1-6-2019

Semiparametric Optimal Estimation With Nonignorable Nonresponse Data

Kosuke Morikawa
Osaka University

Jae Kwang Kim
Iowa State University, jkim@iastate.edu

Semiparametric Optimal Estimation With Nonignorable Nonresponse Data

Abstract
When the response mechanism is believed to be not missing at random (NMAR), a valid analysis requires stronger assumptions on the response mechanism than standard statistical methods would otherwise require. Semiparametric estimators have been developed under the model assumptions on the response mechanism. In this paper, a new statistical test is proposed to guarantee model identifiability without using any instrumental variable. Furthermore, we develop optimal semiparametric estimation for parameters such as the population mean. Specifically, we propose two semiparametric optimal estimators that do not require any model assumptions other than the response mechanism. Asymptotic properties of the proposed estimators are discussed. An extensive simulation study is presented to compare with some existing methods. We present an application of our method using Korean Labor and Income Panel Survey data.

Keywords
Estimating functions, identification, incomplete data, not missing at random (NMAR), semiparametric efficient estimation

Disciplines
Statistical Methodology | Statistical Models

Comments
This is a pre-print made available through arxiv: https://arxiv.org/abs/1612.09207v2.

Semiparametric Optimal Estimation With Nonignorable Nonresponse Data

Kosuke Morikawa[*,‡] **and Jae Kwang Kim**[†,¶]

Osaka University[‡]*, Iowa State Univerisity*[¶] *and KAIST*[¶]
e-mail: morikawa@sigmath.es.osaka-u.ac.jp; jkim@iastate.edu

Abstract: When the response mechanism is believed to be not missing at random (NMAR), a valid analysis requires stronger assumptions on the response mechanism than standard statistical methods would otherwise require. Semiparametric estimators have been developed under the model assumptions on the response mechanism. In this paper, a new statistical test is proposed to guarantee model identifiability without using any instrumental variable. Furthermore, we develop optimal semiparametric estimation for parameters such as the population mean. Specifically, we propose two semiparametric optimal estimators that do not require any model assumptions other than the response mechanism. Asymptotic properties of the proposed estimators are discussed. An extensive simulation study is presented to compare with some existing methods. We present an application of our method using Korean Labor and Income Panel Survey data.

MSC 2010 subject classifications: Primary 62F35, 62G20; secondary 62G10.
Keywords and phrases: Estimating functions, identification, incomplete data, not missing at random (NMAR), semiparametric efficient estimation.

1. Introduction

Handling missing data often requires some assumptions about the response mechanism. If the study variable does not affect the probability of the response, the response mechanism is called missing at random (MAR) [27]. If, on the other hand, the response probability of a study variable depends on that variable directly, the response mechanism is called not missing at random (NMAR) [15]. Under NMAR, the response probability cannot be verified using the observed study variables only, therefore, additional assumptions about the study variable are often required.

Let r be the response indicator of the study variable y with auxiliary variable x, where r takes 1 if y is observed, and takes 0 otherwise. In this paper, we consider a situation where the study variable y is subject to missingness. Ignorable nonresponse or MAR can be understood as the conditional independence of r and y given x, namely $r \perp y \mid x$, which is usually untestable. Greenlees et al. [8] and Diggle & Kenward [5] proposed a fully parametric approach to analyze nonignorable nonresponse data; their method requires two parametric models:

[*]Supported by JSPS KAKENHI Grant Number JP15549977.
[†]Supported by U.S. National Science Foundation, MMS-1733572.

(i) an outcome model, $[y \mid x]$; and (ii) a response model $[r \mid x, y]$. In practice, it is difficult to verify models (i) and (ii), because some of Y are not observed. For the fully parametric approach, model identification and model misspecification can be a problem, and sensitivity analysis becomes necessary [28, 26, 35, 34]. Sverchkov [31] and Riddles et al. [22] proposed a fully parametric approach that uses different model specifications based on (i) $[y \mid x, r = 1]$, and (ii) $[r \mid x, y]$. Their approach is attractive because one can verify a model for $[y \mid x, r = 1]$ from the observed responses; however, because it is a fully parametric approach, it is still subject to model misspecification.

Recently, several semiparametric approaches have been proposed for nonignorable nonresponses. Ma et al. [17] studied identification and parameter estimation for binary study variables. Tang et al. [32] also considered model identification using an instrumental variable and proposed a maximum pseudo likelihood estimator that does not require model specification of the response mechanism. D'Haultfoeuille [4] also used the same instrumental variable assumption and considered a regression analysis using the nonparametric propensity score model. Zhao & Shao [38] extended the method of Tang et al. [32] and relaxed the condition on the instrumental variable, which is called nonresponse instrumental variable [36]. Fitzmaurice et al. [7] and Skrondal & Rabe-Hesketh [30] proposed protective estimators that do not require a model for the response mechanism, but the application of this approach is limited to situations in which Y is binary. In the meantime, Kim & Yu [13] proposed a semiparametric method for estimating $E(Y)$ using a semiparametric response model, but a validation sample is required in order to estimate the parameters in the response mechanism. Tang et al. [33] used the method of empirical likelihood to extend the method of Kim & Yu [13] to estimate more general parameters. In Zhao et al. [39], the method of Qin et al. [21] was used to construct a $n^{1/2}$-consistent estimator without a validation sample. Morikawa et al. [20] used the kernel regression estimator to remove the parametric model assumption on model (i) $[y \mid x, r = 1]$. Chang & Kott [2] and Wang et al. [36] considered a generalized method of moments (GMM) estimator that uses the response model assumption only, but their method is generally lacking in efficiency. Recently, Shao & Wang [29] proposed a semiparametric inverse propensity weighting method using the nonresponse instrumental variable (NIV) assumption of Wang et al. [36]. However, the above papers do not address efficiency of their semiparametric estimation methods. Furthermore, the NIV assumption is difficult to verify from the sample. Developing an optimal semiparametric estimator and a test procedure for model idenitification under NMAR are an important research topics in missing data analysis.

In this paper we use a parametric model for $[r \mid x, y]$, and a fully nonparametric model for $[y \mid x, r = 1]$ to form a semiparametric model and develop a nonparametric test procedure for model identification of the semiparametric model. After that, we construct optimal estimators for parameters both related to the response mechanisms and for the parameter of interest such as population mean. Efficiency under this setup has already been discussed by Rotnitzky & Robins [25] and Robins et al. [24]. However, their estimator requires many working models to achieve the semiparametric efficiency bound. Misspecification of the working models may lead to loss of efficiency. Therefore, we consider an alternative approach and propose two semiparametric estimators that attain the semiparametric lower bound [1] (1) with a working model assumption or (2) without requiring working model assumptions. The first estimator is an adaptive estimator using a working model for $[y \mid x, r = 1]$. If the working model is correct, the first estimator attains the lower bound. The second one is based on the nonparametric regression model which does not require any additional assumptions, but it still attains the lower bound. All technical details are given in Appendix B.

2. Basic setup

Let $(z_i, r_i),\ i = 1, \ldots, n$ be n realizations from a joint distribution $[z, r]$, where $z = (x^{\mathrm{T}}, y)^{\mathrm{T}}$, x is a d-dimensional covariate vector, y is a response variable, and r is a response indicator of y, i.e., it takes 1 if y is observed, and takes 0 otherwise. Also, let $G_r(z)$ be the observed data when the response indicator is r, i.e., $G_1(z) = z$ and $G_0(z) = x$. Suppose that the response model is $\pi(z; \phi)$ with a q-dimensional parameter $\phi \in \Phi$. Let $\theta \in \Theta$ be an one-dimensional parameter satisfying $E\{U(Z; \theta)\} = 0$, where U is a known function of z, which does not prescribe the distribution of $[x, y]$. For example, if we are interested in $E(Y)$, then $U(z; \theta) = y - \theta$. In this paper, we consider semiparametric estimation of (ϕ, θ) from partial observations. In particular, we seek the most efficient estimator among the regular asymptotically linear estimators [1, 34] and propose two adaptive estimators.

For model identification, we need to check

$$\begin{aligned} &\pi(z;\phi) f(y \mid x) = \pi(z;\phi') f'(y \mid x) \quad \text{for almost all } z \\ \Rightarrow\ &\phi = \phi' \text{ and } f(y \mid x) = f'(y \mid x) \quad \text{for almost all } z \end{aligned}, \tag{2.1}$$

where $f(y \mid x)$ and $f'(y \mid x)$ are conditional density functions of $[y \mid x]$. If the above condition does not hold, two different models have the same observed likelihood and cannot be identified. To guarantee the condition, Miao et al. [18] gives a sufficient condition when the outcome models are normal or normal mixture. However, the normality assumption cannot be checked directly from observed data. In the meantime, Wang et al. [36] developed a theory for identification by assuming that there exists a NIV x_1 in the covariate vector $x = (x_1^{\mathrm{T}}, x_2^{\mathrm{T}})^{\mathrm{T}}$such that x_2 is independent of r, given x_1 and y. When x is the single variable, x itself is the NIV. Although the existence of such a NIV is a sufficient condition, it is hard to verify it from the observed data. Therefore, both identification conditions is not testable with observed data. In §3, we propose an alternative condition for the model identification by assuming a restriction on $[y \mid x, r = 1]$, not only on the response mechanism, and develop a test procedure for model identification.

Classical approaches for analyzing nonignorable nonresponse data are based on correct specification for $[y \mid x]$ as well as the response mechanism [8]. This requirement can be challenging because the specification cannot be verified under nonignorable nonresponse [19]. Chang & Kott [2] proposed a semiparametric estimator for ϕ based on the following estimating equation:

$$\sum_{i=1}^{n}\left\{1-\frac{r_i}{\pi(z_i;\phi)}\right\}g(x_i)=0, \tag{2.2}$$

where $g=\{g_1(x),g_2(x),\dots g_q(x)\}^{\mathrm{T}}$, which can be called calibration function, is a function of x whose elements are linearly independent; q is the dimension of ϕ. Note that although this estimator satisfies consistency and asymptotic normality under certain regularity conditions, its efficiency is not guaranteed.

Recently, Riddles et al. [22] proposed an efficient estimator that uses a parametric model for $[y \mid x, r=1]$. Using the mean score theorem [16], the maximum likelihood estimator can be obtained by solving

$$\sum_{i=1}^{n}[r_i s_1(z_i;\phi)+(1-r_i)E_0\{s_0(Z;\phi)\mid x_i\}]=0,$$

where $s_r(z;\phi)$ is the score function of ϕ, that is,

$$s_r(z;\phi)=\frac{\{r-\pi(z;\phi)\}\dot{\pi}(z;\phi)}{\pi(z;\phi)\{1-\pi(z;\phi)\}}, \tag{2.3}$$

$\dot{\pi}(z;\phi)=\partial\pi(z;\phi)/\partial\phi$, and $E_0(\cdot\mid x)$ is the conditional expectation conditional on x and $r=0$. To compute $E_0(\cdot\mid x)$, under Bayes' formula, Riddles et al. [22] proposed using

$$\sum_{i=1}^{n}\left[r_i s_1(z_i;\phi)+(1-r_i)\frac{E_1\{O(Z;\phi)s_0(Z;\phi)\mid x_i\}}{E_1\{O(Z;\phi)\mid x_i\}}\right]=0, \tag{2.4}$$

where $O(z;\phi)=\{1-\pi(z;\phi)\}/\pi(z;\phi)$, and $E_1(\cdot\mid x)$ is the conditional expectation on y given x and $r=1$. The conditional expectation is computed by assuming a parametric model $f_1(y\mid x;\gamma)=f(y\mid x,r=1;\gamma)$. This may increase the efficiency, however, because misspecification of the f_1 model would cause the solution $\hat{\phi}$ to be inconsistent. Morikawa et al. [20] proposed a semiparametric method using a nonparameteric estimator of f_1, assuming that the semiparametric model is identified. We now give more rigorous treatments of the model identification of the semiparametric model.

