Research Article

Statistics in Medicine

Received XXXX

(www.interscience.wiley.com) DOI: 10.1002/sim.0000

Using a monotone single-index model to stabilize the propensity score in missing data problems and causal inference

Jing Qin[1], Tao Yu[2], Pengfei Li[3], Hao Liu[4] and Baojiang Chen [5*]

The augmented inverse weighting method is one of the most popular methods for estimating the mean of the response in causal inference and missing data problems. An important component of this method is the propensity score. Popular parametric models for the propensity score include the logistic, probit, and complementary log-log models. A common feature of these models is that the propensity score is a monotonic function of a linear combination of the explanatory variables. To avoid the need to choose a model, we model the propensity score via a semiparametric single-index model, in which the score is an unknown monotonic nondecreasing function of the given single index. Under this new model, the augmented inverse weighting estimator of the mean of the response is asymptotically linear, semiparametrically efficient, and more robust than existing estimators. Moreover, we have made a surprising observation. The inverse probability weighting and augmented inverse weighting estimators based on a correctly specified parametric model may have worse performance than their counterparts based on a nonparametric model. A heuristic explanation of this phenomenon is provided. A real-data example is used to illustrate the proposed methods. Copyright © 2017 John Wiley & Sons, Ltd.

Keywords: Causal inference; Empirical process; Inverse weighting; Missing data; Pool adjacent violation algorithm; Single-index model.

1. Introduction

Causal inference and missing data problems have been extensively researched in recent decades in medical, social and economical sciences (e.g., [1, 2, 3, 4, 5, 6, 7, 8, 9]). Consider a medical experiment with n subjects where each subject is assigned to either the treatment or the control group. Denote by Y_i (T_i) the outcome variable for subject i, if it was assigned to the treatment (control) group. At the end of the study, we observe Y_i or T_i but not both. Let X_i and Δ_i be the corresponding baseline covariates and the treatment indicator respectively; $\Delta_i = 1$ if the ith patient is assigned to the treatment group and therefore Y_i is observed, and $\Delta_i = 0$ otherwise. In summary, the data are denoted $(Y_i, T_i, \Delta_i, X_i), i = 1, \ldots, n$.

We wish to estimate $\mu = E(Y_i)$ and $\nu = E(T_i)$. In the aforementioned medical study, the meanings of these quantities are clear. For example, μ is the population average of the response for the patients in the treatment group. This estimation problem has applications in social science, medical research, economic studies, and other fields (e.g., [1], [2]). For presentational convenience, we will focus on estimation methods for μ; those for ν can be similarly established. Therefore, the observed data are $(\Delta_i Y_i, \Delta_i, X_i)$, $i = 1, \ldots, n$.

We now briefly review existing methods for the estimation of μ. An important quantity is the propensity score, defined to be $\pi(x) = P(\Delta = 1|X = x)$, which is the probability that a subject will be assigned to the treatment group, given the

[1] *National Institute of Allergy and Infectious Diseases, National Institute of Health, Bethesda, MD 20892, US*
[2] *Department of Statistics & Applied Probability, National University of Singapore, 117546, Singapore*
[3] *Department of Statistics and Actuarial Sciences, University of Waterloo, Waterloo, ON, N2L 3G1, Canada*
[4] *Department of Biostatistics, Indiana University School of Medicine, Indianapolis, IN 46202, USA*
[5] *Department of Biostatistics and Data Science, University of Texas Health Science Center at Houston, School of Public Health in Austin, Austin, 78701 U.S.A.*
[*] *Correspondence to: Baojiang Chen, email: baojiang.chen@uth.tmc.edu*

This is the author's manuscript of the article published in final edited form as:

Qin, J., Yu, T., Li, P., Liu, H., & Chen, B. (2019). Using a monotone single-index model to stabilize the propensity score in missing data problems and causal inference. Statistics in Medicine, 38(8), 1442–1458. https://doi.org/10.1002/sim.8048

observed covariate $X = x$. The importance of $\pi(x)$ in the estimation of μ has been discussed by Rosenbaum and Rubin [11]. In this article, we assume $0 < \pi(x) < 1$ to avoid potential technical difficulties.

One popular estimator of μ is the inverse probability weighting estimator ([12]), defined to be

$$\widehat{\mu}_{HT}(\pi(\cdot)) = \frac{1}{n}\sum_{i=1}^{n}\frac{\Delta_i Y_i}{\pi(X_i)}. \tag{1}$$

However, this estimator has a counterintuitive feature ([13]): if $\pi(x)$ takes the parametric form $\pi(x;\beta)$, with $\pi(x;\beta)$ being a known function up to an unknown parameter β, then $\widehat{\mu}_{HT}(\pi(\cdot;\widehat{\beta}))$ could be a more efficient estimator than $\widehat{\mu}_{HT}(\pi(\cdot;\beta_0))$. Here $\widehat{\beta}$ and β_0 denote the maximum likelihood estimator and the true value of β, respectively.

The *augmented inverse weighting estimator* (AIWE; [14]) improves the performance of $\widehat{\mu}_{HT}$ by augmenting a *working model* $\psi(\cdot)$ of the response on the covariates; it does not have the above counterintuitive feature. The AIWE is given by

$$\widehat{\mu}(\pi(\cdot),\psi(\cdot)) = \frac{1}{n}\sum_{i=1}^{n}\left\{\frac{\Delta_i Y_i}{\pi(X_i)} - \frac{\Delta_i - \pi(X_i)}{\pi(X_i)}\psi(X_i)\right\}. \tag{2}$$

A common choice of the working model is $\psi(x) = E(Y|X = x)$. Clearly, the performance of $\widehat{\mu}(\pi(\cdot),\psi(\cdot))$ relies heavily on the choice of $\pi(\cdot)$ and $\psi(\cdot)$. Scharfstein, Rotnitzky, and Robins [15] noted that this estimator is doubly robust in that it is consistent if one of the models for estimating π and ψ, but not both, is misspecified. Wooldridge [16] and Uysalc[17] applied this method in treatment effects models. Sloczynski and Wooldridge [18] provided a unified framework for various doubly robust estimators of the average treatment effect under unconfoundedness. However, Kang and Schafer [19] demonstrated via numerical studies that this estimator can be severely biased if both models are misspecified. Therefore, it is important to specify a flexible but correct model for at least one of π and ψ. Given the importance of the propensity score, we propose a novel and flexible semiparametric model for $\pi(x)$ and study the asymptotic properties of the AIWE of μ.

Note that $\pi(\cdot)$ can be viewed as a regression model for the binary response data $(\Delta_i, X_i), i = 1,\ldots,n$, where the Δ_i's are the response and the X_i's are the covariates. A popular model for $\pi(\cdot)$ is the well-known logistic model:

$$\log\frac{\pi(x)}{1-\pi(x)} = x^T\beta.$$

The probit and complementary log-log models are also widely used for binary response data. The common feature of these parametric models is that $\pi(x)$ is latently assumed to be a monotonic function of a linear combination of the explanatory variables, i.e., $x^T\beta$. To avoid the need to choose a model, we model $\pi(x)$ as a monotonic function of $h^T(x)\beta$, with $h(x)$ being a user-specified function. In particular, we propose the following semiparametric single-index model:

$$\pi(x) = \theta(h^T(x)\beta), \tag{3}$$

where both $\theta(\cdot)$ and β are unknown, and $\theta(\cdot)$ is a monotonic nondecreasing function. Since the form of $\theta(\cdot)$ is not specified, our model is more flexible than parametric models.

Under the setup (3), we consider the estimation of μ within the framework of the AIWE. We first propose to estimate $\theta(\cdot)$ and β by the maximum likelihood method. Compared with the nonparametric methods for estimating $\pi(x)$ (e.g., [20], [5]), our methods do not need tuning parameters. In our numerical studies, we observe that our methods are more accurate in estimating the AIWEs than the nonparametric methods if the linear predictors in the propensity score is correctly specified. Another limitation of the nonparametric methods is the curse of dimensionality: the estimates for $\pi(x)$ may not perform well when the dimension of x is relatively large; in contrast, our proposed methods do not suffer this problem. Compared with the parametric methods (e.g., the logistic regression method) for estimating $\pi(x)$, our methods are more robust. We then consider the estimation of ψ in broad function classes via the weighted least square principle. We show theoretically that with our proposed π estimator and the general ψ estimator based on broad function classes, the AIWE is asymptotically linear and can achieve semiparametric efficiency. We observe that because of the non-smoothness of our estimators for $\pi(x)$, the existing asymptotic theory in the community is not directly applicable to our estimators; the theoretical developments for the linear expansion and the efficiency are nontrivial. Furthermore, we present extensive numerical studies to demonstrate that our estimator has better accuracy than existing estimators.

The organization of the paper is as follows. In Section 2, we present the methods for estimating $\pi(\cdot)$ under the setup (3) with ψ based on broad function classes. In Section 3, we investigate the asymptotic properties of the AIWE based on the proposed semiparametric model. Sections 4 and 5 present the simulation results and a real application, respectively. Section 6 provides concluding remarks. For convenience of presentation, the technical details are given in the Appendix.

2. Estimation Methods for $\pi(\cdot)$ and $\psi(\cdot)$

In this section, we discuss estimation methods for $\pi(\cdot)$ and $\psi(\cdot)$. We first consider the estimation of $\pi(\cdot)$ by the maximum likelihood method. We then discuss the estimation of ψ in broad function classes with the weighted least square principle.

2.1. Estimation Methods for π

Recall that the estimation of $\pi(\cdot)$ can be established by appropriately modeling the binary response data $(\Delta_i, X_i), i = 1, \ldots, n$. Specifically, we consider the maximum likelihood method. The log-likelihood based on $(\Delta_i, X_i), i = 1, \ldots, n$ is given by

$$\ell(\pi(\cdot)) = \sum_{i=1}^{n} \Big[\Delta_i \log\{\pi(X_i)\} + (1 - \Delta_i) \log\{1 - \pi(X_i)\}\Big], \tag{4}$$

and the maximum likelihood estimator of $\pi(\cdot)$ is defined to be

$$\widehat{\pi} = \text{argmax}_{\pi \in \mathcal{F}} \ell(\pi(\cdot)), \tag{5}$$

where $\mathcal{F}$ is a prespecified function class for π. Clearly, $\mathcal{F}$ plays a central role in determining the asymptotic performance of $\widehat{\pi}(\cdot)$.

