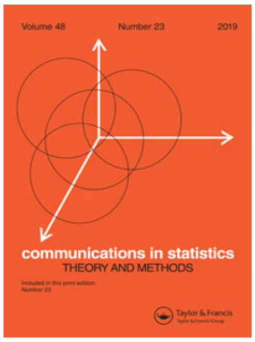

Taylor & Francis
Taylor & Francis Group

Communications in Statistics - Theory and Methods

ISSN: 0361-0926 (Print) 1532-415X (Online) Journal homepage: https://www.tandfonline.com/loi/lsta20

Multiple robustness estimation in causal inference

Lei Wang

To cite this article: Lei Wang (2019) Multiple robustness estimation in causal inference, Communications in Statistics - Theory and Methods, 48:23, 5701-5718, DOI: 10.1080/03610926.2018.1520881

To link to this article: https://doi.org/10.1080/03610926.2018.1520881

Published online: 04 Dec 2018.

COMMUNICATIONS IN STATISTICS—THEORY AND METHODS
2019, VOL. 48, NO. 23, 5701–5718
https://doi.org/10.1080/03610926.2018.1520881

Taylor & Francis
Taylor & Francis Group

Multiple robustness estimation in causal inference

Lei Wang

School of Statistics and Data Science & LPMC, Nankai University, Tianjin, China

ABSTRACT

Estimation of average treatment effect is crucial in causal inference for evaluation of treatments or interventions in biostatistics, epidemiology, econometrics, sociology. However, existing estimators require either a propensity score model, an outcome vector model, or both is correctly specified, which is difficult to verify in practice. In this paper, we allow multiple models for both the propensity score models and the outcome models, and then construct a weighting estimator based on observed data by using two-sample empirical likelihood. The resulting estimator is consistent if any one of those multiple models is correctly specified, and thus provides multiple protection on consistency. Moreover, the proposed estimator can attain the semiparametric efficiency bound when one propensity score model and one outcome vector model are correctly specified, without requiring knowledge of which models are correct. Simulations are performed to evaluate the finite sample performance of the proposed estimators. As an application, we analyze the data collected from the AIDS Clinical Trials Group Protocol 175.

ARTICLE HISTORY
Received 11 July 2017
Accepted 3 September 2018

KEYWORDS
Average treatment effect; causal inference; multiple robust; propensity; semiparametric efficiency

1. Introduction

A primary goal of causal inference is to construct unbiased estimators of treatment effects. In this paper, we focus on estimation of average treatment effect (ATE), which is an important problem in biostatistics, epidemiology, econometrics, sociology and so on. Let T denote a binary treatment assignment ($T=1$ for treated, $T=0$ for untreated), X be a d-dimensional vector of pretreatment covariates, Y_1 and Y_0 be two potential outcome variables when $T=1$ and 0, respectively. For each unit, since only one treatment is applied, either Y_1 or Y_0 is observed, but not both, i.e., we only observe $Y = TY_1 + (1-T)Y_0$. The ATE is defined as $\Delta = E(Y_1) - E(Y_0)$. The problem of interest is to estimate Δ based on n independent and identically distributed samples $(X_i, Y_i, T_i), i = 1, \ldots, n$, from (X, Y, T).

In order to identify the ATE, Rosenbaum and Rubin (1983) assumed that the treatment assignment is independent of the potential outcomes when conditioning on a set of observed covariates, i.e., $(Y_1, Y_0) \perp\!\!\!\perp T | X$, which is usually called ignorability of treatment or unconfoundedness assumption. Here, $\perp\!\!\!\perp$ stands for conditional independence. Under this assumption, there are mainly three ways for estimating the ATE: (i) propensity score,

CONTACT Lei Wang lwangstat@nankai.edu.cn School of Statistics and Data Science & LPMC, Nankai University, Tianjin 300071, China.

(ii) outcome regression, and (iii) combination estimation (see Hahn 1998; Hirano et al. 2003; Tan 2010).

Denote $\pi(X) = \Pr(T = 1|X)$ as the probability of treatment assignment conditional on covariates, which is usually called as propensity. The ATE can be estimated by the inverse probability weighting (IPW; Horvitz and Thompson 1952; Rosenbaum and Rubin 1983) estimator as follows

$$\hat{\Delta}_{ipw} = \frac{1}{n}\sum_{i=1}^{n}\left[\frac{T_iY_i}{\pi(X_i,\tilde{\beta})} - \frac{(1-T_i)Y_i}{1-\pi(X_i,\tilde{\beta})}\right]$$

where $\pi(X,\beta)$ is a parametric model for $\pi(X)$ with unknown parameter vector β, and $\tilde{\beta}$ is an estimated vector of β. Alternatively, denote $m(X) = \{m^1(X), m^0(X)\}^T$ with $m^t(X) = E(Y_t|X)$, $t = 0$, 1. In general, the linear or more complex parametric outcome regression (OR) model $m(X,\gamma) = \{m^1(X,\gamma^1), m^0(X,\gamma^0)\}^T$ with unknown parameter vector $\gamma = (\gamma^1,\gamma^0)^T$ is assumed for predicting unobservable potential outcome vector $m(X)$, such that the ATE can be estimated by averaging predicted outcomes, i.e.,

$$\hat{\Delta}_{or} = \frac{1}{n}\sum_{i=1}^{n}\left\{m^1\left(X_i,\tilde{\gamma}^1\right) - m^0\left(X_i,\tilde{\gamma}^0\right)\right\}$$

where $\tilde{\gamma} = (\tilde{\gamma}^1,\tilde{\gamma}^0)^T$ is an estimated vector of γ. It can be seen that $\hat{\Delta}_{ipw}$ is consistent only if $\pi(X)$ is correctly modeled and $\hat{\Delta}_{or}$ is consistent only if $m(X)$ is correctly modeled. However, $\hat{\Delta}_{ipw}$ may not have enough estimation efficiency, as it does not fully extract the information contained in the auxiliary variables.

To improve the efficiency over $\hat{\Delta}_{ipw}$, Robins et al. (1994) introduced the augmented inverse probability weighting (AIPW) estimator by employing both a propensity score $\pi(X)$ and an outcome vector model $m(X)$ for the treatment and control groups as follows,

$$\hat{\Delta}_{aipw} = \frac{1}{n}\sum_{i=1}^{n}\left[\frac{T_iY_i}{\pi(X_i,\tilde{\beta})} - \frac{T_i-\pi(X_i,\tilde{\beta})}{\pi(X_i,\tilde{\beta})}m^1\left(X_i,\tilde{\gamma}^1\right)\right] - \frac{1}{n}\sum_{i=1}^{n}\left[\frac{(1-T_i)Y_i}{1-\pi(X_i,\tilde{\beta})} + \frac{T_i-\pi(X_i,\tilde{\beta})}{1-\pi(X_i,\tilde{\beta})}m^0\left(X_i,\tilde{\gamma}^0\right)\right]$$

The AIPW estimators are double robust (DR) in the sense that the consistency of $\hat{\Delta}_{aipw}$ is achieved if either $\pi(X)$ or $m(X)$ is correctly modeled and locally efficient when the two working models are all correctly specified (Bang and Robins 2005; Cao et al. 2009).

Notice that all methods mentioned above require specification of either a propensity score model $\pi(X)$, an outcome vector model $m(X)$, or both. Consistency of these estimators requires some underlying models to be correctly specified. In practice, however, it is difficult to verify that either $\pi(X)$ or $m(X)$ is correctly specified. Recently, Han and Wang (2013) introduced the concept of multiple robustness in missing data problem; see also Han (2014). In their procedure, multiple propensity score models and multiple imputation models (or outcome models) may be fitted. The estimator is said to be multiply robust (MR) in the sense that it is consistent if any one of those multiple models, for either the propensity score or the imputation, is correctly specified. Multiple robustness can be viewed as an extension of the concept of double robustness.

In this paper, we extend the MR estimation of Han (2014) from missing data problem to the estimation of causal inference. Similar to Naik et al. (2016), we first allow multiple models for both $\pi(X)$ and $m(X)$, each involving different subsets of covariates and possibly different function forms. However, we propose a new weighting estimator based on two-sample empirical likelihood (EL) framework by matching population moments of model-based fitted values from the treated and untreated subsamples to the full sample (Han and Wang 2013). Under the ignorability of treatment assumption, we show that:

1. The proposed estimator is consistent if any one of those multiple models for $\pi(X)$ or $m(X)$ is correctly specified. Therefore, our estimator provides multiple protection on estimation consistency, which appears superior to the double robustness property.
2. When both $\pi(X)$ and $m(X)$ are correctly modeled, our estimator can attain the semiparametric efficiency bound, without requiring the knowledge of exactly which two models are correct.
3. Simulation results show that the proposed estimators provide extra protection against model misspecification while estimation efficiency is not compromised.

This article is organized as follows. The proposed estimator is introduced in Section 2 based on the constrained two-sample EL approach. The multiple robustness, asymptotic theory and other properties are derived and discussed in Section 3. Sections 4 and 5 cover the numerical studies and real data application, respectively. Some relevant discussions are given in Section 6. The Appendix contains some technical details.

2. Methodology

Notice that the true functional forms of the propensity score $\pi(X)$ and outcome vector $m(X)$ are unknown in general. Thus, we assume two sets of multiple working models

$$\mathcal{P} = \left\{\pi_l(X, \beta_l), l = 1, 2, \ldots, L\right\},$$
$$\mathcal{M} = \left\{m_k(X, \gamma_k) = \left\{m_k^1(X, \gamma_k^1), m_k^0(X, \gamma_k^0)\right\}^T, k = 1, 2, \ldots, K\right\}$$

for propensity score $\pi(X)$ and outcome vector $m(X)$, respectively, where β_l and $\gamma_k = (\gamma_k^1, \gamma_k^0)^T$ are unknown parameter vectors for $\pi_l(X, \beta_l)$ and $m_k(X, \gamma_k)$, $l = 1, 2, \ldots, L, k = 1, 2, \ldots, K$. Denote $\hat{\beta}_l$ and $\hat{\gamma}_k = (\hat{\gamma}_k^1, \hat{\gamma}_k^0)^T$ as the corresponding estimators based on working models $\pi_l(X, \beta_l)$ and $m_k(X, \gamma_k)$ in $\mathcal{P}$ and $\mathcal{M}$, respectively. Since the covariate X_i is observable for each subject, $\hat{\beta}_l$ is taken to be the maximizer of the binomial likelihood

$$\prod_{i=1}^{n} \left\{\pi_l(X_i, \beta_l)\right\}^{T_i}\left\{1 - \pi_l(X_i, \beta_l)\right\}^{1-T_i}, \quad l = 1, 2, \ldots, L \tag{1}$$

and $\hat{\gamma}_k^t$ is derived to be the regression coefficients by fitting the model $m_k^t(X, \gamma_k^t)$ based on data $\{(X_i, Y_i, T_i = t), i = 1, ..., n\}$ for $k = 1, 2, ..., K$.

