SEMIPARAMETRIC OPTIMAL ESTIMATION WITH NONIGNORABLE NONRESPONSE DATA

BY KOSUKE MORIKAWA[1] AND JAE KWANG KIM[2]

[1]*Graduate School of Engineering Science, Osaka University, morikawa@sigmath.es.osaka-u.ac.jp*

[2]*Department of Statistics, Iowa State University, jkim@iastate.edu*

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 parametric model assumptions on the response mechanism. In this paper, a new statistical test is proposed to guarantee model identifiability without using instrumental variable assumption. 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.

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) [32]. 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) [19]. 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. [11] and Diggle and Kenward [7] proposed a fully parametric approach to analyze nonignorable nonresponse data; their method requires two parametric models: (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 [31, 33, 41, 42]. Sverchkov [37] and Riddles et al. [27] 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 error.

Recently, several semiparametric approaches have been proposed for nonignorable nonresponses. Ma et al. [22] studied identification and parameter estimation for binary study variables. Tang et al. [38] 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 [6] also used the same instrumental variable assumption and considered a regression analysis using the nonparametric propensity score model. Zhao and Shao [44] extended the method of Tang et al. [38] and relaxed the condition on the instrumental variable, which is called nonresponse instrumental variable [43]. Fitzmaurice et al. [10] and Skrondal and Rabe-Hesketh [36] 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 and Yu [17] 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. [39] used the method of empirical likelihood to extend the method of Kim and Yu [17] to estimate more general parameters. In Zhao et al. [45], the method of Qin et al. [26] was used to construct a $n^{1/2}$-consistent estimator without a validation sample. Morikawa et al. [25] used the kernel regression estimator to remove the parametric model assumption on model (i) $[y \mid x, r = 1]$. Chang and Kott [3] and Wang et al. [43] 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 and Wang [35] proposed a semiparametric inverse propensity weighting method using the nonresponse instrumental variable (NIV) assumption of Wang et al. [43]. 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 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 mechanism and for the parameter of interest such as population mean. Efficiency under this setup has already been discussed by Rotnitzky and Robins [30] and Robins et al. [28]. However, their estimator requires many working models to achieve the semiparametric efficiency bound. Misspecification of the working models may lead to loss of efficiency. See the simulation study in Section 6 and real data analysis in Section 7 for comparison with the method of Rotnitzky and Robins [30].

Therefore, we consider an alternative approach and propose two semiparametric estimators that attain the semiparametric lower bound [2] (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 $z = (x^T, y)^T$, x is a d-dimensional covariate vector, y is a response variable and r is a response indicator of y, that is, 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, that is, $G_1(z) = z$ 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 a parameter satisfying $E\{U(Z; \theta)\} = 0$, where U is a known function of z. For example, if we are interested in $E(Y)$, then $U(z; \theta) = y - \theta$, and in regression coefficients $E(Y \mid x) = \mu(x; \theta)$, then $U(z; \theta) = a(x)\{y - \mu(x; \theta)\}$, where $a(\cdot)$ is any linearly independent function of x having same dimension as θ. In this paper, we consider semiparametric estimation of (ϕ, θ) from partial observations. In particular, we propose the efficient estimator among the regular

asymptotically linear estimators [2, 41] without specifying the correct outcome model. We assume that the parameter ϕ is distinct from θ in the sense that they are independent a priori [32]. Therefore, we consider a setting where θ and ϕ are irrelevant to each other, and propose two adaptive estimators for (ϕ, θ).

For model identification for a response model, Miao et al. [23] 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. [43] developed a theory for identification by assuming that there exists a NIV x_2 in the covariate vector $x = (x_1^T, x_2^T)^T$ 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 are not testable with observed data. In Section 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 [11]. This requirement can be challenging because the specification cannot be verified under nonignorable nonresponse [24]. Chang and Kott [3] proposed a semiparametric estimator for ϕ based on the following estimating equation:

$$(2.1) \qquad \sum_{i=1}^{n} \Gamma(x_i, y_i, r_i; \phi) = \sum_{i=1}^{n} \left\{ 1 - \frac{r_i}{\pi(z_i; \phi)} \right\} g(x_i; \phi) = 0,$$

where $g = \{g_1(x), g_2(x), \ldots g_q(x)\}^T$, 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.

Riddles et al. [27] proposed an efficient estimator that uses a parametric model for $[y \mid x, r = 1]$. Using the mean score theorem [21], the maximum likelihood estimator can be obtained by solving

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

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

$$(2.2) \qquad s_r(z; \phi) = \frac{\{r - \pi(z; \phi)\} \dot{\pi}(z; \phi)}{\pi(z; \phi)\{1 - \pi(z; \phi)\}},$$

$\dot{\pi}(z; \phi) = \partial \pi(z; \phi)/\partial \phi$, 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. [27] proposed using

$$(2.3) \qquad \sum_{i=1}^{n} \left[r_i s_1(z_i; \phi) + (1 - r_i) \frac{E_1 \{ O(Z; \phi) s_0(Z; \phi) \mid x_i \}}{E_1 \{ O(Z; \phi) \mid x_i \}} \right] = 0,$$

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, misspecification of the f_1 model could cause the solution $\hat{\phi}$ to be inconsistent. Morikawa et al. [25] 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 a new identification condition for estimation of the response model with observed data. Our idea is to define the target parameter ϕ_0 as a unique solution to

$$E\{\Gamma(Z, R; \phi) \mid X\} = 0 \quad \text{a.s.,} \tag{3.1}$$

where Γ is defined in (2.1), though natural definition of the parameter might be through either (i) $E\{\Gamma(Z, R; \phi) \mid Z\} = 0$ or (ii) $E\{\Gamma(Z, R; \phi)\} = 0$. Note that providing a sufficient condition for the parameter defined in (ii) is the strongest (and in (i) is the weakest) since $E\{\Gamma(Z, R; \phi)\} = E[E\{\Gamma(Z, R; \phi) \mid Z\}]$ and $E\{\Gamma(Z, R; \phi) \mid X\} = E[E\{\Gamma(Z, R; \phi) \mid Z\} \mid X]$ hold. This implies a sufficient condition for the parameter (3.1) does not necessarily guarantee the model identification of (ii), which is the probability limit of the estimating equation (2.1). However, it does not matter because even if we face such a problem, it can be solved by constructing an objective function with the integrated regression function [8]. Therefore, we focus on providing a sufficient condition of the model identification for the parameter defined in (3.1).

3.1. *Identification condition with f_1 model.* 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 true conditional expectation given x and $r = 1$. A new identification condition for the semiparametric model is given in the following theorem.

THEOREM 3.1. *The identification condition for a parameter* (3.1) *holds under the following conditions*:

(I1) $E_1\{O(Z; \phi)|x\}$ *exists and is bounded almost surely*;

(I2) *The weight function g in* (2.1) *satisfies* $P(\inf_{\phi \in \Phi} |g(X; \phi)| > 0) > 0$, *and elements of $g(x; \phi)$ are linearly independent functions with respect to x for all ϕ*;

(I3) $E_1\{O(Z; \phi) \mid x\} = E_1\{O(Z; \phi') \mid x\}$ *a.s. implies* $\phi = \phi'$.

As for the condition (I1), it can be shown that the probit models do not satisfy the condition, but the logistic and the robit models [20, 23], which are cumulative functions of t-distributions, meet the condition. Although the probit model is not applicable, Theorem 3.1 is practically useful because the performance with the probit and logistic model is very similar, and thus, misspecification of the response model is not a serious problem in practice (see Section 7 for the performance with misspecified response models). Condition (I2) is required to avoid g becomes identically zero.

The key condition is (I3), which implies that we should check the identification of $E_1\{O(Z; \phi) \mid x\}$. 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 $\phi = (\phi_{x0}, \phi_{x1}, \phi_y)^{\mathrm{T}}$. Then $E_1\{O(Z; \phi)|x\}$ is written as

$$E_1\{O(Z; \phi)|x\} = \exp\{\phi_{x0} + \phi_{x1}x + K_{\phi_y}(x)\}, \tag{3.2}$$

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 only need to verify that $K_{\phi_y}(x)$ is not linear with respect to x. 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 first- and second-order function of x such as log-function (e.g., Gamma distribution), exponential-function (e.g., Poisson distribution), etc. However, when b is a second-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

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

Thus, by Theorem 3.1, ϕ is identifiable unless the mean structure $\tau(x)$ is linear since there are three parameters with two equations. If $\tau(x)$ is linear, we may use a transformation approach which is introduced in Section 3.3.

On the other hand, checking the model identifiability with a nonprametric $f_1(y \mid x)$ model is still challenging, because there is no way to estimate the cumulative function $K_{\phi_y}(x)$ nonparametrically for every ϕ_y up to our knowledge. Therefore, we propose a test statistic to test a reasonable necessary condition for the identification condition.