Under the semiparametric single-index model (3) for $\pi(\cdot)$, the log-likelihood function becomes

$$\ell(\theta, \beta) = \sum_{i=1}^{n} \Big[\Delta_i \log\{\theta(h^T(X_i)\beta)\} + (1 - \Delta_i) \log\{1 - \theta(h^T(X_i)\beta)\}\Big]. \tag{6}$$

Without loss of generality, we assume hereafter that $\theta(\cdot)$ is monotonically increasing. Furthermore, we assume $\beta_1 = 1$ so that the model is identifiable, although in principle other assumptions can be made for identifiability. Then the maximum likelihood estimators for $\theta(\cdot)$ and β are defined as

$$(\widehat{\theta}, \widehat{\beta}) = \text{argmax}_{\theta \in \Theta, \beta \in \Lambda} \ell(\theta, \beta), \tag{7}$$

where $\Theta = \{\theta(\cdot) : 0 \le \theta(x) \le 1$ is monotone in $x \in \mathbb{R}\}$ and $\Lambda = \{1\} \times \Lambda_{-1}$ are the parameter spaces for $\theta(\cdot)$ and β, respectively.

A numerical algorithm for the optimization problem (7) can be established using a similar strategy to that of Cosslett [21]; see also Chen et al. [22] and the references therein. Specifically, we implement the following two-stage algorithm to compute $\widehat{\theta}(\cdot)$ and $\widehat{\beta}$.

Stage 1. For a given β, profile $\theta(\cdot)$ to obtain the profile likelihood of β through the following steps:

(a) Let $(v_1(\beta), \ldots, v_n(\beta))$ be a vector composed of $\{h^T(X_i)\beta : i = 1, \ldots, n\}$, and sort the entries from smallest to largest:

$$v_{(1)}(\beta) \le \ldots \le v_{(n)}(\beta).$$

The corresponding Δ_i's are denoted $\Delta_1(\beta), \ldots, \Delta_n(\beta)$. Substitute $v_{(i)}(\beta)$ and the corresponding $\Delta_i(\beta)$ into (6) to obtain

$$\ell(\theta, \beta) = \sum_{i=1}^{n} \Big[\Delta_i(\beta) \log\{\theta(v_{(i)}(\beta))\} + \{1 - \Delta_i(\beta)\} \log\{1 - \theta(v_{(i)}(\beta))\}\Big].$$

(b) For any β and the $\ell(\theta, \beta)$ given in (a), let

$$\widehat{\theta}_\beta = \text{argmax}_{\theta \in \Theta} \ell(\theta, \beta). \tag{8}$$

Following Dykstra, Kochar, and Robertson [23], solve the maximum problem using the well-known pool-adjacent-violation-algorithm (PAVA; [24]).

(c) Find the profile log-likelihood via

$$\ell(\widehat{\theta}_\beta, \beta) = \sum_{i=1}^{n} \Big[\Delta_i \log\{\widehat{\theta}_\beta(h^T(X_i)\beta)\} + (1 - \Delta_i) \log\{1 - \widehat{\theta}_\beta(h^T(X_i)\beta)\}\Big]. \tag{9}$$

Stage 2. Maximize (9) with respect to β to obtain $\widehat{\beta}$. This step can be implemented using software such as the R function optim() ([25]) with the given initial value of β. Then $\widehat{\theta}(\cdot) = \widehat{\theta}_{\widehat{\beta}}(\cdot)$.

Consequently, we can estimate $\pi(x)$ by $\widehat{\pi}(x) = \widehat{\theta}(h^T(x)\widehat{\beta})$. The following theorem establishes the convergence rate of $\widehat{\pi}(x)$ to its true value $\pi_0(x) = \theta_0(h^T(x)\beta_0)$, where $\theta_0(\cdot)$ is monotonically increasing and is the true value of $\theta(\cdot)$, and β_0 is the true value of β.

Theorem 1 *Let $\mathbb{P}$ be a probability measure. Assuming Conditions 1–3 in the Appendix, we have*

$$\|\widehat{\pi} - \pi_0\|_{2,\mathbb{P}} = O_p(n^{-1/3}),$$

where $\|\cdot\|_{2,\mathbb{P}}$ denotes the L_2 norm under the probability measure $\mathbb{P}$, specifically, for any measurable function f,

$$\|f\|_{2,\mathbb{P}} = \left(\int f^2 d\mathbb{P}\right)^{1/2}.$$

The proposed algorithm performs well when the dimension of β is small or moderate. However, it becomes computationally expensive when this dimension is relatively large. The reason is twofold. First, it involves a two-stage iteration. Second, the profile likelihood given in (8) is not smooth in the neighborhood of $\widehat{\beta}$. On the other hand, in our theoretical development and numerical studies, we observe that the estimate of μ is robust to the estimate of β if it is not too far from $\widehat{\beta}$. Intuitively, this is because $\widehat{\mu}(\pi(\cdot), \psi(\cdot))$ is a smooth function of $\widehat{\pi}(\cdot)$. In practice, when the dimension of β is relatively large, one could implement the following much faster algorithm instead of the two-stage algorithm above.

Step 1. Given the binary response data $(\Delta_i, X_i), i = 1, \ldots, n$, obtain $\widehat{\beta}$ by fitting a parametric model, say a logistic regression model.

Step 2. Similarly to Stage 1(b), estimate θ using the classical PAVA algorithm and the data $(\Delta_i, h^T(X_i)\widehat{\beta}), i = 1, \ldots, n$.

The extensive numerical studies in Section 4 show that the inverse probability weighting estimator $\widehat{\mu}_{HT}(\pi(\cdot))$ with $\widehat{\pi}(\cdot)$ from our method leads to more robust μ estimates than that based on $\widehat{\pi}(\cdot)$ from the parametric methods, even though the assumed parametric model is correct. We now give an intuitive explanation of this observation. Consider the case where $\Delta_i = 1$ but the corresponding $\pi_0(X_i)$ is very close to 0. This may occur occasionally ([26]); in our simulation, with $n = 1000$ observations and 1000 replications, this is likely to occur at least once. In this scenario the parametric estimate of $\pi(\cdot)$, denoted $\pi(\cdot; \widehat{\beta})$, is likely also close to 0, because it is a consistent estimator of $\pi_0(\cdot)$. In our simulation, $\pi_0(\cdot)$ and $\pi(\cdot; \widehat{\beta})$ can be as small as 9×10^{-7}. Since $\pi(\cdot)$ appears in the denominator in the estimator $\widehat{\mu}_{HT}(\pi(\cdot))$—see (1)—these observations may significantly affect the accuracy of $\widehat{\mu}_{HT}(\pi(\cdot; \widehat{\beta}))$. With our proposed $\widehat{\pi}(\cdot)$ estimate, however, $\widehat{\mu}_{HT}(\widehat{\pi}(\cdot))$ is much more robust. This is because when $\Delta_i = 1$, the corresponding $\widehat{\theta}(h^T(X_i)\widehat{\beta})$ from the PAVA algorithm is greater than $1/n$. Therefore, the accuracy of $\widehat{\mu}_{HT}(\widehat{\pi}(\cdot))$ is much less affected by any individual observation.

2.2. *Estimation Methods for ψ*

Estimation methods for $\psi(\cdot)$ have been widely discussed (e.g., [27], [14], [15]). In this paper, we generally consider the weighted least square objective function given by

$$Q(\psi) = \sum_{i=1}^{n} w(\Delta_i, X_i)\{Y_i - \psi(X_i)\}^2,$$

where $w(\Delta_i, X_i)$ is the user-specified weight function. Assume that ψ is estimated by

$$\widehat{\psi} = \text{argmin}_{\psi \in \Psi} Q(\psi), \tag{10}$$

where Ψ denotes the class of functions for the "guessed" working model.

Cao, Tsiatis, and Davidian [27] suggest the following parametric model for $\psi(\cdot)$ and form for $w(\cdot, \cdot)$:

(1) $\psi(x) = h(x; \gamma)$, with h a known function and γ the unknown Euclidean parameter. Therefore, $\Psi = \{\psi : \psi(x) = h(x; \gamma), \gamma \in \mathbb{R}^k\}$, where k is the dimension of γ.
(2) $w(\delta, x) = \frac{\delta\{1-\widehat{\pi}(x)\}}{\widehat{\pi}^2(x)}$, where $\widehat{\pi}$ is obtained from (5).

In our simulation study and our analysis of the real-data example, we follow the above suggestions.

3. Asymptotic Behavior of the Augmented Inverse Weighting Estimator

With the estimators $\widehat{\pi}$ and $\widehat{\psi}$ given in the last section, the AIWE of μ is given by

$$\begin{aligned}\widehat{\mu}(\widehat{\pi}(\cdot),\widehat{\psi}(\cdot)) &= \frac{1}{n}\left\{\sum_{i=1}^{n}\frac{\Delta_i Y_i}{\widehat{\pi}(X_i)}-\frac{\Delta_i-\widehat{\pi}(X_i)}{\widehat{\pi}(X_i)}\widehat{\psi}(X_i)\right\}\\ &= \mathbb{P}_n\phi(\boldsymbol{v};\mu_0,\widehat{\pi}(\cdot),\widehat{\psi}(\cdot))+\mu_0, \qquad (11)\end{aligned}$$

where $\boldsymbol{v}=(\delta,y,x^T)^T$,

$$\phi(\boldsymbol{v};\mu,\pi(\cdot),\psi(\cdot))=\frac{\delta y}{\pi(x)}-\frac{\delta-\pi(x)}{\pi(x)}\psi(x)-\mu,$$

and $\mathbb{P}_n$ is an empirical measure such that $\mathbb{P}_n g(\boldsymbol{v})=\int g(\boldsymbol{v})d\mathbb{P}_n$ for any function $g(\boldsymbol{v})$. For simplicity, we denote this estimator as $\widehat{\mu}$ when the context is clear.