Denote $n_1 = \sum_{i=1}^n T_i$ and $n_0 = n - n_1$ as the numbers of subjects who are treated and not treated. Without loss of generality, assume $T_1 = ... = T_{n_1} = 1$ and $T_{n_1+1} = ... = T_n = 0$. Denote $Y_{1i} = Y_i, X_{1i} = X_i$ for $i = 1, ..., n_1$ and $Y_{0j} = Y_{n_1+j}, X_{0j} = X_{n_1+j}$ for $j = 1, ..., n_0$. Let $f_1(Y, X)$ and $f_0(Y, X)$ be the unconditional joint density functions of (Y_1, X) and (Y_0, X), respectively. As in Qin and Zhang (2007), the joint conditional likelihood of $\{(Y_{1i}, X_{1i}) : i = 1, ..., n_1\}$ and $\{(Y_{0j}, X_{0j}) : j = 1, ..., n_0\}$ is given by

$$L = \prod_{i=1}^{n_1} \frac{\pi(X_{1i}, \beta) w_{1i}}{\theta} \prod_{j=1}^{n_0} \frac{\{1 - \pi(X_{0j}, \beta)\} w_{0j}}{1 - \theta} \tag{2}$$

where $\theta = E\{\pi(X, \beta)\}, w_{1i} = f_1(Y_{1i}, X_{1i})$ and $w_{0j} = f_0(Y_{0j}, X_{0j})$ are non-negative jump sizes with total mass 1, respectively, $i = 1, ..., n_1, j = 1, ..., n_0$. Since the constrained EL estimation is more efficient than the unconstrained EL estimation, the estimation efficiency based on the conditional likelihood (2) can be improved if we can impose some constraints by making use of the availability of all covariates.

Define $\hat{\theta}_l = n^{-1} \sum_{i=1}^n \pi_l(X_i, \hat{\beta}_l)$ for $l = 1, 2, ..., L$ and $\hat{\eta}_k^t = n^{-1} \sum_{i=1}^n m_k^t(X_i, \hat{\gamma}_k^t)$ for $k = 1, 2, ..., K$. With β_l replaced by $\hat{\beta}_l$, θ_l replaced by $\hat{\theta}_l$ and η_k^t replaced by $\hat{\eta}_k^t$, we can maximize the conditional likelihood (2) subject to the constraints

$$\sum_{i=1}^{n_1} w_{1i} = 1, \quad \sum_{i=1}^{n_1} w_{1i} \pi_l\left(X_{1i}, \hat{\beta}_l\right) = \hat{\theta}_l, \sum_{i=1}^{n_1} w_{1i} m_k^1(X_{1i}, \hat{\gamma}_k^1) = \hat{\eta}_k^1 \tag{3}$$

$$\sum_{j=1}^{n_0} w_{0j} = 1, \quad \sum_{j=1}^{n_0} w_{0j} \pi_l\left(X_{0j}, \hat{\beta}_l\right) = \hat{\theta}_l, \sum_{j=1}^{n_0} w_{0j} m_k^0\left(X_{0j}, \hat{\gamma}_k^0\right) = \hat{\eta}_k^0 \tag{4}$$

by matching population moments of model-based fitted values from the treated and untreated subsamples to the full sample. The derivation of such w_{1i} and w_{0j} pertains to a constrained maximization problem. Denote

$$\begin{aligned} g_1(X, \hat{\beta}, \hat{\gamma}^1) = (&\pi_1(X, \hat{\beta}_1) - \hat{\theta}_1, ..., \pi_J(X, \hat{\beta}_J) - \hat{\theta}_J, m_1^1(X, \hat{\gamma}_1^1) - \hat{\eta}_1^1, \\ &..., m_K^1(X, \hat{\gamma}_K^1) - \hat{\eta}_K^1)^T, \\ g_0(X, \hat{\beta}, \hat{\gamma}^0) = (&\hat{\theta}_1 - \pi_1(X, \hat{\beta}_1), ..., \hat{\theta}_J - \pi_J(X, \hat{\beta}_J), m_1^0(X, \hat{\gamma}_1^0) - \hat{\eta}_1^0, \\ &..., m_K^0(X, \hat{\gamma}_K^0) - \hat{\eta}_K^0)^T \end{aligned}$$

where $\hat{\beta} = (\hat{\beta}_1, ..., \hat{\beta}_L)^T$ and $\hat{\gamma}^t = (\hat{\gamma}_1^t, ..., \hat{\gamma}_K^t)^T$ for $t = 0$, 1. For fixed $(\hat{\beta}, \hat{\gamma}^t)$, using the Lagrange multipliers method as in Owen (1988) or Qin and Lawless (1994), it is easy to show that the w_{1i} and w_{0j} maximizing the conditional likelihood (2) subject to the constrains in (3)–(4) are given by

$$\hat{w}_{1i} = \frac{1}{n_1\left\{1 + \hat{\rho}_1^T g_1\left(X_{1i}, \hat{\beta}, \hat{\gamma}^1\right)\right\}}, \qquad \hat{w}_{0j} = \frac{1}{n_0\left\{1 + \hat{\rho}_0^T g_0\left(X_{0j}, \hat{\beta}, \hat{\gamma}^0\right)\right\}}$$

where $\hat{\rho}_t = (\hat{\rho}_{t1}, \hat{\rho}_{t2}, ..., \hat{\rho}_{tS})^T$ is $S = J + K$-dimensional vector satisfying the equations

$$\sum_{i=1}^{n_1} \frac{g_1\left(X_{1i}, \hat{\beta}, \hat{\gamma}^1\right)}{1+\rho_1^T g_1\left(X_{1i}, \hat{\beta}, \hat{\gamma}^1\right)} = 0, \quad \sum_{j=1}^{n_0} \frac{g_0\left(X_{0j}, \hat{\beta}, \hat{\gamma}^0\right)}{1+\rho_0^T g_0\left(X_{0j}, \hat{\beta}, \hat{\gamma}^0\right)} = 0$$

Our proposed estimator of Δ is

$$\hat{\Delta}_{mr} = \sum_{i=1}^{n_1} \hat{w}_{1i} Y_{1i} - \sum_{j=1}^{n_0} \hat{w}_{0j} Y_{0j} \tag{5}$$

The Lagrange multipliers $\hat{\rho}_t$ is essential in the calculation of $\hat{\Delta}_{mr}$. Because of the non-negativity of $\hat{w}_{1i}$ and $\hat{w}_{0j}$, $\hat{\rho}_t$ must satisfy that $1+\hat{\rho}_1^T g_1(X_{1i}, \hat{\beta}, \hat{\gamma}^1) > 0$ and $1+\rho_0^T g_0(X_{0j}, \hat{\beta}, \hat{\gamma}^0) > 0$, $i = 1, ..., n_1, j = 1, ..., n_0$. The modified Newton-Raphson algorithm discussed by Chen et al. (2002) can be applied. On the other hand, The IPW and AIPW estimators are sensitive to near-zero or one values of $\hat{\pi}(X)$, which can yield extremely large weight on $T/\hat{\pi}(X)$ or $(1-T)/\{1-\hat{\pi}(X)\}$. However, the proposed estimator $\hat{\Delta}_{mr}$ does not have this problem, since the values of w_{1i} and w_{0j} will be more evenly distributed rather than concentrating on a few subjects, in order to maximize (2) subject to the constraints (3) and (4). This can prevent the occurrence of the extreme values of w_{1i} and w_{0j}, which can be seen from our simulation results in Section 4.

3. Multiple robustness and asymptotic theory

In this section, we first consider the consistency of $\hat{\Delta}_{mr}$ and show that it is multiply protected in Theorem 1. Then, we establish asymptotic results of the proposed estimator. In particular, Theorem 2 studies the asymptotic efficiency when $\mathcal{P}$ contains a correctly specified model for $\pi(X)$. Theorem 3 shows that $\hat{\Delta}_{mr}$ can attain the semiparametric efficiency bound, when $\mathcal{P}$ contains a correctly specified model for $\pi(X)$ and $\mathcal{M}$ contains a correctly specified model for $m(X)$.

3.1. Multiple robustness

Theorem 1 shows that $\hat{\Delta}_{mr}$ is multiple robust, in the sense of that $\hat{\Delta}_{mr}$ is a consistent estimator as long as one working model is correctly specified, either for $\pi(X)$ or $m(X)$. This is a significant improvement over both the IPW and AIPW estimators. The proof is given in Appendix.

Theorem 1. When $\mathcal{P}$ contains a correctly specified model for $\pi(x)$ or $\mathcal{M}$ contains a correctly specified vector model for m(x), $\hat{\Delta}_{mr}$ is a consistent estimator of Δ_0 as $n \to \infty$, where Δ_0 is the true value of Δ.

Furthermore, we show that $\hat{\Delta}_{mr}$ is also a consistent estimator, even if the true outcome vector model $m(X)$ is a linear combination of the multiple outcome vector models in $\mathcal{M}$.

Corollary 1. *When $m^t(X)$, $t=0, 1$, is a linear combination of functions $\{m_k^t(X, \gamma_k^t),\ k = 1, 2, ..., K\}$ for some unknown coefficients, i.e., $m^t(X) = c_{t0} + \sum_{k=1}^K c_{tk} m_k^t(X, \gamma_k^t)$, $\hat{\Delta}_{mr}$ is a consistent estimator of Δ_0 as $n \to \infty$.*

3.2. Asymptotic theory

Without loss of generality, assume that $\pi_1(X, \beta_1)$ is a correctly specified model for $\pi(X)$ and denote

$$S(X, T, \beta_1) = \frac{T - \pi_1(X, \beta_1)}{\pi_1(X, \beta_1)\{1 - \pi_1(X, \beta_1)\}} \frac{\partial \pi_1(X, \beta_1)}{\partial \beta_1}$$

as the score function of β_1 corresponding to the binomial likelihood (6). We first derive the asymptotic distribution of $\hat{\Delta}_{mr}$ when $\pi(X)$ is correctly modelled.

