Covariate balance is a conventional key diagnostic for methods used 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.

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 (1983) pointed 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

E-mail addresses: qyzhao@stanford.edu, dancsi@google.com.

Date: March 10, 2016.

Key words and phrases. Causal Inference, Double Robustness, Exponential Tilting, Convex Optimization, Survey Sampling.

We would like to thank Jens Hainmueller, Trevor Hastie, Hera He, William Heavlin, Diane Lambert, Daryl Pregibon, James Robins, Jean Steiner and one anonymous reviewer for their very helpful comments.
mean causal effect. The most popular approaches are matching (e.g. Rosenbaum and Rubin, 1985; Abadie and Imbens, 2006), stratification (e.g. Rosenbaum and Rubin, 1984), and weighting (e.g. Robins et al., 1994; Hirano and Imbens, 2001). Theoretically, propensity score weighting is the most attractive among these methods. Hirano et al. (2003) showed that nonparametric propensity score weighting can achieve the semiparametric efficiency bound for the estimation of mean causal effect derived by Hahn (1998). Another desirable property is double robustness. The pioneering work of Robins et al. (1994) pointed out that propensity score weighting can be further augmented 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, which is brought to attention by Kang and Schafer (2007). 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. (2008) 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 (e.g. Diamond and Sekhon, 2013; Imai and Ratkovic, 2014; Zubizarreta, 2015). In this paper, we study a method of this kind called Entropy Balancing (hereafter EB) that is proposed by Hainmueller (2011). 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 (Marcus, 2013; Ferwerda, 2014). However, little do we known about the theoretical properties of EB. The original proposal in Hainmueller (2011) 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 (Tan, 2010), 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. [Robins et al., 1994; Lunceford and Davidian, 2004; Bang and Robins, 2005; Kang and Schafer, 2007]). In contrast, Entropy Balancing achieves this goal through enforcing covariate balance. The primal optimization problem of EB amounts to an empirical calibration estimator (Deville and Särndal, 1992; Särndal and Lundström, 2005), which is widely popular in survey sampling but perhaps not sufficiently recognized in causal inference (Chan et al., 2015). 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 (2007)) 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 (Hastie and Tibshirani, 1990), can also easily be fitted. A forthcoming article by Zhao (2016) offers more discussion and extension to other weighted average treatment effects.

2. Setting

First, we fix some notations for the causal inference problem considered in this paper. We follow the potential outcome language due to [Neyman (1923); Rubin (1974)]. 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 copies from a population, for which we wish to infer the treatment’s effect. However, the “fundamental problem of causal inference” (Holland, 1986) is that only one potential outcome is observed: \( Y_i = T_i Y_i(1) - (1 - T_i) Y_i(0). \)

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

\[
\gamma = E[Y(1)|T = 1] - E[Y(0)|T = 1] \triangleq \mu(1|1) - \mu(0|1).
\]
The counterfactual mean $\mu(0|1) = \mathbb{E}[Y(0)|T = 1]$ also naturally occurs in survey sampling with missing data [Deville and Särndal (1992); Särndal and Lundström (2005)] if $Y(0)$ is the only outcome of interest and $T = 1$ stands for non-response.

Along with the treatment exposure $T_i$ and outcome $Y_i$, each experiment 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 selection bias. The seminal work by Rosenbaum and Rubin (1983) suggested it is possible to correct this selection 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 implies that the observed covariates contain all the information that may cause the selection bias, i.e. there is no confounding variable, and the second assumption ensures this bias-correction information is present across the entire domain of $X$.

Since the covariates $X$ contain all the information of selection 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. Rosenbaum and Rubin (1983) proved that under Assumptions 1 (strong ignorability) and 2 (overlap) $(Y(0), Y(1)) \perp T | e(X)$, i.e. the propensity score function $e(X)$ itself is sufficient to produce unbiased estimates of causal effect. 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 regression functions $g_0(X) = \mathbb{E}[Y(0)|X]$ and $g_1(X) = \mathbb{E}[Y(1)|X]$.