We first explore the asymptotic behavior of $\widehat{\mu}$ when $\widehat{\pi}$ and $\widehat{\psi}$ are assumed to be estimated from (5) and (10) respectively. The asymptotic behavior of $\widehat{\pi}$ and $\widehat{\psi}$ are significantly affected by the complexity of the function classes $\mathcal{F}$ and Ψ. The following entropy conditions, which combine the main parts of Conditions A and B in the Appendix, play key roles in the proof of Theorem 2.

Entropy Conditions: There exist $0<\alpha_1,\alpha_2<2$ such that for every $\epsilon>0$

$$H_{2,B}(\epsilon,\mathcal{F},F_X)<A_1\epsilon^{-\alpha_1} \quad \text{and} \quad H_{2,B}(\epsilon,\Psi,F_X)<A_2\epsilon^{-\alpha_2},$$

where F_X denotes the cumulative distribution function of the covariates X, and A_1 and A_2 are universal constants.

Here $H_{2,B}(\epsilon,\mathcal{F},F_X)$ is the ϵ-entropy with bracketing of $\mathcal{F}$, which is commonly adopted in empirical process texts. We give a quick review of $H_{2,B}(\epsilon,\mathcal{F},F_X)$ in the Appendix.

Theorem 2 *Let $\psi_0(x)$ be the true value of $\psi(x)$. Assuming Conditions A–C in the Appendix, we have*

$$\begin{aligned}&\sqrt{n}(\widehat{\mu}-\mu_0)-\sqrt{n}\mathbb{P}_n\phi(\boldsymbol{v};\mu_0,\pi_0,\psi_0)+\sqrt{n}\mathbb{P}\left[\left\{\frac{\pi_0(x)}{\widehat{\pi}(x)}-1\right\}\left\{\widehat{\psi}(x)-\psi_0(x)\right\}\right]\\ =\;& O_p\left(\|\widehat{\pi}-\pi_0\|_{2,\mathbb{P}}^{1-\alpha_1/2}\right)+O_p\left(\|\widehat{\psi}-\psi_0\|_{2,\mathbb{P}}^{1-\max\{\alpha_1,\alpha_2\}/2}\right)+o_p(1), \qquad (12)\end{aligned}$$

where $\mathbb{P}g(\boldsymbol{v})=\int g(\boldsymbol{v})d\mathbb{P}$ for any function $g(\boldsymbol{v})$.

Corollary 1 *Assuming Conditions A–C in the Appendix, we have*

(P1) if

$$\sqrt{n}\mathbb{P}\left[\left\{\frac{\pi_0(x)}{\widehat{\pi}(x)}-1\right\}\left\{\widehat{\psi}(x)-\psi_0(x)\right\}\right]=O_p(1),$$

then $\widehat{\mu}-\mu_0=O_p(n^{-1/2})$;

(P2) if $\|\widehat{\pi}-\pi_0\|_{2,\mathbb{P}}=o_p(1)$, $\|\widehat{\psi}-\psi_0\|_{2,\mathbb{P}}=o_p(1)$, and

$$\sqrt{n}\mathbb{P}\left[\left\{\frac{\pi_0(x)}{\widehat{\pi}(x)}-1\right\}\left\{\widehat{\psi}(x)-\psi_0(x)\right\}\right]=o_p(1),$$

then $\sqrt{n}(\widehat{\mu}-\mu_0)=\sqrt{n}\mathbb{P}_n\phi(\boldsymbol{v};\mu_0,\pi_0,\psi_0)+o_p(1)$, and $\widehat{\mu}$ achieves the semiparametric information bound.

Remark 1 *In the development of Theorem 2 and Corollary 1, we do not require that $\pi_0\in\mathcal{F}$ and $\psi_0\in\Psi$. That is, Models (5) and (10) need not be the true models of π and ψ.*

Remark 2 *Calculations for entropies of function classes are available in empirical process texts. Based on these, the entropy condition $H_{2,B}\le A\epsilon^{-\alpha}$ for some universal constant A and $0<\alpha<2$ accommodates many "good" function classes. For example, for most parametric models, the corresponding function classes satisfy this condition with $\alpha>0$; the class of bounded monotone functions satisfies this condition with $\alpha=1$; the function class with every element g satisfying $g:[0,1]\to[0,1]$ such that $\int_0^1\{g^{(m)}(x)\}^2dx\le 1$ satisfies this condition with $\alpha=1/m$. See Section 2.2 of van de Geer [28], Sections 2.6 and 2.7 of van der Vaart and Wellner [29], Chapter 9 of Kosorok [30], and the references therein for more classes of functions that satisfy this entropy condition.*

Theorem 3 *Let $\widehat{\pi}(x) = \widehat{\theta}(h^T(x)\widehat{\beta})$, where $(\widehat{\theta}, \widehat{\beta})$ are obtained from (7). Assume Conditions 1–3, and B–E in the Appendix. We have*

$$\sqrt{n}(\widehat{\mu} - \mu_0) - \sqrt{n}\mathbb{P}_n\phi(\boldsymbol{v}; \mu_0, \pi_0, \psi_0) = o_p(1).$$

Remark 3 *From this theorem, we observe that in our method the estimator $\widehat{\mu}$ is semiparametrically efficient under mild regularity conditions. The main regularity conditions are (1) $\pi_0(x)$ is monotone, and (2) $\psi_0(x)$ belongs to a function class that satisfies a mild entropy condition, say Condition* **B1**. *Furthermore, if the parametric model and the single-index model (3) for $\pi(x)$ are both correctly specified, our method and the AIWE based on the correctly specified parametric model have the same efficiency. This is true even when we obtain the estimate of $\pi(x)$ via the second algorithm in Section 2.1, where we first obtain the estimate $\widehat{\beta}$ by fitting a parametric model and then estimate $\theta(\cdot)$ using the PAVA based on the data $(\Delta_i, h(X_i)^T\widehat{\beta}),\ i = 1, \ldots, n$. That is, compared with the AIWE based on the correctly specified parametric model, our method does not lose any efficiency provided the single-index model is correctly specified. If this condition is not satisfied, our estimator is root-n consistent given the even more mild condition required by (P1) of Corollary 1. In particular, our $\widehat{\mu}$ maintains the double robustness property as desired and is more robust than the AIWE based on the parametric model.*

4. Simulation Studies

In this section we conduct simulation studies to explore the performance of the proposed estimators. We consider the following eight estimators:

1) The inverse probability weighting estimator (1) with $\pi(x)$ estimated by the logistic regression model reviewed in Section 2.1; we call this method *HT-par*.
2) The inverse probability weighting estimator with $\pi(\cdot)$ estimated by the semiparametric method proposed in Section 2.1, i.e., $\pi(x) = \theta(h^T(x)\beta)$. Here we use the second algorithm in Section 2.1 to estimate $\pi(\cdot)$. We call this method *HT-pava*.
3) The Hájek inverse probability weighting estimator ([31]), with $\pi(x)$ estimated as in *HT-pava*; we call this method *HAJ-pava*. In particular,
$$\widehat{\mu}_{\text{HAJ-pava}} = \frac{\sum_{i=1}^n \Delta_i Y_i/\widehat{\pi}(X_i)}{\sum_{i=1}^n \Delta_i/\widehat{\pi}(X_i)}.$$
4) The estimator $\widehat{\mu}_{\text{PROJ}}$ proposed in Cao, Tsiatis, and Davidian [27]; we call this method *Cao-proj*.
5) The AIWE with $\pi(x)$ estimated by the using the generalized additive model [10] and $\psi(x)$ estimated by the method presented in Section 2.2; we call this method *NP-AIWE*. Specifically, we consider the following nonparametric model for estimating the propensity score:
$$\text{logit}\,\{P(\Delta = 1|x_1, \ldots, x_p)\} = \sum_{j=1}^p h_j(x_j),$$
where $h_j(\cdot)$, $j = 1, \ldots, p$ are nonparametric functions; the natural cubic-spline is applied to estimate them.
6) The propensity score matching method ([11]) using different number of cutpoints J; we call this method *PSM-J*. In this study, we use $J = 5$, 10, and 20.
7) The AIWE with $\pi(x)$ estimated by the first algorithm in Section 2.1 and $\psi(x)$ estimated by the method presented in Section 2.2; we call this method *New-pava1*.
8) The AIWE with $\pi(x)$ estimated by the second algorithm in Section 2.1 and $\psi(x)$ estimated by the method presented in Section 2.2; we call this method *New-pava2*.

We consider three examples.

Example 1. We first consider the artificial example created by Kang and Schafer [19]. In this scenario, for each i $(i = 1, ..., n)$, $Z_i = (Z_{i1}, Z_{i2}, Z_{i3}, Z_{i4})^T$ is generated as standard multivariate normal random variables, and the elements of $X_i = (X_{i1}, X_{i2}, X_{i3}, X_{i4})^T$ are defined as $X_{i1} = \exp(Z_{i1}/2), X_{i2} = Z_{i2}/\{1 + \exp(Z_{i1})\} + 10, X_{i3} = (Z_{i1}Z_{i3}/25 + 0.6)^3$, and $X_{i4} = (Z_{i2} + Z_{i4} + 20)^2$, so that Z_i may be expressed in terms of X_i . For each i,

$$Y_i = 210 + 27.4Z_{i1} + 13.7Z_{i2} + 13.7Z_{i3} + 13.7Z_{i4} + \epsilon_i,$$

where ϵ_i is independent standard normal, and Δ_i is generated as a Bernoulli random variable with the true propensity score being

$$\text{logit}\{\pi_i(Z)\} = -Z_{i1} + 0.5Z_{i2} - 0.25Z_{i3} - 0.1Z_{i4},$$

Table 1. Comparison of the bias, RMSE, and MAE for the estimates of μ in Example 1. Here "P-C" (or "P-I") indicates that the model for $\pi(\cdot)$ is correctly (or incorrectly) specified; "W-C" (or "W-I") indicates that the model for $\psi(\cdot)$ is correctly (or incorrectly) specified.

