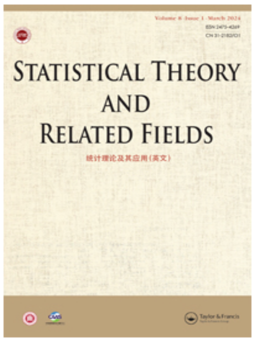

Statistical Theory and Related Fields

ISSN: (Print) (Online) Journal homepage: www.tandfonline.com/journals/tstf20

Multiply robust estimation for average treatment effect among treated

Lu Wang & Peisong Han

STATISTICAL THEORY AND RELATED FIELDS
2024, VOL. 8, NO. 1, 29–39
https://doi.org/10.1080/24754269.2023.2293554

OPEN ACCESS

Multiply robust estimation for average treatment effect among treated

Lu Wang[a] and Peisong Han[b]

[a]Department of Biostatistics, University of Michigan, Ann Arbor, MI, USA; [b]Biostatistics Innovation Group, Gilead Sciences, Foster City, CA, USA

ABSTRACT
We propose a multiply robust estimator for the Average Treatment Effect Among the Treated (ATT). The proposed estimation procedure can simultaneously accommodate multiple working models for both the propensity score and the conditional mean of the counterfactual outcome given covariates. In addition, it can explicitly balance a set of user-specified moments of the covariate distributions between the treatment groups. The resulting estimator is consistent if any working model is correctly specified. With the data generating process typically unknown for observational studies, the proposed method provides substantial robustness against possible model misspecifications compared to existing estimators of the ATT. Simulation results show the excellent finite sample performance of the proposed estimator.

ARTICLE HISTORY
Received 6 March 2023
Revised 28 September 2023
Accepted 10 November 2023

KEYWORDS
Causal inference; covariate balancing; empirical likelihood; multiple robustness; propensity score

1. Introduction

In causal inference, since we only observe what happens to an individual under the treatment condition they actually receive, it is generally impossible to estimate the causal effects for individuals. The causal effects one typically considers involve summary statistics of the individual effects across populations or sub-populations of interest. Two widely considered causal summaries are the average treatment effect (ATE) and the average treatment effect among the treated (ATT). The ATE of a treatment relative to a control is the comparison of the mean outcome which had the entire population been treated versus had the entire population been the control. The ATT is the comparison of the mean outcome under treatment among those who are treated with the mean outcome which had the treated subjects received control instead. A study can estimate both ATE and ATT, but one or the other may be better suited for a particular situation. The ATE may be of more interest if each treatment can potentially be offered to every member of the population. Conversely, if the research question focuses on the effectiveness of an alternative treatment were it to replace the standard treatment, and then the ATT may be of more interest because it measures the relative effectiveness of the two treatment options on the population that is receiving the standard treatment.

There has been a large literature on estimating the ATE and the ATT for observational studies, where a typical consideration is confounding in the sense that individual characteristics are related to both the treatment assignment and the outcome of interest. Propensity score (PS) based methods, where the PS is the probability of receiving a treatment given covariates (Rosenbaum & Rubin, 1983), are commonly used to deal with confounding and to achieve balance of covariate distributions between different treatment groups. Such methods include PS matching (e.g., Abadie & Imbens, 2006; Rosenbaum & Rubin, 1985) and weighting (e.g., Hirano et al., 2003). Refer to Imbens and Rubin (2015) and Hernán and Robins (2018) for more details. Parametric modelling of the PS is common, especially when the dimension of covariates is moderate to large. Misspecification of the PS model is usually a major concern as it may lead to substantial estimation bias.

To mitigate the impact of misspecification of the PS model, substantial interests have been given to doubly robust estimators, which involve models for both the PS and the conditional mean of the counterfactual outcome, or outcome regression (OR), and remain consistent if either model is correctly specified. The original doubly robust estimator was constructed through augmented inverse probability weighting (AIPW) in the missing data context (Robins et al., 1994). Since then, a large number of doubly robust estimators have been proposed in both missing data and causal inference settings (e.g., Bang & Robins, 2005; Cao et al., 2009; Han, 2012; Kang & Schafer, 2007; Qin et al., 2008; Qin & Zhang, 2007; Rotnitzky et al., 2012; Tan, 2010; van der Laan & Gruber, 2010). When both the PS model and the OR model are wrong, these estimators are in general no longer consistent.

CONTACT Peisong Han ✉ peisong.han2@gilead.com, peisong@umich.edu Biostatistics Innovation Group, Gilead Sciences, Foster City, CA 94404, USA

As an improvement over double robustness, in the missing data context, Han and Wang (2013) proposed a multiply robust estimation procedure that can simultaneously accommodate multiple working models for both PS and OR, and the resulting estimators are consistent if any of these models is correctly specified. Since the data generating process for observational studies is typically unknown, it is common that several candidate models all seem reasonable yet none rules out the possibility of others, especially when the dimension of covariates is moderate to large. In such a case, the multiply robust method in Han and Wang (2013) provides a useful tool for data analysis with more protection on estimation consistency. Such a method has attracted considerable interest in both missing data and causal inference research (e.g., Chan & Yam, 2014; Chan et al., 2016; S. Chen & Haziza, 2017; Duan & Yin, 2017; Han, 2014a, 2014b, 2016a, 2016b; Han et al., 2019; Li et al., 2020). Especially, Wang (2019) extended the multiply robust method to the estimation of the ATE. It is worth pointing out that the term " multiply robust" has also been used by other authors in different settings with different meanings. For example, in Molina et al. (2017), it refers to estimation consistency when various combinations of the components of a factorized likelihood are correctly modelled, while in Wang and Tchetgen Tchetgen (2018) and Shi et al. (2020) it refers to estimation consistency being achieved across the union of three different observed data models. We use " multiply robust" to refer to the property that estimation consistency is achieved if one of the multiple working models for the same quantity is correctly specified.

Despite the success in dealing with missing data problems and in estimating the ATE, multiply robust estimators have not been developed for estimating the ATT. In this paper, we construct such a desirable estimator for ATT. In addition to being multiply robust, the proposed estimator can easily achieve certain level of balancing of covariate distributions between treatment groups. Covariate balancing is a highly desired property for causal effect estimation. For the estimation of ATT, Hainmueller (2012) proposed the entropy balancing (EB) method by imposing a set of balance constraints so that certain moments of covariate distributions for different treatment groups match exactly. Zhao and Percival (2017) found that EB implicitly fits a logistic linear regression model for the PS and a linear regression model for the OR, and when either model is correctly specified the EB estimator is consistent. The estimator we propose preserves the same balance of covariate distributions as EB, and in addition it accounts for multiple models for PS and OR simultaneously so that consistency of the resulting estimator is guaranteed if any one model is correctly specified.

The rest of the paper is organized as follows. Section 2 gives the setup and a brief review of some existing methods. Section 3 presents our proposed multiply robust estimator for the ATT. Simulation studies are provided in Section 4, followed by some discussion in Section 5.

2. Setup and some existing methods

Let A denote the treatment indicator, where $A = 1$ for the treatment of interest and $A = 0$ for control. Let Y^1 and Y^0 denote the counterfactual outcomes under treatment and under control, respectively. Let Y denote the observed outcome in the data, for which we make the consistency assumption that $Y = AY^1 + (1 - A)Y^0$. Let $\boldsymbol{X}$ denote the pre-treatment covariates in the dataset, including potential confounders. The potential full data is $(A, Y^1, Y^0, \boldsymbol{X})$ whereas the observed data is $(A, Y, \boldsymbol{X})$. The causal effect we are interested in is the ATT $\tau = E(Y^1 - Y^0 \mid A = 1)$, which is also $\tau_{1,1} - \tau_{1,0}$, where $\tau_{a,b}$ is the mean outcome for subjects who receive treatment a had they instead received treatment b, i.e., $\tau_{a,b} = E[Y^b|A = a]$.

To ensure the identifiability of τ from the observed data, we make some assumptions following the main literature (e.g., Rosenbaum & Rubin, 1983). First, we assume that all potential confounders are included in $\boldsymbol{X}$, or in other words, the treatment assignment process and the counterfactual outcome are independent given all the covariates measured. Second, we assume that every subject has a positive probability of being assigned to either the treatment or the control group. These two assumptions are formalized as follows.

