Computational Statistics and Data Analysis ••• (••••) ••••••

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

www.elsevier.com/locate/csda

Efficient estimation in a partially specified nonignorable propensity score model ☆

Mengyan Li [a], Yanyuan Ma [b], Jiwei Zhao [c,*]

[a] Department of Mathematical Sciences, Bentley University, United States of America
[b] Department of Statistics, The Pennsylvania State University, United States of America
[c] Department of Biostatistics and Medical Informatics, University of Wisconsin-Madison, United States of America

ARTICLE INFO

Article history:
Received 6 January 2021
Received in revised form 11 July 2021
Accepted 12 July 2021
Available online xxxx

Keywords:
Nonignorable missing data
Propensity score
Semiparametric model
Efficient score method

ABSTRACT

Consider the regression setting where the response variable is subject to missing data and the covariates are fully observed. A nonignorable propensity score model, i.e., the probability that the response is observed conditional on all variables depends on the missing values themselves, is assumed throughout the paper. In such problems, model misspecification and model identifiability are two critical issues. A fully parametric approach can produce results that are sensitive to the model assumptions, while a fully nonparametric approach may not be sufficient for model identification. A new flexible semiparametric propensity score model is proposed where the relationship between the missingness indicator and the partially observed response is totally unspecified and estimated nonparametrically, while the relationship between the missingness indicator and the fully observed covariates is modeled parametrically. The proposed estimator is constructed via a semiparametric treatment and is proved to be semiparametrically efficient. Comprehensive simulation studies are conducted to examine the finite-sample performance of the estimators. While the naive parametric method leads to heavily biased estimator and poor coverage results, the proposed method produces estimator with negligible finite-sample biases and also correct inference results. The proposed method is further illustrated via an electronic health records (EHR) data application for the albumin level in the blood sample. The empirical analyses demonstrated that the proposed semiparametric propensity score model is more sensible than a purely parametric model. The proposed method could be very useful to uncover the unknown and possibly nonlinear dependence of the propensity score model to the albumin level, and is recommended for practical use.

1. Introduction

Missing data are ubiquitous in many areas of scientific inquiry, especially in research involving human subjects, such as health related studies and sample surveys. Differentiating the different natures of missingness through the propensity score model (Rosenbaum and Rubin, 1983) is crucial in statistical analysis with missing data. The propensity score model,

☆ The programming code that can reproduce the numerical results of the paper is available at https://github.com/MengyanLi1992/Efficient-Estimation-in-a-Partially-Specified-Nonignorable-Propensity-Score-Model.

* Corresponding author at: WARF Office Building, 610 Walnut Street, Room 285, Madison WI 53726.
E-mail addresses: mengyanli@bentley.edu (M. Li), yzm63@psu.edu (Y. Ma), jiwei.zhao@wisc.edu (J. Zhao).

https://doi.org/10.1016/j.csda.2021.107322
0167-9473/

i.e., the distribution of the missingness indicator conditional on all variables, is called ignorable if it does not depend on the missing values. Otherwise, it is called nonignorable, also known as missing not at random (Little and Rubin, 2002). Extensive literature exists on ignorable missing data (Rubin, 1978; Robins et al., 1994; Schafer, 1997; Little and Rubin, 2002; Tsiatis, 2006; Kim and Shao, 2013; Molenberghs et al., 2014), while nonignorable missingness is more challenging with less research on it.

The assumption on ignorable missingness can be violated in some applications. Specifically, in the electronic health record (EHR) data example that motivated our work, data are collected in a non-prescheduled fashion. Thus, data are only available when a patient seeks care or a physician orders care, therefore the visiting process is potentially reflective of a patient's risk category. In other words, it is most likely that the propensity score will depend on the missing value itself. In such situation, missing value dependent propensity score models are needed for handling nonignorable missingness.

Many earlier literatures model the propensity score parametrically, see Ibrahim and Lipsitz (1996); Rotnitzky and Robins (1997); Qin et al. (2002); Chang and Kott (2008); Wang et al. (2014); Morikawa and Kim (2021). However, parametric model assumptions are generally restrictive and subject to model misspecification. On the other extreme, nonparametric propensity score is also considered in recent literatures (Tang et al., 2003; Shao and Zhao, 2013; Kott, 2014; Wang et al., 2014; Zhao and Shao, 2015; Zhao and Ma, 2021). Although nonparametric models avoid the risk of model misspecification, they also lose the ability to model any aspect of the model more specifically hence to improve subsequent inference results. A middle ground that encompasses the advantages of both parametric and nonparametric propensity scores is semiparametric models. To this end, Kim and Yu (2011) and Shao and Wang (2016) proposed a semiparametric logistic regression model for the propensity, where the relationship between the missingness indicator and the fully observed variables is left unspecified, while the relationship between the missingness indicator and the variable that contains missing values is parametrically modeled. While these works open the door to the semiparametric modeling of the propensity score, the particular strategy encounters two difficulties. First, it encounters a curse of dimensionality issue in handling the nonparametric component in the propensity score (Shao and Wang, 2016). Second, because the relationship between the variable subject to missingness and the missingness indicator is far more difficult to grasp than the relationship between fully observed variables and the missingness indicator, it is likely more sensible to allow more flexibility and consider a complement modeling strategy. In other words, the missingness indicator nonparametrically depends on the variable subject to missing, while parametrically depends on the fully observed variables. This is the propensity model we propose in this work.

Regardless the propensity is modeled parametrically, nonparametrically or semiparametrically, nonignorable missing data problems encounter a universal identification issue. It is known (Robins and Ritov, 1997; Miao et al., 2016) that if we assume that the missingness depends on both the variable subject to missingness and all fully observed variables, then the problem is not identifiable. Two types of additional assumptions are usually made to achieve identifiability, the instrumental variable approach (Tchetgen Tchetgen and Wirth, 2017; Sun et al., 2018) and the shadow variable approach (Shao and Zhao, 2013; Kott, 2014; Wang et al., 2014; Zhao and Shao, 2015). Here, we adopt the shadow variable approach, where we assume the missingness depends on the variable subject to missingness and only part of the fully observed variables. The part that has no involvement in the propensity score model is termed shadow variable.

The semiparametric propensity score with the variable subject to missingness modeled nonparametrically turns out challenging and interesting both methodologically and theoretically, because standard nonparametric procedures no longer apply. To tackle this problem, we devise a new asymptotically consistent likelihood-based estimation method for the nuisance functions in the presence of nonignorable missing data, followed by a nonstandard nonparametric estimation procedure. We show that the proposed estimator is semiparametrically efficient.

In Section 2, we present our model, and devise estimators for the parameters of interest in the parametric parts by deriving efficient scores. Details of implementation and algorithm are given in Section 2.3. In Section 3, we establish the asymptotic properties of the newly proposed semiparametrically efficient estimator. In Section 4, we examine the finite-sample performance of our method through simulation studies. The application of our method to the motivating data is presented in Section 5. The paper is concluded with some discussions in Section 6.

2. Model and estimation

Consider N observations $(\mathbf{x}_i, r_i y_i, r_i)$, $i = 1, \ldots, N$, which are independent and identically distributed realizations of $(\mathbf{X}, RY, R)$. Here $\mathbf{X}$ is a d_x-dimensional fully observed covariate, R is a binary missingness indicator, and the scalar response Y is observed if and only if the indicator $R = 1$. Without loss of generality, assume that for $i = 1, \ldots, n$, $R_i = 1$, while for $i = n+1, \ldots, N$, $R_i = 0$. Following the shadow variable approach, we write $\mathbf{X} = (\mathbf{U}^{\mathrm{T}}, \mathbf{Z}^{\mathrm{T}})^{\mathrm{T}}$, where $\mathbf{U}$ is d_u-dimensional and $\mathbf{Z}$ is $(d_x - d_u)$-dimensional, $d_x > d_u \geq 0$. We term $\mathbf{Z}$ the shadow variable in the sense that, given the variables Y and $\mathbf{U}$, the indicator R and the shadow variable $\mathbf{Z}$ are conditionally independent, i.e., $\mathrm{pr}(R = 1 \mid y, \mathbf{x}) = \mathrm{pr}(R = 1 \mid y, \mathbf{u})$. Thus, the shadow variable $\mathbf{Z}$ does not contribute to the propensity score model. The shadow variable assumption is widely adopted to achieve model identification in nonignorable missing data problems. We use a semiparametric model of the form

$$\mathrm{pr}(R = 1 \mid y, \mathbf{u}) = \pi(y, \mathbf{u}; \boldsymbol{\beta}, g) = \mathrm{expit}\{g(y) + h(\mathbf{u}; \boldsymbol{\beta})\}, \tag{1}$$

for the propensity score, where $\text{expit}(\cdot) = \exp(\cdot)/\{1+\exp(\cdot)\}$, $\boldsymbol{\beta}$ is a q-dimensional unknown parameter, h is a known function, and g is an unspecified function. For identification, we incorporate intercept into $h(\mathbf{u};\boldsymbol{\beta})$ and require $g(0)=0$. On the other hand, we use the familiar exponential family model to describe the fully observed data,

$$f_{Y|\mathbf{X},R=1}(y,\mathbf{x};\boldsymbol{\alpha}) = f_{Y|\mathbf{Z},R=1}(y,\mathbf{z};\boldsymbol{\alpha}) = \exp\{y\eta(\mathbf{z};\boldsymbol{\alpha}) + \rho(\mathbf{z};\boldsymbol{\alpha}) + \tau(y;\boldsymbol{\alpha})\}, \tag{2}$$

where $\boldsymbol{\alpha}$ is a $(p-q)$-dimensional unknown parameter, and η, ρ and τ are all given functions satisfying $\rho(\mathbf{z};\boldsymbol{\alpha}) = -\log[\int \exp\{y\eta(\mathbf{z};\boldsymbol{\alpha}) + \tau(y;\boldsymbol{\alpha})\}dy]$. In Appendix A.1, we show that the model specified in (1) and (2) is identifiable. We summarize the result in Lemma 1.

Lemma 1. *Under the model assumptions (1) and (2), the unknown components $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$ and $g(\cdot)$ are all identifiable.*

For the model described in (1) and (2), the likelihood function of one observation, which is the joint pdf/pmf of (X, RY, R), is given by

$$\begin{aligned} f_{\mathbf{X},RY,R}(\mathbf{x},ry,r) &= f_{\mathbf{X}}(\mathbf{x})\{\pi(y,\mathbf{u};\boldsymbol{\beta},g) f_{Y|\mathbf{X}}(y,\mathbf{x})\}^r \left\{\int \{1-\pi(y,\mathbf{u};\boldsymbol{\beta},g)\} f_{Y|\mathbf{X}}(y,\mathbf{x})dy\right\}^{1-r} \\ &= f_{\mathbf{X}}(\mathbf{x}) w(\mathbf{x};\boldsymbol{\beta},\boldsymbol{\alpha},g) f_{Y|\mathbf{Z},R=1}(y,\mathbf{z};\boldsymbol{\alpha})^r \left\{\frac{1-w(\mathbf{x};\boldsymbol{\alpha},\boldsymbol{\beta},g)}{w(\mathbf{x};\boldsymbol{\alpha},\boldsymbol{\beta},g)}\right\}^{1-r}, \end{aligned} \tag{3}$$

where $f_{\mathbf{X}}(\mathbf{x})$ is the pdf/pmf of $\mathbf{x}$ and

$$\begin{aligned} w(\mathbf{x};\boldsymbol{\alpha},\boldsymbol{\beta},g) &\equiv \text{pr}(R=1 \mid \mathbf{X}=\mathbf{x}) \\ &= \frac{1}{1+\exp\{-h(\mathbf{u};\boldsymbol{\beta})\}E[\exp\{-g(Y)\} \mid \mathbf{z}, R=1]}, \end{aligned}$$

and we used the fact that

$$f_{Y|\mathbf{X}}(y,\mathbf{x}) = \frac{f_{Y|\mathbf{X},R=1}(y,\mathbf{x})/\pi(y,\mathbf{u};\boldsymbol{\beta},g)}{\int f_{Y|\mathbf{X},R=1}(y,\mathbf{x})/\pi(y,\mathbf{u};\boldsymbol{\beta},g)dy} = \frac{f_{Y|\mathbf{Z},R=1}(y,\mathbf{z};\boldsymbol{\alpha})/\pi(y,\mathbf{u};\boldsymbol{\beta},g)}{\int f_{Y|\mathbf{Z},R=1}(y,\mathbf{z};\boldsymbol{\alpha})/\pi(y,\mathbf{u};\boldsymbol{\beta},g)dy}.$$

This likelihood is semiparametric where $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are two unknown parameters, $g(y)$ and $f_{\mathbf{X}}(\mathbf{x})$ are two unknown functions. Following the semiparametric convention, in our following derivation, we name $\boldsymbol{\theta} = (\boldsymbol{\alpha}^{\mathrm{T}}, \boldsymbol{\beta}^{\mathrm{T}})^{\mathrm{T}}$ as the p-dimensional parameter of interest, and name $f_{\mathbf{X}}(\mathbf{x})$ and $g(y)$ as nuisance functions, although we provide estimation of $g(y)$ as well.

In the following, we will take a semiparametric approach (Bickel et al., 1993; Tsiatis, 2006) and derive the nuisance tangent space, and the efficient score with respect to $\boldsymbol{\theta}$, followed by constructing a regular and asymptotically linear estimator for $\boldsymbol{\theta}$. The estimator for $g(y)$ is quite unusual due to the need of handling partially missing y_i values, and will be provided as a by-product. For notation simplicity, we write $w(\mathbf{x};\boldsymbol{\theta},g)$ as $w(\mathbf{x})$ and $E(\cdot \mid \mathbf{z}, R=1)$ as $E(\cdot \mid \mathbf{z}, 1)$.

