Qingyuan Zhao[1] / Daniel Percival[2]

Entropy Balancing is Doubly Robust

[1] Department of Statistics, The Wharton School, University of Pennsylvania, Philadelphia, PA, USA, E-mail:
    qyzhao@wharton.upenn.edu. http://orcid.org/0000-0001-9902-2768.
[2] Google Inc., Mountain View, CA, USA, E-mail: dancsi@google.com

Abstract:
Covariate balance is a conventional key diagnostic for methods estimating causal effects from observational studies. Recently, there is an emerging interest in directly incorporating covariate balance in the estimation. We study a recently proposed entropy maximization method called Entropy Balancing (EB), which exactly matches the covariate moments for the different experimental groups in its optimization problem. We show EB is doubly robust with respect to linear outcome regression and logistic propensity score regression, and it reaches the asymptotic semiparametric variance bound when both regressions are correctly specified. This is surprising to us because there is no attempt to model the outcome or the treatment assignment in the original proposal of EB. Our theoretical results and simulations suggest that EB is a very appealing alternative to the conventional weighting estimators that estimate the propensity score by maximum likelihood.

Keywords: causal inference, double robustness, exponential tilting, convex optimization, survey sampling

1   Introduction

Consider a typical setting of observational study that two conditions ("treatment" and "control") are not randomly assigned to the units. Deriving a causal conclusion from such observational data is essentially difficult because the treatment exposure may be related to some covariates that are also related to the outcome. In this case, those covariates may be imbalanced between the treatment groups and the naive mean causal effect estimator can be severely biased.

To adjust for the covariate imbalance, the seminal work of Rosenbaum and Rubin [1] points out the essential role of *propensity score*, the probability of exposure to treatment conditional on observed covariates. This quantity, rarely known in an observation study, may be estimated from the data. Based on the estimated propensity score, many statistical methods are proposed to estimate the mean causal effect. The most popular approaches are *matching* [2, 3], *stratification* [4], and *weighting* [5, 6]. Theoretically, propensity score weighting is the most attractive among these methods. Hirano et al. [7] show that nonparametric propensity score weighting can achieve the semiparametric efficiency bound for the estimation of mean causal effect derived by Hahn [8]. Another desirable property is double robustness. The pioneering work of Robins et al. [5] augments propensity score weighting by an outcome regression model. The resulting estimator has the so-called *double robustness* property:

Property 1. *If either the propensity score model or the outcome regression model is correctly specified, the mean causal effect estimator is statistically consistent*.

In practice, the success of any propensity score method hinges on the quality of the estimated propensity score. The weighting methods are usually more sensitive to model misspecification than matching and stratification, and furthermore, this bias can even be amplified by a doubly robust estimator, brought to attention by Kang and Schafer [9]. In order to avoid model misspecification, applied researchers usually increase the complexity of the propensity score model until a sufficiently balanced solution is found. This cyclical process of modeling propensity score and checking covariate balance is criticized as the "propensity score tautology" by Imai et al. [10] and, moreover, has no guarantee of finding a satisfactory solution eventually.

Recently, there is an emerging interest, particularly among applied researchers, in directly incorporating covariate balance in the estimation procedure, so there is no need to check covariate balance repeatedly [11–14]. In this paper, we study a method of this kind called *Entropy Balancing* (hereafter EB) proposed in Hainmueller [15]. In a nutshell, EB solves an (convex) entropy maximization problem under the constraint of exact balance of covariate moments. Due to its easy interpretation and fast computation, EB has already gained some popularity in applied fields [16, 17]. However, little do we known about the theoretical properties of EB. The original

Qingyuan Zhao is the corresponding author.

proposal in Hainmueller [15] did not give a condition such that EB is guaranteed to give a consistent estimate of the mean causal effect.

In this paper, we shall show EB is indeed a very appealing propensity score weighting method. We find EB simultaneously fits a logistic regression model for the propensity score and a linear regression model for the outcome. The linear predictors of these regression models are the covariate moments being balanced. We shall prove EB is doubly robust (Property 1), in the sense that if at least one of the two models are correctly specified, EB is consistent for the Population Average Treatment effect for the Treated (PATT), a common quantity of interest in causal inference and survey sampling. Moreover, EB is sample bounded [18], meaning the PATT estimator is always within the range of the observed outcomes, and it is semiparametrically efficient if both models are correctly specified. Lastly, The two linear models have an exact correspondence to the primal and dual optimization problem used to solve EB, revealing an interesting connection between doubly robust estimation and convex optimization.

Our discoveries can be summarized in the diagram in Figure 1. Conventionally, the recipe given by Robins and his coauthors is to fit separate models for propensity score and outcome regression and then combine them by a doubly robust estimator (see e. g. [5, 9, 19, 20]). In contrast, Entropy Balancing achieves this goal through enforcing covariate balance. The primal optimization problem of EB amounts to an empirical calibration estimator [21, 22], which is widely popular in survey sampling but perhaps not sufficiently recognized in causal inference [23]. The balancing constraints in this optimization problem result in unbiasedness of the PATT estimator under linear outcome regression model. The dual optimization problem of EB is fitting a logistic propensity score model with a loss function different from the negative binomial likelihood. The Fisher-consistency of this loss function (also called *proper* scoring rule in statistical decision theory, see e. g. Gneiting and Raftery [24]) ensures the other half of double robustness – consistency under correctly specified propensity score model. Since EB essentially just uses a different loss function, other types of propensity score models, for example the generalized additive models [25] can also easily be fitted. A forthcoming article by Zhao [26] offers more discussion and extension to other weighted average treatment effects.
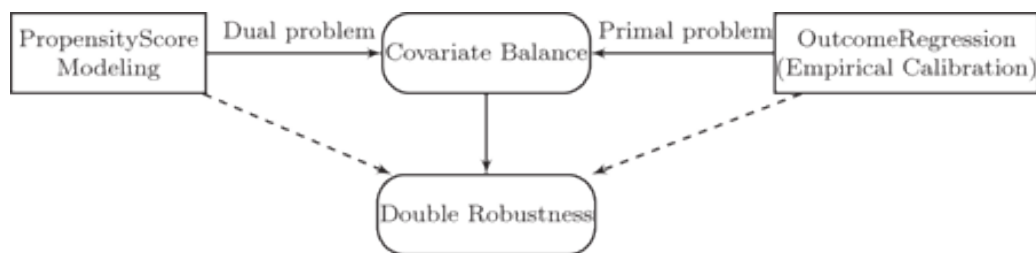

Figure 1: The role of covariate balance in doubly robust estimation. Dashed arrows: conventional procedure to achieve double robustness. Solid arrows: double robustness of Entropy Balancing via covariate balance.

2 Setting

First, we fix some notations for the causal inference problem considered in this paper. We follow the potential outcome language of Neyman [27] and Rubin [28]. In this causal model, each unit i is associated with a pair of potential outcomes: the response $Y_i(1)$ that is realized if $T_i = 1$ (treated), and another response $Y_i(0)$ realized if $T_i = 0$ (control). We assume the observational units are independent and identically distributed samples from a population, for which we wish to infer the treatment's effect. The main obstacle is that only one potential outcome is observed: $Y_i = T_i Y_i(1) - (1 - T_i) Y_i(0)$, which is commonly known as the "fundamental problem of causal inference" [29].

In this paper we focus on estimating the Population Average Treatment effect on the Treated (PATT):

$$\gamma = \mathrm{E}[Y(1)|T = 1] - \mathrm{E}[Y(0)|T = 1] \stackrel{\Delta}{=} \mu(1|1) - \mu(0|1). \tag{1}$$

The counterfactual mean $\mu(0|1) = \mathrm{E}[Y(0)|T = 1]$ also naturally occurs in survey sampling with missing data [21, 22] by viewing $Y(0)$ as the only outcome of interest (so $T = 1$ stands for non-response).

Along with the treatment exposure T_i and outcome Y_i, each unit i is usually associated with a set of covariates denoted by X_i measured prior to the treatment assignment. In a typical observational study, both treatment assignment and outcome may be related to the covariates, which can cause serious confounding bias. The seminal work by Rosenbaum and Rubin [1] suggest that it is possible to correct the confounding bias under the following two assumptions:

Assumption 1 (strong ignorability). $(Y(0), Y(1)) \perp T|X$.

Assumption 2 (overlap). $0 < P(T = 1|X) < 1$.

Intuitively, the first assumption says that the observed covariates contain all the information that may cause the selection bias, i. e. there is no unmeasured confounding variable, and the second assumption ensures that the bias-correction information is available across the entire domain of X.

Since the covariates X contain all the information of confounding bias, it is important to understand the relationship between T, Y and X. Under Assumption 1 (strong ignorability), the joint distribution of (X, Y, T) is determined by the marginal distribution of X and two conditional distributions given X. The first conditional distribution $e(X) = P(T = 1|X)$ is often called the propensity score and plays a central role in causal inference [1]. The second conditional distribution is the density of $Y(0)$ and $Y(1)$ given X. Since we only consider the mean causal effect in this paper, it suffices to study the mean regression functions $g_0(X) = E[Y(0)|X]$ and $g_1(X) = E[Y(1)|X]$.