Assumption 2.1 (no unmeasured confounders assumption): $A \perp (Y^0, Y^1) \mid \boldsymbol{X}$.

Assumption 2.2 (strict positivity): $0 < \sigma_1 < P(A = 1 \mid \boldsymbol{X}) < \sigma_2 < 1$ *with probability one for some positive constants* σ_1 *and* σ_2.

Let $\pi(\boldsymbol{X}) = P(A = 1 \mid \boldsymbol{X})$ denote the PS for treatment assignment. In the following we give a brief review of some widely used estimators of the ATT. Note that $\tau_{1,1} = E(Y^1 \mid A = 1)$ can be consistently estimated by the sample average $\hat{\tau}_{1,1} = (\sum_{i=1}^n A_i Y_i)/(\sum_{i=1}^n A_i)$ over the treatment group. Therefore, the estimation of the ATT reduces to the estimation of $\tau_{1,0} = E(Y^0 \mid A = 1)$, where the counterfactual outcome Y^0 is not observable for individuals in the $A = 1$ group.

One straightforward way to estimate $\tau_{1,0}$ is to impute Y^0 for those individuals in the $A = 1$ group. Suppose $d_0(\boldsymbol{\gamma})$ is a parametric regression model for $E(Y^0 \mid \boldsymbol{X})$ that is parametrized by $\boldsymbol{\gamma}$. Because $E(Y^0 \mid \boldsymbol{X}) = E(Y^0 \mid \boldsymbol{X}, A = 0)$

from Assumption 1 and Y^0 is fully observed in the $A = 0$ group, $\boldsymbol{\gamma}$ can be estimated by $\hat{\boldsymbol{\gamma}}$ based on individuals in the control group alone. Also from Assumption 1, we have $E(Y^0 \mid \boldsymbol{X}) = E(Y^0 \mid \boldsymbol{X}, A = 1)$, and thus the unobserved Y^0 in the treatment group can be imputed by $d_0(\hat{\boldsymbol{\gamma}})$. Therefore, an outcome regression estimator of $\tau_{1,0}$ is $\hat{\tau}_{1,0\mathrm{reg}} = \sum_{i=1}^n A_i d_{0i}(\hat{\boldsymbol{\gamma}})/(\sum_{i=1}^n A_i)$.

A widely used alternative method is inverse probability weighting (IPW). From Assumption 1, since

$$f(Y^1, Y^0, \boldsymbol{X} \mid A = 1) = \frac{\pi(\boldsymbol{X}) f(Y^1, Y^0, \boldsymbol{X})}{P(A = 1)} \quad \text{and} \quad f(Y^1, Y^0, \boldsymbol{X} \mid A = 0) = \frac{\{1 - \pi(\boldsymbol{X})\} f(Y^1, Y^0, \boldsymbol{X})}{P(A = 0)},$$

where $f(\cdot)$ is a generic notation for a probability density function, we have

$$f(Y^1, Y^0, \boldsymbol{X} \mid A = 1) = \frac{\pi(\boldsymbol{X})}{1 - \pi(\boldsymbol{X})} \frac{P(A = 0)}{P(A = 1)} f(Y^1, Y^0, \boldsymbol{X} \mid A = 0). \tag{1}$$

Therefore, an estimator of $\tau_{1,0} = E(Y^0 \mid A = 1)$ can be constructed by properly weighting the observed Y^0 in the $A = 0$ group, and this leads to an IPW estimator

$$\hat{\tau}_{1,0\mathrm{ipw}} = \frac{1}{\sum_{s=1}^n (1 - A_s)} \sum_{\{i: A_i = 0\}} \frac{\hat{\pi}(\boldsymbol{X}_i)}{1 - \hat{\pi}(\boldsymbol{X}_i)} \frac{\sum_{s=1}^n (1 - A_s)/n}{\sum_{s=1}^n A_s/n} Y_i = \frac{1}{\sum_{s=1}^n A_s} \sum_{\{i: A_i = 0\}} \frac{\hat{\pi}(\boldsymbol{X}_i)}{1 - \hat{\pi}(\boldsymbol{X}_i)} Y_i, \tag{2}$$

where $\hat{\pi}(\boldsymbol{X})$ is the estimated value of $\pi(\boldsymbol{X})$.

Consistency of $\hat{\tau}_{1,0\mathrm{reg}}$ and $\hat{\tau}_{1,0\mathrm{ipw}}$ requires correct modelling of $E(Y^0 \mid \boldsymbol{X})$ and $\pi(\boldsymbol{X})$, respectively. To improve the robustness against possible model misspecifications, the AIPW method combines the models for $E(Y^0 \mid \boldsymbol{X})$ and $\pi(\boldsymbol{X})$ so that estimation consistency is guaranteed if either model is correctly specified but not necessarily both. The AIPW estimator for $\tau_{1,0}$ is given as

$$\hat{\tau}_{1,0\mathrm{aipw}} = \hat{\tau}_{1,0\mathrm{reg}} + \frac{1}{\sum_{s=1}^n A_s} \sum_{\{i: A_i = 0\}} \frac{\hat{\pi}(\boldsymbol{X}_i)}{1 - \hat{\pi}(\boldsymbol{X}_i)} \{Y_i - d_{0i}(\hat{\boldsymbol{\gamma}})\}.$$

To achieve certain covariate balancing, the EB method (Hainmueller, 2012) considers matching covariate moments between the treatment and the control groups. Specifically, let w_i be a set of positive weights assigned to the control subjects $\{i : A_i = 0\}$. The EB method imposes covariate balancing constraints

$$w_i > 0, \quad \sum_{\{i: A_i = 0\}} w_i = 1, \quad \sum_{\{i: A_i = 0\}} w_i \boldsymbol{h}(\boldsymbol{X}_i) = \frac{1}{\sum_{s=1}^n A_s} \sum_{\{i: A_i = 1\}} \boldsymbol{h}(\boldsymbol{X}_i) \tag{3}$$

on w_i, where $\boldsymbol{h}(\boldsymbol{X})$ contains some user-specified moments of $\boldsymbol{X}$. These constraints ensure that certain moments of $\boldsymbol{X}$ exactly match between the two groups. The EB estimator of $\tau_{1,0}$ is $\hat{\tau}_{1,0\mathrm{eb}} = \sum_{\{i: A_i = 0\}} \hat{w}_{\mathrm{eb},i} Y_i$ where $\hat{w}_{\mathrm{eb},i}$ minimizes $\sum_{\{i: A_i = 0\}} w_i \log w_i$ subject to the constraints in (3). Although no parametric models are explicitly fitted by the EB method, Zhao and Percival (2017) showed that, implicitly, the EB method fits linear regression models for logit$\{\pi(\boldsymbol{X})\}$ and $E(Y^0 \mid \boldsymbol{X})$ with components of $\boldsymbol{h}(\boldsymbol{X})$ as regressors. The EB estimator $\hat{\tau}_{1,0\mathrm{eb}}$ is consistent if either model is correctly specified, and thus is doubly robust.

3. The proposed multiply robust estimator for ATT

Our goal is to construct an easy-to-implement estimator of the ATT that is multiply robust and achieves explicit covariate balancing as the EB estimator. Let $\mathcal{P} = \{\pi^{(j)}(\boldsymbol{\alpha}^{(j)}) : j = 1, \ldots, J\}$ denote a set of multiple parametric models for $\pi(\boldsymbol{X})$ and $\mathcal{D}_0 = \{d_0^{(k)}(\boldsymbol{\gamma}^{(k)}) : k = 1, \ldots, K\}$ a set of multiple parametric models for $E(Y^0 \mid \boldsymbol{X})$, where $\boldsymbol{\alpha}^{(j)}$ and $\boldsymbol{\gamma}^{(k)}$ are the corresponding parameters. $\boldsymbol{\alpha}^{(j)}$ is typically estimated by $\hat{\boldsymbol{\alpha}}^{(j)}$ that maximizes the binomial likelihood

$$\prod_{i=1}^n \{\pi_i^{(j)}(\boldsymbol{\alpha}^{(j)})\}^{A_i} \{1 - \pi_i^{(j)}(\boldsymbol{\alpha}^{(j)})\}^{1 - A_i}, \tag{4}$$

and $\boldsymbol{\gamma}^{(k)}$ is typically estimated by $\hat{\boldsymbol{\gamma}}^{(k)}$ based on individuals in the control group because $E(Y^0 \mid \boldsymbol{X}) = E(Y^0 \mid \boldsymbol{X}, A = 0)$ from Assumption 1 and Y^0 is fully observed in the control group.

