Robust estimation under a semiparametric propensity model for nonignorable missing data

Samidha Shetty; Yanyuan Ma; Jiwei Zhao

Electronic Journal of Statistics
Vol. 19 (2025) 956–981
ISSN: 1935-7524
https://doi.org/10.1214/25-EJS2355

Robust estimation under a semiparametric propensity model for nonignorable missing data

Samidha Shetty[1], **Yanyuan Ma**[2] **and Jiwei Zhao**[3,4]

[1] *Department of Mathematical Sciences, Montana State University,*
e-mail: samidha.shetty@montana.edu

[2] *Department of Statistics, Pennsylvania State University,*
e-mail: yzm63@psu.edu

[3] *Departments of Statistics, University of Wisconsin-Madison*

[4] *Departments of Biostatistics and Medical Informatics, University of Wisconsin-Madison,*
e-mail: jiwei.zhao@wisc.edu

Abstract: We study the problem of estimating a functional or a parameter in the context where outcome is subject to nonignorable missingness. We completely avoid modeling the regression relation between outcome and covariates, while allowing the propensity to be modeled by a semiparametric relation where the dependence on covariates is unknown and unspecified. This unknown function in the semiparametric propensity model is not directly estimable from the observed data and is the fundamental challenge in the estimation of the parameter of interest. By carefully analyzing the semiparametric structure of the model, we discover that the estimation of the parameter in the propensity model as well as the functional estimation can be carried out without estimating this unknown function. This phenomenon is especially interesting and extremely helpful in our context, because it overcomes the very obstacle that this unknown function cannot be directly estimated from the observed data. The property of the proposed estimator is established rigorously through theoretical derivations, and supported by simulations and a data application.

MSC2020 subject classifications: Primary 62D10; secondary 62G05.
Keywords and phrases: Efficient influence function, nonignorable missing data, robust estimation, semiparametric statistics.

Received November 2023.

1. Introduction

1.1. Modeling Strategies for Nonignorable Missing Data

Handling missing data is an inevitable issue in many empirical studies, especially when the data are directly collected from human beings or strongly linked to the subjects' behaviors. Broadly speaking, there are two types of missing data depending on the structure of the missing data propensity, which is the

arXiv: 2108.04966

probability that a variable is missing. The missing data is called missing at random or ignorable [15] if the propensity of missing data depends only on the observed values. In applications, however, the propensity is usually nonignorable in the sense that its distribution not only depends on observed data but also unobserved ones. Nonignorable missing data are prevalent in surveys in social sciences as well as in patient reported outcomes in biomedical studies [4, 12, 5, 10]. Hence, it is imperative that we closely study estimation under this setting and discuss the difficulties encountered.

In this paper, we use Y to denote an univariate outcome variable and $\mathbf{X}$ to denote a p-dimensional covariate. We consider the situation that $\mathbf{X}$ is fully observed but Y is subject to nonignorable missingness. We encode R as the indicator of observing Y, in that $R = 1$ if Y is observed and $R = 0$ otherwise. The missing data setting involves two models of interest, the missing outcome propensity model $\pi(y, \mathbf{x})$ and the conditional probability distribution function of the outcome given the covariates $f_{Y|\mathbf{X}}(y, \mathbf{x})$. The propensity model is defined as

$$\pi(y, \mathbf{x}) \equiv \text{pr}(R = 1 \mid y, \mathbf{x}),$$

and under the nonignorable missing data setting it depends on Y but does not necessarily depend on all variables in $\mathbf{X}$. In addition to hampering the modeling of $\pi(y, \mathbf{x})$, the incompleteness of Y also causes challenges in the modeling of $f_{Y|\mathbf{X}}(y, \mathbf{x})$. Consider an application where the outcome Y is the pain score of a patient suffering from arthritis or the depression score of a breast cancer patient, then the corresponding scientific interest is to estimate some unknown quantities about the marginal distribution of Y such as $E(Y)$. In these situations, it is not advisable to impose assumptions on $f_{Y|\mathbf{X}}(y, \mathbf{x})$, because an incorrect assumption on $f_{Y|\mathbf{X}}(y, \mathbf{x})$ will easily jeopardize the estimation of $E(Y)$ if one computes it through $E(Y) = E\{E(Y \mid \mathbf{X})\}$. On the other hand, [21] showed that a model with nonignorable missing outcome data is not identifiable without any assumptions on either $f_{Y|\mathbf{X}}(y, \mathbf{x})$ or the propensity model $\pi(y, \mathbf{x})$. Therefore, in order to enable model identifiability it is sensible to impose some assumptions on $\pi(y, \mathbf{x})$ and allow $f_{Y|\mathbf{X}}(y, \mathbf{x})$ to be free of assumptions. The key question is: what assumptions are appropriate for a propensity model $\pi(y, \mathbf{x})$ for nonignorable missing outcome data?

1.2. A Semiparametric Propensity Model

A simplistic approach is to impose a fully parametric assumption on the propensity model, such as logistic regression, as in [7, 22, 20, 2, 28, 18]. However, it is difficult to verify any model on $\pi(y, \mathbf{x})$ under nonignorable missingness, and hence is desirable to have a model assumption that is as weak as possible. To this end, [8] considered the following exponential tilting model for the propensity:

$$\text{logit } \pi(y, \mathbf{x}) = h(y, \beta) + g(\mathbf{x}), \tag{1.1}$$

which is semiparametric with a parametric component $h(y, \beta)$, where $h(0, \beta) = 0$ and β is an unknown d-dimensional parameter, and a nonparametric compo-

nent which is an unknown function of the covariates $g(\mathbf{x})$. The parameter β, characterizing the dependence of the propensity on the outcome Y, indicates how severe the nonignorability is. In contrast to a fully parametric model, the nonparametric component $g(\mathbf{x})$ provides certain flexibility to protect against possible model misspecification. However, β and $g(\mathbf{x})$ in model (1.1) are not identifiable, as shown in Example 1 in [23]. Thus, [8] assumed that β is either known or can be estimated using some external data. While model (1.1) is a great improvement over a parametric model for the propensity, the requirement of a known β severely limits its application.

[23] modified model (1.1) and showed that the parameter β can be consistently estimated if one can find some part of $\mathbf{X}$, say, $\mathbf{Z}$, that is not involved in the propensity model; that is, if $\mathbf{X} = (\mathbf{U}^{\mathrm{T}}, \mathbf{Z}^{\mathrm{T}})^{\mathrm{T}}$ and

$$\text{logit } \pi(y, \mathbf{x}) = \text{logit } \pi(y, \mathbf{u}, \beta, g) = h(y, \beta) + g(\mathbf{u}), \tag{1.2}$$

where the dimensions of $\mathbf{U}$ and $\mathbf{Z}$ are q and $p-q$, respectively. The variable $\mathbf{Z}$ is termed non-response instrument or shadow variable in the nonignorable missing data literature. [23] provides extensive discussions on the choice of shadow variable in practice. Note that, while [23] simply wrote $h(y, \beta) = \beta y$, all of their methods are applicable to the more general function $h(y, \beta)$. While the modification by [23] leads to consistent estimation of β, their estimation method involves first estimating $g(\mathbf{u})$. Similarly, [27] proposed a variety of estimation procedures under the same setting.

This introduces some challenges, because although $g(\mathbf{u})$ per se does not have missing data, its estimation is not stand-alone and involves modeling of missing data or estimation of other unknown components. [23] and [27] adopted the profiling approach, and estimated $\widehat{g}(\mathbf{u}, \beta)$ via the standard kernel estimation for every fixed value of β. Since $\widehat{g}(\mathbf{u}, \beta)$ has to be repeatedly estimated in the algorithm, the procedure becomes computationally expensive. Unfortunately, this approach brings tremendous complexity for end-users and it limits the applicability of the semiparametric propensity model to wider practice.

There has been other related work in the field such as [16] and [17] who followed the idea of pattern-mixture factorization [14] and proposed a semiparametric propensity model for nonignorable missing data. Their model decomposition relies on a parametric odds ratio function that quantifies the difference between the distribution models for the observed data and the missing data, and two other nonparametric components: $f_{Y,\mathbf{Z}|\mathbf{U},R=1}(y, \mathbf{z}, \mathbf{u})$ and $\text{pr}(R = 1 \mid y = 0, \mathbf{u})$. Note that, logit $\text{pr}(R = 1 \mid y = 0, \mathbf{u})$ equals the component $g(\mathbf{u})$ in our setting. In Section 4 of [16], the authors proposed a broad class of estimators that require the correct assessment of $g(\mathbf{u})$. This is a critical drawback of their work since $g(\mathbf{u})$ cannot be directly estimated from the observed data.

1.3. Our Contributions

In this paper, we adopt the model in (1.2) and propose a completely novel estimation framework which consistently estimates the unknown quantities of

interest, without estimating or modeling the nonparametric component $g(\mathbf{u})$. In other words, our method is robust to the misspecification of $g(\mathbf{u})$. This robustness of our method eliminates the need of estimating $g(\mathbf{u})$, in turn addressing the limitations of the papers discussed above.

We first consider the estimation of β. Through extensive and careful derivations, we discover that our framework allows the consistent estimation of β without estimating $g(\mathbf{u})$. We then extend our framework to the estimation of $\boldsymbol{\theta}$ that satisfies $E\{\boldsymbol{\zeta}(\mathbf{X}, Y, \boldsymbol{\theta})\} = \mathbf{0}$ for a given function $\boldsymbol{\zeta}(\cdot)$. We show that, once β can be consistently estimated without estimating $g(\mathbf{u})$, $\boldsymbol{\theta}$ can also be consistently estimated without estimating $g(\mathbf{u})$. Note that, $\boldsymbol{\theta}$, which could be $E(Y)$ or median of Y, is the quantity of interest in most applications, but here we treat β as a quantity of interest as well because its estimation is crucial for estimating $\boldsymbol{\theta}$. In other words, β is of interest due to its supporting role.

Our proposed method uses semiparametric techniques [1, 26] to estimate the parameter β. The nonparametric component $g(\mathbf{u})$ is regarded as a nuisance in a semiparametric model, and its effect to estimating β and $\boldsymbol{\theta}$ is projected to an orthogonal direction via the semiparametric treatment. Our estimation procedure only needs a working model of $g(\mathbf{u})$ for the implementation, and this working model does not have to contain the true $g(\mathbf{u})$. Importantly, we find out that such a robustness does not inherit from the standard robustness in the traditional semiparametric literature, where often the orthogonal projection itself is sufficient to guarantee the robustness. Instead, our robustness is a consequence of the specific modeling structure of the propensity and the unique way in which we construct the estimating equations.

Last but not least, we point out that model (1.2) can be naturally extended to

$$\text{logit } \pi(y, \mathbf{x}) = \text{logit } \pi(y, \mathbf{u}, \beta, g) = h(y, \mathbf{u}, \beta) + g(\mathbf{u}), \tag{1.3}$$

where $h(0, \mathbf{u}; \beta) = 0$, and, for any $\beta_1 \neq \beta_2$, the function $h(y, \mathbf{u}; \beta_1) - h(y, \mathbf{u}; \beta_2)$ is not a function of y alone. This generalization permits the interaction between Y and $\mathbf{U}$, such as $h(y, \mathbf{u}, \beta) = (\beta_0 + \beta_1 u_1 + \beta_2 u_2)y$. Our paper mainly presents the results for model (1.2). We show that, by rigorously characterizing the geometric structure of the model (1.2), one can completely avoid modeling or estimating $g(\mathbf{u})$. This type of robustness can be naturally extended to model (1.3), but we will omit a detailed discussion in this paper.