To estimate the PATT defined in 1, a conventional weighting estimator based on the propensity score is the inverse probability weighting (IPW) defined as

$$\hat{\gamma}^{IPW} = \sum_{T_i=1} \frac{1}{n_1} Y_i - \sum_{T_i=0} \frac{\hat{e}(X_i)(1 - \hat{e}(X_i))^{-1}}{\sum_{T_i=0} \hat{e}(X_i)(1 - \hat{e}(X_i))^{-1}} Y_i. \tag{2}$$

Here $\sum_{T_i=t}$ is a short-hand notation of summation over all units i such that $T_i = t$. This will be repeatedly used throughout this paper. In 2, the control units are weighted proportionally to $\hat{e}(X_i)(1 - \hat{e}(X_i))^{-1}$ to resemble the full population. The most popular choice of propensity score model is the logistic regression, where $\text{logit}(e(x)) = \log[e(x)/(1 - e(x))]$ is modeled by $\sum_{j=1}^{p} \theta_j c_j(x)$ and $c_j(x)$ are functions of the covariates.

3 Entropy balancing

Entropy Balancing (EB) is an alternative weighting method proposed by Hainmueller [15] to estimate PATT. EB operates by maximizing the entropy of the weights under some pre-specified balancing constraints:

$$\begin{aligned}
&\underset{w}{\text{maximize}} && -\sum_{T_i=0} w_i \log w_i \\
&\text{subject to} && \sum_{T_i=0} w_i c_j(X_i) = \bar{c}_j(1) = \frac{1}{n_1} \sum_{T_i=1} c_j(X_i), j = 1, \dots, p, \\
& && \sum_{T_i=0} w_i = 1, \\
& && w_i > 0, i = 1, \dots, n.
\end{aligned} \tag{3}$$

Hainmueller [15] proposes to use the weighted average $\sum_{T_i=0} w_i^{EB} Y_i$ to estimate the counterfactual mean $E[Y(0)|T = 1]$. This gives the Entropy Balancing estimator of PATT

$$\hat{\gamma}^{EB} = \sum_{T_i=1} \frac{Y_i}{n_1} - \sum_{T_i=0} w_i^{EB} Y_i. \tag{4}$$

The functions $\{c_j(\cdot)\}_{j=1}^{p}$ in 3 are called moment functions of the covariates. They can be any transformation of X, not necessarily polynomial functions. We use $c(X)$ and $\bar{c}(1)$ to stand for the vector of $c_j(X)$ and $\bar{c}_j(1), j = 1, \dots, p$. We shall see the functions $\{c_j(\cdot)\}_{j=1}^{p}$ indeed serve as the linear predictors in the propensity score model and the outcome regression model, although at this point it is not even clear that EB attempts to fit any model.

First, we give some heuristics that allows us to view EB as a propensity score weighting method. Since EB seeks to empirically match the control and treatment covariate distributions, we connect EB with density estimation. Let $m(x)$ be the density function of the covariates X for the control population. The minimum relative entropy principle estimates the density of the treatment population by

$$\underset{\tilde{m}}{\text{maximize}} \, H(\tilde{m} \parallel m) \quad \text{subject to} \, E_{\tilde{m}}[c(X)] = \bar{c}(1), \tag{5}$$

where $H(\tilde{m} \parallel m) = \mathrm{E}_{\tilde{m}}[\log(\tilde{m}(X)/m(X))]$ is the relative entropy between $\tilde{m}$ and m. As an estimate of the distribution of the treatment group, the optimal $\tilde{m}$ of 5 is the "closest" to the control distribution among all distributions satisfying the moment constraints. Now let $w(x) = [\mathrm{P}(T = 1) \cdot \tilde{m}(x)]/[\mathrm{P}(T = 0) \cdot m(x)]$ be the population version of the inverse probability weights in 2. Applying a change of measure, we can rewrite 5 as an optimization problem over $w(x)$:

$$\underset{w}{\text{maximize}}\, \mathrm{E}_m[w(X) \log w(X)] \quad \text{subject to} \, \mathrm{E}_m[w(X)c(X)] = \bar{c}(1). \tag{6}$$

The EB optimization problem 3 is the finite sample version of 6, where the population distribution m is replaced by the empirical distribution of the control units.

Using the Lagrangian multipliers, one can show the solution of 5 belongs to the family of exponential titled distributions of m [30]:

$$m_\theta(x) = m(x) \exp(\theta^T c(x) - \psi(\theta)).$$

Here, $\psi(\theta)$ is the moment generating function of this exponential family. Consequently, the solution of the population EB 6 is

$$\frac{e(x)}{1 - e(x)} = \frac{\mathrm{P}(T = 1 | X = x)}{\mathrm{P}(T = 0 | X = x)} = w(x) = \exp(\alpha + \theta^T c(x)),$$

where $\alpha = \log(\mathrm{P}(T = 1)/\mathrm{P}(T = 0))$. This is exactly the logistic regression model with predictors $c(x)$.

Notice that EB is different from the maximum likelihood fit of the logistic regression. The dual optimization problem of 3 is

$$\underset{\theta}{\text{minimize}} \quad \log\left(\sum_{T_i=0} \exp(\sum_{j=1}^{p} \theta_j c_j(X_i))\right) - \sum_{j=1}^{p} \theta_j \bar{c}_j(1), \tag{7}$$

whereas the maximum likelihood solves

$$\underset{\theta}{\text{minimize}} \quad \sum_{i=1}^{n} \log\left(1 + \exp(-(2T_i - 1)\sum_{j=1}^{p} \theta_j c_j(X_i))\right). \tag{8}$$

It is apparent from 7 and 8 that EB and maximum likelihood use different loss functions. As a remark, the estimating equations defined by 7 are used to augment the estimating equations defined by 8 in the covariate balancing propensity score (CBPS) approach of Imai and Ratkovic [13]. We will compare the empirical performance of these methods in Section 5.

The optimization problem 7 is strictly convex and the unique solution $\hat{\theta}^{\mathrm{EB}}$ can be efficiently computed by Newton method. The EB weights (solution to the primal problem 3) are given by the Karush-Kuhn-Tucker (KKT) conditions: for any i such that $T_i = 0$,

$$w_i^{EB} = \frac{\exp\left(\sum_{j=1}^{p} \hat{\theta}_j^{EB} c_j(X_i)\right)}{\sum_{T_i=0} \exp\left(\sum_{j=1}^{p} \hat{\theta}_j^{EB} c_p(X_i)\right)}. \tag{9}$$

As a final remark, Entropy Balancing bridges two existing approaches of estimating the mean causal effect:

(1) The calibration estimator that is very popular in survey sampling [21–23];

(2) The empirical likelihood approach that significantly advances the theory of doubly robust estimation in observation study [18, 31–33].

EB is a special case of these two approaches. The main distinction is that it uses the Shannon entropy $\sum_{i=1}^{n} w_i \log w_i$ as the discrepancy function, resulting in an easy-to-solve convex optimization. Due to its easy interpretation, Entropy Balancing has already gained some ground in practice (e. g. [16, 17]).

4 Properties of entropy balancing

We give some theoretical guarantees of Entropy Balancing to justify its usage in real applications. The following is the main theorem of this paper, which shows EB is doubly robust even though its original form 3 does not contain a propensity score model or a outcome regression model.

Theorem 1. *Let Assumption 1 (strong ignorability) and Assumption 2 (overlap) be given. Additionally, assume the expectation of $c(x)$ exists and $Var(Y(0)) < \infty$. Then Entropy Balancing is doubly robust (Property 1) in the sense that*

(1) If $logit(e(x))$ or $g_0(x)$ is linear in $c_j(x), j = 1, \ldots, R$, then $\hat{\gamma}^{EB}$ is statistically consistent.

(2) Moreover, if $logit(e(x)), g_0(x)$ and $g_1(x)$ are all linear in $c_j(x), j = 1, \ldots, R$, then $\hat{\gamma}^{EB}$ reaches the semiparametric variance bound of γ derived in Hahn [34] with unknown propensity score.

We give two proofs of the first claim in Theorem 1. The first proof reveals an interesting connection between the primal-dual optimization problems 3 and 7 and the statistical property, double robustness, which motivates the interpretation in Figure 1. The second proof uses a stabilization trick in Robins et al. [8].

First proof (sketch) The consistency under the linear model of $logit(P(T = 1|X))$ is a consequence of the dual optimization problem 7. See Section 3 for a heuristic justification via the minimum relative entropy principle and Appendix A for a rigorous proof by using the M-estimation theory.

The consistency under the linear model of $Y(0)$ can be proved by expanding $E[Y(0)|X]$ and $\sum_{T_i=0} w_i Y_i$. Here we provide an indirect proof by showing that augmenting EB with a linear outcome regression does not change the estimator. Given an estimated propensity score model $\hat{e}(x)$, the corresponding weights $\hat{e}(x)/(1 - \hat{e}(x))$ for the control units, and an estimated outcome regression model $\hat{g}_0(x)$, a doubly robust estimator of PATT is given by

$$\hat{\gamma}^{DR} = \sum_{T_i=1} \frac{1}{n_1}(Y_i - \hat{g}_0(X_i)) - \sum_{T_i=0} \frac{\hat{e}(X_i)}{1 - \hat{e}(X_i)}(Y_i - \hat{g}_0(X_i)). \tag{10}$$

This estimator satisfies Property 1, i. e. if $\hat{e}(x) \to e(x)$ or $\hat{g}_0(x) \to g(x)$, then $\hat{\gamma}^{DR}$ is statistically consistent for γ. To see this, in the case that $\hat{g}_0(x) \to g_0(x)$, the first sum in 10 is consistent for γ and the second sum in 10 has mean going to 0 as $n \to \infty$. In the case where $\hat{g}_0(x) g_0(x)$ but $\hat{e}(x) \to e(x)$, the second sum in 10 is consistent for the bias of the first sum (as an estimator of γ).