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

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

Here $\sum_{T_i=t}$ is defined as summation over all units $i$ such that $T_i = t$. This notation will be repeatedly used throughout the paper. In this formula, the control units are associated with weights proportional to $\hat{e}(X_i)(1 - \hat{e}(X_i))^{-1}$ to resemble the full population. The most popular choice of obtaining $\hat{e}(x)$ is via logistic regression, where $\text{logit}(e(x))$ 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 (2011) to estimate PATT. EB operates by maximizing the entropy of the
weights under some pre-specified balancing constraints:

\[
\begin{align*}
\text{maximize} & \quad - \sum_{T_i=0} w_i \log w_i \\
\text{subject to} & \quad \sum_{T_i=0} w_i c_j(X_i) = \tilde{c}_j(1) = \frac{1}{n_1} \sum_{T_i=1} c_j(X_i), \ j = 1, \ldots, p, \\
& \quad \sum_{T_i=0} w_i = 1, \\
& \quad w_i > 0, \ i = 1, \ldots, n.
\end{align*}
\]

(3)

Hainmueller (2011) proposes to use the weighted average \(\sum_{T_i=0} w_{EB}^i Y_i\) to estimate the counterfactual mean \(\mathbb{E}[Y(0)|T = 1]\). This gives the Entropy Balancing estimator of \(\text{PATT}\):

\[
\hat{\gamma}_{EB} = \sum_{T_i=1} Y_i - \sum_{T_i=0} w_{EB}^i Y_i.
\]

(4)

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

First, we give some heuristics allowing us to view EB as a propensity score weighting method. Since EB seeks to empirically match the control and treatment covariate distributions, we can draw connections between EB and density estimation. The 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 group by

\[
\begin{align*}
\text{maximize} & \quad H(\tilde{m}||m) \quad \text{subject to} \quad E_{\tilde{m}}[c(X)] = \tilde{c}(1), \\
& \quad H(\tilde{m}||m) = E_{\tilde{m}}[\log(\tilde{m}(X)/m(X))],
\end{align*}
\]

(5)

where \(H(\tilde{m}||m) = E_{\tilde{m}}[\log(\tilde{m}(X)/m(X))]\) 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 \(w(x) = [P(T = 1) \cdot \tilde{m}(x)]/[P(T = 0) \cdot m(x)]\) 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 \(w(x)\):

\[
\begin{align*}
\text{maximize} & \quad E_m[w(X) \log w(X)] \quad \text{subject to} \quad E_m[w(X)c(X)] = \tilde{c}(1). \\
& \quad E_m[w(X) \log w(X)] = E_m[w(X)c(X)] = \tilde{c}(1).
\end{align*}
\]

(6)

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

By using the Lagrangian multipliers, we can show the solution to (5) belongs to the family of exponential titled distributions of \(m\) (Cover and Thomas, 2012):

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

(6)

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

\[
\begin{align*}
e(x) &= \frac{P(T = 1|X = x)}{P(T = 0|X = x)} = w(x) = \exp(\alpha + \theta^T c(x)),
\end{align*}
\]

where \( \alpha = \log(P(T = 1)/P(T = 0)) \). This is exactly the logistic regression model of the propensity score using \( c(x) \) as predictors.

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

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

whereas the maximum likelihood solves

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

It is apparent from (7) and (8) that EB and maximum likelihood use different loss functions.

The optimization problem (7) can be shown to be strictly convex, and the unique solution \( \hat{\theta}_{EB} \) can be quickly 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 \( T_i = 0 \),

\[
\text{(9)} \quad w_{i}^{EB} = \frac{\exp \left( \sum_{j=1}^{p} \hat{\theta}_{EB}^j c_j(X_i) \right)}{\sum_{T_i = 0} \exp \left( \sum_{j=1}^{p} \hat{\theta}_{EB}^j c_p(X_i) \right)}.
\]

Entropy Balancing bridges two existing approaches of estimating the mean causal effect

1. The calibration estimator that is very popular in survey sampling (Deville and Särndal, 1992; Särndal and Lundström, 2005; Chan et al., 2015);
2. The empirical likelihood approach that significantly advances the theory of doubly robust estimation in observation study (Wang and Rao, 2002; Tan, 2006; Qin and Zhang, 2007; Tan, 2010).

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. Marcus, 2013; Ferwerda, 2014).

4. Properties of Entropy Balancing

We give some theoretical guarantees of Entropy Balancing to justify its usage in real applications. Here is the main theorem of the paper, which suggests EB has a double robustness property 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 \( \text{Var}(Y(0)) < \infty \). Then Entropy Balancing is doubly robust (Property 7) 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 \( \gamma \) derived in Hahn (1998 Theorem 1) 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 and the statistical property, double robustness, which motivates the interpretation in Figure 1. The second proof uses a stabilization trick in Robins et al. (2007).

