Received: Added at production | Revised: Added at production | Accepted: Added at production

DOI: xxx/xxxx

ARTICLE TYPE

Balancing Versus Modeling Approaches to Weighting in Practice

Ambarish Chattopadhyay[1] | Christopher H. Hase[1] | José R. Zubizarreta*[1,2]

[1]Department of Statistics, Harvard University, Cambridge, MA, USA
[2]Department of Health Care Policy, Harvard Medical School, Harvard University, 180 Longwood Ave, MA, USA

Correspondence
*José R. Zubizarreta, Department Health Care Policy, Harvard Medical School, 180 Longwood Ave, Office 307-D, Boston, MA 02115, USA. Email: zubizarreta@hcp.med.harvard.edu

Present Address
180 Longwood Ave, Office 307-D, Boston, MA 02115, USA

Summary

There are two seemingly unrelated approaches to weighting in observational studies. One of them maximizes the fit of a model for treatment assignment to then derive weights — we call this the modeling approach. The other directly optimizes certain features of the weights — we call this the balancing approach. The implementations of these two approaches are related: the balancing approach implicitly models the propensity score, while instances of the modeling approach impose balance conditions on the covariates used to estimate the score. In this paper, we review and compare these two approaches to weighting. Previous review papers have focused on the modeling approach, emphasizing the importance of checking covariate balance, but as we discuss, the dispersion of the weights is another important aspect of the weights to consider, in addition to the representativeness of the weighted sample and the sample boundedness of the weighted estimator. In particular, the dispersion of the weights is important because it translates into a measure of effective sample size which can be used to select between alternative weighting schemes. In this paper, we examine the balancing approach to weighting, discuss recent methodological developments, and compare instances of the balancing and modeling approaches in a simulation study and an empirical study. In practice, unless the treatment assignment model is known, we recommend using the balancing approach to weighting, as it systematically results in better covariate balance with weights that are minimally dispersed. As a result, effect estimates tend to be more accurate and stable.

KEYWORDS:
causal inference, inverse probability weights, propensity scores, observational studies, weighting.

1 | INTRODUCTION

1.1 | The need to weight

In observational studies, it is well-known that the treated and control samples may suffer from imbalances in the distributions of their covariates. These imbalances, in turn, may result in biases when estimating treatment effects, if the covariates are also associated with the outcome. In order to improve balance and remove biases, it is recommended to design the observational study as if it was a randomized experiment, using information only on the observed covariates and not the outcome.[1] This helps to maintain the objectivity of the study and preserve the level of its statistical tests. Weighting is a very general method that can

[0]The authors thank Zach Branson, Avi Feller, Yige Li, Bijan Niknam, and Sherri Rose for helpful conversations and input. This work was supported through a Patient-Centered Outcomes Research Institute (PCORI) Project Program Award (ME-2019C1-16172) and grants from the National Institute of Health (1DP2MD012722) and the Alfred P. Sloan Foundation (G-2018-10118).

be used to design an observational study in this spirit. Other common methods such as matching[2] and subclassification,[3] can be viewed as particular cases of weighting. See Stuart (2010)[4] and Austin and Stuart (2015)[5] for reviews of these methods.

Conventionally, the weights are calculated by first modeling the conditional probability of receiving treatment given the observed covariates — that is, the propensity score —[6] and then inverting the estimated probabilities. The resulting weights are used as estimates of the inverse probability of treatment weights, or simply, inverse probability weights (IPW). Part of the popularity of this approach can be attributed to the propensity score itself, since it provides a one-dimensional summary of the (possibly high-dimensional) vector of observed covariates while being a balancing score. In other words, IPW estimated with a correctly specified propensity score model induces balanced weighted treatment and control groups in expectation and thus can provide unbiased estimators for several common estimands of interest, such as the average treatment effect. Furthermore, resulting estimators have desirable large sample properties under certain smoothness conditions, which make them theoretically appealing.[7]

1.2 | Weighting over the last years

In an interesting paper on IPW and weight diagnostics, Austin and Stuart (2015)[5] review IPW-based methods published in the applied biomedical literature during the period 1987-2014. Figure 1 provides an updated account of the number of published IPW articles both in the applied and methodological literatures in the last twenty years (see the online supplementary materials for more details). From the figure, it is evident that the usage of propensity score-based weighting has increased rapidly over the years. Since 2011, there have been more applied articles using inverse probability weighting than methodological articles. Specifically in the last six years (2013–2018), the number of applied articles has been more than twice the number of methodological articles in this topic. This exemplifies the generality and wide applicability of the propensity score and IPW in practice.