When the estimated propensity score model $\hat{e}(x)$ is obtained by the EB dual problem 7 and the estimated outcome regression model is $\hat{g}_0(x) = \sum_{j=1}^{p} \hat{\beta}_j c_j(x)$, we have

$$
\begin{aligned}
\hat{\gamma}^{DR} - \hat{\gamma}^{EB} \quad &= \sum_{T_i=0} w_i^{EB} \hat{g}_0(X_i) - \frac{1}{n_1} \sum_{T_i=0} \hat{g}_0(X_i) \\
&= \sum_{T_i=0} w_i^{EB} \sum_{j=1}^{p} \hat{\beta}_j c_j(X_i) - \frac{1}{n_1} \sum_{T_i=1} \sum_{j=1}^{p} \hat{\beta}_j c_j(X_i) \\
&= \sum_{j=1}^{p} \hat{\beta}_j \left(\sum_{T_i=0} w_i^{EB} c_j(X_i) - \frac{1}{n_1} \sum_{T_i=1} c_j(X_i) \right) \\
&= 0.
\end{aligned}
$$

Therefore by enforcing covariate balancing constraints, EB implicitly fits a linear outcome regression model and is consistent for γ under this model. $\square$

Second proof. This proof is pointed out by an anonymous reviewer. In a discussion of Kang and Schafer [9], Robins et al. [34] show that one can stabilize the standard doubly robust estimator in a number of ways. Specifically, one trick suggested by Robins et al. [34] is to estimate the propensity score, say $\tilde{e}(x)$, by the following estimating equation

$$\sum_{i=1}^{n} \left[\frac{(1 - T_i)\tilde{e}(X_i)/(1 - \tilde{e}(X_i))}{\sum_{i=1}^{n}(1 - T_i)\tilde{e}(X_i)/(1 - \tilde{e}(X_i))} - \frac{T_i}{\sum_{i=1}^{n} T_i} \right] \hat{g}_0(X_i) = 0. \tag{11}$$

Then one can estimate PATT by the IPW estimator 2 by replacing $\hat{e}(X_i)$ with $\tilde{e}(X_i)$. This estimator is sample bounded (the estimator is always within the range of observed values of Y) and doubly robust with respect to

the parametric specifications of $\tilde{e}(x) = \tilde{e}(x;\theta)$ and $\hat{g}_0(x) = \hat{g}_0(x;\beta)$. The only problem with 11 is it may not have a unique solution. However, when $\text{logit}(e(x))$ and $g_0(x)$ are assumed linear in $c(x)$, 11 corresponds to the first order condition of the EB dual problem 7. Since 7 is strictly convex, it has an unique solution and $\tilde{e}(X;\theta)$ is the same as the EB estimate $\hat{e}(X;\theta)$. As a consequence, $\hat{\gamma}^{EB}$ is also doubly robust.

To prove the second claim in Theorem 1, we compute the asymptotic variance of $\hat{\gamma}^{EB}$ using the M-estimation theory. To state our results, we need to introduce four differently weighted covariance-like functions for two random vectors a_1 and a_2 of length p:

$$
\begin{aligned}
H_{a_1,a_2} &= \text{Cov}(a_1, a_2 | T = 1), \\
G_{a_1,a_2} &= \text{E}\left[\frac{e(X)}{1-e(X)} (a_1 - \text{E}[a_1|T=1])(a_2 - \text{E}[a_2|T=1])^T | T = 1 \right], \\
K_{a_1,a_2} &= \text{E}[(1 - e(X)) a_1 a_2^T | T = 1], \\
K^m_{a_1,a_2} &= \text{E}[(1 - e(X)) a_1 (a_2 - \text{E}[a_2|T=1])^T | T = 1].
\end{aligned}
$$

It is obvious that $H \geq K$ and usually $G \geq H$. To make the notation more concise, $c(X)$ will be abbreviated as c and $Y(0)$ as 0 in subscripts. For example, $H_{c,0} = H_{c(X),Y(0)}$, $G_{c,1} = G_{c(X),Y(1)}$ and $K_c = K_{c(X),c(X)}$.

Theorem 2. *Assume the logistic regression model of propensity score is correct, i. e.* $\text{logit}(\text{P}(T = 1|X))$ *is a linear combination of* $\{c_j(X)\}_{j=1}^p$. *Let* $\pi = \text{P}(T = 1)$, *then we have* $\hat{\gamma}^{\text{EB}} \xrightarrow{d} \text{N}(\gamma, V^{\text{EB}}/n)$ *and* $\hat{\gamma}^{\text{IPW}} \xrightarrow{d} \text{N}(\gamma, V^{\text{IPW}}/n)$ *where*

$$
V^{\text{EB}} = \pi^{-1} \cdot \left\{ H_1 + G_0 - H_{c,0}^T H_c^{-1} \left(2G_{c,0} - H_{c,0} - G_c H_c^{-1} H_{c,0} + 2H_{c,1} \right) \right\}, \tag{12}
$$

$$
V^{\text{IPW}} = \pi^{-1} \cdot \left\{ H_1 + G_0 - H_{c,0}^T K_c^{-1} \left(H_{c,0} - 2K_{c,0}^m + 2K_{c,1}^m \right) \right\}. \tag{13}
$$

The proof of Theorem 2 is given in Appendix A. The H, G and K matrices in Theorem 2 can be estimated from the observed data, yielding approximate sampling variances for $\hat{\gamma}^{EB}$ and $\hat{\gamma}^{IPW}$. Alternatively, variance estimates may be obtained via the empirical sandwich method (e. g. [35]). In practice (particularly in simulations where we compare to a known truth), we find that the empirical sandwich method is more stable than the plug-in method, which is consistent with the suggestion in Lunceford and Davidian [20] for PATE estimators.

To complete the proof of the second claim in Theorem 1, we compare these variances with the semiparametric variance bound of γ with unknown $e(X)$ derived by Hahn ([8], Theorem 1):

$$
V^* = \frac{1}{\pi^2} \text{E}\left[e(X)\text{Var}(Y(1)|X) + \frac{e(X)^2}{1-e(X)} \text{Var}(Y(0)|X) + e(X)(g_1(X) - g_0(X) - \gamma)^2 \right]
$$

After some algebra, one can express V^* in terms of $H_{\cdot,\cdot}$ and $G_{\cdot,\cdot}$ defined above:

$$
V^* = \pi^{-1} \cdot \left\{ H_1 + G_0 - 2H_{g_0,g_1} - G_{g_0} + H_{g_0} \right\}.
$$

Now assume $\text{logit}(\text{P}(T = 1|X)) = \theta^T c(X)$ and $\text{E}[Y(t)|X] = \beta(t)^T c(X), t = 0, 1$, it is easy to verify that

$$
\begin{aligned}
H_{c,t} &= \text{Cov}(c(X), Y(t)|T = 1) \\
&= \text{Cov}(c(X), \beta(t)^T c(X)|T = 1) \\
&= H_c \beta(t), \text{for} t = 0, 1.
\end{aligned}
$$

Similarly, $G_{c,t} = G_c\beta(t), t = 0, 1$. From here it is easy to check V^{EB} and V^* are the same. Since Entropy Balancing reaches the efficiency bound in this case, obviously $V^{EB} < V^{IPW}$ when both models are linear.

If $\text{logit}(\text{P}(T = 1|X)) = \theta^T c(X)$ is true but $\text{E}[Y(t)|X] = \beta(t)^T c(X)$ is not true for some $t = 0, 1$, there is no guarantee that EB has the smaller asymptotic variance. In practice, the features $c(X)$ in the outcome regression models are almost always correlated with Y. This correlation compensates the slight efficiency loss of not maximizing the likelihood function in logistic regression. As a consequence, the variance V^{EB} in 12 is usually smaller than V^{IPW} in 13. This efficiency advantage of EB over IPW is verified in the next section using simulations.

5 Simulations

5.1 Kang-Schafer example

We use the simulation example in Kang and Schafer [9] to compare EB weighting with IPW (after maximum likelihood logistic regression) and the over-identified Covariate Balancing Propensity Score (CBPS) proposed by Imai and Ratkovic [13]. The simulated data consist of $\{X_i, Z_i, T_i, Y_i\}, i = 1, \ldots, n\}$. X_i and T_i are always observed, Y_i is observed only if $T_i = 1$, and Z_i is never observed. To generate this data set, X_i is distributed as $N(0, I_4)$, Z_i is computed by first applying the following transformation:

$$
\begin{aligned}
Z_{i1} &= \exp(X_{i1}/2), \\
Z_{i2} &= X_{i2}/(1 + \exp(X_{i1})) + 10, \\
Z_{i3} &= (X_{i1}X_{i3} + 0.6)^3, \\
Z_{i4} &= (X_{i2} + X_{i4} + 20)^2.
\end{aligned}
$$

Next we normalize each column such that Z_i has mean 0 and standard deviation 1.

In one setting, Y_i is generated by $Y_i = 210 + 27.4X_{i1} + 13.7X_{i2} + 13.7X_{i3} + 13.7X_{i4} + \varepsilon_i, \varepsilon_i \sim N(0, 1)$ and the true propensity scores are $e_i = \text{expit}(-X_{i1} + 0.5X_{i2} - 0.25X_{i3} - 0.1X_{i4})$. In this case, both Y and T can be correctly modeled by (generalized) linear model of the observed covariates X.

In the other settings, at least one of the propensity score model and the outcome regression model is incorrect. In order to achieve this, the data generating process described above is altered such that Y or T (or both) is linear in the unobserved Z instead of the observed X, though the parameters are kept the same.

For each setting (4 in total), we generated 1000 simulated data sets of size $n = 200$ and 1000, then apply various methods discussed earlier including

(1) IPW, CBPS: the IPW estimator in 2 with propensity score estimated by logistic regression or CBPS (since the estimand is overall mean, we use the CBPS weights tailored for estimating PATE);

(2) EB: the Entropy Balancing estimator (the EB weights are used to estimate the unobserved mean $E[Y|T = 0]$);

(3) IPW+DR, CBPS+DR: the doubly robust estimator in 10 with propensity score estimated by logistic regression or CBPS.