	P-I, W-I			P-I, W-C			P-C, W-I			P-C, W-C		
Method	Bias	RMSE	MAE	Bias	RMSE	MAE	Bias	RMSE	MAE	Bias	RMSE	MAE
					$n = 200$							
HT-par	22.28	77.58	9.36	18.46	92.82	9.24	74.83	3735.34	9.38	-0.46	12.73	6.71
HT-pava	11.13	11.61	11.06	-11.18	11.64	11.14	-15.38	15.80	15.29	-10.11	10.66	10.16
HAJ-pava	-2.31	4.18	2.95	-2.38	4.23	3.00	-2.28	4.18	3.01	-0.84	3.55	2.37
Cao-proj	-0.45	3.37	2.21	-0.04	2.56	1.75	-0.00	2.56	1.74	-0.02	2.54	1.72
NP-AIWE	-0.83	3.31	2.21	-0.15	2.55	1.74	-0.55	3.23	2.22	-0.16	2.58	1.77
PSM-5	-4.99	5.94	5.04	-4.99	5.94	5.04	-3.08	4.37	3.32	-3.08	4.37	3.32
PSM-10	-4.49	5.65	4.60	-4.48	5.65	4.60	-2.52	4.15	2.90	-2.52	4.15	2.90
PSM-20	-4.46	5.66	4.47	-4.46	5.66	4.47	-2.50	4.25	2.99	-2.50	4.25	2.99
New-pava1	-0.38	3.29	2.22	-0.04	2.56	1.75	0.00	2.56	1.73	-0.02	2.53	1.71
New-pava2	-1.15	3.13	2.17	-0.04	2.56	1.75	0.00	2.56	1.75	-0.02	2.53	1.71
					$n = 1000$							
HT-par	48.63	643.67	13.63	54.93	530.75	13.87	36.72	177.45	13.93	0.04	4.85	2.86
HT-pava	-13.25	13.33	13.30	-10.17	10.28	10.17	-13.29	13.38	13.28	-10.00	10.13	10.01
HAJ-pava	-1.82	2.37	1.86	-1.83	2.38	1.91	-1.83	2.38	1.88	-0.20	1.51	1.00
Cao-proj	-1.24	1.78	1.32	-0.01	1.15	0.78	-0.02	1.15	0.78	0.00	1.13	0.78
NP-AIWE	-1.57	2.10	1.62	-0.01	1.14	0.77	-0.19	1.46	0.98	-0.02	1.13	0.76
PSM-5	-3.22	3.51	3.24	-3.22	3.51	3.24	-1.44	1.95	1.50	-1.44	1.95	1.50
PSM-10	-2.59	2.96	2.62	-2.59	2.96	2.62	-0.83	1.60	1.11	-0.83	1.60	1.11
PSM-20	-2.36	2.76	2.39	-2.36	2.76	2.39	-0.60	1.54	1.02	-0.60	1.54	1.02
New-pava1	-0.77	1.72	1.18	-0.01	1.15	0.77	-0.02	1.15	0.78	0.00	1.13	0.78
New-pava2	-1.50	1.96	1.51	-0.01	1.15	0.77	-0.02	1.15	0.78	0.00	1.13	0.78

where $\text{logit}(t) = \log\{t/(1-t)\}$ for $t \in (0,1)$. The true value of μ is $\mu_0 = 210$.

If we fit a linear regression model for Y_i over Z_i and a logistic regression model for Δ_i over Z_i, then the models for $\psi(\cdot)$ and $\pi(\cdot)$ are correctly specified. However, if Z_i is replaced by X_i in the above fitted models, then the models for $\psi(\cdot)$ and $\pi(\cdot)$ are incorrectly specified. In total, we have four combinations of model specifications for $\psi(\cdot)$ and $\pi(\cdot)$. For each combination, we calculate the bias, square root mean square error (RMSE), and median absolute error (MAE) of the ten estimates of μ based on 5000 repetitions. We consider two sample sizes, 200 and 1000; the results are reported in Table 1. We make the following observations:

- Clearly, *Cao-proj*, *New-pava1*, and *New-pava2* have similar performance. They perform better than the other estimators.
- NP-AIWE performs comparable or slightly worse than our methods.
- *HT-pava* has a much smaller RMSE than its counterpart *HT-par*. However, the MAEs are comparable for $n = 200$.
- Although *HAJ-pava* does not use the working regression model, its performance is close to that of *Cao-proj* and the new estimators.

To illustrate the robustness of our methods, we give two further examples.

Example 2. The "working" propensity score function is given by the logistic regression $\text{logit}\{P(\Delta = 1|x_1, x_2)\} = \beta_0 + \beta_1 x_1 + \beta_2 x_2$. The true propensity score is given by the following three models:

Model I:

$$\text{logit}\{P(\Delta = 1|x_1, x_2)\} = 2x_1 - x_2.$$

In this case the propensity score is correctly specified.

Model II:

$$\text{logit}\{P(\Delta = 1|x_1, x_2)\} = 2x_1 - x_2 + 2x_1^2;$$

A quadratic term is missing in the propensity score.

Table 2. Comparison of the bias, RMSE, and MAE for the estimates of μ in Example 2.

Method	Model I			Model II			Model III		
	Bias	RMSE	MAE	Bias	RMSE	MAE	Bias	RMSE	MAE
				$n = 200$					
HT-par	-0.05	1.85	0.52	0.00	0.24	0.14	0.00	0.21	0.13
HT-pava	-0.63	0.75	0.65	-0.15	0.25	0.17	-0.11	0.23	0.16
HAJ-pava	-0.21	0.32	0.26	-0.01	0.16	0.11	-0.00	0.16	0.11
Cao-proj	-0.09	0.46	0.24	0.01	0.17	0.11	-0.00	0.16	0.11
NP-AIWE	-0.29	0.41	0.30	0.06	0.17	0.11	0.05	0.17	0.11
PSM-5	-0.24	0.34	0.27	-0.11	0.19	0.14	-0.10	0.19	0.13
PSM-10	-0.24	0.34	0.25	-0.10	0.19	0.13	-0.09	0.18	0.13
PSM-20	-0.15	0.31	0.20	-0.01	0.16	0.11	-0.01	0.16	0.11
New-pava1	-0.14	0.30	0.19	0.00	0.16	0.10	-0.01	0.16	0.10
New-pava2	-0.14	0.30	0.20	-0.00	0.16	0.11	-0.01	0.16	0.10
				$n = 1000$					
HT-par	0.01	1.79	0.32	0.00	0.10	0.07	0.00	0.09	0.06
HT-pava	-0.34	0.41	0.35	-0.05	0.10	0.07	-0.04	0.09	0.06
HAJ-pava	-0.14	0.20	0.16	-0.00	0.08	0.05	-0.00	0.07	0.05
Cao-proj	-0.04	0.15	0.09	0.00	0.07	0.05	-0.00	0.07	0.05
NP-AIWE	-0.15	0.22	0.16	0.06	0.10	0.07	0.05	0.09	0.06
PSM-5	-0.18	0.21	0.18	-0.05	0.09	0.06	-0.04	0.08	0.06
PSM-10	-0.11	0.17	0.12	-0.03	0.08	0.05	-0.03	0.08	0.05
PSM-20	-0.11	0.17	0.12	-0.03	0.08	0.05	-0.03	0.08	0.05
New-pava1	-0.06	0.14	0.09	-0.00	0.07	0.04	-0.01	0.07	0.04
New-pava2	-0.06	0.14	0.09	-0.00	0.07	0.04	-0.00	0.07	0.04

Model III:

$$\text{logit}\{P(\Delta = 1|x_1, x_2)\} = 2x_1 - x_2 - x_1x_2 + 2x_1^2.$$

An interaction term and a quadratic term are missing.

The regression model for Y is given by

$$Y = 3 + x_1^2 + x_2 + \epsilon,$$

where x_1, x_2, and ϵ are independent standard normal random variables. Hence, the true value of μ is $\mu_0 = 4$. The "working model" is

$$\psi(x) = \gamma_0 + \gamma_1 x_1 + \gamma_2 x_2.$$

For the proposed methods, the parameters are estimated by the weighted least square method described in Section 2.2.

We again consider two sample sizes: $n = 200, 1000$. Table 2 gives the bias, RMSE, and MAE of all the estimates of μ for 5000 repetitions in Table 2. We make the following observations:

- For model I, where the propensity score is correctly specified, the proposed *New-pava1, New-pava2* methods perform better than the other methods.
- For models II and III, where the propensity score is misspecified, the proposed methods perform comparable or slightly better than other methods.

Example 3. To evaluate the robustness of the proposed method to the misspecification of the link function in the propensity score, we consider the following example. The setup is the same as in example 2, except that we posit the following propensity models

Model I:

$$\log[-\log\{1 - P(\Delta = 1|x_1, x_2)\}] = 2x_1 - x_2;$$

Model II:

$$\log[-\log\{1 - P(\Delta = 1|x_1, x_2)\}] = 2x_1 - x_2 + 2x_1^2;$$

Table 3. Comparison of the bias, RMSE, and MAE for the estimates of μ in Example 3.

