,R_i,R_iY_i,\boldsymbol\beta,\mathbf{b}^*(\cdot,\boldsymbol\beta)\}}{d\boldsymbol\beta^{\mathrm{T}}} \equiv \frac{\partial\mathbf{S}^*_{\text{eff}}\{\mathbf{X}_i,R_i,R_iY_i,\boldsymbol\beta,\mathbf{b}^*(\cdot,\boldsymbol\beta)\}}{\partial\boldsymbol\beta^{\mathrm{T}}} + \frac{\partial\mathbf{S}^*_{\text{eff}}\{\mathbf{X}_i,R_i,R_iY_i,\boldsymbol\beta,\mathbf{b}^*(\cdot,\boldsymbol\beta)\}}{\partial\mathbf{b}^{*\mathrm{T}}}\frac{\partial\mathbf{b}^*(\cdot,\boldsymbol\beta)}{\partial\boldsymbol\beta^{\mathrm{T}}}.$$

Remark 2. Theorem 2 has a similar $h(\mathbf{x}_i)$ term as in Theorem 1. Similar to Remark 1, this term can also be viewed as the additional cost from replacing the integral in (6) with its approximation in (7).

Remark 3. So far in this section, we have implicitly assumed that $\mathbf{U}$ is continuous. When $\mathbf{U}$ contains discrete component, we only need to stratify the data according to the different discrete values, then construct the corresponding integral equations within each stratum according to either (5) or (7). Solving these integral equations will then provide $\widehat{\mathbf{b}}^*(y,\mathbf{u},\boldsymbol\beta)$ and the remaining estimation procedures are completely identical to the last two steps in Algorithm 2. Specifically, for discrete $\mathbf{U}$, assume that $\mathbf{U}$ can be $\mathbf{u}^0_k, k = 1,\ldots,K$. Then, we replace (5) with

$$\frac{1}{N_k}\sum_{i=1,\mathbf{u}_i=\mathbf{u}^0_k}^N \left[\frac{\mathbf{f}_{\boldsymbol\beta}(y,\mathbf{u}^0_k,\mathbf{z}_i;\boldsymbol\beta)}{f_{Y|\mathbf{X}}(y,\mathbf{u}^0_k,\mathbf{z}_i;\boldsymbol\beta)} - \mathbf{b}(y,\mathbf{u}^0_k) + \frac{\int \mathbf{f}_{\boldsymbol\beta}(t,\mathbf{u}^0_k,\mathbf{z}_i;\boldsymbol\beta)\pi^*(t,\mathbf{u}^0_k)d\mu(t)}{\int f_{Y|\mathbf{X}}(t,\mathbf{u}^0_k,\mathbf{z}_i;\boldsymbol\beta)\{1-\pi^*(t,\mathbf{u}^0_k)\}d\mu(t)} - \frac{\int \mathbf{b}(t,\mathbf{u}^0_k)f_{Y|\mathbf{X}}(t,\mathbf{u}^0_k,\mathbf{z}_i;\boldsymbol\beta)\pi^*(t,\mathbf{u}^0_k)d\mu(t)}{\int f_{Y|\mathbf{X}}(t,\mathbf{u}^0_k,\mathbf{z}_i;\boldsymbol\beta)\{1-\pi^*(t,\mathbf{u}^0_k)\}d\mu(t)}\right] \times f_{Y|\mathbf{X}}(y,\mathbf{u}^0_k,\mathbf{z}_i;\boldsymbol\beta) = \mathbf{0},$$

where $N_k = \sum_{i=1}^N I(\mathbf{u}_i = \mathbf{u}^0_k)$, and solve it to obtain $\widehat{\mathbf{b}}^*(y,\mathbf{u}^0_k,\boldsymbol\beta)$. If $\mathbf{U}$ is a mix of discrete ($\mathbf{U}_d$) and continuous ($\mathbf{U}_c$) variables, say $\mathbf{U} = (\mathbf{U}_d^{\mathrm{T}},\mathbf{U}_c^{\mathrm{T}})^{\mathrm{T}}$. Assume that $\mathbf{U}_d$ can be $\mathbf{u}^0_{dk}, k = 1,\ldots,K$. We then replace (7) with

$$\frac{1}{N_k}\sum_{i=1,\mathbf{u}_{di}=\mathbf{u}^0_{dk}}^N \left[\frac{\mathbf{f}_{\boldsymbol\beta}(y,\mathbf{u}^0_{dk},\mathbf{u}_{ci},\mathbf{z}_i;\boldsymbol\beta)}{f_{Y|\mathbf{X}}(y,\mathbf{u}^0_{dk},\mathbf{u}_{ci},\mathbf{z}_i;\boldsymbol\beta)} - \mathbf{b}(y,\mathbf{u}^0_{dk},\mathbf{u}_{ci}) + \frac{\int \mathbf{f}_{\boldsymbol\beta}(t,\mathbf{u}^0_{dk},\mathbf{u}_{ci},\mathbf{z}_i;\boldsymbol\beta)\pi^*(t,\mathbf{u}^0_{dk},\mathbf{u}_{ci})d\mu(t)}{\int f_{Y|\mathbf{X}}(t,\mathbf{u}^0_{dk},\mathbf{u}_{ci},\mathbf{z}_i;\boldsymbol\beta)\{1-\pi^*(t,\mathbf{u}^0_{dk},\mathbf{u}_{ci})\}d\mu(t)}\right]$$

$$-\frac{\int \mathbf{b}(t,\mathbf{u}_{dk}^0,\mathbf{u}_{ci})f_{Y|\mathbf{X}}(t,\mathbf{u}_{dk}^0,\mathbf{u}_{ci},\mathbf{z}_i;\boldsymbol{\beta})\pi^*(t,\mathbf{u}_{dk}^0,\mathbf{u}_{ci})d\mu(t)}{\int f_{Y|\mathbf{X}}(t,\mathbf{u}_{dk}^0,\mathbf{u}_{ci},\mathbf{z}_i;\boldsymbol{\beta})\{1-\pi^*(t,\mathbf{u}_{dk}^0,\mathbf{u}_{ci})\}d\mu(t)}\Bigg]$$
$$\times f_{Y|\mathbf{X}}(y,\mathbf{u}_{dk}^0,\mathbf{u}_{ci},\mathbf{z}_i;\boldsymbol{\beta})K_h(\mathbf{u}_{ci}-\mathbf{u}_c)=\mathbf{0},$$

where $N_k = \sum_{i=1}^N I(\mathbf{u}_{di} = \mathbf{u}_{dk}^0)$, and solve it to obtain $\widehat{\mathbf{b}}^*(y,\mathbf{u}_{dk}^0,\mathbf{u}_c,\boldsymbol{\beta})$.

5. Other Estimators

In Sections 3 and 4, we proposed estimator $\widehat{\boldsymbol{\beta}}$ with minimum assumption regarding estimating or modeling $f_{\mathbf{X}}(\mathbf{x})$ and $f_{\mathbf{Z}|\mathbf{U}}(\mathbf{z},\mathbf{u})$. If we are willing and able to adopt further modeling and estimation procedures to assess $f_{\mathbf{X}}(\mathbf{x})$ and $f_{\mathbf{Z}|\mathbf{U}}(\mathbf{z},\mathbf{u})$, different estimators for $\boldsymbol{\beta}$ can be obtained. We illustrate two alternative estimators.

First, instead of approximating the expectations empirically, we can use nonparametric kernel method in both Sections 3 and 4. For example, in Section 3, we can approximate $f_{\mathbf{X}}(\cdot)$ via $\widehat{f}_{\mathbf{X}}(\mathbf{x}) = N^{-1}\sum_{i=1}^N K_h(\mathbf{x}_i-\mathbf{x})$, then insert it into (4) to form an approximate integral equation. We denote the resulting estimator $\widehat{\boldsymbol{\beta}}_{\text{non}}$. We summarize its property below, with its proof in Appendix A.12.