As in White (1982) and Han (2014), we know that $\hat{\beta}_l \xrightarrow{p} \beta_{l*}$ and $\hat{\gamma}_k^t \xrightarrow{p} \gamma_{k*}^t$, where $\xrightarrow{p}$ denotes convergence in probability, β_{l*} and γ_{k*}^t minimize the corresponding Kullback–Leibler distances between the probability distributions based on the postulated models and the one generating the data. Denote $\beta_* = (\beta_{1*}, \ldots, \beta_{L*})^T$ and $\gamma_*^t = (\gamma_{1*}^t, \ldots, \gamma_{K*}^t)^T$. Write $L_1 = E\{(Y_1 - \mu_1) g_1(X, \beta_*, \gamma_*^1)/\pi(X)\}$, $G_1 = E\{g_1(X, \beta_*, \gamma_*^1)^{\otimes 2}/\pi(X)\}$, $L_0 = E[(Y_0 - \mu_0) g_0(X, \beta_*, \gamma_*^0)/\{1 - \pi(X)\}]$ and $G_0 = E[g_0(X, \beta_*, \gamma_*^0)^{\otimes 2}/\{1 - \pi(X)\}]$, where $\mu_t = E(Y_t)$ for $t = 0$, 1 and $A^{\otimes 2} = AA^T$ for any vector A. Denote $S = S(X, T, \beta_1^0)$, $R_1(X) = L_1^T G_1^{-1} g_1(X, \beta_*, \gamma_*^1)$ and $R_0(X) = L_0^T G_0^{-1} g_0(X, \beta_*, \gamma_*^0)$. Let

$$U_1 = T(Y - \mu_1)/\pi(X) - \{T - \pi(X)\} R_1(X)/\pi(X),$$
$$U_0 = (1 - T)(Y - \mu_0)/\{1 - \pi(X)\} - \{\pi(X) - T\} R_0(X)/\{1 - \pi(X)\}$$

Theorem 2. *When $\mathcal{P}$ contains a correctly specified model for $\pi(X)$, we have*

$$\sqrt{n}(\hat{\Delta}_{mr} - \Delta_0) \xrightarrow{L} N(0, \sigma^2)$$

where $\sigma^2 = \mathrm{Var}(Z)$, $Z = U - E(US^T) E(S^{\otimes 2})^{-1} S$, $U = U_1 - U_0$, and $\xrightarrow{L}$ denotes convergence in distribution.

It is clear that Z is the residual of the projection of U onto S, so that $\mathrm{Var}(Z) \le \mathrm{Var}(U)$. As in Han and Wang (2013), this helps us to assess the efficiency of $\hat{\Delta}_{mr}$ and identify possible ways to improve the efficiency.

Theorem 3. When $\mathcal{P}$ contains a correctly specified model for $\pi(X)$ and $\mathcal{M}$ contains a correctly specified model for m(X),

$$\sqrt{n}(\hat{\Delta}_{mr} - \Delta_0) \xrightarrow{L} N(0, \sigma_{se}^2)$$

with

$$\sigma_{se}^2 = \mathrm{Var}\{m^1(X) - m^0(X)\} + E\left\{\frac{\mathrm{Var}(Y_1|X)}{\pi(X)}\right\} + E\left\{\frac{\mathrm{Var}(Y_0|X)}{1 - \pi(X)}\right\}$$

From Theorem 3, the proposed estimator can attain the semiparamtric efficiency bound, without knowing which two among the $J + K$ models are correctly specified. This is called as local oracle efficiency in Han and Wang (2013). We also show that the oracle efficiency of $\hat{\Delta}_{mr}$ can be achieved even if the true outcome model $m(X)$ is a linear combination of the multiple outcome models in $\mathcal{M}$.

Corollary 2. *Suppose that the propensity score function $\pi(X)$ is correctly specified, when $m^t(X)$, $t = 0, 1$, is a linear combination of functions $\{m^t_k(X, \gamma^t_k), k = 1, 2, ..., K\}$, then the asymptotic variance of $\hat{\Delta}_{mr}$ attains the semiparametric efficiency bound.*

4. Simulation studies

In this section, we conduct simulation experiments in Han and Wang (2013) with some modifications to evaluate the finite-sample performance of the following five estimators: (a) the proposed estimator $\hat{\Delta}_{mr}$, (b) the IPW estimator $\hat{\Delta}_{ipw}$, (c) the OR estimator $\hat{\Delta}_{or}$, (d) the AIPW estimator $\hat{\Delta}_{aipw}$, and (e) the naive estimator $\hat{\Delta}_{naive} = \sum_{i=1}^n T_i Y_i / \sum_{i=1}^n T_i Y_i - \sum_{i=1}^n (1 - T_i) Y_i / \sum_{i=1}^n (1 - T_i)$ based on the observed Y_i's. In particular, we obtain the simulated relative bias (RB) and root mean squared error (RMSE) of estimators. All results are based on 4,000 simulation replications and the sample sizes $n = 500$ and 1000.

Example 1. In the first simulation, X is generated from an uniform distribution $U(-3, 3)$ and we consider two choices of Y_t as follows,

$$(1)\ Y_t = -0.5 + X - 0.4t + \epsilon_t, \quad \text{or} \quad (2)\ Y_t = -0.5 + 0.2\exp(X) - 0.4t + \epsilon_t$$

where ϵ_t's are independently from $N(0, \sigma^2_t(X))$, $t = 0$, 1. We generate T from the Bernoulli distribution with probability $\pi(X)$ and consider two choices of $\pi(X)$:

$$(1)\ \pi(X) = 1/\{1 + \exp(-0.8 - 0.8X)\}, \ \text{or} \ (2)\ \pi(X) = 1 - \exp\{-\exp(-0.4 - 0.6X)\}$$

We conducted simulations with two error structures: (i) the first generated homoscedastic errors with $\sigma_t(X) = 1$, ϵ_t's and X are independent; (ii) the second errors that are heteroscedastic with $\sigma_t(X) = 1/(1 + 2X^2)$. The true values of ATE are −0.4 in these cases.

Example 2. In the second simulation, we consider X is generated from a 4-dimensional random vector, where X_1 is from $\mathrm{U}(-3, 3)$ and $(X_2, X_3, X_4)^T$ is from a 3-dimensional normal distribution with mean 0 and identity covariance matrix. Two choices of Y_t are considered as follows,

$$\begin{aligned} (1)\ Y_t &= -0.5 + 0.5X_1 + 0.5X_2 - 0.4t + \epsilon_t, \\ (2)\ Y_t &= -0.5 + 0.2\exp(X_1) + 0.1\exp(X_4) - 0.4t + \epsilon_t \end{aligned}$$

where ϵ_t's are independently from $N(0,\ 1)$ and $t = 0$, 1. We generate T from the Bernoulli distribution with probability $\pi(X)$ and consider two choices of $\pi(X)$:

$$\begin{aligned} (1)\ \pi(X) &= 1/\{1 + \exp(-0.8 + 0.8X_1 - 0.2X_2)\}, \\ (2)\ \pi(X) &= 1 - \exp\{-\exp(-0.6 - 0.5X_1 - 0.1X_2)\} \end{aligned}$$

For each combination, we postulate two models for $\pi(X)$ as follows: $\pi_1(X, \beta_1) = 1/\{1 + \exp(\beta_{10} + \beta_{11}X_1 + ... + \beta_{1d}X_d)\}$ and $\pi_2(X, \beta_2) = 1 - \exp\{-\exp(\beta_{20} + \beta_{21}X_1 + ... + \beta_{2d}X_d)\}$, and two models for $m(x)$: $m^t_1(X, \gamma^t_1) = \gamma^t_{10} + \gamma^t_{11}X_1 + ... + \gamma^t_{1d}X_d$ and $m^t_2(X, \gamma^t_2) = \gamma^t_{20} + \gamma^t_{21}\exp(X_1) + ... + \gamma^t_{2d}\exp(X_d)$ for $t = 0$, 1. Following the notation of Han and Wang (2013), we use a four-digit subscript to distinguish estimators constructed using different postulated models, with each digit, from left to right, indicating if $\pi_1(X, \beta_1), \pi_2(X, \beta_2), m_1(X, \gamma_1)$ and $m_2(X, \gamma_2)$ are used. For example, $\hat{\Delta}_{ipw(1000)}$ denotes the IPW estimator based $\pi_1(X, \beta_1)$; $\hat{\Delta}_{or(0010)}$ denotes the OR estimator based $m_1(X, \gamma_1)$; $\hat{\Delta}_{aipw(1010)}$ denotes the AIPW estimator based on $\pi_1(X, \beta_1)$ and $m_1(X, \gamma_1)$; $\hat{\Delta}_{mr(1101)}$ denotes our proposed estimator based on $\pi_1(X, \beta_1), \pi_2(X, \beta_2)$ and $m_2(X, \gamma_2)$. Simulation results are presented in Tables 1–3. The Matlab codes are available from the author on request.

A few conclusions can be drawn from the simulation results.