From (1) it is easy to see that, for any function $b(\boldsymbol{X})$, we have

$$E\{b(\boldsymbol{X}) \mid A=1\} = E\left\{\frac{\pi(\boldsymbol{X})}{1-\pi(\boldsymbol{X})} b(\boldsymbol{X}) \mid A=0\right\} \frac{P(A=0)}{P(A=1)}. \tag{5}$$

In particular, taking $b(\boldsymbol{X}) \equiv 1$ gives

$$1 = E\left\{\frac{\pi(\boldsymbol{X})}{1-\pi(\boldsymbol{X})} \mid A=0\right\} \frac{P(A=0)}{P(A=1)},$$

and thus

$$\begin{aligned} E\{b(\boldsymbol{X}) \mid A=1\} &= E\{b(\boldsymbol{X}) \mid A=1\} E\left\{\frac{\pi(\boldsymbol{X})}{1-\pi(\boldsymbol{X})} \mid A=0\right\} \frac{P(A=0)}{P(A=1)} \\ &= E\left\{\frac{\pi(\boldsymbol{X})}{1-\pi(\boldsymbol{X})} E\{b(\boldsymbol{X}) \mid A=1\} \mid A=0\right\} \frac{P(A=0)}{P(A=1)}. \end{aligned} \tag{6}$$

Subtracting (6) from (5) leads to

$$E\left\{\frac{\pi(\boldsymbol{X})}{1-\pi(\boldsymbol{X})} [b(\boldsymbol{X}) - E\{b(\boldsymbol{X}) \mid A=1\}] \mid A=0\right\} = 0. \tag{7}$$

Our method starts by constructing an empirical version of (7). Let w_i be a set of positive weights assigned on the control subjects with $A_i = 0$ such that $\sum_{\{i:A_i=0\}} w_i = 1$, and then an empirical version of (7) is

$$\sum_{\{i:A_i=0\}} w_i \left\{ b(\boldsymbol{X}_i) - \frac{1}{\sum_{s=1}^n A_s} \sum_{s=1}^n A_s b(\boldsymbol{X}_s) \right\} = 0.$$

To achieve multiple robustness so that $\tau_{1,0}$ is consistently estimated when any one model is correctly specified, we take $b(\boldsymbol{X})$ to be $1/\pi^{(j)}(\boldsymbol{\alpha}^{(j)}), j = 1, \ldots, J$, and $d_0^{(k)}(\boldsymbol{\gamma}^{(k)}), k = 1, \ldots, K$, and consider the constraints on the w_i as

$$w_i > 0, \quad \sum_{\{i:A_i=0\}} w_i = 1, \quad \sum_{\{i:A_i=0\}} w_i \hat{\boldsymbol{g}}_i(\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\gamma}}) = \boldsymbol{0}, \tag{8}$$

where $\hat{\boldsymbol{\alpha}} = (\hat{\boldsymbol{\alpha}}^{(1)}, \ldots, \hat{\boldsymbol{\alpha}}^{(J)})$, $\hat{\boldsymbol{\gamma}} = (\hat{\boldsymbol{\gamma}}^{(1)}, \ldots, \hat{\boldsymbol{\gamma}}^{(J)})$ and

$$\hat{\boldsymbol{g}}(\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\gamma}}) = \begin{pmatrix} \dfrac{1}{\pi^{(1)}(\hat{\boldsymbol{\alpha}}^{(1)})} - \dfrac{1}{\sum_{s=1}^n A_s} \sum_{s=1}^n A_s \dfrac{1}{\pi_s^{(1)}(\hat{\boldsymbol{\alpha}}^{(1)})} \\ \vdots \\ \dfrac{1}{\pi^{(J)}(\hat{\boldsymbol{\alpha}}^{(J)})} - \dfrac{1}{\sum_{s=1}^n A_s} \sum_{s=1}^n A_s \dfrac{1}{\pi_s^{(J)}(\hat{\boldsymbol{\alpha}}^{(J)})} \\ d_0^{(1)}(\hat{\boldsymbol{\gamma}}^{(1)}) - \dfrac{1}{\sum_{s=1}^n A_s} \sum_{s=1}^n A_s d_{0s}^{(1)}(\hat{\boldsymbol{\gamma}}^{(1)}) \\ \vdots \\ d_0^{(K)}(\hat{\boldsymbol{\gamma}}^{(K)}) - \dfrac{1}{\sum_{s=1}^n A_s} \sum_{s=1}^n A_s d_{0s}^{(K)}(\hat{\boldsymbol{\gamma}}^{(K)}) \\ \boldsymbol{h}(\boldsymbol{X}) - \dfrac{1}{\sum_{s=1}^n A_s} \sum_{s=1}^n A_s \boldsymbol{h}(\boldsymbol{X}_s) \end{pmatrix}.$$

Note that here $\hat{\boldsymbol{g}}(\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\gamma}})$ can also contain components based on taking $b(\boldsymbol{X})$ to be $\boldsymbol{h}(\boldsymbol{X})$, a vector of user-specified moments of $\boldsymbol{X}$. This can be used to achieve certain degree of covariate balancing between the treatment and control groups, similar to the EB method.

Subject to the constraints in (8), we consider the weights $\hat{w}_{\mathrm{mr},i}$ on the subjects in the control group $\{i : A_i = 0\}$ that maximize the empirical likelihood $\prod_{\{i:A_i=0\}} w_i$ and propose to estimate $\tau_{1,0}$ by $\hat{\tau}_{1,0\mathrm{mr}} = \sum_{\{i:A_i=0\}} \hat{w}_{\mathrm{mr},i} Y_i$.

Following the standard empirical likelihood technique (e.g., Qin & Lawless, 1994), we have

$$\hat{w}_{\mathrm{mr},i} = \frac{1}{\sum_{i=1}^{n}(1-A_i)} \frac{1}{1+\hat{\boldsymbol{\rho}}^{\top}\hat{\boldsymbol{g}}_i(\hat{\boldsymbol{\alpha}},\hat{\boldsymbol{\gamma}})}, \text{for } i \text{ satisfying } A_i = 0,$$

where $\hat{\boldsymbol{\rho}}$ solves

$$\sum_{\{i:A_i=0\}} \frac{\hat{\boldsymbol{g}}_i(\hat{\boldsymbol{\alpha}},\hat{\boldsymbol{\gamma}})}{1+\boldsymbol{\rho}^{\top}\hat{\boldsymbol{g}}_i(\hat{\boldsymbol{\alpha}},\hat{\boldsymbol{\gamma}})} = \mathbf{0}. \tag{9}$$

For implementation, directly solving (9) for $\hat{\boldsymbol{\rho}}$ is not the ideal way because, as pointed out in Han (2014a), (9) typically has multiple roots but only one of them makes the $\hat{w}_{\mathrm{mr},i}$ positive. We recommend calculating $\hat{\boldsymbol{\rho}}$ by minimizing $F(\boldsymbol{\rho}) \equiv -\sum_{\{i:A_i=0\}} \log\{1+\boldsymbol{\rho}^{\top}\hat{\boldsymbol{g}}_i(\hat{\boldsymbol{\alpha}},\hat{\boldsymbol{\gamma}})\}$, which is the negative antiderivative of the left-hand side of (9). As shown in Han (2014a), this is a convex minimization that always has a unique minimizer when $\mathbf{0}$ is inside the convex hull of $\{\hat{\boldsymbol{g}}_i(\hat{\boldsymbol{\alpha}},\hat{\boldsymbol{\gamma}}): \quad A_i = 0\}$, which indeed holds at least when n is large because of the moment equality (7) and (8) being an empirical version of (7). The unique minimizer of $F(\boldsymbol{\rho})$ is $\hat{\boldsymbol{\rho}}$ needed for calculating $\hat{w}_{\mathrm{mr},i}$ and can be easily found by a Newton-Raphson-type algorithm. Refer to J. Chen et al. (2002) and Han (2014a) for such an algorithm.