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 logit(P(Y|X)) can be proved by expanding E[Y(0)|X] and ∑T=0 w_iY_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 ě(x), the corresponding weights ě(x)/(1-ě(x)) for the control units, and an estimated outcome regression model ̂g(x), a doubly robust estimator of PATT is given by

\[
\hat{γ}^{DR} = \sum_{T_i=1} \frac{1}{n_1} (Y_i - ̂g_0(X_i)) - \sum_{T_i=0} \frac{ê(X_i)}{1-ê(X_i)} (Y_i - ̂g_0(X_i)).
\]

This estimator satisfies Property 1, i.e. if ě(x) → e(x) or ̂g_0(x) → g(x), then ěDR is statistically consistent for γ. To see this, in the case that ̂g_0(x) → 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 → ∞. In the case where ̂g_0(x) ≠ g_0(x) but ě(x) → 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 ě(x) is generated by the EB dual problem (7) and the estimated outcome regression model is ̂g_0(x) = ∑β_ĵc_j(x), we have

\[
\hat{γ}^{DR} - \hat{γ}^{EB} = \sum_{T_i=0} w_i^{EB} ̂g_0(X_i) - \frac{1}{n_1} \sum_{T_i=0} ̂g_0(X_i)
\]

\[
= \sum_{T_i=0} w_i^{EB} \sum_{j=1}^p ̂β_j ̂c_j(X_i) - \frac{1}{n_1} \sum_{T_i=0} \sum_{j=1}^p ̂β_j ̂c_j(X_i)
\]

\[
= \sum_{j=1}^p ̂β_j \left( \sum_{T_i=0} w_i^{EB} ̂c_j(X_i) - \frac{1}{n_1} \sum_{T_i=0} ̂c_j(X_i) \right) = 0.
\]

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

Second proof. This proof is pointed out by an anonymous reviewer. In a discussion of Kang and Schafer (2007), Robins et al. (2007) indicated that one can stabilize the standard doubly robust estimator in a number of ways. Specifically, one trick suggested by Robins et al. (2007, Section 4.1.2) is to estimate the propensity score, say ě(x), by the following estimating equation

\[
\sum_{i=1}^n \left[ \frac{(1-T_i)ê(X_i)/(1-ê(X_i))}{\sum_{i=1}^n (1-T_i) ê(X_i)/(1-ê(X_i))} - \frac{T_i}{\sum_{i=1}^n T_i} \right] ̂g_0(X_i) = 0.
\]

Then one can estimate PATT by the IPW estimator (2) by replacing ě(X_i) with ê(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 ˆθ(x) = ˆθ(x; θ) and ˆg0(x) = ˆg0(x; β). The only problem with (11) is it may not have a unique solution. However, when logit(e(x)) and g0(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 ˆθ(X; θ) is the same as the EB estimate ˆθ(X; θ). As a consequence, ˆγEB is also doubly robust. □

To prove the second claim in Theorem 1, we compute the asymptotic variance of ˆγEB using the M-estimation theory. To state our result, we need to introduce four different kinds of weighted covariance-like functions for two random vectors a1 and a2 of length p:

\[ H_{a_1,a_2} = \text{Cov}(a_1, a_2 | T = 1), \]
\[ G_{a_1,a_2} = E \left[ \frac{e(X)}{1 - e(X)} (a_1 - E[a_1 | T = 1])(a_2 - E[a_2 | T = 1])^T | T = 1 \right], \]
\[ K_{a_1,a_2} = E[(1 - e(X))a_1a_2^T | T = 1], \]
\[ K_{a_1,a_2} = E[(1 - e(X))a_1(a_2 - E[a_2 | T = 1])^T | T = 1]. \]