Theorem 3. For any choice of $\pi^*(y)$, under conditions (A1)–(A3), if $Nh^{2m}\to 0$, then $\widehat{\boldsymbol{\beta}}_{\text{non}}$ satisfies

$$\sqrt{N}(\widehat{\boldsymbol{\beta}}_{\text{non}}-\boldsymbol{\beta})\to N\{\mathbf{0},\mathbf{A}_{\text{non}}^{-1}\mathbf{B}_{\text{non}}(\mathbf{A}_{\text{non}}^{-1})^{\mathrm{T}}\},$$

in distribution when $N\to\infty$, where

$$\mathbf{A}_{\text{non}} = E\left[\frac{d\mathbf{S}_{\text{eff}}^*\{\mathbf{X}_i,R_i,R_iY_i,\boldsymbol{\beta},\mathbf{b}^*(\cdot,\boldsymbol{\beta})\}}{d\boldsymbol{\beta}^{\mathrm{T}}}\right],$$
$$\mathbf{B}_{\text{non}} = \text{var}[\mathbf{S}_{\text{eff}}^*\{\mathbf{X}_i,R_i,R_iY_i,\boldsymbol{\beta},\mathbf{b}^*(\cdot,\boldsymbol{\beta})\}-\mathbf{h}(\mathbf{X}_i)],$$
$$\mathbf{h}(\mathbf{x}_i) = \int\big[\{\pi(y)-1\}\{\mathbf{v}_i(y)-\mathcal{A}_i(\mathbf{b}^*)(y)\}$$
$$+\mathcal{A}^{-1}\{\mathbf{v}_i-\mathcal{A}_i(\mathbf{b}^*)\}(y)u_1(y)\big]d\mu(y).$$

Here

$$\frac{d\mathbf{S}_{\text{eff}}^*\{\mathbf{X}_i,R_i,R_iY_i,\boldsymbol{\beta},\mathbf{b}^*(\cdot,\boldsymbol{\beta})\}}{d\boldsymbol{\beta}^{\mathrm{T}}}$$
$$\equiv\frac{\partial\mathbf{S}_{\text{eff}}^*\{\mathbf{X}_i,R_i,R_iY_i,\boldsymbol{\beta},\mathbf{b}^*(\cdot,\boldsymbol{\beta})\}}{\partial\boldsymbol{\beta}^{\mathrm{T}}}$$
$$+\frac{\partial\mathbf{S}_{\text{eff}}^*\{\mathbf{X}_i,R_i,R_iY_i,\boldsymbol{\beta},\mathbf{b}^*(\cdot,\boldsymbol{\beta})\}}{\partial\mathbf{b}^{*\mathrm{T}}}\frac{\partial\mathbf{b}^*(\cdot,\boldsymbol{\beta})}{\partial\boldsymbol{\beta}^{\mathrm{T}}}.$$

Remark 4. Similar to Theorem 3, under the assumption (1) in Section 4, a pure nonparametric kernel based estimator can also be derived. We omit the details to avoid repetition.

Remark 5. Similar to the discussion in Section 4, in the above analysis of $\widehat{\boldsymbol{\beta}}_{\text{non}}$, we have assumed that all components in $\mathbf{X}$ are continuous. If $\mathbf{X}$ contains discrete components, say $\mathbf{X} = (\mathbf{X}_c^{\mathrm{T}},\mathbf{X}_d^{\mathrm{T}})^{\mathrm{T}}$, where $\mathbf{X}_c$ consists of continuous variables and $\mathbf{X}_d$ is the collection of discrete variables, then we need to slightly adjust the procedure. Specifically, let $\mathbf{X}_d$ have values $\mathbf{x}_{dk}^0, k = 1,\ldots,K$. We would stratify the data into K strata. Within each stratum, we treat $\mathbf{X}_c$ as the new $\mathbf{X}$ variable and write the kernel estimator $\widehat{f}_{\mathbf{X}_c|\mathbf{X}_d=\mathbf{x}_{dk}^0}$ as $\widehat{f}_{\mathbf{X}_c,k}$. The integral Equation (4) is then approximated by

$$\sum_{k=1}^K\widehat{p}_k\int\Bigg\{\mathbf{f}_{\boldsymbol{\beta}}(y,\mathbf{x}_c,\mathbf{x}_{dk};\boldsymbol{\beta})\widehat{f}_{\mathbf{X}_c,k}(\mathbf{x}_c)$$
$$+\frac{\int\mathbf{f}_{\boldsymbol{\beta}}(t,\mathbf{x}_c,\mathbf{x}_{dk};\boldsymbol{\beta})\pi^*(t)d\mu(t)}{1-\int f_{Y|\mathbf{X}}(t,\mathbf{x}_c,\mathbf{x}_{dk};\boldsymbol{\beta})\pi^*(t)d\mu(t)}$$
$$\times f_{Y|\mathbf{X}}(y,\mathbf{x}_c,\mathbf{x}_{dk};\boldsymbol{\beta})\widehat{f}_{\mathbf{X}_c,k}(\mathbf{x}_c)\Bigg\}d\mu(\mathbf{x}_c)$$
$$=\sum_{k=1}^K\widehat{p}_k\int\Bigg\{\mathbf{b}(y)f_{Y|\mathbf{X}}(y,\mathbf{x}_c,\mathbf{x}_{dk};\boldsymbol{\beta})\widehat{f}_{\mathbf{X}_c,k}(\mathbf{x}_c)$$
$$+\frac{\int\mathbf{b}(t)f_{Y|\mathbf{X}}(t,\mathbf{x}_c,\mathbf{x}_{dk};\boldsymbol{\beta})\pi^*(t)d\mu(t)}{1-\int f_{Y|\mathbf{X}}(t,\mathbf{x}_c,\mathbf{x}_{dk};\boldsymbol{\beta})\pi^*(t)d\mu(t)}$$
$$\times f_{Y|\mathbf{X}}(y,\mathbf{x}_c,\mathbf{x}_{dk},\boldsymbol{\beta})\widehat{f}_{\mathbf{X}_c,k}(\mathbf{x}_c)\Bigg\}d\mu(\mathbf{x}_c),$$

where $\widehat{p}_k$ is the empirical frequency of observations in the kth stratum. After solving the integral equation, we still proceed to the same estimating equation in Algorithm 1.

Second, we consider parametric estimation of $f_{\mathbf{X}}(\mathbf{x})$ and $f_{\mathbf{Z}|\mathbf{U}}(\mathbf{z},\mathbf{u})$, that is, $f_{\mathbf{X}}(\mathbf{x};\widehat{\boldsymbol{\alpha}})$ in Section 3 and $f_{\mathbf{Z}|\mathbf{U}}(\mathbf{z},\mathbf{u};\widehat{\boldsymbol{\alpha}})$ in Section 4. This scenario can arise in the situation when one is confident to correctly specify a parametric model, using all fully observed data. For convenience, we assume $N^{1/2}(\widehat{\boldsymbol{\alpha}}-\boldsymbol{\alpha}) = N^{-1/2}\sum_{i=1}^N\boldsymbol{\phi}(\mathbf{x}_i;\boldsymbol{\alpha})+o_p(1)$, which is the typical expansion for most full data parametric estimators. For example, when maximum likelihood estimator (MLE) is used, $\boldsymbol{\phi}(\mathbf{x}_i;\boldsymbol{\alpha}) = -[E\{\partial^2\log f_{\mathbf{X}}(\mathbf{x};\boldsymbol{\alpha})/\partial\boldsymbol{\alpha}\partial\boldsymbol{\alpha}^{\mathrm{T}}\}]^{-1}\partial\log f_{\mathbf{X}}(\mathbf{x}_i;\boldsymbol{\alpha})/\partial\boldsymbol{\alpha}$. We call the corresponding estimator $\widehat{\boldsymbol{\beta}}_{\text{par}}$. For $\widehat{\boldsymbol{\beta}}_{\text{par}}$ we have the following asymptotic result and its proof is in Appendix A.13.

Theorem 4. For both the special assumption (2) with an arbitrary choice of $\pi^*(y)$, and the general assumption (1) with an arbitrary choice of $\pi^*(y,\mathbf{u})$, the corresponding estimator $\widehat{\boldsymbol{\beta}}_{\text{par}}$ satisfies

$$\sqrt{N}(\widehat{\boldsymbol{\beta}}_{\text{par}}-\boldsymbol{\beta})\to N\{\mathbf{0},\mathbf{A}_{\text{par}}^{-1}\mathbf{B}_{\text{par}}(\mathbf{A}_{\text{par}}^{-1})^{\mathrm{T}}\}$$

in distribution when $N\to\infty$, where

$$\mathbf{A}_{\text{par}} = E\left[\frac{d\mathbf{S}_{\text{eff}}^*\{\mathbf{X}_i,R_i,R_iY_i,\boldsymbol{\beta},\mathbf{b}^*(\cdot,\boldsymbol{\beta},\boldsymbol{\alpha})\}}{d\boldsymbol{\beta}^{\mathrm{T}}}\right].$$

Under the special assumption (2),

$$\mathbf{B}_{\text{par}} = \text{var}\Bigg(\mathbf{S}_{\text{eff}}^*\{\mathbf{X}_i,R_i,R_iY_i,\boldsymbol{\beta},\mathbf{b}^*(\cdot,\boldsymbol{\beta},\boldsymbol{\alpha})\}$$
$$+E\left[\mathbf{S}_{\text{eff}}^*\{\mathbf{X}_i,R_i,R_iY_i,\boldsymbol{\beta},\mathbf{b}^*(\cdot,\boldsymbol{\beta},\boldsymbol{\alpha})\}\frac{\partial\log f_{\mathbf{X}}(\mathbf{X}_i;\boldsymbol{\alpha})}{\partial\boldsymbol{\alpha}^{\mathrm{T}}}\right]$$
$$\times\boldsymbol{\phi}(\mathbf{X}_i;\boldsymbol{\alpha})\Bigg).$$

Under the general assumption (1), $\mathbf{B}_{\text{par}}$ has the same form but with $\partial\log f_{\mathbf{X}}(\mathbf{X}_i;\boldsymbol{\alpha})/\partial\boldsymbol{\alpha}^{\mathrm{T}}$ replaced by $\partial\log f_{\mathbf{Z}|\mathbf{U}}(\mathbf{Z}_i,\mathbf{U}_i;\boldsymbol{\alpha})/\partial\boldsymbol{\alpha}^{\mathrm{T}}$.