Method	Model I			Model II			Model III		
	Bias	RMSE	MAE	Bias	RMSE	MAE	Bias	RMSE	MAE
				$n = 200$					
HT-par	1.70	21.45	0.62	0.21	0.61	0.18	0.09	0.37	0.14
HT-pava	-0.65	0.76	0.67	-0.15	0.25	0.17	-0.10	0.21	0.14
HAJ-pava	-0.22	0.33	0.26	-0.02	0.16	0.11	-0.01	0.16	0.10
Cao-proj	-0.09	0.62	0.28	0.00	0.17	0.11	0.00	0.16	0.10
NP-AIWE	-0.31	0.40	0.31	0.05	0.16	0.11	0.04	0.16	0.10
PSM-5	-0.26	0.35	0.27	-0.20	0.26	0.20	-0.20	0.26	0.20
PSM-10	-0.25	0.34	0.26	-0.18	0.25	0.18	-0.18	0.25	0.19
PSM-20	-0.26	0.34	0.26	-0.17	0.24	0.18	-0.18	0.25	0.18
New-pava1	-0.12	0.30	0.19	0.00	0.16	0.11	-0.01	0.16	0.10
New-pava2	-0.12	0.29	0.20	0.00	0.16	0.11	-0.01	0.15	0.10
				$n = 1000$					
HT-par	2.12	7.74	0.71	0.21	0.31	0.16	0.11	0.20	0.09
HT-pava	-0.34	0.41	0.35	-0.05	0.10	0.07	-0.03	0.08	0.06
HAJ-pava	-0.13	0.20	0.16	-0.01	0.07	0.05	-0.01	0.07	0.05
Cao-proj	-0.02	0.19	0.12	0.00	0.07	0.05	0.00	0.07	0.05
NP-AIWE	-0.15	0.21	0.15	0.05	0.09	0.06	0.04	0.08	0.05
PSM-5	-0.19	0.21	0.19	-0.14	0.16	0.14	-0.14	0.16	0.14
PSM-10	-0.11	0.17	0.12	-0.12	0.14	0.11	-0.13	0.15	0.13
PSM-20	-0.10	0.16	0.12	-0.11	0.13	0.10	-0.12	0.14	0.12
New-pava1	-0.05	0.14	0.09	0.00	0.07	0.05	0.00	0.07	0.05
New-pava2	-0.05	0.13	0.09	0.00	0.07	0.05	0.00	0.07	0.05

Model III:

$$\log[-\log\{1 - P(\Delta = 1|x_1, x_2)\}] = 2x_1 - x_2 - x_1x_2 + 2x_1^2.$$

Table 3 gives the results for 5000 repetitions. Similar scenarios to Example 2 are observed; the details are omitted for brevity.

5. Applications

In this section, we apply our methods to the AIDS Clinical Trials Group Study 175 (ACTG 175; [32]). ACTG 175 is a randomized clinical trial comparing monotherapy (zidovudine or didanosine) with combination therapy (zidovudine and didanosine, or zidovudine and zalcitabine) in adults infected with the type-I HIV virus with CD4 T cell counts between 200 and 500 per cubic millimeter. The study included 2139 HIV-positive subjects. These subjects were followed for about 96 weeks, and the CD4 T cell counts were measured at week 20 and week 96. Some cell counts at week 96 were missing because of subject dropout; the missing proportion is 37.3%. The baseline covariates and the week-20 counts were always observed. The covariates include: age at baseline (in years), weight at baseline (in kg), hemophilia (0=no, 1=yes), homosexual activity (0=no, 1=yes), history of intravenous drug use (0=no, 1=yes), Karnofsky score (on a scale of 0–100), non-zidovudine antiretroviral therapy prior to initiation of study treatment (0=no, 1=yes), zidovudine use in the 30 days prior to treatment initiation (0=no, 1=yes), zidovudine use prior to treatment initiation (0=no, 1=yes), number of days of previously received antiretroviral therapy, race (0=white, 1=nonwhite), gender (0=female, 1=male), antiretroviral history (0=naive, 1=experienced), antiretroviral history stratification (1=naive, 2=1 to 52 weeks of prior antiretroviral therapy, 3=more than 52 weeks), symptomatic indicator (0=asymptomatic, 1=symptomatic), treatment indicator (0=zidovudine only, 1=other therapies), and indicator of whether or not patient was taken off treatment before 96 $\pm$ 5 weeks (0=no, 1=yes). The details can be found in https://cran.r-project.org/web/packages/speff2trial/speff2trial.pdf.

We are interested in comparing the cell counts at week 96 for two groups: a) zidovudine only and b) other therapies. Specifically, we are interested in testing $H_0 : \delta = 0$, where $\delta = \mu - \nu$, μ is the mean CD4 T cell count at week 96 in the zidovudine group, and ν is the mean CD4 T cell count at week 96 in the other group. We apply the six methods of Section

Table 4. Estimation of the marginal means of the CD4 T cell counts at week 96 in two groups, the zidovudine-only group and the other-therapy group, and estimation of the difference of the marginal means between these groups

	Zidovudine only		Other therapies		Difference	
Method	Mean	SE	Mean	SE	Mean	SE
HT-par	274.03	11.42	322.55	5.59	-48.52	12.60
HT-pava	287.62	10.28	338.23	6.70	-50.61	12.16
HAJ-pava	272.70	11.37	323.13	5.60	-50.43	12.61
Cao-proj	270.20	10.34	321.87	5.21	-51.67	11.42
NP-AIWE	270.50	10.12	322.81	5.29	-52.31	11.47
PSM-5	274.20	11.36	323.03	5.84	-48.83	12.76
PSM-10	272.60	11.65	321.56	5.84	-48.96	13.09
PSM-20	266.53	12.15	318.78	6.00	-52.26	13.62
New-pava1	269.98	10.11	322.76	5.22	-52.78	11.20
New-pava2	269.05	10.10	321.99	5.20	-52.94	11.19

4. The standard errors are obtained using 3000 bootstrap samples. The covariates listed above were included in the model for the propensity score of the missing probability and the working regression model of the response. Table 4 reports the estimates of the marginal means of the cell counts at week 96 in the two groups and the difference in the cell counts at week 96 between the groups. *Cao-proj*, *NP-AIWE*, *New-pava1*, and *New-pava2* yield similar estimates and standard errors for the cell count differences, and the estimates differ from the estimates of the other methods. All the methods indicate that the marginal mean of the cell counts at week 96 is lower in the zidovudine group.

6. Concluding remarks

In causal inference and the missing data problem, the response mean is important. The statistical, economic, and epidemiological literature has many different estimation methods, such as the inverse probability weighting estimator ([12]) and the AIWE (Robins, Rotnitzky, and Zhao 1994). Other estimation methods such as propensity matching ([11]) are also available. For a comprehensive discussion of the related problems see the monograph by Imbens and Rubin [33].

It is well known that the propensity score $\pi(x)$ and the "working regression model" $\psi(x)$ play extremely important roles in the mean estimation. In this paper we have provided a propensity score method that is almost nonparametric by using the monotonic single-index model. In contrast to doubly robust methods, where both the propensity score and the regression model have to be modeled as accurately as possible, our method requires us to model the regression model $\psi(x) = E(Y|X = x)$ carefully but leaves the propensity score almost nonparametric. It is encouraging that the inverse probability weighting estimators and AIWEs based on the nonparametrically fitted single-index model perform better than their counterparts based on the fitted correctly specified parametric model. To ease the computational burden in the semiparametric maximum likelihood estimation of $\pi(x)$ for high-dimensional covariates X, we propose first obtaining the estimate $\widehat{\beta}$ using a "working parametric propensity score model," say the logistic regression model. We can then estimate the link function $\theta(\cdot)$ using the PAVA based on the data $(\Delta_i, h(X_i)^T\widehat{\beta}),\ i = 1, 2, ..., n$. We can successfully stabilize the propensity score when it is too close to zero. Examples 2 and 3 showed that our method is more robust than the methods based on the fitted correctly specified parametric model. Further investigation of this method in regression parameter estimation based on missing data would be worthwhile.

Acknowledgements

The authors thank the editor, the associate editor, and the referee for constructive comments and suggestions that lead to a significant improvement over the article. Dr. Yu's research is supported in part by Singapore Ministry Education Academic Research Fund Tier 1. Dr. Li was supported in part by the Natural Sciences and Engineering Research Council of Canada, RGPIN-2015-06592. Dr. Liu's research is supported in part by National Institutes of Health grant, NIH P30 CA082709.

Appendix: Proofs of Theorems

Technical Conditions

We need the following conditions in the proof of Theorem 1.

Condition 1: Λ_{-1} is a compact subspace of $\mathbb{R}^{p-1}$.

Condition 2: There exists a constant C_0 independent of u such that

$$\mathbb{P}\left|I\left(\beta_1^\tau h(x) \le u\right) - I\left(\beta_2^\tau h(x) \le u\right)\right| \le C_0||\beta_1 - \beta_2||_1,$$

where $\|\cdot\|_1$ denotes the l_1 norm.

Condition 3: $\theta_0(\cdot) \in \Theta$ and $\inf_x \theta_0(v(x;\beta_0)) > 0$.

We need the following conditions in the proof of Theorem 2.

Condition A: The following technical conditions are for "π":

A1: There exists $0 < \alpha_1 < 2$ such that for every $\epsilon > 0$,

$$H_{2,B}(\epsilon, \mathcal{F}, F_X) < A_1\epsilon^{-\alpha_1},$$

where A_1 is a universal constant.

A2: $\inf_{\pi\in\mathcal{F}} \inf_x |\pi(x)| > 0$.

A3: $\inf_x \pi_0(x) > 0$.

Condition B: The following technical conditions are for "ψ":

B1: There exists $0 < \alpha_2 < 2$ such that for every $\epsilon > 0$,

$$H_{2,B}(\epsilon, \Psi, F_X) < A_2\epsilon^{-\alpha_2},$$

where A_2 is a universal constant.

B2: $\sup_{\psi\in\Psi} \sup_x |\psi(x)| < \infty$.

B3: $\sup_x \psi_0(x) < \infty$.

Condition C: We further assume that the support of $f_Y(y)$ is bounded, where $f_Y(y)$ denotes the marginal density of Y.

Condition D: $\psi_0 \in \Psi$, $\inf_x w(1,x) > 0$, and $\sup_{\delta,x} w(\delta,x) < \infty$.

Condition E: There exists a constant $c > 0$ such that $\inf_x \theta(h^T(x)\beta) > c > 0$.