It is obvious that H ≥ K and usually G ≥ 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 T is correct, i.e. logit(P(T = 1 | X)) is a linear combination of \{c_j(X)\}_{j=1}^p. Let \( \pi = P(T = 1) \), then we have \( \hat{\gamma}_{EB} \xrightarrow{d} N(\gamma, V^{EB}/n) \) and \( \hat{\gamma}_{IPW} \xrightarrow{d} N(\gamma, V^{IPW}/n) \) where

\[ V^{EB} = \pi^{-1} \cdot \left\{ H_1 + G_0 - H_{c,0}^T H^{-1}_{c,0} (2G_{c,0} - H_{c,0} - G_c H_{c,0}^{-1} H_{c,0} + 2H_{c,1}) \right\}, \]
\[ V^{IPW} = \pi^{-1} \cdot \left\{ H_1 + G_0 - H_{c,0}^T K_{c,0}^{-1} (H_{c,0} - 2K_{c,0}^m + 2K_{c,1}^m) \right\}. \]

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. Stefanski and Boos [2002]). 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 Lanceford and Davidian [2004] 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 Häim [1998] Theorem 1:

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

After some algebra, one can express \( V^* \) in terms of H, G, 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 logit(P(T = 1 | X)) = \theta^T c(X) and E[Y(t) | X] = β(t)^T c(X), \( t = 0, 1 \), it is easy to verify that

\[ H_{c,t} = \text{Cov}(c(X), Y(t) | T = 1) \]
\[ = \text{Cov}(c(X), \beta(t)^T c(X) | T = 1) \]
\[ = H_{c,\beta}(t), \text{ for } t = 0, 1. \]
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 $\text{logit}(P(T = 1|X)) = \theta^T c(X)$ but $E[Y(t)|X] = \beta(t)^T c(X)$ is not true for some $t = 0, 1$, there is no guarantee that $V_{EB} < V_{IPW}$. In practice, the features $c(X)$ in the linear models of $Y$ 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

We use the simulation example in Kang and Schafer (2007) 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 (2014). The simulated data consists 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:

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

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

In one setting, $Y_i$ is generated by $Y_i = 210 + 27.4 X_{i1} + 13.7 X_{i2} + 13.7 X_{i3} + 13.7 X_{i4} + \epsilon_i$, $\epsilon_i \sim N(0, 1)$ and the true propensity scores are $e_i = \expit(-X_{i1} + 0.5 X_{i2} - 0.25 X_{i3} - 0.1 X_{i4})$. 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 $n = 1000$ and apply various methods discussed earlier, including

1. IPW, CBPS: the IPW estimator in (2) with propensity score estimated by logistic regression or CBPS;
2. EB: the Entropy Balancing estimator in (4);
3. IPW+DR, CBPS+DR: the doubly robust estimator in (10) with propensity score estimated by logistic regression or CBPS.

Notice that the correct mean of $Y$ is always 210. The simulation results are shown in Figure 2. The numbers printed at the top of the figure are standard deviations of each method.

The reader may have noticed some unusual facts in this plot. First, the doubly robust estimator “IPW+DR” performs poorly when both models are misspecified (bottom-right panel in Figure 2). In fact, all the three doubly robust methods are worse than just using IPW. Second, the three doubly robust estimators have
Figure 2. Results of the Kang-Schafer Example. 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). Both propensity score model and outcome regression model can be correct or incorrect, so there are four scenarios in total. We generate 1000 simulations of 1000 sample from the example in (Kang and Schafer, 2007) and apply five different estimators. Target mean is 210 and is marked as a black horizontal line to compare the biases of the methods. Numbers printed at $Y = 230$ are the sample standard deviation of each method to compare their efficiency.

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 (2007) 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.
and RIDGWAY AND MCCAFFREY (2007) 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.

Although the Kang-Schafer example is artificial and is arguably not very likely to occur in practice, we make two comments here about entropy balancing in this unfavorable setting for doubly robust estimators:

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), EB has the smallest variance among all estimators. This supports the conclusion of our efficiency comparison of IPW and EB in Section 4.

Finally we want to notice that the same simulation setting is used in Tan (2010) to study the performance of a number of doubly robust estimators. The reader can compare the Figure 2 with the results there. Overall, the performance of Entropy Balancing is comparable to the best estimator in Tan (2010).

References

Abadie, A. and G. W. Imbens (2006). Large sample properties of matching estimators for average treatment effects. *Econometrica* 74(1), 235–267.

Albert, A. and J. A. Andersen (1984). On the existence of maximum likelihood estimates in logistic regression models. *Biometrika* 71(1), 1–10.

Bang, H. and J. M. Robins (2005). Doubly robust estimation in missing data and causal inference models. *Biometrics* 61(4), 962–973.