Remark 6. In practice, a potential obstacle to using the parametric model and the result of Theorem 4 is the possible model misspecification. Theorem 4 shows that, when this parametric model is indeed correct, the variance of the estimator contains an additional term, which resembles $h(\mathbf{x}_i)$ in Theorems 1 and 2. Furthermore, if the working model π^* happens to be correctly specified, this term is zero.

Remark 7. In the fortunate situation when the missingness mechanism model $\pi(y, \mathbf{u})$ in the partial shadow case in Section 4 is correct, then we can misspecify the model $f_{\mathbf{Z}|\mathbf{U}}(\mathbf{z}, \mathbf{u})$ and our estimator based on $\mathbf{S}^*_{\text{eff}}\{\mathbf{u}, \mathbf{z}, r, ry, \boldsymbol{\beta}, \mathbf{b}^*(\cdot, \mathbf{u}, \boldsymbol{\beta})\}$ will remain consistent. In fact, in this case, replacing $\mathbf{b}^*(\cdot, \mathbf{u}, \boldsymbol{\beta})$ with an arbitrary function of $(y, \mathbf{u})$ will lead to a consistent estimator. This is because we can easily verify that

$$
\begin{aligned}
&E\{\mathbf{S}^*_{\text{eff}}(\mathbf{u}, \mathbf{z}, R, RY) \mid \mathbf{x}\} \\
&= E\left[\frac{R\mathbf{f}_{\boldsymbol{\beta}}(Y, \mathbf{u}, \mathbf{z}; \boldsymbol{\beta})}{f_{Y|\mathbf{X}}(Y, \mathbf{u}, \mathbf{z}; \boldsymbol{\beta})} - \frac{(1-R)\int \mathbf{f}_{\boldsymbol{\beta}}(t, \mathbf{u}, \mathbf{z}; \boldsymbol{\beta})\pi(t, \mathbf{u})d\mu(t)}{\int f_{Y|\mathbf{X}}(t, \mathbf{u}, \mathbf{z}; \boldsymbol{\beta})\{1-\pi(t, \mathbf{u})\}d\mu(t)}\right. \\
&\quad - R\mathbf{b}^*(Y, \mathbf{u}) \\
&\quad \left. + \frac{(1-R)\int \mathbf{b}^*(t, \mathbf{u}) f_{Y|\mathbf{X}}(t, \mathbf{u}, \mathbf{z}; \boldsymbol{\beta})\pi(t, \mathbf{u})d\mu(t)}{\int f_{Y|\mathbf{X}}(t, \mathbf{u}, \mathbf{z}; \boldsymbol{\beta})\{1-\pi(t, \mathbf{u})\}d\mu(t)} \mid \mathbf{x}\right] \\
&= \mathbf{0}
\end{aligned}
$$

regardless of the form of $\mathbf{b}^*(y, \mathbf{u})$. Similar double robustness property also holds for the complete shadow case in Section 3 when we simply eliminate the variable $\mathbf{u}$ from $\mathbf{b}^*(y, \mathbf{u})$ and $\pi(y, \mathbf{u})$.

6. Simulation Studies

We conduct comprehensive simulation studies to evaluate the finite sample performance of our proposed estimators $\widehat{\boldsymbol{\beta}}$ in Theorems 1 and 2 and the alternative ones $\widehat{\boldsymbol{\beta}}_{\text{non}}$ and $\widehat{\boldsymbol{\beta}}_{\text{par}}$ presented in Section 5. We compare them with the estimator derived using all simulated data without missingness (`FullData`), and the estimator derived using all completely observed subjects, the so-called complete-case analysis (`CC`). We also compare our proposed estimators with the pseudo likelihood estimator (`Pseudo`) studied in Tang et al. (2003) under the special assumption (2) and later in Zhao and Shao (2015) under the general assumption (1). Note that the pseudo likelihood estimators are members of our proposed estimator family, corresponding to the result of adopting a degenerated working mechanism model $\pi^*(y) = 0$ or $\pi^*(y, \mathbf{u}) = 0$. To evaluate the performance against the theoretical optimal limit, we also implement the oracle estimator $\widehat{\boldsymbol{\beta}}_{\text{ora}}$, obtained when the true $f_{\mathbf{X}}(\mathbf{x})$ in (4), or $f_{\mathbf{Z}|\mathbf{U}}(\mathbf{z}, \mathbf{u})$ in (6), is used. We first present the results under the special assumption (2), then the results under the general assumption (1).

6.1. Scenarios Under Special Assumption (2)

We experiment two situations under the special assumption (2). In each situation, we implement eight different methods, where the working mechanism $\pi^*(y)$ is correct or misspecified, in combination with $\widehat{f}_{\mathbf{X}}(\mathbf{x})$ being obtained by one of the four approaches: its truth, the proposal in Section 3, and the two alternatives in Section 5. In addition, we compare them with the `FullData` method, the `CC` method, and the `Pseudo` method studied in Tang et al. (2003).

For the first situation, we generate X from a univariate normal distribution with mean 0.5 and variance $\sigma^2 = 0.25$. The response Y is generated from the model $Y = \beta_0 + \beta_1 X + \epsilon$, where the parameter of interest $\boldsymbol{\beta} = (\beta_0, \beta_1)^{\mathrm{T}} = (0.25, -0.5)^{\mathrm{T}}$, and ϵ follows the standard normal distribution. The true model of the missingness mechanism is

$$
\text{pr}(R = 1 \mid y) = \pi(y) = \frac{\exp(1+y)}{1+\exp(1+y)}.
$$

This leads to about 1/3 subjects with missing response. The misspecified working mechanism model is

$$
\pi^*(y) = \frac{\exp(1-y)}{1+\exp(1-y)}.
$$

In terms of numerical implementation, we use the Gauss–Hermite quadrature with 15 points to approximate the integrations. We adopt the Epanechnikov kernel function $K(u) = 0.75(1-u^2)I(|u| \le 1)$ in the nonparametric density estimation. We choose the bandwidth $CN^{-1/3}$ with $C = 1.5$ in our simulations. We find that the results are robust in the situations where C ranges from 1 to 2.

We consider the total sample size $N = 500$ and the results summarized in Table 1 are based on 1000 simulation replicates. For each estimator, we compute its sample bias (`bias`), sample standard deviation (`std`), estimated standard deviation using the asymptotic distribution ($\widehat{\texttt{std}}$), and the coverage probability (`cvg`) at the nominal level 95%.

In the second situation, we consider a higher dimensional $\mathbf{X}$ and we generate $\mathbf{X}$ from a three-dimensional multivariate normal distribution with mean zero and covariance matrix $\boldsymbol{\Sigma} = (0.5^{|i-j|})_{1\le i,j\le 3}$, and generate Y from the linear model $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \epsilon$. Here $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2, \beta_3)^{\mathrm{T}} = (0, 0.1, -0.2, -0.3)^{\mathrm{T}}$ and ϵ has the standard normal distribution. We adopt the same missingness mechanism model as in the univariate X case and it also leads to around 1/3 missingness. We use the same misspecified working mechanism model and the kernel function. In implementing the multivariate Gauss–Hermite quadrature (Liu and Pierce 1994) to approximate the integrations, we adopt 6 quadrature points in each dimension which generates $6^3 = 216$ points in total. We set the bandwidth to $2N^{-2/7}$, and find the results robust if the constant 2 varies between 1.5 and 2.5. We consider sample size $N = 1000$ and the results summarized in Table 2 are also based on 1000 simulation replicates.

We reach the following conclusions from summarizing the results in Tables 1 and 2. First, for all the eight estimators that we propose in each scenario, the biases are very close to zero, the sample standard deviation and the estimated standard deviation are rather close to each other, and the sample coverage rates of the estimated 95% confidence intervals are close to the nominal level. Hence, regardless of how $f_{\mathbf{X}}(\mathbf{x})$ is estimated and whether the working mechanism model $\pi^*(y)$ is specified correctly or not, our methodology always produces consistent estimator and the inference results based on the asymptotic results are sufficiently precise. Second, the `CC` method is clearly biased,

Table 1. Under assumption (2), one-dimensional X.