3. Identification

We consider an identification condition which can be checked with observed data. In this section, we assume that the dimension of the covariate d is one, and the support of the covariate is $[0,1]$ for simplicity. Let $O(z;\phi)=1/\pi(z;\phi)-1$ be the odds function of the response model, $E_1(\cdot\mid x)$ be the operator for the conditional expectation given x and $r=1$. Then the identification condition for the semiparametric model is given in the next theorem.

Theorem 3.1. *Suppose that $E_1\{O(Z;\phi) \mid x\}$ is bounded for almost all x. If $f_1(y \mid x)$ is identifiable, and $E_1\{O(Z;\phi) \mid x\} = E_1\{O(Z;\phi') \mid x\}$ for almost all x implies $\phi = \phi'$, then this is a necessary and sufficient condition for model identification.*

This theorem indicates that we have only to check the identification of $E_1\{O(Z;\phi) \mid x\}$ as long as the f_1 model is identifiable. Checking the identification of $E_1\{O(Z;\phi) \mid x\}$ is relatively easy and feasible with observed data. For example, if the response mechanism is specified as $\pi(z;\phi) = 1/\{1 + \exp(\phi_{x0} + \phi_{x1}x + \phi_y y)\}$, where $\phi = (\phi_{x0}, \phi_{x1}, \phi_y)^T$. Then, $E_1\{O(Z;\phi) \mid x\}$ is written as

$$E_1\{O(Z;\phi) \mid x\} = \exp\{\phi_{x0} + \phi_{x1}x + K_{\phi_y}(x)\}, \tag{3.1}$$

where $K_{\phi_y}(x) = \log E_1\{\exp(\phi_y Y) \mid x\}$ is the cumulant-generating function of $[y \mid x, r = 1]$. Therefore, we have only to check whether $K_{\phi_y}(x)$ is linear with respect to x or not. If f_1 is a parametric model, the model identification for ϕ is easy to check. For example, if $[y \mid x, r = 1]$ belongs to an exponential family with the density function

$$f_1(y \mid x; \tau, \psi) = \exp\left[\frac{y\tau(x) - b\{\tau(x)\}}{\psi} + c(y,\psi)\right],$$

where ψ is the dispersion parameter and τ, b, c are known functions, then the cumulant-generating function reduces to $K_{\phi_y}(x) = \{b(\phi_y\psi + \tau(x)) - b(\tau(x))\}/\psi$, from which we can verify the model identification. For example, for model identification, b is allowed to be any polynomial function except for the 1st- and 2nd-order function of x such as log-function (e.g. Gamma distribution), exponential-function (e.g. Poisson distribution), etc. However, when b is a 2nd-order polynomial function, for example, $b(\tau) = \tau^2/2$, which means f_1 follows normal distribution, then $K_{\phi_y}(x) = \tau(x)\phi_y + \phi_y^2\psi^2/2$. Also, we obtain

$$E_1\{O(Z;\phi) \mid x\} = \exp\{\phi_{x0} + \phi_{x1}x + \tau(x)\phi_y + \phi_y^2\psi^2/2\}.$$

Thus, by Theorem 3.1, ϕ is identifiable unless the mean structure $\tau(x)$ is linear since there are three parameters. The identifiability for other distributions of $[y \mid x, r = 1]$ can be checked in a similar way. If $\tau(x)$ is linear, we may use a transformation approach which is introduced in §7.

On the other hand, checking the model identifiability with a nonprametric $f_1(y \mid x)$ model is hard because, there is no method to estimate the cumulative function $K_{\phi_y}(x)$ nonparametrically for every ϕ_y as far as we know. Therefore, we propose a test statistic to test a necessary condition. In view of (3.1), the model is unidentifiable when the cumulative function is linear with respect to x for all ϕ_y, i.e., the null hypothesis H_0: $K_{\phi_y}(x) = c_1(\phi_y) + c_2(\phi_y)x$ holds, where $c_1(\phi_y)$ and $c_1(\phi_y)$ are infinitely differentiable at $\phi_y = 0$. Under the null hypothesis, we have $K_0^{(\ell)}(x) = c_1^{(\ell)}(0) + c_2^{(\ell)}(0)x$ for all $\ell = 1, 2, \ldots$, where the superscript stands for the ℓ-th partial derivative with respect to ϕ_y. As is well known, the cumulant-generating function can be expanded as

$$K_{\phi_y}(x) = \sum_{k=0}^{\infty} \frac{\phi_y^\ell}{\ell!} K_0^{(\ell)}(x) = \sum_{\ell=0}^{\infty} \frac{\phi_y^k}{k!}\{c_1^{(\ell)}(0) + c_2^{(\ell)}(0)x\}.$$

Thus, the linearity of the cumulant-generating function can be checked by that of $K_0^{(\ell)}(x)$ for all ℓ. Based on this idea, we construct a test of a null hypothesis $H_0 : K_0^{(\ell)}(x) = c_1^{(\ell)}(0) + c_2^{(\ell)}(0)x,\ \ell = 1, \dots, L$, for a positive integer L. When $L = 1$, this corresponds to a goodness-of-fit test of a simple linear regression with a normal distribution f_1. Thus, the test with $L > 1$ is a generalization of the f_1 model to more general models.

The partial derivative of the cumulant-generating function can be computed by the moment function $\mu_\ell(x) = E(Y^\ell \mid x)$ $(\ell = 1, \dots, L)$ because there is one to one relationship between the two functions. Specifically, we can express

$$\mu_\ell(x) = \sum_{m=1}^{\ell} B_{\ell,m}(K_0^{(1)}(x), \dots, K_0^{(\ell)}(x)), \tag{3.2}$$

where $B_{\ell,m}$ are incomplete Bell polynomials. For example, when $L = 3$,

$$\mu_1(x) = c_1^{(1)}(0) + c_2^{(1)}(0)x, \tag{3.3}$$

$$\mu_2(x) = c_1^{(2)}(0) + c_2^{(2)}(0)x + \{c_1^{(1)}(0) + c_2^{(1)}(0)x\}^2, \tag{3.4}$$

$$\begin{aligned}\mu_3(x) = c_1^{(3)}(0) + c_2^{(3)}(0)x &+ 3\{c_1^{(2)}(0) + c_2^{(2)}(0)x\}\{c_1^{(1)}(0) + c_2^{(1)}(0)x\} \\ &+ \{c_1^{(1)}(0) + c_2^{(1)}(0)x\}^3.\end{aligned} \tag{3.5}$$

In (3.3), $(c_1^{(1)}(0), c_2^{(1)}(0))$ is estimated by regressing Y on $(1,\ x)$. With estimated $(c_1^{(1)}(0), c_2^{(1)}(0))$, $(c_1^{(2)}(0), c_2^{(2)}(0))$ in (3.4) can be estimated by regressing $Y^2 - \{c_1^{(1)}(0) + c_2^{(1)}(0)x\}^2$ on $(1,\ x)$. In a similar way, all pairs of $(c_1^{(\ell)}(0), c_2^{(\ell)}(0))$ $(\ell = 1, \dots, L)$ can be estimated recursively. Then the sum of the right-hand side of (3.3) to (3.5) is a regressor of $Y + Y^2 + Y^3$. Thus, by checking the goodness-of-fit for the regression $E(Y + Y^2 + Y^3 \mid x)$, the null hypothesis H_0 can be tested. We adopt Eubank & Hart [6]'s nonparametric test because it does not require any nonparametric smoother such as a bandwidth and knots of splines.

Theorem 3.2. *Suppose that there exists a positive integer L, and the cumulant-generating function up to L-th order is linear, i.e., $H_0 : K_0^{(\ell)}(x) = c_1^{(\ell)}(0) + c_2^{(\ell)}(0)x,\ \ell = 1, \dots, L,$, where c_1 and c_2 are L-th differntiable functions of ϕ_{y} at $\phi_{\mathrm{y}} = 0$. Consider a regression model $E(\sum_{\ell=1}^{L} Y^\ell \mid x)$ which is estimated from the relationship (3.2), and let $e_1, \dots, e_n$ be the residuals,*

$$\hat{b}_j = \frac{1}{n}\sum_{i=1}^{n} e_i \cos(\pi j x_i), \quad j = 1, 2, \dots,$$

and $\hat{\sigma}^2$ is the variance based on the residuals $e_1, \dots, e_n$. Also let S_n be a test statistics of the form

$$S_n = \max_{1 \le m \le n-1} \frac{1}{m}\sum_{j=1}^{m} \frac{2n\hat{b}_j^2}{\hat{\sigma}^2}.$$

Then, it holds that

$$S_n \xrightarrow{\mathcal{L}} \sup_{m\geq 1} \frac{1}{m} \sum_{j=1}^{m} Z_j^2, \tag{3.6}$$

as $n \to \infty$, *where* $Z_1, Z_2, \ldots,$ *be a Gaussian process with* $E(Z_j) = 0, j \geq 1$, *and covariances*

$$\mathrm{Cov}(Z_i, Z_j) = \begin{cases} 1 - \beta_i^2, & i = j \\ -\beta_i \beta_j, & i \neq j \end{cases},$$

and the β_j*'s are given in the Appendix. Also, the statistical test based on* S_n *has consistency.*

Remark 3.1. *Note that we have confined the support of the covariate* X *in* $[0, 1]$. *If the dataset does not satisfy this condition, location and scale are to be modified appropriately before data analysis. Also, when the dimension of the covariates is more than one, some modification is needed [see 11, §9.3].*

Remark 3.2. *Instead of using a parametric limiting distribution in* (3.6), *a bootstrap test can be also considered. However, the bootstrap test requires computing the variance of* Y^L, *which needs the computation of* $2L$*-order Bell polynomial in* (3.2). *Therefore the computation for the bootstrap test can be quite heavy.*

4. Efficiency Bound

In this section, we provide an optimal influence function for the true parameter $(\phi_0^{\mathrm{T}}, \theta_0)^{\mathrm{T}}$ that is the most efficient among all regular and asymptotically linear estimators. If the optimal influence function φ_{eff} is found, the semiparametric lower bound is given as $E(\varphi_{\mathrm{eff}} \varphi_{\mathrm{eff}}^{\mathrm{T}})$. We begin by presenting the efficient influence function in Lemma 4.1. Although θ is a parameter not prescribing the distribution of $[x, y]$ as defined in §2, this limitation is just for simplicity and can be removed. For example, Rotnitzky & Robins [25] derived the semiparametric efficiency bound for regression parameters, which prescribe the first moment of the distribution of $[y \mid x]$. However, the ideas used for adaptive estimators expressed in §5 are still applicable for such parameters.

In the following discussion, we abbreviate the parameter value or random variable, for example, $\pi(z; \phi_0) = \pi(z) = \pi(\phi_0)$, unless this would lead to ambiguity.

Lemma 4.1. *Let* $S_{\mathrm{eff}} = (S_1^{\mathrm{T}}, S_2)^{\mathrm{T}}$, *where* $S_1 = S_1(R, G_R(Z))$ *and* $S_2 = S_2(R, G_R(Z))$ *be defined as*

$$S_1(R, G_R(Z); \phi) = \left\{1 - \frac{R}{\pi(Z; \phi)}\right\} g^{\star}(X; \phi_0), \tag{4.1}$$

$$S_2(R, G_R(Z); \phi, \theta) = \frac{R}{\pi(Z; \phi)} U(Z; \theta) + \left\{1 - \frac{R}{\pi(Z; \phi)}\right\} U^{\star}(X; \phi_0, \theta)\}, \tag{4.2}$$

$g^\star(x;\phi_0) = E^\star\{s_0(Z;\phi_0) \mid x;\phi_0\}$, $U^\star(x;\phi_0,\theta) = E^\star\{U(Z;\theta) \mid x;\phi_0\}$, and

$$E^\star\{g(Z) \mid x;\phi_0\} = \frac{E\{O(Z;\phi_0)g(Z) \mid x\}}{E\{O(Z;\phi_0) \mid x\}} \tag{4.3}$$

with $O(z;\phi_0) = \{1-\pi(z;\phi_0)\}/\pi(z;\phi_0)$. Then, the efficient influence function is $\varphi_{\text{eff}} = H^{-1}S_{\text{eff}}$, where $H = E(S_{\text{eff}}^{\otimes 2}) = E\left\{\partial S_{\text{eff}}(\phi_0,\theta_0)/\partial(\phi^{\text{T}},\theta)^{\text{T}}\right\}$ and $B^{\otimes 2} = BB^{\text{T}}$. Therefore, the semiparametric efficiency bound is given by $\{E(S_{\text{eff}}^{\otimes 2})\}^{-1}$.