2.1. Nuisance tangent space and efficient score

Consider the Hilbert space $\mathcal{H}$ of all p-dimensional zero-mean measurable functions of the observed data with finite variance, equipped with the inner product

$$\langle h_1, h_2\rangle = E\{h_1^{\mathrm{T}}(\mathbf{X}, RY, R) h_2(\mathbf{X}, RY, R)\},$$

where $h_1, h_2 \in \mathcal{H}$. Nuisance tangent space is defined as the mean squared closure of the nuisance tangent spaces of parametric sub-models spanned by the nuisance score vectors. By simple calculations, we can show that the nuisance tangent space for $f_{\mathbf{X}}(\mathbf{x})$ is

$$\Lambda_f = \left[\mathbf{a}(\mathbf{x}) \in \mathbb{R}^p : E\{\mathbf{a}(\mathbf{X})\} = \mathbf{0}\right],$$

and the nuisance tangent space for $g(y)$ is

$$\begin{aligned} \Lambda_g &= \left(E[\mathbf{a}(Y)\{1-\pi(Y,\mathbf{u};\boldsymbol{\beta},g)\} \mid \mathbf{x}]\frac{r-w(\mathbf{x})}{1-w(\mathbf{x})} : \mathbf{a}(y) \in \mathbb{R}^p\right) \\ &= \left(E[\mathbf{a}(Y)\{\pi^{-1}(Y,\mathbf{u};\boldsymbol{\beta},g)-1\} \mid \mathbf{z}, 1]\frac{w(\mathbf{x})\{r-w(\mathbf{x})\}}{1-w(\mathbf{x})} : \mathbf{a}(y) \in \mathbb{R}^p\right). \end{aligned}$$

The detailed derivation of Λ_g is provided in Appendix A.2. Since $E[\{R-w(\mathbf{x})\}/\{1-w(\mathbf{x})\} \mid \mathbf{x}] = 0$, we can easily verify that $\Lambda_f \perp \Lambda_g$. Thus, the nuisance tangent space is $\Lambda = \Lambda_f \oplus \Lambda_g$, where $\oplus$ stands for the addition of two spaces that are orthogonal to each other. Let $\Lambda^{\perp}$ denote the orthogonal complement of Λ, then $\mathcal{H} = \Lambda_f \oplus \Lambda_g \oplus \Lambda^{\perp}$.

The efficient score is defined as the orthogonal projection of the score vector $\mathbf{S}_{\boldsymbol{\theta}}$ onto $\Lambda^{\perp}$. Let $\mathbf{S}(y,\mathbf{z};\boldsymbol{\alpha}) \equiv \partial \log f_{Y|\mathbf{Z},R=1}(y,\mathbf{z};\boldsymbol{\alpha})/\partial\boldsymbol{\alpha}$. In Appendix A.3 we show that the score vector $\mathbf{S}_{\boldsymbol{\theta}} = (\mathbf{S}_{\boldsymbol{\alpha}}^{\mathrm{T}}, \mathbf{S}_{\boldsymbol{\beta}}^{\mathrm{T}})^{\mathrm{T}}$ is given as

$$\mathbf{S}_{\boldsymbol{\alpha}}(\mathbf{x}, ry, r; \boldsymbol{\theta}, g) = r\mathbf{S}(y,\mathbf{z};\boldsymbol{\alpha}) - E\{\mathbf{S}(Y,\mathbf{z};\boldsymbol{\alpha}) \mid \mathbf{x}\}\frac{r-w(\mathbf{x})}{1-w(\mathbf{x})}, \text{ and}$$
$$\mathbf{S}_{\boldsymbol{\beta}}(\mathbf{x}, ry, r; \boldsymbol{\theta}, g) = \mathbf{h}'_{\boldsymbol{\beta}}(\mathbf{u};\boldsymbol{\beta})\{r-w(\mathbf{x})\},$$

where $\mathbf{h}'_{\boldsymbol{\beta}}(\mathbf{u};\boldsymbol{\beta})$ is the derivative of $h(\mathbf{u};\boldsymbol{\beta})$ with respect to $\boldsymbol{\beta}$. Since $E\{\mathbf{S}_{\boldsymbol{\alpha}}(\mathbf{x}, RY, R; \boldsymbol{\theta}, g) \mid \mathbf{x}\} = \mathbf{0}$ and $E\{\mathbf{S}_{\boldsymbol{\beta}}(\mathbf{x}, RY, R; \boldsymbol{\theta}, g) \mid \mathbf{x}\} = \mathbf{0}$, it can be easily seen that, for any $p \times (p-q)$ constant matrix $\mathbf{G}_1$, $\mathbf{G}_1\mathbf{S}_{\boldsymbol{\alpha}}(\mathbf{x}, ry, r; \boldsymbol{\theta}, g) \in \Lambda_g \oplus \Lambda^{\perp}$, and for any $p \times q$ constant matrix $\mathbf{G}_2$, $\mathbf{G}_2\mathbf{S}_{\boldsymbol{\beta}}(\mathbf{x}, ry, r; \boldsymbol{\theta}, g) \in \Lambda_g \oplus \Lambda^{\perp}$.

Define $\mathbf{S}_{\boldsymbol{\theta},\mathrm{eff}} = (\mathbf{S}_{\boldsymbol{\alpha},\mathrm{eff}}^{\mathrm{T}}, \mathbf{S}_{\boldsymbol{\beta},\mathrm{eff}}^{\mathrm{T}})^{\mathrm{T}}$, where $\mathbf{S}_{\boldsymbol{\alpha},\mathrm{eff}}(\mathbf{x}, ry, r; \boldsymbol{\theta}, g)$ and $\mathbf{S}_{\boldsymbol{\beta},\mathrm{eff}}(\mathbf{x}, ry, r; \boldsymbol{\theta}, g)$ are the projections of $\mathbf{S}_{\boldsymbol{\alpha}}(\mathbf{x}, ry, r; \boldsymbol{\theta}, g)$ and $\mathbf{S}_{\boldsymbol{\beta}}(\mathbf{x}, ry, r; \boldsymbol{\theta}, g)$ onto the space $\Lambda^{\perp}$, respectively. By simple calculations, we obtain

$$\begin{aligned}
&\mathbf{S}_{\boldsymbol{\alpha},\mathrm{eff}}(\mathbf{x}, ry, r; \boldsymbol{\theta}, g)\\
&= r\mathbf{S}(y,\mathbf{z};\boldsymbol{\alpha}) - \frac{r-w(\mathbf{x})}{1-w(\mathbf{x})}\left(E\{\mathbf{S}(Y,\mathbf{z};\boldsymbol{\alpha}) \mid \mathbf{x}\} + E[\mathbf{a}_0(Y)\{1-\pi(Y,\mathbf{u};\boldsymbol{\beta},g)\} \mid \mathbf{x}]\right)\\
&= r\mathbf{S}(y,\mathbf{z};\boldsymbol{\alpha})\\
&\quad - \frac{r-w(\mathbf{x})}{E[\exp\{-g(Y)\} \mid \mathbf{z},1]}\left(E[\mathbf{S}(Y,\mathbf{z};\boldsymbol{\alpha})\exp\{-g(Y)\} \mid \mathbf{z},1] + E[\mathbf{a}_0(y)\exp\{-g(Y)\} \mid \mathbf{z},1]\right),
\end{aligned} \tag{4}$$

where $\mathbf{a}_0(y)$ satisfies

$$\begin{aligned}
&E\left(E[\mathbf{a}_0(Y)\{1-\pi(Y,\mathbf{U};\boldsymbol{\beta},g)\} \mid \mathbf{X}]\frac{w(\mathbf{X})\{1-\pi(y,\mathbf{U};\boldsymbol{\beta},g)\}}{1-w(\mathbf{X})} \mid y\right)\\
&= -E\left[E\{\mathbf{S}(Y,\mathbf{Z};\boldsymbol{\alpha}) \mid \mathbf{X}\}\frac{w(\mathbf{X})\{1-\pi(y,\mathbf{U};\boldsymbol{\beta},g)\}}{1-w(\mathbf{X})} \mid y\right].
\end{aligned} \tag{5}$$

The efficient score for $\boldsymbol{\beta}$ is

$$\begin{aligned}
&\mathbf{S}_{\boldsymbol{\beta},\mathrm{eff}}(\mathbf{x}, ry, r; \boldsymbol{\theta}, g)\\
&= \frac{r-w(\mathbf{x})}{1-w(\mathbf{x})}(\{1-w(\mathbf{x})\}\mathbf{h}'_{\boldsymbol{\beta}}(\mathbf{u};\boldsymbol{\beta}) - E[\mathbf{a}_1(Y)\{1-\pi(Y,\mathbf{u};\boldsymbol{\beta},g)\} \mid \mathbf{x}])\\
&= \{r-w(\mathbf{x})\}\mathbf{h}'_{\boldsymbol{\beta}}(\mathbf{u};\boldsymbol{\beta}) - \frac{r-w(\mathbf{x})}{E[\exp\{-g(Y)\} \mid \mathbf{z},1]}E[\mathbf{a}_1(Y)\exp\{-g(Y)\} \mid \mathbf{z},1],
\end{aligned} \tag{6}$$

where $\mathbf{a}_1(y)$ satisfies

$$\begin{aligned}
&E\left(E[\mathbf{a}_1(Y)\{1-\pi(Y,\mathbf{U};\boldsymbol{\beta},g)\} \mid \mathbf{X}]\frac{w(\mathbf{X})\{1-\pi(y,\mathbf{U};\boldsymbol{\beta},g)\}}{1-w(\mathbf{X})} \mid y\right)\\
&= E[w(\mathbf{X})\{1-\pi(y,\mathbf{U};\boldsymbol{\beta},g)\}\mathbf{h}'_{\boldsymbol{\beta}}(\mathbf{U};\boldsymbol{\beta}) \mid y].
\end{aligned} \tag{7}$$

In Appendix A.4, we verify that $\mathbf{S}_{\boldsymbol{\alpha},\mathrm{eff}}(\mathbf{x}, ry, r; \boldsymbol{\theta}, g)$ and $\mathbf{S}_{\boldsymbol{\beta},\mathrm{eff}}(\mathbf{x}, ry, r; \boldsymbol{\theta}, g)$ given in (4) and (6) are projections of $\mathbf{S}_{\boldsymbol{\alpha}}(\mathbf{x}, ry, r; \boldsymbol{\theta}, g)$ and $\mathbf{S}_{\boldsymbol{\beta}}(\mathbf{x}, ry, r; \boldsymbol{\theta}, g)$ onto $\Lambda^{\perp}$, respectively.

2.2. Nuisance function estimation

Despite of the results above, the efficient score $\mathbf{S}_{\boldsymbol{\theta},\mathrm{eff}}$ is not readily implementable because it contains the unknown quantities $f_{\mathbf{X}}(\mathbf{x})$ and $g(y)$. For $f_{\mathbf{X}}(\mathbf{x})$, due to the key observations

$$E\{R - w(\mathbf{x}) \mid \mathbf{x}\} = 0 \text{ and } E\{R\mathbf{S}(Y,\mathbf{z};\boldsymbol{\alpha}) \mid \mathbf{x}\} = \mathbf{0},$$

any working model $f^*_{\mathbf{X}}(\mathbf{x})$ can be used to construct estimating equations and the mean-zero property retains. On the other hand, the estimation of $f_{\mathbf{X}}(\mathbf{x})$ does not involve any missing data, and it only appears in conditional expectations in (5) and (7). Therefore, we recommend to simply approximate those expectations using their corresponding empirical versions. This corresponds to employing the empirical estimator $\widehat{f}_{\mathbf{X}}(\mathbf{x})$.

For $g(y)$, we opt a local constant approximation; i.e., at a fixed y_0, we employ γ to replace $g(y_0)$. Specifically, for any fixed y_0, we have

$$\pi(y_0,\mathbf{u};\boldsymbol{\beta},\gamma) \equiv \mathrm{pr}(R=1 \mid \mathbf{X}=\mathbf{x}, Y=y_0) = \mathrm{expit}\{\gamma + h(\mathbf{u};\boldsymbol{\beta})\}, \text{ and}$$
$$w(\mathbf{x}, y_0) \equiv \mathrm{pr}(R=1 \mid \mathbf{X}=\mathbf{x}) = \mathrm{expit}\{\gamma + h(\mathbf{u};\boldsymbol{\beta})\}.$$

The score function for γ is given as

$$S_\gamma(\mathbf{x}, ry_0, r; \boldsymbol{\beta}, \gamma) = r - w(\mathbf{x}, y_0) = r - \pi(y_0, \mathbf{u}; \boldsymbol{\beta}, \gamma).$$

One intuitive way to estimate γ, i.e., $g(y_0)$, is to solve the following estimating equation

$$\frac{1}{N}\sum_{i=1}^{N}\{r_i - \pi(y_0, \mathbf{u}_i; \boldsymbol{\beta}, \gamma)\}K_h(y_i - y_0) = 0,$$

where $K_h(\cdot)$ is a kernel function with bandwidth h, i.e. $K_h(\cdot) = K(\cdot/h)/h$, $K(\cdot)$ is a kernel function and h is a bandwidth. It is equivalent to

$$N^{-1}\sum_{i=1}^{n}\{1 - \pi(y_0, \mathbf{u}_i; \boldsymbol{\beta}, \gamma)\}K_h(y_i - y_0) = N^{-1}\sum_{i=n+1}^{N}\pi(y_0, \mathbf{u}_i; \boldsymbol{\beta}, \gamma)K_h(y_i - y_0). \tag{8}$$

However, y_i is missing when $i = n+1, \ldots, N$. We use $f_{\mathbf{X}|R}(\mathbf{x}, r)$ to denote the conditional pdf/pmf of $\mathbf{X}$ given R and the right-hand-side of (8) can be approximated by

$$\begin{aligned}
&\mathrm{pr}(R=0)E\left\{\pi(y_0, \mathbf{U}; \boldsymbol{\beta}, \gamma)K_h(Y - y_0) \mid R = 0\right\}\\
&= \mathrm{pr}(R=0)E\left[\pi(y_0, \mathbf{U}; \boldsymbol{\beta}, \gamma)E\left\{K_h(Y - y_0) \mid \mathbf{X}, R = 0\right\} \mid R = 0\right]\\
&\approx N^{-1}\sum_{i=n+1}^{N}\frac{E[K_h(Y - y_0)\exp\{-g(Y)\} \mid \mathbf{z}_i, 1]}{E[\exp\{-g(Y)\} \mid \mathbf{z}_i, 1][1 + \exp\{-\gamma - h(\mathbf{u}_i, \boldsymbol{\beta})\}]}.
\end{aligned}$$