In the following we provide arguments and explanations for the multiple robustness of $\hat{\tau}_{1,0\mathrm{mr}}$. We do not include detailed technical conditions and mathematical proofs for these arguments so that the main ideas are easier to follow. To see the consistency of $\hat{\tau}_{1,0\mathrm{mr}}$ when $\mathcal{P}$ contains a correctly specified model for $\pi(\boldsymbol{X})$, say $\pi^{(1)}(\boldsymbol{\alpha}^{(1)})$ without loss of generality, let

$$\hat{\vartheta} = \frac{1}{\sum_{s=1}^{n} A_s} \sum_{i=1}^{n} A_i \frac{1}{\pi_i^{(1)}(\hat{\boldsymbol{\alpha}}^{(1)})}$$

and define $\hat{\boldsymbol{\lambda}}$ in such a way that $\hat{\rho}_1 = (\hat{\lambda}_1+1)/(\hat{\vartheta}-1)$ and $\hat{\boldsymbol{\rho}}_{-1} = \hat{\boldsymbol{\lambda}}_{-1}/(\hat{\vartheta}-1)$, where $\hat{\rho}_1$ and $\hat{\lambda}_1$ are the first component of $\hat{\boldsymbol{\rho}}$ and $\hat{\boldsymbol{\lambda}}$, respectively, and $\hat{\boldsymbol{\rho}}_{-1}$ and $\hat{\boldsymbol{\lambda}}_{-1}$ are vectors of the rest components. Then some simple algebra shows that

$$\hat{w}_{\mathrm{mr},i} = \frac{\hat{\vartheta}-1}{\sum_{s=1}^{n}(1-A_s)} \frac{\pi_i^{(1)}(\hat{\boldsymbol{\alpha}}^{(1)})}{1-\pi_i^{(1)}(\hat{\boldsymbol{\alpha}}^{(1)})} \frac{1}{1+\hat{\boldsymbol{\lambda}}^{\top}\frac{\pi_i^{(1)}(\hat{\boldsymbol{\alpha}}^{(1)})}{1-\pi_i^{(1)}(\hat{\boldsymbol{\alpha}}^{(1)})}\hat{\boldsymbol{g}}_i(\hat{\boldsymbol{\alpha}},\hat{\boldsymbol{\gamma}})}, \quad \text{for } i \text{ satisfying } A_i = 0$$

and (9) becomes an equation for $\hat{\boldsymbol{\lambda}}$

$$\sum_{\{i:A_i=0\}} \frac{\frac{\pi_i^{(1)}(\hat{\boldsymbol{\alpha}}^{(1)})}{1-\pi_i^{(1)}(\hat{\boldsymbol{\alpha}}^{(1)})}\hat{\boldsymbol{g}}_i(\hat{\boldsymbol{\alpha}},\hat{\boldsymbol{\gamma}})}{1+\hat{\boldsymbol{\lambda}}^{\top}\frac{\pi_i^{(1)}(\hat{\boldsymbol{\alpha}}^{(1)})}{1-\pi_i^{(1)}(\hat{\boldsymbol{\alpha}}^{(1)})}\hat{\boldsymbol{g}}_i(\hat{\boldsymbol{\alpha}},\hat{\boldsymbol{\gamma}})} = \mathbf{0}.$$

From the Z-estimator theory (e.g., van der Vaart, 1998) and because of the moment equality (7), we must have $\hat{\boldsymbol{\lambda}} = O_p(n^{-1/2})$, which leads to

$$\hat{w}_{\mathrm{mr},i} = \frac{\hat{\vartheta}-1}{\sum_{s=1}^{n}(1-A_s)} \frac{\pi_i^{(1)}(\hat{\boldsymbol{\alpha}}^{(1)})}{1-\pi_i^{(1)}(\hat{\boldsymbol{\alpha}}^{(1)})}\{1+O_p(n^{-1/2})\}, \quad \text{for } i \text{ satisfying } A_i = 0.$$

In other words, when $\mathcal{P}$ contains a correctly specified model for $\pi(\boldsymbol{X})$, this correct model is implicitly accounted for by the empirical likelihood procedure to make $\hat{w}_{\mathrm{mr},i}$ the form of the IPW weight as used in (2), and this leads to the consistency of $\hat{\tau}_{1,0\mathrm{mr}}$:

$$\begin{aligned}\hat{\tau}_{1,0\mathrm{mr}} = \sum_{\{i:A_i=0\}} \hat{w}_{\mathrm{mr},i}Y_i &\xrightarrow{p} E\left[\left(E\left\{\frac{1}{\pi(\boldsymbol{X})} \mid A=1\right\}-1\right)\frac{\pi(\boldsymbol{X})}{1-\pi(\boldsymbol{X})}Y \mid A=0\right] \\ &= E\left[\frac{\pi(\boldsymbol{X})}{1-\pi(\boldsymbol{X})}Y^0 \mid A=0\right]\frac{P(A=0)}{P(A=1)} = E(Y^0 \mid A=1),\end{aligned}$$

where the second last equality follows from

$$E\left[\frac{1-\pi(\boldsymbol{X})}{\pi(\boldsymbol{X})} \mid A=1\right] = \frac{P(A=0)}{P(A=1)}$$

that can be shown by taking $b(\boldsymbol{X}) = \{1-\pi(\boldsymbol{X})\}/\pi(\boldsymbol{X})$ in (5), and the last equality follows from (1).

A difference between the constraints (8) and those used by Han and Wang (2013) and Wang (2019) for estimating ATE is that in (8) it is $1/\pi^{(j)}(\boldsymbol{\alpha}^{(j)})$ to be balanced whereas in Han and Wang (2013) and Wang (2019) it is $\pi^{(j)}(\boldsymbol{\alpha}^{(j)})$. The reason for this difference is that, for consistent estimation of ATT, the calibration weight $\hat{w}_{\mathrm{mr}}$ needs to implicitly be of the form $\pi(\boldsymbol{X})/\{1-\pi(\boldsymbol{X})\}$ as in (2) whereas, for consistent estimation of ATE, the calibration weight needs to implicitly be of the IPW form $1/\pi(\boldsymbol{X})$. From the above algebra it is seen that balancing $1/\pi^{(j)}(\boldsymbol{\alpha}^{(j)})$ in (8) ensures $\pi(\boldsymbol{X})/\{1-\pi(\boldsymbol{X})\}$ implicitly appear, whereas the algebra in Han and Wang (2013) shows that balancing $\pi^{(j)}(\boldsymbol{\alpha}^{(j)})$ ensures $1/\pi(\boldsymbol{X})$ implicitly appear. Thus, consistency of our estimator for ATT would not hold if in (8) $1/\pi^{(j)}(\boldsymbol{\alpha}^{(j)})$ is replaced by $\pi^{(j)}(\boldsymbol{\alpha}^{(j)})$.