The simulation results are presented in Figure 2 and Figure 3. Figure 2 and Figure 3 show the covariate imbalance before adjustment in terms of standardized difference. Figure 2 and Figure 3 show the mean estimates given by the five different methods. First, notice that the doubly robust estimator "IPW+DR" performs poorly when both models are misspecified (bottom-right panel in Figure 2) and Figure 3. In fact, all the three doubly robust methods are worse than just using IPW. Second, the three doubly robust estimators have exactly the same variance if the Y model is correct (top two panels in Figure 2 and Figure 4). It seems that how one fits the propensity score model has no impact on the final estimate. This is related to the observation in Kang and Schafer [9] that, in this example, the plain OLS estimate of Y actually outperforms any method involving the propensity score model. Discussion articles such as Robins et al. [34] and Ridgeway and McCaffrey [36] find this phenomenon very uncommon in practice and is most likely due to the estimated inverse probability weights are highly variable, which is a bad setting for doubly robust estimators.
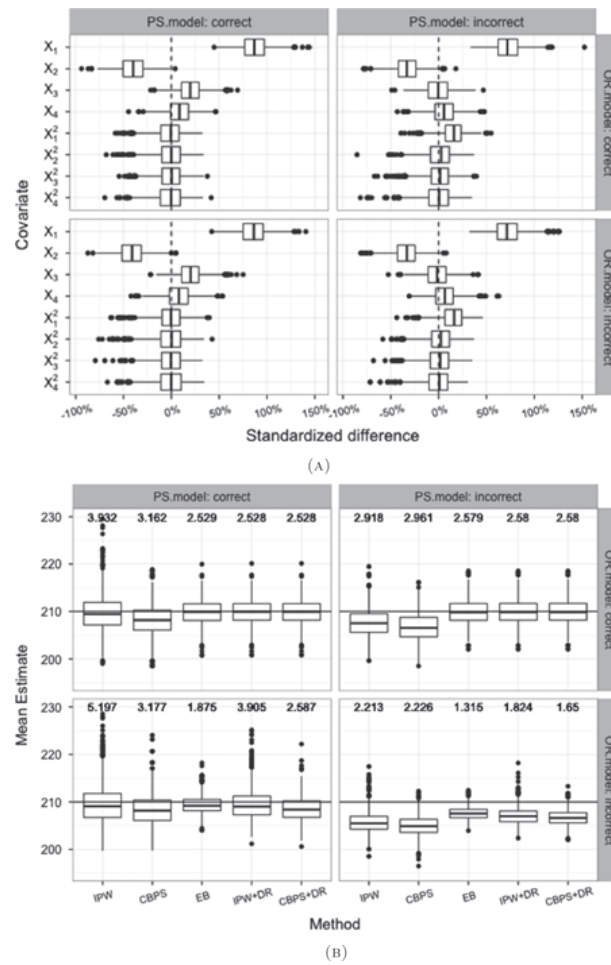
Figure 2: Kang-Schafer example: sample size n = 200. Both propensity score model and outcome regression model can be correct or incorrect, so there are four scenarios in total. We generate 1,000 simulations in each scenario. (a) Covariate imbalance before adjustment. (b) Mean estimates. The methods are: Inverse Propensity Weighting (IPW), Covariate Balancing Propensity Score (CBPS), Entropy Balancing (EB), and doubly robust versions of the first two (IPW+DR, CBPS+DR). Target mean is 210 and is marked as a black horizontal line to compare the biases. Numbers printed at Y = 230 are the sample standard deviations to compare efficiency.
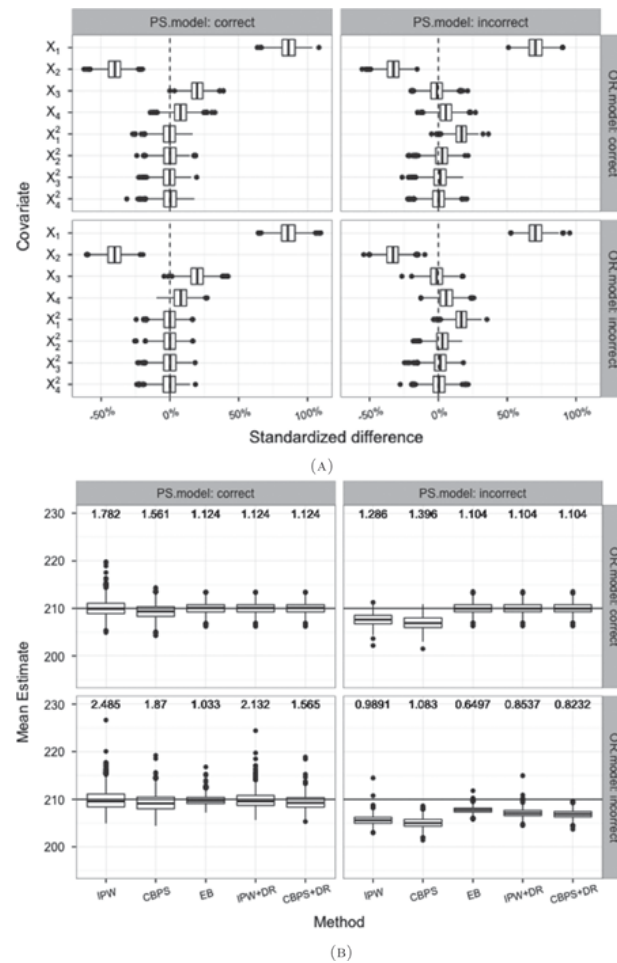
(A)

(B)

Figure 3: Kang-Schafer example: sample size $n = 200$. Both propensity score model and outcome regression model can be correct or incorrect, so there are four scenarios in total. We generate 1,000 simulations in each scenario. (a) Covariate imbalance before adjustment. (b) Mean estimates. The methods are: Inverse Propensity Weighting (IPW), Covariate Balancing Propensity Score (CBPS), Entropy Balancing (EB), and doubly robust versions of the first two (IPW+DR, CBPS+DR). Target mean is 210 and is marked as a black horizontal line to compare the biases. Numbers printed at Y = 230 are the sample standard deviations to compare efficiency.

Regarding Entropy Balancing (EB), we find that:

(1) If both T and Y models are misspecified, EB has smaller bias than the conventional "IPW+DR" or "CBPS+DR". So EB seems to be less affected by such unfavorable setting.

(2) When T model is correct but Y model is wrong (bottom-left panel in Figure 2 and Figure 3), EB has the smallest variance among all estimators. This supports the conclusion of our efficiency comparison of IPW and EB in Section 4.

Finally notice that the same simulation setting is used in Tan [18] to study the performance of a number of doubly robust estimators. The reader can compare the Figure 2 and Figure 3 with the results there. The performance of Entropy Balancing is comparable to the best estimator in Tan [18].

5.2 Lunceford-Davidian example

We provide another simulation example by Lunceford and Davidian [20] to verify claims in Theorems 1 and 2. In this simulation, the data still consist of $\{(X_i, Z_i, T_i, Y_i), i = 1, \ldots, n\}$, but all of them are observed. Both X_i and Z_i are three dimensional vectors. The propensity score is only related to X through:

$$\text{logit}(P(T_i = 1)) = \beta_0 + \sum_{j=1} \beta_j X_{ij}.$$

Note the above does not involve elements of Z_i. The response Y is generated according to

$$Y_i = \nu_0 + \sum_{j=1}^{3} \nu_j X_{ij} + \nu_4 T_i + \sum_{j=1}^{3} \xi_j Z_{ij} + \varepsilon_i; \ \varepsilon_i \sim \mathrm{N}(0,1).$$

The parameters here are set to be $\nu = (0, -1, 1, -1, 2)^T$, and β is set as:

$$\beta^{no} = (0, 0, 0, 0)^T,$$
$$\beta^{moderate} = (0, 0.3, -0.3, 0.3)^T, \ or$$
$$\beta^{moderate} = (0, 0.3, -0.3, 0.3)^T, \ or$$

The choice of β depends on the level of association of T and X. ξ is based on a similar choice on the level of association of Y and Z:

$$\xi^{no} = (0, 0, 0)^T,$$
$$\xi^{moderate} = (-0.5, 0.5, 0.5)^T, \ or$$
$$\xi^{strong} = (-1, 1, 1)^T.$$

The joint distribution of (X_i, Z_i) is specified by taking $X_{i3} \sim$ Bernoulli(0.2) and then generate Z_{i3} as Bernoulli with

$$\mathrm{P}\,(Z_{i3} = 1 | X_{i3}) = 0.75 X_{i3} + 0.25(1 - X_{i3}).$$

Conditional on X_{i3}, $(X_{i1}, Z_{i1}, X_{i2}, Z_{i2})$ is then generated as multivariate normal $\mathrm{N}(a_{X_{i3}}, B_{X_{i3}})$, where $a_1 = (1, 1, -1, -1)^T$, $a_0 = (-1, -1, 1, 1)^T$ and

$$B_0 = B_1 = \begin{pmatrix} 1 & 0.5 & -0.5 & -0.5 \\ 0.5 & 1 & -0.5 & -0.5 \\ -0.5 & -0.5 & 1 & 0.5 \\ -0.5 & -0.5 & 0.5 & 1 \end{pmatrix}.$$