Chan, K. C. G., S. C. P. Yam, and Z. Zhang (2015). Globally efficient nonparametric inference of average treatment effects by empirical balancing calibration weighting. *Journal of Royal Statistical Society, Series B (Methodology)* to appear.

Cover, T. M. and J. A. Thomas (2012). *Elements of information theory*. John Wiley & Sons.

Deville, J.-C. and C.-E. Särndal (1992). Calibration estimators in survey sampling. *Journal of the American Statistical Association* 87(418), 376–382.

Diamond, A. and J. S. Sekhon (2013). Genetic matching for estimating causal effects: A general multivariate matching method for achieving balance in observational studies. *Review of Economics and Statistics* 95(3), 932–945.

Ferwerda, J. (2014). Electoral consequences of declining participation: A natural experiment in Austria. *Electoral Studies* 35(0), 242–252.

Gneiting, T. and A. E. Raftery (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association* 102(477), 359–378.

Hahn, J. (1998). On the role of the propensity score in efficient semiparametric estimation of average treatment effects. *Econometrica* 66(2), 315–332.

Hainmueller, J. (2011). Entropy balancing for causal effects: A multivariate reweighting method to produce balanced samples in observational studies. *Political Analysis* 20, 25–46.

Hastie, T. J. and R. J. Tibshirani (1990). *Generalized additive models*, Volume 43. CRC Press.

Hirano, K. and G. Imbens (2001). Estimation of causal effects using propensity score weighting: An application to data on right heart catheterization. *Health Services and Outcomes Research Methodology* 2, 259–278.
Hirano, K., G. W. Imbens, and G. Ridder (2003). Efficient estimation of average treatment effects using the estimated propensity score. *Econometrica* 71(4), 1161–1189.

Holland, P. W. (1986). Statistics and causal inference. *Journal of the American Statistical Association* 81, 945–960.

Imai, K., G. King, and E. A. Stuart (2008). Misunderstandings between experimentalists and observationalists about causal inference. *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 171(2), 481–502.

Imai, K. and M. Ratkovic (2014). Covariate balancing propensity score. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 76(1), 243–263.

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

Lunceford, J. K. and M. Davidian (2004). Stratification and weighting via the propensity score in estimation of causal treatment effects: a comparative study. *Statistics in Medicine* 23(19), 2937–2960.

Marcus, J. (2013). The effect of unemployment on the mental health of spouses evidence from plant closures in Germany. *Journal of Health Economics* 32(3), 546–558.

Neyman, J. (1923). Sur les applications de la thar des probabilities aux experiences agaricales: Essay des principle. excerpts reprinted (1990) in english. *Statistical Science* 5, 463–472.

Qin, J. and B. Zhang (2007). Empirical-likelihood-based inference in missing response problems and its application in observational studies. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 69(1), 101–122.

Ridgeway, G. and D. F. McCaffrey (2007). Comment: Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data. *Statistical Science* 22(4), 540–543.

Robins, J., M. Sued, Q. Lei-Gomez, and A. Rotnitzky (2007). Comment: Performance of double-robust estimators when inverse probability weights are highly variable. *Statistical Science* 22(4), 544–559.

Robins, J. M., A. Rotnitzky, and L. Zhao (1994). Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association* 89, 846–866.

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

Rosenbaum, P. and D. Rubin (1984). Reducing bias in observational studies using subclassification on the propensity score. *Journal of the American Statistical Association* 79, 516–524.

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

Rubin, D. (1974). Estimating causal effects of treatments in randomized and non-randomized studies. *Journal of Educational Psychology* 66(5), 688–701.

Särndal, C.-E. and S. Lundström (2005). *Estimation in surveys with nonresponse*. John Wiley & Sons.

Silvapulle, M. J. (1981). On the existence of maximum likelihood estimators for the binomial response models. *Journal of the Royal Statistical Society. Series B*
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

\[
\sum_{T_i=0} w_i c_j(X_i) = \bar{c}_j(1), \quad j = 1, \ldots, p, \quad w > 0, \quad \sum_{T_i=0} w_i = 1.
\]

As one may expect, this is closely related to the existence condition of maximum likelihood estimate of logistic regression (Silvapulle, 1981; Albert and Andersen, 1984). An easy way to obtain such condition is through the dual problem of (8)

\[
\begin{align*}
\text{maximize} & \quad w - \sum_{i=1}^n \left[ w_i \log w_i + (1 - w_i) \log(1 - w_i) \right] \\
\text{subject to} & \quad \sum_{T_i=0} w_i c_j(X_i) = \sum_{T_i=1} w_i c_j(X_i), \quad j = 1, \ldots, p, \\
& \quad 0 < w_i < 1, \quad i = 1, \ldots, n.
\end{align*}
\]

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 (1981).