When $\mathcal{D}_0$ contains a correctly specified model for $E(Y^0 \mid \boldsymbol{X})$, say $d_0^{(1)}(\boldsymbol{\gamma}^{(1)})$ without loss of generality, the consistency of $\hat{\tau}_{1,0\mathrm{mr}}$ follows from

$$\begin{aligned}
\sum_{\{i:A_i=0\}} \hat{w}_{\mathrm{mr},i} Y_i &= \sum_{\{i:A_i=0\}} \hat{w}_{\mathrm{mr},i} \left\{ Y_i - d_{0i}^{(1)}(\hat{\boldsymbol{\gamma}}^{(1)}) \right\} + \frac{1}{\sum_{s=1}^n A_s} \sum_{i=1}^n A_i d_{0i}^{(1)}(\hat{\boldsymbol{\gamma}}^{(1)}) \\
&= \frac{1}{\sum_{s=1}^n (1-A_s)} \sum_{\{i:A_i=0\}} \frac{Y_i - d_{0i}^{(1)}(\hat{\boldsymbol{\gamma}}^{(1)})}{1+\hat{\boldsymbol{\rho}}^\top \hat{\boldsymbol{g}}_i(\hat{\boldsymbol{\alpha}},\hat{\boldsymbol{\gamma}})} + E\{E(Y^0 \mid \boldsymbol{X}) \mid A=1\} + o_p(1) \\
&= E\left\{ \frac{Y^0 - E(Y^0 \mid \boldsymbol{X})}{1+\boldsymbol{\rho}_*^\top \boldsymbol{g}_*(\boldsymbol{\alpha}_*,\boldsymbol{\gamma}_*)} \mid A=0 \right\} + E\{E(Y^0 \mid \boldsymbol{X}, A=1) \mid A=1\} + o_p(1) \\
&= E\left[E\left\{ \frac{Y^0 - E(Y^0 \mid \boldsymbol{X}, A=0)}{1+\boldsymbol{\rho}_*^\top \boldsymbol{g}_*(\boldsymbol{\alpha}_*,\boldsymbol{\gamma}_*)} \mid \boldsymbol{X}, A=0 \right\} \mid A=0 \right] + E(Y^0 \mid A=1) + o_p(1) \\
&= E(Y^0 \mid A=1) + o_p(1),
\end{aligned}$$

where $\boldsymbol{\rho}_*$ and $\boldsymbol{g}_*(\boldsymbol{\alpha}_*,\boldsymbol{\gamma}_*)$ are the probability limits of $\hat{\boldsymbol{\rho}}$ and $\hat{\boldsymbol{g}}(\hat{\boldsymbol{\alpha}},\hat{\boldsymbol{\gamma}})$, respectively.

In summary, we have the following result on the multiple robustness property for the proposed estimator of the ATT.

Proposition: *If $\mathcal{P}$ contains a correctly specified model for $\pi(\boldsymbol{X})$ or $\mathcal{D}_0$ contains a correctly specified model for $E\{Y^0|\boldsymbol{X}\}$, then $\frac{1}{\sum_{i=1}^n A_i}\sum_{i=1}^n A_i Y_i - \sum_{\{i:A_i=0\}} \hat{w}_i Y_i \xrightarrow{p} E(Y^1|A=1) - E(Y^0|A=1)$ as $n \to \infty$.*

The proposed estimator $\hat{\tau}_{1,0\mathrm{mr}}$ and the EB estimator $\hat{\tau}_{1,0\mathrm{eb}}$ are both weighted averages of the observed outcomes for the control group subjects, and both are calibration-type estimators originally considered in survey sampling (Deville & Särndal, 1992) that were later extended to missing data analysis and causal inference (e.g., Chan & Yam, 2014; S. Chen & Haziza, 2017; Han & Wang, 2013; Kim, 2010; Kim & Park, 2010; Qin & Zhang, 2007; Tan, 2010; Wu & Sitter, 2001; Zhang et al., 2022). Some of the constraints in (8) based on the $\boldsymbol{h}(\boldsymbol{X})$ component in $\hat{\boldsymbol{g}}(\hat{\boldsymbol{\alpha}},\hat{\boldsymbol{\gamma}})$ are actually the constraints (3) used by the EB method, and thus our proposed method achieves the same degree of covariate balancing between the treatment and the control groups. The other constraints in (8) based on parametric models are for multiple robustness purpose.

Note that the constraints in (8) are for estimating ATT and are determined by (7), which balance the control group with the treatment group. This is different from estimating ATE where the constraints balance the control group with the full sample or the treatment group with the full sample (e.g., Fan et al., 2023). Adding the latter type of constraints into (8) would not work, since the weight matching control to treatment is different from the weight matching control to the full sample.

Standard error of the proposed estimator is needed to make inference, such as constructing confidence intervals. Due to the presence of multiple models and the lack of knowledge of which one is correct, deriving the asymptotic distribution of the proposed estimator as an approximation to the finite sample distributions is challenging. Therefore, we recommend using bootstrap to calculate the standard error. The excellent performance of the bootstrap method for multiply robust estimators in missing data context has been demonstrated through comprehensive simulation studies (e.g., Han, 2014a, 2014b), and its effectiveness for the proposed estimator will be shown in the next section.

4. Simulation studies

In this section we conduct simulation studies to evaluate the finite sample performance of the proposed multiply robust estimator of the ATT. The simulation setting mimics that in Han and Wang (2013). The data are generated in the following way: $X \sim \text{Uniform}(-2.5, 2.5)$, $Y^0 \mid X \sim N\{d_0(X), 4X^2+2\}$, $Y^1 \mid X \sim N\{d_1(X), 4X^2+2\}$

Table 1. Comparison of different estimators based on 1000 replications and $n = 300$. Each digit of the four-digit number inside an estimator's name, from left to right, indicates if $\pi^{(1)}(\boldsymbol{\alpha}^{(1)})$, $\pi^{(2)}(\boldsymbol{\alpha}^{(2)})$, $d_0^{(1)}(\boldsymbol{\gamma}^{(1)})$ and $d_0^{(2)}(\boldsymbol{\gamma}^{(2)})$ are used, respectively. The results have been multiplied by 100.

	$\pi^{(1)}$						$\pi^{(2)}$					
	$(d_0^{(1)}, d_1^{(1)})$			$(d_0^{(2)}, d_1^{(2)})$			$(d_0^{(1)}, d_1^{(1)})$			$(d_0^{(2)}, d_1^{(2)})$		
	rbias	rmse	mae	rbias	rmse	mae	rbias	rmse	mae	rbias	rmse	mae
IPW-1000	−6	99	60	−9	79	53	11	38	26	10	38	26
IPW-0100	69	93	64	−42	68	43	−1	44	28	−1	37	24
OR-0010	−8	57	37	−134	99	68	−2	35	23	−25	47	32
OR-0001	224	131	114	−7	51	34	−19	50	34	−2	36	24
AIPW-1010	−9	65	42	−16	80	54	−2	35	23	−34	51	39
AIPW-1001	−3	85	56	−8	65	41	34	53	38	−2	35	23
AIPW-0110	−9	62	40	−69	82	53	−2	35	23	−2	40	25
AIPW-0101	63	84	56	−8	61	39	−1	42	27	−2	35	24
MR-1000	−1	71	46	−10	76	50	14	40	27	12	39	26
MR-0100	51	77	52	−59	73	48	−1	49	32	−1	39	25
MR-0010	−8	57	37	−158	108	80	−2	35	23	−30	49	35
MR-0001	229	134	116	−6	52	34	21	47	32	−2	36	24
MR-1100	−24	76	49	−56	86	52	−3	39	26	−8	41	27
MR-1010	−10	65	42	−17	73	50	−1	35	23	−27	46	32
MR-1001	−1	73	50	−9	65	42	29	51	35	−2	35	23
MR-0110	−10	64	41	−68	80	52	−2	35	24	−2	38	24
MR-0101	60	79	53	−8	61	39	−2	40	26	−2	35	24
MR-0011	−13	83	51	−12	80	43	−2	35	24	−1	38	26
MR-1110	−9	65	42	−22	72	47	−2	36	24	−1	36	25
MR-1101	4	70	46	−9	65	42	−3	38	26	−2	35	24
MR-1011	−11	73	45	−10	73	44	−2	36	24	−2	36	24
MR-0111	−11	77	47	−10	78	48	−1	36	24	−2	35	24
MR-1111	−10	75	46	−10	75	46	−2	36	24	−2	35	24
EB-1	523	282	266	352	194	177	−317	321	311	−275	279	272
EB-2	−11	65	43	−19	71	46	−2	35	23	−31	49	36
MR-1	231	166	128	197	136	107	−318	323	313	−276	280	272
MR-2	−12	91	56	57	99	64	−1	34	23	−36	52	39