1. If $\pi_l(X)$, $l = 1$, 2, is identical to the true propensity $\pi(X)$, the corresponding IPW estimator based on $\pi_l(X)$ is unbiased; if $m_k(X, \gamma_k)$, $k = 1$, 2, is identical to the true outcome $m(X)$, the corresponding OR estimator based on $m_k(X, \gamma_k)$ is unbiased. In these cases, as expected, all the MR estimators using $\pi_l(X)$ or $m_k(X, \gamma_k)$ are also unbiased. However, it can be seen that the biases of the IPW and OR estimators are much larger when the wrong models are used.
2. To assess the efficiency of our proposed estimators, we use the RMSEs of $\hat{\Delta}_{aipw(opt)}$ as the benchmark, where $\hat{\Delta}_{aipw(opt)}$ denote the corresponding AIPW estimator that attains the semiparametric efficiency bound under each specific data-generating model; that is $\hat{\Delta}_{aipw(1010)}$ under the first model (both π_1 and m_1 are true), $\hat{\Delta}_{mr(1001)}$ under the second model (both π_1 and m_2 are true), $\hat{\Delta}_{mr(0110)}$ under the third model (both π_2 and m_1 are true), $\hat{\Delta}_{mr(0101)}$ under the fourth model (both π_2 and m_2 are true). However, we found that the RMSEs of the corresponding MR estimators are smaller than or identical to that of $\hat{\Delta}_{aipw(opt)}$. On the other hand, under each model, it can be seen that $\hat{\Delta}_{aipw(0101)}, \hat{\Delta}_{aipw(0110)}, \hat{\Delta}_{aipw(1001)}, \hat{\Delta}_{aipw(1010)}$ are the worst AIPW estimators, respectively, while the corresponding $\hat{\Delta}_{mr(0101)}, \hat{\Delta}_{mr(0110)}, \hat{\Delta}_{mr(1001)}, \hat{\Delta}_{mr(1010)}$ perform better.
3. The RMSEs of the corresponding $\hat{\Delta}_{mr(1011)}$ are identical to that of $\hat{\Delta}_{aipw(opt)}$ under the first two models, which is in full agreement with our asymptotic theory that $\hat{\Delta}_{mr(1011)}$ attains the semiparametric efficiency bound in these two cases. Compared to $\hat{\Delta}_{aipw(opt)}$ under the last two models, $\hat{\Delta}_{mr(0111)}$ have more samller RMSEs and the conclusions also hold.
4. Under these four models, the proposed $\hat{\Delta}_{mr(1010)}, \hat{\Delta}_{mr(1001)}, \hat{\Delta}_{mr(0110)}, \hat{\Delta}_{mr(0101)}$ are doubly robust because only one propensity score model and one outcome regression model were being assumed. However, under the same occasions, it can be seen that the MR estimators $\hat{\Delta}_{mr(1010)}, \hat{\Delta}_{mr(1001)}, \hat{\Delta}_{mr(0110)}, \hat{\Delta}_{mr(0101)}$ are more efficient than the AIPW estimators counterparts $\hat{\Delta}_{aipw(1010)}, \hat{\Delta}_{aipw(1001)}, \hat{\Delta}_{aipw(0110)}, \hat{\Delta}_{aipw(0101)}$, respectively.
5. The multiple robustness of the proposed estimator is well demonstrated by inspecting the bias of $\hat{\Delta}_{mr(1110)}, \hat{\Delta}_{mr(1101)}, \hat{\Delta}_{mr(1011)}, \hat{\Delta}_{mr(0111)}$, and $\hat{\Delta}_{mr(1111)}$, all of which are consistent under the four data generating scenarios according to our theory. Neither the AIPW estimator nor any existing doubly robust estimator can achieve such robustness.
6. According to our theory, $\hat{\Delta}_{mr(1111)}$ attains the semiparametric efficiency bound under all four data-generating models; this is confirmed by comparing its RMESs to that of $\hat{\Delta}_{mr(1010)}, \hat{\Delta}_{mr(1001)}, \hat{\Delta}_{mr(0110)}, \hat{\Delta}_{mr(0101)}$, respectively, under each

Table 1. Results of Example 1 when $\sigma_t(X) = 1$; values are multiplied by 100.

	π_1 is true				π_2 is true			
	m_1 is true		m_2 is true		m_1 is true		m_2 is true	
Method	RB	RMSE	RB	RMSE	RB	RMS	RB	RMSE
				$n = 500$				
$\hat{\Delta}_{naive}$	474.84	190.66	229.32	92.63	506.79	203.30	212.47	85.83
$\hat{\Delta}_{ipw(1000)}$	1.89	20.07	0.67	13.96	30.47	19.18	−20.24	18.14
$\hat{\Delta}_{ipw(0100)}$	−55.93	40.33	13.54	21.06	3.01	23.25	0.84	16.92
$\hat{\Delta}_{or(0010)}$	1.62	11.34	23.40	15.81	1.95	11.86	−22.69	15.60
$\hat{\Delta}_{or(0001)}$	167.22	68.30	0.39	10.67	235.87	95.32	0.49	10.33
$\hat{\Delta}_{aipw(1010)}$	0.97	13.04	0.22	14.56	0.56	12.36	−27.35	18.63
$\hat{\Delta}_{aipw(1001)}$	1.94	19.82	0.68	12.97	43.33	22.77	0.84	12.06
$\hat{\Delta}_{aipw(0110)}$	0.98	17.52	28.54	22.94	0.95	14.39	0.49	16.50
$\hat{\Delta}_{aipw(0101)}$	−63.87	39.18	1.13	18.71	2.52	21.40	0.80	14.21
$\hat{\Delta}_{mr(1010)}$	1.18	14.13	1.44	14.39	1.12	15.04	−11.57	16.43
$\hat{\Delta}_{mr(1001)}$	1.98	13.75	0.75	12.93	5.36	13.79	0.64	13.05
$\hat{\Delta}_{mr(0110)}$	1.32	15.33	14.74	16.76	0.63	14.42	−3.21	15.10
$\hat{\Delta}_{mr(0101)}$	−2.29	13.97	0.75	13.33	2.13	14.29	0.73	13.50
$\hat{\Delta}_{mr(1110)}$	1.19	14.92	0.05	14.72	1.07	15.05	−2.36	15.64
$\hat{\Delta}_{mr(1101)}$	1.87	13.63	0.76	13.18	2.14	17.08	0.13	15.21
$\hat{\Delta}_{mr(1011)}$	1.23	14.24	0.71	13.81	1.28	15.83	0.49	15.82
$\hat{\Delta}_{mr(0111)}$	1.49	15.02	0.57	14.27	2.26	16.79	0.23	16.05
$\hat{\Delta}_{mr(1111)}$	0.89	15.11	0.46	14.64	2.02	17.48	0.10	16.23
				$n = 1000$				
$\hat{\Delta}_{naive}$	474.28	190.06	229.93	92.44	507.58	202.93	212.27	85.32
$\hat{\Delta}_{ipw(1000)}$	0.76	13.85	−0.20	9.84	29.96	16.01	−20.27	13.88
$\hat{\Delta}_{ipw(0100)}$	−56.68	31.78	12.48	14.09	1.74	14.61	0.24	11.74
$\hat{\Delta}_{or(0010)}$	0.91	8.34	23.31	12.89	1.82	8.76	−22.85	12.74
$\hat{\Delta}_{or(0001)}$	166.45	67.38	−0.39	7.61	236.57	94.72	0.23	7.35
$\hat{\Delta}_{aipw(1010)}$	−0.17	9.13	−0.71	10.27	0.63	8.62	−27.32	15.14
$\hat{\Delta}_{aipw(1001)}$	0.58	13.78	−0.16	9.15	43.71	20.34	0.37	8.43
$\hat{\Delta}_{aipw(0110)}$	−0.23	11.97	27.97	17.24	0.66	9.68	−0.18	11.54
$\hat{\Delta}_{aipw(0101)}$	−64.89	32.98	−0.34	12.38	1.39	13.82	0.45	9.92
$\hat{\Delta}_{mr(1010)}$	−0.02	9.22	−0.01	9.78	0.44	10.01	−12.24	11.32
$\hat{\Delta}_{mr(1001)}$	0.34	9.35	−0.15	9.14	5.92	9.54	0.28	9.02
$\hat{\Delta}_{mr(0110)}$	−0.31	9.84	14.24	11.83	0.22	10.12	−0.91	10.06
$\hat{\Delta}_{mr(0101)}$	−2.41	9.54	−0.20	9.41	1.08	9.46	0.23	9.56
$\hat{\Delta}_{mr(1110)}$	−0.16	10.72	−0.31	9.81	0.16	10.21	−0.73	10.32
$\hat{\Delta}_{mr(1101)}$	0.19	9.24	−0.17	9.22	0.51	10.32	0.27	9.68
$\hat{\Delta}_{mr(1011)}$	0.03	9.31	−0.13	9.45	0.24	10.36	0.42	10.25
$\hat{\Delta}_{mr(0111)}$	0.09	9.39	−0.09	9.20	0.21	10.52	0.46	10.10
$\hat{\Delta}_{mr(1111)}$	−0.30	9.75	−0.14	9.69	0.27	10.54	0.55	10.78

Notes: RB, relative bias; RMSE, root mean squared error.

model. Adding more models did not lead to any noticeable increase in RMSE, consistent with what the theoretical results suggested.

In conclusion, the simulation results suggested that the MR estimators provide extra protection against model misspecification while estimation efficiency is not compromised.

5. ACTG 175 data

In this section, we illustrate the proposed method using data collected on 2139 HIV positive patients enrolled in AIDS Clinical Trials Group Protocol 175 (ACTG 175) (Hammer et al. 1996). In this HIV clinical trial, the patients were randomized into four

Table 2. Results of Example 1 when $\sigma_t(X) = (1 + 2X^2)^{-1}$; values are multiplied by 100.