3.2. *Nonparametric test statistics.*

In view of (3.2), the model is unidentifiable when the cumulant-generating function is linear with respect to x for all ϕ_y. In this section, we show that linearity of the cumulant-generating function implies linearity of the mean function in a data generation process, and propose a simple nonparametric test statistics for identifiability by checking the linearity of the mean function. Let a general data-generating process be $y = \mu(x) + \varepsilon(x)$, where $\mu(x)$ is the conditional expectation of y given x, and $\varepsilon(x)$ is the conditional mean-zero error. Consider a class of error functions $\mathcal{E}$: for $\varepsilon \in \mathcal{E}$, $\varepsilon(x) = \sum_{j=0}^{\infty} \xi_j e_j(x)$, where ξ_j $(j \geq 0)$ are mean-zero random variables, which are independent of x, and e_j $(j \geq 0)$ are any measurable functions of x satisfying $E[\{\sum_{j=0}^{\infty} |\xi_j e_j(X)|\}^k] < \infty$ for any positive integer k, and $e_l \neq e_m$ for $l \neq m$. This class of error functions include many functions with mean-zero conditional expectation such as the infinite normal mixture distribution. Under this setup, we can show the following proposition.

PROPOSITION 3.1. *Suppose that $\varepsilon \in \mathcal{E}$, then linearity of the conditional cumulant-generating function $K_{\phi_y}(x)$ implies that of the conditional mean function $\mu(x)$.*

For the above reasons, we test a data-generation structure

$$H_0 : y = \mu(x; c) + \varepsilon, \tag{3.3}$$

where $\mu(x; c) = c_1 + c_2^\top x$, $c = (c_1, c_2^\top)^\top$, and ε is a mean-zero random variable and independent of x. Denoting the true conditional mean function by $m(x) = E(Y \mid X = x)$, the null hypothesis can be also represented by H_0: there exists c such that $m(x) = \mu(x; c)$. It is desirable that the statistical test enjoys two properties: (i) dimension-free for x; (ii) no parametric assumption on ε. The first property is practically useful because classical nonparametric tests such as Eubank and Hart [9] suffer from curse of dimensionality. The second property can avoid subjectivity imposing some parametric assumption on the error variable. Recently, some nonparametric methods to check a goodness-of-fit have been proposed with Hilbert–Schmidt independence criterion (HSIC) [12, 13, 34, 40] and mutual information [1]. In this paper, we utilize an idea of HSIC proposed by [12, 13]. With HSIC, Sen and Sen [34] and Hidalgo et al. [40] proposed a test statistics to check goodness-of-fit of a (parametric/nonparametric) model, which has the two desirable properties. Their idea is based on the fact that independence of X and ε implies correctness of the mean function

$\mu(x; c) = c_1 + c_2^\top x$ because ε is independent of x. The HSIC can be used to check the independence.

We propose using a nonparametric test statistic by using the bootstrap method [34] to check the linearity of $m(x) = E(Y \mid X = x)$. Let $H = I_n - n^{-1} 1_n 1_n^\top$ where I_n is the $n \times n$ identity matrix and 1_n is the $n \times 1$ vector of ones. Also let each k and l be a characteristic kernel that prescribes RKHS of the random variables X and Y, respectively, such as a Gaussian kernel $k(x, \tilde{x}) = \exp(-\sigma^{-1} \|x - \tilde{x}\|)$, where σ is a tuning parameter and the sample median is often used as a heuristic estimate of σ. In the algorithm, let $x = (x_1, \ldots, x_{n1})^\top$ and $y = (y_1, \ldots, y_{n1})$ be observed covariate variables and response variables. To make the algorithm simple and clear, a vector is used instead of each element, that is, $\mu(x; c)$ implies the vector $(\mu(x_1; c), \ldots, \mu(x_{n_1}; c))^\top$. Also for a matrix A, $(A)_{ij}$ implies (i, j)th element.

1: **procedure** COMPUTATION FOR NONPARAMETRIC TEST STATISTICS $\hat{M}_{X\varepsilon}$ AND ITS
    BOOTSTRAP SAMPLES $\hat{M}_{X\varepsilon}^{(b)}(b = 1, \ldots, B)$
2:       $\hat{c} \leftarrow \text{argmin}_c \sum_{i=1}^{n_1} \{y_i - \mu(x_i; c)\}^2$
3:       $\hat{\varepsilon} \leftarrow y - \mu(x; \hat{c})$
4:       $(K)_{ij} \leftarrow k(x_i, x_j); (E)_{ij} = l(\hat{\varepsilon}_i, \hat{\varepsilon}_j)$ for all i, j
5:       $\hat{M}_{X\varepsilon} = n_1^{-2} \text{tr}(KHEH)$
6:     **for** $b = 1$ **to** B **do**
7:         $x^{(b)} \leftarrow$ bootstrap sample from observed data x
8:         $\varepsilon^{(b)} \leftarrow$ bootstrap sample from $\hat{\varepsilon}$
9:         $y^{(b)} \leftarrow \mu(x^{(b)}; \hat{c}) + \hat{\varepsilon}^{(b)}$
10:       $\hat{c}^{(b)} \leftarrow \text{argmin}_c \sum_{i=1}^{n_1} \{y_i^{(b)} - \mu(x_i^{(b)}; c)\}^2$
11:       $\hat{\varepsilon}^{(b)} \leftarrow y^{(b)} - \mu(x^{(b)}; \hat{c}^{(b)})$
12:       $(K^{(b)})_{ij} \leftarrow k(x_i^{(b)}, x_j^{(b)}); (E^{(b)})_{ij} = l(\hat{\varepsilon}_i^{(b)}, \hat{\varepsilon}_j^{(b)})$ for all i, j
13:       $\hat{M}_{X\varepsilon}^{(b)} = n_1^{-2} \text{tr}(K^{(b)} H E^{(b)} H)$
14:     **end for**

To obtain the limiting distribution of the test statistics $\hat{M}_{X\varepsilon}$ and its bootstrap samples $\hat{M}_{X\varepsilon}^{(b)}(b = 1, \ldots, B)$, we need additional conditions (T1)–(T4) given in Appendix A. Using Sen and Sen [34], we have the following asymptotic result.

PROPOSITION 3.2. *Suppose that conditions (T1)–(T3) hold. Then, under the null hypothesis H_0 in (3.3), the limiting distribution of $n_1 \hat{M}_{X\varepsilon}$ is a quadratic function of a Gaussian field χ. Also, if the conditions (T1), (T2), and (T4) hold, the distribution of the bootstrap samples $n_1 \hat{M}_{X\varepsilon}^{(b)}(b = 1, \ldots, B)$, given observed data almost surely, converge in law to the same distribution χ.*

REMARK 3.1. Proposition 3.2 guarantees that the bootstrap samples converge to the same distribution χ given observed data almost surely. Therefore, we do not need to know the explicit form of the limiting distribution χ, which is given in equation (14) in Sen and Sen [34] by setting $g(x) = (1, x^\top)^\top$. Thus, the p-value of the test statistics is computed by $B^{-1} \sum_{b=1}^{B} I(\hat{M}_{X\varepsilon} \leq \hat{M}_{X\varepsilon}^{(b)})$.

3.3. *Doubly-normalized exponential transformation.* When the null hypothesis (3.3) is not rejected, an instrumental variable is required to make the estimator, which is the unique solution to (2.1), consistent. However, selecting the instrumental variable is very difficult even if it exists. Because the problem comes from using the same covariate between the response model and mean function, we can make the model identifiable artificially by transforming

covariate variable x in the response model to a nonlinear variable $\mathcal{T}(x)$ such as $\exp(x)$ and x^2, at the sacrifice of consistency. Although there are many choices of such functions, it would be desirable that the transformation enjoys three properties: (i) "nonlinearity" can be adjusted through a tuning parameter a such that $\lim_{a \to 0} \mathcal{T}_a(x) = x$; (ii) the value a does not depend on range/scale of x; (iii) range of $\mathcal{T}_a(x)$ is same as that of x. The first condition is necessary to adjust "nonlinearity": small a-value holds the original data structure, and large a-value breaks the structure, but provides stronger identification. For example, one may come up with a transformation $\mathcal{T}_a(x) = \log\{a + \exp(x)\}$. However, nonlinearity of such a transformation may heavily depend on both a and range/scale of x so that it is necessary to find an appropriate value a (which is close to 0) for every covariate or data set, hence, the second condition is required. The third condition is requisite to retain the value of response probability to some extent. Considerably large (small) value of $\mathcal{T}_a(x)$ may damage the bounded condition $\pi(\mathcal{T}_a(x), y) > 0$, which is often assumed in this field.