This theorem implies that if we can compute $E^\star(\cdot \mid x)$, then estimating functions (4.1) and (4.2) will provide an optimal estimator. The optimal estimator is the solution to

$$\sum_{i=1}^n S_{\text{eff},i}(\phi,\theta) = \sum_{i=1}^n \{S_1^{\text{T}}(r_i, G_{r_i}(z_i);\phi), S_2(r_i, G_{r_i}(z_i);\phi,\theta)\}^{\text{T}} = 0. \tag{4.4}$$

The equation based on $S_1(\phi)$ in (4.1) gives an optimal estimator for ϕ, say $\hat{\phi}$. Then, by using $\hat{\phi}$, $S_2(\hat{\phi},\theta)$ in (4.2) can provide an optimal estimator for θ. However, the expectation $E^\star(\cdot \mid x)$ and the parameter ϕ_0 are unknown and need to be estimated. Also, to compute the conditional expectation, we may need to correctly specify the distribution of $[y \mid x]$, which is subjective and unverifiable, as is stated in §1. In the next section, two adaptive estimators are proposed to work around the problem and to attain the lower bound derived in Lemma 4.1.

Remark 4.1. *Equation (4.1) can be viewed as a special case of the estimator of Chang & Kott [2] defined in (2.2). Thus, the optimal g function in (2.2) for the Chang & Kott [2] method is given by $g^\star(x,\phi_0)$ in (4.1) although ϕ_0 is unknown. One might think that the efficiency can be improved with a larger dimension of g because the above two methods can handle over-identified models with $q > d+1$. However, according to Lemma 4.1, there is no need to use more g functions and it is enough to consider only $g^\star(x,\phi_0)$ (i.e., $q = d+1$) as the calibration function.*

5. Adaptive Estimators

We now propose two adaptive estimators for (ϕ_0,θ_0): (i) with a parametric working model for $f_1(y \mid x)$; (ii) with a nonparametric estimator for $f_1(y \mid x)$, where $f_1(y \mid x) = f(y \mid x, r=1)$.

To discuss the first method, let $f_1(y \mid x)$ be known up to the parameter $\gamma \in \Gamma$, and let $\hat{\gamma}$ be the maximizer of $\sum_{i=1}^n r_i \log f_1(y_i \mid x_i;\gamma)$. This can be easily implemented, and its validity can be checked by using information criteria such as the Akaike information criterion (AIC) and the Bayesian information criterion (BIC). By using the idea similar to that used to derive (2.4), we can show that, for any function $g(z)$,

$$E^\star\{g(Z) \mid x;\phi_0,\gamma\} = \frac{E_1\{\pi^{-1}(Z;\phi_0)O(Z;\phi_0)g(Z) \mid x;\gamma\}}{E_1\{\pi^{-1}(Z;\phi_0)O(Z;\phi_0) \mid x;\gamma\}}, \tag{5.1}$$

where $E_1(\cdot \mid x) = E(\cdot \mid x, r = 1)$. Thus, the expectation can be estimated by using $f_1(y \mid x; \hat{\gamma})$ and $\pi(z; \phi_0)$. However, since ϕ_0 is unknown, we propose an estimating equation $\sum_{i=1}^n S_{\text{eff},i}(\phi, \theta, \hat{\gamma}) = 0$, where

$$S_{\text{eff},i}(\phi, \theta, \hat{\gamma}) = \{S_1^{\text{T}}(r_i, G_{r_i}(z_i); \phi, \hat{\gamma}), S_2(r_i, G_{r_i}(z_i); \phi, \theta, \hat{\gamma})\}^{\text{T}}, \tag{5.2}$$

with

$$S_1(r, G_r(z); \phi; \hat{\gamma}) = \left\{1 - \frac{r}{\pi(z; \phi)}\right\} E^\star\{s_0(z; \phi) \mid x_i; \phi, \hat{\gamma}\},$$

$$S_2(r, G_r(z); \phi, \theta, \hat{\gamma}) = \frac{r}{\pi(z; \phi)} U(z; \theta) + \left\{1 - \frac{r_i}{\pi(z; \phi)}\right\} E^\star\{U(z; \theta) \mid x_i; \phi, \hat{\gamma}\}.$$

What if $f_1(y \mid x)$ is misspecified? One might expect the solution to the estimating equation with (5.2) to be inconsistent as a result. To our surprise, however, the estimator that uses the function on the right-hand side of (5.1) is consistent even when the assumed model for $f_1(y \mid x)$ is misspecified. Also, if the model is correctly specified, the estimator attains the lower bound. This leads to Theorem 5.1.

Theorem 5.1. *Let $(\hat{\phi}^{\text{T}}, \hat{\theta})^{\text{T}}$ be the solution to $\sum_{i=1}^n S_{\text{eff},i}(\phi, \theta, \hat{\gamma}) = 0$ in (5.2). Under conditions (C1)–(C8) given in Appendix A and the identification conditions assumed in Theorem 3.1, $(\hat{\phi}^{\text{T}}, \hat{\theta})^{\text{T}}$ satisfies consistency and asymptotic normality with variance*

$$E\left\{\frac{\partial S^*_{\text{eff}}}{\partial(\phi^{\text{T}}, \theta)}\right\}^{-1} E(S^{*\otimes 2}_{\text{eff}}) E\left\{\frac{\partial S^*_{\text{eff}}}{\partial(\phi^{\text{T}}, \theta)^{\text{T}}}\right\}^{-1},$$

even if $f_1(y \mid x; \hat{\gamma})$ is misspecified, where γ^* is the probability limit of $\hat{\gamma}$, and $S^*_{\text{eff}} = \{S_1(\phi_0, \gamma^*)^{\text{T}}, S_2(\phi_0, \theta_0, \gamma^*)\}^{\text{T}}$ is defined in (5.2). In particular, the asymptotic variance of $\hat{\theta}$ is given as

$$V^* = \text{var}\left[\tau_{\text{U}}^{-1}\{S_2(\phi_0, \theta_0, \gamma^*) - \kappa^* S_1(\phi_0, \gamma^*)\}\right], \tag{5.3}$$

where $\kappa^* = \kappa_1^*(\kappa_2^*)^{-1}$, $\kappa_1^* = E[\{U^\star(\phi_0, \theta_0, \gamma^*) - U(\theta_0)\}\dot{\pi}(\phi_0)^{\text{T}}/\pi(\phi_0)\}]$, $\kappa_2^* = E\{g^\star(\phi_0, \gamma^*)\dot{\pi}(\phi_0)^{\text{T}}/\pi(\phi_0)\}$, and $\tau_{\text{U}} = E\{\partial U(\theta_0)/\partial\theta\}$. In addition, if the model is correctly specified, the estimator attains the semiparametric efficiency bound.

Unlike the estimator of Riddles et al. [22], the parametric model f_1 is irrelevant to the consistency and asymptotic normality of the estimator here. Therefore, we call f_1 a working model, as in Liang & Zeger [14]. Also, though equation (4.2) has a form similar to that of the doubly robust estimator under MAR [23], our estimator does not have the doubly robustness property. This is because the computation for $E^\star(\cdot \mid x)$ relies on the correct response mechanism.

Numerical computation is needed to calculate the conditional expectation in (5.1). The expectation-maximization (EM) algorithm considered in Riddles et al. [22] can be used with a minor modification. We can directly apply their method, once the weights w_{ij}^* defined in (15) of Riddles et al. [22] are changed to

$$w_{ij}^* = \frac{r_j \pi^{-1}(x_i, y_j; \phi) O(x_i, y_j; \phi) f_1(y_j \mid x_i, \gamma)/C(y_j; \gamma)}{\sum_{k=1}^n r_k \pi^{-1}(x_i, y_k; \phi) O(x_i, y_k; \phi) f_1(y_k \mid x_i, \gamma)/C(y_k; \gamma)},$$

where $C(y; \gamma) = \sum_{l=1}^n r_l f_1(y \mid x_l; \gamma)$. The weight w_{ij}^* can be called fractional weights in the context of fractional imputation [12]. With these weights, $E^\star\{g(x_i, Y) \mid x_i; \gamma, \phi\}$ can be computed by $\sum_{j=1}^n w_{ij}^* g(x_i, y_j)$.

We now discuss the second adaptive estimator based on nonparametric estimation for $f_1(y \mid x)$. When x is discrete, such as when x is a binary variable, the expectation can be computed by averaging the data conditioned by $X = x$ and $R = 1$, e.g., for $x = 0, 1$,

$$\hat{E}^\star\{g(x, Y) \mid x; \phi\} = \frac{\sum_{j \in I_x} r_j \pi^{-1}(x, y_j; \phi) O(x, y_j; \phi) g(x, y_j)}{\sum_{j \in I_x} r_j \pi^{-1}(x, y_j; \phi) O(x, y_j; \phi)} \tag{5.4}$$

is a consistent estimator of (5.1), where $I_x = \{j \in \{1, \dots, n\} \mid X_j = x\}$.

When x is continuous, the Nadaraya-Watson estimator can be employed. That is, for any function $g(z)$,

$$\hat{E}^\star\{g(x, Y) \mid x; \phi\} = \frac{\sum_{j=1}^n K_h(x - x_j) r_j \pi^{-1}(x, y_j; \phi) O(x, y_j; \phi) g(x, y_j)}{\sum_{j=1}^n K_h(x - x_j) r_j \pi^{-1}(x, y_j; \phi) O(x, y_j; \phi)} \tag{5.5}$$

is consistent under the regularity conditions given in Appendix A. Here, $K_h(x - w) = K\{(x - w)/h\}$, where K is a kernel function, and h is the bandwidth. We have the following result for the adaptive estimators obtained with the Nadaraya-Watson estimation.