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 \( \epsilon > 0 \), \( P(\|\bar{c}(1) - \bar{c}^*(1)\|_\infty \geq \epsilon) \to 0 \) as \( n \to \infty \).
Now condition on $\|\tilde{c}(1) - \tilde{c}^*(1)\|_\infty \geq \epsilon$, i.e. $\tilde{c}(1)$ is in the box of side length $2\epsilon$ centered at $\tilde{c}^*(1)$, we want to prove that with probability going to 1 there exists $w$ such that $w_i > 0$, $\sum_{i=0}^{T_i = 0} w_i = 1$ and $\sum_{i=0}^{T_i = 0} w_i c(X_i) = \tilde{c}(1)$. Equivalently, this is saying the convex hull generated by $\{c(X_i)\}_{T_i = 0}$ contains $\tilde{c}(1)$. 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 $B_\epsilon(\tilde{c}^*(1)) = \{c(x) : \|c(x) - \tilde{c}^*(1)\|_\infty \leq \epsilon \}$ for some $\epsilon > 0$.

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 $\tilde{c}^*(1)$ hence $B_\epsilon(\tilde{c}^*(1))$ is in the interior of the convex hull of $\Omega(X)$ for sufficiently small $\epsilon$. Let $R_i$, $i = 1, \ldots, 3^p$, be the $3^p$ boxes centered at $\tilde{c}^*(1) + \frac{2}{\sqrt{p}}\tilde{b}$, where $\tilde{b} \in \mathbb{R}^p$ is a vector that each entry can be $-1, 0, 1$. It is easy to check that the sets $R_i$ are disjoint and the convex hull generated by $\{x_i\}_{i=1}^{3^p}$ contains $B_\epsilon(\tilde{c}^*(1))$ if $x_i \in R_i$, $i = 1, \ldots, 3^p$.

Since $0 < P(T = 0|X) < 1$, $\rho = \min_i P(X \in R_i|T = 0) > 0$. This implies

$$P(3X_i \in R_i \text{ and } T_i = 0, \forall i = 1, \ldots, 3^p) \geq 1 - \sum_{i=1}^{3^p} P(X \notin R_i|T = 0)^n$$

(16)

$$\rightarrow 1$$

as $n \rightarrow \infty$. 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 $w_{\text{opt}}$.

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 $n = O(p \cdot 3^p)$.

Now we turn to the second claim of the proposition, i.e. $\sum_{T_i = 0} w_i^2 \rightarrow 0$. To prove this, we only need to find a sequence (with respect to growing $n$) of feasible solutions to (16) such that $\max_i w_i \rightarrow 0$. 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 $B_\epsilon(\tilde{c}^*(1))$ 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 $B_\epsilon(\tilde{c}^*(1))$ with probability at least $1 - 3^i \delta$. This means for each $\{x_i\}_{i=n_i+1}^{n_{i+1}}$, $i = 0, 1, \ldots$, we have a set of weights $\{\tilde{w}_i\}_{i=n_i+1}^{n_{i+1}}$ such that $\sum_{i=n_i+1}^{n_{i+1}} \tilde{w}_i x_i = \tilde{c}(1)$. 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 $n \rightarrow \infty$, this implies $\max_i w_i \rightarrow 0$ and hence $\sum_i w_i^2 \rightarrow 0$ 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 $\logit(P|T = 1|X) = \sum_{j=1}^{p} \theta_j^* c_j(X)$, then as $n \rightarrow \infty$, $\hat{\theta}^{EB} \rightarrow 0$. As a consequence,

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

\[1\] Note that this naive rate can actually be greatly improved by Wendel’s theorem in geometric probability theory.
Proof: The proof is a standard application of M-estimation (more precisely Z-estimation) theory. We will follow the estimating equations approach described in [Stefanski and Boos 2002] to derive consistency of $\hat{\theta}_{EB}$. First we note the first order optimality condition of (7) is