We propose a simple nonlinear transformation having three desirable properties called doubly-normalized exponential transformation (DNET). Let $\mathcal{S}_a(x)$ be a normalized exponential transformation $\mathcal{S}_a(x) = \{\mathrm{Var}(aX)\}^{-1/2}\{\exp(ax) - E(\exp(aX))\}$. By letting $a \to 0$, we obtain

$$
\begin{aligned}
\lim_{a \to 0} \mathcal{S}_a(x) &= \lim_{a \to 0} \frac{a^{-1}\{\exp(ax) - 1\} + a^{-1}\{1 - E(\exp(aX))\}}{\{\mathrm{Var}(X)\}^{1/2}} \\
&= \frac{x - E(X)}{\{\mathrm{Var}(X)\}^{1/2}}.
\end{aligned}
$$

This indicates that the normalized exponential transformation $\mathcal{S}_a$ after data normalization is an identity map as $a \to 0$, that is, with $\mathcal{Z} : x \mapsto \{\mathrm{Var}(X)\}^{-1/2}(x - E(X))$, a map $\mathcal{S}_a \circ \mathcal{Z}$ becomes identity as $a \to 0$. Finally, after some minor modification to satisfy the third condition above, we have our proposed transformation method:

1: **procedure** COMPUTE DNET(a)
2: $\quad z \leftarrow \{\mathrm{var}(x)\}^{-1/2}(x - \mathrm{mean}(x))$
3: $\quad s \leftarrow \{\mathrm{var}(az/5)\}^{-1/2}\{\exp(az/5) - \mathrm{mean}(\exp(az/5))\}$
4: $\quad r_x \leftarrow \max(x) - \min(x); \ r_s \leftarrow \max(s) - \min(s)$
5: $\quad \mathcal{T}_a(x) \leftarrow \min(x) + \{s - \min(s)\} \times r_x/r_s$

In the algorithm, each mean, var, max and min is sample mean, variance, maximum and minimum value of $x = (x_1, \ldots, x_n)$. Obtained $\mathcal{T}_a(x_i)(i = 1, \ldots, n)$ is the proposed nonlinear transformation. The reason divided by 5 is just for scale adjustment. We call the transformation with a-value 0.5 (weak), 1 (moderate) and 2 (strong) nonlinearity. In Figure 1, we illustrate the scatterplot of $\mathcal{T}_a(x_i)$ versus y_i, for $a = 0$ (original), 0.5, 1, 2, where $(x_i, y_i)(i = 1, \ldots, 500)$ are independently generated from a bivariate normal distribution with both mean 0, variance 1 and correlation 0.5. It can be seen that the transformation enjoys the three desirable properties.

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 under the constraint that the nuisance tangent space of θ and ϕ are orthogonal. Rotnitzky and Robins [30] derived the semiparametric efficiency bound for regression parameters, which prescribe the first moment of the distribution of $[y \mid x]$. However, their adaptive estimators require many working models, and misspecification of either a regression model or a response model leads to a biased estimator, but in most cases, we do not expect the regression model is true and assume a simple function such as a linear regression model.
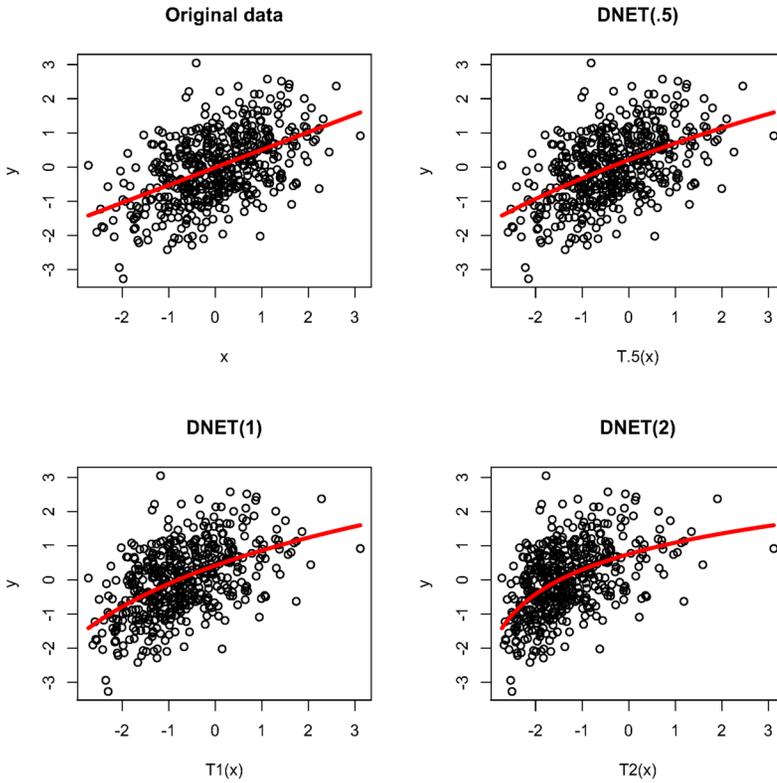
FIG. 1. *Illustration of DNET: each top left, top right, bottom left and bottom right shows the scatterplot of $\mathcal{T}_a(x)$ v.s. y, for $a = 0$ (original), 0.5, 1, 2, respectively. The red curve is conditional mean function of y given $\mathcal{T}_a(x)$.*

In this section, we first provide the efficiency bound under the response model only, without relying on the information of U-function because the most difficult task in nonignorable nonresponse missing data analysis is to obtain a consistent estimator of the response model. Optimal estimators achieving this lower bound will be considered in the next section.

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

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

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

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

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

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 $H = E(S_{\text{eff}}^{\otimes 2}) = E\{\partial S_{\text{eff}}(\phi_0, \theta_0)/\partial(\phi^{\mathrm{T}}, \theta)^{\mathrm{T}}\}$ and $B^{\otimes 2} = BB^{\mathrm{T}}$. Therefore, the semiparametric efficiency bound is given by $\{E(S_{\text{eff}}^{\otimes 2})\}^{-1}$.

This lemma 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

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

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 Section 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 and Kott [3] defined in (2.1). Thus, the optimal g function in (2.1) for the Chang and Kott [3] 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.

REMARK 4.2. The optimal score function in (4.1) can be derived differently as follows. Consider the class of estimating equations in (2.1) indexed by g. For given g, the asymptotic variance of the solution $\hat{\phi}_g$ to (2.1) can be written as

$$V(\hat{\phi}_g) = \frac{1}{n} A_g^{-1} B_g A_g^{-1},$$

where

$$A_g = E[E\{O(Z; \phi_0) \cdot s_0(Z; \phi_0) \mid X\} g(X; \phi_0)^{\mathrm{T}}],$$
$$B_g = E[E\{O(Z; \phi_0) \mid X\} g(X; \phi_0) g(X; \phi_0)^{\mathrm{T}}].$$

Using the Cauchy–Schwarz inequality, the asymptotic variance is minimized at $g^\star(x; \phi_0) = E^\star\{s_0(Z; \phi_0) \mid x; \phi_0\}$. Similarly, we can obtain the optimal estimating function in (4.2) by considering a class of estimating equations of the form

$$(4.5) \qquad \sum_{i=1}^{n} \left[\frac{r_i}{\pi(z_i; \phi)} U(z_i; \theta) + \left\{1 - \frac{r_i}{\pi(z_i; \phi)}\right\} h(z_i) \right] = 0,$$

indexed by h. The asymptotic variance of the solution to (4.5) is minimized at $h = E^\star\{U(Z; \theta) \mid x; \phi_0\}$.

REMARK 4.3. In our estimation steps, ϕ and θ are separately estimated. Thus, it follows from the identifiability of ϕ that θ is also identifiable. This is because, under assumptions (I1)–(I3), ϕ is identifiable; thus, the identification problem of θ reduces to that of the probability limit of (4.4) or expectation of (4.2), that is, $E\{U(Z; \theta)\}$.

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 $f_1(y \mid x) = f(y \mid x, r = 1)$. Although the optimality result in Lemma 4.1 has already been discussed by Rotnitzky and Robins [30], the adaptive estimators proposed here are different from those of Rotnitzky and Robins [30]. See Appendix C for some discussion of Rotnitzky and Robins [30] estimator.