$\pi^{(1)}$: $\pi(X) = \{1 + \exp(0.8 + 0.5X - 0.3X^2)\}^{-1}$.
$\pi^{(2)}$: $\pi(X) = 1 - \exp[-\exp\{0.5 + 0.5X - 0.3\exp(X)\}]$.
$(d_0^{(1)}, d_1^{(1)})$: $\{d_0(X), d_1(X)\} = (1 + 2X + 3X^2, 2 + 3X + 3X^2)$.
$(d_0^{(2)}, d_1^{(2)})$: $\{d_0(X), d_1(X)\} = \{1 + 2X + 3\exp(X), 2 + 3X + 3\exp(X)\}$.
rbias: relative bias, bias divided by the true value. rmse: root mean square error. mae: median absolute error. IPW: inverse probability weighting. OR: outcome regression. AIPW: augmented inverse probability weighting. MR: multiply robust. EB: entropy balancing. EB-1: EB with $\boldsymbol{h}(X) = X$. EB-2: EB with $\boldsymbol{h}(X) = (X, X^2)$. MR-1: MR with $\boldsymbol{h}(X) = X$ and no working models. MR-2: MR with $\boldsymbol{h}(X) = (X, X^2)$ and no working models.

and $A \mid X \sim \text{Bernoulli}\{\pi(X)\}$, where $\pi(X)$ is either $\{1 + \exp(0.8 + 0.5X - 0.3X^2)\}^{-1}$ or $1 - \exp[-\exp\{0.5 + 0.5X - 0.3\exp(X)\}]$ and $\{d_0(X), d_1(X)\}$ is either $(1 + 2X + 3X^2, 2 + 3X + 3X^2)$ or $\{1 + 2X + 3\exp(X), 2 + 3X + 3\exp(X)\}$. Therefore, we have in total four data generating processes. For each of them, we postulate two propensity score models $\pi^{(1)}(\boldsymbol{\alpha}^{(1)}) = \{1 + \exp(\alpha_1^{(1)} + \alpha_2^{(1)}X + \alpha_3^{(1)}X^2)\}^{-1}$ and $\pi^{(2)}(\boldsymbol{\alpha}^{(2)}) = 1 - \exp[-\exp\{\alpha_1^{(2)} + \alpha_2^{(2)}X + \alpha_3^{(2)}\exp(X)\}]$ and two regression models for $d_0(X)$, $d_0^{(1)}(\boldsymbol{\gamma}^{(1)}) = \gamma_1^{(1)} + \gamma_2^{(1)}X + \gamma_3^{(1)}X^2$ and $d_0^{(2)}(\boldsymbol{\gamma}^{(2)}) = \gamma_1^{(2)} + \gamma_2^{(2)}X + \gamma_3^{(2)}\exp(X)$.

Tables 1 and 2 contain some simulation results summarized based on 1000 replications for $n = 300$ and $n = 800$, respectively. To show the flexibility of the proposed procedure in providing a unified framework for constructing estimators, both our proposed estimators based on each possible combination of models from $\{\pi^{(1)}(\boldsymbol{\alpha}^{(1)}), \pi^{(2)}(\boldsymbol{\alpha}^{(2)}), d_0^{(1)}(\boldsymbol{\gamma}^{(1)}), d_0^{(2)}(\boldsymbol{\gamma}^{(2)})\}$ and the corresponding OR, IPW and/or AIPW estimators, when available, are listed for comparison. Each digit of the four-digit number inside an estimator's name, from left to right, indicates if $\pi^{(1)}(\boldsymbol{\alpha}^{(1)})$, $\pi^{(2)}(\boldsymbol{\alpha}^{(2)})$, $d_0^{(1)}(\boldsymbol{\gamma}^{(1)})$ and $d_0^{(2)}(\boldsymbol{\gamma}^{(2)})$ are used, respectively. For example, MR-1101 denotes our proposed multiply robust estimator based on models $\pi^{(1)}(\boldsymbol{\alpha}^{(1)})$, $\pi^{(2)}(\boldsymbol{\alpha}^{(2)})$ and $d_0^{(2)}(\boldsymbol{\gamma}^{(2)})$. Compared to the OR, IPW and AIPW estimators, our proposed estimators using the same parametric models have similar numerical performance. The multiple robustness property is well demonstrated by the negligible biases, especially with $n = 800$, of the proposed estimators MR-1100, MR-0011, MR-1110, MR-1101, MR-1011, MR-0111 and MR-1111, which are consistent under all four data generating processes. The numerical performance of MR-1100 with a small sample size may be occasionally unstable, as shown by the relatively large bias when $n = 300$ with data generated based on $\pi(X) = \{1 + \exp(0.8 + 0.5X - 0.3X^2)\}^{-1}$, due to a high correlation between $\pi^{(1)}(\hat{\boldsymbol{\alpha}}^{(1)})$ and $\pi^{(2)}(\hat{\boldsymbol{\alpha}}^{(2)})$, but it improves dramatically as the sample size increases to $n = 800$.

The estimator EB-1 by balancing only the first moment of X has a very large bias under all four data generating processes. Replacing the exponential tilting with the empirical likelihood, the estimator MR-1 has a significantly improved performance compared to EB-1 when the data are generated based on $\pi(X) = \{1 + \exp(0.8 + 0.5X - 0.3X^2)\}^{-1}$. By balancing the first two moments of X, the EB method implicitly fits linear regression models for

Table 2. Comparison of different estimators based on 1000 replications and $n = 800$. Each digit of the four-digit number inside an estimator's name, from left to right, indicates if $\pi^{(1)}(\boldsymbol{\alpha}^{(1)})$, $\pi^{(2)}(\boldsymbol{\alpha}^{(2)})$, $d_0^{(1)}(\boldsymbol{\gamma}^{(1)})$ and $d_0^{(2)}(\boldsymbol{\gamma}^{(2)})$ are used, respectively. The results have been multiplied by 100.

	$\pi^{(1)}$						$\pi^{(2)}$					
	$(d_0^{(1)},d_1^{(1)})$			$(d_0^{(2)},d_1^{(2)})$			$(d_0^{(1)},d_1^{(1)})$			$(d_0^{(2)},d_1^{(2)})$		
	rbias	rmse	mae	rbias	rmse	mae	rbias	rmse	mae	rbias	rmse	mae
IPW-1000	2	58	39	1	46	32	14	26	18	13	26	17
IPW-0100	84	66	50	−32	41	28	0	27	18	0	23	15
OR-0010	2	35	25	−127	76	62	0	22	15	−24	34	26
OR-0001	244	126	119	2	32	22	−18	34	23	0	23	15
AIPW-1010	3	40	27	0	46	30	0	22	14	−33	40	33
AIPW-1001	4	52	35	3	40	27	37	44	35	0	22	15
AIPW-0110	3	38	26	−57	52	36	0	22	15	0	25	16
AIPW-0101	75	60	43	3	37	25	0	26	17	0	22	15
MR-1000	5	44	30	1	46	32	16	28	19	15	27	18
MR-0100	63	54	39	−50	46	32	0	30	19	0	24	16
MR-0010	2	35	25	−152	86	75	0	22	14	−28	37	29
MR-0001	246	127	120	2	32	22	24	35	26	0	23	15
MR-1100	−6	46	32	−18	49	32	−1	24	16	−3	24	15
MR-1010	3	40	28	0	43	30	0	22	15	−26	34	26
MR-1001	5	46	31	3	40	28	32	41	32	0	22	15
MR-0110	3	39	28	−56	51	35	0	22	15	0	23	15
MR-0101	73	57	41	3	37	25	0	25	16	0	22	15
MR-0011	5	55	36	8	68	30	0	22	15	1	24	15
MR-1110	3	40	27	−1	42	28	0	22	15	0	22	15
MR-1101	6	44	30	3	40	28	0	24	15	0	22	15
MR-1011	4	42	29	4	42	29	0	22	15	0	22	15
MR-0111	4	46	31	7	61	31	0	22	15	0	22	15
MR-1111	4	42	29	4	42	29	0	22	15	0	22	15
EB-1	564	281	277	383	193	189	−322	316	314	−279	274	271
EB-2	2	40	28	−1	42	29	0	22	15	−30	38	29
MR-1	251	142	126	220	121	112	−323	316	314	−279	274	272
MR-2	2	73	42	92	81	52	0	21	14	−35	41	34