Theorem 5.2. *Let $(\hat{\phi}^{\mathrm{T}}, \hat{\theta})^{\mathrm{T}}$ be the solution to $\sum_{i=1}^n \hat{S}_{\mathrm{eff},i}(\phi, \theta) = 0$, where $\hat{S}_{\mathrm{eff},i}(\phi, \theta)$ is defined in (4.4) with the estimated conditional expectation in (5.5). Under Conditions (C1)–(C3), (C9)–(C14) given in Appendix A, $(\hat{\phi}^{\mathrm{T}}, \hat{\theta})^{\mathrm{T}}$ satisfies consistency and asymptotic normality, and the estimator attains the semiparametric efficiency bound.*

Remark 5.1. *The proposed estimator is attractive because it does not require any model assumptions on f_1, but it would not work well when the dimension of x is high, as is common in any nonparametric estimation.*

Variance estimation is also a difficult problem in semiparametric estimation. When we consider a parametric working model $f_1(y \mid x; \gamma)$,

$$\hat{V} = n^{-1} \sum_{i=1}^n \left[\hat{\tau}_{\mathrm{U}}^{-1} \{ S_2(r_i, G_{r_i}(z_i); \hat{\phi}, \hat{\theta}, \hat{\gamma}) - \hat{\kappa} S_1(r_i, G_{r_i}(z_i); \hat{\phi}, \hat{\gamma}) \} \right]^{\otimes 2} \tag{5.6}$$

converges to V^* in probability as defined in (5.3), where $\hat{\tau}_{\mathrm{U}}$ and $\hat{\kappa}$ are consistent estimators for τ_{U} and $\kappa^* = \kappa_1^* (\kappa_2^*)^{-1}$, respectively, for κ_1^* and κ_2^* as defined in Theorem 5.1. To estimate κ_1^*, we propose using the same method that we used to compute θ_0, i.e., let $\mathcal{U}(\phi_0, k_1, \gamma^*) = k_1 - (U^\star(\gamma^*) - U) \dot{\pi}(\phi_0)^{\mathrm{T}} / \pi(\phi_0)$ be our new U-function and let the solution to $E\{\mathcal{U}(\phi_0, k_1, \gamma^*)\} = 0$ with respect to k_1 be our target parameter; solve the following equation:

$$\sum_{i=1}^{n} \left[\frac{r_i}{\pi(z_i; \hat{\phi})} \mathcal{U}(z_i; \hat{\phi}, k_1, \hat{\gamma}) + \left\{ 1 - \frac{r_i}{\pi(z_i; \hat{\phi})} \right\} E^{\star}\{\mathcal{U}(Z; \hat{\phi}, k_1, \hat{\gamma}) \mid x_i; \hat{\gamma}\} \right] = 0.$$

This is the optimal estimator for (ϕ_0, κ_1^*) in terms of the asymptotic variance, because $\mathcal{U}$ is a known function and Theorem 5.1 is applicable. The best estimator for κ_2^* can be obtained in the same way. When we use the nonparametric method stated in Theorem 5.2 to estimate θ_0, the variance can be also estimated by using the nonparametric method (5.4) and (5.5), instead of using the parametric model $f_1(y \mid x; \gamma)$ in (5.6).

6. Simulation Study

In order to evaluate the performance of our proposed estimators and compare their efficiency with other methods in finite samples, we conducted a Monte Carlo simulation study with four scenarios. In each scenario, we used a covariate $X \sim U(-1, 1)$, set the response mechanism to a Bernoulli distribution with $\text{logit}\{\pi_{\text{y}}(y)\} = \phi_{\text{x}0} + \phi_{\text{y}} y$, and generated the response outcome variable from $Y \mid (x, r = 1) \sim N(\mu_s(x), \sigma_s^2)$. In Scenarios 1–3, $\mu_s(x)$ is defined as the s-th order polynomial: $\mu_1(x) = x - 0.121$, $\sigma_1^2 = 1/3$; $\mu_2(x) = 0.8x^2 - 0.3415$, $\sigma_2^2 = 1/4$; $\mu_3(x) = 2x(x - 3/4)(x + 3/4) - 0.0802$, $\sigma_3^2 = 1/3$. In Scenario 4, $\mu_4(x) = \{\cos(x\pi) + 2\sin(2x\pi)\}/2 - 0.06$, $\sigma_4^2 = 1/4$. We generated missing data by the response mechanism with $(\phi_{\text{y}0}, \phi_{\text{y}}) = (1.03, -1.2),\ (0.91, -1),\ (0.9, -0.8),\ (0.91, -0.8)$ in Scenarios 1–4 respectively, so that the response rate is about 70 % and $E(Y) = 0$.

We note that x is a NIV [36]; thus the parameters are identifiable in all scenarios. We also consider the case when the response mechanism is over-specified as $\text{logit}\{\pi_{\text{xy}}(x, y)\} = \phi_{\text{x}0} + \phi_{\text{x}1} x + \phi_{\text{y}} y$. In this case, there is no instrumental variable, but all the parameters are identifiable except for Scenario 1 by Theorem 3.1. However, by using Theorem 3.1, it is possible to make the response model in Scenario 1 identifiable at the risk of misspecification of response mechanism. This problem is covered in the next section. We estimate the parameters for the two response mechanisms $\pi_{\text{y}}(y; \phi)$ and $\pi_{\text{xy}}(x, y; \phi)$, as well as $\theta = E(Y)$. For the response mechanisms, only ϕ_{y} is reported.

From each sample, we computed four estimators, as follows:

[1] MAR: A naive estimator based on the assumption that the missing data are missing-at-random:

$$\sum_{i=1}^{n} \delta_i(\theta - y_i)/\hat{\pi}_i = 0, \tag{6.1}$$

where $\hat{\pi}_i$ is an estimated response mechanism, that is, $\hat{\pi}_i = \{1 + \exp(\hat{\phi}_{\text{x}0} + \hat{\phi}_{\text{x}} x_i)\}^{-1}$, where $(\hat{\phi}_{\text{x}0}, \hat{\phi}_{\text{x}})$ is the maximum likelihood estimator.

[2] CK: The estimator of Chang & Kott [2]. We use the estimating equation (2.2), setting g as $(1, x)$ for $\pi_{\rm y}(y)$ and $(1, x, x^2)$ for $\pi_{\rm xy}(x, y)$; θ is estimated by using (6.1) with the estimated response mechanism.

[3] RKI: The estimator of Riddles et al. [22]. In all scenarios, we specified a parametric model on f_1 based on normal distribution with mean structure $\mu(x) = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3$. A best model among $2^4 - 1$ candidate models was chosen by using AIC; θ is estimated by using (6.1) with the estimated response mechanism. We note that misspecified model was used in Scenario 4.

[4] New: Our proposed estimator. As for the working model for f_1, the same model specification as in the RKI method was used. We also consider the nonparametric estimator proposed in Theorem 5.2. We call the parametric method "P" and the nonparametric method "NP" in this section.

Monte Carlo samples of size $n = 500$ and 2,000 were independently generated 2,000 times. We used the correct models $\pi_{\rm y}(y)$ and $\pi_{\rm xy}(x, y)$ for the response mechanism, except for MAR.

Before estimating the parameters, we check the model identification of the over-specified response model $\pi_{\rm xy}(x, y; \phi)$ by using Theorem 3.2. We set $L = 1$ (for parametric models) and $L = 5$ (for nonparametric models). In this setup, the model of Scenario 1 is unidentifiable, but the others are identifiable (see §3 for details). For all Scenarios, both test statistics with $L = 1, 5$ can detect the identifiable model because f_1 model is normally distributed with a linear mean structure in Scenario 1. The asymptotic distribution of the test statistic is approximated by

$$\sup_{m\geq 1} \frac{1}{m}\sum_{j=1}^{m} Z_j^2 \approx \max_{m=1,\ldots,M} \frac{1}{m}\sum_{j=1}^{m} Z_j^2,$$

with $M = 2,000$, where $Z_1, \ldots, Z_M$ are defined in Theorem 3.2. The Z's are generated 2,000 times for each datset. For Scenario 1, mean of the p-value of the test statistic with sample size $n = 500, 2000$ is 0.566 and 0.579, respectively, which means the test works well. While, for the other Scenarios, the mean of the p-value is almost 0 (less than 0.001) and the standard error is less than 0.01, which means identifiable models are correctly detected for Scenarios 2–4.

Figure 1 shows the Monte Carlo simulation results with the response mechanism $\pi_{\rm y}(y)$ in all scenarios; Figure 2 shows the results with $\pi_{\rm xy}(x, y)$. In Figure 2, only the results for Scenarios 2-4 are shown because the parameters are not identifiable in Scenario 1; the result of MAR is not presented because it is already shown in Figure 1. In the CK method with $\pi_{\rm y}(y)$ mechanism, we encountered some numerical problems in Scenarios 2–4 and there was no solution because the estimate of ϕ did not converge. The following is the summary of the simulation results shown in Figure 1 and Figure 2:

[1] In all scenarios, the naive estimator using the MAR assumption is significantly biased, since this assumption does not hold.

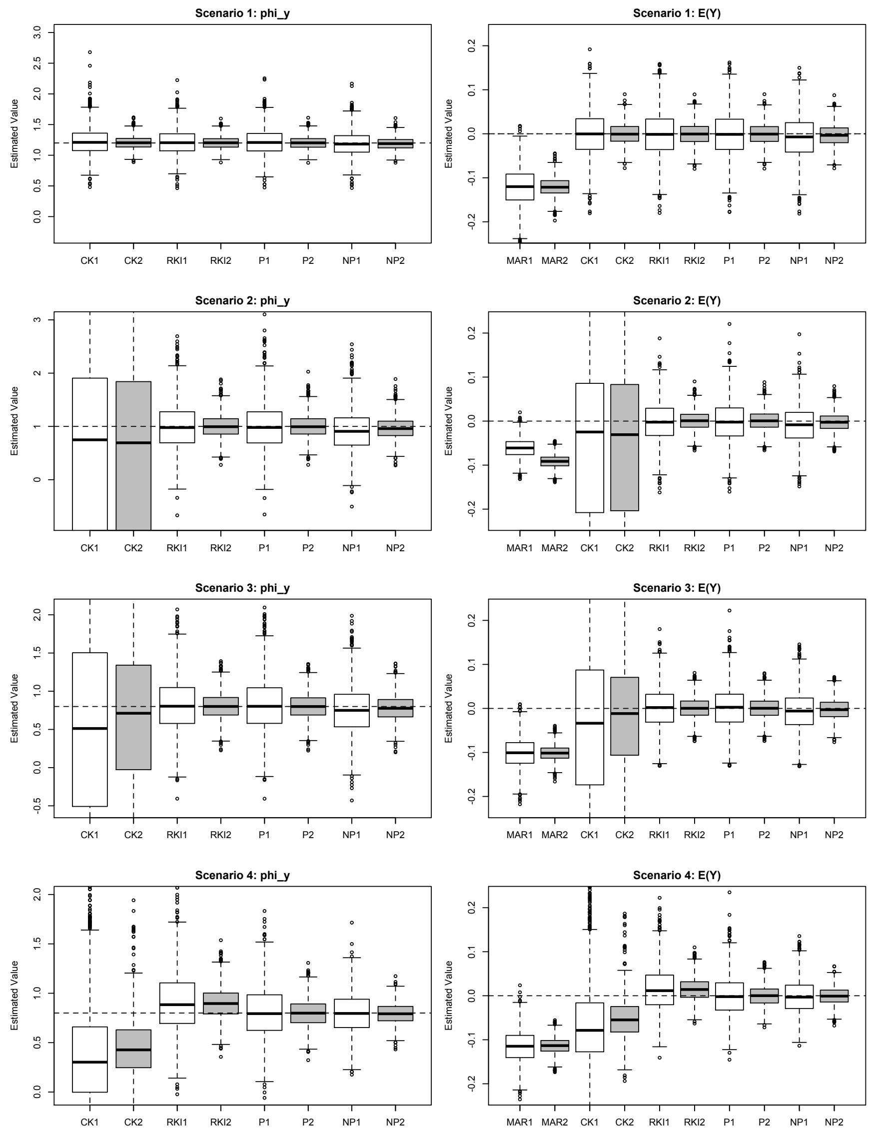

FIG 1. *Boxplot of Monte Carlo results for* ϕ_{y} *and* $\theta\{= E(Y)\}$ *under four scenarios when* $\phi_{\mathrm{x}1}$ *is set to be 0. The four estimators are MAR (missing at random), CK (Chang & Kott's estimator), RKI (Riddles' estimator), P (our proposed estimator with parametric* f_1 *model) NP (our proposed estimator with nonparametric method). Numbers 1 and 2 stand for* $n = 500$ *and* $n = 2,000$*, respectively. The broken line shows the true value.*