To discuss the first proposed 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 the model selection can be implemented 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 (2.3), we can show that, for any function $g(z)$,

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

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 efficient estimating equation $\sum_{i=1}^{n} S_{\text{eff},i}(\phi, \theta, \hat{\gamma}) = 0$, where

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

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. Note that 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 $(\hat{\phi}^{\mathrm{T}}, \hat{\theta})^{\mathrm{T}}$ be the solution to $\sum_{i=1}^{n} S_{\text{eff},i}(\phi, \theta, \hat{\gamma}) = 0$ in (5.2). Under conditions (I1)–(I3) and (C1)–(C7) given in Appendix A and the identification conditions assumed in Theorem 3.1, $(\hat{\phi}^{\mathrm{T}}, \hat{\theta})^{\mathrm{T}}$ satisfies consistency and asymptotic normality with variance*

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

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

$$(5.3) \qquad V^* = \text{var}[\tau_U^{-1}\{S_2(\phi_0, \theta_0, \gamma^*) - \kappa^* S_1(\phi_0, \gamma^*)\}],$$

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

Note that Theorem 5.1 does not require that f_1 be correctly specified. Unlike the estimator of Riddles et al. [27], 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 and Zeger

[18]. Also, though equation (4.2) has a form similar to that of the doubly robust estimator under MAR [29], 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. [27] 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. [27] 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 of [16]. 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 $R = 1$, for example, for $x = 0, 1$,

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

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

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

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

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 $(\hat{\phi}^T, \hat{\theta})^T$ be the solution to $\sum_{i=1}^n \hat{S}_{\text{eff},i}(\phi, \theta) = 0$, where $\hat{S}_{\text{eff},i}(\phi, \theta)$ is defined in (4.4) with the estimated conditional expectation in (5.5). Under Conditions (I1)–(I3), (C1)–(C4) and, (C8)–(C13) given in Appendix A, $(\hat{\phi}^T, \hat{\theta})^T$ satisfies consistency and asymptotic normality, and the estimator attains the semiparametric efficiency bound.*

REMARK 5.1. The second proposed estimator is robust 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 $f_1(y \mid x; \gamma)$,

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

converges to V^* in probability as defined in (5.3), where $\hat{\tau}_U$ 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, that is, let

$\mathcal{U}(\phi_0, k_1, \gamma^*) = k_1 - (U^*(\gamma^*) - U)\dot{\pi}(\phi_0)^T/\pi(\phi_0)$ 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:

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

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 to compare their efficiency with other methods in finite samples, we conduct a Monte Carlo simulation study with three scenarios. In each scenario, two covariates $X_1 \sim N(0, 1/\sqrt{2}^2)$ and $X_2 \mid X_1 = x_1 \sim N(-x_1/3, 1/2^2)$ are used. For each scenario $s(= 1, 2, 3)$, the response mechanism is set to a Bernoulli distribution with parameter $\pi^{[s]}(x_1, x_2, y)$, where $\pi^{[s]}(x_1, x_2, y) = 1/\{1 + \exp(\phi_{x0}^{[s]} + 0.5x_1 + 0.5x_2 + \phi_y^{[s]}y)\}$, and the response outcome variables are generated from $Y \mid (x, r = 1) \sim N(\mu^{[s]}(x), 1/2^2)$, where $\mu^{[s]}(x) = a_0^{[s]} + 0.4x_1 + 0.4x_2 + a_1^{[s]}x_1x_2$. The coefficients of the nonlinear term, which is the degree of nonlinearity, are set to $a_1^{[1]} = 0, a_1^{[2]} = 0.3, a_1^{[3]} = 0.6$, and the other parameters are set, so that the expectation of the outcome variable is zero and the marginal response probability is 70%, to $\phi_{x0}^{[1]} = -0.959, \phi_{x0}^{[2]} = -0.914, \phi_{x0}^{[3]} = -0.904, \phi_y^{[1]} = 0.75, \phi_y^{[2]} = 0.4, \phi_y^{[3]} = 0.3, a_0^{[1]} = -0.0563, a_0^{[2]} = 0.02$ and $a_0^{[3]} = 0.0775$. Note that the scenarios 2 and 3 are identifiable without using any instrumental variable because of the nonlinear term x_1x_2 in f_1, on the other hand, Scenario 1 is unidentifiable, and Scenario 2 is weakly identified than Scenario 3. We estimate $\theta = E(Y)$; thus, $U(\theta; Z) = \theta - Y$, with two different Monte Carlo samples of size $n = 500$ and $n = 2000$ being independently generated 2000 times.

In Scenario 1, however, it is still possible to make the response model identifiable at the risk of misspecification of the response mechanism by using DNET. In this article, we change the variable $x_1 \to \mathcal{T}_a(x_1)$ and $x_2 \to \mathcal{T}_a(x_2)$ ($a = 0.5, 1, 2$).

From each sample, we compute seven estimators, as follows:

1. CK: The estimator of Chang and Kott [3]. We use the estimating equation (2.1), setting g as $(1, x_1, x_2)$; θ is estimated by solving

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

where $\hat{\pi}$ is the estimated response model.

2. RR: The estimator of Rotnitzky and Robins [30]. This estimator is defined through four steps (i)–(iv) in Appendix C. In the first step, a consistent estimator is set to be the CK estimator, and in the second step, each of (C.1), (C.3)–(C.6) is modeled by at most third-order polynomial function of x_1 and x_2.

3. RRC: The estimator of Rotnitzky and Robins [30] with correct (C.1), (C.3)–(C.6) models. This estimator is exactly the same as the RR estimator except for using the correct specification of (C.1), (C.3)–(C.6) models.

4. RKI: The estimator of Riddles et al. [27]. In all scenarios, we specify a parametric model on f_1 based on normal distribution with the correct mean structure $\mu(x) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2$.

5. P: Our proposed estimator with parametric f_1 model. As for the working model for f_1, the same model specification as in the RKI method is used.

6. NP: Our proposed estimator with nonparametric f_1 model. As for the kernel function and its bandwidth, Gaussian kernel and a rule-of-thumb bandwidth $h_j = n_1^{-1/5} \hat{\sigma}_{x_j}$ $(j = 1, 2)$ is used, where n_1 is the sample size of observed outcome variable and $\hat{\sigma}_j$ is the square root of the sample variance of x_j for $j = 1, 2$.

7. DNET(a): Same method as P and NP with the nonlinearly transformed data $\mathcal{T}_a(x_1)$ ($a = 0.5, 1, 2$) for the variables in response models.

Suppose that the response model is correctly specified in our proposed methods, and except for our proposed methods, for identifiability, suppose that x_2 is specified as the instrumental variable, that is, the response model is specified as

$$\text{logit}\{\pi(x_1, y)\} = \phi_{x0} + \phi_{x1} x_1 + \phi_y y.$$

Before estimating the parameters, we first check the model identifiability of our proposed method. The right panel in Figure 2 shows the p-values of the statistical tests proposed in Section 3.2 under the three scenarios with different sample sizes. The p-values in scenario 1 spread around $1/2$ because the null hypothesis is correct or the model is unidentifiable. On the other hand, as the nonlinearity increases, p-values are close to zero. In particular, when $n = 2000$, model identification can be judged with the probability almost 1 even for Scenario 2 which has a small degree of the nonlinearity 0.3.

The left panel in Figure 2 shows the Monte Carlo simulation results with sample size $n = 500$. The results with sample size $n = 2000$ are omitted because they are almost the same. In some Monte Carlo samples, we encounter some numerical problems and there is no solution because the estimate of the response model does not converge due to weak identifiability. The rates of data sets not having converging estimators are reported at the bottom right in Figure 2. The following is a summary of the simulation results:

1. The CK method estimates the parameter stably, but it is biased due to the misspecification of the response model. The standard error of the CK estimators is a little larger than RKI, P, NP and DNET methods due to the lack of efficiency.

2. In many cases, the RR estimators do not converge. This comes from the difficulty of finding a good starting value of ϕ in the first step and of modeling the working models defined in Appendix C.

3. The RRC estimators relatively improve the performance of RR estimators due to the correct working models, though they are unknown in practice.

4. Performance of RKI method is similar to RRC because these two estimators use the correct f_1 model.

5. When the model is identifiable, the proposed P method works well. However, when it is unidentifiable, it is hard even to get a convergent sequence of estimators, though that can be inferred by testing linearity of the mean function.

6. Surprisingly, the NP method can estimate the parameter stably despite of the unidentifiability of the model, and the estimates are biased according to the degree of linearity of $\mu^{[s]}(x)$.

7. Proposed DNET works well for all the transformations $\mathcal{T}_a(x)$. In Scenario 1, when the model is not identifiable, the rate obtaining a nonconvergent estimator and bias increase as the nonlinearity increases, in the meantime, the standard error decreases.