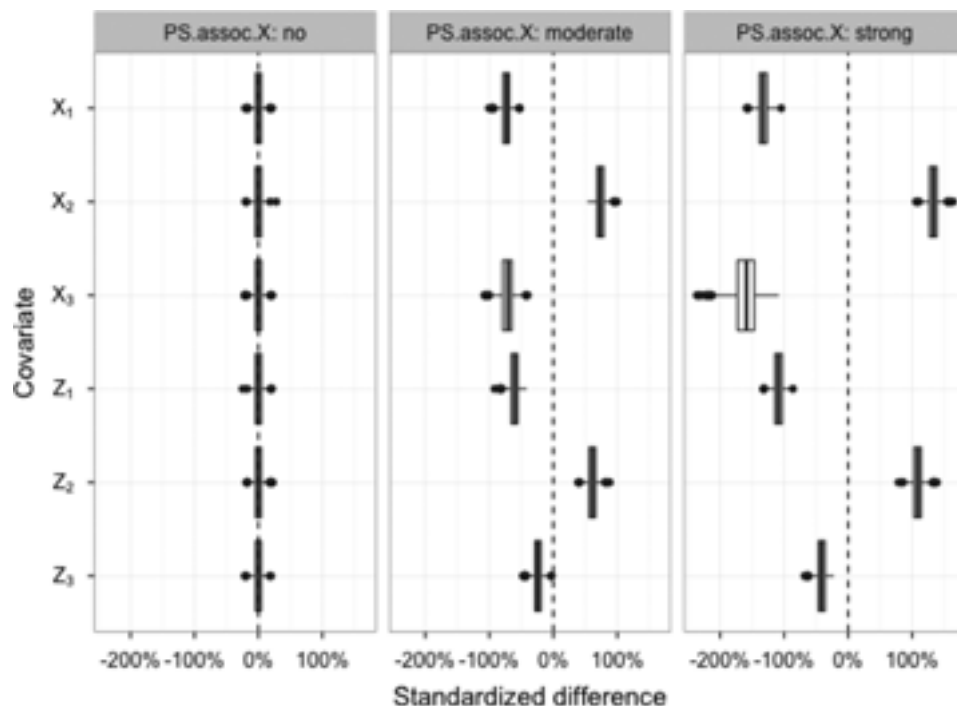Figure 4 shows the covariate imbalance in the three settings before adjustment.

Figure 4: Lunceford-Davidian example: covariate imbalance before adjustment.

The data generating model implies that the true PATT is $\gamma = 2$. Since the outcome Y depends on both X and Z, we always fit a full linear model of Y using X and Z, if such model is needed. T only depends on X, so it is not necessary to include Z in propensity score modeling. However, as pointed out by Lunceford and Davidian [20], it is actually beneficial to "overmodel" the propensity score by including Z in the model. Here we will try both possibilities, the "full" modeling of T using both X and Z, and the "partial" modeling of T using only X. Since the estimand is PATT in this case, we use the over-identified CBPS weights tailored for estimating PATT.

We generated 1000 simulated data sets and the results are shown in Figure 4 for "full" propensity score modeling and Figure 6 for "partial" propensity score modeling. We make the following comments about these two plots:

(1) IPW and all other estimators are always consistent, no matter what level of association is specified. This is because the propensity score model is always correctly specified.

(2) When using the "full" propensity score modeling, all doubly robust estimators (EB, IPW+DR, CBPS+DR and EB+DR) have almost the same sample variance. This is because all of them are asymptotically efficient.

(3) CBPS, to our surprise, does not perform very well in this simulation. It has smaller variance than IPW but this comes with the price of some bias. If we use the partial propensity score model (only involve X, Figure 6), this bias is smaller but still not negligible. While it is not clear what causes this bias, one possible reason is that the optimization problem of CBPS is nonconvex, so the local solution which is used to construct γ estimator could be far from the global solution. Another possibility is that CBPS uses GMM or Empirical Likelihood to combine likelihood with imbalance penalty, which is less efficient than maximum likelihood directly. Thus, although the estimator is asymptotically unbiased, the convergence spend to the true γ is quite slower than IPW. CBPS combined with outcome regression (CB+DR) fixes the bias and inefficiency issue occurred in CBPS without outcome regression.

(4) EB, in contrast, performs quite well in this simulation. It has relatively small variance, particularly if we use the "full" model in which both X and Z are balanced.

(5) The difference between EB and EB+DR is that while EB only balances "partial" or "full" covariates, EB+DR additionally combines a outcome linear regression model on all the covariates. As shown in the first proof of Theorem 1, when the "full" covariates are used, EB is exactly the same as EB+DR. We can observe this from Figure 4. When EB only balances "partial" covariates, the two methods are different and indeed EB+DR is more efficient in Figure 6 since it fits the correct Y model.

(6) Using the "full" propensity score model improves the efficiency of pure weighting estimators (IPW, CBPS and EB) a lot, but has very little impact on estimators that involves an outcome regression model (IPW+DR and CBPS+DR) compared to "partial" propensity score modeling. Although EB could be viewed as fitting a outcome model implicitly, the "partial" EB estimator only uses X in the outcome model, that is precisely the reason why it is not efficient. Thus there are both robustness and efficiency reasons that one should include all relevant covariates in EB, even if the covariates affect only one of T and Y.
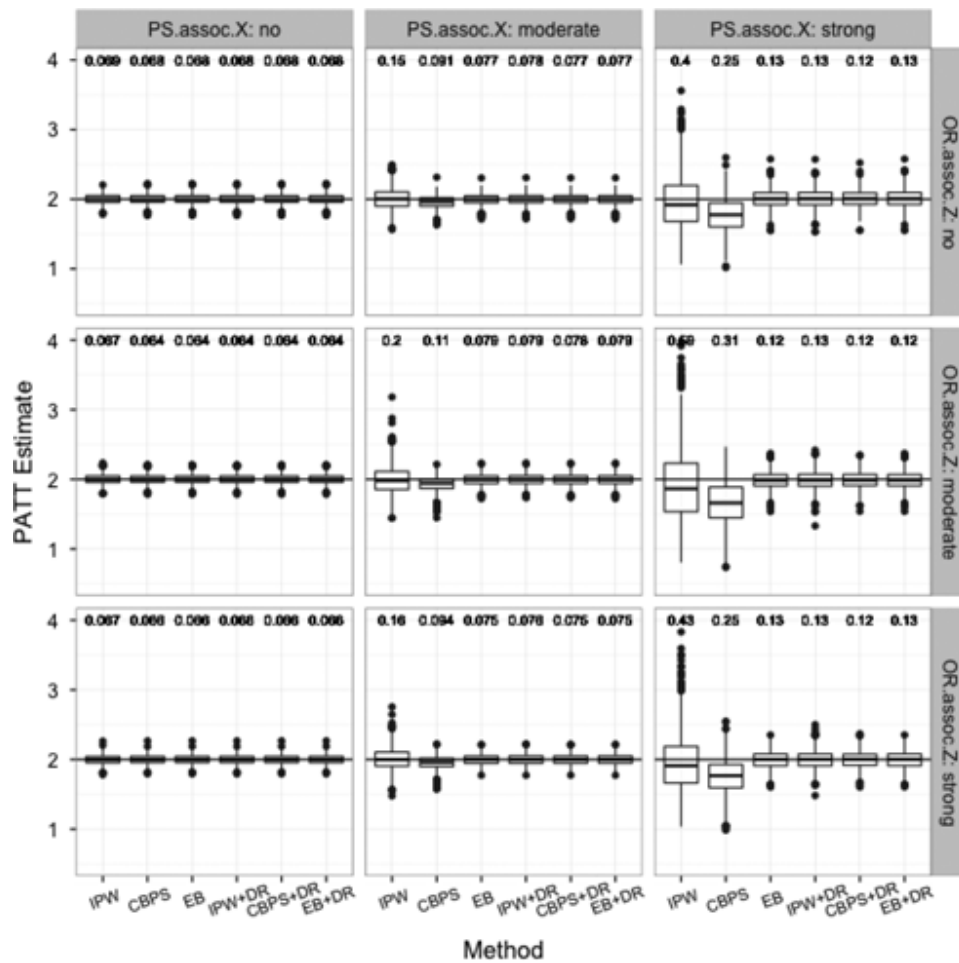
Figure 5: Results of the Lunceford-Davidian example (full propensity score modeling). The propensity score model and outcome regression model, if applies, are always correctly specified, but the level of association between T or Y with X or Z could be different, ended up with 9 different scenarios. X are confounding covariates and Z only affects the outcome. We generate 1,000 simulations of 1,000 in each scenario and apply five different estimators. The true PATT is 2 and is marked as a black horizontal line to compare the biases of the methods. Numbers printed at Y = 5 are the sample standard deviation of each method, in order to compare their efficiency.
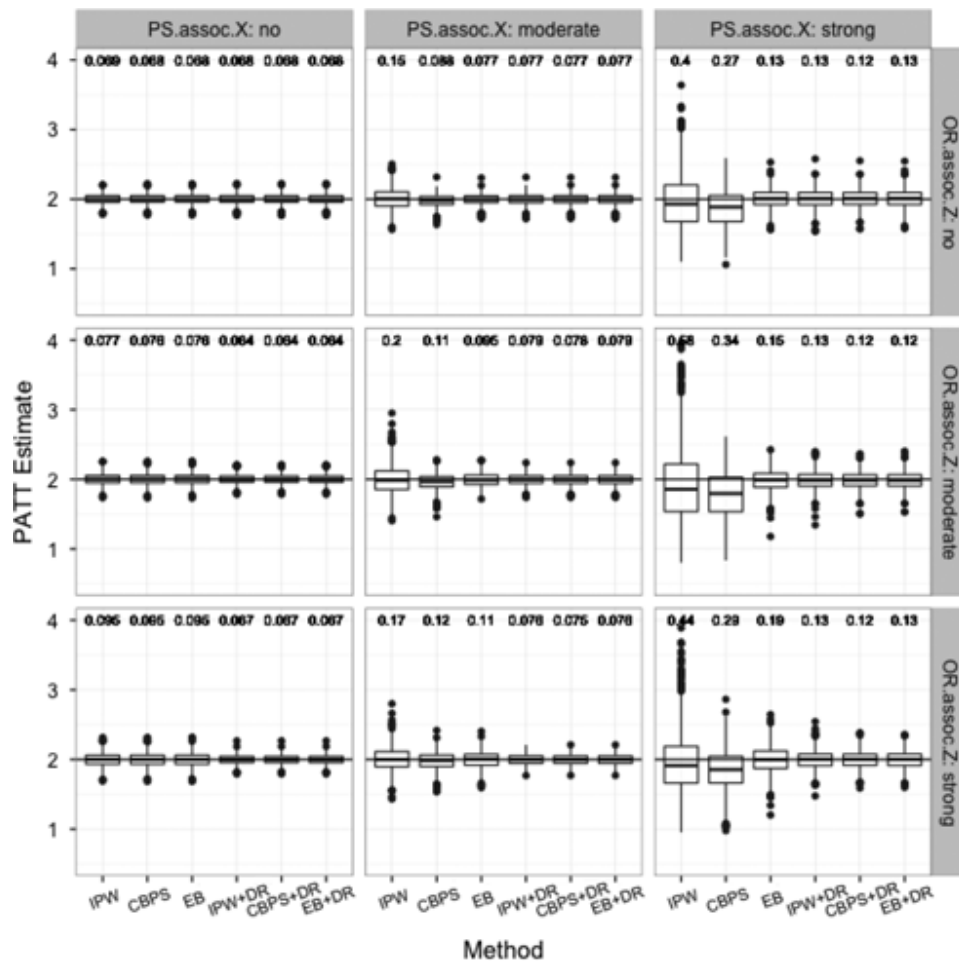
Figure 6: Results of the Lunceford-Davidian example (partial propensity score modeling). The settings are exactly the same as Figure 4 except the methods here don't use Z in their propensity score models.