The remainder of the paper is structured as follows. In Section 2, we briefly introduce all of our results on model identifiability. We then present the geometric structure of the model in Section 3, including the detailed description of the nuisance tangent space Λ, its complement, and the efficient score for estimating β. In Section 4, we construct the proposed estimator of β, and provide its theoretical properties. We then extend our results to the estimation and inference for $\boldsymbol{\theta}$ in Section 5. The simulation studies are reported in Section 6, followed by a data application in Section 7. The paper is concluded with a discussion in Section 8, where we summarize more nuanced differences from some related work in this field, such as [23], [27], [16], [17], [18] and [29]. All the technical details are contained in the Supplement.

2. Model Identifiability

Throughout this paper, we denote the marginal density function by $f_{\cdot}(\cdot)$ such as $f_{\mathbf{X}}(\mathbf{x})$ and the conditional density function by $f_{\cdot|\cdot}(\cdot,\cdot)$ such as $f_{Y|\mathbf{X}}(y,\mathbf{x})$ and $f_{Y|\mathbf{X},R=1}(y,\mathbf{x})$. We consider N independent and identically distributed samples from the joint distribution of $(\mathbf{X}, R, RY)$. We adopt the model in (1.2) for the propensity $\pi(y,\mathbf{u},\beta,g)$, and introduce

$$\begin{aligned} w(\mathbf{x}) \equiv \mathrm{pr}(R=1 \mid \mathbf{x}) &= \int \pi(y,\mathbf{u},\beta,g) f_{Y|\mathbf{X}}(y,\mathbf{x}) dy \\ &= [E\{\pi^{-1}(Y,\mathbf{u},\beta,g) \mid \mathbf{x}, 1\}]^{-1}, \end{aligned}$$

where we write $E(\cdot \mid \cdot, R=1)$ as $E(\cdot \mid \cdot, 1)$ for notational simplicity. Note that, $w(\cdot)$ also depends on β and g, but we write it as $w(\mathbf{x})$ for simplicity. We emphasize that in our context $w(\mathbf{x}) \neq \pi(y,\mathbf{u},\beta,g)$ because of the nonignorable missingness, since $\pi(y,\mathbf{u},\beta,g)$ depends on y and $\mathbf{u}$ but not on $\mathbf{z}$, whereas $w(\mathbf{x})$ depends on $\mathbf{u}$ and $\mathbf{z}$ but not on y. However, we can show that $\pi(y,\mathbf{u},\beta,g) f_{Y|\mathbf{X}}(y,\mathbf{x}) = w(\mathbf{x}) f_{Y|\mathbf{X},R=1}(y,\mathbf{x})$.

We write the likelihood function from one single observation, $f_{\mathbf{X},R,RY}(\mathbf{x},r,ry)$, as

$$\begin{aligned} & f_{\mathbf{X}}(\mathbf{x}) \left\{ \frac{f_{Y|\mathbf{X},R=1}(y,\mathbf{x})}{\int f_{Y|\mathbf{X},R=1}(t,\mathbf{x})/\pi(t,\mathbf{u},\beta,g)dt} \right\}^{r} \\ & \times \left\{ 1 - \frac{1}{\int f_{Y|\mathbf{X},R=1}(t,\mathbf{x})/\pi(t,\mathbf{u},\beta,g)dt} \right\}^{1-r}. \end{aligned} \tag{2.1}$$

This is a semiparametric model indexed by one d-dimensional parameter β and three nonparametric functions $f_{\mathbf{X}}(\mathbf{x})$, $f_{Y|\mathbf{X},R=1}(y,\mathbf{x})$ and $g(\mathbf{u})$. Unfortunately, as pointed out in [23] and the references therein, without imposing further assumptions, (2.1) is not fully identifiable in the sense that two different sets of unknown components can result in the same likelihood function.

Next, we study conditions under which the model (2.1) is identifiable. One approach is to impose the completeness condition on $f_{Y|\mathbf{X},R=1}(y,\mathbf{x})$. The completeness condition has been widely used in a variety of research topics to ensure the model identifiability, such as measurement error, instrumental variable and nonignorable missing data; see, e.g., [19, 3, 6, 16, 29, 13]. For nonignorable missing data, [16] showed that one can identify a nonparametric odds ratio function by solving a type-I Fredholm equation, even though the estimation does need a parametric odds ratio function. In addition, it is even possible to identify some functionals of the population, e.g., the outcome mean, without identifying the whole distribution of the population [13]. In our context, we present the model identifiability result in Lemma 1 by assuming the completeness condition on $f_{Y|\mathbf{X},R=1}(y,\mathbf{x})$. The proof of Lemma 1 is in Section 1 of the Supplementary Material [25].

Lemma 1. *Assume that for any function $\kappa(y,\mathbf{u})$ with finite mean, $E\{\kappa(Y,\mathbf{U}) \mid \mathbf{X}, R=1\} = 0$ implies $\kappa(y,\mathbf{u}) = 0$ almost everywhere. Then, all unknown components β, $g(\mathbf{u})$, $f_{Y|\mathbf{X},R=1}(y,\mathbf{x})$ and $f_{\mathbf{X}}(\mathbf{x})$ in (2.1) are identifiable.*

Now we elaborate the result in Lemma 1, and compare it with some existing literature. It has been well documented, e.g., Theorem 4.3.1 in [9], that the completeness condition is satisfied if $f_{Y|\mathbf{X},R=1}(y,\mathbf{x})$ is an exponential family with full rank. Hence, commonly-seen regression models, such as the linear regression for continuous Y and the logistic regression for binary Y, satisfy this condition. In a special situation when both Y and $\mathbf{Z}$ are discrete with finite support $\{y_1,\ldots,y_S\}$ and $\{\mathbf{z}_1,\ldots,\mathbf{z}_L\}$, the condition of Lemma 1 implicitly requires $L \geqslant S$, i.e., the support of the nonresponse instrument is no smaller than that of the outcome. Different from [29], the model $f_{Y|\mathbf{X},R=1}(y,\mathbf{x})$ in our context is free of missing data, hence the completeness condition can be adequately assessed in empirical studies.

The identifiability result presented in [23] is different yet related to what has been discussed above. [23] did not work on the likelihood function (2.1) directly. Instead, they aimed to find enough estimating equations to estimate β. Once the estimating equations can be identified, the parameter β is estimable and hence identifiable. Similarly, to create enough estimating equations, [23] required $L > d$, i.e., the support of the nonresponse instrument $\mathbf{Z}$ is greater than the dimension of the unknown parameter β.

We would like to point out that, the identifiability conditions presented in Lemma 1 are sufficient, but not necessary. It is possible to develop a slightly different set of conditions under which all the components in the model (2.1) are also identifiable. For example, the result of Theorem 3.1 in [18] may be generalized here under the model (2.1). It is certainly worthwhile to explore the nuanced differences among various identifiability conditions per se; however, we choose to focus on one and then move forward to presenting the estimation results below.

3. Space Decomposition and Efficient Score for Estimating β

As discussed earlier, the likelihood function in (2.1) is a semiparametric model. We regard β as the parameter of interest and $f_{\mathbf{X}}(\cdot), f_{Y|\mathbf{X},R=1}(\cdot), g(\cdot)$ as nuisance components. For semiparametric models, the influence functions of regular and asymptotically linear estimators, lying in the Hilbert space $\mathcal{H}$ of all d-dimensional zero-mean measurable functions of the observed data with finite variance, provide a way of constructing robust and efficient estimators for β. According to the theory of semiparametrics [1, 26], influence functions belong to the linear space orthogonal to the nuisance tangent space Λ, denoted as $\Lambda^{\perp}$ throughout. Here, the nuisance tangent space Λ is defined as the mean squared closure of nuisance tangent spaces associated with all parametric submodels, and the inner product of $\mathcal{H}$ is defined as $\langle \mathbf{h}_1, \mathbf{h}_2 \rangle = E(\mathbf{h}_1^{\mathrm{T}}\mathbf{h}_2)$. Next, we derive the following result, regarding the mutually orthogonal decomposition of the Hilbert space $\mathcal{H}$, which is crucial for devising the proposed estimator.

Proposition 1. *The mutually orthogonal decomposition of the space* $\mathcal{H}$ *is*

$$\mathcal{H} \quad = \quad \Lambda_1 \oplus \Lambda_2^{*} \oplus \Lambda_3 \oplus \Lambda^{\perp}, \ \textit{where}$$

$$\begin{aligned}\Lambda_1 &= \left[\mathbf{a}(\mathbf{x}) \in \mathbb{R}^d : E\{\mathbf{a}(\mathbf{X})\} = \mathbf{0}\right] \text{ and}\\ \Lambda_3 &= \left[\mathbf{a}(\mathbf{u})\{r - w(\mathbf{x})\} : \forall \mathbf{a}(\mathbf{u}) \in \mathbb{R}^d\right]\end{aligned}$$

are the nuisance tangent spaces with respect to $f_{\mathbf{X}}(\mathbf{x})$ *and* $g(\mathbf{u})$ *respectively,*

$$\begin{aligned}\Lambda_2^* = &\left\{r\mathbf{b}(y,\mathbf{x}) - \{r - w(\mathbf{x})\}\left(\frac{E\{\mathbf{b}(Y,\mathbf{x}) \mid \mathbf{x}\}}{1 - w(\mathbf{x})} - \frac{E\left[E\{\mathbf{b}(Y,\mathbf{x}) \mid \mathbf{x}\} \mid \mathbf{u}, 1\right]}{E\{1 - w(\mathbf{X}) \mid \mathbf{u}, 1\}}\right) :\right.\\ &\left.\mathbf{b}(y,\mathbf{x}) \in \mathbb{R}^d, E\{\mathbf{b}(Y,\mathbf{x}) \mid \mathbf{x}, 1\} = \mathbf{0}\right\},\end{aligned}$$

is the residual space of projecting Λ_2 *onto* Λ_3*; i.e.,* $\Lambda_2^* = \Pi(\Lambda_2 \mid \Lambda_3^{\perp})$*, where*

$$\Lambda_2 = \left[\frac{\{w(\mathbf{x}) - r\}E\{\mathbf{b}(Y,\mathbf{x}) \mid \mathbf{x}\}}{1 - w(\mathbf{x})} + r\mathbf{b}(y,\mathbf{x}) : \mathbf{b}(y,\mathbf{x}) \in \mathbb{R}^d, E\{\mathbf{b}(Y,\mathbf{x}) \mid \mathbf{x}, 1\} = \mathbf{0}\right]$$

is the nuisance tangent space with respect to $f_{Y|\mathbf{X},R=1}(y,\mathbf{x})$*, and*

$$\Lambda^{\perp} = \left(\mathbf{g}_0(\mathbf{x})\{1 - r\pi^{-1}(Y,\mathbf{u},\beta,g)\} : E[\{1 - w^{-1}(\mathbf{X})\}\mathbf{g}_0(\mathbf{X}) \mid \mathbf{u}, 1] = \mathbf{0}\right). \quad (3.1)$$

Here, the notation $\oplus$ *stands for the direct sum of two spaces that are orthogonal to each other.*

The detailed proof of Proposition 1 is in Section 2 of the Supplementary Material [25]. The space $\Lambda^{\perp}$ in (3.1) contains all the influence functions for estimating β. Because of its complexity, an arbitrarily chosen element in $\Lambda^{\perp}$ does not lead to a feasible estimator of β, unfortunately. Next, we derive the efficient score $\mathbf{S}_{\text{eff}}$ for β, defined as the residual of the score vector $\mathbf{S}_{\boldsymbol{\beta}}$ after projecting it on to the nuisance tangent space Λ. We expect that $\mathbf{S}_{\text{eff}}$ may provide us more guidance on constructing estimators for β.