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 [17, 35, 43]. The data contain $n = 2506$ Korean wage earners; the response variable y is total
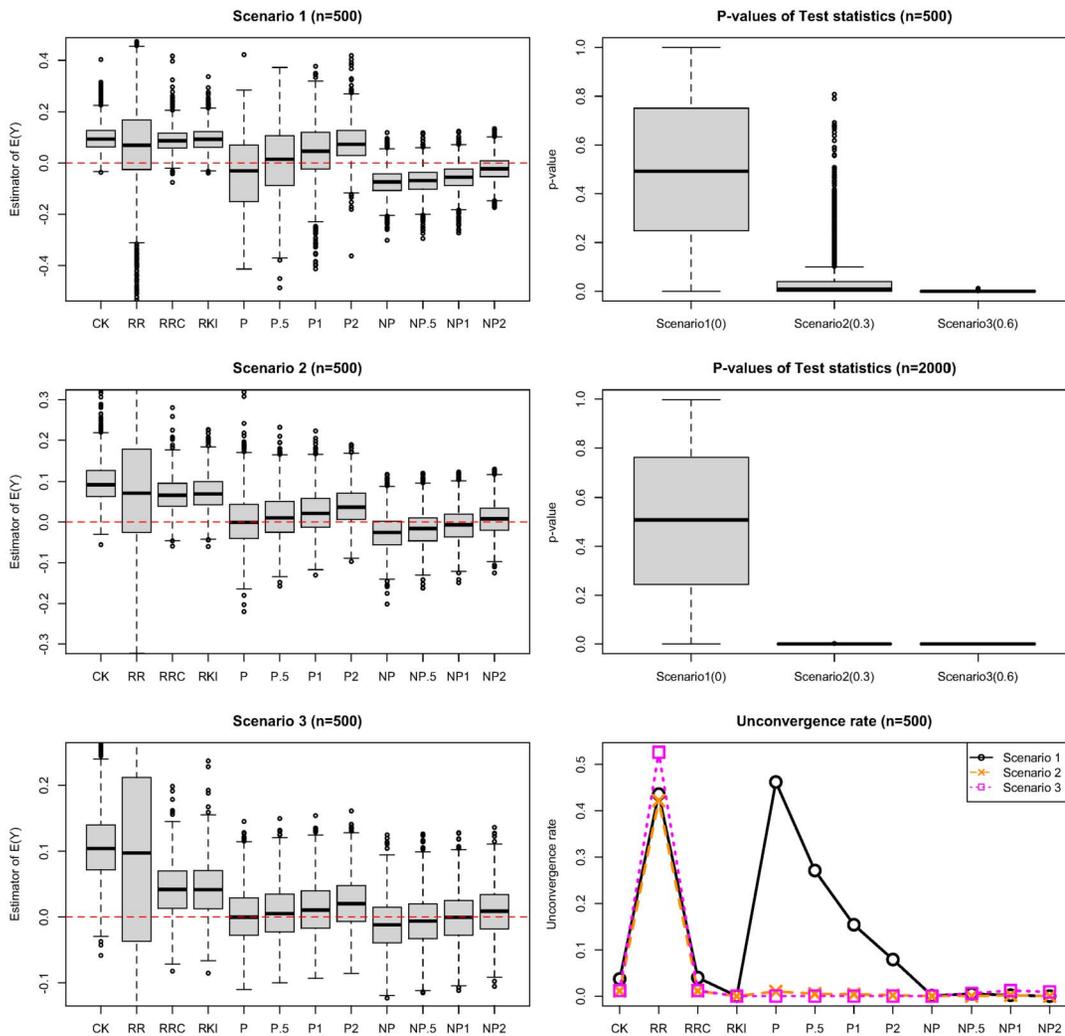
FIG. 2.   *Left panel*: *Boxplot of Monte Carlo results for converged estimators for* $\theta = E(Y)$ *under three scenarios with sample size* 500. *The six estimators are CK* (*Chang & Kott's estimator*), *RR* (*Rotnitzky & Robins's estimator*), *RRC* (*Rotnitzky & Robins's estimator with correct working models*), *RKI* (*Riddles' estimator*), *P* (*our proposed estimator with parametric* f_1), *NP* (*our proposed estimator with nonparametric* f_1) *and P and NP method with DNET* (*doubly-normalized exponential transformation*) *where the suffix stands for a value of* $\mathcal{T}_a(x_1)$. *The broken line shows the true value. Right panel*: *Test statistics for identifiability for data sets with each sample size* 500 *and* 2000, *and the bottom right panel shows the rate of estimators failed to obtain an estimator.*

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, respectively, and x_3 has three categories 1–3: $x_3 < 35, 35 \leq x_3 < 51$ 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 and Yu [17] and created 1000 incomplete data sets with the following eight response mechanisms: M1 (linear nonignorable without (x_2, x_3)): logit(π) $= 0.48 - 0.3x_1 - 0.5y$; M2 (linear nonignorable): logit(π) $= -0.85 - 0.2x_1 + 0.5x_2 + 0.2x_3 - 0.4y$; M3 (nonlinear nonignorable, quadratic in x_1 without (x_2, x_3)): logit(π) $= 0.33 - 0.3x_1 - 0.1x_1^2 - 0.3y$; M4 (nonlinear nonignorable, quadratic in x_1): logit(π) $= -0.89 - 0.4x_1 - 0.1x_1^2 + 0.5x_2 + 0.2x_3 - 0.4y$; M5 (nonlinear nonignorable, quadratic in y without (x_2, x_3)): logit(π) $= 0.24 - 0.25x_1 - 0.25y - 0.1y^2$;

M6 (nonlinear nonignorable, quadratic in y): $\text{logit}(\pi) = -0.93 - 0.2x_1 + 0.5x_2 + 0.2x_3 - 0.2y - 0.1y^2$; M7 (probit nonignorable) $\pi = \Phi(-0.55 + 0.3x_1 + 0.4y)$; 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 $\Phi(\cdot)$ is the cumulative distribution function of the standard normal distribution, and $I(A)$ is the indicator function that takes 1(0) if 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 $\text{logit}\{\pi(x, y; \phi)\} = \phi_{x0} + \phi_{x1}x_1 + \phi_{x2}x_2 + \phi_{x3}x_3 + \phi_y y$. Therefore, M1 and M2 are correctly specified while M3–M8 are misspecified.

We specified unknown f_1 models as normal distribution $Y \mid (x_1, x_2 = i, x_3 = j, r = 1) \sim N(\mu_{i,j}(x_1), \sigma^2_{i,j})$ $(i = 1, 2; j = 1, 2, 3)$, 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. Using Theorem 3.1, one can show that this model is identifiable as one of the 6 mean structures are nonlinear, or all the structures are linear but all of them are not the same. One simple sufficient condition is to check whether the conditional mean of y given x_1 is linear with respect to x_1. In the real data, the correlation between x_1 and y is too high because wage income does not change considerably within one year; the mean structure is almost linear. However, the p-values of the test statistics are almost zero in all data sets with M1–M8, therefore, all the response models are identifiable without using any instrumental variable nor transformation. In Table 1, Bias, S.E. (standard error) and RMSE (root mean square error) with five methods, CK, RR, RKI, P, NP methods same as in Section 5, are reported. The following are a summary of the results:

1. The CK method estimates the parameter stably, but it is inefficient compared to our proposed methods.

2. As in Section 5, the RR estimators do not converge in many data sets.

3. RKI methods can obtain estimates stably, but it is severely biased due to the misspecification of f_1 model, which is generally unknown in real.

4. The proposed P method works well, but for some data sets, we encounter some numerical problems due to the misspecification of the response model. As for such data sets, we may get a reasonable estimator by using DNET. Note that our method is effective for the probit response mechanism (M7), even though the use of the probit model makes it hard to identify the parameter as stated in Section 3.1.

5. Performance of the proposed NP method is the best among the five methods considered. However, the results with data set M5 and M6 implies the difficulty of obtaining the estimator with misspecified response models.

8. Discussion. We have presented a test statistic for model identification, semiparametric efficiency bound for $(\phi_0^T, \theta_0)^T$ under nonignorable nonresponse; proposed two types of adaptive semiparametric estimators that attain the semiparametric lower bound. Identifiability is a challenging problem in nonignorable nonresponse [23]; previous methods require nonignorable NIVs to guarantee model identification [43]. Our new identifiability condition is not on the response mechanism, but on the distribution of $[y \mid x, r = 1]$.

In this paper, we have mainly used a logistic response model, but as discussed in Section 3, it may be possible to use other response models such as the probit model. Finding a sufficient condition for model identifiability of the response models such as done in Miao et al. [23] will be a future work. We adopted a kernel regression to construct an adaptive estimator. However, as recently shown in Chernozhukov et al. [5], we expect that other nonparametric methods

Model		Methods				
		CK	RR	RKI	P	NP
M1	Bias	73	118	737	22	17
	S.E.	130	388	676	48	31
	RMSE	149	406	1000	52	35
	NA rate(%)	0.7	31.9	0	1.8	1.6
M2	Bias	66	90	599	23	13
	S.E.	119	217	580	49	31
	RMSE	136	235	834	54	33
	NA rate(%)	1.1	27.9	0	1.2	1.0
M3	Bias	205	197	879	11	-6
	S.E.	128	694	810	49	31
	RMSE	241	722	1195	50	32
	NA rate(%)	8.1	48.4	0	4.3	2.5
M4	Bias	-103	10	245	55	29
	S.E.	100	179	295	51	26
	RMSE	144	179	383	75	39
	NA rate(%)	15.8	40.9	0	1.1	0.3
M5	Bias	-39	78	840	59	0
	S.E.	71	288	780	211	56
	RMSE	81	298	1146	219	56
	NA rate(%)	2.9	24.2	0	1.9	32.7
M6	Bias	-68	57	776	44	13
	S.E.	87	550	734	51	49
	RMSE	111	553	1068	68	51
	NA rate(%)	7.4	34.0	0	1.7	37.4
M7	Bias	158	125	1131	15	11
	S.E.	155	472	844	42	34
	RMSE	221	489	1412	45	36
	NA rate(%)	2.2	36.7	0	5.8	2.0
M8	Bias	175	136	689	27	9
	S.E.	115	475	579	48	33
	RMSE	209	494	900	55	34
	NA rate(%)	0.6	37.2	0	1.4	1.1

such as the kernel unconstrained least squares importance fitting (KuLSIF) [15] for estimating the $f_1(y \mid x)$ distribution, which have a certain mild convergence rate, are also applicable because the estimating equations (4.1) and (4.2) have, as it can be easily shown, the Neyman orthogonality defined in [5]. The proposed method is based on the correct specification of the response model. There may be various other 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 under the ignorable nonresponse setup [4, 14]. Extension of multiple robustness to the nonignorable nonresponse case will also be a topic of our future research.