Hence, the approximate estimating equation for γ is

$$\begin{aligned}
&N^{-1}\sum_{i=1}^{n}\frac{1}{1 + \exp\{\gamma + h(\mathbf{u}_i; \boldsymbol{\beta})\}}K_h(y_i - y_0)\\
&= N^{-1}\sum_{i=n+1}^{N}\frac{E[K_h(Y - y_0)\exp\{-g(Y)\} \mid \mathbf{z}_i, 1]}{E[\exp\{-g(Y)\} \mid \mathbf{z}_i, 1][1 + \exp\{-\gamma - h(\mathbf{u}_i, \boldsymbol{\beta})\}]}.
\end{aligned}$$

We propose to estimate $g(y)$ on L distinct points $(d_1, \ldots, d_L)$ approximately evenly distributed in the range of Y. Then estimate other $g(y)$ values by polynomial interpolation with degree $m-1$. Note that L goes to infinity when N goes to infinity, but $L \ll N$. Let $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_L)^{\mathrm{T}} = \{g(d_1), \ldots, g(d_L)\}^{\mathrm{T}}$, and we use $\widehat{g}(y; \boldsymbol{\gamma})$ to denote the approximate function. Employing the idea of profiling, at any $\boldsymbol{\theta}$, the L-dimensional vector $\boldsymbol{\gamma}(\boldsymbol{\theta})$ can be solved from the approximate estimating equation set that consists of L equations

$$N^{-1}\sum_{i=1}^{N}\mathbf{S}_g\{\mathbf{x}_i, r_i y_i, r_i; \boldsymbol{\theta}, \widehat{\boldsymbol{\gamma}}(\boldsymbol{\theta})\} = \mathbf{0}, \tag{9}$$

where the l-th component of $\mathbf{S}_g\{\mathbf{x}_i, r_i y_i, r_i; \boldsymbol{\theta}, \widehat{\boldsymbol{\gamma}}(\boldsymbol{\theta})\}$ is

$$\frac{r_i K_h(y_i - d_l)}{1 + \exp\{\gamma_l + h(\mathbf{u}_i; \boldsymbol{\beta})\}} + \frac{(r_i - 1)E[K_h(Y - y_0)\exp\{-\widehat{g}(Y; \boldsymbol{\gamma})\} \mid \mathbf{z}_i, 1]}{E[\exp\{-\widehat{g}(Y; \boldsymbol{\gamma})\} \mid \mathbf{z}_i, 1][1 + \exp\{-\gamma_l - h(\mathbf{u}_i, \boldsymbol{\beta})\}]}.$$

2.3. Implementation and algorithm

Based on the methodology presented above, we now summarize the algorithm below, followed by some elaborations.

In each iteration of step (b) of Algorithm 1, we need to solve integral equations (5) and (7) to obtain $\mathbf{a}_0(y)$ and $\mathbf{a}_1(y)$, respectively. Note that $E[a(\mathbf{X}, Y)\{1 - \pi(Y, \mathbf{U}; \boldsymbol{\beta}, g)\} \mid Y] = 0$ is equivalent to $E[a(\mathbf{X}, Y)\exp\{-h(\mathbf{U}; \boldsymbol{\beta})\} \mid Y, 1] = 0$ for any function $a(\mathbf{x}, y)$, as well as the facts that $f_{\mathbf{X}|R=1}(\mathbf{x}, 1) = f_{\mathbf{X}}(\mathbf{x})w(\mathbf{x})/\mathrm{pr}(R=1)$, and

$$f_{\mathbf{X}|Y,R=1}(\mathbf{x}, y) = \frac{f_{Y|\mathbf{Z},R=1}(y, \mathbf{z}; \boldsymbol{\alpha})f_{\mathbf{X}}(\mathbf{x})w(\mathbf{x})}{\int f_{Y|\mathbf{Z},R=1}(y, \mathbf{z}; \boldsymbol{\alpha})f_{\mathbf{X}}(\mathbf{x})w(\mathbf{x})d\mathbf{x}},$$

then (5) can be written as

Algorithm 1

Input: data $(\mathbf{u}_i, \mathbf{z}_i, r_i y_i, r_i)$, $i = 1, \ldots, N$; parameter initial value $\boldsymbol{\theta}^0$; error $\epsilon > 0$.
$t \leftarrow 0$;
while $\|\boldsymbol{\theta}^{t+1} - \boldsymbol{\theta}^t\| > \epsilon$ **do**
(a) solve the approximate estimating equation set (9) to obtain $\widehat{\boldsymbol{\gamma}}(\boldsymbol{\theta}^t) = \{\widehat{g}(d_1; \boldsymbol{\theta}^t), \ldots \widehat{g}(d_L; \boldsymbol{\theta}^t)\}^{\mathrm{T}}$;
(b) solve for the functions $\widehat{\mathbf{a}}_0(y)$ and $\widehat{\mathbf{a}}_1(y)$ from the following integral equations

$$\mathcal{L}(\mathbf{a}_0, \widehat{f}_{\mathbf{X}}; \boldsymbol{\theta}^t, \widehat{g})(y) = \boldsymbol{\phi}_0(\widehat{f}_{\mathbf{X}}; \boldsymbol{\theta}^t, \widehat{g})(y), \text{ and}$$

$$\mathcal{L}(\mathbf{a}_1, \widehat{f}_{\mathbf{X}}; \boldsymbol{\theta}^t, \widehat{g})(y) = \boldsymbol{\phi}_1(\widehat{f}_{\mathbf{X}}; \boldsymbol{\theta}^t, \widehat{g})(y);$$

(c) solve the estimating equation

$$\sum_{i=1}^{N} \widehat{\mathbf{S}}_{\boldsymbol{\theta},\mathrm{eff}}[\mathbf{x}_i, r_i y_i, r_i; \boldsymbol{\theta}, \widehat{g}\{\cdot; \widehat{\boldsymbol{\gamma}}(\boldsymbol{\theta})\}] = \mathbf{0}, \tag{10}$$

to obtain $\boldsymbol{\theta}^{t+1}$, where $\widehat{\mathbf{S}}_{\boldsymbol{\theta},\mathrm{eff}}[\mathbf{x}_i, r_i y_i, r_i; \boldsymbol{\theta}, \widehat{g}\{\cdot; \widehat{\boldsymbol{\gamma}}(\boldsymbol{\theta})\}]$ is $\mathbf{S}_{\boldsymbol{\theta},\mathrm{eff}}[\mathbf{x}_i, r_i y_i, r_i; \boldsymbol{\theta}, \widehat{g}\{\cdot; \widehat{\boldsymbol{\gamma}}(\boldsymbol{\theta})\}]$ with $\mathbf{a}_0(y)$ and $\mathbf{a}_1(y)$ replaced by $\widehat{\mathbf{a}}_0(y)$ and $\widehat{\mathbf{a}}_1(y)$, respectively;
(d) $t \leftarrow t + 1$;
end while
$\widehat{\boldsymbol{\theta}} \leftarrow \boldsymbol{\theta}^{t+1}$;
Output: $\widehat{\boldsymbol{\theta}}$.

$$\mathcal{L}(\mathbf{a}_0, f_{\mathbf{X}}; \boldsymbol{\theta}, g)(y) = \boldsymbol{\phi}_0(f_{\mathbf{X}}; \boldsymbol{\theta}, g)(y), \tag{11}$$

where $\mathcal{L}$ is a bilinear operator defined as

$$\mathcal{L}(\mathbf{a}_0, f_{\mathbf{X}}; \boldsymbol{\theta}, g)(y) = \iint \frac{w^2(\mathbf{x}) f_{Y|\mathbf{Z},R=1}(t, \mathbf{z}; \boldsymbol{\alpha}) f_{Y|\mathbf{Z},R=1}(y, \mathbf{z}; \boldsymbol{\alpha})}{\exp\{h(\mathbf{u}; \boldsymbol{\beta})\} E[\exp\{-g(Y)\} \mid \mathbf{z}, 1]} \exp\{-g(t)\} \mathbf{a}_0(t) f_{\mathbf{X}}(\mathbf{x}) dt d\mathbf{x},$$

and

$$\boldsymbol{\phi}_0(f_{\mathbf{X}}; \boldsymbol{\theta}, g)(y) = -\int E[\mathbf{S}(Y, \mathbf{z}; \boldsymbol{\alpha}) \exp\{-g(Y)\} \mid \mathbf{z}, 1] \frac{w^2(\mathbf{x}) f_{Y|\mathbf{Z},R=1}(y, \mathbf{z}; \boldsymbol{\alpha})}{\exp\{h(\mathbf{u}; \boldsymbol{\beta})\} E[\exp\{-g(Y)\} \mid \mathbf{z}, 1]} f_{\mathbf{X}}(\mathbf{x}) d\mathbf{x}.$$

Similarly, (7) can be written as

$$\mathcal{L}(\mathbf{a}_1, f_{\mathbf{X}}; \boldsymbol{\theta}, g)(y) = \boldsymbol{\phi}_1(f_{\mathbf{X}}; \boldsymbol{\theta}, g)(y), \tag{12}$$

where

$$\boldsymbol{\phi}_1(f_{\mathbf{X}}; \boldsymbol{\theta}, g)(y) = \int \frac{\mathbf{h}'_{\boldsymbol{\beta}}(\mathbf{u}; \boldsymbol{\beta}) w^2(\mathbf{x})}{\exp\{h(\mathbf{u}; \boldsymbol{\beta})\}} f_{Y|\mathbf{Z},R=1}(y, \mathbf{z}; \boldsymbol{\alpha}) f_{\mathbf{X}}(\mathbf{x}) d\mathbf{x}.$$

In practice, we discretize (11) and (12), and solve for $\mathbf{a}_0(y)$ and $\mathbf{a}_1(y)$ by solving two linear systems, respectively.

3. Theoretical property

In this section, we establish the asymptotic properties of $\widehat{\boldsymbol{\theta}}$, which is constructed with $\widehat{f}_{\mathbf{X}}$. To facilitate the proof, we assume that $\widehat{f}_{\mathbf{X}}$ is independent of the data used to estimate $\boldsymbol{\theta}$ and $g(y)$. Sample splitting can be used to create the independence. Specifically, we randomly split the data into two subsets $\mathcal{D}_1$ and $\mathcal{D}_2$, where $\mathcal{D}_k$ contains N_k observations and n_k out of N_k observations' responses are not missing. For simplicity, assume $n_k/N_K = n/N$, $k = 1, 2$. We use $\mathcal{I}_k$ to denote the indices of observations in the k-th subset. Subset $\mathcal{D}_1$ is used to estimate $f_{\mathbf{X}}(\mathbf{x})$, while $\mathcal{D}_2$ is used to estimate $\boldsymbol{\theta}$ and $g(y)$.

We further need the following regularity conditions.

(C1) Let $N_1 = CN^{\delta}$, where $\delta \in [0.5, 1)$ and C is a positive constant. Then $N_2 = N - CN^{\delta}$.
(C2) The degree of the polynomial interpolation is $m - 1$, where $m \geq 2$.
(C3) Let $L = ch^{-a}$, where c is a positive constant and $0 < a < 2$. The bandwidth $h = o(1)$ satisfies $Nh^{4a+2} \to \infty$, $Nh^{8-4a} \to 0$, $Nh^{4am-4a} \to 0$, and $Nh^{2am} \to 0$.
(C4) Let $A_0(\mathbf{z}) = E[\mathbf{a}_0(y) \exp\{-g(y)\} \mid \mathbf{z}, 1]$, and $A_1(\mathbf{z}) = E[\mathbf{a}_1(y) \exp\{-g(y)\} \mid \mathbf{z}, 1]$, both of which are functionals of $f_{\mathbf{X}}$. Assume that the Fréchet derivative of $A_0(\mathbf{z})$ with respect to $f_{\mathbf{X}}$ and the Fréchet derivative of $A_1(\mathbf{z})$ with respect to $f_{\mathbf{X}}$ are bounded.
(C5) Function $g(y) \in C^{m-1}$ is bounded, and its $(m-1)$th derivative is Lipschitz continuous.

Remark 1. To guarantee that (C1), (C2) and (C3) hold simultaneously, by simple calculations, we need $m \geq 4$, and

$$\frac{1}{m-2} < a < \frac{3}{4}.$$

We set $m = 4$ in implementation, which means that cubic interpolation is employed.

The asymptotic normality of $\widehat{\boldsymbol{\theta}}$ is stated in Theorem 1, and the detailed proof can be found in Appendix A.5.

Theorem 1. *Assume that Conditions* (C1) - (C5) *hold. Let*

$$\mathbf{Q} \equiv E\left\{\frac{\partial \mathbf{S}_{\boldsymbol{\theta},\mathrm{eff}}(\mathbf{X}, RY, R; \boldsymbol{\theta}_0, g)}{\partial \boldsymbol{\theta}_0^{\mathrm{T}}}\right\} = -var\{\mathbf{S}_{\boldsymbol{\theta},\mathrm{eff}}(\mathbf{X}, RY, R; \boldsymbol{\theta}_0, g)\}.$$

Then

$$N^{1/2}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) = -\mathbf{Q}^{-1} N^{-1/2} \sum_{i \in \mathcal{I}_2} \mathbf{S}_{\boldsymbol{\theta},\mathrm{eff}}(\mathbf{x}_i, r_i y_i, r_i; \boldsymbol{\theta}_0, g) + o_p(1).$$

Consequently, $N^{1/2}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \to Normal(\mathbf{0}, \mathbf{V})$ *in distribution when* $N \to \infty$, *where*

$$\mathbf{V} = \mathbf{Q}^{-1} var\{\mathbf{S}_{\boldsymbol{\theta},\mathrm{eff}}(\mathbf{X}, RY, R; \boldsymbol{\theta}_0, g)\}(\mathbf{Q}^{-1})^{\mathrm{T}} = [E\{\mathbf{S}_{\boldsymbol{\theta},\mathrm{eff}}(\mathbf{X}, RY, R; \boldsymbol{\theta}_0, g)^{\otimes 2}\}]^{-1}.$$

Remark 2. Typically, estimating nuisance parameters will alter, often inflate, the variance of the estimator for the parameter of interest. However, in our construction, this is not the case. In other words, even if we knew the true functions $g(y)$ and $f_{\mathbf{X}}(\mathbf{x})$ and used them in the estimating equation (10), the variance of $\widehat{\boldsymbol{\theta}}$ would not be further reduced. Indeed, the asymptotic variance of our $\widehat{\boldsymbol{\theta}}$ achieves the optimal estimation variance bound, and thus, it is semiparametrically efficient.