Proposition 2. *Under model (2.1), the efficient score for estimating* β *is*

$$\begin{aligned}\mathbf{S}_{\text{eff}}(\mathbf{x}, r, ry) &= \mathbf{c}(\mathbf{x})[1 - r\{1 + e^{-g(\mathbf{u}) - h(y,\boldsymbol{\beta})}\}], \text{ where}\\ \mathbf{c}(\mathbf{x}) &= \frac{\mathbf{a}(\mathbf{u})E\{e^{-h(Y,\boldsymbol{\beta})} \mid \mathbf{x}, 1\} - E\{e^{-h(Y,\boldsymbol{\beta})}\mathbf{h}'_{\boldsymbol{\beta}}(Y,\beta) \mid \mathbf{x}, 1\}}{E\{e^{-h(Y,\boldsymbol{\beta})} \mid \mathbf{x}, 1\} + e^{-g(\mathbf{u})}E\{e^{-2h(Y,\boldsymbol{\beta})} \mid \mathbf{x}, 1\}},\\ \mathbf{a}(\mathbf{u}) &= \frac{E\left[E\{e^{-h(Y,\boldsymbol{\beta})}\mathbf{h}'_{\boldsymbol{\beta}}(Y,\beta) \mid \mathbf{X}, 1\}E\{e^{-h(Y,\boldsymbol{\beta})} \mid \mathbf{X}, 1\}/d(\mathbf{X}) \mid \mathbf{u}, 1\right]}{E\left([E\{e^{-h(Y,\boldsymbol{\beta})} \mid \mathbf{X}, 1\}]^2/d(\mathbf{X}) \mid \mathbf{u}, 1\right)},\\ &\quad \text{and}\\ d(\mathbf{x}) &= E\{e^{-h(Y,\boldsymbol{\beta})} \mid \mathbf{x}, 1\} + e^{-g(\mathbf{u})}E\{e^{-2h(Y,\boldsymbol{\beta})} \mid \mathbf{x}, 1\}.\end{aligned}$$

The proof of Proposition 2 is in Section 3 of the Supplementary Material [25]. The efficient score $\mathbf{S}_{\text{eff}}(\mathbf{x}, r, ry)$ is an unique element in $\Lambda^{\perp}$. Solving an estimating equation formed by the sample average of the efficient score leads to the efficient estimator of β. To do so, one would need to evaluate the conditional expectations of various functions of $\mathbf{Z}$ given $\mathbf{U}, R = 1$, i.e., $E(\cdot \mid \mathbf{u}, 1)$, the conditional expectations of various functions of Y given $\mathbf{X}, R = 1$, i.e., $E(\cdot \mid \mathbf{x}, 1)$, and to estimate $g(\mathbf{u})$.

Since $\mathbf{X}$ is fully observed, the conditional expectations of functions of $\mathbf{Z}$ given $\mathbf{U}, R = 1$ is a standard full data estimation problem and we consider this feasible. For example, we could use results based on exploration of the relation between $\mathbf{Z}$ and $\mathbf{U}$, either known or parametrically modeled, or even employing standard nonparametric regression techniques if necessary. Similarly, the conditional expectations of functions of Y given $\mathbf{X}, R = 1$ is also a full data estimation problem and well studied. Both these estimations can be done via the existing parametric, nonparametric or semiparametric modeling and estimation procedures.

In contrast, the estimation of $g(\mathbf{u})$ is not a full data problem and is hence far from conventional. As noted before, in [16], the authors also derived the efficient score under their model factorization. Recognizing the difficulty of directly estimating $g(\mathbf{u})$, in their Section 4.2 on inverse probability weighted estimation, they proposed estimators where a correctly specified parametric model for $\text{pr}(R = 1 \mid y = 0, \mathbf{u}) = \text{expit}\{g(\mathbf{u})\}$ is mandatory. This is a severe drawback and taints the true robustness. In contrast, the methods we propose for estimating β and $\boldsymbol{\theta}$ in Sections 4 and 5 respectively, do not need to estimate $g(\mathbf{u})$ or specify a parametric model on $g(\mathbf{u})$. Instead, our proposal only needs a working model on $g(\mathbf{u})$, which could be arbitrarily misspecified.

4. Proposed Method for β

In investigating the estimation of β, we make an important discovery, that the difficult task of estimating $g(\mathbf{u})$ can be completely avoided. We have the freedom of adopting a working model, say, $g^*(\mathbf{u})$, in the efficient score, and this still leads to a consistent estimator of β. The discovery is a surprise to us as well because the original model (2.1) does not promise such a feature.

In our analysis, we find that this property benefits from the structure of the propensity model, specifically the logistic propensity form and the additive form inside the logistic link function. It bears the resemblance to the case-control study, where the retrospective sampling feature can be ignored without hampering the covariate effect estimation. We point out that this nice feature does not inherit from the standard robustness in semiparametric literature such as [29]. Semiparametric methods can sometimes be robust to misspecification of the nuisance parameters but it is not always promised. When the robustness holds, it often happens due to model convexity [1] and is readily verifiable. However, the robustness discovered here is a result of the specific form of the propensity in combination with the semiparametric treatment and it does not reveal itself without further investigation. For example, if we had replaced the logit link with probit link, the robustness will not hold. This kind of link function dependent robustness property has not been seen before as far as we are aware. In addition, because logit and probit models are the most often used link functions for binary data, and it has been shown that in practice the logit and probit models behave similarly up to a constant scaling, this discovery means that we can indeed use this robustness to our advantage without giving up much in terms of flexibility of modeling.

We write this important discovery as Theorem 1 below, and provide its proof in Section 4 of the Supplementary Material [25]. The key idea of the proof is that, the construction of $\mathbf{a}(\mathbf{u})$ actually ensures a certain adaptivity to $d^*(\mathbf{x})$ hence to $g^*(\mathbf{u})$ so that $\mathbf{S}^*_{\text{eff}}(\mathbf{x}, r, ry)$ has conditional mean zero given $\mathbf{u}$.

Theorem 1. *Define*

$$\begin{aligned}
\mathbf{S}^*_{\text{eff}}(\mathbf{x}, r, ry) &= \mathbf{c}^*(\mathbf{x})[1 - r\{1 + e^{-g^*(\mathbf{u})-h(y,\boldsymbol{\beta})}\}], \text{ where} \qquad (4.1)\\
\mathbf{c}^*(\mathbf{x}) &= \frac{\mathbf{a}^*(\mathbf{u})E\{e^{-h(Y,\boldsymbol{\beta})} \mid \mathbf{x}, 1\} - E\{e^{-h(Y,\boldsymbol{\beta})}\mathbf{h}'_{\boldsymbol{\beta}}(Y,\beta) \mid \mathbf{x}, 1\}}{E\{e^{-h(Y,\boldsymbol{\beta})} \mid \mathbf{x}, 1\} + e^{-g^*(\mathbf{u})}E\{e^{-2h(Y,\boldsymbol{\beta})} \mid \mathbf{x}, 1\}},\\
\mathbf{a}^*(\mathbf{u}) &= \frac{E\left[E\{e^{-h(Y,\boldsymbol{\beta})}\mathbf{h}'_{\boldsymbol{\beta}}(Y,\beta) \mid \mathbf{X}, 1\}E\{e^{-h(Y,\boldsymbol{\beta})} \mid \mathbf{X}, 1\}/d^*(\mathbf{X}) \mid \mathbf{u}, 1\right]}{E\left([E\{e^{-h(Y,\boldsymbol{\beta})} \mid \mathbf{X}, 1\}]^2/d^*(\mathbf{X}) \mid \mathbf{u}, 1\right)},\\
d^*(\mathbf{x}) &= E\{e^{-h(Y,\boldsymbol{\beta})} \mid \mathbf{x}, 1\} + e^{-g^*(\mathbf{u})}E\{e^{-2h(Y,\boldsymbol{\beta})} \mid \mathbf{x}, 1\},
\end{aligned}$$

and $g^*(\mathbf{u})$ is an arbitrary working model of $g(\mathbf{u})$. Then, we have

$$E\{\mathbf{S}^*_{\text{eff}}(\mathbf{X}, R, RY)\} = \mathbf{0}.$$

Using Theorem 1, to obtain a consistent estimator of β, we can actually bypass estimating $g(\mathbf{u})$ completely. Instead, we can adopt an arbitrary working function $g^*(\mathbf{u})$, and construct an estimating equation $\sum_{i=1}^N \widehat{\mathbf{S}}^*_{\text{eff}}(\mathbf{x}_i, r_i, r_i y_i) = \mathbf{0}$, i.e.,

$$\sum_{i=1}^{N} \widehat{\mathbf{c}}^*(\mathbf{x})[1 - r_i\{1 + e^{-g^*(\mathbf{u}_i)-h(y_i,\boldsymbol{\beta})}\}] = \mathbf{0}, \qquad (4.2)$$

to solve for β, where at any $\mathbf{u}, \mathbf{x}$,

$$\begin{aligned}
\widehat{\mathbf{c}}^*(\mathbf{x}) &= \frac{\widehat{\mathbf{a}}^*(\mathbf{u}_i)\widehat{E}\{e^{-h(Y,\boldsymbol{\beta})} \mid \mathbf{x}_i, 1\} - \widehat{E}\{e^{-h(Y,\boldsymbol{\beta})}\mathbf{h}'_{\boldsymbol{\beta}}(Y,\beta) \mid \mathbf{x}_i, 1\}}{\widehat{E}\{e^{-h(Y,\boldsymbol{\beta})} \mid \mathbf{x}_i, 1\} + e^{-g^*(\mathbf{u}_i)}\widehat{E}\{e^{-2h(Y,\boldsymbol{\beta})} \mid \mathbf{x}_i, 1\}},\\
\widehat{\mathbf{a}}^*(\mathbf{u}) &= \frac{\widehat{E}\left[\widehat{E}\{e^{-h(Y,\boldsymbol{\beta})}\mathbf{h}'_{\boldsymbol{\beta}}(Y,\beta) \mid \mathbf{X}, 1\}\widehat{E}\{e^{-h(Y,\boldsymbol{\beta})} \mid \mathbf{X}, 1\}/\widehat{d}^*(\mathbf{X}) \mid \mathbf{u}, 1\right]}{\widehat{E}\left([\widehat{E}\{e^{-h(Y,\boldsymbol{\beta})} \mid \mathbf{X}, 1\}]^2/\widehat{d}^*(\mathbf{X}) \mid \mathbf{u}, 1\right)}, \text{ and}\\
\widehat{d}^*(\mathbf{x}) &= \widehat{E}\{e^{-h(Y,\boldsymbol{\beta})} \mid \mathbf{x}, 1\} + e^{-g^*(\mathbf{u})}\widehat{E}\{e^{-2h(Y,\boldsymbol{\beta})} \mid \mathbf{x}, 1\}.
\end{aligned}$$

In the above construction, we use $\widehat{E}(\cdot \mid \mathbf{x}, 1)$ and $\widehat{E}(\cdot \mid \mathbf{u}, 1)$ to denote an estimator of the corresponding conditional expectations. Note that these are all standard operations based on fully observed data, hence in practice one can use any traditional parametric or nonparametric methods or various machine learning methods. In this paper, we study three options which lead to three different estimators:

1. Estimator $\widehat{\beta}^*$, which satisfies (4.2) where $E(\cdot \mid \mathbf{x}, 1)$ and $E(\cdot \mid \mathbf{u}, 1)$ are known. Although $\widehat{\beta}^*$ is infeasible in practice, it is included here for comparison purposes.

2. Estimator $\breve{\beta}^*$, which satisfies (4.2) where $E(\cdot \mid \mathbf{x}, 1)$ and $E(\cdot \mid \mathbf{u}, 1)$ are estimated parametrically, where the parameters are collected as $\boldsymbol{\alpha}$.
3. Estimator $\widetilde{\beta}^*$, which satisfies (4.2) where $E(\cdot \mid \mathbf{x}, 1)$ and $E(\cdot \mid \mathbf{u}, 1)$ are estimated via nonparametric kernel regression. More details of estimator $\widetilde{\beta}^*$ is contained in Section 5 of the Supplementary Material [25].