APPENDIX A: REGULARITY CONDITIONS

Conditions for consistency and asymptotic normality of the proposed estimator.

(C1) Φ and Θ are compact.

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

(C3) 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 $\|E\{\sup_{\gamma \in \Gamma_\mathcal{N}} \partial S_\gamma(\gamma)/\partial \gamma^{\mathrm{T}}\}\| < \infty$.

(C4) Identifiability of θ for complete data: there exists a unique $\theta_0 \in \Theta$ such that $E\{U(Z; \theta_0)\} = 0$.

(C5) $\partial S_{\mathrm{eff}}(\phi, \theta, \gamma)/\partial(\phi^{\mathrm{T}}, \theta, \gamma^{\mathrm{T}})$ 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

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

(C6) $S_{\mathrm{eff}}(\phi, \theta, \gamma)$ 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 $E\{d_1(W)\} < \infty$.

(C7) $E\{\partial S_{\mathrm{eff}}(\phi, \theta, \gamma^*)/\partial(\phi^{\mathrm{T}}, \theta, \gamma^{\mathrm{T}})\}$ is nonsingular at $(\phi_0, \theta_0, \gamma^*)$.

(C8) The conditions (C5)–(C7) hold at the true value $\gamma^* = \gamma_0$.

(C9) Let $\mathcal{X}$ be the support of x. Then $f_1(x) > 0$ and $E_1\{\pi(x, Y; \phi_0) \mid x\} > 0$ for all $x \in \mathcal{X}$.

(C10) The kernel $K(u)$ has bounded derivatives of order k, satisfies $\int K(u)\,du = 1$, has zero moments of order $\le m - 1$, and has a nonzero mth order moment.

(C11) 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 $\mathcal{X}$.

(C12) 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 $x \in \mathcal{X}$.

(C13) 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 $n^{1/2}h^{2m} \to 0$.

Conditions for consistency of the test statistics.

(T1) The matrix $E\{(1, X^{\mathrm{T}})^{\mathrm{T}}(1, X^{\mathrm{T}})\}$ is invertible;

(T2) Let k and l be characteristic kernels, k is continuous, and l is twice continuously differentiable. Denoting the partial derivatives of l by $l_x(x, y) = \partial_x l(x, y)$, $l_{xy}(x, y) = \partial_x \partial_y l(x, y)$, etc. Assume that l_{xx}, l_{xy} and l_{yy} are Lipschitz continuous with Lipschitz constant L with respect to the l_∞-norm, where the l_∞-norm for a vector $u = (u_1, \ldots, u_p)^{\mathrm{T}}$ is defined by $|u|_\infty = \max_{1 \le i \le p} |u_i|$.

(T3) Let each $(X_1, \ldots, X_4)$ and $(\varepsilon_1, \varepsilon_2)$ be indenpendent and identically distributed (i.i.d.) copies of X and ε, respectively. Then (a) $E|X|_\infty^2 < \infty$, (b) $E(\varepsilon^2) < \infty$, (c) $E\{v_2(X_q, X_r, X_s, X_t)\} < \infty$ ($1 \le q, r, s, t \le 4$), and (d) $E\{f^2(\varepsilon_q, \varepsilon_r)\} < \infty$ for $f = l, l_x, l_y, l_{xx}, l_{yy}, l_{xy}$ ($1 \le q, r \le 2$), where $v_p(X_q, X_r, X_s, X_t) = k^p(X_q, X_r)\{1 + |X_s|_\infty^p\}\{1 + |X_t|_\infty^p\}$.

(T4) Let $\epsilon = m(X) - \mu(X; \tilde{c}) + \varepsilon$, $\epsilon^0 = \epsilon - E(\epsilon)$, $\tilde{c}$ be the probability limit of the least squared estimator $\hat{c}$, and $(\epsilon_1^0, \epsilon_2^0)$ be i.i.d. copies of ϵ^0. There exists $\delta > 0$ such that (a) $E\{|X|_\infty^{4+2\delta}\} < \infty$, (b) $E\{|\mu(X)^{2+\delta}|\} < \infty$, (c) $E\{|\varepsilon|^{2+\delta}\} < \infty$, (d) $E\{v_{2+\delta}(X_q, X_r, X_s, X_t)\} < \infty$ ($1 \le q, r, s, t \le 4$), (e) for $1 \le q, r \le 2$,

$$E\{|l(\epsilon_q^0, \epsilon_r^0)|^{2+\delta}\} < \infty, \qquad E[\{1 + |X_q|_\infty^{2+\delta}\}|f(\epsilon_q^0, \epsilon_r^0)|^{2+\delta}] < \infty \quad (f = l_x, l_y),$$

$$E[\{1 + |X_q|_\infty^{2+\delta} + |X_q|_\infty^{4+2\delta}\}|f(\epsilon_q^0, \epsilon_r^0)|^{2+\delta}] < \infty \quad (f = l_{xx}, l_{yy}, l_{xy}).$$

APPENDIX B: PROOFS OF THE TECHNICAL RESULTS

PROOF OF THEOREM 3.1. Let $f_1(y \mid x)$ be the true density function of $[y \mid x, r = 1]$. Here, the distribution of $[y \mid x]$ can be represented through the observed outcome density and the response model, because by using Bayes' formula, we have

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

Suppose that ϕ_0 is the true value of the response model so that the true distribution of $[y \mid x]$ is $f(y \mid x; \phi_0)$. Then it follows from (B.1) that the probability limit of the estimating equation is

$$E\{\Gamma(Z, R; \phi) \mid x\} = g(x; \phi) \int \left\{1 - \frac{\pi(Z; \phi_0)}{\pi(Z; \phi)}\right\} f(y \mid x; \phi_0)\,dy$$

$$= g(x; \phi)\left\{1 - \frac{\int \pi(Z; \phi)^{-1} f_1(y \mid x; \phi)\,dy}{\int \pi(Z; \phi_0)^{-1} f_1(y \mid x; \phi_0)\,dy}\right\}.$$

By using (I2) and (I3), the conditional expectation cannot be vanished unless $\phi = \phi_0$. Therefore, the solution is unique. □

PROOF OF PROPOSITION 3.1. For any error function $\varepsilon \in \mathcal{E}$, under the null hypothesis $H_0^{(\infty)}$, there exist $c_1^{(\ell)}, c_2^{(\ell)}(\ell = 2, 3)$ such that $E(\varepsilon^2 \mid x) = c_1^{(2)} + (c_2^{(2)})^\top x$ and $E(\varepsilon^3 \mid x) = c_1^{(3)} + (c_2^{(3)})^\top x$. On the other hand, it holds that

$$\varepsilon^2 = \sum_{j=0}^\infty \xi_j^2 e_j^2(x) + \sum_{j \neq k} \xi_j \xi_k e_j(x) e_k(x).$$

It follows from $e_j \neq e_k$ $(j \neq k)$ that there must exist a positive integer j such that $e_j = \{E(\xi_j^2)\}^{-1/2}(c_1^{(2)} + (c_2^{(2)})^\top x)^{1/2}$ and $e_k \equiv 0$ for $k \neq j$, so that $\varepsilon = (c_1^{(2)} + (c_2^{(2)})^\top x)^{1/2}\xi_j$. Let such j be 1 without loss of generality. In a similar way, it follows from the third moment condition of ε that $\varepsilon = (c_1^{(3)} + (c_2^{(3)})^\top x)^{1/3}\xi_1$, which implies $c_2^{(2)} = c_2^{(3)} = 0$ and $c_1^{(3)} = (c_1^{(2)})^{3/2}$. By using the induction, under the null hypothesis $H_0^{(\infty)}$, it can be shown that $c_2^{(\ell)} \equiv 0$ for $\ell \geq 2$. As a result, ε is a random variable, which is independent of x. Therefore, $H_0^{(1)}$ or testing linearity of mean function is enough to check the model identification. □

Next, we provide a proof of Lemma 4.1 and Theorems 5.1 and 5.2. In order to prove Lemma 4.1, we will assume $U(z) = y$ just for simplicity. We specify the joint distribution $z = (x^\top, y)^\top$ 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 $\langle h_1, h_2 \rangle = E(h_1^\top h_2)$, where the expectation is taken under the true model. See Bickel et al. [2] and Tsiatis [41] for more details. When U is comprised of other functions, the proof is almost the same.