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

We can rewrite (17) as estimating equations. Let $\phi_j(X; T; m) = T(c_j(X) - m_j)$, $j = 1, \ldots, R$ 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

$$
\sum_{i=1}^{n} \phi_j(X_i, T_i; m) = 0, \quad j = 1, \ldots, R,
$$

$$
\sum_{i=1}^{n} \psi_j(X_i, T_i; \theta, m) = 0, \quad j = 1, \ldots, R.
$$

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

$$
E[\psi_j(X; T; \theta, m^*) | X] = (1 - e(X))e^{\sum_{k=1}^{p} \theta_k c_k(X)}(c_j(X) - m^*_j)
$$

$$
= \left(1 - \frac{e^{\sum_{k=1}^{p} \theta_k c_k(X)}}{1 + e^{\sum_{k=1}^{p} \theta_k c_k(X)}}\right) e^{\sum_{k=1}^{p} \theta_k c_k(X)}(c_j(X) - m^*_j)
$$

$$
= \frac{e^{\sum_{k=1}^{p} \theta_k c_k(X)}}{1 + e^{\sum_{k=1}^{p} \theta_k c_k(X)}}(c_j(X) - E[c_j(X)|T = 1]).
$$

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

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

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

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

$$
w^*_{\gamma_{EB}} = \frac{\exp(\sum_{j=1}^{p} \hat{\theta}^E_j c_j(X_i))}{\sum_{T_i=0}^{p} \exp(\sum_{j=1}^{p} \hat{\theta}^E_j c_j(X_i))} \frac{P(T_i = 1|X_i)}{1 - P(T_i = 1|X_i)},
$$

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 $\xi^*$. Throughout this section we assume $\logit(e(X)) = \sum_{j=1}^{p} \theta^*_j c_j(X)$. Denote $\tilde{c}(X) = c(X) - \tilde{c}^*(1)$, $e^*(X) = e(X; \theta^*), l^*(X) = \exp(\sum_{j=1}^{p} \theta^*_j c_j(X)) = e^*(X)/(1 - e^*(X))$. Let

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

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

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

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

$$
\varphi_1(X; T, Y; \mu(1|1)) = T(Y - \mu(1|1)).
$$
and $\zeta(X, T, Y; m, \theta, \mu(1|1), \gamma) = (\varphi^T, \psi^T, \varphi_{11}, \varphi)^T$ be all the estimating equations. The Entropy Balancing estimator $\hat{\gamma}^{EB}$ is the solution to

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

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

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

$$= \left( \mathbb{E} \left[ -\frac{\partial}{\partial \theta^T} \zeta(X; \zeta^*) \right] \right)$$

$$= \left( \mathbb{E} \left[ -\frac{\partial}{\partial \mu(1|1)} \zeta(X; \zeta^*) \right] \right)$$

$$= \mathbb{E} \left[ \begin{pmatrix}
T^T \cdot I_R \\
(1-T)^*I^*(X) - I_R \\
0
\end{pmatrix}
\begin{pmatrix}
0 & 0 & 0 \\
0 & 0 & 0 \\
0 & 0 & 0
\end{pmatrix}
\begin{pmatrix}
T^T \cdot I_R \\
(1-T)^*I^*(X) - I_R \\
0
\end{pmatrix}
\mathbb{E} \left[ \begin{pmatrix}
0 & 0 & 0 \\
0 & 0 & 0 \\
0 & 0 & 0
\end{pmatrix}
\begin{pmatrix}
0 & 0 & 0 \\
0 & 0 & 0 \\
0 & 0 & 0
\end{pmatrix}
\begin{pmatrix}
T^T \cdot I_R \\
(1-T)^*I^*(X) - I_R \\
0
\end{pmatrix}
\right].$$

A very useful identity in the computation of the expectation is

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

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

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

$$= \mathbb{E} \left[ \begin{pmatrix}
T^T \cdot I_R \\
(1-T)^*I^*(X) - I_R \\
0
\end{pmatrix}
\begin{pmatrix}
0 & 0 & 0 \\
0 & 0 & 0 \\
0 & 0 & 0
\end{pmatrix}
\begin{pmatrix}
T^T \cdot I_R \\
(1-T)^*I^*(X) - I_R \\
0
\end{pmatrix}
\right].$$