Method	$f_{\mathbf{X}}(\mathbf{x})$	$\pi(y)$	Measure	β_0	β_1
	FullData		bias	−0.0105	−0.0038
			std	0.0838	0.0905
			$\widehat{\text{std}}$	0.0817	0.0996
			cvg	0.9490	0.9500
	CC		bias	0.1307	0.0942
			std	0.2059	0.1534
			$\widehat{\text{std}}$	0.2030	0.1628
			cvg	0.6880	0.7310
	Pseudo		bias	0.0287	−0.0156
			std	0.3427	0.2526
			$\widehat{\text{std}}$	0.3535	0.2359
			cvg	0.9580	0.9370
$\widehat{\boldsymbol{\beta}}_{\text{ora}}$	True	Correct	bias	−0.0186	−0.0040
			std	0.2033	0.1217
			$\widehat{\text{std}}$	0.1943	0.1240
			cvg	0.9530	0.9500
		Incorrect	bias	−0.0198	−0.0049
			std	0.2089	0.1281
			$\widehat{\text{std}}$	0.2079	0.1274
			cvg	0.9520	0.9460
$\widehat{\boldsymbol{\beta}}$	Empirical	Correct	bias	−0.0156	−0.0040
			std	0.2088	0.1271
			$\widehat{\text{std}}$	0.2121	0.1289
			cvg	0.9530	0.9510
		Incorrect	bias	−0.0185	−0.0052
			std	0.2210	0.1362
			$\widehat{\text{std}}$	0.2176	0.1408
			cvg	0.9580	0.9560
$\widehat{\boldsymbol{\beta}}_{\text{non}}$	Nonparametric	Correct	bias	−0.0180	−0.0044
			std	0.2067	0.1257
			$\widehat{\text{std}}$	0.2109	0.1308
			cvg	0.9530	0.9520
		Incorrect	bias	−0.0136	−0.0042
			std	0.2248	0.1433
			$\widehat{\text{std}}$	0.2201	0.1439
			cvg	0.9580	0.9680
$\widehat{\boldsymbol{\beta}}_{\text{par}}$	Parametric	Correct	bias	−0.0177	−0.0045
			std	0.1968	0.1253
			$\widehat{\text{std}}$	0.2015	0.1226
			cvg	0.9510	0.9450
		Incorrect	bias	−0.0140	−0.0057
			std	0.2365	0.1481
			$\widehat{\text{std}}$	0.2355	0.1357
			cvg	0.9560	0.9370

NOTE: Sample bias (bias), sample standard deviation (std), estimated standard deviation ($\widehat{\text{std}}$), and coverage probability (cvg) of 95% confidence interval of FullData (the estimator using all simulated data), CC (the estimator using only completely observed subjects), Pseudo (the estimator proposed in Tang et al. (2003)), as well as the oracle estimator $\widehat{\boldsymbol{\beta}}_{\text{ora}}$, the mainly proposed estimator $\widehat{\boldsymbol{\beta}}$ studied in Theorem 1, the estimator $\widehat{\boldsymbol{\beta}}_{\text{non}}$ studied in Theorem 3, and the estimator $\widehat{\boldsymbol{\beta}}_{\text{par}}$ studied in Theorem 4.

resulting in empirical coverage far from the nominal level. The `Pseudo` method which completely eschews the missingness mechanism model, is much less efficient than any of the eight proposed estimators. The `FullData` method, which is not realistic in applications, is merely used here as a benchmark. Third, among the eight proposed estimators, in each of the scenarios considered, although the estimator with a misspecified mechanism $\pi^*(y)$ is less efficient than its counterpart with the true $\pi(y)$, the inflation of the standard deviation is not large. This indicates a certain robustness of our method to the working mechanism model in terms of estimation efficiency, in addition to the established estimation consistency. This seems to be an added advantage of our estimator because the true form of $\pi(y)$ is difficult to obtain in practice. Our observation here helps to alleviate the burden of extensive efforts to identify a proper missingness mechanism description $\pi(y)$ to reach sufficiently small estimation variability. Last but not least, when the true $\pi(y)$ is used, all estimators have similar numerical performance, especially in the $p = 3$ case. Similar phenomenon is also observed when $\pi(y)$ is misspecified. Therefore, considering the possible model misspecification of $f_{\mathbf{X}}(\mathbf{x})$ in $\widehat{\boldsymbol{\beta}}_{\text{par}}$ and the potential difficulty of nonparametric estimation in implementing $\widehat{\boldsymbol{\beta}}_{\text{non}}$, we highly recommend the use of $\widehat{\boldsymbol{\beta}}$ in practice.

6.2. Scenarios Under General Assumption (1)

Under the general assumption (1), we perform three different simulation studies to examine the finite sample performance of our proposed estimators.

In the first study, we consider a two-dimensional $\mathbf{X}$ where both U and Z are univariate and are continuous variables so the theory established in Theorem 2 applies. We consider treating the conditional expectation related to the unknown quantity $f_{\mathbf{Z}|\mathbf{U}}(\mathbf{x})$ via nonparametric regression, parametric modeling or adopting the truth, in combination with the mechanism model being correct or misspecified. Thus, we implement six different estimators. We also compare them with the `FullData` method, the `CC` method, and the `Pseudo` method studied in Zhao and Shao (2015).

The data generation process is as follows. We first generate $\mathbf{X}$ from a bivariate normal distribution with mean zero and covariance matrix $\boldsymbol{\Sigma} = (0.5^{|i-j|})_{1\le i,j\le 2}$. Then we generate the outcome Y from

$$\text{logit}\{\text{pr}(Y = 1 \mid u, z)\} = \beta_0 + \beta_1 u + \beta_2 z$$

with the parameter of interest $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2)^{\mathrm{T}} = (0, 0.3, -0.3)^{\mathrm{T}}$. The missing data indicator R is generated from

$$\text{pr}(R = 1 \mid y, u) = \pi(y, u) = \frac{\exp(1 + y + u)}{1 + \exp(1 + y + u)},$$

which yields approximately 20% missingness in Y. We adopt the misspecified working mechanism model as

$$\pi^*(y, u) = \frac{\exp(1 - y - u)}{1 + \exp(1 - y - u)}.$$

With the total sample size $N = 1000$, we implement the estimator $\widehat{\boldsymbol{\beta}}$ following Algorithm 2 in Section 4 and the estimators $\widehat{\boldsymbol{\beta}}_{\text{ora}}$ and $\widehat{\boldsymbol{\beta}}_{\text{par}}$ following the discussion in Section 5. We adopt the Epanechnikov kernel in (7). Similar to Section 6.1, we use the Gauss–Hermite quadrature with 15 bases to approximate the integrals. The bandwidth is chosen as $CN^{-1/3}$ with $C = 2$. Results based on 1000 simulation are summarized in Table 3.

In the second study, we consider a three-dimensional $\mathbf{X}$ with a bivariate $\mathbf{U}$ and a univariate Z, to gauge the complexity brought by the increasing dimensionality of the covariate $\mathbf{X}$. Similar to the first study, we implement six different proposed estimators, as well as the `FullData`, `CC` and `Pseudo` estimators. Specifically, we first generate U_1 and U_2 independently from

Table 2. Under assumption (2), three-dimensional $\mathbf{X}$.