4. Simulation studies

In this section, we conduct simulation studies to evaluate the finite sample performance of our proposed method, as well as its comparison with some naive method. We consider the following data generating process. We first generate U from a uniform distribution on $(0, 1)$, then given $U = u$, we generate Z from Normal$(u, 0.5^2)$. The conditional distribution of Y given $Z = z$ and $R = 1$ is Normal$(\alpha z, 1)$ with $\alpha = 2$. We then generate R by $\mathrm{pr}(R = 1 \mid y, u) = \mathrm{expit}\{g(y) + h(u; \beta)\}$, where $g(y) = 3\mathrm{expit}(y)$, a commonly used bounded smooth function, and $h(u; \beta) = \beta u$ with $\beta = -1$. We use the rejection sampling method (Gilks and Wild, 1992) to generate the independent and identically distributed data $(\mathbf{x}_i, r_i y_i, r_i)$, $i = 1, \ldots, N$. This results around 20% missingness. We implement the Algorithm 1 in Section 2.3, where the nuisance function is estimated on 12 distinct points, i.e., the dimension of $\boldsymbol{\gamma}$ is 12.

We compare the proposed method to the approach with a purely parametric propensity score model (Little and Rubin, 2002), where $g(y)$ is parameterized as a specified value g_0, and the same $f_{Y|\mathbf{Z},R=1}$ model is adopted. In this naive method, we estimate α, the unknown parameter in $f_{Y|\mathbf{Z},R=1}$, by the ordinary least squares based on the complete cases, and estimate g_0 and β by logistic regression using r_i and u_i, $i = 1, \ldots, N$.

In addition to estimating model parameters, we also estimate $E(Y)$ under different model settings to further illustrate the performance of our method. Under the same data generating process, we generate 10^6 observations and use the average response as the true value of $E(Y)$. We consider two methods to estimate $E(Y)$ corresponding to the two comparison approaches above. Under the proposed approach, we estimate $E(Y)$ using

$$\frac{1}{N}\sum_{i=1}^{N}\left\{r_i y_i + (1 - r_i)\frac{\widehat{E}[Y \exp\{-\widehat{g}(Y)\} \mid z_i, 1]}{\widehat{E}[\exp\{-\widehat{g}(Y)\} \mid z_i, 1]}\right\}.$$

The detailed derivations of the above procedure are given in Appendix A.6. Under the parametric approach, we estimate $E(Y)$ using $N^{-1}\sum_{i=1}^{N} r_i y_i / \pi(u_i; \widehat{\beta}, \widehat{g}_0)$.

In our analysis, we consider both $N = 1000$ and $N = 2000$. Based on 1000 simulation replications, we summarize the estimation and inference results of parameters of interest in Table 1, the estimation results of $g(y)$ or g_0 in Fig. 1, and the estimation results of $E(Y)$ in Table 2. In Table 1, we can see that, the estimation and inference results for β using our proposed method are satisfactory. Specifically, the biases of $\widehat{\beta}$ are much smaller than those using the parametric method, the means of the estimated standard deviations closely approximate the empirical ones, and the empirical coverages of the estimated 95% confidence intervals are close to the nominal level. To the contrary, the naive parametric method produces tremendous estimation biases, especially for $\widehat{\beta}$, so that the coverage probability is not reliable at all. In Fig. 1, our estimates for $g(y)$ also match with the true curve very well. With a larger sample size $N = 2000$, the 95% confidence band of $\widehat{g}(y)$ becomes narrower. Finally, Table 2 clearly demonstrates that the estimate of $E(Y)$ using the proposed approach is asymptotically unbiased; however, the parametric approach, which estimates the function $g(y)$ as a constant, is heavily biased. Hence, the parametric method should be used with great caution.

Table 1
Estimation and inference results of the parameters of interest.

	Parametric Method				Proposed Method			
	N=1000		N=2000		N=1000		N=2000	
	α	β	α	β	α	β	α	β
Bias	-0.0006	1.0749	-0.0008	1.0639	-0.0005	0.0368	-0.0008	0.0138
SD	0.0449	0.2646	0.0305	0.1963	0.0449	0.3148	0.0306	0.2278
$\widehat{\text{SD}}$	0.0399	0.2674	0.0282	0.1889	0.0452	0.3114	0.0319	0.2195
cvg	91.2%	1.9%	93.1%	0.0%	94.9%	95.2%	96.1%	94.4%

In Table 1, "SD" denotes the empirical standard deviation of 1000 estimates; "$\widehat{\text{SD}}$" denotes the mean of 1000 estimated asymptotic standard deviations; "cvg" denotes empirical coverage of the estimated 95% CI.

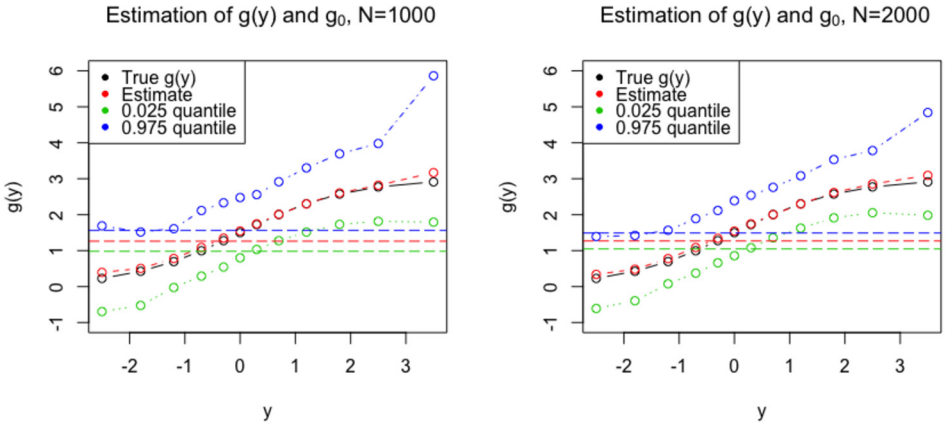

Fig. 1. The three dotted curves (red, green and blue) represent the estimations of function $g(y)$ (estimate, 2.5% quantile and 97.5% quantile) using the proposed approach, respectively. The three dashed horizontal lines (red, green and blue) represent the estimations of constant g_0 (estimate, 2.5% quantile and 97.5% quantile) using the parametric approach, respectively. (For interpretation of the colors in the figure(s), the reader is referred to the web version of this article.)

Table 2
Comparison of different $E(Y)$ estimates (true value equals 0.8879).

	Parametric Method		Proposed Method	
	N=1000	N=2000	N=1000	N=2000
Bias	0.2190	0.2192	0.0051	0.0028
SD	0.0529	0.0360	0.0513	0.0360

In Table 2, "SD" denotes the empirical standard deviation of 1000 estimates.

5. Real data application

In this section, we analyze a data set from a publicly available electronic health records (EHR) database (Johnson et al., 2016), the Medical Information Mart for Intensive Care III (MIMIC-III). This database comprises de-identified health-related data associated with intensive care unit patients with rich information including demographics, vital signs, laboratory test, medications, among others.

When we analyze this database, we encounter different types of missing values for laboratory test biomarkers. As we explained in Section 1, it is very plausible that these missing data are nonignorable (Hu et al., 2017; Li et al., 2018). Therefore, we would like to investigate what the corresponding propensity score models look like and what are the effective factors in those models. In this application, we focus on the setting that the response Y is the albumin level in the blood sample, a highly indicative biomarker associated with different types of diseases (Phillips et al., 1989).

For illustrative purpose and to control the effect of race and marital status, our analysis concentrates on a subgroup of the whole data set, where the subjects are white and divorced, with sample size $N = 1476$ in which 537 samples (around 36%) have missing values in the albumin level. Our data set consists of six fully observed covariates. One of them, the calcium level in the blood sample, has been shown in the biomedical literature to have high correlation with the albumin level (Katz and Klotz, 1953; Butler et al., 1984; Hossain et al., 2015); therefore, we adopt the calcium level as the shadow variable Z following the literature (Zhao and Chen, 2020). Five other variables comprise the vector $\mathbf{U}$, which are age, gender,

Table 3
Estimation and inference results of parameters of interest in the parametric method and in the proposed method.

	Parametric Method			Proposed Method		
	Estimate	$\widehat{\text{SD}}$	p-value	Estimate	$\widehat{\text{SD}}$	p-value
α_1	-2.1801	0.2710	< 0.0001	-2.1492	0.2966	< 0.0001
α_2	0.6394	0.0321	< 0.0001	0.6348	0.0349	< 0.0001
α_3	0.5378	0.0124	< 0.0001	0.5739	0.0146	< 0.0001
β_1 (age)	-0.0019	0.0017	0.2739	-0.0005	0.0017	0.7564
β_2 (gender)	0.2347	0.1107	0.0339	0.2168	0.1107	0.0503
β_3 (systolic blood pressure)	0.0039	0.0036	0.2778	0.0091	0.0037	0.0145
β_4 (body temperature)	0.3149	0.1166	0.0069	0.3928	0.0471	< 0.0001
β_5 (SpO2)	0.1278	0.0355	0.0003	0.0855	0.0174	< 0.0001

In Table 3, "$\widehat{\text{SD}}$" denotes the estimated asymptotic standard deviation.

systolic blood pressure, body temperature, and peripheral capillary oxygen saturation (SpO2). The assumption in (2) that Y is conditionally independent of $\mathbf{U}$ given $\mathbf{Z}$ and $R = 1$, can be empirically validated. In our analysis, there is no statistically significant evidence that we can reject this null hypothesis. The kernel-based conditional independence test (Heinze-Deml et al., 2018) gives us p-value as 0.5670.

We use a linear regression model with homoscedastic normal regression error as $f_{Y|Z,R=1}$, i.e.,

$$Y_i = \alpha_1 + \alpha_2 Z_i + \epsilon_i,$$

where ϵ_i is i.i.d from Normal$(0, \alpha_3^2)$, and consider the following nonignorable propensity score model

$$\pi(y, \mathbf{u}; \boldsymbol{\beta}, g) = \text{expit}\{g(y) + \boldsymbol{\beta}^{\mathrm{T}}\mathbf{u}\},$$

where function $g(y)$ is unknown and $\boldsymbol{\beta} \in \mathbb{R}^5$.

Similar to Section 4, we compare our approach to a parametric propensity score method, where the same $f_{Y|\mathbf{Z},R=1}$ model is used but the propensity is assumed to be

$$\pi(\mathbf{u}; \boldsymbol{\beta}, g_0) = \text{expit}(g_0 + \boldsymbol{\beta}^{\mathrm{T}}\mathbf{u}).$$

We estimate $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \alpha_3)^{\mathrm{T}}$ by the ordinary least squares method based on the complete cases, and estimate g_0 and $\boldsymbol{\beta}$ by logistic regression using r_i and $\mathbf{u}_i$, $i = 1, \ldots, 1476$.

The estimation and inference results of parameters of interest in both methods are summarized in Table 3. We can see that the results for $\widehat{\boldsymbol{\alpha}}$, the unknown parameter in the model $f_{Y|\mathbf{Z},R=1}$, are quite close to each other. This is not surprising. In contrast, the results for $\boldsymbol{\beta}$ are generally different. The variable age (β_1) is insignificant, whereas the variables body temperature (β_4) and SpO2 (β_5) are both significant, based on either of the two methods. The variable gender (β_2) is significant based on the parametric method but only marginally significant based on the nonparametric method. More importantly, the variable systolic blood pressure (β_3) is insignificant according to the parametric method but becomes statistically significant based on the nonparametric method.

Finally, using the proposed method, we provide the estimated curve of $g(y)$ as well as its 95% bootstrapped confidence band in Fig. 2. Although $\widehat{g}(y)$ does not change drastically over the range of albumin level (from 3g/dL to 5g/dL), it is clear that $\widehat{g}(y)$ decreases when the albumin level changes from 4.3g/dL to 5g/dL. This demonstrates that naively assuming $g(y)$ is a constant hence the propensity score model does not depend on the albumin level is unrealistic and should be avoided in applications. Instead, our proposed nonparametric method can be very useful to uncover the unknown and possibly nonlinear dependence of the propensity score model to the albumin level, and should be recommended for use in practice.

6. Discussion

In this paper, we propose a new nonignorable propensity score model where the relationship between the missingness indicator and the partially observed response is totally unspecified and estimated nonparametrically. This new propensity is flexible in modeling the dependence of the missingness indicator on the partially observed response, and avoids the use of the multivariate kernel estimation which suffers from the curse of dimensionality. By employing a semiparametric treatment, our estimator for the parameters of interest is not only asymptotically unbiased but also semiparametrically efficient. In terms of estimating the nuisance function, we devise a new likelihood-based consistent estimator when the input variable of the nuisance function is subject to missing data.

The proposed method in this paper can be applied to any regression settings where the response variable is subject to nonignorable missing data, the covariates are fully observed, and the choice of the shadow variable is appropriate. Our simulation studies and a real data application to modeling the albumin level in the blood sample have demonstrated that the proposed method could be very useful to uncover the unknown and possibly nonlinear dependence of the missingness

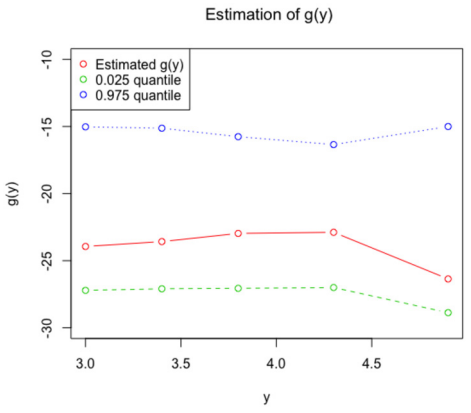

Fig. 2. The estimation of the unknown function $g(y)$ with 95% bootstrapped confidence band.

indicator on the partially observed response in the propensity score model. Our programming code has been provided online https://github.com/MengyanLi1992/Efficient-Estimation-in-a-Partially-Specified-Nonignorable-Propensity-Score-Model, and is ready to be used in other similar applications.