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

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

$$G_{a_1, a_2} = \mathbb{E} \left[ l^*(X)(a_1 - E[a_1|T = 1])(a_2 - E[a_2|T = 1])^T|T = 1 \right]^T,$$

and $H_a = H_{a, a}$, $G_a = G_{a, a}$. So

$$A^{EB}(\zeta^*) = \pi \cdot \begin{pmatrix}
I_R \\
I_R \\
0^T
\end{pmatrix}
\begin{pmatrix}
0 & 0 & 0 \\
0 & 0 & 0 \\
1 & 0 & 1
\end{pmatrix},$$

$$A^{EB}(\zeta^*)^{-1} = \pi^{-1} \cdot \begin{pmatrix}
I_R \\
H_{c(X)}^{-1} \\
H_{c(X)}^{-1}
\end{pmatrix}
\begin{pmatrix}
0 & 0 & 0 \\
0 & 0 & 0 \\
1 & 0 & 1
\end{pmatrix}.$$
Let the propensity score is only related to $X$ and all of them are observed. Both $(2004)$ to verify claims in Theorems 1 and 2.

Here we provide an additional simulation example by Lunceford and Davidian $(2004)$ to verify claims in Theorems 1 and 2.

Thus

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

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

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

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

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

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

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

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

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

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

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

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

$V^{IPW}$ can thus be computed consequently and we omit the details.

**Appendix B. Additional Simulation Example**

Here we provide an additional simulation example by Lunceford and Davidian $(2004)$.

In this simulation, the data still consists 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:

$$\text{logit}(P(T_i = 1)) = \beta_0 + \sum_{j=1}^{3} \beta_j X_{ij}. $$
Note the above does not involve elements of \( Z_i \). The response \( Y \) is generated according to

\[
Y_i = \nu_0 + \sum_{j=1}^{3} \nu_j X_{ij} + \nu_4 T_i + \sum_{j=1}^{3} \xi_j Z_{ij} + \epsilon_i; \quad \epsilon_i \sim N(0,1).
\]

The parameters here are set to be \( \nu = (0, -1, 1, -1, 2)^T \);

\( \beta \) is set as:

\[
\beta_{\text{no}} = (0, 0, 0)^T,
\beta_{\text{moderate}} = (0, 0.3, -0.3, 0.3)^T, \text{ or }
\beta_{\text{strong}} = (0, 0.6, -0.6, 0.6)^T.
\]

The choice of \( \beta \) depends on the level of association of \( T \) and \( X \). \( \xi \) is based on a similar choice on the level of association of \( Y \) and \( Z \):

\[
\xi_{\text{no}} = (0, 0, 0)^T,
\xi_{\text{moderate}} = (-0.5, 0.5, 0.5)^T, \text{ or }
\xi_{\text{strong}} = (-1, 1, 1)^T.
\]

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

\[
P(Z_{i3} = 1|X_{i3}) = 0.75 X_{i3} + 0.25 (1 - X_{i3}).
\]

Conditional on \( X_{i3} \), \((X_{i1}, Z_{i1}, X_{i2}, Z_{i2})\) is then generated as multivariate normal

\( \mathcal{N}(a_{X_{i3}}, B_{X_{i3}}) \), where \( a_1 = (1, 1, -1, -1)^T \), \( a_0 = (-1, -1, 1, 1)^T \) and

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

The data generating process implies the true \( \gamma \) is 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 (2004, Sec. 3.3), 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 \( X \).

We generated 1000 simulated data sets and the results are shown in Figure 3 for “full” propensity score modeling and Figure 4 for “partial” propensity score modeling. We make the following comments about these two plots:

1. IPW and all the 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 full propensity score modeling, all the 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.
Figure 3. 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 1000 simulations of 1000 in each scenario and apply five different estimators. Target PATE 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.

(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 4), this bias is a little smaller but still not negligible. While it is not clear what causes this bias, one possible reason is that the optimization problem of CBPS
20 ENTROPY BALANCING

Figure 4. Results of the Lunceford-Davidian example (partial propensity score modeling). The settings are exactly the same as Figure 3 except the methods here don’t use $Z$ in their propensity score models.

is nonconvex, so the local solution which is used to construct $\gamma$ 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 $\gamma$ 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 overall, especially if we use the “full” model, i.e. both $X$ and $Z$ are being balanced in $\beta$. 

(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 3. When EB only balances “partial” covariates, the two methods are different and indeed EB+DR is more efficient in Figure 4 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 $Y$. 