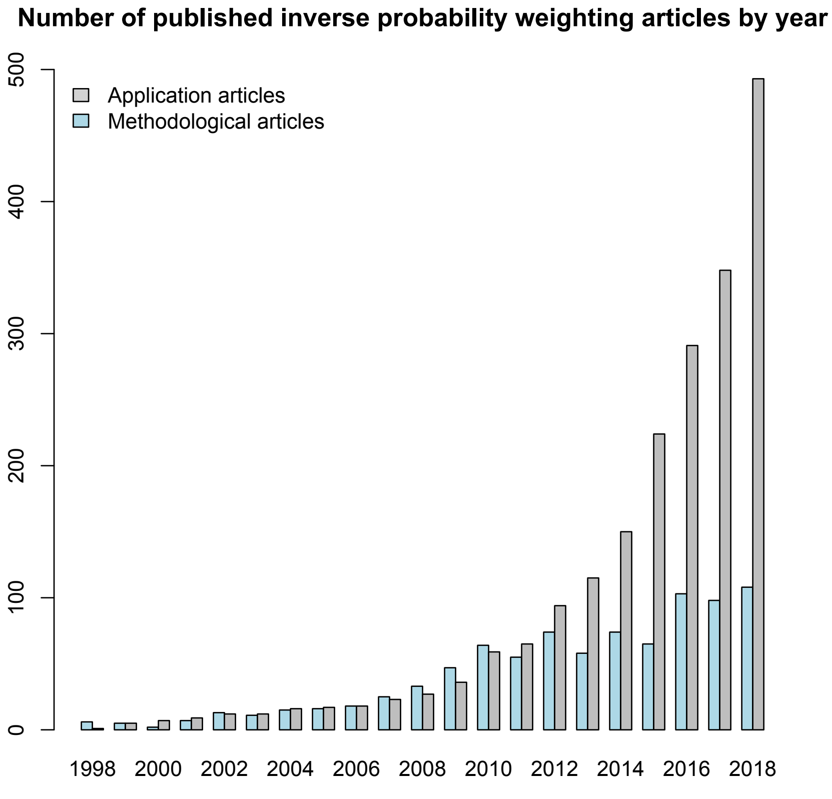


FIGURE 1 Number of published IPW articles by year

However, in most observational studies, the propensity score is unknown to the investigator. Thus, the use of IPW requires estimating the propensity score, often through a model that explicitly relates the treatment indicator to the observed covariates. We term this approach as the 'modeling approach' to weighting.[8] A common criticism of the modeling approach is that weighting by the estimated propensity score may not produce adequate balance on the observed covariate distributions.[9,10] Moreover, the weights under the modeling approach can be highly variable, and highly variable weights bear the risk of producing highly unstable estimators.[11,12] There are several recently proposed weighting schemes which have attempted to address one or both of these two issues. One one hand, some methods estimate the propensity score under the requirement that certain prespecified covariate balance conditions are satisfied,[13,14,15,16] instead of estimating the propensity score simply by maximizing the fit of the model. On the other hand, other methods find the weights that directly optimize towards certain attributes of the weights, without explicitly specifying a functional form for the underlying propensity score model.[17,10,18,19] These methods target covariate balance directly while simultaneously minimizing a convex function of the weights (usually, a measure of dispersion of the weights). Since both sets of methods directly address the balancing aspect requirement of weighting, we term them 'balancing approaches' to weighting.

1.3 | Contribution and outline

In this paper, we take a step back from the traditional modeling approaches to weighting and ask, "what should a good weighting strategy be?" As alluded to earlier, the weights should balance the observed covariate distributions. Also, it is important for the weights to have small dispersion, as this translates into smaller standard errors and larger effective sample sizes. In this paper, we provide a formal motivation for focusing on balance and the dispersion of the weights and augment the set of diagnostics in Austin and Stuart (2015).[5] In particular, we propose a targeted measure of covariate balance and discuss that the effective sample size of the weighted sample is an intuitive and practical diagnostic for the dispersion of the weights. In addition, we argue that the weights should be non-negative to allow standard weighted estimators to satisfy the sample-boundedness property discussed by Robins et al. (2007),[20] which results in an interpolation of the available data as opposed to an extrapolation beyond the support of the data.