Due to an unspecified function $g(y)$ in the propensity score model, model identifiability is challenging to achieve. Motivated by the real data application in Section 5 where a shadow variable has been identified in the literature (Zhao and Chen, 2020), we establish our model identifiability based on the existence of a shadow variable and a testable conditional independence assumption about the model $f_{Y|\mathbf{X},R=1}$. The shadow variable assumption is standard in the literature of nonignorable missing data and is commonly used in practice (Shao and Zhao, 2013; Kott, 2014; Wang et al., 2014; Zhao and Shao, 2015; Zhao and Ma, 2018, 2021). Empirically, we identify the shadow variable based on the domain knowledge and then check the conditional independence assumption using the observed data. In some applications where covariates are from different sources, there may exist a natural determination of $\mathbf{U}$ and $\mathbf{Z}$. In the simulation studies of this paper, we consider the "ideal setting" where all the assumptions of our method are fully met. In more extended numerical studies, it is worth studying the performance of our method when the assumptions are not fully met. Additionally, to use the proposed method in other applications, it might also be helpful, if the choice of the shadow variable is not certain from the domain knowledge, to consider different choices of shadow variable and regard as a sensitivity analysis.

Finally, as a first attempt to study the proposed semiparametric propensity score model where the dependence on the partially observed variable is nonparametric, many directions warrant further research. For example, one might like to consider other regression families instead of the simple exponential family, or more sophisticated structure in the propensity score model instead of the additive structure considered in this paper. Multivariate response or high dimensional covariates can also be of interest to explore. Additionally, if the choice of the instrumental variable (Tchetgen Tchetgen and Wirth, 2017), instead of the shadow variable, is clear from the domain knowledge, it will be interesting and worthwhile to study the model identifiability with the instrumental variable.

Acknowledgement

This research was partially supported by the National Science Foundation (2122074) of the United States.

Appendix A

A.1. Proof of Lemma 1

Proof. Since $\mathbf{X}$ is fully observed, then $f_{\mathbf{X}}(\mathbf{x})$ is identifiable. Since when $R=1$, both $\mathbf{X}$ and Y are fully observed, then $\boldsymbol{\alpha}$ is identifiable. We need to show that $\boldsymbol{\beta}$ and $g(y)$ are identifiable. Note that $w(\mathbf{x})$ is identifiable by (3), and

$$w(\mathbf{x})^{-1} = 1 + \exp\{-h(\mathbf{u};\boldsymbol{\beta})\}E[\exp\{-g(Y)\} \mid \mathbf{z}, 1].$$

We only need to show that for any $\mathbf{x}$, if

$$\exp\{-h(\mathbf{u};\boldsymbol{\beta})\}E[\exp\{-g(Y)\} \mid \mathbf{z}, 1] = \exp\{-h(\mathbf{u};\widetilde{\boldsymbol{\beta}})\}E[\exp\{-\widetilde{g}(Y)\} \mid \mathbf{z}, 1],$$

then $\boldsymbol{\beta} = \widetilde{\boldsymbol{\beta}}$ and $g(\cdot) = \widetilde{g}(\cdot)$. There exists a constant c such that

$$\frac{\exp\{-h(\mathbf{u};\boldsymbol{\beta})\}}{\exp\{-h(\mathbf{u};\widetilde{\boldsymbol{\beta}})\}} = \frac{E[\exp\{-\widetilde{g}(Y)\} \mid \mathbf{z}, 1]}{E[\exp\{-g(Y)\} \mid \mathbf{z}, 1]} = c, \quad \forall \mathbf{u}, \mathbf{z},$$

because the left hand side is a function of $\mathbf{u}$ only while the right hand side is a function of $\mathbf{z}$ only. Then $h(\mathbf{u};\boldsymbol{\beta}) = h(\mathbf{u};\widetilde{\boldsymbol{\beta}}) - \log(c)$ for all $\mathbf{u}$, and $\widetilde{g}(y) = g(y) - \log(c)$ for all y due to the invertibility of the Laplace transform. Taking into account the requirement that $g(0) = 0$, we obtain $c = 1$. Hence, $\boldsymbol{\beta}$ and g are also identifiable. □

A.2. Derivation of Λ_g

Consider a parametric sub-model $g(y) = g(y;\boldsymbol{\zeta})$. Then the score vector w.r.t $\boldsymbol{\zeta}$ is given as $\mathbf{S}_{\boldsymbol{\zeta}}(\boldsymbol{\theta},\boldsymbol{\zeta};\mathbf{x},ry,r) \equiv \partial\log\{\mathbf{f}_{X,RY,R}(\mathbf{x},ry,r;\boldsymbol{\theta},\boldsymbol{\zeta})\}/\partial\boldsymbol{\zeta}$, where $\mathbf{f}_{X,RY,R}(\mathbf{x},ry,r;\boldsymbol{\theta},\boldsymbol{\zeta})$ is the likelihood function in (3) with $g(y)$ replaced by $g(y;\boldsymbol{\zeta})$. Since

$$\frac{\partial w(\mathbf{x})}{\partial\boldsymbol{\zeta}} = w^2(\mathbf{x})E\left[\frac{\partial g(Y;\boldsymbol{\zeta})}{\partial\boldsymbol{\zeta}}\{\pi^{-1}(Y,\mathbf{u};\boldsymbol{\beta},\boldsymbol{\zeta}) - 1\} \mid \mathbf{z}, 1\right],$$

then

$$\mathbf{S}_{\boldsymbol{\zeta}}(\mathbf{x},ry,r;\boldsymbol{\theta},\boldsymbol{\zeta}) = E\left[\frac{\partial g(Y;\boldsymbol{\zeta})}{\partial\boldsymbol{\zeta}}\{\pi^{-1}(Y,\mathbf{u};\boldsymbol{\beta},\boldsymbol{\zeta}) - 1\} \mid \mathbf{z}, 1\right]\frac{w(\mathbf{x})\{r - w(\mathbf{x})\}}{1 - w(\mathbf{x})},$$

where $\partial g(y;\boldsymbol{\zeta})/\partial\boldsymbol{\zeta}$ can be any function of y. Note that Λ_g is the mean squared closure of nuisance tangent spaces of parametric sub-models spanned by the corresponding nuisance score vectors. Then

$$\Lambda_g = \left(E[\mathbf{a}(Y)\{\pi^{-1}(Y,\mathbf{u};\boldsymbol{\beta},g) - 1\} \mid \mathbf{z}, 1]\frac{w(\mathbf{x})\{r - w(\mathbf{x})\}}{1 - w(\mathbf{x})} : \mathbf{a}(y) \in \mathbb{R}^p\right).$$

A.3. Derivation of score functions

The log-likelihood function is

$$l(\boldsymbol{\alpha},\boldsymbol{\beta},g;\mathbf{x},ry,r) = \log\{f_X(\mathbf{x})\} + r\log\{f_{Y|\mathbf{Z},R=1}(y,\mathbf{z};\boldsymbol{\alpha})\} + r\log\{w(\mathbf{x})\} + (1-r)\log\{1 - w(\mathbf{x})\}.$$

Since

$$\begin{aligned}
\frac{\partial w(\mathbf{x})}{\partial\boldsymbol{\beta}} &= w^2(\mathbf{x})\int f_{Y|\mathbf{Z},R=1}(t,\mathbf{z};\boldsymbol{\alpha})\exp\{-g(t) - h(\mathbf{u};\boldsymbol{\beta})\}\mathbf{h}'_{\boldsymbol{\beta}}(\mathbf{u};\boldsymbol{\beta})dt \\
&= w^2(\mathbf{x})\int f_{Y|\mathbf{Z},R=1}(t,\mathbf{z};\boldsymbol{\alpha})\{\pi^{-1}(t,\mathbf{u};\boldsymbol{\beta},g) - 1\}\mathbf{h}'_{\boldsymbol{\beta}}(\mathbf{u};\boldsymbol{\beta})dt \\
&= w(\mathbf{x})\{1 - w(\mathbf{x})\}\mathbf{h}'_{\boldsymbol{\beta}}(\mathbf{u};\boldsymbol{\beta}), \\
\frac{\partial w(\mathbf{x})}{\partial\boldsymbol{\alpha}} &= -w^2(\mathbf{x})\frac{\partial}{\partial\boldsymbol{\alpha}}\int f_{Y|\mathbf{Z},R=1}(t,\mathbf{z};\boldsymbol{\alpha})[1 + \exp\{-g(t) - h(\mathbf{u};\boldsymbol{\beta})\}]dt \\
&= -w^2(\mathbf{x})\exp\{-h(\mathbf{u};\boldsymbol{\beta})\}E\left[\mathbf{S}(Y,\mathbf{z};\boldsymbol{\alpha})\exp\{-g(Y)\} \mid \mathbf{z}, 1\right],
\end{aligned}$$

then

$$\begin{aligned}
\mathbf{S}_{\boldsymbol{\beta}}(\mathbf{x},ry,r;\boldsymbol{\beta},\boldsymbol{\alpha},g) &= \frac{\partial w(\mathbf{x})}{\partial\boldsymbol{\beta}}\frac{r - w(\mathbf{x})}{w(\mathbf{x})\{1 - w(\mathbf{x})\}} \\
&= \{r - w(\mathbf{x})\}\mathbf{h}'_{\boldsymbol{\beta}}(\mathbf{u}),
\end{aligned}$$

and

$$\begin{aligned}
&\mathbf{S}_{\boldsymbol{\alpha}}(\mathbf{x},ry,r;\boldsymbol{\beta},\boldsymbol{\alpha},g) \\
&= r\mathbf{S}(y,\mathbf{z};\boldsymbol{\alpha}) + \frac{\partial w(\mathbf{x})}{\partial\boldsymbol{\alpha}}\frac{r - w(\mathbf{x})}{w(\mathbf{x})\{1 - w(\mathbf{x})\}} \\
&= r\mathbf{S}(y,\mathbf{z};\boldsymbol{\alpha}) - E\left[\mathbf{S}(Y,\mathbf{z};\boldsymbol{\alpha})\exp\{-g(Y)\} \mid \mathbf{z}, 1\right]\frac{w(\mathbf{x})\{r - w(\mathbf{x})\}}{1 - w(\mathbf{x})}\exp\{-h(\mathbf{u};\boldsymbol{\beta})\} \\
&= r\mathbf{S}(y,\mathbf{z};\boldsymbol{\alpha}) - E\left[\mathbf{S}(Y,\mathbf{z};\boldsymbol{\alpha})\{\pi^{-1}(Y,\mathbf{u};\boldsymbol{\beta},g) - 1\} \mid \mathbf{z}, 1\right]\frac{w(\mathbf{x})\{r - w(\mathbf{x})\}}{1 - w(\mathbf{x})} \\
&= r\mathbf{S}(y,\mathbf{z};\boldsymbol{\alpha}) - E\left[\mathbf{S}(Y,\mathbf{z};\boldsymbol{\alpha})\pi^{-1}(Y,\mathbf{u};\boldsymbol{\beta},g) \mid \mathbf{z}, 1\right]\frac{w(\mathbf{x})\{r - w(\mathbf{x})\}}{1 - w(\mathbf{x})} \\
&= r\mathbf{S}(y,\mathbf{z};\boldsymbol{\alpha}) - E\left\{\mathbf{S}(Y,\mathbf{z};\boldsymbol{\alpha}) \mid \mathbf{x}\right\}\frac{r - w(\mathbf{x})}{1 - w(\mathbf{x})}.
\end{aligned}$$

A.4. Efficient score

Proof. Recall that $\mathbf{S}_{\boldsymbol{\theta}}$ is in $\Lambda_g \oplus \Lambda^{\perp}$. It is obvious that $\mathbf{S}_{\boldsymbol{\theta}} - \mathbf{S}_{\boldsymbol{\theta},\text{eff}} \in \Lambda_g$. We only need to verify that $\mathbf{S}_{\boldsymbol{\theta},\text{eff}}$ is in $\Lambda^{\perp}$. With $\mathbf{a}_0(\mathbf{x})$ and $\mathbf{a}_1(\mathbf{x})$ satisfying (5) and (7), respectively, it is easy to show that

$$E\left(\mathbf{S}_{\boldsymbol{\theta},\text{eff}}(\mathbf{X}, RY, Y; \boldsymbol{\theta}, g)^{\mathrm{T}} E[\mathbf{a}(Y)\{1-\pi(Y,\mathbf{U};\boldsymbol{\beta},g)\} \mid \mathbf{X}]\frac{R-w(\mathbf{X})}{1-w(\mathbf{X})}\right) = 0,\ \forall \mathbf{a}(Y) \in \mathbb{R}^p.$$

In other words, $\mathbf{S}_{\boldsymbol{\theta},\text{eff}}$ is orthogonal to Λ_g. Hence, $\mathbf{S}_{\boldsymbol{\theta},\text{eff}}$ is in $\Lambda^{\perp}$ and is the projection of $\mathbf{S}_{\boldsymbol{\theta}}$ onto $\Lambda^{\perp}$. □

A.5. Proof of Theorem 1

Proof. Let $\mathbf{o}_i = (\mathbf{x}_i, r_i y_i, r_i)$. Define $f_{\mathbf{X},RY,R}(\mathbf{o}_i; \boldsymbol{\theta}_0, \boldsymbol{\gamma}_0)$ as the likelihood of a parametric submodel with true parameters $\boldsymbol{\theta}_0$ and $\boldsymbol{\gamma}_0$. Note that any parametric submodel contains the true model, which implies that

$$f_{\mathbf{X},RY,R}(\mathbf{o}_i; \boldsymbol{\theta}_0, \boldsymbol{\gamma}_0) = f_{\mathbf{X},RY,R}(\mathbf{o}_i; \boldsymbol{\theta}_0, g).$$