4.1. Theoretical Properties

For estimators $\widehat{\beta}^*$ and $\breve{\beta}^*$, we have the following results with their proofs in Sections 6 and 7 of the Supplementary Material [25] respectively.

Theorem 2. *Estimator $\widehat{\beta}^*$ satisfies*

$$N^{1/2}(\widehat{\beta}^* - \beta) = N^{-1/2} \sum_{i=1}^{N} \phi_{1\boldsymbol{\beta}}(\mathbf{x}_i, r_i, r_i y_i, \beta, g^*) + o_p(1),$$

where

$$\phi_{1\boldsymbol{\beta}}(\mathbf{x}, r, ry, \beta, g^*) = -\left[E\left\{\frac{\partial \mathbf{S}^*_{\text{eff}}(\mathbf{X}, R, RY, \beta)}{\partial \beta^{\mathrm{T}}}\right\}\right]^{-1} \mathbf{S}^*_{\text{eff}}(\mathbf{x}, r, ry, \beta),$$

and $\widehat{\beta}^*$ satisfies $N^{1/2}(\widehat{\beta}^* - \beta) \to N(\mathbf{0}, \mathbf{A}_{1\boldsymbol{\beta}}^{-1}\mathbf{B}_{1\boldsymbol{\beta}}\mathbf{A}_{1\boldsymbol{\beta}}^{-1^{\mathrm{T}}})$ in distribution when $N \to \infty$, where

$$\begin{aligned}
\mathbf{A}_{1\boldsymbol{\beta}} &= E\left\{\frac{\partial \mathbf{S}^*_{\text{eff}}(\mathbf{X}_i, R_i, R_i Y_i, \beta)}{\partial \beta^{\mathrm{T}}}\right\}, \\
\mathbf{B}_{1\boldsymbol{\beta}} &= E\left\{\mathbf{S}^*_{\text{eff}}(\mathbf{X}_i, R_i, R_i Y_i, \beta)^{\otimes 2}\right\}.
\end{aligned}$$

Theorem 3. *Estimator $\breve{\beta}^*$ satisfies*

$$N^{1/2}(\breve{\beta}^* - \beta) = N^{-1/2} \sum_{i=1}^{N} \phi_{2\boldsymbol{\beta}}(\mathbf{x}_i, r_i, r_i y_i, \beta, g^*) + o_p(1),$$

where

$$\begin{aligned}
\phi_{2\boldsymbol{\beta}}(\mathbf{x}, r, ry, \beta, g^*) &= -\left[E\left\{\frac{\partial \mathbf{S}^*_{\text{eff}}(\mathbf{X}, R, RY, \beta)}{\partial \beta^{\mathrm{T}}}\right\}\right]^{-1} \\
&\quad \times \left[\mathbf{S}^*_{\text{eff}}(\mathbf{x}, r, ry, \beta)\right. \\
&\quad \left. + E\left\{\frac{\partial \mathbf{S}^*_{\text{eff}}(\mathbf{X}, R, RY, \beta, \boldsymbol{\alpha})}{\partial \boldsymbol{\alpha}^{\mathrm{T}}}\right\} r\phi_{\boldsymbol{\alpha}}(\mathbf{x}, y, \boldsymbol{\alpha})\right],
\end{aligned}$$

and $\breve{\beta}^*$ satisfies $N^{1/2}(\breve{\beta}^* - \beta) \to N(\mathbf{0}, \mathbf{A}_{2\boldsymbol{\beta}}^{-1}\mathbf{B}_{2\boldsymbol{\beta}}\mathbf{A}_{2\boldsymbol{\beta}}^{-1^{\mathrm{T}}})$ in distribution when $N \to \infty$, where

$$\mathbf{A}_{2\boldsymbol{\beta}} = E\left\{\frac{\partial \mathbf{S}^*_{\text{eff}}(\mathbf{X}_i, R_i, R_i Y_i, \beta)}{\partial \beta^{\mathrm{T}}}\right\},$$

$$\mathbf{B}_{2\boldsymbol{\beta}} = E\left([\mathbf{S}^*_{\text{eff}}(\mathbf{X}_i, R_i, R_iY_i, \beta) + E\left\{\frac{\partial \mathbf{S}^*_{\text{eff}}(\mathbf{X}_i, R_i, R_iY_i, \beta, \boldsymbol{\alpha})}{\partial \boldsymbol{\alpha}^{\mathrm{T}}}\right\} R_i \boldsymbol{\phi}_{\boldsymbol{\alpha}}(\mathbf{X}_i, Y_i, \boldsymbol{\alpha})\right]^{\otimes 2}\right).$$

Here $\boldsymbol{\phi}_{\boldsymbol{\alpha}}(\mathbf{x}, y, \boldsymbol{\alpha})$ is the influence function associated with $\hat{\boldsymbol{\alpha}}$ in the full data parametric estimation of the conditional expectations.

For both $\hat{\beta}^*$ and $\breve{\beta}^*$, we see that they are $\sqrt{N}$-consistent with a loss of efficiency, due to the use of a working model $g^*(\mathbf{u})$ when compared to the semiparametric efficiency bound which is given by $\mathbf{A}_{\boldsymbol{\beta}}^{-1}\mathbf{B}_{\boldsymbol{\beta}}\mathbf{A}_{\boldsymbol{\beta}}^{-1^{\mathrm{T}}}$ where

$$\mathbf{A}_{\boldsymbol{\beta}} = E\left\{\frac{\partial \mathbf{S}_{\text{eff}}(\mathbf{X}_i, R_i, R_iY_i, \beta)}{\partial \beta^{\mathrm{T}}}\right\},$$
$$\mathbf{B}_{\boldsymbol{\beta}} = E\left[\{\mathbf{S}_{\text{eff}}(\mathbf{X}_i, R_i, R_iY_i, \beta)\}^{\otimes 2}\right].$$

To present the theoretical property of $\tilde{\beta}^*$, we need the following regularity conditions.

- C1 The univariate kernel function $K(u)$ is symmetric, has order m, has smooth mth derivative and is supported on $[-1, 1]$. The multivariate kernel function $K(\mathbf{u})$ is the product of the univariate kernel function of each component, i.e. for a q-dimensional vector $\mathbf{u}$, the multivariate kernel function can be written as $K(\mathbf{u}) = \prod_{j=1}^{q} K(u_j)$. Here, with a slight abuse of notation, we use the notation K for both univariate and multivariate kernel functions. Similar assumptions hold for a p-dimensional vector $\mathbf{X}$ as well.
- C2 The functions $f_{Y|\mathbf{X},R=1}(y, \mathbf{x})$, $f_{\mathbf{X}}(\mathbf{x})$ and $\pi(y, \mathbf{u}, \beta, g)$ have mth smooth derivatives with respect to $\mathbf{x}$.
- C3 The bandwidth h satisfies $N^{1/2}h^m \to 0$ and $N^{1/2}h^p \to \infty$ as $N \to \infty$.

Theorem 4. *Under the regularity conditions C1, C2 and C3, the estimator $\tilde{\beta}^*$ satisfies*

$$N^{1/2}(\tilde{\beta}^* - \beta) = N^{-1/2} \sum_{i=1}^{N} \boldsymbol{\phi}_{3\boldsymbol{\beta}}(\mathbf{x}_i, r_i, r_iy_i, \beta, g^*) + o_p(1),$$

where

$$\boldsymbol{\phi}_{3\boldsymbol{\beta}}(\mathbf{x}, r, ry, \beta, g^*) = -\left[E\left\{\frac{\partial \mathbf{S}^*_{\text{eff}}(\mathbf{X}, R, RY, \beta)}{\partial \beta^{\mathrm{T}}}\right\}\right]^{-1} \{\mathbf{S}^*_{\text{eff}}(\mathbf{x}, r, ry, \beta) + r\mathbf{k}(\mathbf{x}, y, \beta, g^*)\},$$

and $\breve{\beta}^*$ satisfies $N^{1/2}(\tilde{\beta}^* - \beta) \to N(\mathbf{0}, \mathbf{A}_{3\boldsymbol{\beta}}^{-1}\mathbf{B}_{3\boldsymbol{\beta}}\mathbf{A}_{3\boldsymbol{\beta}}^{-1^{\mathrm{T}}})$ in distribution when $N \to \infty$, where

$$\mathbf{A}_{3\boldsymbol{\beta}} = E\left\{\frac{\partial \mathbf{S}^*_{\text{eff}}(\mathbf{X}_i, R_i, R_iY_i, \beta)}{\partial \beta^{\mathrm{T}}}\right\},$$

$$\mathbf{B}_{3\boldsymbol{\beta}} = E\left[\{\mathbf{S}^*_{\text{eff}}(\mathbf{X}_i, R_i, R_iY_i, \beta) + R_i\mathbf{k}(\mathbf{X}_i, Y_i, \beta, g^*)\}^{\otimes 2}\right].$$

Here

$$\begin{aligned} & k(\mathbf{x}_i, y_i, \beta, g^*) \\ = & \frac{\{e^{-g(\mathbf{u}_i)} - e^{-g^*(\mathbf{u}_i)}\}[\mathbf{a}^*(\mathbf{u}_i)E\{e^{-h(Y_i)} \mid \mathbf{x}_i, 1\} - E\{e^{-h(Y_i)}\mathbf{h}'_{\boldsymbol{\beta}}(Y_i, \beta) \mid \mathbf{x}_i, 1\}]}{d^*(\mathbf{x}_i)} \\ & \times \Bigg[2E\{e^{-h(Y_i)} \mid \mathbf{x}_i, 1\} - e^{-h(y_i)} \\ & - \frac{E\{e^{-h(Y_i)} \mid \mathbf{x}_i, 1\}\{e^{-h(y_i)} + e^{-g^*(\mathbf{u}_i)}e^{-2h(y_i)}\}}{d^*(\mathbf{x}_i)}\Bigg]. \end{aligned}$$

The proof of Theorem 4 is given in Section 8 of the Supplementary Material [25]. We see that even when the conditional expectations are nonparametrically estimated and a working model for $g^*(\mathbf{u})$ is used, the resulting estimator is $\sqrt{N}$-consistent, albeit, with a loss of efficiency.

Under the special situation that $g^*(\mathbf{u}) = g(\mathbf{u})$, $\mathbf{S}^*_{\text{eff}}(\mathbf{x}_i, r_i, r_iy_i, \beta)$ becomes the true efficient score $\mathbf{S}_{\text{eff}}(\mathbf{x}_i, r_i, r_iy_i, \beta)$. This leads to

$$E\{\partial \mathbf{S}^*_{\text{eff}}(\mathbf{x}_i, r_i, r_iy_i, \beta, \boldsymbol{\alpha})/\partial \boldsymbol{\alpha}^{\mathrm{T}}\} = \mathbf{0}$$

and the function $\mathbf{k}(\mathbf{x}_i, y_i, \beta, g^*)$ also vanishes. Consequently, the corresponding estimators $\widehat{\beta}$, $\breve{\beta}$ and $\widetilde{\beta}$ all achieve the semiparametric efficiency bound and are all efficient estimators.