In summary, EB outperforms IPW in all the simulations, making it an appealing alternative to the conventional propensity score weighting methods.

Acknowledgement

Appendix A: Theoretical proofs

We first describe the conditions under which the EB problem 3 admits a solution. The existence of w^{EB} depends on the solvability of the moment matching constraints

$$\sum_{T_i=0} w_i c_j(X_i) = \bar{c}_j(1), j = 1, \ldots, p, w > 0, \sum_{T_i=0} w_i = 1. \tag{14}$$

As one may expect, this is closely related to the existence condition of maximum likelihood estimate of logistic regression [37, 38]. An easy way to obtain such condition is through the dual problem of 8

$$\begin{aligned}
&\underset{w}{\text{maximize}} && -\sum_{i=1}^{n} \left[w_i \log w_i + (1 - w_i) \log(1 - w_i) \right] \\
&\text{subject to} && \sum_{T_i=0} w_i c_j(X_i) = \sum_{T_i=1} w_i c_j(X_i), j = 1, \dots, p, \\
& && 0 < w_i < 1, i = 1, \dots, n.
\end{aligned} \tag{15}$$

Thus, the existence of $\hat{\theta}^{MLE}$ is equivalent to the solvability of the constraints in 15, which is the overlap condition first given by Silvapulle [37].

Intuitively, in the space of $c(X)$, the solvability of 14 or the existence of w^{EB} means there is no hyperplane separating $\{c(X_i)\}_{T_i=0}$ and $\bar{c}(1)$. In contrast, the solvability of 15 or the existence of w^{MLE} means there is no hyperplane separating $\{c(X_i)\}_{T_i=0}$ and $\{c(X_i)\}_{T_i=1}$. Hence the existence of EB requires a stronger condition than the logistic regression MLE.

The next proposition suggests that the existence of w^{EB} and hence w^{MLE} is guaranteed by Assumption 2 (overlap) with high probability.

Proposition 1. *Suppose Assumption 2 (overlap) is satisfied and the expectation of $c(X)$ exist, then $P(w^{EB} \text{exists}) \to 1$ as $n \to \infty$. Furthermore, $\sum_{i=1}^{n} (w_i^{EB})^2 \to 0$ in probability as $n \to \infty$.*

Proof. Since the expectation of $c(X)$ exist, the weak law of large number says $\bar{c}(1) \xrightarrow{p} \bar{c}^*(1) = E[c(X)|T = 1]$. Therefore

Lemma 1. *For any $\varepsilon > 0$, $P(\| \bar{c}(1) - \bar{c}^*(1) \|_\infty \geq \varepsilon) \to 0$ as $n \to \infty$.*

Now condition on $\| \bar{c}(1) - \bar{c}^*(1) \|_\infty \geq \varepsilon$, i. e. $\bar{c}(1)$ is in the box of side length 2ε centered at $\bar{c}^*(1)$, we want to prove that with probability going to 1 there exists w such that $w_i > 0$, $\sum_{T_i=0} w_i = 1$ and $\sum_{T_i=0} w_i c(X_i) = \bar{c}(1)$. Equivalently, this is saying the convex hull generated by $\{c(X_i)\}_{T_i=0}$ contains $\bar{c}(1)$. We indeed prove a stronger result:

Lemma 2. *With probability going to 1 the convex hull generated by $\{c(X_i)\}_{T_i=0}$ contains the box $B_\varepsilon(\bar{c}^*(1)) = \{c(x) :\| c(x) - \bar{c}^*(1) \|_\infty \leq \varepsilon\}$ for some $\varepsilon > 0$.*

Proposition 1 follow immediately from Lemma 1 and Lemma 2. Now we prove Lemma 2. Denote the sample space of X by $\Omega(X)$. Assumption 2 (overlap) implies $\bar{c}^*(1)$ hence $B_\varepsilon(\bar{c}^*(1))$ is in the interior of the convex hull of $\Omega(X)$ for sufficiently small ε. Let $R_i, i = 1, \dots, 3^p$, be the 3^p boxes centered at $\bar{c}^*(1) + \frac{3}{2}\varepsilon b$, where $b \in {}^p$ is a vector that each entry can be -1, 0, or 1. It is easy to check that the sets R_i are disjoint and the convex hull of $\{x_i\}_{i=1}^{3^p}$ contains $B_\varepsilon(\bar{c}^*(1))$ if $x_i \in R_i, i = 1, \dots, 3^p$. Since $0 < P(T = 0|X) < 1$, $\rho = \min_i P(X \in R_i|T = 0) > 0$. This implies

$$\begin{aligned}
P(\exists X_i \in R_i \text{and} T_i = 0, \forall i = 1, \dots, 3^p) &\geq 1 - \sum_{i=1}^{3^p} P(XR_i|T = 0)^n \\
&\geq 1 - 3^p(1 - \rho)^n \\
&\to 1
\end{aligned} \tag{16}$$

as $n \to \infty$. This proves the lemma because the event in the left hand side implies the convex hull generated by $\{c(X_i)\}_{T_i=0}$ contains the desired box. Note that 16 also tells us how many samples we actually need to ensure the existence of w^{EB}. Indeed if $n \geq \rho^{-1}(p \log 2 + \log \delta^{-1}) \geq \log_{(1-\rho)}(\delta 2^{-p})$, then the probability in 16 is greater than $1 - \delta$. Usually we expect $\delta = O(3^{-p})$. If this is the case, the number of samples needed is $n = O(p \cdot 3^p)$. [1]

Now we turn to the second claim of the proposition, i. e. $\sum_{T_i=0} w_i^2 \xrightarrow{p} 0$. To prove this, we only need to find a sequence (with respect to growing n) of feasible solutions to 3 such that $\max_i w_i \to 0$. This is not hard to show, because the probability in 16 is exponentially decaying as n increases. We can pick $n_1 \geq N(\delta, p, \rho)$ such that the probability of the convex hull of $\{x_i\}_{i=1}^{n_1}$ contains $B_\varepsilon(\bar{c}^*(1))$ is at least $1 - \delta$, then pick $n_{i+1} \geq n_i + 3^i N(\delta, p, \rho)$ so the convex hull of $\{x_i\}_{i=n_i+1}^{n_{i+1}}$ contains $B_\varepsilon(\bar{c}^*(1))$ with probability at least $1 - 3^i \delta$. This means for each $\{x_i\}_{i=n_i+1}^{n_{i+1}}, i = 0, 1, \dots$, we have a set of weights $\{\tilde{w}_i\}_{i=n_i+1}^{n_{i+1}}$ such that $\sum_{i=n_i+1}^{n_{i+1}} \tilde{w}_i x_i = \bar{c}(1)$. Now suppose $n_k \leq n < n_{k+1}$, the choice $w_i = \tilde{w}_i/k$ if $i \leq n_k$ and $w_i = 0$ if $i > n_k$ satisfies the constraints and $\max_i w_i \leq k$. As $n \to \infty$, this implies $\max_i w_i \to 0$ and hence $\sum_i w_i^2 \to 0$ with probability tending to 1. □

Now we turn to the main theorem of the paper (Theorem 1). The first claim in Theorem 1 follows immediately from the following lemma:

Lemma 3. *Under the assumptions in Theorem 1 and suppose* $\text{logit}(P[T = 1|X]) = \sum_{j=1}^{p} \theta_j^* c_j(X)$, *then as* $n \to \infty$, $\hat{\theta}^{\text{EB}} \xrightarrow{p} \theta^*$. *As a consequence,*

$$\text{E}\left[\sum_{T_i=0} w_i^{\text{EB}} Y_i\right] \xrightarrow{p} \text{E}[Y(0)|T = 1].$$