For term $\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}$, we expand the estimating equation (10) as a function for $\boldsymbol{\theta}$ and $\boldsymbol{\gamma}$ about the truth $\boldsymbol{\theta}_0$ and $\boldsymbol{\gamma}_0$ to obtain

$$\begin{aligned}
\mathbf{0} &= N_2^{-1/2}\sum_{i\in\mathcal{I}_2}\widehat{\mathbf{S}}_{\boldsymbol{\theta},\text{eff}}[\mathbf{o}_i; \widehat{\boldsymbol{\theta}}, \widehat{g}\{\cdot; \widehat{\boldsymbol{\gamma}}(\widehat{\boldsymbol{\theta}})\}] \\
&= N_2^{-1/2}\sum_{i\in\mathcal{I}_2}\widehat{\mathbf{S}}_{\boldsymbol{\theta},\text{eff}}\{\mathbf{o}_i; \boldsymbol{\theta}_0, \widehat{g}(\cdot; \boldsymbol{\gamma}_0)\} + N_2^{-1}\sum_{i\in\mathcal{I}_2}\frac{\partial\widehat{\mathbf{S}}_{\boldsymbol{\theta},\text{eff}}\{\mathbf{o}_i; \widetilde{\boldsymbol{\theta}}, \widehat{g}(\cdot; \widetilde{\boldsymbol{\gamma}})\}}{\partial\widetilde{\boldsymbol{\theta}}^{\mathrm{T}}}N_2^{1/2}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \\
&\quad + N_2^{-1}\sum_{i\in\mathcal{I}_2}\frac{\partial\widehat{\mathbf{S}}_{\boldsymbol{\theta},\text{eff}}\{\mathbf{o}_i; \widetilde{\boldsymbol{\theta}}, \widehat{g}(\cdot; \widetilde{\boldsymbol{\gamma}})\}}{\partial\widetilde{\boldsymbol{\gamma}}^{\mathrm{T}}}N_2^{1/2}\{\widehat{\boldsymbol{\gamma}}(\widehat{\boldsymbol{\theta}}) - \boldsymbol{\gamma}_0\},
\end{aligned}$$

where $\widetilde{\boldsymbol{\theta}}$ is on the line connecting $\widehat{\boldsymbol{\theta}}$ and $\boldsymbol{\theta}_0$. We will show that under regularity conditions (C1), (C2), (C3) and (C4)

$$N_2^{-1}\sum_{i\in\mathcal{I}_2}\frac{\partial\widehat{\mathbf{S}}_{\boldsymbol{\theta},\text{eff}}\{\mathbf{o}_i; \widetilde{\boldsymbol{\theta}}, \widehat{g}(\cdot; \widetilde{\boldsymbol{\gamma}})\}}{\partial\widetilde{\boldsymbol{\gamma}}^{\mathrm{T}}}N_2^{1/2}\{\widehat{\boldsymbol{\gamma}}(\widehat{\boldsymbol{\theta}}) - \boldsymbol{\gamma}_0\} = o_p(1),$$

element-wise.

First, for $j = 1, \ldots, p$ and $k = 1, \ldots, L$ we have

$$\begin{aligned}
&N_2^{-1}\sum_{i\in\mathcal{I}_2}\left[\frac{\partial\widehat{\mathbf{S}}_{\boldsymbol{\theta},\text{eff}}\{\mathbf{o}_i; \widetilde{\boldsymbol{\theta}}, \widehat{g}(\cdot; \widetilde{\boldsymbol{\gamma}})\}}{\partial\widetilde{\boldsymbol{\gamma}}^{\mathrm{T}}}\right]_{j,k} \\
&= N_2^{-1}\sum_{i\in\mathcal{I}_2}\left[\frac{\partial\widehat{\mathbf{S}}_{\boldsymbol{\theta},\text{eff}}\{\mathbf{o}_i; \boldsymbol{\theta}_0, g(\cdot; \boldsymbol{\gamma}_0)\}}{\partial\boldsymbol{\gamma}_0^{\mathrm{T}}}\right]_{j,k} + O_p\{\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\|_1 + \|\widehat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}_0\|_1 + h_L^m\} \\
&= E\left(\left[\frac{\partial\widehat{\mathbf{S}}_{\boldsymbol{\theta},\text{eff}}\{\mathbf{O}_i; \boldsymbol{\theta}_0, g(\cdot; \boldsymbol{\gamma}_0)\}}{\partial\boldsymbol{\gamma}_0^{\mathrm{T}}} \mid \mathcal{D}_1\right]_{j,k}\right) + O_p\{\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\|_1 + \|\widehat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}_0\|_1 + h_L^m + N_2^{-1/2}\}. \qquad \text{(A.1)}
\end{aligned}$$

Note that in $\widehat{\mathbf{S}}_{\boldsymbol{\theta},\text{eff}}\{\mathbf{o}_i; \boldsymbol{\theta}_0, g(\cdot; \boldsymbol{\gamma}_0)\}$, $\widehat{\mathbf{a}}_0(y)$ and $\widehat{\mathbf{a}}_1(y)$ are estimated at the true $\boldsymbol{\theta}_0$ and $g(\cdot)$. Then $\widehat{\mathbf{a}}_0(\cdot)$ and $\widehat{\mathbf{a}}_1(\cdot)$ are deterministic conditional on $\mathcal{D}_1$. The first equality holds because

$$\|E\left[\frac{\partial}{\partial\boldsymbol{\theta}_0}\frac{\partial\widehat{\mathbf{S}}_{\boldsymbol{\theta},\text{eff}}\{\mathbf{O}_i; \boldsymbol{\theta}_0, g(\cdot; \boldsymbol{\gamma}_0)\}}{\partial\boldsymbol{\gamma}_0^{\mathrm{T}}} \mid \mathcal{D}_1\right]\|_{\max} \text{ and } \|E\left[\frac{\partial}{\partial\boldsymbol{\gamma}_0}\frac{\partial\widehat{\mathbf{S}}_{\boldsymbol{\theta},\text{eff}}\{\mathbf{O}_i; \boldsymbol{\theta}_0, g(\cdot; \boldsymbol{\gamma}_0)\}}{\partial\boldsymbol{\gamma}_0^{\mathrm{T}}} \mid \mathcal{D}_1\right]\|_{\max}$$

are bounded, and

$$\begin{aligned}
&\sum_{l=1}^{L}N_2^{-1}\sum_{i\in\mathcal{I}_2}\frac{\partial}{\partial\gamma_{0,l}}\left[\frac{\partial\widehat{\mathbf{S}}_{\boldsymbol{\theta},\text{eff}}\{\mathbf{o}_i; \boldsymbol{\theta}_0, g(\cdot; \boldsymbol{\gamma}_0)\}}{\partial\boldsymbol{\gamma}_0^{\mathrm{T}}}\right]_{j,k}(\widehat{\gamma}_l - \gamma_{0,l}) \qquad \text{(A.2)} \\
&= \sum_{l=1}^{L}\left\{E\left(\frac{\partial}{\partial\gamma_{0,l}}\left[\frac{\partial\widehat{\mathbf{S}}_{\boldsymbol{\theta},\text{eff}}\{\mathbf{O}_i; \boldsymbol{\theta}_0, g(\cdot; \boldsymbol{\gamma}_0)\}}{\partial\boldsymbol{\gamma}_0^{\mathrm{T}}} \mid \mathcal{D}_1\right]_{j,k}\right) + O_p(N_2^{-1/2})\right\}(\widehat{\gamma}_l - \gamma_{0,l}) \\
&= O_p(\|\widehat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}_0\|_1).
\end{aligned}$$

Further, since the dimension of $\widehat{\boldsymbol{\theta}}$ is fixed, then any norm of $\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0$ is of the same order.

We can also show that

$$
\begin{aligned}
& E\left[\frac{\partial \widehat{\mathbf{S}}_{\boldsymbol{\theta},\mathrm{eff}}\{\mathbf{O}_i;\boldsymbol{\theta}_0,g(\cdot;\boldsymbol{\gamma}_0)\}}{\partial \boldsymbol{\gamma}_0^{\mathrm{T}}} \mid \mathcal{D}_1\right] \\
&= \int \frac{\partial \widehat{\mathbf{S}}_{\boldsymbol{\theta},\mathrm{eff}}\{\mathbf{o}_i;\boldsymbol{\theta}_0,g(\cdot;\boldsymbol{\gamma}_0)\}}{\partial \boldsymbol{\gamma}_0^{\mathrm{T}}} f_{\mathbf{X},R,RY}(\mathbf{o}_i;\boldsymbol{\theta}_0,g)d\mathbf{o}_i \\
&= \int \frac{\partial \widehat{\mathbf{S}}_{\boldsymbol{\theta},\mathrm{eff}}\{\mathbf{o}_i;\boldsymbol{\theta}_0,g(\cdot;\boldsymbol{\gamma}_0)\}}{\partial \boldsymbol{\gamma}_0^{\mathrm{T}}} f_{\mathbf{X},R,RY}(\mathbf{o}_i;\boldsymbol{\theta}_0,\boldsymbol{\gamma}_0)d\mathbf{o}_i \\
&= \frac{\partial}{\partial \boldsymbol{\gamma}_0^{\mathrm{T}}}\int \widehat{\mathbf{S}}_{\boldsymbol{\theta},\mathrm{eff}}\{\mathbf{o}_i;\boldsymbol{\theta}_0,g(\cdot;\boldsymbol{\gamma}_0)\} f_{\mathbf{X},R,RY}(\mathbf{o}_i;\boldsymbol{\theta}_0,\boldsymbol{\gamma}_0)d\mathbf{o}_i \\
&\quad - \int \widehat{\mathbf{S}}_{\boldsymbol{\theta},\mathrm{eff}}\{\mathbf{o}_i;\boldsymbol{\theta}_0,g(\cdot;\boldsymbol{\gamma}_0)\}\frac{\partial f_{\mathbf{X},R,RY}(\mathbf{o}_i;\boldsymbol{\theta}_0,\boldsymbol{\gamma}_0)}{\partial \boldsymbol{\gamma}_0^{\mathrm{T}}}d\mathbf{o}_i \\
&= -\int \widehat{\mathbf{S}}_{\boldsymbol{\theta},\mathrm{eff}}\{\mathbf{o}_i;\boldsymbol{\theta}_0,g(\cdot;\boldsymbol{\gamma}_0)\}\frac{\partial f_{\mathbf{X},R,RY}(\mathbf{o}_i;\boldsymbol{\theta}_0,\boldsymbol{\gamma}_0)}{\partial \boldsymbol{\gamma}_0^{\mathrm{T}}}d\mathbf{o}_i \\
&= -E\left[\widehat{\mathbf{S}}_{\boldsymbol{\theta},\mathrm{eff}}\{\mathbf{O}_i;\boldsymbol{\theta}_0,g(\cdot;\boldsymbol{\gamma}_0)\}\frac{\partial \log\{f_{\mathbf{X},R,RY}(\mathbf{O}_i;\boldsymbol{\theta}_0,\boldsymbol{\gamma}_0)\}}{\partial \boldsymbol{\gamma}_0^{\mathrm{T}}} \mid \mathcal{D}_1\right] \\
&= -E\left[\mathbf{S}_{\boldsymbol{\theta},\mathrm{eff}}\{\mathbf{O}_i;\boldsymbol{\theta}_0,g(\cdot;\boldsymbol{\gamma}_0)\}\frac{\partial \log\{f_{\mathbf{X},R,RY}(\mathbf{O}_i;\boldsymbol{\theta}_0,\boldsymbol{\gamma}_0)\}}{\partial \boldsymbol{\gamma}_0^{\mathrm{T}}}\right] \\
&\quad -E\left(\left[\widehat{\mathbf{S}}_{\boldsymbol{\theta},\mathrm{eff}}\{\mathbf{O}_i;\boldsymbol{\theta}_0,g(\cdot;\boldsymbol{\gamma}_0)\} - \mathbf{S}_{\boldsymbol{\theta},\mathrm{eff}}\{\mathbf{O}_i;\boldsymbol{\theta}_0,g(\cdot;\boldsymbol{\gamma}_0)\}\right]\frac{\partial \log\{f_{\mathbf{X},R,RY}(\mathbf{O}_i;\boldsymbol{\theta}_0,\boldsymbol{\gamma}_0)\}}{\partial \boldsymbol{\gamma}_0^{\mathrm{T}}} \mid \mathcal{D}_1\right) \\
&= -E\left(\left[\widehat{\mathbf{S}}_{\boldsymbol{\theta},\mathrm{eff}}\{\mathbf{O}_i;\boldsymbol{\theta}_0,g(\cdot;\boldsymbol{\gamma}_0);\} - \mathbf{S}_{\boldsymbol{\theta},\mathrm{eff}}\{\mathbf{O}_i;\boldsymbol{\theta}_0,g(\cdot;\boldsymbol{\gamma}_0)\}\right]\frac{\partial \log\{f_{\mathbf{X},R,RY}(\mathbf{O}_i;\boldsymbol{\theta}_0,\boldsymbol{\gamma}_0)\}}{\partial \boldsymbol{\gamma}_0^{\mathrm{T}}} \mid \mathcal{D}_1\right) \\
&= O_p(N_1^{-1/2}),
\end{aligned}
$$

element-wise. Due to $f_{\mathbf{X}}$-robustness, we have

$$\int \widehat{\mathbf{S}}_{\boldsymbol{\theta},\mathrm{eff}}[\mathbf{o}_i;\boldsymbol{\theta},g\{\cdot;\boldsymbol{\gamma}(\boldsymbol{\theta})\}]f_{\mathbf{X},R,RY}\{\mathbf{o}_i;\boldsymbol{\theta},\boldsymbol{\gamma}(\boldsymbol{\theta})\}d\mathbf{o}_i = \mathbf{0}$$

for any parameters $\boldsymbol{\theta}$ and $\boldsymbol{\gamma}(\boldsymbol{\theta})$. So the fourth equality holds. The seventh equality holds because $\mathbf{S}_{\boldsymbol{\theta},\mathrm{eff}}$ is orthogonal to the nuisance tangent space Λ_g and the nuisance score for $\boldsymbol{\gamma}$ of any parametric submodel is in Λ_g. For the last equality, we have

$$
\begin{aligned}
& E\left[\{\widehat{\mathbf{S}}_{\boldsymbol{\beta},\mathrm{eff}}(\mathbf{O}_i;\boldsymbol{\theta}_0,g) - \mathbf{S}_{\boldsymbol{\beta},\mathrm{eff}}(\mathbf{O}_i;\boldsymbol{\theta}_0,g)\}\frac{\partial \log\{f_{\mathbf{X},R,RY}(\mathbf{O}_i;\boldsymbol{\theta}_0,\boldsymbol{\gamma}_0)\}}{\partial \boldsymbol{\gamma}_0^{\mathrm{T}}} \mid \mathcal{D}_1\right] \\
&= E\left(\frac{w^3(\mathbf{X}_i)}{1-w(\mathbf{X}_i)}\exp\{-2h(\mathbf{U}_i;\boldsymbol{\beta}_0)\}E\left[\exp\{-g(Y;\boldsymbol{\gamma}_0)\}\frac{\partial g(Y;\boldsymbol{\gamma}_0)}{\partial \boldsymbol{\gamma}_0} \mid \mathbf{Z}_i,1\right]\{\widehat{\mathbf{A}}_1(\mathbf{Z}_i) - \mathbf{A}_1(\mathbf{Z}_i)\} \mid \mathcal{D}_1\right),
\end{aligned}
$$

where

$$\mathbf{A}_1(\mathbf{Z}_i) = E[\mathbf{a}_1(Y)\exp\{-g(Y)\} \mid \mathbf{Z}_i,1], \text{ and } \widehat{\mathbf{A}}_1(\mathbf{Z}_i) = E[\widehat{\mathbf{a}}_1(Y;\boldsymbol{\theta}_0,g)\exp\{-g(Y)\} \mid \mathbf{Z}_i,1].$$