Method	$f_{\mathbf{X}}(\mathbf{x})$	$\pi(y)$	Measure	β_0	β_1	β_2	β_3
	FullData		bias	−0.0019	0.0027	0.0082	−0.0088
			std	0.0585	0.0539	0.0559	0.0651
			$\widehat{\text{std}}$	0.0555	0.0564	0.0613	0.0544
			cvg	0.9478	0.9526	0.9489	0.9431
	CC		bias	0.1885	−0.2628	−0.1718	0.0855
			std	0.1034	0.0955	0.0871	0.0923
			$\widehat{\text{std}}$	0.0919	0.0965	0.0941	0.0923
			cvg	0.2590	0.4877	0.3859	0.8166
	Pseudo		bias	−0.0118	0.0159	0.0271	0.0144
			std	0.2014	0.1855	0.2371	0.2519
			$\widehat{\text{std}}$	0.2278	0.1945	0.2388	0.2611
			cvg	0.9397	0.9379	0.9618	0.9573
$\widehat{\boldsymbol{\beta}}_{\text{ora}}$	True	Correct	bias	−0.0053	−0.0018	−0.0038	0.0079
			std	0.0892	0.0776	0.0750	0.0835
			$\widehat{\text{std}}$	0.0905	0.0742	0.0768	0.0863
			cvg	0.9541	0.9613	0.9469	0.9541
		Incorrect	bias	0.0249	0.0006	0.0070	0.0081
			std	0.0982	0.0857	0.0935	0.0924
			$\widehat{\text{std}}$	0.0846	0.1010	0.1035	0.0954
			cvg	0.9558	0.9573	0.9624	0.9639
$\widehat{\boldsymbol{\beta}}$	Empirical	Correct	bias	−0.0028	−0.0024	−0.0035	0.0047
			std	0.0951	0.0872	0.0802	0.0854
			$\widehat{\text{std}}$	0.0886	0.0781	0.0855	0.0877
			cvg	0.9566	0.9586	0.9519	0.9494
		Incorrect	bias	0.0167	0.0076	0.0018	0.0024
			std	0.1085	0.1011	0.1043	0.0946
			$\widehat{\text{std}}$	0.0958	0.1038	0.1010	0.1017
			cvg	0.9604	0.9624	0.9586	0.9543
$\widehat{\boldsymbol{\beta}}_{\text{non}}$	Nonparametric	Correct	bias	−0.0060	−0.0087	−0.0028	0.0081
			std	0.0909	0.0981	0.0814	0.1013
			$\widehat{\text{std}}$	0.0823	0.0975	0.0855	0.1046
			cvg	0.9652	0.9675	0.9530	0.9617
		Incorrect	bias	0.0275	0.0043	0.0019	0.0070
			std	0.1043	0.1075	0.0970	0.1033
			$\widehat{\text{std}}$	0.1085	0.1132	0.1067	0.1052
			cvg	0.9659	0.9692	0.9670	0.9626
$\widehat{\boldsymbol{\beta}}_{\text{par}}$	Parametric	Correct	bias	−0.0018	−0.0031	−0.0016	0.0052
			std	0.0833	0.0761	0.0779	0.0815
			$\widehat{\text{std}}$	0.0796	0.0748	0.0779	0.0802
			cvg	0.9426	0.9464	0.9559	0.9447
		Incorrect	bias	0.0249	0.0044	0.0021	0.0097
			std	0.1049	0.1011	0.0981	0.0921
			$\widehat{\text{std}}$	0.0949	0.1033	0.0909	0.0900
			cvg	0.9650	0.9556	0.9540	0.9492

NOTE: Sample bias (bias), sample standard deviation (std), estimated standard deviation ($\widehat{\text{std}}$), and coverage probability (cvg) of 95% confidence interval of FullData (the estimator using all simulated data), CC (the estimator using only completely observed subjects), Pseudo (the estimator proposed in Tang et al. (2003)), as well as the oracle estimator $\widehat{\boldsymbol{\beta}}_{\text{ora}}$, the mainly proposed estimator $\widehat{\boldsymbol{\beta}}$ studied in Theorem 1, the estimator $\widehat{\boldsymbol{\beta}}_{\text{non}}$ studied in Theorem 3, and the estimator $\widehat{\boldsymbol{\beta}}_{\text{par}}$ studied in Theorem 4.

the uniform$(-1, 1)$ distribution, then generate Z following $Z = 0.3U_1 + 0.3U_2 + \varepsilon$ where ε is a normal error with mean zero and standard deviation 0.1. The response variable Y is generated from

$$Y = \beta_0 + \beta_1 U_1 + \beta_2 U_2 + \beta_3 Z + \epsilon,$$

with $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2, \beta_3)^{\mathrm{T}} = (0, -1, -1, 2)^{\mathrm{T}}$ and ϵ is the noise following the standard normal distribution. The missing data indicator R is generated from

$$\text{pr}(R = 1 \mid y, \mathbf{u}) = \pi(y, \mathbf{u}) = \frac{\exp(2 - y + 0.5u_1 + 0.5u_2)}{1 + \exp(2 - y + 0.5u_1 + 0.5u_2)},$$

which yields approximately 20% missingness in Y. We adopt the misspecified working mechanism model as

$$\pi^*(y, \mathbf{u}) = \frac{\exp(2 + y - 0.5u_1 - 0.5u_2)}{1 + \exp(2 + y - 0.5u_1 - 0.5u_2)}.$$

Based on the sample size $N = 2000$ and 1000 simulation replicates, the results for this setting are summarized in Table 4. The ingredients during the estimator implementation are similar to the first study, except that we adopt the fourth order Epanechnikov kernel function $K(t) = \frac{45}{32}(1-\frac{7}{3}t^2)(1-t^2)I_{\{|t|\le 1\}}$, and the bandwidth is chosen as $CN^{-1/5}$ with $C = 3$.

The third simulation study is designed to mimic the real data example presented in Section 7. We first generate binary

Table 3. Under assumption (1), two-dimensional $\mathbf{X}$.

Method	$f_{\mathbf{Z}\|\mathbf{U}}(\mathbf{z},\mathbf{u})$	$\pi(y,\mathbf{u})$	Measure	β_0	β_1	β_2
	FullData		bias	−0.0149	0.0158	0.0083
			std	0.0526	0.0859	0.0692
			$\widehat{\text{std}}$	0.0508	0.0953	0.0727
			cvg	0.9480	0.9550	0.9510
	CC		bias	1.2571	−0.3069	0.0305
			std	0.3516	0.2877	0.2681
			$\widehat{\text{std}}$	0.3367	0.2910	0.2583
			cvg	0.0000	0.2674	0.9316
	Pseudo		bias	0.0349	−0.0255	0.0203
			std	0.2057	0.2751	0.2893
			$\widehat{\text{std}}$	0.1949	0.2684	0.2700
			cvg	0.9290	0.9530	0.9610
$\widehat{\boldsymbol{\beta}}_{\text{ora}}$	True	Correct	bias	−0.0179	0.0167	0.0045
			std	0.0813	0.1214	0.1175
			$\widehat{\text{std}}$	0.0751	0.1288	0.1209
			cvg	0.9520	0.9460	0.9580
		Incorrect	bias	−0.0134	0.0089	0.0036
			std	0.0901	0.1287	0.1195
			$\widehat{\text{std}}$	0.0974	0.1335	0.1107
			cvg	0.9490	0.9510	0.9530
$\widehat{\boldsymbol{\beta}}$	Empirical	Correct	bias	−0.0158	0.0123	0.0043
			std	0.0825	0.1231	0.1147
			$\widehat{\text{std}}$	0.0800	0.1189	0.1161
			cvg	0.9500	0.9450	0.9530
		Incorrect	bias	−0.0167	0.0155	0.0055
			std	0.0916	0.1262	0.1200
			$\widehat{\text{std}}$	0.0979	0.1283	0.1240
			cvg	0.9480	0.9510	0.9530
$\widehat{\boldsymbol{\beta}}_{\text{par}}$	Parametric	Correct	bias	−0.0140	0.0150	0.0064
			std	0.0910	0.1281	0.1164
			$\widehat{\text{std}}$	0.0875	0.1310	0.1294
			cvg	0.9600	0.9520	0.9560
		Incorrect	bias	−0.0156	0.0077	0.0038
			std	0.1096	0.1307	0.1340
			$\widehat{\text{std}}$	0.0984	0.1284	0.1259
			cvg	0.9610	0.9540	0.9570

NOTE: Sample bias (bias), sample standard deviation (std), estimated standard deviation ($\widehat{\text{std}}$), and coverage probability (cvg) of 95% confidence interval of FullData (the estimator using all simulated data), CC (the estimator using only completely observed subjects), Pseudo (the estimator proposed in Zhao and Shao (2015)), as well as the oracle estimator $\widehat{\boldsymbol{\beta}}_{\text{ora}}$, the mainly proposed estimator $\widehat{\boldsymbol{\beta}}$ studied in Theorem 2, and the estimator $\widehat{\boldsymbol{\beta}}_{\text{par}}$ studied in Theorem 4.

covariate U from a Bernoulli distribution with $\text{pr}(U = 1) = 0.5$. Then we generate Z following

$$\text{logit}\{\text{pr}(Z = 1 \mid u)\} = -1.5 + 0.2u.$$

The outcome variable Y is generated from

$$\text{logit}\{\text{pr}(Y = 1 \mid u, z)\} = \beta_0 + \beta_1 u + \beta_2 z$$

with $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2)^{\text{T}} = (-0.5, 0.2, 0.7)^{\text{T}}$. We then generate the missing data indicator R following

$$\text{pr}(R = 1 \mid y, u) = \pi(y, u) = \frac{\exp(1 - 2y + 0.3u)}{1 + \exp(1 - 2y + 0.3u)}.$$

We use the working model

$$\pi^*(y, u) = \frac{\exp(1 + 2y + 0.3u)}{1 + \exp(1 + 2y + 0.3u)}$$

as the misspecified mechanism model. We also implement the six different estimators, respectively, $\widehat{\boldsymbol{\beta}}$, $\widehat{\boldsymbol{\beta}}_{\text{ora}}$, and $\widehat{\boldsymbol{\beta}}_{\text{par}}$ in combination with a correct or misspecified mechanism model, as well as the `FullData`, `CC`, and `Pseudo` estimators. Results based on sample size $N = 2000$ and 1000 simulation replications are provided in Table 5.