Proof. The proof is a standard application of M-estimation [35] theory. We will follow the estimating equations approach described in [35] to derive consistency of $\hat{\theta}^{EB}$. First we note the first order optimality condition of 7 is

$$\sum_{i=1}^{n} (1 - T_i) e^{\sum_{k=1}^{p} \theta_k c_k(X_i)} (c_j(X_i) - \bar{c}_j(1)) = 0, j = 1, \dots, R. \tag{17}$$

We can rewrite 17 as estimating equations. Let $\phi_j(X, T; m) = T(c_j(X) - m_j), j = 1, \dots, R$ and $\psi_j(X, T; \theta, m) = (1 - T) \exp\{\sum_{k=1}^{p} \theta_k c_k(X)\}(c_j(X) - m_j)$, then 17 is equivalent to

$$\sum_{i=1}^{n} \phi_j(X_i, T_i; m) = 0, j = 1, \dots, R,$$
$$\sum_{i=1}^{n} \psi_j(X_i, T_i; \theta, m) = 0, j = 1, \dots, R. \tag{18}$$

Since $\phi(\cdot)$ and $\psi(\cdot)$ are all smooth functions of θ and m, all we need to verify is that $m_j^* = \text{E}[c_j(X)|T = 1]$ and θ^* is the unique solution to the population version of 18. It is obvious that m^* is the solution to $\text{E}[\phi_j(X, T; m)] = 0, j = 1, \dots, R$. Now take conditional expectation of ψ_j given X:

$$
\begin{aligned}
\text{E}[\psi_j(X, T; \theta, m^*)|X] \quad &= (1 - e(X)) e^{\sum_{k=1}^{p} \theta_k c_k(X)} (c_j(X) - m_j^*) \\
&= \left(1 - \frac{e^{\sum_{k=1}^{p} \theta_k^* c_k(X)}}{1 + e^{\sum_{k=1}^{p} \theta_k^* c_k(X)}}\right) e^{\sum_{k=1}^{p} \theta_k c_k(X)} (c_j(X) - m_j^*) \\
&= \frac{e^{\sum_{k=1}^{p} \theta_k c_k(X)}}{1 + e^{\sum_{k=1}^{p} \theta_k^* c_k(X)}} (c_j(X) - \text{E}[c_j(X)|T = 1]).
\end{aligned}
$$

The only way to make $\text{E}[\psi_j(X, T; \hat{\theta}, m^*)] = 0$ is to have

$$\frac{e^{\sum_{k=1}^{p} \hat{\theta}_k c_k(X)}}{1 + e^{\sum_{k=1}^{p} \theta_k^* c_k(X)}} = \text{const} \cdot P(T = 1|X),$$

i. e. $\hat{\theta} = \theta^*$. This proves the consistency of $\hat{\theta}^{EB}$.

The consistency of $\hat{\gamma}^{EB}$ is proved by noticing

$$w_i^{EB} = \frac{\exp(\sum_{j=1}^{p} \hat{\theta}_j^{EB} c_j(X_i))}{\sum_{T_i=0} \exp(\sum_{j=1}^{p} \hat{\theta}_j^{EB} c_j(X_i))} \xrightarrow{p} \frac{P(T_i = 1|X_i)}{1 - P(T_i = 1|X_i)},$$

which is the IPW-NR weight defined in 2.

The second claim is a corollary of Theorem 2, which is proved below. For simplicity we denote $\xi = (m^T, \theta^T, \mu(1|1), \gamma)^T$ and the true parameter as ξ^*. Throughout this section we assume $\text{logit}(e(X)) = \sum_{j=1}^{p} \theta_j^* c_j(X)$. Denote $\tilde{c}(X) = c(X) - \bar{c}^*(1)$, $e^*(X) = e(X; \theta^*)$, $l^*(X) = \exp\{\sum_{j=1}^{p} \theta_j^* c_j(X)\} = e^*(X)/(1 - e^*(X))$. Let

$$\phi_j(X,T;m) = T(c_j(X) - m_j), j = 1, \ldots, p,$$

$$\psi_j(X,T;\theta,m) = (1-T)e^{\sum_{k=1}^{p} \theta_k c_k(X)}(c_j(X) - m_j), j = 1, \ldots, p,$$

$$\varphi_{1|1}(X,T,Y;\mu(1|1)) = T(Y - \mu(1|1)),$$

$$\varphi(X,T,Y;\theta,\mu(1|1),\gamma) = (1-T)e^{\sum_{j=1}^{p} \theta_j c_j(X)}(Y + \gamma - \mu(1|1)),$$

and $\zeta(X,T,Y;m,\theta,\mu(1|1),\gamma) = (\phi^T, \psi^T, \varphi_{1|1}, \varphi)^T$ be all the estimating equations. The Entropy Balancing estimator $\hat{\gamma}^{EB}$ is the solution to

$$\frac{1}{n}\sum_{i=1}^{n} \zeta(X_i, T_i, Y_i; m, \theta, \mu(1|1), \gamma) = 0. \tag{19}$$

There are two forms of "information" matrix that need to be computed. The first is

$$A^{EB}(\xi^*) = \mathrm{E}\left[-\frac{\partial}{\partial \xi^T}\zeta(X,T,Y;\xi^*)\right]$$

$$= \left(\mathrm{E}\left[-\frac{\partial}{\partial m^T}\zeta(\xi^*)\right] \mathrm{E}\left[-\frac{\partial}{\partial \theta^T}\zeta(\xi^*)\right] \mathrm{E}\left[-\frac{\partial}{\partial \mu(1|1)}\zeta(\xi^*)\right] \mathrm{E}\left[-\frac{\partial}{\partial \gamma}\zeta(\xi^*)\right]\right)$$

$$= \mathrm{E}\left[\begin{pmatrix} T \cdot I_R & 0 & 0 & 0 \\ (1-T)l^*(X) \cdot I_R & -(1-T)l^*(X)(c(X) - \bar{c}^*(1))c(X)^T & 0 & 0 \\ 0^T & 0^T & T & 0 \\ 0 & -(1-T)l^*(X)(Y(0) - \mu^*(0|1))c(X)^T & (1-T)l^*(X) & -(1-T)l^*(X) \end{pmatrix}\right]$$

$$= \pi \cdot \begin{pmatrix} I_R & 0 & 0 & 0 \\ I_R & -\mathrm{Cov}[c(X)|T=1] & 0 & 0 \\ 0^T & 0^T & 1 & 0 \\ 0 & -\mathrm{Cov}(Y(0), c(X)|T=1) & 1 & -1 \end{pmatrix}.$$

A very useful identity in the computation of the expectation is

$$\mathrm{E}[f(X,Y)|T=1] = \pi^{-1}\mathrm{E}[e(X)f(X,Y)]$$

$$= \frac{\mathrm{P}(T=0)}{\pi} \cdot \mathrm{E}\left[\frac{e(x)}{1-e(X)}f(X,Y)|T=0\right].$$

The second information matrix is the covariance of $\zeta(X,T,Y;\xi^*)$. Denote $\tilde{Y}(t) = Y(t) - \mu^*(t|1), t = 0,1$

$$B^{EB}(\xi^*) = \mathrm{E}[\zeta(X,Y,T;\xi^*)\zeta(X,Y,T;\xi^*)^T]$$

$$= \mathrm{E}\left[\begin{pmatrix} T\tilde{c}(X)\tilde{c}(X)^T & 0 & T\tilde{Y}(1)\tilde{c}(X) & 0 \\ 0 & (1-T)l^*(X)^2\tilde{c}(X)\tilde{c}(X)^T & 0 & (1-T)l^*(X)^2\tilde{Y}(0)\tilde{c}(X) \\ T\tilde{Y}(1)\tilde{c}(X)^T & 0^T & T\tilde{Y}^2(1) & 0 \\ 0 & (1-T)l^*(X)^2\tilde{Y}(0)\tilde{c}(X)^T & 0 & (1-T)l^*(X)^2\tilde{Y}^2(0) \end{pmatrix}\right].$$

The asymptotic distribution of $\hat{\gamma}^{EB}$ is $N(\gamma, V^{EB}(\xi^*)/n)$ where $V^{EB}(\xi^*)$ is the bottom right entry of $A^{EB}(\xi^*)^{-1}B^{EB}(\xi^*)A^{EB}(\xi^*)^{-T}$. Let's denote

$$H_{a_1, a_2} = \mathrm{Cov}(a_1, a_2|T=1),$$

$G_{a_1,a_2} = \mathrm{E}\left[l^*(X)(a_1 - \mathrm{E}[a_1|T=1])(a_2 - \mathrm{E}[a_2|T=1])^T|T=1\right]^T$, and $H_a = H_{a,a}, G_a = G_{a,a}$. So

$$A^{EB}(\xi^*) = \pi \cdot \begin{pmatrix} I_R & 0 & 0 & 0 \\ I_R & -H_{c(X)} & 0 & 0 \\ 0^T & 0^T & 1 & 0 \\ 0 & -H_{Y(0),c(X)} & 1 & -1 \end{pmatrix},$$

$$A^{EB}(\xi^*)^{-1} = \pi^{-1} \cdot \begin{pmatrix} I_R & 0 & 0 & 0 \\ H_{c(X)}^{-1} & -H_{c(X)}^{-1} & 0 & 0 \\ 0^T & 0^T & 1 & 0 \\ -H_{c(X),Y(0)}^T H_{c(X)}^{-1} & H_{c(X),Y(0)}^T H_{c(X)}^{-1} & 1 & -1 \end{pmatrix},$$

and

$$B^{EB}(\xi^*) = \pi \cdot \begin{pmatrix} H_{c(X)} & 0 & H_{c(X),Y(1)} & 0 \\ 0 & G_{c(X)} & 0 & G_{c(X),Y(0)} \\ H_{Y(1),c(X)} & 0^T & H_{Y(1)} & 0 \\ 0 & G_{Y(0),c(X)}^T & 0 & G_{Y(0)} \end{pmatrix}.$$