Corollary 1. *When $g^*(\mathbf{u}) = g(\mathbf{u})$, the corresponding estimators $\widehat{\beta}$, $\breve{\beta}$ and $\widetilde{\beta}$ are identical to the first order. In addition, they are all efficient and reach the semiparametric efficiency bound.*

Our main message in Theorems 2, 3, 4 is that the difficult task of estimating $g(\mathbf{u})$ can be avoided due to a very unusual robustness property discovered in Theorem 1. When we use a working model g^* that is not the same as g, the various parametric or nonparametric estimation procedures of the conditional means bring an efficiency loss but are consistent. However, Corollary 1 further reveals an interesting fact that, this efficiency loss disappears if the working model $g^*(\mathbf{u})$ happens to be identical to g. This phenomenon is a direct consequence of the fact that $\mathbf{S}^*_{\text{eff}}$ is no longer an element in $\Lambda^{\perp}$ unless $g^* = g$. In practice, one can fit very flexible logistic regression model and use the resulting coefficients as a starting point for g^*. We emphasize here that this result is different from that in [29]. The estimator proposed in [29], even if the truth of the nuisance function is known, still cannot achieve the semiparametric efficiency bound. One can check their Remarks 1 and 2 for details.

5. Proposed Method for θ

Once β is estimated without the need of estimating $g(\mathbf{u})$, we move forward to propose estimators for $\boldsymbol{\theta}$ which also do not require the estimation of $g(\mathbf{u})$. Similar

to Theorem 1 from Section 4, this nice feature for $\boldsymbol{\theta}$ is also a benefit from the structure of the propensity score model. Since the estimators proposed in this section are not built upon the efficient influence function, this property for $\boldsymbol{\theta}$ also does not inherit from the standard robustness in semiparametric literature.

To estimate $\boldsymbol{\theta}$, we make use of the relation

$$E\{\zeta(\mathbf{X},Y,\boldsymbol{\theta})\} = E\{\zeta(\mathbf{X},Y,\boldsymbol{\theta}) \mid R=1\}\text{pr}(R=1)+E\{\zeta(\mathbf{X},Y,\boldsymbol{\theta}) \mid R=0\}\text{pr}(R=0).$$

Since it is based on fully observed data, $E\{\zeta(\mathbf{X},Y,\boldsymbol{\theta}) \mid R=1\}\text{pr}(R=1)$ can be estimated by $N^{-1}\sum_{i=1}^{N} r_i\zeta(\mathbf{x}_i,y_i,\boldsymbol{\theta})$. Further, $E\{\zeta(\mathbf{X},Y,\boldsymbol{\theta}) \mid R=0\} = \int E\{\zeta(\mathbf{X},Y,\boldsymbol{\theta}) \mid R=0,\mathbf{x}\}f_{\mathbf{X}|R}(\mathbf{x},0)d\mathbf{x}$ and

$$\begin{aligned} E\{\zeta(\mathbf{x},Y,\boldsymbol{\theta}) \mid R=0,\mathbf{x}\} &= \frac{E\{\zeta(\mathbf{x},Y,\boldsymbol{\theta})(\pi^{-1}-1) \mid \mathbf{x},1\}}{E\{(\pi^{-1}-1) \mid \mathbf{x},1\}} \\ &= \frac{E[\zeta(\mathbf{x},Y,\boldsymbol{\theta})\exp\{-h(Y,\beta)\} \mid \mathbf{x},1]}{E[\exp\{-h(Y,\beta)\} \mid \mathbf{x},1]}. \end{aligned}$$

Hence $E\{\zeta(\mathbf{X},Y,\boldsymbol{\theta}) \mid R=0)\text{pr}(R=0)$ can be estimated by

$$N^{-1}\sum_{i=1}^{N}(1-r_i)\frac{\widehat{E}[\zeta(\mathbf{x}_i,Y,\boldsymbol{\theta})\exp\{-h(Y,\beta^{\text{est}})\} \mid \mathbf{x}_i,1]}{\widehat{E}[\exp\{-h(Y,\beta^{\text{est}})\} \mid \mathbf{x}_i,1]},$$

where β^{est} is a generic notation for an estimator of β. Summarizing the above analysis, we propose to estimate $\boldsymbol{\theta}$ by solving

$$\begin{aligned} \mathbf{0} &= \widehat{E}\{\zeta(\mathbf{X},Y,\boldsymbol{\theta})\} \\ &= N^{-1}\sum_{i=1}^{N}\left(r_i\zeta(\mathbf{x}_i,y_i,\boldsymbol{\theta}) + (1-r_i)\frac{\widehat{E}[\zeta(\mathbf{x}_i,Y,\boldsymbol{\theta})\exp\{-h(Y,\beta^{\text{est}})\} \mid \mathbf{x}_i,1]}{\widehat{E}[\exp\{-h(Y,\beta^{\text{est}})\} \mid \mathbf{x}_i,1]}\right). \end{aligned} \tag{5.1}$$

Importantly, (5.1) does not involve the unknown function $g(\mathbf{u})$, hence it shares the same robust property as the estimation for β. The two expectations in (5.1) are both complete data conditional expectations, hence can be assessed similarly as in Section 4. Also, any estimator of β obtained in Section 4 can be used as β^{est}.

We simply denote the estimator from solving (5.1) as $\widehat{\boldsymbol{\theta}}$. In a special case, if the interest is in estimating $E\{\zeta(\mathbf{X},Y)\}$, then (5.1) can be simplified to

$$\begin{aligned} &\widehat{E}\{\zeta(\mathbf{X},Y)\} \\ &\quad = N^{-1}\sum_{i=1}^{N}\left(r_i\zeta(\mathbf{x}_i,y_i) + (1-r_i)\frac{\widehat{E}[\zeta(\mathbf{x}_i,Y)\exp\{-h(Y,\beta^{\text{est}})\} \mid \mathbf{x}_i,1]}{\widehat{E}[\exp\{-h(Y,\beta^{\text{est}})\} \mid \mathbf{x}_i,1]}\right). \end{aligned} \tag{5.2}$$

5.1. *Theoretical Properties*

The theoretical property of $\widehat{E}\{\zeta(\mathbf{X},Y)\}$ is relatively simple to obtain given the results in Theorems 2, 3, and 4, as we summarize below.

Lemma 2. *When $N \to \infty$, the estimator given in (5.2) satisfies*

$$N^{1/2}[\widehat{E}\{\zeta(\mathbf{X},Y)\} - E\{\zeta(\mathbf{X},Y)\}] = N^{-1/2} \sum_{i=1}^{N} \phi_{\boldsymbol{\zeta}}(\mathbf{x}_i, r_i, r_i y_i, \beta, g^*) + o_p(1),$$

where

$$\begin{aligned}\phi_{\boldsymbol{\zeta}}(\mathbf{x}, r, ry, \beta, g^*) &= r\zeta(\mathbf{x}, y) + (1-r)\frac{E[\zeta(\mathbf{x}, Y)\exp\{-h(Y,\beta)\} \mid \mathbf{x}, 1]}{E[\exp\{-h(Y,\beta)\} \mid \mathbf{x}, 1]} \\ &\quad - E\{\zeta(\mathbf{X}, Y)\} \\ &\quad + E\left(\frac{\partial}{\partial \beta^{\mathrm{T}}} \frac{E[\zeta(\mathbf{X}, Y)\exp\{-h(Y,\beta)\} \mid \mathbf{X}, 1]}{E[\exp\{-h(Y,\beta)\} \mid \mathbf{X}, 1]}\right) \\ &\quad \times \phi_{\boldsymbol{\beta}}(\mathbf{x}, r, ry, \beta, g^*) + \mathbf{k}_{\boldsymbol{\alpha}}(\mathbf{x}, r, ry, \beta).\end{aligned}$$

Here, $\mathbf{k}_{\boldsymbol{\alpha}}(\mathbf{x}, r, ry, \beta) = \mathbf{0}$ when the conditional expectations are known as in Theorem 2,

$$\begin{aligned}\mathbf{k}_{\boldsymbol{\alpha}}(\mathbf{x}, r, ry, \beta) &= E\left(\frac{\partial}{\partial \boldsymbol{\alpha}^{\mathrm{T}}} \frac{E[\zeta(\mathbf{X}, Y)\exp\{-h(Y,\beta)\} \mid \mathbf{X}, 1, \alpha]}{E[\exp\{-h(Y,\beta)\} \mid \mathbf{X}, 1, \alpha]}\right) \\ &\quad \times N^{-1/2} \sum_{i=1}^{N} \phi_{\boldsymbol{\alpha}}(\mathbf{x}, r, ry)\end{aligned}$$

when the conditional expectations are parametrically estimated as in Theorem 3, and

$$\begin{aligned}\mathbf{k}_{\boldsymbol{\alpha}}(\mathbf{x}, r, ry, \beta) &= \frac{r\{1 - E(R \mid \mathbf{x})\}\exp\{-h(y,\beta)\}}{E[\exp\{-h(Y,\beta)\} \mid \mathbf{x}, 1]E(R \mid \mathbf{x})} \\ &\quad \times \left(\zeta(\mathbf{x}, y) - \frac{E[\zeta(\mathbf{X}, Y)\exp\{-h(Y,\beta)\} \mid \mathbf{x}, 1]}{E[\exp\{-h(Y,\beta)\} \mid \mathbf{x}, 1]}\right)\end{aligned}$$

when the conditional expectations are nonparametrically estimated as in Theorem 4. Further, $N^{1/2}[\widehat{E}\{\zeta(\mathbf{X},Y)\} - E\{\zeta(\mathbf{X},Y)\}] \to N(\mathbf{0}, \mathbf{V}_{\boldsymbol{\zeta}})$, where $\mathbf{V}_{\boldsymbol{\zeta}} = E\{\phi_{\boldsymbol{\zeta}}(\mathbf{X}, R, RY, \beta, g^*)^{\otimes 2}\}$.

Consequently, for $\widehat{\boldsymbol{\theta}}$, we have

Theorem 5. *When $N \to \infty$, $\widehat{\boldsymbol{\theta}}$ satisfies*

$$N^{1/2}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}) = -\mathbf{A}_{\boldsymbol{\theta}}^{-1} N^{-1/2} \sum_{i=1}^{N} \phi_{\boldsymbol{\zeta}}(\mathbf{x}_i, r_i, r_i y_i, \beta, g^*) + o_p(1),$$

therefore,

$$N^{1/2}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}) \to N(\mathbf{0}, \mathbf{A}_{\boldsymbol{\theta}}^{-1}\mathbf{V}_{\boldsymbol{\zeta}}\mathbf{A}_{\boldsymbol{\theta}}^{-1^{\mathrm{T}}}),$$

where $\phi_{\boldsymbol{\zeta}}(\mathbf{x}_i, r_i, r_i y_i, \beta, g^*)$, $\mathbf{V}_{\boldsymbol{\zeta}}$ are given in Lemma 2, and

$$\mathbf{A}_{\boldsymbol{\theta}} = \partial E\{\zeta(\mathbf{X}_i, Y_i, \boldsymbol{\theta})\}/\partial \boldsymbol{\theta}^{\mathrm{T}}.$$

The proofs of Lemma 2 and Theorem 5 are given in the Sections 9 and 10 of the Supplementary Material [25] respectively.

6. Simulation Studies

We conduct some simulation studies to examine the finite-sample performance of our proposed estimators as well as their comparison with some existing methods in the literature. We first present the simulation results for the estimators of β and then for the estimators of $\boldsymbol{\theta}$.

6.1. Estimators of β

We first compare the performance of the three proposed estimators of β. We consider a discrete shadow variable Z with two categories such that $\text{pr}(Z = -1) = \text{pr}(Z = 1) = 0.5$. Given $Z = z$, $U \sim N(z, 1)$. Given $Z = z$, $U = u$ and $R = 1$, $Y \sim N\{(u - z)^2, 1\}$. The propensity model is defined as $\pi(y, u, \beta, g) = \text{expit}\{h(y, \beta) + g(u)\}$ with $h(y) = -\beta y$ and $g(u) = -a_1 - a_2 u$ where $(\beta, a_1, a_2) = (-0.2, 0.4, -0.3)$. This is similar to a simulation setting from [23] and we modify it to have the ability to evaluate the case where the conditional expectations are known.