	π_1 is true				π_2 is true			
	m_1 is true		m_2 is true		m_1 is true		m_2 is true	
Method	RB	RMSE	RB	RMSE	RB	RMS	RB	RMSE
				$n = 500$				
$\hat{\Delta}_{naive}$	474.77	190.48	229.48	92.47	507.17	203.32	213.15	85.86
$\hat{\Delta}_{ipw(1000)}$	1.08	17.88	0.46	7.97	30.41	16.11	−19.37	14.27
$\hat{\Delta}_{ipw(0100)}$	−56.70	40.21	13.35	10.31	2.41	18.96	0.64	10.92
$\hat{\Delta}_{or(0010)}$	1.73	8.74	23.53	13.80	1.45	8.29	−21.99	12.61
$\hat{\Delta}_{or(0001)}$	166.60	67.67	−0.46	6.43	235.52	94.85	0.48	6.78
$\hat{\Delta}_{aipw(1010)}$	−0.22	6.20	−0.76	8.87	−0.36	5.78	−26.42	15.35
$\hat{\Delta}_{aipw(1001)}$	1.19	16.48	0.06	5.95	42.61	20.16	0.52	5.89
$\hat{\Delta}_{aipw(0110)}$	−0.56	6.63	27.50	14.87	−0.51	5.93	−0.15	10.83
$\hat{\Delta}_{aipw(0101)}$	−64.84	38.09	−0.04	6.51	1.76	17.04	0.43	6.02
$\hat{\Delta}_{mr(1010)}$	−0.20	6.33	0.52	7.37	−0.37	6.19	−13.06	9.22
$\hat{\Delta}_{mr(1001)}$	0.67	7.12	0.03	6.00	6.19	6.89	0.41	5.96
$\hat{\Delta}_{mr(0110)}$	−0.17	6.61	14.75	9.78	−0.67	6.25	−3.76	6.78
$\hat{\Delta}_{mr(0101)}$	−2.16	6.79	0.01	5.96	1.12	6.33	0.39	6.04
$\hat{\Delta}_{mr(1110)}$	−0.29	6.83	−0.71	7.01	−0.40	6.36	−2.85	7.11
$\hat{\Delta}_{mr(1101)}$	0.56	6.67	0.03	6.05	1.49	7.42	−0.07	7.53
$\hat{\Delta}_{mr(1011)}$	−0.30	6.73	0.05	6.18	−0.35	6.52	0.26	6.39
$\hat{\Delta}_{mr(0111)}$	−0.42	6.97	−0.01	6.25	1.18	7.53	−0.02	6.64
$\hat{\Delta}_{mr(1111)}$	−0.07	7.15	−0.07	6.44	0.41	7.10	−0.25	7.11
				$n = 1000$				
$\hat{\Delta}_{naive}$	474.29	189.72	229.72	92.20	506.14	202.69	211.98	85.24
$\hat{\Delta}_{ipw(1000)}$	1.39	12.12	−0.19	5.67	30.27	13.92	−20.10	11.37
$\hat{\Delta}_{ipw(0100)}$	−55.70	29.37	12.60	8.06	1.09	11.75	−0.07	7.29
$\hat{\Delta}_{or(0010)}$	2.19	6.30	23.37	11.27	1.66	6.53	−22.04	11.48
$\hat{\Delta}_{or(0001)}$	166.59	67.18	−0.24	4.63	235.26	94.48	−0.29	4.93
$\hat{\Delta}_{aipw(1010)}$	0.39	4.33	−0.29	6.38	−0.06	4.07	−27.09	12.89
$\hat{\Delta}_{aipw(1001)}$	1.37	11.08	−0.05	4.36	42.91	18.79	0.09	3.99
$\hat{\Delta}_{aipw(0110)}$	0.17	4.48	28.35	13.27	−0.04	4.12	−0.51	6.95
$\hat{\Delta}_{aipw(0101)}$	−63.76	30.53	−0.11	4.58	0.91	10.77	0.07	4.14
$\hat{\Delta}_{mr(1010)}$	0.36	4.38	0.15	5.21	0.05	4.24	−12.55	6.78
$\hat{\Delta}_{mr(1001)}$	0.72	4.93	−0.04	4.32	5.83	4.97	0.07	4.08
$\hat{\Delta}_{mr(0110)}$	−0.38	5.54	14.62	6.97	−0.11	4.29	−1.09	4.89
$\hat{\Delta}_{mr(0101)}$	−2.19	4.77	−0.04	4.27	0.65	4.38	0.06	4.12
$\hat{\Delta}_{mr(1110)}$	0.22	5.72	0.05	4.70	−0.02	4.68	−0.84	4.62
$\hat{\Delta}_{mr(1101)}$	0.59	4.60	−0.03	4.33	0.72	5.21	0.04	4.17
$\hat{\Delta}_{mr(1011)}$	0.37	4.38	−0.02	4.35	0.06	4.32	−0.01	4.28
$\hat{\Delta}_{mr(0111)}$	0.36	4.41	−0.01	4.34	0.08	4.46	−0.05	4.33
$\hat{\Delta}_{mr(1111)}$	0.31	4.62	−0.01	4.40	0.21	4.73	−0.06	4.35

Notes: RB, relative bias; RMSE, root mean squared error.

arms to receive the respective antiretroviral regimen. Due to death and dropout, 532 subjects with zidovudine or ZDV, 522 subjects with didanosine or ddi, 524 subjects with ZDV + ddi, 561 subjects with ZDV + zalcitabine were collected.

In this study, the CD4 cell count is of prime interest which decreases as HIV progresses. Let response Y_s be the CD4 count at 96 ± 5 weeks which receiving the sth antiretroviral regimen, where $s = 0, 1, 2, 3$ denotes for ZDV, ddi, ZDV + ddi and ZDV + zalcitabine, respectively. There are six continues baseline covariates X: age, weight, CD4 cell counts at baseline and 20 ± 5 weeks, and CD8 cell counts at baseline and 20 ± 5 weeks. The ZDC is a traditional regimen, while the ddi, ZDV + ddi and ZDV + zalcitabine are new regimens. Thus, we compute $\Delta_s = E(Y_s) - E(Y_0)$ to see whether the new regimens work or not.

Table 3. Results of Example 2; values are multiplied by 100.

	π_1 is true				π_2 is true			
	m_1 is true		m_2 is true		m_1 is true		m_2 is true	
Method	RB	RMSE	RB	RMSE	RB	RMS	RB	RMSE
				$n = 500$				
$\hat{\Delta}_{naive}$	219.21	88.58	228.63	92.36	229.95	92.86	178.62	72.43
$\hat{\Delta}_{ipw(1000)}$	0.45	17.01	0.18	14.57	6.48	12.27	−15.53	16.69
$\hat{\Delta}_{ipw(0100)}$	−32.51	32.94	16.93	22.83	1.20	12.76	0.94	14.94
$\hat{\Delta}_{or(0010)}$	7.21	18.54	−4.75	23.72	14.11	19.03	−29.31	24.53
$\hat{\Delta}_{or(0001)}$	76.74	44.86	−2.93	26.67	100.36	51.45	1.53	23.54
$\hat{\Delta}_{aipw(1010)}$	0.17	13.56	0.29	14.97	−0.13	11.25	−14.39	14.95
$\hat{\Delta}_{aipw(1001)}$	0.26	16.24	0.19	13.20	10.41	12.98	0.47	11.33
$\hat{\Delta}_{aipw(0110)}$	0.63	23.98	31.53	27.62	−0.06	11.31	0.24	13.54
$\hat{\Delta}_{aipw(0101)}$	−31.02	29.64	−0.08	21.79	0.63	12.38	0.57	11.46
$\hat{\Delta}_{mr(1010)}$	−0.75	13.42	0.49	14.27	0.06	10.92	−9.01	12.94
$\hat{\Delta}_{mr(1001)}$	0.90	13.92	−0.05	13.14	4.12	11.21	0.34	11.21
$\hat{\Delta}_{mr(0110)}$	−0.03	13.86	15.46	16.15	−0.28	11.48	0.25	12.37
$\hat{\Delta}_{mr(0101)}$	−7.28	14.98	−0.69	13.97	1.18	11.66	0.46	11.37
$\hat{\Delta}_{mr(1110)}$	−1.45	16.52	1.36	15.27	−0.42	11.66	−0.43	12.25
$\hat{\Delta}_{mr(1101)}$	0.31	14.21	−0.42	13.64	0.57	11.57	0.04	12.18
$\hat{\Delta}_{mr(1011)}$	−0.46	15.54	−1.03	14.26	0.01	11.60	0.24	11.30
$\hat{\Delta}_{mr(0111)}$	−0.95	15.29	−0.10	14.13	0.01	11.45	0.01	11.45
$\hat{\Delta}_{mr(1111)}$	−0.22	15.49	−0.19	13.84	−0.46	11.85	0.34	11.48
				$n = 1000$				
$\hat{\Delta}_{naive}$	228.18	87.88	227.96	91.64	229.85	92.33	177.71	71.57
$\hat{\Delta}_{ipw(1000)}$	0.26	11.18	−0.28	10.47	5.97	8.89	−16.43	12.36
$\hat{\Delta}_{ipw(0100)}$	−37.57	26.15	16.45	17.36	0.34	9.28	1.08	10.49
$\hat{\Delta}_{or(0010)}$	6.39	18.06	−8.22	22.34	16.71	17.64	−29.40	23.19
$\hat{\Delta}_{or(0001)}$	80.61	43.47	1.92	24.90	92.52	49.17	3.50	21.59
$\hat{\Delta}_{aipw(1010)}$	0.19	9.16	−0.24	10.87	0.21	8.07	−15.38	11.10
$\hat{\Delta}_{aipw(1001)}$	0.25	10.79	−0.08	9.43	9.94	9.48	−0.31	7.91
$\hat{\Delta}_{aipw(0110)}$	−0.18	14.85	32.52	21.11	0.31	7.92	−0.48	9.47
$\hat{\Delta}_{aipw(0101)}$	−36.15	27.08	−1.10	14.70	0.52	8.79	−0.24	7.98
$\hat{\Delta}_{mr(1010)}$	−0.38	9.44	0.33	9.96	0.27	7.72	−9.44	9.41
$\hat{\Delta}_{mr(1001)}$	0.53	9.73	−0.25	9.21	3.86	8.73	−0.46	8.01
$\hat{\Delta}_{mr(0110)}$	−0.63	9.88	16.21	12.56	0.23	8.00	−0.51	8.45
$\hat{\Delta}_{mr(0101)}$	−8.40	10.72	0.17	9.91	1.18	8.34	−0.45	9.57
$\hat{\Delta}_{mr(1110)}$	−1.63	11.61	1.25	11.20	−0.53	8.74	−0.61	8.36
$\hat{\Delta}_{mr(1101)}$	0.08	9.84	−0.72	9.96	0.39	8.16	−0.22	7.84
$\hat{\Delta}_{mr(1011)}$	−0.24	10.12	−0.36	9.74	0.16	7.86	−0.51	7.78
$\hat{\Delta}_{mr(0111)}$	−1.43	13.23	−0.09	10.53	0.19	8.19	−0.18	7.73
$\hat{\Delta}_{mr(1111)}$	−0.13	10.86	−0.93	10.95	−0.35	8.42	−0.74	8.18

Notes: RB, relative bias; RMSE, root mean squared error.

We think it is reasonable to assume that, given the six baseline covariates, the treatment assignment does not depend on the CD4 count at 96 ± 5 weeks. Thus, the unconfoundedness assumption holds. The two propensity score models $\pi_1(X, \beta_1), \pi_2(X, \beta_2)$ and two outcome vector models $m_1(X, \gamma_1), m_2(X, \gamma_2)$ in Section 4 are used. The point estimates and their standard errors based on the bootstrap with replication size 200 are reported in Table 4.

Under each occasion, it can be seen that the proposed MR estimates are close except $\hat{\Delta}_{mr}^{0011}$, and smaller than $\hat{\Delta}_{naive}$ in all cases. However, the IPW and AIPW estimates are not robust. On the other hand, the standard errors of the MR estimates are smaller than those of $\hat{\Delta}_{naive}$, the IPW estimates and the AIPW estimates. All these estimates indicate that the last three regimens have significantly higher CD4 counts at 96 ± 5 weeks than the first regimen.

Table 4. Estimates (with standard errors in parentheses) for the ACTG 175 data.