By (C4), we know that given $\mathcal{D}_1$, there exists a positive constant c such that

$$\|[\mathbf{A}_1(\mathbf{z})]_k - [\widehat{\mathbf{A}}_1(\mathbf{z})]_k\|_2 \le cN_1^{-1/2}, \quad k = 1,\ldots,p.$$

Hence,

$$E\left[\{\widehat{\mathbf{S}}_{\boldsymbol{\beta},\mathrm{eff}}(\mathbf{O}_i;\boldsymbol{\theta}_0,g) - \mathbf{S}_{\boldsymbol{\beta},\mathrm{eff}}(\mathbf{O}_i;\boldsymbol{\theta}_0,g)\}\frac{\partial \log\{f_{\mathbf{X},R,RY}(\mathbf{O}_i;\boldsymbol{\theta}_0,\boldsymbol{\gamma}_0)\}}{\partial \boldsymbol{\gamma}_0^{\mathrm{T}}} \mid \mathcal{D}_1\right] = O_p(N_1^{-1/2}).$$

Similarly, we can show that

$$E\left[\{\widehat{\mathbf{S}}_{\boldsymbol{\alpha},\text{eff}}(\mathbf{O}_i;\boldsymbol{\theta}_0,g)-\mathbf{S}_{\boldsymbol{\alpha},\text{eff}}(\mathbf{O}_i;\boldsymbol{\theta}_0,g)\}\frac{\partial\log\{f_{\mathbf{X},R,RY}(\mathbf{O}_i;\boldsymbol{\theta}_0,\boldsymbol{\gamma}_0)\}}{\partial\boldsymbol{\gamma}_0^{\mathrm{T}}}\mid\mathcal{D}_1\right]=O_p(N_1^{-1/2}).$$

Therefore, by (A.1) we have

$$N_2^{-1}\sum_{i\in\mathcal{I}_2}\left[\frac{\partial\widehat{\mathbf{S}}_{\boldsymbol{\theta},\text{eff}}\{\mathbf{o}_i;\widetilde{\boldsymbol{\theta}},\widehat{g}(\cdot;\widetilde{\boldsymbol{\gamma}})\}}{\partial\widetilde{\boldsymbol{\gamma}}^{\mathrm{T}}}\right]_{j,k}=O_p\{\|\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0\|_1+\|\widehat{\boldsymbol{\gamma}}-\boldsymbol{\gamma}_0\|_1+h_L^m+N_2^{-1/2}+N_1^{-1/2}\}. \tag{A.3}$$

Second, we consider the term $\widehat{\boldsymbol{\gamma}}(\widehat{\boldsymbol{\theta}})-\boldsymbol{\gamma}_0$. Since for any $\boldsymbol{\theta}$, we have

$$N_2^{-1}\sum_{i\in\mathcal{I}_2}\mathbf{S}_g\{\mathbf{o}_i;\boldsymbol{\theta},\widehat{\boldsymbol{\gamma}}(\boldsymbol{\theta})\}=\mathbf{0},$$

then

$$\frac{\partial\widehat{\boldsymbol{\gamma}}(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}^{\mathrm{T}}}=-\{\mathbf{T}_{11}(\boldsymbol{\theta})\}^{-1}\mathbf{T}_{12}(\boldsymbol{\theta}),$$

where

$$\mathbf{T}_{11}(\boldsymbol{\theta})=N_2^{-1}\sum_{i\in\mathcal{I}_2}\frac{\partial\mathbf{S}_g\{\mathbf{o}_i;\boldsymbol{\theta},\widehat{\boldsymbol{\gamma}}(\boldsymbol{\theta})\}}{\partial\widehat{\boldsymbol{\gamma}}(\boldsymbol{\theta})^{\mathrm{T}}},\ \text{and}\ \mathbf{T}_{12}(\boldsymbol{\theta})=N_2^{-1}\sum_{i\in\mathcal{I}_2}\frac{\partial\mathbf{S}_g(\mathbf{o}_i;\boldsymbol{\theta},\widehat{\boldsymbol{\gamma}})}{\partial\boldsymbol{\theta}^{\mathrm{T}}}.$$

By Taylor expansion we have

$$\widehat{\boldsymbol{\gamma}}(\widehat{\boldsymbol{\theta}})-\widehat{\boldsymbol{\gamma}}(\boldsymbol{\theta}_0)=\frac{\partial\widehat{\boldsymbol{\gamma}}(\widetilde{\boldsymbol{\theta}})}{\partial\widetilde{\boldsymbol{\theta}}^{\mathrm{T}}}(\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0)=-\{\mathbf{T}_{11}(\widetilde{\boldsymbol{\theta}})\}^{-1}\mathbf{T}_{12}(\widetilde{\boldsymbol{\theta}})(\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0),$$

where $\widetilde{\boldsymbol{\theta}}$ is between $\widehat{\boldsymbol{\theta}}$ and $\boldsymbol{\theta}_0$. Further,

$$\widehat{\boldsymbol{\gamma}}(\boldsymbol{\theta}_0)-\boldsymbol{\gamma}_0=-\{\mathbf{T}_{21}(\boldsymbol{\theta}_0)\}^{-1}\mathbf{T}_{22}(\boldsymbol{\theta}_0),$$

where

$$\mathbf{T}_{21}(\boldsymbol{\theta}_0)=N_2^{-1}\sum_{i\in\mathcal{I}_2}\frac{\partial\mathbf{S}_g\{\mathbf{o}_i;\boldsymbol{\theta}_0,\widetilde{\boldsymbol{\gamma}}(\boldsymbol{\theta}_0)\}}{\partial\widetilde{\boldsymbol{\gamma}}(\boldsymbol{\theta}_0)^{\mathrm{T}}},\ \text{and}\ \mathbf{T}_{22}(\boldsymbol{\theta}_0)=N_2^{-1}\sum_{i\in\mathcal{I}_2}\mathbf{S}_g(\mathbf{o}_i;\boldsymbol{\theta}_0,\boldsymbol{\gamma}_0),$$

and $\widetilde{\boldsymbol{\gamma}}(\boldsymbol{\theta}_0)$ is in the line connecting $\widehat{\boldsymbol{\gamma}}(\boldsymbol{\theta}_0)$ and $\boldsymbol{\gamma}_0$. We first consider the order of $\mathbf{T}_{22}(\boldsymbol{\theta}_0)$. Since

$$\sup_y|\widehat{g}(y;\boldsymbol{\gamma}_0)-g(y)|=O_p(h_L^m),$$

then we have

$$E[\exp\{-\widehat{g}(Y;\boldsymbol{\gamma}_0)\}\mid\mathbf{z}_i,1]=E[\exp\{-g(Y)\}\mid\mathbf{z}_i,1]+O_p(h_L^m).$$

Due to the local linear approximation, we have

$$\|\mathbf{T}_{22}(\boldsymbol{\theta}_0)\|_\infty=O_p\{h_L^m+h^2+(n_2h)^{-1/2}\}.$$

Then we consider the order of off-diagonal elements in matrices $\mathbf{T}_{11}(\boldsymbol{\theta})$ and $\mathbf{T}_{21}(\boldsymbol{\theta}_0)$. For $j\neq l$, we have

$$\begin{aligned}&N_2^{-1}\sum_{i\in\mathcal{I}_2}\frac{\partial S_{g,l}(\mathbf{o}_i;\widehat{\boldsymbol{\theta}},\widehat{\boldsymbol{\gamma}})}{\partial\widehat{\gamma}_j}\\&=N_2^{-1}\sum_{i\in\mathcal{I}_2}\frac{(r_i-1)\exp(-\widehat{\gamma}_l)f_{Y|\mathbf{Z},R=1}(d_l,\mathbf{z}_i;\widehat{\boldsymbol{\alpha}})E[\exp\{-\widehat{g}(Y;\widehat{\boldsymbol{\gamma}})\}\{\partial\widehat{g}(Y;\widehat{\boldsymbol{\gamma}})/\partial\widehat{\gamma}_j\}\mid\mathbf{z}_i,1]}{[1+\exp\{-\widehat{\gamma}_l-h(\mathbf{u}_i;\widehat{\boldsymbol{\beta}})\}](E[\exp\{-\widehat{g}(Y;\widehat{\boldsymbol{\gamma}})\}\mid\mathbf{z}_i,1])^2}\\&=O_p(h_L).\end{aligned}$$

The last equality holds because

$$\frac{\partial\widehat{g}(y;\widehat{\boldsymbol{\gamma}})}{\partial\widehat{\gamma}_j}\equiv0,\quad\text{for } y\notin(d_{j-m+1},d_{j+m-1}),$$

where $m-1$ is the degree of polynomial interpolation. Then we have

$$\|\widehat{\boldsymbol{\gamma}}(\widehat{\boldsymbol{\theta}})-\boldsymbol{\gamma}_0\|_\infty=O_p\{h_L^m+h^2+(n_2h)^{-1/2}\}.$$

Under conditions (C1), (C2) and (C3), by (A.3), we have

$$
\begin{aligned}
&N_2^{-1}\sum_{i\in\mathcal{I}_2}\frac{\partial\widehat{\mathbf{S}}_{\boldsymbol{\theta},\mathrm{eff}}\{\mathbf{o}_i;\widetilde{\boldsymbol{\theta}},\widehat{g}(\cdot;\widetilde{\boldsymbol{\gamma}})\}}{\partial\widetilde{\boldsymbol{\gamma}}^{\mathrm{T}}}N_2^{1/2}\{\widehat{\boldsymbol{\gamma}}(\widehat{\boldsymbol{\theta}})-\boldsymbol{\gamma}_0\}\\
&=O_p(LN_2^{1/2}[L\{h_L^m+h^2+(n_2h)^{-1/2}\}+N_1^{-1/2}+N_2^{-1/2}]\{h_L^m+h^2+(n_2h)^{-1/2}\})\\
&=O_p[L^2N_2^{1/2}\{h_L^m+h^2+(n_2h)^{-1/2}\}^2+L(N_2^{1/2}N_1^{-1/2}+1)\{h_L^m+h^2+(n_2h)^{-1/2}\}]\\
&=o_p(1).
\end{aligned}
$$

Hence, we have

$$
N_2^{-1/2}\sum_{i\in\mathcal{I}_2}\widehat{\mathbf{S}}_{\boldsymbol{\theta},\mathrm{eff}}\{\mathbf{o}_i;\widetilde{\boldsymbol{\theta}},\widehat{g}(\cdot;\widetilde{\boldsymbol{\gamma}})\}+N_2^{-1}\sum_{i\in\mathcal{I}_2}\frac{\partial\widehat{\mathbf{S}}_{\boldsymbol{\theta},\mathrm{eff}}\{\mathbf{o}_i;\widetilde{\boldsymbol{\theta}},\widehat{g}(\cdot;\widetilde{\boldsymbol{\gamma}})\}}{\partial\widetilde{\boldsymbol{\theta}}^{\mathrm{T}}}N_2^{1/2}(\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0)=o_p(1).
$$

Note that

$$
\begin{aligned}
&N_2^{-1}\sum_{i\in\mathcal{I}_2}\frac{\partial\widehat{\mathbf{S}}_{\boldsymbol{\theta},\mathrm{eff}}\{\mathbf{o}_i;\widetilde{\boldsymbol{\theta}},\widehat{g}(\cdot;\widetilde{\boldsymbol{\gamma}})\}}{\partial\widetilde{\boldsymbol{\theta}}^{\mathrm{T}}}\\
&=N_2^{-1}\sum_{i\in\mathcal{I}_2}\frac{\partial\widehat{\mathbf{S}}_{\boldsymbol{\theta},\mathrm{eff}}\{\mathbf{o}_i;\boldsymbol{\theta}_0,g(\cdot;\boldsymbol{\gamma}_0)\}}{\partial\boldsymbol{\theta}_0^{\mathrm{T}}}+O_p(\|\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0\|_1+\|\widehat{\boldsymbol{\gamma}}-\boldsymbol{\gamma}_0\|_1+h_L^m)\\
&=E\left[\frac{\partial\widehat{\mathbf{S}}_{\boldsymbol{\theta},\mathrm{eff}}\{\mathbf{O}_i;\boldsymbol{\theta}_0,g(\cdot;\boldsymbol{\gamma}_0)\}}{\partial\boldsymbol{\theta}_0^{\mathrm{T}}}\mid\mathcal{D}_1\right]+O_p(\|\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0\|_1+\|\widehat{\boldsymbol{\gamma}}-\boldsymbol{\gamma}_0\|_1+h_L^m+N_2^{-1/2})\\
&=E\left\{\frac{\partial\mathbf{S}_{\boldsymbol{\theta},\mathrm{eff}}(\mathbf{O}_i;\boldsymbol{\theta}_0,g)}{\partial\boldsymbol{\theta}_0^{\mathrm{T}}}\right\}+O_p(\|\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0\|_1+\|\widehat{\boldsymbol{\gamma}}-\boldsymbol{\gamma}_0\|_1+h_L^m+N_2^{-1/2}+N_1^{-1/2})\\
&=E\left\{\frac{\partial\mathbf{S}_{\boldsymbol{\theta},\mathrm{eff}}(\mathbf{O}_i;\boldsymbol{\theta}_0,g)}{\partial\boldsymbol{\theta}_0^{\mathrm{T}}}\right\}+o_p(1).
\end{aligned}
$$