We implement the estimator $\widehat{\beta}^*$ where all the conditional expectations are replaced with their truth, the estimator $\breve{\beta}^*$ where the conditional expectations are evaluated parametrically with the conditional densities estimated using the maximum likelihood method, and the estimator $\widetilde{\beta}^*$ where the conditional expectations are estimated via nonparametric kernel regression. In order to satisfy the conditions C1, C2 and C3 for the 1-dimensional U which leads to $p = 1$ (since the 1-dimensional Z is discrete) and $m = 2$, we use a Gaussian kernel with bandwidth $1.5N^{-1/3}$. We consider two different situations for $g^*(u)$ that it is correctly specified with $g(u) = -0.4 + 0.3u$ and misspecified with $g^*(u) = -0.4u$, for three sample sizes $N = 200, 500, 1000$.

Based on 1000 simulation replicates, Table 1 summarizes the estimation bias, the Monte Carlo standard deviation, the estimated standard error, and the 95% coverage probability. From Table 1, it is clear that all the estimators have very small estimation bias, no matter the working model $g^*(u)$ is correct or not. In all scenarios, the estimated standard errors calculated through the asymptotic theory are fairly close to the Monte Carlo standard deviations. Compared to the situation with the correct model of $g(u)$, although not drastic at all, the misspecified situation has a little bit efficiency loss, which matches with the theoretical properties of these estimators presented in Section 4.1. The coverage probability, in almost all scenarios, is quite close to the nominal level 95%. The only exception is the estimator $\widetilde{\beta}^*$, where it is slightly under-covered but gets better when the sample size increases. The potential reason is the relatively more restrictive requirement on the sample size for nonparametric kernel estimation. Overall the results presented in Table 1 are in accordance to that found in our theoretical derivations. They are convincing in that the proposed estimators of β with an arbitrarily misspecified working model $g^*(u)$ are still consistent and asymptotically normally distributed.

TABLE 1
Estimation bias (Bias), standard deviation (SD), estimated standard error (SE) and 95% coverage probability (CVP) of the proposed estimators of β. All values have been multiplied by 100.

$g^*(u)$	N	Estimator	Bias	SD	SE	CVP
Correct	200	$\widehat{\beta}$	−0.90	18.23	16.49	94.9
$g(u) = -0.4 + 0.3u$		$\breve{\beta}$	−0.91	18.16	16.31	94.9
		$\widetilde{\beta}$	−3.96	16.61	15.25	93.4
	500	$\widehat{\beta}$	−0.25	10.41	9.72	93.5
		$\breve{\beta}$	−0.25	10.39	9.73	93.4
		$\widetilde{\beta}$	−1.20	9.89	9.29	94.3
	1000	$\widehat{\beta}$	−0.23	6.94	6.80	95.7
		$\breve{\beta}$	−0.23	6.93	6.80	95.9
		$\widetilde{\beta}$	−0.69	6.71	6.54	95.8
Incorrect	200	$\widehat{\beta}^*$	−0.88	19.11	16.72	93.3
$g^*(u) = -0.4u$		$\breve{\beta}^*$	−0.79	18.96	16.53	93.4
		$\widetilde{\beta}^*$	−4.40	17.20	14.34	90.4
	500	$\widehat{\beta}^*$	−0.16	11.03	10.27	94.3
		$\breve{\beta}^*$	−0.10	10.94	10.14	94.0
		$\widetilde{\beta}^*$	−1.40	10.36	8.50	90.7
	1000	$\widehat{\beta}^*$	−0.24	7.45	7.22	94.9
		$\breve{\beta}^*$	−0.20	7.43	7.12	94.3
		$\widetilde{\beta}^*$	−0.80	6.90	5.95	91.4

6.2. *Estimators of $\theta = E(Y)$*

To compare the proposed estimators with the method in [23], we focus on the estimation of the outcome mean, i.e., $\theta = E(Y)$. From Section 5, the proposed estimators $\widehat{\theta}^*$, $\breve{\theta}^*$ or $\widetilde{\theta}^*$ estimate $E(Y)$ as

$$N^{-1}\sum_{i=1}^{N}\left(r_i y_i + (1-r_i)\frac{\widehat{E}[Y\exp\{-h(Y,\beta^{\text{est}})\}\mid \mathbf{x}_i, 1]}{\widehat{E}[\exp\{-h(Y,\beta^{\text{est}})\}\mid \mathbf{x}_i, 1]}\right),$$

where $\beta^{\text{est}} = \widehat{\beta}^*$, $\breve{\beta}^*$ or $\widetilde{\beta}^*$, respectively. Besides the method in [23], we also compare with the oracle estimator $\sum_{i=1}^{N} y_i/N$, which assumes no missing data, and the naïve estimator $\sum_{i=1}^{N} r_i y_i/\sum_{i=1}^{N} r_i$, which is the complete-case estimator that uses only the fully observed data.

We consider two simulation settings. In the first setting, the data generation is largely identical to that in Section 6.1 except that we use a quadratic form of the nonparametric component $g(u) = -a_1 - a_2 u^2$ where $(\beta, a_1, a_2) = (-0.1, 0.2, -0.3)$. When we implement the estimator $\widetilde{\beta}^*$, we use the Gaussian kernel with bandwidth $1.5N^{-1/3}$, as in Section 6.1. We consider the misspecified model $g^*(u) = -0.4u$ in our implementations. Based on 1000 simulation replicates, we report the estimation bias, the Monte Carlo standard deviation, the mean squared error, the estimated standard error based on 200 bootstrap samples, and the 95% coverage probability in Table 2.

TABLE 2
Estimation bias (Bias), standard deviation (SD), mean squared error (MSE), estimated standard error (SE) and 95% coverage probability (CVP) of the proposed estimators of $\theta = E(Y)$, with one-dimensional U. All values have been multiplied by 100.

$g^*(u)$	N	Estimator	Bias	SD	MSE	SE	CVP
Correct	200	Oracle	0.57	12.32	12.32	12.22	94.7
$g(u) = -0.2 + 0.3u^2$		Naïve	23.46	17.41	29.21	17.14	74.6
		SW	−5.70	14.89	15.94		
		$\widehat{\theta}$	−0.14	13.63	13.62	15.69	97.3
		$\breve{\theta}$	−0.17	14.03	14.02	15.87	97.3
		$\widetilde{\theta}$	−2.74	15.74	15.97	16.61	95.8
	500	Oracle	0.63	7.56	7.58	7.77	95.5
		Naïve	23.84	11.09	26.29	10.93	41.6
		SW	−6.14	9.44	11.26		
		$\widehat{\theta}$	0.46	8.20	8.21	8.45	95.9
		$\breve{\theta}$	0.48	8.49	8.50	8.70	95.7
		$\widetilde{\theta}$	−0.73	9.64	9.66	9.80	95.5
	1000	Oracle	0.29	5.59	5.60	5.49	94.3
		Naïve	23.60	7.82	24.86	7.72	12.0
		SW	−6.28	6.59	9.10		
		$\widehat{\theta}$	0.15	5.87	5.87	5.88	94.9
		$\breve{\theta}$	0.16	6.05	6.05	6.05	95.1
		$\widetilde{\theta}$	−0.44	6.83	6.84	6.82	94.9
Incorrect	200	Oracle	0.57	12.32	12.32	12.22	94.4
$g^*(u) = -0.4u$		Naïve	23.46	17.41	29.21	17.11	74.0
		SW	−5.70	14.89	15.94		
		$\widehat{\theta}^*$	0.31	14.66	14.67	18.22	97.6
		$\breve{\theta}^*$	0.09	14.47	14.46	17.48	97.8
		$\widetilde{\theta}^*$	−4.35	15.97	16.54	16.62	95.0
	500	Oracle	0.63	7.56	7.58	7.80	95.6
		Naïve	23.84	11.09	26.29	10.95	41.8
		SW	−6.14	9.44	11.26		
		$\widehat{\theta}^*$	0.64	7.97	7.99	8.52	95.9
		$\breve{\theta}^*$	0.64	8.24	8.26	8.59	96.1
		$\widetilde{\theta}^*$	−1.70	9.78	9.92	9.90	95.6
	1000	Oracle	0.29	5.59	5.60	5.49	94.2
		Naïve	23.60	7.82	24.86	7.72	12.2
		SW	−6.28	6.59	9.10		
		$\widehat{\theta}^*$	0.24	5.71	5.71	5.73	94.6
		$\breve{\theta}^*$	0.23	5.88	5.88	5.90	94.8
		$\widetilde{\theta}^*$	−1.10	6.86	6.95	6.88	94.4

The second setting concerns a two-dimensional $\mathbf{U}$. Specifically, we consider variable Z with two categories such that $\text{pr}(Z = 1) = \text{pr}(Z = 2) = 0.5$. We generate $U_1|Z = z \sim N(z, 1)$ and $U_2|Z = z \sim N(z, 1)$. Also, given $Z = z$, $\mathbf{U} = (u_1, u_2)$ and $R = 1$, $Y \sim N\{(u_1 - z)^2 + (u_2 - z)^2, 1\}$. The propensity model is defined as $\pi(y, \mathbf{u}, \beta, g) = \text{expit}\{h(y, \beta) + g(\mathbf{u})\}$ with $h(y) = -\beta y$ and $g(\mathbf{u}) = -a_1 - a_2u_1^2 - a_3u_2^2$ where $(\beta, a_1, a_2, a_3) = (-0.1, 0.8, -0.2, -0.2)$. When we implement the estimator $\widetilde{\theta}^*$, we use a fourth order Tri-weight kernel with bandwidth $1.0N^{-1/6}$ when estimating $\widetilde{\beta}^*$ and we use a second order Gaussian

TABLE 3
Estimation bias (Bias), standard deviation (SD), mean squared error (MSE), estimated standard error (SE) and 95% coverage probability (CVP) of the proposed estimators of $\theta = E(Y)$, with two-dimensional $\mathbf{U}$. All values have been multiplied by 100.

$g^*(\mathbf{u})$	N	Estimator	Bias	SD	MSE	SE	CVP
Correct	200	Oracle	−1.89	15.22	15.32	15.64	95.2
$g(\mathbf{u}) = -0.8 + 0.2u_1^2$		Naïve	25.37	19.87	32.22	20.87	77.9
$+0.2u_2^2$		SW	−11.00	20.71	23.44		
		$\widehat{\theta}$	−1.63	16.21	16.29	16.82	94.7
		$\breve{\theta}$	−2.11	16.70	16.82	17.28	94.4
		$\widetilde{\theta}$	−5.44	17.00	17.84	18.09	95.6
	500	Oracle	−2.09	9.85	10.07	9.96	94.6
		Naïve	24.82	12.63	27.85	13.29	55.8
		SW	−8.60	13.88	16.32		
		$\widehat{\theta}$	−1.72	10.41	10.54	10.61	94.8
		$\breve{\theta}$	−2.20	10.71	10.93	10.92	95.1
		$\widetilde{\theta}$	−4.14	10.74	11.51	11.18	94.7
	1000	Oracle	−1.59	6.91	7.09	7.06	94.6
		Naïve	25.65	9.16	27.23	9.42	22.5
		SW	−6.38	8.72	10.80		
		$\widehat{\theta}$	−1.22	7.33	7.43	7.48	95.2
		$\breve{\theta}$	−1.65	7.55	7.73	7.70	94.7
		$\widetilde{\theta}$	−2.50	7.72	8.11	7.81	94.3
Incorrect	200	Oracle	−1.82	15.27	15.37	15.66	95.5
$g^*(\mathbf{u}) = -0.4u_1 - 0.6u_2$		Naïve	25.44	20.02	32.36	20.88	78.4
		SW	−11.00	20.71	23.44		
		$\widehat{\theta}^*$	−2.33	15.69	15.85	16.43	94.4
		$\breve{\theta}^*$	−2.56	16.27	16.46	16.65	94.0
		$\widetilde{\theta}^*$	−5.48	17.36	18.19	18.24	94.6
	500	Oracle	−2.10	9.84	10.05	9.95	94.4
		Naïve	24.84	12.59	27.85	13.28	55.8
		SW	−8.60	13.88	16.32		
		$\widehat{\theta}^*$	−2.13	10.43	10.64	10.60	93.8
		$\breve{\theta}^*$	−2.46	10.69	10.96	10.79	93.4
		$\widetilde{\theta}^*$	−4.22	10.63	11.43	11.25	94.2
	1000	Oracle	−1.61	6.91	7.09	7.03	94.3
		Naïve	25.63	9.16	27.21	9.40	21.9
		SW	−6.38	8.72	10.80		
		$\widehat{\theta}^*$	−1.35	7.48	7.59	7.43	94.5
		$\breve{\theta}^*$	−1.69	7.65	7.83	7.59	93.9
		$\widetilde{\theta}^*$	−2.44	7.69	8.07	7.83	94.7

kernel with bandwidth $1.0N^{-1/3}$ for the $E(Y)$ estimation. The bandwidth for $\widetilde{\beta}^*$ estimation is modified in order to satisfy the conditions C1, C2 and C3 for the 2-dimensional U which leads to $p = 2$ (since the 1-dimensional Z is discrete) and $m = 4$. We consider the misspecified model $g^*(\mathbf{u}) = -0.4u_1 - 0.6u_2$ in our experiments. The corresponding results are summarized in Table 3.