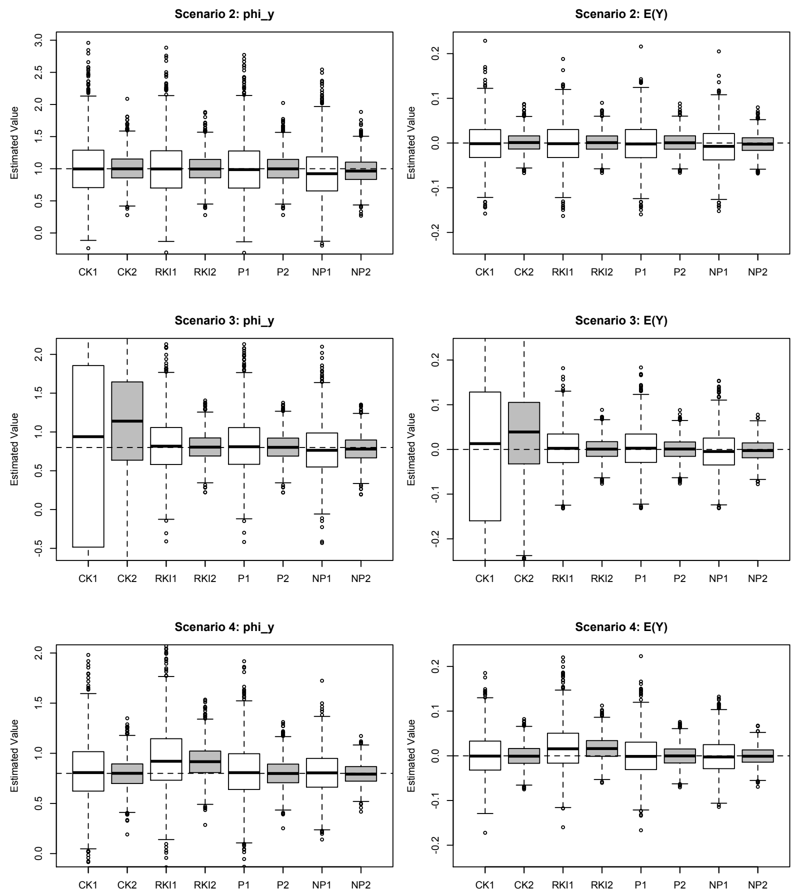

FIG 2. *Boxplot of Monte Carlo results for* ϕ_{y} *and* $\theta\{= E(Y)\}$ *under four scenarios when* $\phi_{\mathrm{x}1}$ *is estimated. The four estimators are MAR (missing at random), CK (Chang & Kott's estimator), RKI (Riddles' estimator), P (our proposed estimator with parametric* f_1 *model) NP (our proposed estimator with nonparametric method). Numbers 1 and 2 stand for* $n = 500$ *and* $n = 2,000$*, respectively. The broken line shows the true value.*

Table 1

The coverage probability of the confidence interval with 95% coverage rate for our proposed estimators with sample size $n = 500$ and $n = 2,000$ when ϕ_{x1} is set to be 0 (fix) and estimated (est).

Scenario	Method	n	ϕ_{x1} fix	ϕ_{x1} est	Scenario	Method	n	ϕ_{x1} fix	ϕ_{x1} est
1	P	500	0.939	–	3	P	500	0.958	0.950
		2000	0.944	–			2000	0.953	0.946
	NP	500	0.930	–		NP	500	0.964	0.940
		2000	0.937	–			2000	0.943	0.944
2	P	500	0.953	0.949	4	P	500	0.948	0.953
		2000	0.943	0.946			2000	0.941	0.943
	NP	500	0.942	0.959		NP	500	0.943	0.942
		2000	0.946	0.951			2000	0.946	0.949

P: method using a parametric working model for f_1, NP: nonparametric method.

[2] The CK method with $\pi_{\mathrm{y}}(y)$ model works well in Scenario 1, but the performance suffers from numerical problems in the other scenarios. However, the CK method with $\pi_{\mathrm{xy}}(x, y)$ model works well even in Scenarios 2 and 4 (though less efficient compared to RKI and our proposed estimators). This is because the calibration condition on $g(x) = [1, x]$ falls short of estimating the parameters when the relationship between x and y becomes more complicated.

[3] The RKI method performs quite well in Scenarios 1–3 for both response mechanisms, but the estimators in Scenario 4 are somewhat positively biased in RKI due to the misspecification of the f_1 model.

[4] In all scenarios, our proposed estimators perform better than any other methods. We note that in Scenario 4, the estimator using parametric f_1 is still consistent despite misspecification of f_1. However it is less efficient compared to the nonparametric method because of the misspecification.

Table 1 shows the estimated coverage probability with 95% coverage confidence interval for our proposed estimators. We applied (5.6) to estimate the variance of our estimators both using the parametric f_1 model and the nonparametric model (see §5). Our proposed variance estimator works well in all scenarios.

7. Real data analysis

In this section, our proposed estimators are applied to the Korea Labor and Income Panel Survey (KLIPS) data, which have been analyzed multiple times [13, 36, 29]. The data contain $n = 2,506$ Korean wage earners; the response variable y is total wage income (10^6 Korean Won) in year 2008. There are three fully observed covariates: x_1: total wage income in the previous year (2007); x_2: gender; x_3: age. While x_1 is a continuous variable, x_2 has two categories 1 and 2 for male and female, and x_3 has three categories 1-3: $x_3 < 35$, $35 \le x_3 < 51$, and $x_3 \geq 51$. We also identified three data points as outliers and excluded them from further analysis.

Although the data are completely observed, we took the approach of Kim & Yu [13] and made eight artificial incomplete datasets by assuming the following eight response mechanisms: M1 (linear nonignorable without (x_2, x_3)): $\mathrm{logit}(\pi) = 0.48 - 0.3x_1 - 0.5y$; M2 (linear nonignorable): $\mathrm{logit}(\pi) = -0.85 - 0.2x_1 + 0.5x_2 + 0.2x_3 - 0.4y$; M3 (nonlinear nonignorable, quadratic in x_1 without (x_2, x_3)): $\mathrm{logit}(\pi) = 0.33 - 0.3x_1 - 0.1x_1^2 - 0.3y$; M4 (nonlinear nonignorable, quadratic in x_1): $\mathrm{logit}(\pi) = -0.89 - 0.4x_1 - 0.1x_1^2 + 0.5x_2 + 0.2x_3 - 0.4y$; M5 (nonlinear nonignorable, quadratic in y without (x_2, x_3)): $\mathrm{logit}(\pi) = 0.24 - 0.25x_1 - 0.25y - 0.1y^2$; M6 (nonlinear nonignorable, quadratic in y): $\mathrm{logit}(\pi) = -0.93 - 0.2x_1 + 0.5x_2 + 0.2x_3 - 0.2y - 0.1y^2$; M7 (jump nonignorable without x) $\pi = 0.5I(y \leq 1.7) + 0.9(y > 1.7)$; M8 (jump nonignorable) $\pi = 0.5I(0.5x_2 + 0.2x_3 + y \leq 2.6) + 0.9(0.5x_2 + 0.2x_3 + y > 2.6)$, where $I(A)$ is the indicator function that takes 1(0) if an event A is true (false). Note that there are NIVs for models M2, M4, M6, and M8. For all data sets, the response rate is about 70%. We estimated $\theta = E(Y)$ as considered in the simulation. The "true" average income in 2008 is $\hat{\theta}_n = 1.846$ as calculated using the complete data. In order to estimate the parameters, we assumed a response mechanism $\mathrm{logit}\{\pi(x, y; \phi)\} = \phi_{\mathrm{x}0} + \phi_{\mathrm{x}1}x_1 + \phi_{\mathrm{x}2}x_2 + \phi_{\mathrm{x}3}x_3 + \phi_{\mathrm{y}}y$. Therefore M1 and M2 are correctly specified while M3-M8 are misspecified.

We specified unknown f_1 models as normal distribution $Y \mid (x_1, x_2 = i, x_3 = j, r = 1) \sim N(\mu_{i,j}(x_1), \sigma_{i,j}^2)$ $(i = 1, 2;\ j = 1, 2, 3)$, where $\mu_{i,j}(x_1) = \gamma_{0i,j} + \gamma_{1i,j}x_1 + \gamma_{2i,j}x_1^2 + \gamma_{3i,j}x_1^3 + \gamma_{4i,j}x_1^4$; $(\gamma_{1i,j}, \gamma_{2i,j}, \gamma_{3i,j}, \gamma_{4i,j})$ is the regression parameter when $(x_2, x_3) = (i, j)$. We chose the best model by AIC among $2^5 - 1$ models for each (x_2, x_3)'s 2×3 pattern. By using Theorem 3.1, this model is identifiable as one of the 6 mean structures with $\mu_{i,j}$ being nonlinear with respect to x_1. However, in the real data, the correlation between x_1 and y is too high because wage income does not change considerably in one year; the mean structure is almost linear even when stratified by x_2 and x_3. Therefore, to obtain valid estimator of θ, we considered two different approaches: [1] find NIVs used; [2] transform x_1 in the response model so that the relationship can be nonlinear. For the first approach, we specified x_2, x_3, and (x_2, x_3) as instrumental variables in applying our proposed method, which will lead to inconsistency for models M2 because there is actually no instrumental variable. For the second approach, we transformed x_1 to $\log(x_1)$ so that (3.1) becomes identifiable even if the cumulant-generating function is normally distributed with a linear mean structure. Although this transformation made the model identifiable, this also changed the assumed mechanism to $\mathrm{logit}\{\pi(x, y; \phi)\} = \phi_{\mathrm{x}0} + \phi_{\mathrm{x}1}\log(x_1) + \phi_{\mathrm{x}2}x_2 + \phi_{\mathrm{x}3}x_3 + \phi_{\mathrm{y}}y$. This may be a potential cause of biased estimation. On the flip side, this approach uses all information of covariates, which helps to reduce bias and gain efficiency. We show the result of this approach under both parametric and nonparametric f_1 models.

In Table 2, deviation of estimators from the full sample estimate $\hat{\theta}_n$ and estimated standard errors are shown. The methods using instrumental variable

Table 2

$\hat{\theta} - \hat{\theta}_n$ (S.E. ($\hat{\theta}$)): deviation of our proposed estimator $\hat{\theta}$ from the full sample estimate $\hat{\theta}_n = 1.846$ (and estimated their standard error) for datasets M1–M8 by two approaches: [1] using instrumental variable (IV); [2] using transformed x_1 with parametric (P) and nonparametric (NP) f_1 model. NA stands for not applicable due to numerical problems. All values are multiplied by 1,000.

	Approach				
	[1]			[2]	
IV	x_2	x_3	(x_2, x_3)	None	None
method	P	P	P	P	NP
M1	-8 (24)	16 (59)	14 (35)	-5 (23)	-6 (23)
M2	-73 (25)	-8 (23)	-25 (28)	-9 (23)	-4 (23)
M3	-25 (38)	NA (NA)	NA (NA)	-22 (23)	-20 (22)
M4	-19 (27)	13 (27)	-1 (24)	-4 (23)	5 (23)
M5	41 (206)	56 (373)	54 (563)	-10 (23)	-9 (23)
M6	23 (158)	-10 (31)	-57 (25)	-13 (24)	9 (23)
M7	26 (9500)	32 (7366)	36 (664)	10 (22)	-9 (23)
M8	50 (3985)	183 (NA)	128 (NA)	15 (25)	-18 (23)

encountered some numerical problems even for correctly specified models: M3, M5, and M7. This is because the effect of the instrumental variables on the outcome variable is not so strong; the instrumental variable used is not useful enough. In terms of efficiency, the methods with transformed x_1 outperform by far those using instrumental variables. They are also more robust against misspecification of the response model.

8. Discussion

We have presented a test statistic for model identification, semiparametric efficiency bound for $(\phi_0^{\mathrm{T}}, \theta_0)^{\mathrm{T}}$ under nonignorable nonresponse; proposed two types of adaptive semiparametric estimators that attain the semiparametric lower bound. Identification is a challenging problem in nonignorable nonresponse [18]; previous methods require nonignorable NIVs to guarantee model identification [36]. Our new identifiability condition is not on the response mechanism, but on the distribution of $[y \mid x, r = 1]$.