$\pi^{(1)}$: $\pi(X) = \{1 + \exp(0.8 + 0.5X - 0.3X^2)\}^{-1}$.
$\pi^{(2)}$: $\pi(X) = 1 - \exp[-\exp\{0.5 + 0.5X - 0.3\exp(X)\}]$.
$(d_0^{(1)},d_1^{(1)})$: $\{d_0(X),d_1(X)\} = (1 + 2X + 3X^2, 2 + 3X + 3X^2)$.
$(d_0^{(2)},d_1^{(2)})$: $\{d_0(X),d_1(X)\} = \{1 + 2X + 3\exp(X), 2 + 3X + 3\exp(X)\}$.
rbias: relative bias, bias divided by the true value. rmse: root mean square error. mae: median absolute error.IPW: inverse probability weighting. OR: outcome regression. AIPW: augmented inverse probability weighting. MR: multiply robust. EB: entropy balancing. EB-1: EB with $\boldsymbol{h}(X) = X$. EB-2: EB with $\boldsymbol{h}(X) = (X, X^2)$. MR-1: MR with $\boldsymbol{h}(X) = X$ and no working models. MR-2: MR with $\boldsymbol{h}(X) = (X, X^2)$ and no working models.

logit$\{\pi(X)\}$ and $d_0(X)$ with regressors X and X^2, and this makes the estimator EB-2 consistent under the first three data generating processes and inconsistent under the last one where $\pi(X) = 1 - \exp[-\exp\{0.5 + 0.5X - 0.3\exp(X)\}]$ and $\{d_0(X), d_1(X)\} = \{1 + 2X + 3\exp(X), 2 + 3X + 3\exp(X)\}$. This theoretical conclusion is well confirmed by inspecting the bias of EB-2. The estimator MR-2 with exponential tilting replaced by empirical likelihood is inconsistent under all four data generating processes.

Table 3 contains a summary of the performance of the bootstrap method for standard error calculation. Due to similarity of the performance under different settings, we only include the results for $\pi(X) = \{1 + \exp(0.8 + 0.5X - 0.3X^2)\}^{-1}$ and $\{d_0(X), d_1(X)\} = (1 + 2X + 3X^2, 2 + 3X + 3X^2)$ with $n = 300$. The results are summarized based on 1000 replications with the bootstrap resampling size 200. It is seen that the average of the standard errors based on boostrap is very close to the empirical standard error. In addition, for the estimators that are consistent under this data generating process, i.e., MR-1000, MR-0010, MR-1100, MR-1010, MR-1001, MR-0110, MR-0011, MR-1110, MR-1101, MR-1011, MR-0111 and MR-1111, the coverage percentage of the 95% confidence intervals constructed based on bootstrap standard errors is close to the nominal level. Both observations show that the bootstrap method is reliable to calculate the standard errors for the proposed estimators.

5. Discussion

In this paper we have proposed a multiply robust estimator for the causal effect ATT. The estimator provides more protection on estimation consistency compared to existing doubly robust estimators. It is worth pointing out that, although the proposed method can simultaneously account for multiple working models, there does not have to be multiple models to apply this method. The construction of constraints is very flexible and the method can be applied when only a single working model is available. Therefore, the proposed method provides a unified framework as an alternative to various existing methods, including the IPW and AIPW methods.

A major difference between the proposed method and the EB method is the objective function being optimized. The EB method minimizes the Shannon entropy $\sum_{\{i:A_i=0\}} w_i \log w_i$, or equivalently the exponential tilting function

Table 3. Performance of the bootstrap method in calculating the standard errors for the proposed estimators in the setting of $\pi(X) = \{1 + \exp(0.8 + 0.5X - 0.3X^2)\}^{-1}$ and $\{d_0(X), d_1(X)\} = (1 + 2X + 3X^2, 2 + 3X + 3X^2)$. Results are summarized based on 1000 replications with $n = 300$ and the bootstrap resampling size is 200. Each digit of the four-digit number inside an estimator's name, from left to right, indicates if $\pi^{(1)}(\boldsymbol{\alpha}^{(1)})$, $\pi^{(2)}(\boldsymbol{\alpha}^{(2)})$, $d_0^{(1)}(\boldsymbol{\gamma}^{(1)})$ and $d_0^{(2)}(\boldsymbol{\gamma}^{(2)})$ are used, respectively.

	bias	se-emp	se-bp	cp-95%
MR-1000	0.05	0.71	0.67	92.50
MR-0100	0.31	0.72	0.69	89.00
MR-0010	−0.01	0.56	0.58	95.30
MR-0001	1.21	0.65	0.64	52.60
MR-1100	−0.08	0.77	0.74	94.10
MR-1010	0.01	0.64	0.62	93.60
MR-1001	0.05	0.74	0.70	92.40
MR-0110	0.00	0.62	0.61	94.00
MR-0101	0.36	0.74	0.70	87.50
MR-0011	0.03	0.81	0.75	93.30
MR-1110	0.01	0.63	0.61	93.60
MR-1101	0.08	0.70	0.67	92.00
MR-1011	0.03	0.72	0.70	94.40
MR-0111	0.03	0.76	0.73	93.90
MR-1111	0.04	0.76	0.73	94.10

se-emp: empirical standard error. se-bp: averaged bootstrap standard error. cp-95%: coverage percentage of the 95% confidence interval constructed based on the bootstrap standard error. MR: multiply robust.

in the empirical likelihood literature (e.g., Newey & Smith, 2004), whereas we proposed to maximize the empirical likelihood $\prod_{\{i:A_i=0\}} w_i$. The advantage of considering an empirical likelihood is manifold. First, it improves the robustness of estimation consistency by using constraints based on parametric working models. Although the same constraints can be used by the EB method as well, a correct model does not make the EB estimator consistent, and consistency of the EB estimator is achieved only when $\text{logit}\{\pi(\boldsymbol{X})\}$ or $E(Y^0 \mid \boldsymbol{X})$ is a linear model with regressors $\boldsymbol{h}(\boldsymbol{X})$. Second, by dividing the constraints into two parts, one based on parametric working models and one based on moments of $\boldsymbol{X}$, the proposed method is more flexible when the dimension of $\boldsymbol{X}$ is large. With $\pi(\boldsymbol{X})$ and/or $E(Y^0 \mid \boldsymbol{X})$ depending on a high dimensional $\boldsymbol{X}$, consistency of the EB estimator requires a large number of constraints to include all the regressors for $\text{logit}\{\pi(\boldsymbol{X})\}$ and/or $E(Y^0 \mid \boldsymbol{X})$, and this may jeopardize the numerical performance in practice. The proposed method, on the contrary, has a separate model building step where complex working models can be built and only the ones that are deemed to be close to the truth are used in the constraints. In this case, the parametric working models help to achieve consistency, and thus the other part of constraints based on moments of $\boldsymbol{X}$ can be chosen exclusively for the goal of achieving covariate balancing. This may considerably reduce the number of constraints compared to the EB method and thus improve the numerical performance. Third, it is well known in the empirical likelihood literature (e.g., Newey & Smith, 2004) that the empirical likelihood has smaller higher-order bias, which translates to a better numerical performance under a finite sample size, compared to exponential tilting and other alternatives. As an empirical version of the moment equality (7), the constraints in (8) are legitimate, at least when the sample size is large. Therefore, the 'model misspecification' problem that typically exists in the empirical likelihood literature under which the performance of empirical likelihood may be worse than that of the exponential tilting (Schennach, 2007) is not of a concern here.

Acknowledgments

We thank the Editor, Associate Editor and a referee for their valuable comments that have helped improve the quality of this paper.

Disclosure statement

No potential conflict of interest was reported by the author(s).

References

Abadie, A., & Imbens, G. W. (2006). Large sample properties of matching estimators for average treatment effects. *Econometrica*, *74*(1), 235–267. https://doi.org/10.1111/ecta.2006.74.issue-1

Bang, H., & Robins, J. M. (2005). Doubly robust estimation in missing data and causal inference models. *Biometrics*, *61*(4), 962–973. https://doi.org/10.1111/biom.2005.61.issue-4

Cao, W., Tsiatis, A. A., & Davidian, M. (2009). Improving efficiency and robustness of the doubly robust estimator for a population mean with incomplete data. *Biometrika*, *96*, 723–734. https://doi.org/10.1093/biomet/asp033

Chan, K. C. G., & Yam, S. C. P. (2014). Oracle, multiple robust and multipurpose calibration in a missing response problem. *Statistical Science*, *29*(3), 380–396.