The results in Tables 2 and 3 deliver the following information. Firstly, the naïve method has very large bias and unreliable coverage probability, hence should not be used in practice. Similar to Section 6.1, the three proposed meth-

ods all perform well in terms of small bias, close Monte Carlo standard deviation and estimated standard error, as well as close to nominal coverage probability. Relatively, the nonparametric estimator $\widetilde{\theta}^*$ has slightly larger bias, but the bias quickly vanishes when the sample size increases. In terms of estimation efficiency, the proposed estimators possess very close, albeit slightly larger, Monte Carlo standard deviation and estimated standard error compared to the oracle estimator, especially when the model $g(\mathbf{u})$ is correct and the sample size is large.

The comparison to the method in [23] is also obvious. With a one-dimensional U, although all the standard deviations are close, the proposed estimators have a smaller bias hence a smaller MSE. With a two-dimensional $\mathbf{U}$, the bias, standard deviation and the MSE are all much smaller than the method in [23]. One may notice that we do not report the estimated standard error and the coverage probability results for the method in [23]. This concerns the computation time. For the two-dimensional $\mathbf{U}$ case, our method with sample size 200 only takes less than a minute to finish 1000 simulation replicates whereas the method in [23] takes about 25 minutes. When the sample size increases to 1000, our method takes about 10 minutes but the method in [23] would need about 16 hours, not even to mention that the standard error estimation needs 200 bootstrap samples. The computational complexity of the method in [23] comes from the fact that their kernel estimation has to be conducted for every value of β in every iteration. Due to the facts that our goal of the paper is not to replicate the results in [23] and that the superior performance of our proposed methods is already very apparent without comparing the estimated standard error, we decide not to report the standard error estimates of the method in [23].

7. Data Application

In this section, we analyze a data set from the Korean Labor & Income Panel Study with its details at https://www.kli.re.kr/klips_eng/index.do. This data set contains $N = 2506$ regular wage earners, and the scientific interest is to estimate the mean of the monthly income in the year of 2006 (outcome Y), which has around 35% missing values. The fully observed covariates $\mathbf{X}$ in this data set include: monthly income in 2005 (continuous), gender (discrete with two categories: male, female), age (discrete with three categories: less than 35, between 35 and 51, greater than 51), education (discrete with two categories: up to high school and beyond high school). This data set was analyzed in [28] and [23] where the former simply assumed a parametric logistic regression model for the propensity.

According to [23], the missingness of the 2006 income will definitely depend on the 2006 income itself as well as the 2005 income. Further, it is reasonable to assume that, given 2005 income and 2006 income, the missingness does not depend on some or all of gender, age and education. We follow the strategy in [23] to consider different instruments $\mathbf{Z}$ in our analysis.

Regarding the working model of $g(u)$, we first fit a simple logistic regression between R and U, where U is the 2005 income, then use its logit as the working

model. That is,

$$g^*(u) = 0.67 + 0.69u. \tag{7.1}$$

For a comparison, we also arbitrarily pick another working model

$$g^*(u) = 0.3 + 0.1u. \tag{7.2}$$

We implement the proposed estimator $\widetilde{\theta}^*$ presented in Section 5. The estimation results as well as the corresponding estimated standard errors based on 200 bootstrap samples are summarized in Table 4. Table 4 also outlines the results from the naïve method $(\sum_{i=1}^N r_i y_i / \sum_{i=1}^N r_i)$, the parametric method proposed in [28] and [23].

From Table 4, the naïve method, which does not take into account the nonignorable missingness, clearly over-estimates the monthly income in 2006, therefore the naïve method is severely biased. The parametric method proposed in [28] can largely alleviate the estimation bias, but it seems sensitive to the choice of the instrument, i.e., the model misspecification. Additionally, the estimated standard errors from this method is much larger than the methods that assume the semiparametric propensity. One potential reason is that the parametric propensity is very fragile to the misspecified $g(\mathbf{u})$. In general, [23] and the proposed estimator $\widetilde{\theta}^*$ give very similar results to both parameter estimation and standard error estimation. Interestingly, $\widetilde{\theta}^*$ always estimates the monthly income in 2006 just slightly larger than the method in [23].

TABLE 4
Estimates of monthly income in 2006 (standard error estimates in parentheses) from the Korean Labor & Income Panel Study data set.

Instrument $\mathbf{z}$	Naïve	Parametric	Method in [23]	$\widetilde{\theta}^*$ with (7.1)	$\widetilde{\theta}^*$ with (7.2)
all		183.9(2.6)	186.0 (2.4)	188.6 (2.6)	187.4 (2.4)
age,education		185.4(3.0)	186.0 (2.4)	188.4 (2.7)	187.4 (2.4)
age,gender		183.1(3.8)	186.3 (2.4)	188.9 (2.6)	187.8 (2.5)
education,gender	205.7(2.8)	183.2(4.5)	185.4 (2.1)	188.1 (2.6)	186.8 (2.6)
age		196.2(6.3)	186.4 (2.3)	189.7 (2.9)	188.2 (2.5)
education		188.2(5.9)	185.1 (2.2)	189.5 (2.8)	187.4 (2.8)
gender		186.1(5.5)	185.5 (2.3)	187.6 (2.5)	185.7 (2.5)

8. Discussion

In this paper, we propose a novel estimation procedure for the parameter β in the semiparametric propensity model for nonignorable missing data, as well as for a general functional $\boldsymbol{\theta}$. In this model, there are three nonparametric nuisance components: $f_{Y|\mathbf{X},R=1}(y,\mathbf{x})$, $f_{\mathbf{X}}(\mathbf{x})$ and $g(\mathbf{u})$, where the first two do not involve any missing data so their estimates are stand-alone and can be flexible, whereas the estimation of $g(\mathbf{u})$ is generally not straightforward.

We treat it as a semiparametric model and characterize its geometric structure. This geometric representation is useful because it will enable us to identify

influence functions of regular and asymptotically linear estimators, which, in turn, will motivate various estimating equations for the parameter of interest. We derive the efficient score $\mathbf{S}_{\text{eff}}$ for estimating β. Instead of directly employing this efficient score, we exploit different forms of $\mathbf{S}_{\text{eff}}$ with a working model $g^*(\mathbf{u})$ bearing in mind that the estimation of $g(\mathbf{u})$ is cumbersome. Fortunately, our paper shows that, this "working" efficient score function $\mathbf{S}^*_{\text{eff}}$ remains to be mean zero, for reasons pertaining to the logistic propensity form and its additive form inside the link function. This allows us to use a simpler, albeit possibly incorrect, model for $g(\mathbf{u})$ to form estimating equations whose solution will still yield consistent estimators of β. We also find out that the similar spirit applies to the estimation of $\boldsymbol{\theta}$ as well. In other words, our estimation procedure does not need the estimation of $g(\mathbf{u})$ at all. It is exciting that, compared to other methods in the literature, our proposal would potentially bring tremendous convenience for end-users, and greatly broaden the applicability of this semiparametric propensity model for real applications. This is the key innovation of our work.

Analyzing nonignorable missing data has been a challenging but active research area over the years. Most recently, [11] adopted a different semiparametric propensity model through an unspecified function of Y and a parametric form of $\mathbf{U}$. This modeling difference leads to completely different statistical treatments from us. Also, [13] developed some very interesting results in that mean functionals can still be identified and estimated even without fully identifying the whole data distribution.

To conclude our paper, we think it imperative to clearly summarize and contrast some nuanced differences of our work with some relevant literature.

8.1. Comparison with [23] and [27]

The model (1.2) was proposed in [23]. Their main idea of estimation is to construct sufficient estimating equations for β. Their estimating equations are based on

$$E[w_l(\mathbf{Z})\{R/\pi(Y,\mathbf{u},\beta,g)-1\}\mid \mathbf{u}]=0, l=1,\ldots,L,$$

where $w_l(\mathbf{z})$ is an arbitrary function of $\mathbf{z}$. If $\mathbf{Z}$ is discrete with the number of categories L satisfying $L \geqslant d+1$, the authors simply advocated using $w_l(\mathbf{z}) = I(\mathbf{z}=l)$; while when $\mathbf{Z}$ is continuous, the authors suggested replacing them by some moment functions. Inserting the form of $\pi(Y,\mathbf{u},\beta,g)$, one obtains that

$$\exp\{-g(\mathbf{u})\}E[w_l(\mathbf{Z})R\exp\{-h(Y,\beta)\}\mid \mathbf{u}]=E\{w_l(\mathbf{Z})(1-R)\mid \mathbf{u}\}.$$

Further, to estimate $\exp\{-g(\mathbf{u})\}$, the authors simply chose $w_l(\mathbf{z})$ as a constant and then adopted the profiling technique, i.e., for each fixed β, they estimated $\exp\{-\widehat{g}(\mathbf{u},\beta)\}$ by

$$\exp\{-\widehat{g}(\mathbf{u},\beta)\}=\frac{\sum_{i=1}^N(1-r_i)K_t(\mathbf{u}-\mathbf{u}_i)}{\sum_{i=1}^N r_i\exp\{-h(y_i,\beta)\}K_t(\mathbf{u}-\mathbf{u}_i)}, \tag{8.1}$$

where K is a kernel function and t is a bandwidth. In other words, the authors estimated the nonparametric component $g(\mathbf{u})$ via the kernel estimation. In their method, this step of estimating $g(\mathbf{u})$ is important and inevitable.

Motivated by enhancing the estimation efficiency, [27] also studied the same model (1.2) and proposed a variety of estimators for both β and $\boldsymbol{\theta}$ via maximum likelihood or calibration. In [27], the authors also relied on the above relation (8.1) for iteratively estimating and updating the nonparametric component $g(\mathbf{u})$.

Given that $g(\mathbf{u})$ is not at the center of scientific interest, in this paper, we explore the possibility of avoiding estimating it. The key message of our paper is, it is feasible to achieve this goal by carefully characterizing the semiparametric structure of the model.