Method	Δ_1	Δ_2	Δ_3
$\hat{\Delta}_{naive}$	39.53 (13.79)	39.60 (13.83)	39.47 (13.24)
$\hat{\Delta}_{ipw(1000)}$	32.53 (25.39)	32.49 (27.10)	32.48 (16.16)
$\hat{\Delta}_{ipw(0100)}$	24.00 (25.84)	24.24 (20.41)	24.08 (16.75)
$\hat{\Delta}_{aipw(1010)}$	29.63 (17.37)	29.75 (25.20)	29.66 (12.82)
$\hat{\Delta}_{aipw(1001)}$	18.27 (18.67)	18.52 (20.24)	18.33 (16.40)
$\hat{\Delta}_{aipw(0110)}$	25.86 (21.52)	26.02 (16.85)	25.84 (13.87)
$\hat{\Delta}_{aipw(0101)}$	14.56 (18.69)	14.80 (19.93)	14.55 (15.94)
$\hat{\Delta}_{mr(1010)}$	29.25 (10.92)	29.33 (12.49)	29.26 (10.35)
$\hat{\Delta}_{mr(1001)}$	25.69 (11.46)	25.79 (12.33)	25.70 (10.95)
$\hat{\Delta}_{mr(0110)}$	27.68 (12.98)	27.83 (11.63)	27.71 (10.33)
$\hat{\Delta}_{mr(0101)}$	24.75 (12.50)	24.91 (12.12)	24.75 (10.81)
$\hat{\Delta}_{mr(1110)}$	29.05 (11.91)	29.15 (11.20)	29.06 (10.13)
$\hat{\Delta}_{mr(1101)}$	26.07 (12.13)	26.17 (11.92)	26.05 (10.98)
$\hat{\Delta}_{mr(1011)}$	25.44 (11.69)	25.54 (12.97)	25.42 (10.87)
$\hat{\Delta}_{mr(0111)}$	24.51 (13.15)	24.66 (12.37)	24.51 (10.82)
$\hat{\Delta}_{mr(1111)}$	26.06 (12.52)	26.16 (11.96)	26.04 (10.74)

6. Discussion

In this paper, we consider the estimation of ATE and allow multiple models for both the propensity score models and the outcome models, and then propose a new estimator, which is shown to be more robust than doubly robust estimators. The proposed methodology is a general method that can be readily adopted for estimating other unknown parameters. For example, the average treatment effect on the treated, i.e., $E(Y_1 - Y_0|T = 1)$, or the weighted average treatment effect (Hahn 1998). Another interesting question is estimate the quantile treatment effect, i.e., $q_1 - q_0$, where q_t is the τ-th quantile of Y_t, $t = 0$, 1, with a fixed $\tau \in (0, 1)$, e.g., $\tau = 0.5$, 0.25 and 0.75 give the difference of medians, lower quartiles and upper quartiles, respectively. For example, when the outcomes are highly skewed medical cost and utilization, the ATE may be highly sensitive to outliers and it is more reasonable to learn about distributional impacts beyond the average effect of treatment.

In theory, the number of models postulated in $\mathcal{P} = \{\pi_l(X, \beta_l), l = 1, 2, ..., L\}$ and $\mathcal{M} = \{m_k(X, \gamma_k), k = 1, 2, ..., K\}$ has no effect on the multiple robustness, as long as L and K are fixed. However, when the sample size is small and/or L and K are large, the convex hulls of $g_1(X, \hat{\beta}, \hat{\gamma}^1)$ and $g_0(X, \hat{\beta}, \hat{\gamma}^0)$ may not contain the vector 0 such that the set for $\hat{w}_{1i}$ and $\hat{w}_{0j}$ are empty. To solve the empty set problem, the adjusted EL (Chen et al. 2008), the balanced augmented EL (Emerson and Owen 2009) and the extended EL (Tsao and Wu 2013) methods can be used. In addition, Owen (2001), Tsao and Zhou (2001) pointed out that the EL method lacks robustness, which is sensitive to outliers. To address this issue, we can tilt the empirical likelihood function by assigning smaller weights to outliers to yield a more robust estimator. Further, when the dimension of covariate X is high, the problem of variable selection in causal inference and the EL should also be considered. The penalized EL can be used to select the subset of influential covariates and obtain the more efficient MR estimators.

Throughout this paper, we assume the unconfoundedness condition holds, i.e., $(Y_1, Y_0) \perp\!\!\!\perp T|X$. However, in observational studies, there often exist unmeasured confounding variables $\mathcal{U}$, such that $(Y_1, Y_0) \perp\!\!\!\perp T|(X, \mathcal{U})$. In this case, the differences in outcome measures between treated and untreated, may be due to differences in selecting treatment status, that also predict the outcome. Such confounding may cause some problems for evaluating the ATE, i.e., the estimates may have large biases or population parameters are not identifiable (Fitzmaurice et al. 1995; Shao and Wang 2016). Extension to this case will also be a topic of our future research.

Acknowledgments

We are grateful to the Editor, the Associate Editor and two anonymous referee for their insightful comments and suggestions on this article, which have led to significant improvements. This work was supported by the National Natural Science Foundation of China (11501208, 11871287) and the Fundamental Research Funds for the Central Universities.

Disclosure statement

The author have declared no conflict of interest.

References

Bang, H., and J. M. Robins. 2005. Doubly robust estimation in missing data and causal inference models. *Biometrics* 61 (4):962–73. doi:10.1111/j.1541-0420.2005.00377.x.

Cao, W., A. A. Tsiatis, and M. Davidian. 2009. Improving efficiency and robustness of the doubly robust estimator for a population mean with incomplete data. *Biometrika* 96 (3):723–34. doi:10.1093/biomet/asp033.

Chen, J., R. R. Sitter, and C. Wu. 2002. Using empirical likelihood methods to obtain range restricted weights in regression estimators for surveys. *Biometrika* 89 (1):230–7. doi:10.1093/biomet/89.1.230.

Chen, J., A. M. Variyath, and B. Abraham. 2008. Adjusted empirical likelihood and its properties. *Journal of Computational and Graphical Statistics* 17 (2):426–43. doi:10.1198/106186008X321068.

Emerson, S. C., and A. B. Owen. 2009. Calibration of the empirical likelihood method for a vector mean. *Electronic Journal of Statistics* 3 (0):1161–92. doi:10.1214/09-EJS518.

Fitzmaurice, G. M., G. Molenberghs, and S. R. Lipsitz. 1995. Regression models for longitudinal binary responses with informative drop-outs. *Journal of the Royal Statistical Society. Series B (Methodological)* 57:691–704.

Hahn, J. 1998. On the role of the propensity score in efficient semiparametric estimation of average treatment effects. *Econometrica* 66 (2):315–31. doi:10.2307/2998560.

Hammer, S. M., D. A. Katzenstein, M. D. Hughes, H. Gundacker, R. T. Schooley, R. H. Haubrich, W. K. Henry, M. M. Lederman, J. P. Phair, M. Niu, et al. 1996. A trial comparing nucleoside monotherapy with combination therapy in HIV-infected adults with CD4 cell counts from 200 to 500 per cubic millimeter. *New England Journal of Medicine* 335 (15):1081–9. doi:10.1056/NEJM199610103351501.

Han, P. 2014. Multiply robust estimation in regression analysis with missing data. *Journal of the American Statistical Association* 109 (507):1159–73. doi:10.1080/01621459.2014.880058.

Han, P., and L. Wang. 2013. Estimation with missing data: beyond double robustness. *Biometrika* 100 (2):417–30. doi:10.1093/biomet/ass087.

Hirano, K., G. W. Imbens, and G. Ridder. 2003. Efficient estimation of average treatment effects using the estimated propensity score. *Econometrica* 71 (4):1161–89. doi:10.1111/1468-0262.00442.

Horvitz, D. G., and D. J. Thompson. 1952. A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association* 47 (260):663–85. doi:10.1080/01621459.1952.10483446.

Naik, C., E. J. McCoy, and D. J. Graham. 2016. Multiply robust dose-response estimation for multivalued causal inference problems. arXiv:1611.02433.

Owen, A. B. 1988. Empirical likelihood ratio confidence intervals for a single functional. *Biometrika* 75 (2):237–49. doi:10.1093/biomet/75.2.237.

Owen, A. B. 2001. *Empirical likelihood*. New York, NY: Chapman and Hall/CRC Press.

Qin, J., and J. Lawless. 1994. Empirical likelihood and general estimating equations. *The Annals of Statistics* 22 (1):300–25. doi:10.1214/aos/1176325370.

Qin, J., and B. Zhang. 2007. Empirical-likelihood-based inference in missing response problems and its application in observational studies. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 69 (1):101–22. doi:10.1111/j.1467-9868.2007.00579.x.

Robins, J. M., A. Rotnitzky, and L. P. Zhao. 1994. Estimation of regression conficients when some regressors are not always observed. *Journal of the American Statistical Association* 89 (427):846–66. doi:10.1080/01621459.1994.10476818.

Rosenbaum, P. R., and D. B. Rubin. 1983. The Central role of the propensity score in observational studies for causal effects. *Biometrika* 70 (1):41–55. doi:10.1093/biomet/70.1.41.

Shao, J., and L. Wang. 2016. Semiparametric inverse propensity weighting for nonignorable missing data. *Biometrika* 103 (1):175–87. doi:10.1093/biomet/asv071.

Tan, Z. 2010. Bounded, efficient and doubly robust estimation with inverse weighting. *Biometrika* 97 (3):661–82. doi:10.1093/biomet/asq035.

Tsao, M., and J. Zhou. 2001. On the robustness of empirical likelihood ratio confidence intervals for location. *Canadian Journal of Statistics* 29 (1):129–40. doi:10.2307/3316056.

Tsao, M., and F. Wu. 2013. Empirical likelihood on the full parameter space. *The Annals of Statistics* 41 (4):2176–96. doi:10.1214/13-AOS1143.

Tsiatis, A. A. 2006. *Semiparametric theory and missing data*. New York, NY: Springer.

White, H. 1982. Maximum likelihood estimation of misspecified models. *Econometrica* 50 (1): 1–25. doi:10.2307/1912526.