The proposed method is based on the correct specification of the response model. There may be various models for the true response mechanism, and thus the appropriate information criteria for choosing the response mechanism will be a topic of future research. Instead of specifying a single response model, one can consider multiple response models, and obtain consistency when one of the specified response models is correct. This multiple robustness property has been investigated in the ignorable nonresponse setup [10, 3]. Extension of multiple robustness to the nonignorable nonresponse case will also be a topic of our future research.

Appendix A: Regularity conditions

C1. Φ and Θ are compact.

C2. $W_i = (X_i, Y_i, R_i)$ are independently and identically distributed.

C3. The response probability $\pi(x_i, y_i)$ is bounded below. That is, $\pi(x_i, y_i) > K$ for some $K > 0$ for all $i = 1, \ldots, n$, uniformly in n.

C4. Γ is compact, $S_\gamma(\gamma) = \partial \log f_1(y \mid x; \gamma)/\partial\gamma$ is continuously differentiable at $\gamma \in \Gamma$ with probability one, there exists $e(W)$ such that $\|S_\gamma(\gamma)\| \le e(W)$ for all $\gamma \in \Gamma$ and $E\{e(W)\} < \infty$, $E\{S_\gamma(\gamma)\} = 0$ has a unique solution $\gamma^* \in \Gamma$, $\partial S_\gamma(\gamma)/\partial\gamma^{\mathrm{T}}$ is continuous at γ^* with probability one, and there is a neighborhood $\Gamma_{\mathcal{N}}$ of γ^* such that $\|E\{\sup_{\gamma\in\Gamma_{\mathcal{N}}} \partial S_\gamma(\gamma)/\partial\gamma^{\mathrm{T}}\}\| < \infty$.

C5. $E\{S_{\mathrm{eff}}(\phi, \theta, \gamma^*)\} = 0$ has a unique solution $(\phi_0, \theta_0) \in \Phi \times \Theta$, where $S_{\mathrm{eff}}(\phi, \theta, \gamma) = (S_1(\phi, \gamma)^{\mathrm{T}}, S_2(\phi, \theta, \gamma))^{\mathrm{T}}$ defined in (10).

C6. $\partial S_{\mathrm{eff}}(\phi, \theta, \gamma)/\partial(\phi^{\mathrm{T}}, \theta, \gamma^{\mathrm{T}})$ is continuous at $(\phi_0, \theta_0, \gamma^*)$ with probability one, and there is a neighborhood $\Phi_{\mathcal{N}} \times \Theta_{\mathcal{N}} \times \Gamma_{\mathcal{N}}$ of $(\phi_0, \theta_0, \gamma^*)$ such that

$$\|E\{\sup_{(\phi,\theta,\gamma^*)\in\Phi_{\mathcal{N}}\times\Theta_{\mathcal{N}}\times\Gamma_{\mathcal{N}}} \partial S_{\mathrm{eff}}(\phi, \theta, \gamma)/\partial(\phi^{\mathrm{T}}, \theta, \gamma^{\mathrm{T}})\}\| < \infty.$$

C7. $S_{\mathrm{eff}}(\phi, \theta, \gamma)$ is continuously differentiable at each $(\phi, \theta, \gamma) \in \Phi\times\Theta\times\Gamma$ with probability one, and there exists $d_1(W)$ such that $\|S_{\mathrm{eff}}(\phi, \theta, \gamma)\| \le d_1(W)$ for all $(\phi, \theta, \gamma) \in \Phi \times \Theta \times \Gamma$ and $E\{d_1(W)\} < \infty$.

C8. $E\{\partial S_{\mathrm{eff}}(\phi, \theta, \gamma^*)/\partial(\phi^{\mathrm{T}}, \theta, \gamma^{\mathrm{T}})\}$ is nonsingular at $(\phi_0, \theta_0, \gamma^*)$.

C9. The conditions (C5)-(C8) hold for known distribution $f_1(y \mid x; \gamma_0)$, i.e., $E\{S_{\mathrm{eff}}(\phi, \theta, \gamma_0)\} = 0$ has a unique solution $(\phi_0, \theta_0) \in \Phi \times \Theta$, where $S_{\mathrm{eff}}(\phi, \theta) = (S_1(\phi, \gamma_0)^{\mathrm{T}}, S_2(\phi, \theta, \gamma_0))^{\mathrm{T}}$; $\partial S_{\mathrm{eff}}(\phi, \theta)/\partial(\phi^{\mathrm{T}}, \theta)$ is continuous at (ϕ_0, θ_0) with probability one, and there is a neighborhood $\Phi_{\mathcal{N}} \times \Theta_{\mathcal{N}}$ of (ϕ_0, θ_0) such that

$$\|E\{\sup_{(\phi,\theta)\in\Phi_{\mathcal{N}}\times\Theta_{\mathcal{N}}} \partial S_{\mathrm{eff}}(\phi, \theta)/\partial(\phi^{\mathrm{T}}, \theta)\}\| < \infty;$$

$S_{\mathrm{eff}}(\phi, \theta)$ is continuously differentiable at each $(\phi, \theta) \in \Phi \times \Theta$ with probability one, and there exists $d_2(W)$ such that $\|S_{\mathrm{eff}}(\phi, \theta)\| \le d_2(W)$ for all $(\phi, \theta) \in \Phi\times\Theta$ and $E\{d_2(W)\} < \infty$; $E\{\partial S_{\mathrm{eff}}(\phi, \theta)/\partial(\phi^{\mathrm{T}}, \theta)\}$ is nonsingular at (ϕ_0, θ_0).

C10. Let $\mathcal{X} = [0, 1]$ be a compact set that is contained in the support of x, let $f_1(x) > 0$, and let $E_1\{\pi(x, Y; \phi_0) \mid x\} > 0$ for all $x \in \mathcal{X}$.

C11. The kernel $K(u)$ has bounded derivatives of order k, satisfies $\int K(u)\mathrm{d}u = 1$, has zero moments of order $\le m - 1$, and has a nonzero m-th order moment.

C12. For all y, $\pi(\cdot, y; \phi_0)$, $\dot{\pi}(\cdot, y; \phi_0)$, and $U(\cdot, y; \theta_0)$ are differentiable to order k and are bounded on an open set containing $\mathcal{X}$.

C13. Let $a_1(z) = 1$, $a_2(z) = s_0(z; \phi_0)$, and $a_3(z) = U(z)$. Then, there exists $v \ge 4$ such that $E_1\{|\pi^{-1}(Z; \phi_0)O(Z; \phi_0)a_i(Z)|^v\}$ and $E_1\{\|\pi^{-1}(Z; \phi_0)O(Z; \phi_0)a_i(Z)\|^v \mid x\}f_1(x)$ are bounded for all $x \in \mathcal{X}$.

C14. As $h \to 0$, $n^{1-(2/v)}h^d/\ln n \to \infty$, $n^{1/2}h^{d+2k}/\ln n \to \infty$, and $n^{1/2}h^{2m} \to 0$.

Sufficient conditions for uniqueness of ϕ in C5 are (i) the condition assumed in Theorem 3.1 and (ii) completeness for $\{1 - \pi(z; \phi_0)/\pi(z; \phi)\}\, g^\star(x; \phi, \gamma^*)$ with respect to ϕ, that is, it holds that

$$E\left[\left\{1-\frac{\pi(z;\phi_0)}{\pi(z;\phi)}\right\}g^\star(x;\phi,\gamma^*)\right]=0 \tag{A.1}$$

implies

$$\left\{1-\frac{\pi(z;\phi_0)}{\pi(z;\phi)}\right\}g^\star(x;\phi,\gamma^*)=0$$

for almost all z. The last equation means $\phi=\phi_0$. Therefore, the completeness condition (ii) assures, (A.1) does not hold unless $\phi=\phi_0$. Similar completeness conditions are also assumed in D'Haultfoeuille [4] and Yang et al. [37]. Sufficient conditions for uniqueness of θ are obtained in a similar way.

Appendix B: Proofs of the technical results

Proof of Theorem 3.1. Let γ be an infinite dimensional parameter of $f_1(y\mid x)$ and the true parameter be γ_0. Denote $\gamma=\gamma'$ if $f_1(y\mid x;\gamma)=f_1(y\mid x;\gamma')$ holds for almost all z. Here, the distribution of $[y\mid x]$ can be represented with (γ,ϕ), because by using Bayes' formula, we have

$$f(y\mid x;\gamma,\phi)=\frac{f_1(y\mid x;\gamma)\pi^{-1}(x,y;\phi)}{\int f_1(y\mid x;\gamma)\pi^{-1}(x,y;\phi)dy}. \tag{B.1}$$

We give a proof for Theorem 3.1 by proving the identification condition for $f(y\mid x;\gamma,\phi)\pi(x,y;\phi)$ is equivalent to that for $\int f_1(y\mid x;\gamma)\pi^{-1}(x,y;\phi)dy$ and the uniqueness of $f_1(y\mid x;\gamma)$. It follows from the definition of $O(x,y;\phi)=\pi^{-1}(x,y;\phi)-1$ that the identification of $\int f_1(y\mid x;\gamma)\pi^{-1}(x,y;\phi)dy$ and $\int f_1(y\mid x;\gamma)O(x,y;\phi)dy$ are equivalent. Therefore, we have only to show that

$$f(y\mid x;\gamma,\phi)\pi(x,y;\phi)=f(y\mid x;\gamma',\phi')\pi(x,y;\phi') \tag{B.2}$$

is equivalent to

$$f_1(y\mid x;\gamma)=f_1(y\mid x;\gamma') \tag{B.3}$$

and

$$\int f_1(y\mid x;\gamma)\pi^{-1}(x,y;\phi)dy=\int f_1(y\mid x;\gamma')\pi^{-1}(x,y;\phi')dy \tag{B.4}$$

It follows from (B.1) that

$$f(y\mid x;\gamma,\phi)\pi(x,y;\phi)=\frac{f_1(y\mid x;\gamma)}{\int f_1(y\mid x;\gamma)\pi^{-1}(x,y;\phi)dy}. \tag{B.5}$$

Hence, (B.3) and (B.4) imply (B.2). On the contrary, by taking integration with respect to y in (B.2) by using (B.5), we have (B.4) and then, (B.3) follows from (B.4) and (B.5). ☐

Proof of Theorem 3.2. If the null hypothesis $H_0 : K_0^{(\ell)}(x) = c_1^{(\ell)}(0)+c_2^{(\ell)}(0)x,\ \ell = 1,\dots,L$ is correct, the conditional expectation of $\sum_{\ell=1}^{L} Y^\ell$ given x is given as $r_1(x) := \sum_{\ell=1}^{L}\mu_\ell(x)$. Let $\bar{r}_1$ be $\int r_1(x)dx$. Because $\mu_\ell(x)$ is the ℓ-th order polynomial, denote $\sum_{\ell=0}^{L}\mu_\ell(x)$ by $\sum_{\ell=0}^{L} a_\ell x^\ell$. By the result of Hart [11, §8.3], the asymptotic distribution of the test statistic S_n is

$$S_n \xrightarrow{\mathcal{L}} \sup_{m\geq 1}\frac{1}{m}\sum_{j=1}^{m} Z_j^2,$$

as $n\to\infty$, where $Z_1, Z_2, \dots,$ be a Gaussian process with $E(Z_j) = 0, j \geq 1$, and covariances

$$\mathrm{Cov}(Z_i, Z_j) = \begin{cases} 1-\beta_i^2, & i = j \\ -\beta_i\beta_j, & i \neq j \end{cases},$$

and

$$\beta_j = \frac{\sqrt{2}\int_0^1 r_1(x)\cos(\pi j x)dx}{\left\{\int_0^1 (r_1(x)-\bar{r}_1)^2dx\right\}^{1/2}}, \quad (j = 1,\dots,n-1).$$