Chan, K. C. G., Yam, S. C. P., & Zhang, Z. (2016). Globally efficient non-parametric inference of average treatment effects by empirical balancing calibration weighting. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *78*(3), 673–700. https://doi.org/10.1111/rssb.12129

Chen, J., Sitter, R. R., & Wu, C. (2002). Using empirical likelihood methods to obtain range restricted weights in regression estimators for surveys. *Biometrika*, *89*(1), 230–237. https://doi.org/10.1093/biomet/89.1.230

Chen, S., & Haziza, D. (2017). Multiply robust imputation procedures for the treatment of item nonresponse in surveys. *Biometrika*, *104*(2), 439–453.

Deville, J., & Särndal, C.-E. (1992). Calibration estimators in survey sampling. *Journal of the American Statistical Association*, *87*(418), 376–382. https://doi.org/10.1080/01621459.1992.10475217

Duan, X., & Yin, G. (2017). Ensemble approaches to estimating the population mean with missing response. *Scandinavian Journal of Statistics*, *44*(4), 899–917. https://doi.org/10.1111/sjos.v44.4

Fan, J., Imai, K., Lee, I., Liu, H., Ning, Y., & Yang, X. (2023). Optimal covariate balancing conditions in propensity score estimation. *Journal of Business and Economic Statistics*, *41*(1), 97–110. https://doi.org/10.1080/07350015.2021.2002159

Hainmueller, J. (2012). Entropy balancing for causal effects: A multivariate reweighting method to produce balanced samples in observational studies. *Political Analysis*, *20*(1), 25–46. https://doi.org/10.1093/pan/mpr025

Han, P. (2012). A note on improving the efficiency of inverse probability weighted estimator using the augmentation term. *Statistics and Probability Letters*, *82*(12), 2221–2228. https://doi.org/10.1016/j.spl.2012.08.005

Han, P. (2014a). A further study of the multiply robust estimator in missing data analysis. *Journal of Statistical Planning and Inference*, *148*, 101–110. https://doi.org/10.1016/j.jspi.2013.12.006

Han, P. (2014b). Multiply robust estimation in regression analysis with missing data. *Journal of the American Statistical Association*, *109*(507), 1159–1173. https://doi.org/10.1080/01621459.2014.880058

Han, P. (2016a). Combining inverse probability weighting and multiple imputation to improve robustness of estimation. *Scandinavian Journal of Statistics*, *43*(1), 246–260. https://doi.org/10.1111/sjos.v43.1

Han, P. (2016b). Intrinsic efficiency and multiple robustness in longitudinal studies with drop-out. *Biometrika*, *103*(3), 683–700. https://doi.org/10.1093/biomet/asw024

Han, P., Kong, L., Zhao, J., & Zhou, X. (2019). A general framework for quantile estimation with incomplete data. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *81*(2), 305–333. https://doi.org/10.1111/rssb.12309

Han, P., & Wang, L. (2013). Estimation with missing data: Beyond double robustness. *Biometrika*, *100*(2), 417–430. https://doi.org/10.1093/biomet/ass087

Hernán, M. A., & Robins, J. M. (2018). *Causal inference*. Chapman & Hall/CRC.

Hirano, K., Imbens, G. W., & Ridder, G. (2003). Efficient estimation of average treatment effects using the estimated propensity score. *Econometrica*, *71*(4), 1161–1189. https://doi.org/10.1111/ecta.2003.71.issue-4

Imbens, G. W., & Rubin, D. B. (2015). *Causal inference for statistics, social, and biomedical sciences: An introduction*. Cambridge University Press.

Kang, J. D. Y., & Schafer, J. L. (2007). Demystifying double robustness a comparison of alternative strategies for estimating a population mean from incomplete data. *Statistical Science*, *22*(4), 523–539.

Kim, J. K. (2010). Calibration estimation using exponential tilting in sample surveys. *Survey Methodology*, *36*(2), 145–155.

Kim, J. K., & Park, M. (2010). Calibration estimation in survey sampling. *International Statistical Review*, *78*(1), 21–39. https://doi.org/10.1111/insr.2010.78.issue-1

Li, W., Yang, S., & Han, P. (2020). Robust estimation for moment condition models with data missing not at random. *Journal of Statistical Planning and Inference*, *207*, 246–254. https://doi.org/10.1016/j.jspi.2020.01.001

Molina, J., Rotnitzky, A., Sued, M., & Robins, J. (2017). Multiple robustness in factorized likelihood models. *Biometrika*, *104*, 561–581. https://doi.org/10.1093/biomet/asx027

Newey, W. K., & Smith, R. J. (2004). Higher order properties of GMM and generalized empirical likelihood estimators. *Econometrica*, *72*(1), 219–255. https://doi.org/10.1111/ecta.2004.72.issue-1

Qin, J., & Lawless, J. (1994). Empirical likelihood and general estimating equations. *The Annals of Statistics*, *22*(1), 300–325. https://doi.org/10.1214/aos/1176325370

Qin, J., Shao, J., & Zhang, B. (2008). Efficient and doubly robust imputation for covariate-dependent missing responses. *Journal of the American Statistical Association*, *103*(482), 797–810. https://doi.org/10.1198/016214508000000238

Qin, J., & Zhang, B. (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), 101–122. https://doi.org/10.1111/j.1467-9868.2007.00579.x

Robins, J. M., Rotnitzky, A., & Zhao, L. P. (1994). Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association*, *89*(427), 846–866. https://doi.org/10.1080/01621459.1994.10476818

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

Rosenbaum, P. R., & Rubin, D. B. (1985). Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. *The American Statistician*, *39*(1), 33–38.

Rotnitzky, A., Lei, Q., Sued, M., & Robins, J. M. (2012). Improved double-robust estimation in missing data and causal inference models.. *Biometrika*, *99*, 439–456. https://doi.org/10.1093/biomet/ass013

Schennach, S. (2007). Point estimation with exponentially tilted empirical likelihood. *Annals of Statistics*, *35*(2), 634–672. https://doi.org/10.1214/009053606000001208

Shi, X., Miao, W., Nelson, J. C., & Tchetgen Tchetgen, E. (2020). Multiply robust causal inference with double-negative control adjustment for categorical unmeasured confounding. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *82*(2), 521–540. https://doi.org/10.1111/rssb.12361

Tan, Z. (2010). Bounded, efficient and doubly robust estimation withinverse weighting. *Biometrika*, *97*(3), 661–682. https://doi.org/10.1093/biomet/asq035

van der Laan, M. J., & Gruber, S. (2010). Collaborative double robust targeted maximum likelihood estimation. *The International Journal of Biostatistics*, *6*(1), Article 17.

van der Vaart, A. W. (1998). *Asymptotic statistics*. Cambridge University Press.

Wang, L. (2019). Multiple robustness estimation in causal inference. *Communications in Statistics - Theory and Methods*, *48*(23), 5701–5718. https://doi.org/10.1080/03610926.2018.1520881

Wang, L., & Tchetgen Tchetgen, E. (2018). Bounded, efficient and multiply robust estimation of average treatment effects using instrumental variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *80*(3), 531–550. https://doi.org/10.1111/rssb.12262

Wu, C., & Sitter, R. R. (2001). A model-calibration approach to using complete auxiliary information from survey data. *Journal of the American Statistical Association*, *96*(453), 185–193. https://doi.org/10.1198/016214501750333054

Zhang, S., Han, P., & Wu, C. (2022). Calibration techniques encompassing survey sampling, missing data analysis and causal inference. *International Statistical Review*. https://doi.org/10.1111/insr.12518 .

Zhao, Q., & Percival, D. (2017). Entropy balancing is doubly robust. *Journal of Causal Inference*, *5*(1). https://doi.org/10.1515/jci-2016-0010