Appendix

Proof of Theorem 1: (a) Suppose that $\mathcal{P}$ contains a correctly specified model for $\pi(X)$, say $\pi_1(X, \beta_1)$ without loss of generality. Let β_1^0 denote the true value of β_1 such that $\pi_1(X, \beta_1^0) = \pi(X)$ and denote $\hat{\beta}_1$ as the maximizer of (1). Denote p_i and q_j as the conditional empirical probability mass on (Y_{1i}, X_{1i}) and (Y_{0j}, X_{0j}), respectively, $i = 1, \ldots, n_1, j = 1, \ldots, n_0$. As in Han (2014), we maximize $\prod_{i=1}^{n_1} p_i \prod_{j=1}^{n_0} q_j$ subjects to $p_i \geq 0, q_j \geq 0$, and

$$\sum_{i=1}^{n_1} p_i = \sum_{j=1}^{n_0} q_j = 1, \sum_{i=1}^{n_1} p_i \frac{g_1(X_{1i}, \hat{\beta}, \hat{\gamma}^1)}{\pi_1\left(X_{1i}, \hat{\beta}_1\right)} = 0, \sum_{j=1}^{n_0} q_j \frac{g_0\left(X_{0j}, \hat{\beta}, \hat{\gamma}^0\right)}{1 - \pi_1\left(X_{0j}, \hat{\beta}_1\right)} = 0.$$

For $i = 1, 2, \ldots, n_1$ and $j = 1, 2, \ldots, n_0$, applying Lagrange multipliers, we have

$$\hat{p}_i = \frac{1}{n_1} \frac{1}{1 + \hat{\lambda}_1^T g_1(X_{1i}, \hat{\beta}, \hat{\gamma}^1)/\pi_1\left(X_{1i}, \hat{\beta}_1\right)},$$
$$\hat{q}_j = \frac{1}{n_0} \frac{1}{1 + \hat{\lambda}_0^T g_0\left(X_{0j}, \hat{\beta}, \hat{\gamma}^0\right)/\left\{1 - \pi_1\left(X_{0j}, \hat{\beta}_1\right)\right\}},$$

where $\hat{\lambda}_t = (\hat{\lambda}_{t1}, \ldots, \hat{\lambda}_{tS})^T$, $t = 0,\ 1$, is S-dimensional vector that solves the equations

$$\begin{aligned}&\sum_{i=1}^{n_1}\frac{g_1\left(X_{1i},\hat{\beta},\hat{\gamma}^1\right)/\pi_1\left(X_{1i},\hat{\beta}_1\right)}{1+\hat{\lambda}_1^T g_1\left(X_{1i},\hat{\beta},\hat{\gamma}^1\right)/\pi_1\left(X_{1i},\hat{\beta}_1\right)}=0,\\ &\sum_{j=1}^{n_0}\frac{g_0\left(X_{0j},\hat{\beta},\hat{\gamma}^0\right)/\left(1-\pi_1\left(X_{0j},\hat{\beta}_1\right)\right)}{1+\hat{\lambda}_0^T g_0\left(X_{0j},\hat{\beta},\hat{\gamma}^0\right)/\left\{1-\pi_1\left(X_{0j},\hat{\beta}_1\right)\right\}}=0.\end{aligned}$$

It can be seen that

$$\hat{\rho}_{11}=\frac{\hat{\lambda}_{11}+1}{\hat{\theta}_1},\ \ \hat{\rho}_{01}=\frac{\hat{\lambda}_{01}+1}{1-\hat{\theta}_1},\ \ \hat{\rho}_{1h}=\frac{\hat{\lambda}_{1h}}{\hat{\theta}_1},\ \ \hat{\rho}_{0h}=\frac{\hat{\lambda}_{0h}}{1-\hat{\theta}_1},$$

for $h=2,\dots,J+K$. Therefore, for $i=1,2,\dots,n_1$ and $j=1,2,\dots,n_0$,

$$\begin{aligned}\hat{w}_{1i}&=\frac{1}{n_1}\frac{\hat{\theta}_1/\pi_1\left(X_{1i},\hat{\beta}_1\right)}{1+\hat{\lambda}_1^T g_1\left(X_{1i},\hat{\beta},\hat{\gamma}^1\right)/\pi_1\left(X_{1i},\hat{\beta}_1\right)}=\frac{\hat{p}_i\hat{\theta}_1}{\pi_1\left(X_{1i},\hat{\beta}_1\right)},\\ \hat{w}_{0j}&=\frac{1}{n_0}\frac{\left(1-\hat{\theta}_1\right)/\left(1-\pi_1\left(X_{0j},\hat{\beta}_1\right)\right)}{1+\hat{\lambda}_0^T g_0\left(X_{0j},\hat{\beta},\hat{\gamma}^0\right)/\left\{1-\pi_1\left(X_{0j},\hat{\beta}_1\right)\right\}}=\frac{\hat{q}_j\left(1-\hat{\theta}_1\right)}{1-\pi_1\left(X_{0j},\hat{\beta}_1\right)}.\end{aligned}$$

As in Owen (2001), we have $\hat{\lambda}_1^T=O_p(n^{-1/2})$ and $\hat{\lambda}_0^T=O_p(n^{-1/2})$. Using the fact $n_1/n\xrightarrow{p}\hat{\theta}_1$ and $\pi_1(X,\beta_1)$ is correctly specified, we have

$$\begin{aligned}&\sum_{i=1}^{n_1}\hat{p}_i y_{1i}-\sum_{j=1}^{n_0}\hat{q}_j y_{0j}\\ &=\frac{\hat{\theta}_1}{n_1}\sum_{i=1}^{n}\frac{T_iY_i/\pi_1\left(X_i,\hat{\beta}_1\right)}{1+\hat{\lambda}_1^T g_1\left(X_i,\hat{\beta},\hat{\gamma}^1\right)/\pi_1\left(X_i,\hat{\beta}_1\right)}-\frac{1-\hat{\theta}_1}{n_0}\sum_{j=1}^{n}\frac{\left(1-T_j\right)Y_j/\left\{1-\pi_1\left(X_j,\hat{\beta}_1\right)\right\}}{1+\hat{\lambda}_0^T g_0\left(X_j,\hat{\beta},\hat{\gamma}^0\right)/\left\{1-\pi_1\left(X_j,\hat{\beta}_1\right)\right\}}\\ &=\frac{1}{n}\sum_{i=1}^{n}\frac{T_iY_i}{\pi_1(X_i,\beta_1)}-\frac{1}{n}\sum_{j=1}^{n}\frac{\left(1-T_j\right)Y_j}{1-\pi_1\left(X_j,\beta_1\right)}+o_p(1)\\ &\xrightarrow{p}E\left[\frac{TY}{\pi(X)}-\frac{(1-T)Y}{1-\pi(X)}\right].\end{aligned}$$

(b) Suppose that $\mathcal{M}$ contains a correctly specified vector model for $m(X)$. Without loss of generality, let $m_1(X,\gamma_1)=\{m_1^1(X,\gamma_1^1),m_1^0(X,\gamma_1^0)\}^T$ be this model and denote the true value of γ_1^t as γ_{10}^t such that $m_1^t(X,\gamma_{10}^t)=E(Y_t|X)$. As in White (1982) and Han (2014), we know that

$$\hat{\beta}_l\xrightarrow{p}\beta_{l*}\quad\text{and}\quad\hat{\gamma}_k^t\xrightarrow{p}\gamma_{k*}^t,$$

where β_{l*} and γ_{k*}^t minimize the corresponding Kullback–Leibler distance between the probability distribution based on the postulated models and the one generating the data. Let $\beta_*=(\beta_{1*},\dots,\beta_{L*})^T$, $\gamma_*^t=(\gamma_{1*}^t,\dots,\gamma_{K*}^t)^T$ and λ_{t*} be the probability limit of $\hat{\lambda}_t$. According (3) and (4), we have

$$\begin{aligned}\sum_{i=1}^{n_1}\hat{p}_i m_1^1\left(X_{1i},\hat{\gamma}_1^1\right)&=n^{-1}\sum_{i=1}^{n}m_1^1\left(X_i,\hat{\gamma}_1^1\right)\xrightarrow{p}\mu_1,\\ \sum_{j=1}^{n_0}\hat{q}_j m_1^0\left(X_{0j},\hat{\gamma}_1^0\right)&=n^{-1}\sum_{j=1}^{n}m_1^0\left(X_j,\hat{\gamma}_1^0\right)\xrightarrow{p}\mu_0.\end{aligned}$$

Thus,

$$\Delta_{mr} = \sum_{i=1}^{n_1}\hat{p}_i\{Y_{1i} - m_1^1(X_{1i},\hat{\gamma}_1^1)\} - \sum_{j=1}^{n_0}\hat{q}_j\left\{Y_{0j} - m_1^0\left(X_{0j},\hat{\gamma}_1^0\right)\right\} + \mu_1 - \mu_0 + o_p(1).$$

It can be shown that

$$\begin{aligned}\sum_{i=1}^{n_1}\hat{p}_i\{Y_{1i} - m_1^1(X_{1i},\hat{\gamma}_1^1)\} &= \frac{n}{n_1}\frac{1}{n}\sum_{i=1}^{n}\frac{T_i\{Y_i - m_1^1(X_i,\hat{\gamma}_1^1)\}/\pi_1(X_i,\beta_1)}{1+\hat{\lambda}_1^T g_1(X_i,\hat{\beta},\hat{\gamma}^1)/\pi_1(X_i,\beta_1)}\\ &\xrightarrow{p}\frac{1}{\Pr(T=1)}E\left[\frac{T\{Y - m_1^1(X,\gamma_1^1)\}}{1+\lambda_{1*}^T g_1(X,\beta_*,\gamma_*^1)}\right] = 0.\end{aligned}$$

Similarly, we can prove that

$$\sum_{j=1}^{n_0}\hat{q}_j\left\{Y_{0j} - m_1^0\left(X_{0j},\hat{\gamma}_1^0\right)\right\}\xrightarrow{p}0.$$

Thus, we have $\Delta_{mr}\xrightarrow{p}\Delta_0$.