Thus

$$V^{EB} = \pi^{-1} \cdot \left\{ H_{c,0}^T H_c^{-1} \left(H_{c,0} + G_c H_c^{-1} H_{c,0} - 2G_{c,0} - 2H_{c,1} \right) + H_1 + G_0 \right\}.$$

It would be interesting to compare $V^{EB}(\xi^*)$ with $V^{IPW}(\xi^*)$, the asymptotic variance of $\hat{\gamma}^{IPW}$. The IPW PATT estimator 2 is equivalent to solving the following estimating equations

$$\sum_{i=1}^n \left(T_i - \frac{1}{1 + e^{-\sum_{k=1}^p \theta_k c_k(X_i)}} \right) c_j(X_i) = 0, r = 1, \ldots, R,$$

$$\frac{1}{n} \sum_{i=1}^n \varphi_{1|1}(X_i, T_i, Y_i; \theta, \mu(1|1), \gamma) = 0,$$

$$\frac{1}{n} \sum_{i=1}^n \varphi(X_i, T_i, Y_i; \theta, \mu(1|1), \gamma) = 0.$$

If we call $K_{a_1,a_2} = E[(1 - e(X))a_1 a_2^T | T = 1]$, we have

$$A^{IPW}(\xi^*) = E\left[\begin{pmatrix} e^*(X)(1-e^*(X))c(X)c(X)^T & 0 & 0 \\ 0^T & T & 0 \\ -(1-T)l^*(X)\tilde{Y}(0)c(X)^T & (1-T)l^*(X) & -(1-T)l^*(X) \end{pmatrix} \right]$$

$$= \pi \cdot \begin{pmatrix} K_{c(X)} & 0 & 0 \\ 0^T & 1 & 0 \\ -H_{Y(0),c(X)} & 1 & -1 \end{pmatrix}.$$

$$A^{IPW}(\xi^*)^{-1} = \pi^{-1} \cdot \begin{pmatrix} K_{c(X)}^{-1} & 0 & 0 \\ 0^T & 1 & 0 \\ -H_{Y(0),c(X)}K_{c(X)}^{-1} & 1 & -1 \end{pmatrix}.$$

Let $q^*(X) = e^*(X)l^*(X)$,

$$B^{IPW}(\xi^*) = E\left[\begin{pmatrix} (T-e^*(X))^2 c(X)c(X)^T & T(T-e^*(X))\tilde{Y}(1)c(X) & -(1-T)q^*(X)\tilde{Y}(0)c(X) \\ T(T-e^*(X))\tilde{Y}(1)c(X)^T & T\tilde{Y}^2(1) & 0 \\ -(1-T)q^*(X)\tilde{Y}(0)c(X) & 0 & (1-T)l^*(X)^2\tilde{Y}^2(0) \end{pmatrix} \right]$$

$$= \pi \cdot \begin{pmatrix} K_{c(X)} & K_{c(X),\tilde{Y}(1)} & K_{c(X),\tilde{Y}(0)} - H_{c(X),Y(0)} \\ K_{c(X),\tilde{Y}(1)}^T & H_{Y(1)} & 0 \\ K_{c(X),\tilde{Y}(0)}^T - H_{c(X),Y(0)}^T & 0 & G_{Y(0)} \end{pmatrix}.$$

V^{IPW} can thus be computed consequently and the details are omitted.

Notes

[1]Note that this naive rate can actually be greatly improved by Wendel's theorem in geometric probability theory.

References

[1] Rosenbaum P, Rubin D. The central role of the propensity score in observational studies for causal effects. Biometrika. 1983;70(1):41–55.

[2] Abadie A, Imbens GW. Large sample properties of matching estimators for average treatment effects. Econometrica. 2006;74(1):235–267.

[3] Rosenbaum PR, Rubin DB. Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. Am Stat. 1985;39(1):33–38.

[4] Rosenbaum P, Rubin D. Reducing bias in observational studies using subclassification on the propensity score. J Am Stat Assoc. 1984;79:516–524.

[5] Robins JM, Rotnitzky A, Zhao L. Estimation of regression coefficients when some regressors are not always observed. J Am Stat Assoc. 1994;89:846–866.

[6] Hirano K, Imbens G. Estimation of causal effects using propensity score weighting: an application to data on right heart catheterization. Health Serv Outcomes Res Method. 2001;2:259–278.

[7] Hirano K, Imbens GW, Ridder G. Efficient estimation of average treatment effects using the estimated propensity score. Econometrica. 2003;71(4):1161–1189.

[8] Hahn J. On the role of the propensity score in efficient semiparametric estimation of average treatment effects. Econometrica. 1998;66(2):315–332.

[9] Kang JD, Schafer JL. Demystifying double robustness: a comparison of alternative strategies for estimating a population mean from incomplete data. Stat Sci. 2007;22(4):523–539.

[10] Imai K, King G, Stuart EA. Misunderstandings between experimentalists and observationalists about causal inference. J Royal Stat Soc Ser A (Statistics in Society). 2008;171(2):481–502.

[11] Diamond A, Sekhon JS. Genetic matching for estimating causal effects: a general multivariate matching method for achieving balance in observational studies. Rev Econ Stat. 2013;95(3):932–945.

[12] Graham BS, Pinto CC, Egel D. Inverse probability tilting for moment condition models with missing data. Rev Econ Stud. 2012;79(3):1053–1079.

[13] Imai K, Ratkovic M. Covariate balancing propensity score. J Royal Stat Soc Ser B (Statistical Methodology). 2014;76(1):243–263.

[14] Zubizarreta JR. Stable weights that balance covariates for estimation with incomplete outcome data. J Am Stat Assoc. 2015;110(511):910–922.

[15] Hainmueller J. Entropy balancing for causal effects: a multivariate reweighting method to produce balanced samples in observational studies. Political Anal. 2011;20:25–46.

[16] Ferwerda J. Electoral consequences of declining participation: a natural experiment in Austria. Elect Stud. 2014;35:242–252.

[17] Marcus J. The effect of unemployment on the mental health of spouses evidence from plant closures in Germany. J Health Econ. 2013;32(3):546–558.

[18] Tan Z. Bounded, efficient and doubly robust estimation with inverse weighting. Biometrika. 2010;97(3):661–682.

[19] Bang H, Robins JM. Doubly robust estimation in missing data and causal inference models. Biometrics. 2005;61(4):962–973.

[20] Lunceford JK, Davidian M. Stratification and weighting via the propensity score in estimation of causal treatment effects: a comparative study. Stat Med. 2004;23(19):2937–2960.

[21] Särndal C-E, Lundström S. Estimation in surveys with nonresponse John Wiley & Sons, 2005.

[22] Deville J-C, Särndal C-E. Calibration estimators in survey sampling. J Am Stat Assoc. 1992;87(418):376–382.

[23] Chan KC, Yam SC, Zhang Z. Globally efficient non-parametric inference of average treatment effects by empirical balancing calibration weighting. J Royal Stat Soc Ser B (Statistical Methodology). 2016;78:673–700.

[24] Gneiting T, Raftery AE. Strictly proper scoring rules, prediction, and estimation. J Am Stat Assoc. 2007;102(477):359–378.

[25] Hastie TJ, Tibshirani RJ. Generalized additive models CRC Press, 1990 43.

[26] Zhao Q. Covariate balancing propensity score by tailored loss functions. 2016;http://arxiv.org/abs/1601.05890.

[27] Neyman J. Sur les applications de la thar des probabilities aux experiences agaricales: essay des principle. excerpts reprinted (1990) in English. Stat Sci. 1923;5:463–472.

[28] Rubin D. Estimating causal effects of treatments in randomized and nonrandomized studies. J Educ Psychol. 1974;66(5):688–701.

[29] Holland PW. Statistics and causal inference. J Am Stat Assoc. 1986;81:945–960.

[30] Cover TM, Thomas JA. Elements of information theory John Wiley & Sons, 2012.

[31] Qin J, Zhang B. Empirical-likelihood-based inference in missing response problems and its application in observational studies. J Royal Stat Soc Ser B (Statistical Methodology). 2007;69(1):101–122.

[32] Tan Z. A distributional approach for causal inference using propensity scores. J Am Stat Assoc. 2006;101:1619–1637.

[33] Wang Q, Rao JNK. Empirical likelihood-based inference under imputation for missing response data. Ann Stat. 2002;30(3):896–924.

[34] Robins J, Sued M, Lei-Gomez Q, Rotnitzky A. Comment: performance of double-robust estimators when inverse probability weights are highly variable. Stat Sci. 2007;22(4):544–559.

[35] Stefanski LA, Boos DD. The calculus of M-Estimation. Am Stat. 2002;56(1):29–38.

[36] Ridgeway G, McCaffrey DF. Comment: demystifying double robustness: a comparison of alternative strategies for estimating a population mean from incomplete data. Stat Sci. 2007;22(4):540–543.

[37] Silvapulle MJ. On the existence of maximum likelihood estimators for the binomial response models. J Royal Stat Soc Ser B (Methodological). 1981;43(3):310–313.

[38] Albert A, Andersen JA. On the existence of maximum likelihood estimates in logistic regression models. Biometrika. 1984;71(1):1–10.