At first, we introduce a proposition of Rotnitzky and Robins [30], 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: $g(B^*) = E(B^* \mid R, G_R(Z))$, $m(B^*) = E\{g(B^*) \mid Z\}$, and $u(B^*) = RB^*/\pi(Z)$. Then the efficient score for (ϕ, θ) can be derived by the following lemma. See Proposition A1 in Rotnitzky and Robins [30] for the proof.

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

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

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

$$(\text{B.3}) \qquad \Pi[m^{-1}(D^*) \mid \Lambda^{F\perp}] = (Q, S_{\text{eff},\theta}^{F\perp}),$$

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

This lemma implies that the efficient score can be represented by (B.2) with D_{eff}^* satisfying condition (B.3). 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 [41].

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

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

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*

$$(\text{B.4}) \qquad \Pi(h \mid \Lambda_2) = \left\{1 - \frac{R}{\pi(Z)}\right\} \frac{E[\{1 - \pi(Z)\}\{h_2(X) - E\{h_1(Z)\} \mid X]}{E\{O(Z) \mid X\}}.$$

PROOF OF LEMMA B.3. Obviously, the right-hand side of (B.4) 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. $\square$

We now give a proof of Lemma 4.1.

PROOF OF LEMMA 4.1. Note that $S_{\text{eff},\theta}^{F\perp} = Y - \theta_0$ 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 [41], Chapter 5). By the projection theorem, there exists a unique $k = (k_1, k_2^{\text{T}})^{\text{T}}$ such that $D_{\text{eff}}^* = k(Y - \theta_0)$.

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

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

where $s_r(\phi)$ is defined in (2.2). It follows from Lemma B.3 with $h_1(z) = s_1(\phi)$ and $h_2(x) = \bar{s}_0(x; \phi)$ in (B.4) 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.2) 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 $\tilde{S}_{\mathrm{eff}} = (\tilde{S}_1, \tilde{S}_2^{\mathrm{T}}) = A S_{\mathrm{eff}}$,

$$\tilde{S}_1 = \left\{ 1 - \frac{R}{\pi(\phi_0)} \right\} g^\star(X),$$

$$\tilde{S}_2 = \frac{R(Y - \theta_0)}{\pi(\phi_0)} + \left\{ 1 - \frac{R}{\pi(\phi_0)} \right\} E^\star(Y - \theta_0 \mid X)$$

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. $\square$

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 (5.2). 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

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

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

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 obtain the asymptotic variance of $\hat\theta$ as given in (5.3). □

PROOF OF THEOREM 5.2. Consistency and asymptotic normality of our proposed estimator are similar to proving Lemma 4.1 of Morikawa et al. [25]. 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. [25], 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

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

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

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

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 $\mathrm{Var}\{E(\dot\varphi(\phi_0))^{-1}\varphi(\phi_0)\}$, where $\dot\varphi(\phi_0) = \partial\varphi(\phi_0)/\partial\phi^{\mathrm{T}}$. 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. □

APPENDIX C: COMPARISON WITH ROTNITZKY AND ROBINS (1997)'S ESTIMATOR

In Rotnitzky and Robins [30], the semiparametric efficiency bound for NMAR data was derived in more general settings in Proposition A1 and A2, and an adaptive estimator for

regression coefficients was proposed. However, to attain the efficiency bound, the estimator requires many working models to be correctly specified, and it would be practically impossible to correctly specify all of the models. For example, for the case of nonignorable nonresponse, seven working models, equations (32)–(38) in Rotnitzky and Robins [30], have to be specified.

In particular, if $\theta = E(Y)$ is our parameter of interest, three working models are required:

(C.1) $$E_1\{\pi^{-1}(Z; \phi_0)O(Z; \phi_0) \mid x\} =: \nu_1(x; \zeta_1),$$

(C.2) $$E_1\{\pi^{-1}(Z; \phi_0)O(Z; \phi_0)s_0(Z; \phi_0) \mid x\} =: \nu_2(x; \zeta_2),$$

(C.3) $$E_1\{Y\pi^{-1}(Z; \phi_0)O(Z; \phi_0) \mid x\} =: \nu_3(x; \zeta_3).$$

Note that (C.2) is a multidimensional function. For example, in the same setup as Section 6, that is, $\text{logit}\{\pi(x, y; \phi)\} = \phi_{x0} + \phi_{x1}x_1 + \phi_y y$, where $x = (x_1, x_2)$, (C.2) can be written as

(C.4) $$E_1\{O(Z; \phi_0) \mid x\} =: \nu_4(x; \zeta_4),$$

(C.5) $$E_1\{x_1 O(Z; \phi_0) \mid x\} =: \nu_5(x; \zeta_5),$$

(C.6) $$E_1\{YO(Z; \phi_0) \mid x\} =: \nu_5(x; \zeta_6),$$

where $\nu_2(x; \zeta_2) = \{\nu_4(x; \zeta_4), \nu_5(x; \zeta_5), \nu_6(x; \zeta_6)\}^\top$.

Then an adaptive estimator of ϕ and θ can be obtained from the following four steps:

(i) Find a consistent estimator $\tilde{\phi}$ of ϕ_0 by, for example, Chang and Kott's [3] method;

(ii) Estimate $\zeta_k (k = 1, 2, 3)$ in (C.1)-(C.3) by the least square method with the estimated $\tilde{\phi}$;

(iii) Let $\hat{\phi}$ be a solution to

$$\sum_{i=1}^{n}\left\{1 - \frac{r_i}{\pi(z_i; \phi)}\right\}\frac{\nu_2(x_i; \hat{\zeta}_2)}{\nu_1(x_i; \hat{\zeta}_1)} = 0.$$

(iv) Let $\hat{\theta}$ be the solution to

$$\sum_{i=1}^{n}\left[\frac{r_i(y_i - \theta)}{\pi(z_i; \hat{\phi})} + \left\{1 - \frac{r_i}{\pi(z_i; \hat{\phi})}\right\}\left\{\frac{\nu_3(x_i; \hat{\zeta}_3)}{\nu_1(x_i; \hat{\zeta}_1)} - \theta\right\}\right] = 0.$$

Therefore, their adaptive estimator is similar to the two-step estimator in GMM. However, as shown in Section 6, it may be practically difficult to find a valid consistent estimator of ϕ for NMAR data. Also, giving reasonable parametric models for (C.1)–(C.3) are challenging because the left-hand side of them are nonlinear functions.

REFERENCES

[1] BERRETT, T. B. and SAMWORTH, R. J. (2019). Nonparametric independence testing via mutual information. *Biometrika* **106** 547–566. MR3992389 https://doi.org/10.1093/biomet/asz024

[2] BICKEL, P. J., KLAASSEN, C. A. J., RITOV, Y. and WELLNER, J. A. (1998). *Efficient and Adaptive Estimation for Semiparametric Models*. Springer, New York. MR1623559

[3] CHANG, T. and KOTT, P. S. (2008). Using calibration weighting to adjust for nonresponse under a plausible model. *Biometrika* **95** 555–571. MR2443175 https://doi.org/10.1093/biomet/asn022

[4] CHEN, S. and HAZIZA, D. (2017). Multiply robust imputation procedures for the treatment of item nonresponse in surveys. *Biometrika* **104** 439–453. MR3698264 https://doi.org/10.1093/biomet/asx007

[5] CHERNOZHUKOV, V., CHETVERIKOV, D., DEMIRER, M., DUFLO, E., HANSEN, C., NEWEY, W. and ROBINS, J. (2018). Double/debiased machine learning for treatment and structural parameters. *Econom. J.* **21** C1–C68. MR3769544 https://doi.org/10.1111/ectj.12097

[6] D'HAULTFOEUILLE, X. (2010). A new instrumental method for dealing with endogenous selection. *J. Econometrics* **154** 1–15. MR2558947 https://doi.org/10.1016/j.jeconom.2009.06.005

[7] DIGGLE, P. and KENWARD, M. G. (1994). Informative drop-out in longitudinal data analysis. *J. R. Stat. Soc. Ser. C. Appl. Stat.* **43** 49–93.