The conclusions from summarizing Tables 3–5 are also very clear. First, similar to Section 6.1, among the six implemented estimators that we propose, regardless of how $f_{\mathbf{X}}(\mathbf{u}, \mathbf{z})$ is estimated and whether $\pi(y, \mathbf{u})$ is specified correctly or not, our methods always produce consistent estimators and the inference results based on the asymptotic results are sufficiently precise. Second, the `CC` estimator is severely biased, and the `Pseudo` estimator is clearly less efficient than any of our proposed estimators even with a highly misspecified mechanism model. Both estimators should be avoided based on their performance. The `FullData` estimator is of course not available in application and only serves as a benchmark. Third, among the six proposed estimators, in each of the scenarios considered, the estimator with an incorrect $\pi(y, \mathbf{u})$ is less efficient than its counterpart with the correct $\pi(y, \mathbf{u})$ model, while the efficiency loss is not large. Finally, when the true $\pi(y, \mathbf{u})$ model is used, the estimators $\widehat{\boldsymbol{\beta}}$ and $\widehat{\boldsymbol{\beta}}_{\text{par}}$ perform similarly and they are both slightly less efficient than $\widehat{\boldsymbol{\beta}}_{\text{ora}}$. The same phenomenon is observed when the misspecified $\pi(u, \mathbf{u})$ model is used. All of these phenomena reflect our theory investigated in Sections 4 and 5 very closely.

7. Real Data Analysis

Ibrahim, Lipsitz, and Horton (2001) analyzed a dataset of mental health of children in Connecticut (Zahner et al. 1992, 1993; Zahner and Daskalakis 1997), where the binary outcome of interest is the teacher's report of the psychopathology status of the student (a score of 1 indicates borderline or clinical psychopathology, and a score of 0 indicates normal). The three covariates of interest are `father`, the parental status of the household (0 indicates father figure present, and 1 no father figure present), `health`, the physical health of the child (0 means no health problems, and 1 means fair or poor health, a chronic condition or a limitation in activity), and `parent's report`, the psychopathology status of the child reported from the parent (a score of 1 indicates borderline or clinical psychopathology, and a score of 0 indicates normal). In this study, a child's possibly unobserved psychopathology status may be related to missingness because a teacher is more likely to fill out the psychopathology status when the teacher feels that the child is not normal. Hence it is highly suspected that the missingness mechanism is nonignorable. There are 2486 subjects in this dataset and 1061 of them have missing outcome values. The dataset is available in Ibrahim, Lipsitz, and Horton (2001).

As discussed in Miao et al. (2019), the missing indicator of the teacher's report may be related to her assessment of the student but is unlikely to be related to a separate parent's report after conditioning on the teacher's assessment and all other fully observed covariates; moreover, the parent's report is likely highly correlated with that of the teacher. In this case,

Table 4. Under assumption (1), three-dimensional $\mathbf{X}$.

Method	$f_{\mathbf{Z}\|\mathbf{U}}(\mathbf{z},\mathbf{u})$	$\pi(y,\mathbf{u})$	Measure	β_0	β_1	β_2	β_3
	FullData		bias	−0.0054	0.0008	0.0007	0.0015
			std	0.0556	0.0582	0.0544	0.0353
			$\widehat{\text{std}}$	0.0559	0.0600	0.0585	0.0469
			cvg	0.9480	0.9530	0.9630	0.9540
	CC		bias	−0.1636	0.0537	−0.0853	−0.0420
			std	0.0869	0.0905	0.0972	0.0638
			$\widehat{\text{std}}$	0.0795	0.0887	0.0920	0.0575
			cvg	0.4830	0.8180	0.7420	0.8780
	Pseudo		bias	−0.0044	0.0043	0.0058	0.0097
			std	0.1035	0.0878	0.1059	0.0655
			$\widehat{\text{std}}$	0.1058	0.0979	0.0823	0.0772
			cvg	0.9530	0.9490	0.9560	0.9500
$\widehat{\boldsymbol{\beta}}_{\text{ora}}$	True	Correct	bias	−0.0055	−0.0027	0.0034	0.0029
			std	0.0744	0.0711	0.0826	0.0510
			$\widehat{\text{std}}$	0.0729	0.0794	0.0727	0.0505
			cvg	0.9500	0.9390	0.9480	0.9600
		Incorrect	bias	−0.0008	−0.0019	−0.0030	0.0010
			std	0.0879	0.0868	0.1017	0.0621
			$\widehat{\text{std}}$	0.0856	0.0965	0.1129	0.0617
			cvg	0.9430	0.9520	0.9530	0.9530
$\widehat{\boldsymbol{\beta}}$	Empirical	Correct	bias	−0.0025	0.0053	−0.0027	0.0015
			std	0.0784	0.0738	0.0869	0.0515
			$\widehat{\text{std}}$	0.0802	0.0793	0.0934	0.0627
			cvg	0.9450	0.9550	0.9510	0.9630
		Incorrect	bias	0.0016	−0.0030	−0.0050	0.0011
			std	0.0809	0.0979	0.1042	0.0553
			$\widehat{\text{std}}$	0.0976	0.1096	0.1033	0.0562
			cvg	0.9440	0.9470	0.9460	0.9520
$\widehat{\boldsymbol{\beta}}_{\text{par}}$	Parametric	Correct	bias	0.0040	−0.0056	−0.0005	−0.0016
			std	0.0779	0.0724	0.0869	0.0411
			$\widehat{\text{std}}$	0.0798	0.0759	0.0853	0.0438
			cvg	0.9530	0.9570	0.9510	0.9460
		Incorrect	bias	−0.0030	−0.0001	0.0016	−0.0030
			std	0.0822	0.1063	0.0953	0.0540
			$\widehat{\text{std}}$	0.0812	0.1062	0.1014	0.0660
			cvg	0.9450	0.9440	0.9350	0.9450

NOTE: Sample bias (bias), sample standard deviation (std), estimated standard deviation ($\widehat{\text{std}}$), and coverage probability (cvg) of 95% confidence interval of FullData (the estimator using all simulated data), CC (the estimator using only completely observed subjects), Pseudo (the estimator proposed in Zhao and Shao (2015)), as well as the oracle estimator $\widehat{\boldsymbol{\beta}}_{\text{ora}}$, the mainly proposed estimator $\widehat{\boldsymbol{\beta}}$ studied in Theorem 2, and the estimator $\widehat{\boldsymbol{\beta}}_{\text{par}}$ studied in Theorem 4.

`parent's report` constitutes a valid shadow variable in our context.

We first follow Ibrahim, Lipsitz, and Horton (2001) to implement a parametric EM algorithm (the method `parEM`) where the mechanism is a logistic regression model. Then we implement the proposed estimator $\widehat{\boldsymbol{\beta}}$, and the estimator $\widehat{\boldsymbol{\beta}}_{\text{par}}$ where $f_{\mathbf{Z}|\mathbf{U}}(\cdot)$ is modeled as

$$\begin{aligned}&\text{logit}\{\text{pr}(\texttt{parent's report=1} \mid \texttt{health}, \texttt{father})\}\\ &\quad = -2.106 + 0.890\,\texttt{health} + 0.623\,\texttt{father}.\end{aligned}$$

The posited missingness mechanism model $\pi^*(y, \texttt{health}, \texttt{father})$ used in both $\widehat{\boldsymbol{\beta}}$ and $\widehat{\boldsymbol{\beta}}_{\text{par}}$ is

$$\begin{aligned}&\text{logit}\{\text{pr}(R=1 \mid y, \texttt{health}, \texttt{father})\}\\ &\quad = 1.058 - 2.037\,y + 0.298\,\texttt{health} - 0.002\,\texttt{father},\end{aligned}$$

the same as found in the method `parEM`. For comparison, we also implement the naive method using only completely observed subjects (the method `CC`), and the pseudo likelihood estimator of Zhao and Shao (2015) (the method `Pseudo`). For each parameter, we report the estimate, its standard error, and the corresponding z-statistic and p-value from the five methods in Table 6.