Preliminaries

The proofs of the theorems rely heavily on empirical process theory. We adopt the usual notation in the literature. In particular, let "$\lesssim$" ("$\gtrsim$") denote smaller (greater) than, up to a universal constant. Recall that $\mathbb{P}_n$ and $\mathbb{P}$ are empirical and probability measures, and for any function $g(\boldsymbol{v})$ and independent and identically distributed observations $\boldsymbol{V}_1, \ldots, \boldsymbol{V}_n$,

$$\begin{aligned}\mathbb{P}_n(g(\boldsymbol{v})) &= \int g(\boldsymbol{v})d\mathbb{P}_n = \frac{1}{n}\sum_{i=1}^n g(\boldsymbol{V}_i);\\ \mathbb{P}(g(\boldsymbol{v})) &= \int g(\boldsymbol{v})d\mathbb{P}.\end{aligned}$$

Furthermore, we denote by $\|g\|_{q,\mathbb{P}}$ the L_q normal of g under $\mathbb{P}$. In particular, $\|g\|_{q,\mathbb{P}} = \left\{\int g^q d\mathbb{P}\right\}^{1/q}$.

We need the following definition of entropy for function classes, which plays a key role in modern empirical process theory. It is adapted from Definition 2.2 in van de Geer [28].

Definition 1 *For any $\epsilon > 0$ and $q > 0$, let $N_{q,B}(\epsilon, \mathcal{G}, \mathbb{P})$ be the smallest value of N for which there exists a set of pairs of functions $\{(g_j^L, g_j^U)\}_{j=1}^N$ such that (i) $\|g_j^U - g_j^L\|_{q,\mathbb{P}} \le \epsilon$, where $\|g_j^U - g_j^L\|_{q,\mathbb{P}} = \left\{\int |g_j^U - g_j^L|^q d\mathbb{P}\right\}^{1/q}$ and (ii) for any $g \in \mathcal{G}$, there exists a $j = j(g)$ such that*

$$g_j^L \le g \le g_j^U.$$

$N_{q,B}(\epsilon, \mathcal{G}, \mathbb{P})$ is called the ϵ-bracketing number of $\mathcal{G}$, and $H_{q,B}(\epsilon, \mathcal{G}, \mathbb{P}) = \log N_{q,B}(\epsilon, \mathcal{G}, \mathbb{P})$ is called the ϵ-entropy with bracketing of $\mathcal{G}$.

Proof of Theorem 1

The proof of this theorem follows the same lines as part (a) of Theorem 1 in Chen et al. [22]; we omit it for brevity.

Proof of Theorem 2

To facilitate our proof, we need the following lemma, which is a direct application of Lemma 5.13 in van de Geer [28].

Lemma 1 *Assume*

$$\sup_{g\in\mathcal{G}}|g-g_0|_\infty\le 1,\qquad H_{2,B}(\epsilon,\mathcal{G},\mathbb{P})\le A\epsilon^{-\alpha}, \tag{13}$$

for every $\epsilon>0$ and some $0<\alpha<2$ and some constant A. Then, for some constant c and n_0 depending on α and A, we have for all $T\ge c$ and $n\ge n_0$,

$$P\left(\sup_{g\in\mathcal{G},\|g-g_0\|_{2,\mathbb{P}}\le n^{-1/(2+\alpha)}}\left|\mathbb{P}_n(g-g_0)-\mathbb{P}(g-g_0)\right|\ge Tn^{-2/(2+\alpha)}\right)\le c\exp\left\{-\frac{Tn^{\alpha/(2+\alpha)}}{c^2}\right\} \tag{14}$$

and

$$P\left(\sup_{g\in\mathcal{G},\|g-g_0\|_{2,\mathbb{P}}>n^{-1/(2+\alpha)}}\frac{\sqrt{n}\,|\mathbb{P}_n(g-g_0)-\mathbb{P}(g-g_0)|}{\|g-g_0\|_{2,\mathbb{P}}^{1-\alpha/2}}\ge T\right)\le c\exp\left(-\frac{T}{c^2}\right). \tag{15}$$

We now prove Theorem 2. By (11) in the main text and straightforward manipulations, we immediately have

$$\begin{aligned}
&\sqrt{n}(\widehat{\mu}_n-\mu_0)-\sqrt{n}\mathbb{P}_n\phi(\boldsymbol{v},\mu_0,\beta_0,\pi_0(\cdot),\psi_0(\cdot))\\
=&\frac{1}{\sqrt{n}}\sum_{i=1}^n\left[\left\{\frac{1}{\widehat{\pi}(X_i)}-\frac{1}{\pi_0(X_i)}\right\}\Delta_i\{Y_i-\psi_0(X_i)\}-\left\{\frac{\Delta_i}{\widehat{\pi}(X_i)}-1\right\}\left\{\widehat{\psi}(X_i)-\psi_0(X_i)\right\}\right]\\
=&\sqrt{n}\mathbb{P}_n\widehat{g}_1(\boldsymbol{v})-\sqrt{n}\mathbb{P}_n\widehat{g}_2(\boldsymbol{v}),
\end{aligned} \tag{16}$$

where $\widehat{g}_1(\boldsymbol{v})=\left\{\frac{1}{\widehat{\pi}(x)}-\frac{1}{\pi_0(x)}\right\}\delta\{y-\psi_0(x)\}$, $\widehat{g}_2(\boldsymbol{v})=\left\{\frac{\delta}{\widehat{\pi}(x)}-1\right\}\left\{\widehat{\psi}(x)-\psi_0(x)\right\}$, and $\widehat{g}_1\in\mathcal{G}_1$, $\widehat{g}_2\in\mathcal{G}_2$ with

$$\begin{aligned}
\mathcal{G}_1&=\left\{g_1(\boldsymbol{v})=\left\{\frac{1}{\pi(x)}-\frac{1}{\pi_0(x)}\right\}\delta\{y-\psi_0(x)\}:\pi\in\mathcal{F}\right\}\\
\mathcal{G}_2&=\left\{g_2(\boldsymbol{v})=\left\{\frac{\delta}{\pi(x)}-1\right\}\{\psi(x)-\psi_0(x)\}:\pi\in\mathcal{F};\psi\in\Psi\right\}.
\end{aligned}$$

Using Conditions A–C, we can easily verify that

$$\sup_{g_1\in\mathcal{G}_1}|g_1|_\infty\lesssim 1,\qquad H_{2,B}(\epsilon,\mathcal{G}_1,\mathbb{P})\lesssim\epsilon^{-\alpha_1}; \tag{17}$$

$$\sup_{g_2\in\mathcal{G}_2}|g_2|_\infty\lesssim 1,\qquad H_{2,B}(\epsilon,\mathcal{G}_2,\mathbb{P})\lesssim\epsilon^{-\max\{\alpha_1,\alpha_2\}}. \tag{18}$$

With Lemma 1, (17) implies that

$$P\left(\sup_{g_1\in\mathcal{G}_1,\|g_1\|_{2,\mathbb{P}}\le n^{-1/(2+\alpha_1)}}\left|\mathbb{P}_ng_1-\mathbb{P}g_1\right|\ge Tn^{-2/(2+\alpha_1)}\right)\le c\exp\left\{-\frac{Tn^{\alpha_1/(2+\alpha_1)}}{c^2}\right\}$$

$$P\left(\sup_{g_1\in\mathcal{G}_1,\|g_1\|_{2,\mathbb{P}}>n^{-1/(2+\alpha_1)}}\frac{\sqrt{n}\,|\mathbb{P}_ng_1-\mathbb{P}g_1|}{\|g_1\|_{2,\mathbb{P}}^{1-\alpha_1/2}}\ge T\right)\le c\exp\left(-\frac{T}{c^2}\right).$$

As a consequence, we have

$$\sqrt{n}|\mathbb{P}_n\widehat{g}_1-\mathbb{P}\widehat{g}_1|=O_p\left(n^{-\frac{2-\alpha_1}{2(2+\alpha_1)}}\right)\vee O_p\left(\|\widehat{g}_1\|_{2,\mathbb{P}}^{1-\alpha_1/2}\right).$$

Here $a\vee b=\max(a,b)$. With (18) and Lemma 1, we similarly have

$$\sqrt{n}|\mathbb{P}_n\widehat{g}_2-\mathbb{P}\widehat{g}_2|=O_p\left(n^{-\frac{2-\max\{\alpha_1,\alpha_2\}}{2(2+\max\{\alpha_1,\alpha_2\})}}\right)\vee O_p\left(\|\widehat{g}_2\|_{2,\mathbb{P}}^{1-\max\{\alpha_1,\alpha_2\}/2}\right).$$

Noting that $\alpha_1 < 2$ and $\alpha_2 < 2$, and $\mathbb{P}\widehat{g}_1 = 0$, we immediately have

$$\begin{aligned}\sqrt{n}|\mathbb{P}_n\widehat{g}_1| &= O_p\left(\|\widehat{g}_1\|_{2,\mathbb{P}}^{1-\alpha_1/2}\right) + o_p(1),\\ \sqrt{n}|\mathbb{P}_n\widehat{g}_2 - \mathbb{P}\widehat{g}_2| &= O_p\left(\|\widehat{g}_2\|_{2,\mathbb{P}}^{1-\max\{\alpha_1,\alpha_2\}/2}\right) + o_p(1),\end{aligned}$$

which, together with (16), Conditions A–C, and the fact that

$$\mathbb{P}\widehat{g}_2 = \mathbb{P}\left[\left\{\frac{\pi_0(x)}{\widehat{\pi}(x)} - 1\right\}\left\{\widehat{\psi}(x) - \psi_0(x)\right\}\right],$$

leads to (12) in the main text. This completes the proof of Theorem 2.