[8] DOMÍNGUEZ, M. A. and LOBATO, I. N. (2004). Consistent estimation of models defined by conditional moment restrictions. *Econometrica* **72** 1601–1615. MR2078215 https://doi.org/10.1111/j.1468-0262.2004.00545.x

[9] EUBANK, R. L. and HART, J. D. (1992). Testing goodness-of-fit in regression via order selection criteria. *Ann. Statist.* **20** 1412–1425. MR1186256 https://doi.org/10.1214/aos/1176348775

[10] FITZMAURICE, G. M., LIPSITZ, S. R., MOLENBERGHS, G. and IBRAHIM, J. G. (2005). A protective estimator for longitudinal binary data subject to non-ignorable non-monotone missingness. *J. Roy. Statist. Soc. Ser. A* **168** 723–735. MR2205403 https://doi.org/10.1111/j.1467-985X.2005.00374.x

[11] GREENLEES, J. S., REECE, W. S. and ZIESCHANG, K. D. (1982). Imputation of missing values when the probability of response depends on the variable being imputed. *J. Amer. Statist. Assoc.* **77** 251–261.

[12] GRETTON, A., FUKUMIZU, K., TEO, C. H., SONG, L., SCHÖLKOPF, B. and SMOLA, A. (2008). A kernel statistical test of independence. *NeurIPS* **20** 585–592.

[13] GRETTON, A., HERBRICH, R., SMOLA, A., BOUSQUET, O. and SCHÖLKOPF, B. (2005). Kernel methods for measuring independence. *J. Mach. Learn. Res.* **6** 2075–2129. MR2249882

[14] HAN, P. (2014). Multiply robust estimation in regression analysis with missing data. *J. Amer. Statist. Assoc.* **109** 1159–1173. MR3265688 https://doi.org/10.1080/01621459.2014.880058

[15] KANAMORI, T., SUZUKI, T. and SUGIYAMA, M. (2012). Statistical analysis of kernel-based least-squares density-ratio estimation. *Mach. Learn.* **86** 335–367. MR2897527 https://doi.org/10.1007/s10994-011-5266-3

[16] KIM, J. K. (2011). Parametric fractional imputation for missing data analysis. *Biometrika* **98** 119–132. MR2804214 https://doi.org/10.1093/biomet/asq073

[17] KIM, J. K. and YU, C. L. (2011). A semiparametric estimation of mean functionals with nonignorable missing data. *J. Amer. Statist. Assoc.* **106** 157–165. MR2816710 https://doi.org/10.1198/jasa.2011.tm10104

[18] LIANG, K. Y. and ZEGER, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika* **73** 13–22. MR0836430 https://doi.org/10.1093/biomet/73.1.13

[19] LITTLE, R. J. A. and RUBIN, D. B. (2002). *Statistical Analysis with Missing Data*, 2nd ed. *Wiley Series in Probability and Statistics*. Wiley, Hoboken, NJ. MR1925014 https://doi.org/10.1002/9781119013563

[20] LIU, C. (2004). Robit regression: A simple robust alternative to logistic and probit regression. In *Applied Bayesian Modeling and Causal Inference from Incomplete-Data Perspectives. Wiley Ser. Probab. Stat.* 227–238. Wiley, Chichester. MR2138259 https://doi.org/10.1002/0470090456.ch21

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

[22] MA, W.-Q., GENG, Z. and HU, Y.-H. (2003). Identification of graphical models for nonignorable nonresponse of binary outcomes in longitudinal studies. *J. Multivariate Anal.* **87** 24–45. MR2007260 https://doi.org/10.1016/S0047-259X(03)00043-5

[23] MIAO, W., DING, P. and GENG, Z. (2016). Identifiability of normal and normal mixture models with nonignorable missing data. *J. Amer. Statist. Assoc.* **111** 1673–1683. MR3601726 https://doi.org/10.1080/01621459.2015.1105808

[24] MOLENBERGHS, G., BEUNCKENS, C., SOTTO, C. and KENWARD, M. G. (2008). Every missingness not at random model has a missingness at random counterpart with equal fit. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **70** 371–388. MR2424758 https://doi.org/10.1111/j.1467-9868.2007.00640.x

[25] MORIKAWA, K., KIM, J. K. and KANO, Y. (2017). Semiparametric maximum likelihood estimation with data missing not at random. *Canad. J. Statist.* **45** 393–409. MR3729977 https://doi.org/10.1002/cjs.11340

[26] QIN, J., LEUNG, D. and SHAO, J. (2002). Estimation with survey data under nonignorable nonresponse or informative sampling. *J. Amer. Statist. Assoc.* **97** 193–200. MR1947279 https://doi.org/10.1198/016214502753479338

[27] RIDDLES, M. K., KIM, J. K. and IM, J. (2016). Propensity-score-adjustment method for nonignorable nonresponse. *J. Surv. Stat. Methodol*. **97** 215–245.

[28] ROBINS, J. M., ROTNITZKY, A. and 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* (*Minneapolis, MN*, 1997). *IMA Vol. Math. Appl.* **116** 1–94. Springer, New York. MR1731681 https://doi.org/10.1007/978-1-4612-1284-3_1

[29] ROBINS, J. M., ROTNITZKY, A. and ZHAO, L. P. (1994). Estimation of regression coefficients when some regressors are not always observed. *J. Amer. Statist. Assoc*. **89** 846–866. MR1294730

[30] ROTNITZKY, A. and ROBINS, J. M. (1997). Analysis of semi-parametric regression models with non-ignorable non-response. *Stat. Med*. **16** 81–102.

[31] ROTNITZKY, A., SCHARFSTEIN, D., SU, T.-L. and ROBINS, J. (2001). Methods for conducting sensitivity analysis of trials with potentially nonignorable competing causes of censoring. *Biometrics* **57** 103–113. MR1833295 https://doi.org/10.1111/j.0006-341X.2001.00103.x

[32] RUBIN, D. B. (1976). Inference and missing data. *Biometrika* **63** 581–592. MR0455196 https://doi.org/10.1093/biomet/63.3.581

[33] SCHARFSTEIN, D. O., ROTNITZKY, A. and ROBINS, J. M. (1999). Adjusting for nonignorable drop-out using semiparametric nonresponse models. *J. Amer. Statist. Assoc*. **94** 1096–1146. MR1731478 https://doi.org/10.2307/2669923

[34] SEN, A. and SEN, B. (2014). Testing independence and goodness-of-fit in linear models. *Biometrika* **101** 927–942. MR3286926 https://doi.org/10.1093/biomet/asu026

[35] SHAO, J. and WANG, L. (2016). Semiparametric inverse propensity weighting for nonignorable missing data. *Biometrika* **103** 175–187. MR3465829 https://doi.org/10.1093/biomet/asv071

[36] SKRONDAL, A. and RABE-HESKETH, S. (2014). Protective estimation of mixed-effects logistic regression when data are not missing at random. *Biometrika* **101** 175–188. MR3180664 https://doi.org/10.1093/biomet/ast054

[37] SVERCHKOV, M. (2014). A new approach to estimation of response probabilities when missing data are not missing at random. In *Proc. Survey Res. Meth. Sect*., *Am. Statist. Ass*. 867–874. Amer. Statist. Assoc., Alexandria, VA.

[38] TANG, G., LITTLE, R. J. A. and RAGHUNATHAN, T. E. (2003). Analysis of multivariate missing data with nonignorable nonresponse. *Biometrika* **90** 747–764. MR2024755 https://doi.org/10.1093/biomet/90.4.747

[39] TANG, N., ZHAO, P. and ZHU, H. (2014). Empirical likelihood for estimating equations with nonignorably missing data. *Statist. Sinica* **24** 723–747. MR3235396

[40] TERAN HIDALGO, S. J., WU, M. C., ENGEL, S. M. and KOSOROK, M. R. (2018). Goodness-of-fit test for nonparametric regression models: Smoothing spline ANOVA models as example. *Comput. Statist. Data Anal*. **122** 135–155. MR3765820 https://doi.org/10.1016/j.csda.2018.01.004

[41] TSIATIS, A. A. (2006). *Semiparametric Theory and Missing Data*. *Springer Series in Statistics*. Springer, New York. MR2233926

[42] VERBEKE, G., MOLENBERGHS, G., THIJS, H., LESAFFRE, E. and KENWARD, M. G. (2001). Sensitivity analysis for nonrandom dropout: A local influence approach. *Biometrics* **57** 7–14. MR1833286 https://doi.org/10.1111/j.0006-341X.2001.00007.x

[43] WANG, S., SHAO, J. and KIM, J. K. (2014). An instrumental variable approach for identification and estimation with nonignorable nonresponse. *Statist. Sinica* **24** 1097–1116. MR3241279

[44] ZHAO, J. and SHAO, J. (2015). Semiparametric pseudo-likelihoods in generalized linear models with nonignorable missing data. *J. Amer. Statist. Assoc*. **110** 1577–1590. MR3449056 https://doi.org/10.1080/01621459.2014.983234

[45] ZHAO, P., TANG, N., QU, A. and JIANG, D. (2017). Semiparametric estimating equations inference with nonignorable missing data. *Statist. Sinica* **27** 89–113. MR3618161