Interestingly all methods produce roughly the same coefficient estimate for the shadow variable `parent's report`, while the estimator $\widehat{\boldsymbol{\beta}}$ has the smallest standard error hence is the most efficient. The primary differences among the five methods occur in the coefficients of `intercept`, `health`, and `father`. The method `CC` which only uses completely observed subjects and the method `parEM` which is confined to a purely parametric model specification are both highly suspected to result in estimation biases. The estimator $\widehat{\boldsymbol{\beta}}_{\text{par}}$ where the parametric $f_{\mathbf{Z}|\mathbf{U}}(\cdot)$ model could be misspecified, and the estimator `Pseudo` where the nonignorable missingness mechanism model is bypassed or is set to zero, provide very similar estimates as the estimator $\widehat{\boldsymbol{\beta}}$. However, both $\widehat{\boldsymbol{\beta}}_{\text{par}}$ and `Pseudo` have relatively larger standard errors. In contrast, the estimator $\widehat{\boldsymbol{\beta}}$ takes into account the effect of the missingness mechanism model and is not prone to any possible $f_{\mathbf{Z}|\mathbf{U}}(\cdot)$ model misspecification, hence is much more efficient than the estimators $\widehat{\boldsymbol{\beta}}_{\text{par}}$ and `Pseudo` in this application.

Table 5. Under assumption (1), two-dimensional $\mathbf{X}$ (discrete U).

Method	$f_{\mathbf{Z}\|\mathbf{U}}(\mathbf{z},\mathbf{u})$	$\pi(y,\mathbf{u})$	Measure	β_0	β_1	β_2
	FullData		bias	0.0254	−0.0015	0.0088
			std	0.0733	0.0795	0.0494
			$\widehat{\text{std}}$	0.0705	0.0808	0.0554
			cvg	0.9500	0.9520	0.9430
	CC		bias	1.0641	−0.6033	0.0270
			std	0.3046	0.3240	0.3146
			$\widehat{\text{std}}$	0.2898	0.3139	0.2963
			cvg	0.0000	0.5200	0.9320
	Pseudo		bias	0.0397	−0.0224	0.0170
			std	0.3157	0.2837	0.2417
			$\widehat{\text{std}}$	0.3034	0.2729	0.2293
			cvg	0.9480	0.9460	0.9560
$\widehat{\boldsymbol{\beta}}_{\text{ora}}$	True	Correct	bias	0.0302	−0.0020	0.0108
			std	0.1188	0.1250	0.0975
			$\widehat{\text{std}}$	0.1258	0.1194	0.1024
			cvg	0.9460	0.9580	0.9460
		Incorrect	bias	0.0459	0.0134	0.0121
			std	0.1371	0.1387	0.1069
			$\widehat{\text{std}}$	0.1429	0.1495	0.1091
			cvg	0.9570	0.9670	0.9680
$\widehat{\boldsymbol{\beta}}$	Empirical	Correct	bias	0.0215	−0.0022	0.0091
			std	0.1194	0.1266	0.1037
			$\widehat{\text{std}}$	0.1205	0.1231	0.0975
			cvg	0.9510	0.9520	0.9610
		Incorrect	bias	0.0533	0.0185	0.0081
			std	0.1327	0.1455	0.1003
			$\widehat{\text{std}}$	0.1395	0.1438	0.0994
			cvg	0.9490	0.9600	0.9590
$\widehat{\boldsymbol{\beta}}_{\text{par}}$	Parametric	Correct	bias	0.0247	−0.0018	0.0097
			std	0.1143	0.1288	0.1018
			$\widehat{\text{std}}$	0.1209	0.1265	0.0905
			cvg	0.9610	0.9540	0.9630
		Incorrect	bias	0.0386	0.0097	0.0089
			std	0.1475	0.1487	0.1121
			$\widehat{\text{std}}$	0.1382	0.1547	0.1094
			cvg	0.9620	0.9580	0.9640

NOTE: Sample bias (bias), sample standard deviation (std), estimated standard deviation ($\widehat{\text{std}}$), and coverage probability (cvg) of 95% confidence interval of FullData (the estimator using all simulated data), CC (the estimator using only completely observed subjects), Pseudo (the estimator proposed in Zhao and Shao (2015)), as well as the oracle estimator $\widehat{\boldsymbol{\beta}}_{\text{ora}}$, the mainly proposed estimator $\widehat{\boldsymbol{\beta}}$ studied in Theorem 2, and the estimator $\widehat{\boldsymbol{\beta}}_{\text{par}}$ studied in Theorem 4.

Table 6. Comparison of the real data analysis results in the children's mental health study.

Method	Measure	intercept	health	father	parent's report
CC	estimate	−1.9307	−0.0516	0.3652	1.4621
	standard error	0.1132	0.1480	0.1690	0.1583
	z-statistic	−17.0618	−0.3487	2.1608	9.2380
	p-value	0.0000	0.7273	0.0307	0.0000
parEM	estimate	−1.7938	−0.0641	0.1610	1.4538
	standard error	0.1258	0.2213	0.1380	0.1646
	z-statistic	−14.2591	−0.2897	1.1667	8.8323
	p-value	0.0000	0.7721	0.2433	0.0000
Pseudo	estimate	−1.3750	−0.9814	−0.0699	1.4687
	standard error	0.2867	0.3622	0.4460	0.1366
	z-statistic	−4.7966	−2.7098	−0.1567	10.7523
	p-value	0.0000	0.0067	0.8755	0.0000
$\widehat{\boldsymbol{\beta}}$	estimate	−1.3585	−0.9817	−0.0718	1.4623
	standard error	0.1823	0.1470	0.1320	0.1194
	z-statistic	−7.4520	−6.6782	−0.5439	12.2471
	p-value	0.0000	0.0000	0.5865	0.0000
$\widehat{\boldsymbol{\beta}}_{\text{par}}$	estimate	−1.3624	−0.9703	−0.0728	1.4651
	standard error	0.2654	0.3221	0.3372	0.1378
	z-statistic	−5.1334	−3.0124	−0.2159	10.6321
	p-value	0.0000	0.0026	0.8291	0.0000

NOTE: CC is the method using only completely observed subjects. parEM is the method using the EM algorithm with a purely parametric model specification. Pseudo is the method proposed in Zhao and Shao (2015). $\widehat{\boldsymbol{\beta}}$ is the mainly proposed estimator studied in Theorem 2. $\widehat{\boldsymbol{\beta}}_{\text{par}}$ is the estimator studied in Theorem 4 but with possible $f_{\mathbf{Z}|\mathbf{U}}(\cdot)$ model misspecification.

8. Discussion

In this article, to bypass the difficulty of correctly specifying and directly estimating the nonignorable missingness mechanism, we propose a class of estimators which only need a working mechanism model. Our procedure guarantees a consistent estimator for the parameter of interest regardless of the working model being correct or not.

In practice, a working model for the missingness mechanism closer to the truth is likely beneficial. To obtain such a model, one can first adopt a rich yet pure parametric mechanism model and use maximum likelihood estimator via the EM algorithm to determine the parameters in it. This allows us to identify a plausible nonignorable missingness mechanism model. This mechanism model can then be used as the working model $\pi^*(y,\mathbf{u})$ in our procedure. Further, when the working model is sufficiently rich or even nonparametric, the resulting estimator from our procedure is likely efficient. While intuitively sensible, the issue of how to achieve and rigorously prove the optimal efficiency requires more in-depth investigation.

To achieve identifiability, a major assumption in our estimation procedure is the existence of the shadow variable $\mathbf{Z}$. From the example we show in this article and some other similar situations, the existence of such a variable is clinically reasonable and practically useful. How to statistically validate a shadow variable is also of interest and it warrants further research.

We also would like to point out that a correct specification or estimation of the conditional pdf/pmf of $f_{\mathbf{Z}|\mathbf{U}}(\mathbf{x})$, which only involves completely observed data, is needed in our implementation. When the dimension of $\mathbf{X}$ is relatively small, one can use kernel estimation or other types of nonparametric techniques. When the dimension becomes larger, because of the curse of dimensionality issue, one might need to concentrate on parametric or semiparametric specification. In principle, since there is no missing data involved in this step, any available statistical methods can be explored and investigated. Finally, despite of a few investigation in Fang and Shao (2016) and Zhao, Yang, and Ning (2018), the issue of high dimensionality in the nonignorable missing data context remains a challenging topic and worth further research.

It is worth emphasizing that our framework is based on a correctly specified regression model $f_{Y|\mathbf{X}}(y \mid \mathbf{x};\boldsymbol{\beta})$, hence is suitable for studying the relation between Y and $\mathbf{X}$. In practice, the research interest may be different. For example, one may be interested in learning $E(Y)$ or some other summary of the outcome. In this case, one can choose to construct models differently and proceed with the statistical analysis. For example, Miao et al. (2019) chose to model the odds ratio function $\text{OR}(y,\mathbf{u}) \equiv \{f(R=0 \mid Y,\mathbf{U})f(R=1 \mid Y=0,\mathbf{U})\}/\{f(R=1 \mid Y,\mathbf{U})f(R=0 \mid Y=0,\mathbf{U})\}$ and one of $f(Y,\mathbf{Z} \mid R=1,\mathbf{U})$

and $f(R \mid Y = 0, \mathbf{U})$, hence by pass the direct modeling of $f_{Y|\mathbf{X}}(y \mid \mathbf{x}; \boldsymbol{\beta})$. These two modeling approaches are suitable in their respect context, and are complementary to each other. In applications, one can use the suitable approach depending on the practical need.