Proof of Corollary 1

By Theorem 2, (P1) immediately follows. Furthermore, by Theorem 2 and the conditions in (P2), we have

$$\sqrt{n}(\widehat{\mu} - \mu_0) = \sqrt{n}\mathbb{P}_n\phi(\boldsymbol{v}; \mu_0, \pi_0, \psi_0) + o_p(1). \tag{19}$$

Therefore, it remains to show that $\widehat{\mu}$ achieves the information bound. By (19), $\phi(\cdot; \mu_0, \pi_0(\cdot), \psi_0(\cdot))$ is the influence function. Referring to the established theory for the semiparametric efficiency bound—e.g., Chapter 3 of Bickel et al. [34], Newey [35], Chapters 3 and 18 of Kosorok [30], and the references therein—we need to show only the following two parts:

(i) $\widehat{\mu}$ is a regular estimator of μ_0.
Let $\mathbb{P}_\eta$ be a submodel indexed by η such that $\mathbb{P}_0$ is the true model. Further, let $\mu_\eta = E_\eta(Y)$, $\pi_\eta(x) = E_\eta(\Delta|X = x)$, and $\psi_\eta(x) = E_\eta(Y|X = x)$, where E_η indicates that the expectation is taken under $\mathbb{P}_\eta$. By Theorem 2.2 in Newey [35], arguing that $\widehat{\mu}$ is a regular estimator of μ_0 is equivalent to showing that

$$\left.\frac{\partial\mu_\eta}{\partial\eta}\right|_{\eta=0} = E_0\{\phi(\boldsymbol{V}; \mu_0, \pi_0, \psi_0)S_0(\boldsymbol{V})\}, \tag{20}$$

where

$$S_0(\boldsymbol{v}) = \left.\frac{\partial \log f_\eta(y, \delta, x)}{\partial\eta}\right|_{\eta=0},$$

$f_\eta(y, \delta, x)$ is the joint density of (Y, Δ, X) under $\mathbb{P}_\eta$, and E_0 indicates that the expectation is taken under the true distribution $f_0(y, \delta, x)$.

(ii) There exists a submodel $\mathbb{P}_{\eta^*}$ with $f^*_\eta(y, \delta, x)$ being the joint density of (Y, Δ, X) under $\mathbb{P}_{\eta^*}$ such that $\mathbb{P}_0$ is the true model and

$$\phi(\boldsymbol{v}; \mu_0, \pi_0, \psi_0) = \left.\frac{\partial \log f^*_\eta(y, \delta, x)}{\partial\eta^*}\right|_{\eta^*=0}.$$

We show Parts (i) and (ii) separately. To show Part (i), we need to verify (20). On the one hand,

$$\mu_\eta = E_\eta(Y) = E_\eta\{\psi_\eta(X)\},$$

which leads to

$$\left.\frac{\partial\mu_\eta}{\partial\eta}\right|_{\eta=0} = E_0\left\{\left.\frac{\partial\psi_\eta(X)}{\partial\eta}\right|_{\eta=0}\right\} + E_0\left\{\psi_0(X)S_0(V)\right\}. \tag{21}$$

On the other hand,

$$E\{\phi(\boldsymbol{V}; \mu_0, \pi_0, \psi_0)S_0(\boldsymbol{V})\} = \left.\frac{\partial E_\eta\left\{\phi(\boldsymbol{V}; \mu_0, \pi_0, \psi_0)\right\}}{\partial\eta}\right|_{\eta=0}.$$

It can be verified that

$$E_\eta\left\{\phi(\boldsymbol{V}; \mu_0, \pi_0, \psi_0)\right\} = E_\eta\left[\frac{\{\pi_\eta(X) - \pi_0(X)\}\{\psi_\eta(X) - \psi_0(X)\}}{\pi_0(X)} + \psi_\eta(X)\right] - \mu_0.$$

Note that $\{\pi_\eta(X)-\pi_0(X)\}\{\psi_\eta(X)-\psi_0(X)\}$ and its first derivative with respect to η are both equal to 0 at $\eta=0$. Thus,

$$E\{\phi(\boldsymbol{V};\mu_0,\pi_0,\psi_0)S_0(\boldsymbol{V})\}=\left.\frac{\partial E_\eta\{\phi(\boldsymbol{V};\mu_0,\pi_0,\psi_0)\}}{\partial\eta}\right|_{\eta=0}=E_0\left\{\left.\frac{\partial\psi_\eta(X)}{\partial\eta}\right|_{\eta=0}\right\}+E_0\{\psi_0(X)S_0(V)\},$$

which together with (21) implies (20). This proves Part (i).

We proceed to show Part (ii). Let $f_{Y|X;0}(y|x)$ be the true conditional density of Y given $X=x$ and $f_{X;0}(\cdot)$ the true marginal density of X. Consider

$$f_{\eta^*}(y,\delta,x)=\left[1+\eta^*\frac{\delta\{y-\psi_0(x)\}}{\pi_0(x)}+\eta^*\{\psi_0(x)-\mu_0\}\right]f_{Y|X;0}(y|x)\{\pi_0(x)\}^\delta\{1-\pi_0(x)\}^{1-\delta}f_{X;0}(x).$$

It is easy to check that for sufficiently small η^* this $f_{\eta^*}(y,\delta,x)$ is a parametric submodel. The tangent of this submodel is given by

$$\left.\frac{\partial\log f_{\eta^*}(y,\delta,x)}{\partial\eta^*}\right|_{\eta^*=0}=\frac{\delta\{y-\psi_0(x)\}}{\pi_0(x)}+\psi_0(x)-\mu_0=\phi(\boldsymbol{v};\mu_0,\pi_0,\psi_0),$$

which completes our proof of Part (ii). Therefore, we have completed the proof of this Corollary.

Proof of Theorem 3

Since we do not assume Condition A in this theorem, we first verify that Condition A is satisfied given Conditions 1–3 and B–E; then we prove the theorem by verifying the conditions in (P2) of Corollary 1. In fact, Conditions A2 and A3 are satisfied by reviewing Conditions 3 and E, so we need to verify only Condition A1. The following Lemma adapted from Lemma 2 of Chen et al. [22] ensures that Condition A1 is satisfied with $\alpha_1=1$.

Lemma 2 *Suppose Conditions 1 and 2 are satisfied. For any $\epsilon>0$ and integer $q>0$ we have*

$$H_{q,B}(\epsilon,\mathcal{F},F_X)\lesssim 1/\epsilon,$$

where $\mathcal{F}=\{\theta(h^T(x)\beta):\theta\in\Theta;\beta\in\Lambda\}$.

Therefore, to prove this theorem, we need to verify only that

$$\|\widehat{\pi}-\pi_0\|_{2,\mathbb{P}} = o_p(1) \tag{22}$$

$$\|\widehat{\psi}-\psi_0\|_{2,\mathbb{P}} = o_p(1) \tag{23}$$

$$\sqrt{n}\mathbb{P}\left[\left\{\frac{\pi_0(x)}{\widehat{\pi}(x)}-1\right\}\left\{\widehat{\psi}(x)-\psi_0(x)\right\}\right] = o_p(1). \tag{24}$$

Note that (22) immediately follows by Theorem 1. Furthermore, by Conditions B and E, we have

$$\left|\sqrt{n}\mathbb{P}\left[\left\{\frac{\pi_0(x)}{\widehat{\pi}(x)}-1\right\}\left\{\widehat{\psi}(x)-\psi_0(x)\right\}\right]\right|\lesssim\sqrt{n}\|\widehat{\pi}-\pi_0\|_{2,\mathbb{P}}\|\widehat{\psi}-\psi_0\|_{2,\mathbb{P}},$$

which together with Theorem 1 implies that both (23) and (24) hold if we can show that

$$\|\widehat{\psi}-\psi_0\|_{2,\mathbb{P}}=o_P(n^{-1/6}). \tag{25}$$

We proceed in two steps. First, note that Conditions D and E imply that

$$\|\widehat{\psi}-\psi_0\|_{2,\mathbb{P}}^2\lesssim\mathbb{P}\Big[w(1,x)\pi_0(x)\{\widehat{\psi}(x)-\psi_0(x)\}^2\Big]=\mathbb{P}\Big[w(\delta,x)\{\widehat{\psi}(x)-\psi_0(x)\}^2\Big]. \tag{26}$$

Second, by the definition of $\widehat{\psi}$, if $\psi_0\in\Psi$, we have $Q(\widehat{\psi})\le Q(\psi_0)$. Hence,

$$\begin{aligned}
\mathbb{P}\Big[w(\delta,x)\{\widehat{\psi}(x)-\psi_0(x)\}^2\Big] &\le Q(\psi_0)-Q(\widehat{\psi})+\mathbb{P}\Big[w(\delta,x)\{\widehat{\psi}(x)-\psi_0(x)\}^2\Big]\\
&= \mathbb{P}_n\Big[w(\delta,x)\{y-\psi_0(x)\}^2\Big]-\mathbb{P}_n\Big[w(\delta,x)\{y-\widehat{\psi}(x)\}^2\Big]\\
&\quad+\mathbb{P}\Big[w(\delta,x)\{\widehat{\psi}(x)-\psi_0(x)\}^2\Big]\\
&= 2\mathbb{P}_n\Big[w(\delta,x)\{\widehat{\psi}(x)-\psi_0(x)\}\{y-\psi_0(x)\}\Big]\\
&\quad-(\mathbb{P}_n-\mathbb{P})\Big[w(\delta,x)\{\widehat{\psi}(x)-\psi_0(x)\}^2\Big]\\
&= 2\mathbb{P}_n\widehat{g}_3-(\mathbb{P}_n-\mathbb{P})\widehat{g}_4,
\end{aligned} \tag{27}$$

where $\widehat{g}_3(\delta, y, x) = w(\delta, x)\{\widehat{\psi}(x) - \psi_0(x)\}\{y - \psi_0(x)\} \in \mathcal{G}_3$, $\widehat{g}_4(\delta, x) = w(\delta, x)\{\widehat{\psi}(x) - \psi_0(x)\}^2 \in \mathcal{G}_4$, with

$$\begin{aligned} \mathcal{G}_3 &= \{g_3(\boldsymbol{v}) = w(\delta, x)\{\psi(x) - \psi_0(x)\}\{y - \psi_0(x)\} : \psi \in \Psi\}, \\ \mathcal{G}_4 &= \left\{g_4(\boldsymbol{v} = w(\delta, x)\{\psi(x) - \psi_0(x)\}^2 : \psi \in \Psi\right\}. \end{aligned}$$