Further, under conditions (C1), (C2) and (C3), we have

$$
\begin{aligned}
N_2^{-1/2}\sum_{i\in\mathcal{I}_2}\widehat{\mathbf{S}}_{\boldsymbol{\theta},\mathrm{eff}}\{\mathbf{o}_i;\boldsymbol{\theta}_0,\widehat{g}(\cdot;\boldsymbol{\gamma}_0)\}&=N_2^{-1/2}\sum_{i\in\mathcal{I}_2}\widehat{\mathbf{S}}_{\boldsymbol{\theta},\mathrm{eff}}\{\mathbf{o}_i;\boldsymbol{\theta}_0,g(\cdot;\boldsymbol{\gamma}_0)\}+O_p(N_2^{1/2}h_L^m)\\
&=N_2^{-1/2}\sum_{i\in\mathcal{I}_2}\mathbf{S}_{\boldsymbol{\theta},\mathrm{eff}}(\mathbf{o}_i;\boldsymbol{\theta}_0,g)+O_p(N_1^{-1/2})+O_p(N_2^{1/2}h_L^m)\\
&=N_2^{-1/2}\sum_{i\in\mathcal{I}_2}\mathbf{S}_{\boldsymbol{\theta},\mathrm{eff}}(\mathbf{o}_i;\boldsymbol{\theta}_0,g)+o_p(1).
\end{aligned}
$$

The second equality holds because $E\{R-w(\mathbf{X})\mid\mathbf{X}\}=0$ and

$$
\begin{aligned}
&N_2^{-1/2}\sum_{i\in\mathcal{I}_2}\{\widehat{\mathbf{S}}_{\boldsymbol{\theta},\mathrm{eff}}(\mathbf{o}_i;\boldsymbol{\theta}_0,g)-\mathbf{S}_{\boldsymbol{\theta},\mathrm{eff}}(\mathbf{o}_i;\boldsymbol{\theta}_0,g)\}\\
&=N_2^{-1/2}\sum_{i\in\mathcal{I}_2}\frac{r_i-w(\mathbf{x}_i)}{1-w(\mathbf{x}_i)}w(\mathbf{x}_i)\exp\{-h(\mathbf{u}_i;\boldsymbol{\beta}_0)\}\begin{pmatrix}\mathbf{A}_0(\mathbf{z}_i)-\widehat{\mathbf{A}}_0(\mathbf{z}_i)\\ \mathbf{A}_1(\mathbf{z}_i)-\widehat{\mathbf{A}}_1(\mathbf{z}_i)\end{pmatrix}\\
&=O_p(N_1^{-1/2}),
\end{aligned}
$$

element-wise, where

$$
\mathbf{A}_0(\mathbf{z}_i)=E[\mathbf{a}_0(Y)\exp\{-g(Y)\}\mid\mathbf{z}_i,1],\ \text{and}\ \widehat{\mathbf{A}}_0(\mathbf{z}_i)=E[\widehat{\mathbf{a}}_0(Y;\boldsymbol{\theta}_0,g)\exp\{-g(Y)\}\mid\mathbf{z}_i,1].
$$

Then we obtain that

$$
N_2^{1/2}(\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0)=-\left[E\left\{\frac{\partial\mathbf{S}_{\boldsymbol{\theta},\mathrm{eff}}(\mathbf{O}_i;\boldsymbol{\theta}_0,g)}{\partial\boldsymbol{\theta}_0^{\mathrm{T}}}\right\}\right]^{-1}N_2^{-1/2}\sum_{i\in\mathcal{I}_2}\mathbf{S}_{\boldsymbol{\theta},\mathrm{eff}}(\mathbf{o}_i;\boldsymbol{\theta}_0,g)+o_p(1).\quad\square
$$

A.6. Estimation of $E(Y)$

We have

$$
\begin{aligned}
E(Y) &= E\{E(Y \mid R)\} \\
&= E(Y \mid R=1)\mathrm{pr}(R=1) + E(Y \mid R=0)\mathrm{pr}(R=0) \\
&= E(Y \mid R=1)\mathrm{pr}(R=1) + E\{E(Y \mid \mathbf{X}, R=0) \mid R=0\}\mathrm{pr}(R=0) \\
&= E(Y \mid R=1)\mathrm{pr}(R=1) + \mathrm{pr}(R=0)\int E(Y \mid \mathbf{X}=\mathbf{x}, R=0)\mathrm{pr}(\mathbf{X}=\mathbf{x} \mid R=0)d\mathbf{x} \\
&= E(Y \mid R=1)\mathrm{pr}(R=1) \\
&\quad + \mathrm{pr}(R=0)\int \frac{\int y\mathrm{pr}(R=0 \mid Y=y, \mathbf{X}=\mathbf{x}) f_{Y|\mathbf{X}}(y,\mathbf{x})dy}{\int \mathrm{pr}(R=0 \mid Y=y, \mathbf{X}=\mathbf{x}) f_{Y|\mathbf{X}}(y,\mathbf{x})dy}\mathrm{pr}(\mathbf{X}=\mathbf{x} \mid R=0)d\mathbf{x} \\
&= E(Y \mid R=1)\mathrm{pr}(R=1) \\
&\quad + \mathrm{pr}(R=0)\int \frac{\int y\{\pi(y,\mathbf{u})^{-1}-1\} f_{Y|\mathbf{X}}(y,\mathbf{x})\pi(y,\mathbf{u})dy}{\int \{\pi(y,\mathbf{u})^{-1}-1\} f_{Y|\mathbf{X}}(y,\mathbf{x})\pi(y,\mathbf{u})dy}\mathrm{pr}(\mathbf{X}=\mathbf{x} \mid R=0)d\mathbf{x} \\
&= E(Y \mid R=1)\mathrm{pr}(R=1) \\
&\quad + \mathrm{pr}(R=0)\int \frac{\int y\{\pi(y,\mathbf{u})^{-1}-1\} f_{Y|\mathbf{X},R=1}(y,\mathbf{x})dy}{\int \{\pi(y,\mathbf{u})^{-1}-1\} f_{Y|\mathbf{X},R=1}(y,\mathbf{x})dy}\mathrm{pr}(\mathbf{X}=\mathbf{x} \mid R=0)d\mathbf{x} \\
&= E(Y \mid R=1)\mathrm{pr}(R=1) \\
&\quad + \mathrm{pr}(R=0)\int \frac{E\{Y\{\pi(Y,\mathbf{U})^{-1}-1\} \mid \mathbf{X}=\mathbf{x}, R=1\}}{E\{\pi(Y,\mathbf{U})^{-1}-1 \mid \mathbf{X}=\mathbf{x}, R=1\}}\mathrm{pr}(\mathbf{X}=\mathbf{x} \mid R=0)d\mathbf{x} \\
&= E(Y \mid R=1)\mathrm{pr}(R=1) \\
&\quad + \mathrm{pr}(R=0)\int \frac{E[Y\exp\{-g(Y)\} \mid \mathbf{Z}=\mathbf{z}, R=1]}{E[\exp\{-g(Y)\} \mid \mathbf{Z}=\mathbf{z}, R=1]}\mathrm{pr}(\mathbf{X}=\mathbf{x} \mid R=0)d\mathbf{x}.
\end{aligned}
$$

References

Bickel, J.P., Klaassen, C.A.J., Ritov, Y., Wellner, J.A., 1993. Efficient and Adaptive Estimation for Semiparametric Models. Springer.

Butler, S., Payne, R., Gunn, I., Burns, J., Paterson, C., 1984. Correlation between serum ionised calcium and serum albumin concentrations in two hospital populations. Br. Med. J. (Clin Res Ed) 289, 948–950.

Chang, T., Kott, P.S., 2008. Using calibration weighting to adjust for nonresponse under a plausible model. Biometrika 95, 555–571.

Gilks, W.R., Wild, P., 1992. Adaptive rejection sampling for Gibbs sampling. Appl. Stat. 41, 337–348.

Heinze-Deml, C., Peters, J., Meinshausen, N., 2018. Invariant causal prediction for nonlinear models. J. Causal Inference 6.

Hossain, A., Mostafa, G., Mannan, K., Prosad Deb, K., Hossain, M., et al., 2015. Correlation between serum albumin level and ionized calcium in idiopathic nephrotic syndrome in children. Urol. Nephrol. Open Access J. 3, 70–71.

Hu, Z., Melton, G.B., Arsoniadis, E.G., Wang, Y., Kwaan, M.R., Simon, G.J., 2017. Strategies for handling missing clinical data for automated surgical site infection detection from the electronic health record. J. Biomed. Inform. 68, 112–120.

Ibrahim, J.G., Lipsitz, S.R., 1996. Parameter estimation from incomplete data in binomial regression when the missing data mechanism is nonignorable. Biometrics 52, 1071–1078.

Johnson, A.E., Pollard, T.J., Shen, L., Li-wei, H.L., Feng, M., Ghassemi, M., Moody, B., Szolovits, P., Celi, L.A., Mark, R.G., 2016. Mimic-iii, a freely accessible critical care database. Sci. Data 3, 160035.

Katz, S., Klotz, I.M., 1953. Interactions of calcium with serum albumin. Arch. Biochem. Biophys. 44, 351–361.

Kim, J.K., Shao, J., 2013. Statistical Methods for Handling Incomplete Data. Chapman & Hall/CRC.

Kim, J.K., Yu, C.L., 2011. A semiparametric estimation of mean functionals with nonignorable missing data. J. Am. Stat. Assoc. 106, 157–165.

Kott, P.S., 2014. Calibration weighting when model and calibration variables can differ. In: Mecatti, F., Conti, L.P., Ranalli, G.M. (Eds.), Contributions to Sampling Statistics. Springer, Cambridge, pp. 1–18.

Li, J., Wang, M., Steinbach, M.S., Kumar, V., Simon, G.J., 2018. Don't do imputation: dealing with informative missing values in ehr data analysis. In: Procedings of the 2018 IEEE International Conference on Big Knowledge (ICBK). IEEE, pp. 415–422.

Little, R.J.A., Rubin, D.B., 2002. Statistical Analysis with Missing Data. Wiley.

Miao, W., Ding, P., Geng, Z., 2016. Identifiability of normal and normal mixture models with nonignorable missing data. J. Am. Stat. Assoc. 111, 1673–1683.

Molenberghs, G., Fitzmaurice, G., Kenward, M.G., Tsiatis, A.A., Verbeke, G., 2014. Handbook of Missing Data Methodology. Chapman & Hall/CRC.

Morikawa, K., Kim, J.K., 2021. Semiparametric optimal estimation with nonignorable nonresponse data. Ann. Stat. In press.

Phillips, A., Shaper, A.G., Whincup, P.H., 1989. Association between serum albumin and mortality from cardiovascular disease, cancer, and other causes. Lancet 2, 1434–1436.

Qin, J., Leung, D., Shao, J., 2002. Estimation with survey data under nonignorable nonresponse or informative sampling. J. Am. Stat. Assoc. 97, 193–200.

Robins, J.M., Ritov, Y., 1997. Toward a curse of dimensionality appropriate (coda) asymptotic theory for semi-parametric models. Stat. Med. 16, 285–319.

Robins, J.M., Rotnitzky, A., Zhao, L.P., 1994. Estimation of regression coefficients when some regressors are not always observed. J. Am. Stat. Assoc. 89, 846–866.

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

Rotnitzky, A., Robins, J.M., 1997. Analysis of semi-parametric regression models with non-ignorable non-response. Stat. Med. 16, 81–102.

Rubin, D.B., 1978. Multiple imputations in sample surveys-a phenomenological Bayesian approach to nonresponse. In: Proceedings of the Survey Research Methods Section of the American Statistical Association. American Statistical Association.

Schafer, J.L., 1997. Analysis of Incomplete Multivariate Data. Chapman & Hall.
Shao, J., Wang, L., 2016. Semiparametric inverse propensity weighting for nonignorable missing data. Biometrika 103, 175–187.
Shao, J., Zhao, J., 2013. Estimation in longitudinal studies with nonignorable dropout. Stat. Interface 6, 303–313.
Sun, B., Liu, L., Miao, W., Wirth, K., Robins, J., Tchetgen Tchetgen, E.J., 2018. Semiparametric estimation with data missing not at random using an instrumental variable. Stat. Sin. 28, 1965–1983.
Tang, G., Little, R.J.A., Raghunathan, T.E., 2003. Analysis of multivariate missing data with nonignorable nonresponse. Biometrika 90, 747–764.
Tchetgen Tchetgen, E.J., Wirth, K.E., 2017. A general instrumental variable framework for regression analysis with outcome missing not at random. Biometrics 73, 1123–1131.
Tsiatis, A.A., 2006. Semiparametric Theory and Missing Data. Springer, New York.
Wang, S., Shao, J., Kim, J.K., 2014. An instrumental variable approach for identification and estimation with nonignorable nonresponse. Stat. Sin. 24, 1097–1116.
Zhao, J., Chen, C., 2020. A nuisance-free inference procedure accounting for the unknown missingness with application to electronic health records. Entropy 22, 1154.
Zhao, J., Ma, Y., 2018. Optimal pseudolikelihood estimation in the analysis of multivariate missing data with nonignorable nonresponse. Biometrika 105, 479–486.
Zhao, J., Ma, Y., 2021. A versatile estimation procedure without estimating the nonignorable missingness mechanism. J. Am. Stat. Assoc. https://doi.org/10.1080/01621459.2021.1893176. In press.
Zhao, J., Shao, J., 2015. Semiparametric pseudo-likelihoods in generalized linear models with nonignorable missing data. J. Am. Stat. Assoc. 110, 1577–1590.