Proof of Corollary 1: According (3) and (4), it can be verified that

$$\begin{aligned}\sum_{i=1}^{n_1}\hat{p}_i m_k^1(X_{1i},\hat{\gamma}_k^1) &= n^{-1}\sum_{i=1}^{n} m_k^1(X_i,\hat{\gamma}_k^1),\\ \sum_{j=1}^{n_0}\hat{q}_j m_k^0\left(X_{0j},\hat{\gamma}_k^0\right) &= n^{-1}\sum_{j=1}^{n} m_k^0\left(X_j,\hat{\gamma}_k^0\right),\end{aligned}$$

which leads to

$$\begin{aligned}\sum_{i=1}^{n_1}\hat{p}_i m^1(X_{1i}) &= \sum_{i=1}^{n_1}\hat{p}_i\left\{c_{10} + \sum_{k=1}^{K}c_{1k}m_k^t(X_{1i},\gamma_k^1)\right\} = n^{-1}\sum_{i=1}^{n}m^1(X_i)\xrightarrow{p}\mu_1,\\ \sum_{j=1}^{n_0}\hat{q}_j m^0(X_{0j}) &= \sum_{j=1}^{n_0}\hat{q}_j\left\{c_{00} + \sum_{k=1}^{K}c_{0k}m_k^t(X_{1i},\gamma_k^1)\right\} = n^{-1}\sum_{j=1}^{n}m^0(X_i)\xrightarrow{p}\mu_0.\end{aligned}$$

Thus,

$$\Delta_{mr} = \sum_{i=1}^{n_1}\hat{p}_i\{Y_{1i} - m^1(X_{1i})\} - \sum_{j=1}^{n_0}\hat{q}_j\left\{Y_{0j} - m^0(X_{0j})\right\} + \mu_1 - \mu_0 + o_p(1).$$

It can be shown that

$$\begin{aligned}\sum_{i=1}^{n_1}\hat{p}_i\{Y_{1i} - m^1(X_{1i})\} &= \frac{n}{n_1}\frac{1}{n}\sum_{i=1}^{n}\frac{T_i\{Y_i - m^1(X_i)\}/\pi_1(X_i,\beta_1)}{1+\hat{\lambda}_1^T g_1(X_i,\hat{\beta},\hat{\gamma}^1)/\pi_1(X_i,\beta_1)}\\ &\xrightarrow{p}\frac{1}{\Pr(T=1)}E\left[\frac{T\{Y - m^1(X)\}}{1+\lambda_{1*}^T g_1(X,\beta_*,\gamma_*^1)}\right] = 0.\end{aligned}$$

In the similar manner, we can prove that $\sum_{j=1}^{n_0}\hat{q}_j\{Y_{0j} - m^0(X_{0j})\}\xrightarrow{p}0$. Thus, we have $\Delta_{mr}\xrightarrow{p}\Delta_0$.

Proof of Theorem 2: When $\pi_1(X;\beta_1)$ is a correctly specified model for $\pi(X)$, as in Han and Wang (2013),

$$
\begin{aligned}
n^{1/2}\hat{\lambda}_1 &= G_1^{-1}\left[n^{-1/2}\sum_{i=1}^{n}\frac{T_i-\pi(X_i)}{\pi(X_i)}g_1\left(X_i,\beta_*,\gamma_*^1\right)-D_1 n^{1/2}\left(\hat{\beta}_1-\beta_1^0\right)\right]+o_p(1),\\
n^{1/2}\hat{\lambda}_0 &= G_0^{-1}\left[n^{-1/2}\sum_{j=1}^{n}\frac{\pi(X_j)-T_j}{1\;-\;\pi(X_j)}g_0\left(X_j,\beta_*,\gamma_*^0\right)-D_0 n^{1/2}\left(\hat{\beta}_1-\beta_1^0\right)\right]+o_p(1),
\end{aligned}
$$

where

$$
\begin{aligned}
D_1 &= E\left[\left\{g_1\left(X,\beta_*,\gamma_*^1\right)/\pi(X)\right\}\left\{\partial\pi_1\left(X,\beta_1^0\right)/\partial\beta_1\right\}^T\right],\\
D_0 &= -E\left[\left\{g_0\left(X,\beta_*,\gamma_*^0\right)/\left\{1-\pi(X)\right\}\right\}\left\{\partial\pi_1\left(X,\beta_1^0\right)/\partial\beta_1\right\}^T\right].
\end{aligned}
$$

Taylor expansion of $\sqrt{n}(\hat{\Delta}_{mr}-\Delta_0)$ about $(\hat{\lambda}_1,\hat{\lambda}_0,\hat{\beta}_1,\hat{\gamma}_1^1,\hat{\gamma}_1^0)$ at $(0,0,\beta_{1*},\gamma_{1*}^1,\gamma_{1*}^0)^T$, we have

$$
\begin{aligned}
&\sqrt{n}\left(\hat{\Delta}_{mr}-\Delta_0\right)\\
&= n^{1/2}\sum_{i=1}^{n}T_i\hat{p}_i(Y_i-\mu_1)-n^{1/2}\sum_{j=1}^{n}(1\;-\;T_j)\hat{q}_j(Y_j-\mu_0)\\
&= n^{1/2}\sum_{i=1}^{n}\frac{T_i(Y_i-\mu_1)/\pi_1\left(X_{1i},\hat{\beta}_1\right)}{1+\hat{\lambda}_1^T g_1\left(X_{1i},\hat{\beta},\hat{\gamma}^1\right)/\pi_1\left(X_{1i},\hat{\beta}_1\right)}-n^{1/2}\sum_{j=1}^{n}\frac{(1-T_j)(Y_j-\mu_0)/\left\{1-\pi_1\left(X_{0j},\hat{\beta}_1\right)\right\}}{1+\hat{\lambda}_0^T g_0\left(X_{0j},\hat{\beta},\hat{\gamma}^0\right)/\left\{1\;-\;\pi_1\left(X_{0j},\hat{\beta}_1\right)\right\}}\\
&= n^{-1/2}\sum_{i=1}^{n}\left[\frac{T_i(Y_i-\mu_1)}{\pi(X_{1i})}-\frac{T_i-\pi(X_{1i})}{\pi(X_{1i})}R_1(X_{1i})\right]\\
&\quad -n^{-1/2}\sum_{i=1}^{n}\left[\frac{(1-T_i)(Y_i-\mu_0)}{1\;-\;\pi(X_{1i})}-\frac{\pi(X_{1i})-T_i}{1\;-\;\pi(X_{1i})}R_0(X_{1i})\right]\\
&\quad -E\left[\left\{\frac{Y_1-\mu_1-Q_1(X)}{\pi(X)}+\frac{Y_0-\mu_0-Q_0(X)}{1\;-\;\pi(X)}\right\}\frac{\partial\pi_1\left(X,\beta_1^0\right)}{\partial\beta_1}\right]n^{1/2}\left(\hat{\beta}_1-\beta_1^0\right)+o_p(1).
\end{aligned}
$$

According to Lemma 9.1 in Tsiatis (2006), it can be verified that

$$
\begin{aligned}
&E\left[\left\{\frac{Y_1-\mu_1-Q_1(X)}{\pi(X)}+\frac{Y_0-\mu_0-Q_0(X)}{1\;-\;\pi(X)}\right\}\frac{\partial\pi_1\left(X,\beta_1^0\right)}{\partial\beta_1}\right]\\
&= -E\left[\partial\left\{\frac{T(Y-\mu_1)}{\pi(X)}-\frac{T-\pi(X)}{\pi(X)}R_1(X)-\frac{(1-T)(Y-\mu_0)}{1\;-\;\pi(X)}+\frac{\pi(X)-T}{1\;-\;\pi(X)}R_0(X)\right\}/\partial\beta_1\right]\\
&= E\left(US^T\right).
\end{aligned}
$$

Therefore, by noting that

$$
n^{1/2}\left(\hat{\beta}_1-\beta_1^0\right)=n^{1/2}E\left(S^{\otimes 2}\right)^{-1}\sum_{i=1}^{n}S\left(X_i,T_i,\beta_1^0\right),
$$

the desired result follows.

Proof of Theorem 3: Let $H_1=T(Y-\mu_1)/\pi(X)$, $Q_1=Tg_1(X,\beta_*,\gamma_*^1)/\pi(X)$, $H_0=(1-T)(Y-\mu_0)/\{1-\pi(X)\}$, $Q_0=(1-T)g_0(X,\beta_*,\gamma_*^0)/\{1-\pi(X)\}$. It is clear that

$$
L_1^T G_1^{-1}=E[H_1 Q_1]E^{-1}\left[Q_1^{\otimes 2}\right],\qquad L_0^T G_0^{-1}=E[H_0 Q_0]E^{-1}\left[Q_0^{\otimes 2}\right].
$$

Note that

$$E\begin{pmatrix} \left[H_1 - T\{E(Y_1-\mu_1|X)\}/\pi(X)\right]Q_1 \\ \left[H_0 - (1-T)\{E(Y_0-\mu_0|X)\}/\{1-\pi(X)\}\right]Q_0 \end{pmatrix} = 0.$$

When $\mathcal{M}$ contains a correctly specified model for $m(x)$, $m^1(X)-\mu_1$ and $m^0(X)-\mu_0$ are the components of $g_1(X,\beta_*,\gamma_*^1)$ and $g_0(X,\beta_*,\gamma_*^0)$, respectively. Thus, we have that $T\{m^1(X-\mu_1\}/\pi(X)$ and $(1-T)\{m^0(X-\mu_0\}/\{1-\pi(X)\}$ are in the linear space spanned by Q_1 and Q_0. This fact yields

$$\begin{aligned} L_1^T G_1^{-1} g_1(X,\beta_*,\gamma_*^1) &= E[H_1Q_1]E^{-1}\left[Q_1^{\otimes 2}\right]g_1(X,\beta_*,\gamma_*^1) = m^1(X)-\mu_1, \\ L_0^T G_0^{-1} g_0(X,\beta_*,\gamma_*^0) &= E[H_0Q_0]E^{-1}\left[Q_0^{\otimes 2}\right]g_0(X,\beta_*,\gamma_*^0) = m^0(X)-\mu_0. \end{aligned}$$

In this case, we have

$$\begin{aligned} U_1^{opt} &= \frac{T(Y-\mu_1)}{\pi(X)} - \frac{T-\pi(X)}{\pi(X)}\{m^1(X)-\mu_1\}, \\ U_0^{opt} &= \frac{(1-T)(Y-\mu_0)}{1-\pi(X)} + \frac{T-\pi(X)}{1-\pi(X)}\{m^0(X)-\mu_0\}. \end{aligned}$$

A simple calculation shows that $E(U^{opt}S^T)=0$ and it is easy to show

$$\mathrm{Var}(U^{opt}) = \mathrm{Var}\{m^1(X)-m^0(X)\} + E\left\{\frac{\mathrm{Var}(Y_1|X)}{\pi(X)}\right\} + E\left\{\frac{\mathrm{Var}(Y_0|X)}{1-\pi(X)}\right\}.$$

Proof of Corollary 2: The proof is similar to Corollary 1.