The other key feature of our work, compared to [23] and [27], is that we rigorously investigated the semiparametric structure of the model (1.2) and derived the efficient score function $\mathbf{S}_{\text{eff}}$ in Proposition 2 for estimating β. Though our proposed estimators in this paper were not directly implemented according to $\mathbf{S}_{\text{eff}}$, Corollary 1 showed that, if $g(\mathbf{u})$ were known, our proposed estimators achieve the semiparametric efficiency bound, the best possible efficiency among all regular asymptotically linear estimators. To the contrary, neither [23] nor [27] theoretically studied the estimation efficiency of β. In [27], though their methods are numerically more efficient than that in [23], there is no theoretical support provided. In addition, building on the theoretical investigations in this paper, [24] also derived the efficient score for estimating the functional $\boldsymbol{\theta}$. There is an interesting tradeoff between efficiency and robustness in estimating $\boldsymbol{\theta}$, and readers of interest in this topic are encouraged to refer to [24] directly.

8.2. Comparison with [16] and [17]

The work of [16] and [17] followed the pattern-mixture factorization idea, and their model decomposition rests on the odds ratio function

$$\text{OR}(y,\mathbf{u}) = \frac{\text{pr}(R=0\mid y,\mathbf{u})\text{pr}(R=1\mid y=0,\mathbf{u})}{\text{pr}(R=1\mid y,\mathbf{u})\text{pr}(R=0\mid y=0,\mathbf{u})}$$

with $\text{OR}(y=0,\mathbf{u})=1$, and two other nonparametric components: $f_{Y,\mathbf{Z}|\mathbf{U},R=1}(y,\mathbf{z},\mathbf{u})$ and $\text{pr}(R=1\mid y=0,\mathbf{u})$. This set of decomposition is equivalent to what we consider in this paper with

$$\text{pr}(R=1\mid y,\mathbf{u}) = \text{expit}\{-\log\text{OR}(y,\mathbf{u}) + g(\mathbf{u})\}$$

where $g(\mathbf{u}) = \text{logit}\{\text{pr}(R=1\mid y=0,\mathbf{u})\}$. It is worthwhile to mention that, the model identifiability result in Section 3 of [16] does not need a parametric $\text{OR}(y,\mathbf{u})$, provided that it can be identified by solving a type-I Fredholm equation. Nevertheless, all other parts that follow, their Sections 4-6, all need a parametric $\text{OR}(y,\mathbf{u})$, which is essentially the same as the model (1.3). Especially, in the numerical part of [16], their Section 5, they assumed $\text{OR}(y,\mathbf{u}) = \text{OR}(y)$, which is actually the same as the model (1.2) of this paper.

Similar to [23], the difficulty in the estimation of [16] and [17] is how to estimate $\text{pr}(R = 1 \mid y = 0, \mathbf{u})$: out of the two other nonparametric components $f_{Y,\mathbf{Z}|\mathbf{U},R=1}(y, \mathbf{z}, \mathbf{u})$ and $\text{pr}(R = 1 \mid y = 0, \mathbf{u})$, $f_{Y,\mathbf{Z}|\mathbf{U},R=1}(y, \mathbf{z}, \mathbf{u})$ can be fitted and estimated using the observed data whereas $\text{pr}(R = 1 \mid y = 0, \mathbf{u})$ cannot. The doubly robust estimation presented in [16] and [17], especially the inverse probability weighted estimation, imposed a correctly specified parametric model of $\text{pr}(R = 1 \mid y = 0, \mathbf{u})$. To the contrary, the main goal of our paper is to propose estimators that can avoid estimating this nuisance component. This is the fundamental difference of our work with [16] and [17].

It is also helpful to clarify that the semiparametric efficiency theory in Section 6 of [16] did not exist in an earlier version of their paper. In our paper, the derivation of efficient influence function relies on the correct characterization of the nuisance tangent space and its complement. The process of our derivation is different and independent to Section 6 of [16]. More importantly, although the efficient influence function was derived in [16], their estimation strategy does not rely on it at all. The doubly robust estimator presented in Section 4 of [16] needs a user-specified function, and it is NOT semiparametrically efficient even if both $f_{Y,\mathbf{Z}|\mathbf{U},R=1}(y, \mathbf{z}, \mathbf{u})$ and $\text{pr}(R = 1 \mid y = 0, \mathbf{u})$ are correctly specified. One can check equations (13) and (14) in [16] for more details.

8.3. Comparison with [18] and [29]

In the literature, there exist other relevant papers that relied on different modeling assumptions than this work.

Instead of considering the semiparametric mechanism, [18] studied the efficient estimation based on a purely parametric mechanism $\pi(y, \mathbf{u})$. The efficient score for estimating β based on the parametric $\pi(y, \mathbf{u})$, presented in Lemma 4.1 of [18], can be regarded as a special case of Proposition 2 presented in Section 3 of this paper, though the derivation of Proposition 2 is quite more involved because of the nonparametric component $g(\mathbf{u})$. More importantly, using the efficient score in [18], which has a much simpler form, one can immediately propose adaptive and efficient estimators for β, that were presented in Section 5 of [18].

Further, instead of modeling the mechanism either parametrically or semiparametrically, [29] did not impose any assumption on $\pi(y, \mathbf{u})$ while as a price for this, they adopted a purely parametric model on the regression relation $f_{Y|\mathbf{X}}(y, \mathbf{x})$. They studied the estimation and inference of the regression parameter. Their approach is suitable for understanding the regression relation between Y and $\mathbf{X}$. Here, our regression relation between Y and $\mathbf{X}$ is totally unspecified, and we impose the weakest possible assumption which only specifies the dependence of the propensity score model on the outcome while allowing all other models completely unspecified. The scientific goal here is to estimate the parameter $\boldsymbol{\theta}$ that satisfies $E\{\boldsymbol{\zeta}(\mathbf{X}, Y, \boldsymbol{\theta})\} = \mathbf{0}$, instead of any regression parameters.

Technically, there exist some intrinsic difference, in terms of efficiency, of the estimators proposed in these two papers. The estimator proposed in [29], even

if the truth of the nuisance function is known, still cannot achieve the semiparametric efficiency bound. One can check their Remarks 1 and 2 for details. In this paper, Corollary 1 shows that, if the corresponding true nuisance is available, our proposed estimators are all efficient and reach the semiparametric efficiency bound.

Further, the robustness property established here is not due to the same reason as in [29]. In fact, the robustness property here is directly linked to the structure of the propensity score model and is a unique property.

Finally, the framework of [29] is simpler, where the ultimate goal is to estimate the parameter in $f_{Y|\mathbf{X}}(y,\mathbf{x})$ which can be achieved by directly working on its efficient score. The framework of this paper is two-step. Our ultimate goal is to estimate a functional $\boldsymbol{\theta}$ that satisfies $E\{\boldsymbol{\zeta}(\mathbf{X},Y,\boldsymbol{\theta})\}=\mathbf{0}$, but we need to first derive the efficient score of β, the parameter in the propensity score model. Our proposed estimators can avoid estimating the unknown function $g(\mathbf{u})$—this nice feature refers to not only estimating β but also estimating the functional of interest $\boldsymbol{\theta}$.

Acknowledgments

The authors would like to thank the Editor, an Associate Editor, and the two reviewers for their insightful comments which have helped improve the manuscript substantially.

Funding

The research is supported in part by NSF (DMS 1953526, 2122074, 2310942), NIH (R01DC021431) and the American Family Funding Initiative of UW-Madison.

Supplementary Material

Supplementary materials for "Robust estimation under a semiparametric propensity model for nonignorable missing data"
(doi: 10.1214/25-EJS2355SUPP; .pdf). The supplementary materials include the technical proofs, some more detailed theoretical results and discussions.

References

[1] Bickel, P. J., Klaassen, J., Ritov, Y. and Wellner, J. A. (1993). *Efficient and Adaptive Estimation for Semiparametric Models*. Johns Hopkins University Press Baltimore. MR1245941

[2] Chang, T. and Kott, P. S. (2008). Using calibration weighting to adjust for nonresponse under a plausible model. *Biometrika* **95** 555–571. MR2443175

[3] D'HAULTFOEUILLE, X. (2010). A new instrumental method for dealing with endogenous selection. *Journal of Econometrics* **154** 1–15. MR2558947
[4] FIELDING, S., FAYERS, P. M., MCDONALD, A., MCPHERSON, G. and CAMPBELL, M. K. (2008). Simple imputation methods were inadequate for missing not at random (MNAR) quality of life data. *Health and Quality of Life Outcomes* **6** 57.
[5] GOMES, M., GUTACKER, N., BOJKE, C. and STREET, A. (2016). Addressing missing data in patient-reported outcome measures (PROMS): Implications for the use of PROMS for comparing provider performance. *Health Economics* **25** 515–528.
[6] HU, Y. and SHIU, J.-L. (2018). Nonparametric identification using instrumental variables: sufficient conditions for completeness. *Econometric Theory* **34** 659–693. MR3803171
[7] 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.
[8] 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. MR2816710
[9] LEHMANN, E. L. and ROMANO, J. P. (2006). *Testing Statistical Hypotheses*. Springer Science & Business Media. MR2135927
[10] LEURENT, B., GOMES, M., FARIA, R., MORRIS, S., GRIEVE, R. and CARPENTER, J. R. (2018). Sensitivity analysis for not-at-random missing data in trial-based cost-effectiveness analysis: a tutorial. *PharmacoEconomics* **36** 889–901.
[11] LI, M., MA, Y. and ZHAO, J. (2022). Efficient estimation in a partially specified nonignorable propensity score model. *Computational Statistics & Data Analysis* **174** 107322. MR4437646
[12] LI, T., HUTFLESS, S., SCHARFSTEIN, D. O., DANIELS, M. J., HOGAN, J. W., LITTLE, R. J., ROY, J. A., LAW, A. H. and DICKERSIN, K. (2014). Standards should be applied in the prevention and handling of missing data for patient-centered outcomes research: a systematic review and expert consensus. *Journal of Clinical Epidemiology* **67** 15–32.
[13] LI, W., MIAO, W. and TCHETGEN, E. T. (2021). Identification and estimation of nonignorable missing outcome mean without identifying the full data distribution. *arXiv preprint arXiv:2110.05776*.
[14] LITTLE, R. J. (1993). Pattern-mixture models for multivariate incomplete data. *Journal of the American Statistical Association* **88** 125–134.
[15] LITTLE, R. J. and RUBIN, D. B. (2019). *Statistical Analysis with Missing Data* **793**. John Wiley & Sons. MR1925014
[16] 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 preprint arXiv:1509.02556*. MR3839004
[17] 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. MR3509900

[18] Morikawa, K. and Kim, J. K. (2021). Semiparametric optimal estimation with nonignorable nonresponse data. *The Annals of Statistics* **49** 2991–3014. MR4338901

[19] Newey, W. K. and Powell, J. L. (2003). Instrumental variable estimation of nonparametric models. *Econometrica* **71** 1565–1578. MR2000257

[20] 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. MR1947279

[21] 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.

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

[23] Shao, J. and Wang, L. (2016). Semiparametric inverse propensity weighting for nonignorable missing data. *Biometrika* **103** 175–187. MR3465829

[24] Shetty, S., Ma, Y. and Zhao, J. (2023). The Pursuit of Efficiency versus Robustness: A Learning Experience from Analyzing a Semiparametric Nonignorable Propensity Score Model. *Observational Studies* **9** 97–104.

[25] Shetty, S., Ma, Y. and Zhao, J. (2025). Supplementary materials for "Robust estimation under a semiparametric propensity model for nonignorable missing data". 10.1214/25-EJS2355SUPP.

[26] Tsiatis, A. A. (2006). *Semiparametric Theory and Missing Data*. New York: Springer. MR2233926

[27] Uehara, M., Lee, D. and Kim, J.-K. (2023). Statistical inference with semiparametric nonignorable nonresponse models. *Scandinavian Journal of Statistics* **50** 1795–1817. MR4677376

[28] 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. MR3241279

[29] Zhao, J. and Ma, Y. (2021). A versatile estimation procedure without estimating the nonignorable missingness mechanism. *Journal of the American Statistical Association* 1–15. MR4528480