For the regression model $r_1(x) = \sum_{\ell=0}^{L} a_\ell x^\ell$, the β's have an easier form. Let

$$A_j^\ell = \frac{\sin(\pi j) - kB_j^{\ell-1}}{\pi j}, \quad B_j^\ell = -\frac{\cos(\pi j) - kA_j^{\ell-1}}{\pi j},$$

where $A_j^0 = \sin(\pi j)/\pi j$ and $B_j^0 = \{1-\cos(\pi j)\}/(\pi j)$. With A_j^ℓ, B_j^ℓ $(j = 1,\dots,m;\ \ell = 0,1,\dots,L)$, it can be written by

$$\beta_j = \frac{\sum_{\ell=0}^{L} a_\ell A_j^\ell}{\left(\sum_{k=1}^{L}\sum_{\ell=1}^{L}\frac{a_k a_\ell k\ell}{2(k+\ell+1)(k+1)(\ell+1)}\right)^{1/2}}.$$

□

Next, we provide a proof of Lemma 4.1 and Theorem 5.1 and 5.2. In order to prove Lemma 4.1, we will assume $U(z) = y$ for simplicity. We specify the joint distribution $z = (x^{\mathrm{T}}, y)^{\mathrm{T}}$ by $f(z;\eta)$, where η is an infinite-dimensional nuisance parameter, and η_0 is the true value. By "full model" we refer to the class of models in which the data are completely observed, and by "obs model" we refer to those in which some Y are missing; that is, a full model consists of functions $h(Z)$ and an obs model consists of $h(R, G_R(Z))$. Furthermore, for each full and obs model, denote the nuisance tangent space by Λ^F and Λ, respectively, and its orthogonal complement by $\Lambda^{F\perp}$ and $\Lambda^\perp$, respectively. Let S_ϕ be the score function with respect to ϕ. Consider a Hilbert space $\mathcal{H} = \{h^{(q+1)\times 1} \mid E(h) = 0; \|h\| < \infty\}$ with inner product $\langle h_1, h_2\rangle = E(h_1^{\mathrm{T}} h_2)$, where the expectation is taken under the true model. See Bickel et al. [1] and Tsiatis [34] for more details. When U is comprised of other functions, the proof is almost the same.

At first, we introduce a proposition of Rotnitzky & Robins [25], which provides the efficient score for (ϕ, θ), as follows. Let B and D be functions of

$(R, G_R(Z))$, and let B^* and D^* be functions of Z. Also, let us define the following three linear operators: $\mathrm{g}(B^*) = E(B^* \mid R, G_R(Z))$, $\mathrm{m}(B^*) = E\{\mathrm{g}(B^*) \mid Z\}$, and $\mathrm{u}(B^*) = RB^*/\pi(Z)$. Then, the efficient score for (ϕ, θ) can be derived by the following Lemma. See Proposition A1 in Rotnitzky & Robins [25] for the proof.

Lemma B.1. *The efficient score for (ϕ, θ) can be written as*

$$S_{\text{eff}} = \mathrm{u}(D^*_{\text{eff}}) - \Pi[\mathrm{u}(D^*_{\text{eff}}) \mid \Lambda_2] + A_{2,\text{eff}} = \mathrm{g}\{\mathrm{m}^{-1}(D^*_{\text{eff}})\} + A_{2,\text{eff}}, \tag{B.6}$$

where $\Pi[h \mid \Lambda_2]$ is the projection of h onto Λ_2, $\Lambda_2 = [h(R, G_R(Z)) : E(h(R, G_R(Z)) \mid Z) = 0]$, and D^*_{eff} is a unique solution to

$$\Pi[\mathrm{m}^{-1}(D^*) \mid \Lambda^{F\perp}] = (Q, S^{F\perp}_{\text{eff},\theta}), \tag{B.7}$$

where $Q = \Pi[\mathrm{m}^{-1}[E\{\mathrm{g}(S^F_\phi) \mid L\}] \mid \Lambda^{F\perp}]$, $A_{2,\text{eff}} = (\Pi[S_\phi \mid \Lambda_2]^{\mathrm{T}}, 0)^{\mathrm{T}} = (\mathrm{g}(S^F_\phi) - \mathrm{g}[\mathrm{m}^{-1}[E\{\mathrm{g}(S^F_\phi) \mid L\}]]^{\mathrm{T}}, 0)^{\mathrm{T}}$, and $S^{F\perp}_{\text{eff},\theta}$ is the efficient score function of θ in the full model.

This Lemma implies that the efficient score can be represented by (B.6) with D^*_{eff} satisfying condition (B.7). Thus, in the nonignorable nonresponse case, $\Lambda^{F\perp}$ needs to be calculated, and it can be done in a way similar to that shown in Section 4.5 of Tsiatis [34].

Lemma B.2. *The nuisance tangent space Λ^F and its orthogonal complement $\Lambda^{F\perp}$ in the full model are written as follows:*

$$\begin{aligned} \Lambda^F &= [h(Z) \in \mathcal{H} \text{ such that } E\{Yh(Z)\} = 0], \\ \Lambda^{F\perp} &= [k(Y - \theta_0), \text{where } k \text{ is any } q+1 \text{ dimensionalvector}]\,. \end{aligned}$$

Finally, we give an explicit formula to calculate the projection onto Λ_2.

Lemma B.3. *For $h(R, G_R(Z)) = Rh_1(Z) + (1 - R)h_2(X)$, it holds that*

$$\Pi(h \mid \Lambda_2) = \left\{1 - \frac{R}{\pi(Z)}\right\} \frac{h_2(X) - E(h_1(Z) \mid X)}{E\{O(Z) \mid X\}}. \tag{B.8}$$

Proof of Lemma B.3. Obviously, the right-hand side of (B.8) belongs to Λ_2. Thus, it remains to check that for any g,

$$\left\langle h - \left\{1 - \frac{R}{\pi(Z)}\right\} \frac{h_2(X) - E\{h_1(Z) \mid X\}}{E\{O(Z) \mid X\}}, \left\{1 - \frac{R}{\pi(Z)}\right\} g(X) \right\rangle = 0,$$

which can be proved easily. □

We now give a proof of Lemma 4.1.

Proof of Lemma 4.1. Note that $S^{F\perp}_{\text{eff},\theta} = Y - \theta_0$ by Lemma B.2, since there exists only one influence function, and it is the efficient one under the assumption that θ does not require any assumptions on the distribution of Z [see 34, Chap. 5]. By the projection theorem, there exists a unique $k = (k_1, k_2^{\mathrm{T}})^{\mathrm{T}}$ such that $D^*_{\text{eff}} = k(Y - \theta_0)$.

Then, we calculate $A_{2,\mathrm{eff}}$. The score function of ϕ is

$$S_\phi = \mathrm{g}(S_\phi^F) = Rs_1(Z;\phi) + (1-R)s_0(X;\phi),$$

where $s_r(\phi)$ is defined in (3). It follows from Lemma B.3 with $h_1(z) = s_1(\phi)$ and $h_2(x) = \bar{s}_0(x;\phi)$ in (B.8) that $\Pi(S_\phi \mid \Lambda_2) = -\{1 - R/\pi(Z)\}g^\star(X)$. Thus, $A_{2,\mathrm{eff}} = [0, -\{1-R/\pi(Z)\}g^\star(X)]$. Again, by using Lemma B.3, it follows that $\Pi[\mathrm{u}(D^*_{\mathrm{eff}}) \mid \Lambda_2] = -\{1-R/\pi(Z)\}E^\star(Y-\theta_0 \mid X)$, by which (B.6) becomes

$$S_1 = k_2\left[\frac{R(Y-\theta_0)}{\pi(\phi_0)} + \left\{1 - \frac{R}{\pi(\phi_0)}\right\}E^\star(Y-\theta_0 \mid X)\right] - \left\{1 - \frac{R}{\pi(Z)}\right\}g^\star(X)$$

and

$$S_2 = k_1\left[\frac{R(Y-\theta_0)}{\pi(\phi_0)} + \left\{1 - \frac{R}{\pi(\phi_0)}\right\}E^\star(Y-\theta_0 \mid X)\right].$$

This $S_{\mathrm{eff}} = (S_1, S_2^{\mathrm{T}})$ can be transformed into $\tilde{S}_{\mathrm{eff}} = (\tilde{S}_1, \tilde{S}_2^{\mathrm{T}}) = AS_{\mathrm{eff}}$,

$$\begin{aligned}\tilde{S}_1 &= \left\{1 - \frac{R}{\pi(\phi_0)}\right\}g^\star(X),\\ \tilde{S}_2 &= \frac{R(Y-\theta_0)}{\pi(\phi_0)} + \left\{1 - \frac{R}{\pi(\phi_0)}\right\}E^\star(Y-\theta_0 \mid X)\end{aligned}$$

with a nonsingular matrix A,

$$A = \begin{bmatrix} -I_q & -k_2^{\mathrm{T}}/k_1 \\ 0^{\mathrm{T}} & k_1^{-1}\end{bmatrix},$$

where I_q is a q-dimensional identity matrix. The score function multiplied by a nonsingular constant matrix does not have an influence on the asymptotic distribution. Thus, we have the desired efficient score. □

Proof of Theorem 5.1. Consistency and asymptotic normality are proved under the assumptions (C1)-(C8) by using the standard argument for GMM. Next, we give the explicit form of the asymptotic variance. Let $\xi = (\phi^{\mathrm{T}}, \theta)^{\mathrm{T}}$. Recall that each $\hat{\gamma}$ and $\hat{\xi}$ is a solution to $\sum_{i=1}^n \partial \log f_1(y_i \mid x_i;\gamma)/\partial\gamma = \sum_{i=1}^n S_{\gamma i}(\gamma) = 0$ and $\sum_{i=1}^n S_{\mathrm{eff},i}(\hat{\gamma},\xi) = 0$, respectively, where $S_{\mathrm{eff},i}(\gamma,\xi)$ is defined in (10). By using standard asymptotic theory,

$$\begin{bmatrix}\hat{\gamma} - \gamma^* \\ \hat{\xi} - \xi_0\end{bmatrix} = -\mathcal{I}^{-1}n^{-1}\sum_{i=1}^n \begin{bmatrix} S_{\gamma i}(\gamma^*) \\ S_{\mathrm{eff},i}(\gamma^*,\xi_0)\end{bmatrix},$$

where

$$\begin{aligned}\mathcal{I} &= E\begin{bmatrix}\partial S_\gamma(\gamma^*)/\gamma^{\mathrm{T}} & \partial S_\gamma(\gamma^*)/\xi^{\mathrm{T}} \\ \partial S_{\mathrm{eff}}(\gamma^*,\xi_0)/\gamma^{\mathrm{T}} & \partial S_{\mathrm{eff}}(\gamma^*,\xi_0)/\xi^{\mathrm{T}}\end{bmatrix}\\ &= E\begin{bmatrix}\partial S_\gamma(\gamma^*)/\gamma^{\mathrm{T}} & O \\ \partial S_{\mathrm{eff}}(\gamma^*,\xi_0)/\gamma^{\mathrm{T}} & \partial S_{\mathrm{eff}}(\gamma^*,\xi_0)/\xi^{\mathrm{T}}\end{bmatrix}.\end{aligned}$$

Let the (i,j) block of $\mathcal{I}$ be $\mathcal{I}_{ij}$. Then,

$$\mathcal{I}^{-1} = \begin{bmatrix} \mathcal{I}_{11}^{-1} & O \\ -\mathcal{I}_{22}^{-1}\mathcal{I}_{21}\mathcal{I}_{2}^{-1} & \mathcal{I}_{22}^{-1} \end{bmatrix}.$$