Supplementary Materials

The supplementary materials contain all the detailed technical derivations and proofs.

Acknowledgments

The authors would like to thank the editor, an associate editor, and three reviewers for their insightful comments which have helped improve the manuscript substantially.

Funding

This research was partially supported by the National Science Foundation under award numbers 1953526 and 2122074.

References

Atkinson, K. (1976), "An Automatic Program for Linear Fredholm Integral Equations of the Second Kind," *ACM Transactions on Mathematical Software (TOMS)*, 2, 154–171. [5,6]

Bickel, P. J., Klaassen, J., Ritov, Y., and Wellner, J. A. (1993), *Efficient and Adaptive Estimation for Semiparametric Models*, Baltimore, MD: Johns Hopkins University Press. [2]

Chang, T., and Kott, P. S. (2008), "Using Calibration Weighting to Adjust for Nonresponse Under a Plausible Model," *Biometrika*, 95, 555–571. [1]

d'Haultfoeuille, X. (2010), "A New Instrumental Method for Dealing With Endogenous Selection," *Journal of Econometrics*, 154, 1–15. [1,3]

Fang, F., and Shao, J. (2016), "Model Selection With Nonignorable Nonresponse," *Biometrika*, 103, 861–874. [14]

Hu, Y., and Shiu, J.-L. (2018), "Nonparametric Identification Using Instrumental Variables: Sufficient Conditions for Completeness," *Econometric Theory*, 34, 659–693. [3]

Ibrahim, J. G., and Lipsitz, S. R. (1996), "Parameter Estimation From Incomplete Data in Binomial Regression When the Missing Data Mechanism Is Nonignorable," *Biometrics*, 1071–1078. [1]

Ibrahim, J. G., Lipsitz, S. R., and Horton, N. (2001), "Using Auxiliary Data for Parameter Estimation With Non-Ignorably Missing Outcomes," *Journal of the Royal Statistical Society*, Series C, 50, 361–373. [3,12,13]

Kim, J. K., and Shao, J. (2013), *Statistical Methods for Handling Incomplete Data*, Boca Raton, FL: Chapman & Hall/CRC. [1]

Kim, J. K., and Yu, C. L. (2011), "A Semiparametric Estimation of Mean Functionals With Nonignorable Missing Data," *Journal of the American Statistical Association*, 106, 157–165. [2]

Kott, P. (2014), "Calibration Weighting When Model and Calibration Variables Can Differ," in *Contributions to Sampling Statistics*, eds. F. Mecatti, F., Conti, L. P., and Ranalli, G. M., Cambridge: Springer International Publishing, pp. 1–18. [1,3]

Lehmann, E. L., and Romano, J. P. (2006), *Testing Statistical Hypotheses*, New York: Springer. [3]

Little, R. J., and Rubin, D. B. (2002), *Statistical Analysis With Missing Data* (2nd ed.), New York: Wiley. [1]

Liu, Q., and Pierce, D. A. (1994), "A Note on Gauss–Hermite Quadrature," *Biometrika*, 81, 624–629. [9]

Miao, W., Liu, L., Tchetgen Tchetgen, E., and Geng, Z. (2019), "Identification, Doubly Robust Estimation, and Semiparametric Efficiency Theory of Nonignorable Missing Data With a Shadow Variable," arXiv no. 1509.02556. [1,2,3,12,14]

Miao, W., and Tchetgen Tchetgen, E. J. (2016), "On Varieties of Doubly Robust Estimators Under Missingness Not at Random With a Shadow Variable," *Biometrika*, 103, 475–482. [1,3]

Molenberghs, G., Fitzmaurice, G., Kenward, M. G., Tsiatis, A. A., and Verbeke, G. (2014), *Handbook of Missing Data Methodology*, Boca Raton, FL: Chapman & Hall/CRC Press. [1]

Morikawa, K., and Kim, J. K. (2016), "Semiparametric Adaptive Estimation With Nonignorable Nonresponse Data," arXiv no. 1612.09207. [1]

Newey, W. K., and Powell, J. L. (2003), "Instrumental Variable Estimation of Nonparametric Models," *Econometrica*, 71, 1565–1578. [1,3]

Qin, J., Leung, D., and Shao, J. (2002), "Estimation With Survey Data Under Nonignorable Nonresponse or Informative Sampling," *Journal of the American Statistical Association*, 97, 193–200. [1]

Robins, J. M., and Ritov, Y. (1997), "Toward a Curse of Dimensionality Appropriate (CODA) Asymptotic Theory for Semi-Parametric Models," *Statistics in Medicine*, 16, 285–319. [1]

Robins, J. M., Rotnitzky, A., and Zhao, L. P. (1994), "Estimation of Regression Coefficients When Some Regressors Are Not Always Observed," *Journal of the American statistical Association*, 89, 846–866. [1]

Rotnitzky, A., and Robins, J. (1997), "Analysis of Semi-Parametric Regression Models With Non-Ignorable Non-Response," *Statistics in Medicine*, 16, 81–102. [1]

Rubin, D. B. (1987), *Multiple Imputation for Nonresponse in Surveys*, New York: Wiley. [1]

Schafer, J. L. (1997), *Analysis of Incomplete Multivariate Data*, London: Chapman & Hall. [1]

Shao, J., and Wang, L. (2016), "Semiparametric Inverse Propensity Weighting for Nonignorable Missing Data," *Biometrika*, 103, 175–187. [2]

Shao, J., and Zhao, J. (2013), "Estimation in Longitudinal Studies With Nonignorable Dropout," *Statistics and Its Interface*, 6, 303–313. [1,3]

Sun, B., Liu, L., Miao, W., Wirth, K., Robins, J., and Tchetgen Tchetgen, E. J. (2018), "Semiparametric Estimation With Data Missing Not at Random Using an Instrumental Variable," *Statistica Sinica*, 28, 1965–1983. [1]

Tang, G., Little, R. J., and Raghunathan, T. E. (2003), "Analysis of Multivariate Missing Data With Nonignorable Nonresponse," *Biometrika*, 90, 747–764. [1,2,3,5,9,10,11]

Tchetgen Tchetgen, E. J., and Wirth, K. E. (2017), "A General Instrumental Variable Framework for Regression Analysis With Outcome Missing Not at Random," *Biometrics*, 73, 1123–1131. [1]

Tsiatis, A. A. (2006), *Semiparametric Theory and Missing Data*, New York: Springer. [1,2]

Wang, S., Shao, J., and Kim, J. K. (2014), "An Instrumental Variable Approach for Identification and Estimation With Nonignorable Nonresponse," *Statistica Sinica*, 24, 1097–1116. [1,3]

Zahner, G. E., and Daskalakis, C. (1997), "Factors Associated With Mental Health, General Health, and School-Based Service Use for Child Psychopathology," *American Journal of Public Health*, 87, 1440–1448. [12]

Zahner, G. E., Jacobs, J. H., Freeman, D. H., and Trainor, K. F. (1993), "Rural-Urban Child Psychopathology in a Northeastern US State: 1986–1989," *Journal of the American Academy of Child & Adolescent Psychiatry*, 32, 378–387. [12]

Zahner, G. E., Pawelkiewicz, W., DeFrancesco, J. J., and Adnopoz, J. (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, 951–960. [12]

Zhao, J. (2017), "Reducing Bias for Maximum Approximate Conditional Likelihood Estimator With General Missing Data Mechanism," *Journal of Nonparametric Statistics*, 29, 577–593. [2]

——— (2018), "Statistical Methods Without Estimating the Missingness Mechanism: A Discussion of 'Statistical Inference for Nonignorable Missing Data Problems: A Selective Review' by Niansheng Tang and Yuanyuan Ju," *Statistical Theory and Related Fields*, 2, 143–145. [2]

Zhao, J., and Ma, Y. (2018), "Optimal Pseudolikelihood Estimation in the Analysis of Multivariate Missing Data With Nonignorable Nonresponse," *Biometrika*, 105, 479–486. [1,2,3]

Zhao, J., and Shao, J. (2015), "Semiparametric Pseudo-Likelihoods in Generalized Linear Models With Nonignorable Missing Data," *Journal of the American Statistical Association*, 110, 1577–1590. [1,2,3,6,9,10,12,13,14]

——— (2017), "Approximate Conditional Likelihood for Generalized Linear Models With General Missing Data Mechanism," *Journal of Systems Science and Complexity*, 30, 139–153. [2]

Zhao, J., Yang, Y., and Ning, Y. (2018), "Penalized Pairwise Pseudo Likelihood for Variable Selection With Nonignorable Missing Data," *Statistica Sinica*, 28, 2125–2148. [14]