By Conditions A–D, we can easily verify that

$$\begin{aligned} \sup_{g_3 \in \mathcal{G}_3} |g_3|_\infty \lesssim 1, \qquad & H_{2,B}(\epsilon, \mathcal{G}_3, \mathbb{P}) \lesssim \epsilon^{-\alpha_2}; \\ \sup_{g_4 \in \mathcal{G}_4} |g_4|_\infty \lesssim 1, \qquad & H_{2,B}(\epsilon, \mathcal{G}_4, \mathbb{P}) \lesssim \epsilon^{-\alpha_2}. \end{aligned}$$

Therefore, Lemma 1 applies to both $\mathcal{G}_3$ and $\mathcal{G}_4$. That is,

$$P\left(\sup_{g_3 \in \mathcal{G}_3, \|g_3\|_{2,\mathbb{P}} \le n^{-1/(2+\alpha_2)}} \left|\mathbb{P}_n g_3 - \mathbb{P} g_3\right| \ge T n^{-2/(2+\alpha_2)}\right) \le c \exp\left\{-\frac{T n^{\alpha_2/(2+\alpha_2)}}{c^2}\right\}$$

$$P\left(\sup_{g_3 \in \mathcal{G}_3, \|g_3\|_{2,\mathbb{P}} > n^{-1/(2+\alpha_2)}} \frac{\sqrt{n}\,|\mathbb{P}_n g_3 - \mathbb{P} g_3|}{\|g_3\|_{2,\mathbb{P}}^{1-\alpha_2/2}} \ge T\right) \le c \exp\left(-\frac{T}{c^2}\right)$$

and

$$P\left(\sup_{g_4 \in \mathcal{G}_4, \|g_4\|_{2,\mathbb{P}} \le n^{-1/(2+\alpha_2)}} \left|\mathbb{P}_n g_4 - \mathbb{P} g_4\right| \ge T n^{-2/(2+\alpha_2)}\right) \le c \exp\left\{-\frac{T n^{\alpha_2/(2+\alpha_2)}}{c^2}\right\}$$

$$P\left(\sup_{g_4 \in \mathcal{G}_4, \|g_4\|_{2,\mathbb{P}} > n^{-1/(2+\alpha_2)}} \frac{\sqrt{n}\,|\mathbb{P}_n g_4 - \mathbb{P} g_4|}{\|g_4\|_{2,\mathbb{P}}^{1-\alpha_2/2}} \ge T\right) \le c \exp\left(-\frac{T}{c^2}\right).$$

These together with the fact that $\mathbb{P}\widehat{g}_3 = 0$ lead to

$$\begin{aligned} |\mathbb{P}_n \widehat{g}_3| &= O_p(n^{-2/(2+\alpha_2)}) \vee \left\{n^{-1/2} \cdot O_p(\|\widehat{g}_3\|_{2,\mathbb{P}}^{1-\alpha_2/2})\right\} \\ |(\mathbb{P}_n - \mathbb{P})\widehat{g}_4| &= O_p(n^{-2/(2+\alpha_2)}) \vee \left\{n^{-1/2} \cdot O_p(\|\widehat{g}_4\|_{2,\mathbb{P}}^{1-\alpha_2/2})\right\}. \end{aligned} \tag{28}$$

Furthermore, by Conditions B–D we have

$$\|\widehat{g}_3\|_{2,\mathbb{P}} \lesssim \|\widehat{\psi} - \psi_0\|_{2,\mathbb{P}}, \quad \|\widehat{g}_4\|_{2,\mathbb{P}} \lesssim \|\widehat{\psi} - \psi_0\|_{2,\mathbb{P}}. \tag{29}$$

Combining (27)–(29), we immediately have

$$\mathbb{P}\left[w(\delta, x)\{\widehat{\psi}(x) - \psi_0(x)\}^2\right] \lesssim O_p(n^{-2/(2+\alpha_2)}) \vee \left\{n^{-1/2} \cdot O_p(\|\widehat{\psi} - \psi_0\|_{2,\mathbb{P}}^{1-\alpha_2/2})\right\},$$

which together with (26) and the condition that $0 < \alpha_2 < 2$ in Condition B leads to

$$\|\widehat{\psi} - \psi_0\|_{2,\mathbb{P}} = O_P(n^{-\frac{1}{2+\alpha_2}}).$$

This implies that (25) is correct, and this completes our proof.

References

1. Rubin D. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology* 1974; 66: 688–701.
2. Rubin D. Direct and indirect causal effects via potential outcomes. *Scandinavian Journal of Statistics* 2004; 31: 161–170.
3. Small DS, Joffe MM, Lynch KG, Roy JA, Localio AR. Tom Ten Haves contributions to causal inference and biostatistics: review and future research directions. *Statistics in Medicine* 2014; 33: 3421-3433.
4. Burgess S, Small DS. Predicting the direction of causal effect based on an instrumental variable analysis: a cautionary tale. *Journal of Causal Inference* 2016; 4: 45-59.
5. Hahn J. On the role of propensity score in efficient semiparametric estimation of average treatment effects. *Econometrica* 1998; 66: 315–332.
6. Mincer J. *Schooling, Experience, and Earnings* 1974. New York: National Bureau of Economic Research.

7. Dehejia RH, Wahba S. Causal effects in nonexperimental studies: Reevaluating the evaluation of training programs. *Journal of the American Statistical Association* 1999; 94: 1053–1062.
8. Lalonde RJ. Evaluating the econometric evaluations of training programs with experimental data. *American Economic Review* 1986; 76: 604–620.
9. Graham BS, Pinto C, Egel D. Inverse Probability Tilting for Moment Condition Models With Missing Data. *Review of Economic Studies* 2012; 79: 1052-1079.
10. Hastie TJ, Tibshirani RJ. *Generalized Additive Models* 1990. Chapman & Hall/CRC.
11. Rosenbaum PR, Rubin DB. The central role of the propensity score in observational studies for causal effects. *Biometrika* 1983; 70: 41-55.
12. Horvitz DG, Thompson DJ. A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association* 1952; 47: 663–685.
13. Henmi M, Eguchi S. A paradox concerning nuisance parameters and projected estimating functions. *Biometrika* 2004; 91: 929–941.
14. Robins JM, Rotnitzky A, Zhao LP. Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association* 1994; 89: 846-866.
15. Scharfstein DO, Rotnitzky A, Robins JM. Adjusting for nonignorable drop-out using semiparametric nonresponse models (with discussion). *Journal of the American Statistical Association* 1999; **94**: 1096–1120.
16. Wooldridge J. Inverse Probability Weighted Estimation for General Missing Data Problems. *Journal of Econometrics* 2007; 141: 1281-1301.
17. Uysal SD. Doubly Robust Estimation of Causal Effects With Multivalued Treatments: An Application to the Returns to Schooling. *Journal of Applied Econometrics* 2015; 30: 763-786.
18. Sloczynski T, Wooldridge JM. A General Double Robustness Result for Estimating Average Treatment Effects 2014. IZA Discussion Paper 8084, Institute for the Study of Labor, Bonn, Germany.
19. Kang JD, Schafer JL. Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data (with discussion). *Statistical Science* 2007; 22: 523–539.
20. Cheng PE. Nonparametric estimation of mean functionals with data missing at random. *Journal of the American Statistical Association* 1994; 89: 81–87.
21. Cosslett SR. Distribution-free maximum likelihood estimator of the binary choice model. *Econometrica* 1983; 51: 765–782.
22. Chen B, Li P, Qin J, Yu T. Using a monotonic density ratio model to find the asymptotically optimal combination of multiple diagnostic tests. *Journal of the American Statistical Association* 2016; 111: 861-874.
23. Dykstra R, Kochar S, Robertson T. Inference for likelihood ratio ordering in the two-sample problem. *Journal of the American Statistical Association* 1995; 90: 1034–1040.
24. Ayer M, Brunk HD, Ewing GM, Reid WT, Silverman E. An empirical distribution function for sampling with incomplete information. *The Annals of Mathematical Statistics* 1955; 26: 641–647.
25. R Development Core Team. R: A language and environment for statistical computing 2011. R Foundation for Statistical Computing, Vienna, Austria. URL: https://www.r-project.org/. ISBN: 3-900051-07-0.
26. Khan S, Tamer E. Irregular Identification, Support Conditions, and Inverse Weight Estimation. *Econometrica* 2010; 78: 2021-2042.
27. Cao WH, Tsiatis AA, Davidian M. Improving efficiency and robutness of the doubly robust estimator for a population mean with incomplete data. *Biometrika* 2009; **96**: 723–734.
28. van de Geer SA. *Empirical Processes in M-Estimation* 2000. Cambridge University Press, New York.
29. van der Vaart AW, Wellner JA. *Weak Convergence and Empirical Processes with Applications to Statistics* 1996. Springer, New York.
30. Kosorok MR. *Introduction to Empirical Processes and Semiparametric Inference* 2008. Springer, New York.
31. Hájek J. Comment on "an essay on the logical foundations of survey sampling, part one." In V. Godambe and D. Sprott (Eds.), *Foundations of Statistical Inference* 1971. Toronto: Holt, Rinehart and Winston.
32. Hammer SM, Katzenstein DA, Hughes MD, Gundacker H, Schooley RT, Haubrich RH et al. A trial comparing nucleoside monotherapy with combination therapy in HIV-infected adults with CD4 cell counts from 200 to 500 per cubic millimeter. AIDS Clinical Trials Group Study 175 Study Team. *New England Journal of Medicine* 1996; 335: 1081–1090.
33. Imbens GW, Rubin DB. *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction* 2015. Cambridge University Press, New York.
34. Bickel P, Klaassen CAJ, Ritov Y, Wellner JA. *Efficient and Adaptive Estimation for Semiparametric Models* 1993. Johns Hopkins University Press, Baltimore.
35. Newey WK. Semiparametric efficiency bounds. *Journal of Applied Econometrics* 1990; 5: 99–135.