Here, it follows that $\mathcal{I}_{21} = O$ because

$$E\left[\left\{1 - \frac{R}{\pi(\phi_0)}\right\}\frac{\partial g^\star(\gamma^*, \xi_0)}{\partial \gamma^{\mathrm{T}}}\right] = O$$

and

$$E\left[\left\{1 - \frac{R}{\pi(\phi_0)}\right\}\frac{\partial U^\star(\gamma^*, \xi_0)}{\partial \gamma^{\mathrm{T}}}\right] = 0^{\mathrm{T}}.$$

Therefore, we have,

$$\mathcal{I}^{-1} = \begin{bmatrix} \mathcal{I}_{11}^{-1} & O \\ O & \mathcal{I}_{22}^{-1} \end{bmatrix}.$$

By applying exactly the same arguments for $\mathcal{I}_{22}^{-1}$ used for $\mathcal{I}^{-1}$, we got the asymptotic variance of $\hat{\theta}$ as given in (11). □

Proof of Theorem 5.2. Consistency and asymptotic normality of our proposed estimator are similar to proving Lemma 4.1 of Morikawa et al. [20]. We herein show our estimator attains the semiparametric lower bound derived in Lemma 4.1. Let $f_1(x)$ be the conditional distribution of $[x \mid r = 1]$. From the same arguments that were used to prove Lemma A.1 in Morikawa et al. [20], it can be shown that the estimating equation in Theorem 5.2, $\hat{S}_{\mathrm{eff}}(\phi, \theta) = \{\hat{S}_1(\phi)^{\mathrm{T}}, \hat{S}_2(\phi, \theta)\}^{\mathrm{T}}$ is expanded as

$$\hat{S}_1(\phi) = n^{-1}\sum_{i=1}^{n}\left[\left\{1 - \frac{r_i}{\pi(\phi; z_i)}\right\} g^\star(\phi; x_i) + r_i G(z_i; \phi)\right] + o_p(n^{-1/2})$$

$$\hat{S}_2(\phi, \theta) = n^{-1}\sum_{i=1}^{n}\left[\frac{r_i}{\pi(\phi; z_i)} U(\theta; z_i) + \left\{1 - \frac{r_i}{\pi(\phi; z_i)}\right\} U^\star(\theta, \phi; x_i) + r_i H(\theta, \phi; z_i)\right] + o_p(n^{-1/2}),$$

where $G(\phi; z_i) = G_1(\phi; x_i) G_2(\phi; z_i)$, $H(\theta, \phi; z_i) = G_1(\phi; x_i) H_2(\theta, \phi; z_i)$, and

$$G_1(\phi; x_i) = 1 - E\left\{\frac{\pi(\phi_0; Z)}{\pi(\phi; Z)} \,\middle|\, x_i\right\},$$

$$G_2(\phi; z_i) = \frac{\pi^{-1}(\phi; z_i) O(\phi; z_i)\{s_0(\phi; z_i) - g^\star(\phi; x_i)\}}{E_1\{\pi^{-1}(\phi; Z) O(\phi; Z) \mid x_i\} P(R = 1 \mid x_i)},$$

$$H_2(\theta, \phi; z_i) = \frac{\pi^{-1}(\phi; z_i) O(\phi; z_i)\{U(\theta; z_i) - U^\star(\theta, \phi; x_i)\}}{E_1\{\pi^{-1}(\phi; Z) O(\phi; Z) \mid x_i\} P(R = 1 \mid x_i)}.$$

Therefore, the asymptotic variance may increase due to the additional terms $rG(\phi)$ and $rH(\phi)$, but this solution also attains the lower bound. At first, we focus on the estimator for ϕ. Once we get an unbiased estimating equation

$\sum_{i=1}^{n} \varphi(z_i; \phi) = 0$, the asymptotic variance can be given as $\text{Var}\{E(\dot{\varphi}(\phi_0))^{-1}\varphi(\phi_0)\}$, where $\dot{\varphi}(\phi_0) = \partial\varphi(\phi_0)/\partial\phi^{\text{T}}$. Thus, for the proving purpose, it suffices to show that $G(\phi_0) = 0$ and $E(R\dot{G}(\phi_0)) = O$. The former equation is trivial, so we only need to work on the latter equation, which can be written as $E(R\dot{G}(\phi_0)) = E(RG_1(\phi_0)\dot{G}_2(\phi_0)) + E(RG_2(\phi_0)\dot{G}_1(\phi_0))$. The first term is zero from $G_1(\phi_0) = 0$. Also, the second term is $E(RG_2(\phi_0)\dot{G}_1(\phi_0)) = E\{E(RG_2(\phi_0) \mid X)\dot{G}_1(\phi_0)\} = O$. Hence, the last equation holds by the definition of $g^\star(\phi; x)$. Therefore, $rG(\phi)$ has no effect on the asymptotic variance and our estimator also attains the semiparametric efficiency bound. The same conclusion can be made when estimating θ. □

References

[1] Bickel, P. J., Klaassen, C. A. J., Ritov, Y., & Wellner, J. A. (1998). *Efficient and Adaptive Estimation for Semiparametric Models.* New York: Springer-Verlag. MR1623559

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

[3] Chen, S. & Haziza, D. (2017). Multiply robust imputation procedures for the treatment of item nonresponse in surveys. *Biometrika* **104**, 439–453. MR3698264

[4] D'Haultfoeuille, X. (2010). A new instrumental method for dealing with endogeneous selection. *J. Econometrics* **154**, 1–15. MR2558947

[5] Diggle, P. & Kenward, M. G. (1994). Informative drop-out in longitudinal data analysis. *J. R. Statist. Soc.* C **43**, 49–93.

[6] Eubank, R. L. & Hart, J. D. (1992). Testing Goodness-of-fit in regression via order selection criteria. *Ann. Statist.* **20**, 1412–1425. MR1186256

[7] Fitzmaurice, G. M., Lipsitz, S. R., Molenberghs, G. & Ibrahim, J. G. (2005). A protective estimator for longitudinal binary data subject to non-ignorable non-monotone missingness. *J. R. Statist. Soc.* A **168**, 723–735. MR2205403

[8] Greenlees, J. S., Reece, W. S. & Zieschang, K. D. (1982). Imputation of missing values when the probability of response depends on the variable being imputed. *J. Am. Statist. Assoc.* **77**, 251–261.

[9] Hájek, J. (1970). A characterization of limiting distributions of regular estimates. *Z. Wahrscheinlichkeitstheorie verw. Gebiete.* **14**, 323–330. MR0283911

[10] Han, P. (2014). Multiply robust estimation in regression analysis with missing data. *J. Am. Statist. Assoc.* **109**, 1159–1173. MR3265688

[11] Hart, J D. (1997). *Nonparametric Smoothing and Lack-of-fit tests.* New York: Springer-Verlag. MR1461272

[12] Kim, J. K. (2011). Parametric fractional imputation for missing data analysis *Biometrika* **98**, 119–132. MR2804214

[13] Kim, J. K. & Yu, C. L. (2011). A semiparametric estimation of mean functionals with nonignorable missing data. *J. Am. Statist. Assoc.* **106**, 157–165. MR2816710

[14] Liang, K.-Y. & Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models *Biometrika* **73**, 13–22. MR0836430

[15] Little, R. J. A. & Rubin, D. B. (2002). *Statistical Inference with Missing Data. Second edition.* New York: Wiley. MR1925014

[16] Louis, T. A. (1982). Finding the observed information matrix when using the EM algorithm. *J. R. Statist. Soc.* B **44**, 226–233. MR0676213

[17] Ma, W. Q., Geng, Z. & Hu, Y. H. (2003). Identification of graphical models for nonignorable nonresponse of binary outcomes in longitudinal studies. *J. Multivar. Anal.* **87**, 24–45. MR2007260

[18] Miao, W., Ding, P. & Geng, Z. (2016). Identifiability of normal and normal mixture models with nonignorable missing data. *J. Am. Statist. Assoc.* **111**, 1673–1683. MR3601726

[19] Molenberghs, G., Beunckens, C., Sotto, C. & Kenward, M. G. (2008). Every missingness not at random model has a missingness at random counterpart with equal fit. *J. R. Statist. Soc.* B **70**, 371–388. MR2424758

[20] Morikawa, K., Kim, J. K. & Kano, Y. (2017). Semiparametric maximum likelihood estimation with data missing not at random. *Canad. J. Statist.* **45**, 393–409. MR3729977

[21] Qin, J., Leung, D. & Shao, J. (2002). Estimation with survey data under nonignorable nonresponse or informative sampling. *J. Am. Statist. Assoc.* **97**, 193–200. MR1947279

[22] Riddles, M. K., Kim, J. K. & Im, J. (2016). Propensity-score-adjustment method for nonignorable nonresponse. *J. Surv. Stat. Methodol.* **97**, 215–245.

[23] Robins, J. M., Rotnitzky, A. & Zhao, L. P. (1994). Estimation of regression coefficients when some regressors are not always observed. *J. Am. Statist. Assoc.* **89**, 846–866. MR1294730

[24] Robins, J. M., Rotnitzky, A. & Scharfstein, D. O. (2000). *Sensitivity Analysis for Selection Bias and Unmeasured Confounding in Missing Data and Causal Inference Models. In Statistical Models in Epidemiology: The Environment and Clinical Trials.* New York: Springer-Verlag, 1–92. MR1731681

[25] Rotnitzky, A. & Robins, J. M. (1997). Analysis of semi-parametric regression models with non-ignorable non-response. *Stat. Med.* **16**, 81–102.

[26] Rotnitzky, A., Scharfstein, D., Su, T.-L. & Robins, J. M. (2001). Methods for conducting sensitivity analysis of trials with potentially nonignorable competing causes of censoring. *Biometrics* **57**, 103–113. MR1833295

[27] Rubin, D. B. (1976). Inference and missing data. *Biometrika* **61**, 581–592. MR0455196

[28] Scharfstein, D. O., Rotnitzky, A. & Robins, J. M. (1999). Adjusting for nonignorable drop-out using semiparametric nonresponse models. *J. Am. Statist. Ass.* **94**, 1096–1146. MR1731478

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

[30] Skrondal, A. & Rabe-Hesketh, S. (2014). Protective estimation of mixed-effects logistic regression when data are not missing at random. *Biometrika* **101**, 175–188. MR3180664

[31] Sverchkov, M. (2014). A new approach to estimation of response prob-

abilities when missing data are not missing at random. *In Proc. Survey Res. Meth. Sect., Am. Statist. Ass.* Washington DC: American Statistical Association, 867–874.

[32] Tang, G., Little, R. J. A. & Raghunathan, T. E. (2003). Analysis of multivariate missing data with nonignorable nonresponse. *Biometrika* **90**, 747–764. MR2024755

[33] Tang, N., Zhao, P. & Zhu, H. (2014). Empirical likelihood for estimating equations with nonignorably missing data. *Statist. Sinica* **24**, 723–747. MR3235396

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

[35] Verbeke, G., Molenberghs, G., Thijs, H., Lesaffre, E. & Kenward, M. G. (2002). Sensitivity analysis for nonrandom dropout: A local influence approach. *Biometrics* **57**, 7–14. MR1833286

[36] Wang, S., Shao, J. & Kim, J. K. (2014). An instrumental variable approach for identification and estimation with nonignorable nonresponse. *Statist. Sinica* **24**, 1097–1116. MR3241279

[37] Yang, S., Wang, L. & Ding, P. (2017). Nonparametric identification of causal effects with confounders subject to instrumental missingness. *arXiv:1702.03951v1.*

[38] Zhao, J. & Shao, J. (2015). Semiparametric pseudo-likelihoods in generalized linear models with nonignorable missing data. *J. Am. Statist. Ass.* **110**, 1577–1590. MR3449056

[39] Zhao, P., Tang, N., Qu, A. & Jiang, D. (2017). Semiparametric estimating equations inference with nonignorable missing data. *Statist. Sinica* **27**, 89–113. MR3618161