In view of these desirable features, we discuss and review the modeling and balancing approaches to weighting. With regard to the balancing approach, we focus on a broad class of weighting methods discussed by Wang and Zubizarreta (2020)[21] called *minimal dispersion approximately balancing weights*, or *minimal weights* for short. These weights are minimal in that they have minimum dispersion (e.g., variance) and balance approximately (not necessarily exactly) statistics of the distribution of the observed covariates. While this balancing approach differs from the modeling approach in its original motivation and implementation (the modeling approach optimizes towards the fit of the propensity score model, whereas the balancing approach optimizes towards balance and dispersion of the weights), we show that these two approaches are connected. In the context of estimation with incomplete outcome data, Wang and Zubizarreta (2020)[21] show that, although the optimization process of the balancing approach does not explicitly involve the propensity score, the balancing approach implicitly models the propensity score by balancing functions of the covariates, and the resulting weights are proportional to the corresponding inverse probability weights. In this paper, we extend their result to observational study settings (see the Appendix for a proof). Under certain assumptions, minimal weights asymptotically behave like the true IPW. Nonetheless, instead of maximizing the standard log-likelihood function, the balancing approach estimates the underlying propensity score by minimizing a loss function that explicitly restrains some form of dispersion of the weights and directly balances given functions of the covariates. Therefore, despite their connection, the modeling and balancing approach can perform much differently in finite samples.

In this paper, we compare the finite sample performances of these two approaches through an extensive simulation study followed by an empirical study. We focus on one particular instance of minimal weights, which are the non-negative weights of minimum variance that approximately balance the covariates.[10] The results based on the two studies reflect that this balancing approach outperforms common modeling approaches in terms of covariate balance and dispersion of the weights, yielding more accurate and stable, sample bounded estimates of treatment effects.

Finally, the balancing approach to weighting with minimal weights requires selection of a tuning parameter which controls the degree of approximate balance between the weighted treated and the weighted control groups, relative to the target. We provide an algorithm for selecting this parameter, which is a modified version of a similar bootstrap-based tuning algorithm given by Wang and Zubizarreta (2020).[21]

The paper is structured as follows. In Section 2, we formally describe the framework and the estimation problem. In Section 3, we delineate a general desiderata for weighting and provide a collection of diagnostics which, in principle, should apply to any form of weighting in practice. We then proceed to explain the modeling and balancing approaches and formally establish their connection in Section 4. Section 5 demonstrates the comparative performance of these two weighting approaches in finite samples using a simulation study. The weighting methods are also compared in terms of weight diagnostics using an empirical study on air quality in Section 6. Finally, Section 7 concludes with a summary and remarks on some open questions regarding weighting in practice.

2 | SETTING

2.1 | Framework

We are interested in estimating the average treatment effect of a time-invariant binary treatment (such as the point exposure to a drug) on an outcome (such as blood pressure) in a given population. Suppose we have access to a sample of n individuals or units, drawn randomly from that population. Corresponding to the ith unit in the sample, let $X_i \in \mathbb{R}^p$ be the vector of p baseline covariates. Let $Z_i \in \{0,1\}$ be the treatment indicator; i.e., $Z_i = 1$ if the unit belongs to the treatment group (labelled as t) and $Z_i = 0$ if the unit belongs to the control group (labelled as c). Let $Y_i^{\text{obs}} \in \mathbb{R}$ be the observed value of the outcome variable. Further, let n_t and n_c be the size of the treatment group and control group respectively: $n_t = \sum_{i=1}^n Z_i$, $n_c = \sum_{i=1}^n (1 - Z_i)$, $n_c + n_t = n$. The set of triplets $\{(X_i, Z_i, Y_i^{\text{obs}}); i = 1, 2, ..., n\}$ constitutes the full set of observed data, and under the assumption of random sampling from a large enough population, each triplet is independently and identically distributed.

We base our discussion on the potential outcome framework for causal inference.[22,23] We require the Stable Unit Treatment Value Assumption (SUTVA)[24] which states that there is no interference between units and no versions of the treatment beyond those encoded by Z. This results in a well-defined set of two potential outcomes, denoted by $Y(1)$ (corresponding to treatment) and $Y(0)$ (corresponding to control) for each unit in the population, and $Y_i^{\text{obs}} = Z_i Y(1) - (1 - Z_i)Y(0)$. Under the potential outcome framework, we define causal effects by contrasting these two potential outcomes as follows.

2.2 | Estimands

We consider two causal effects which are of interest in practice. The first one is the Average Treatment Effect defined as ATE $:= \mathbb{E}\left[Y_i(1) - Y_i(0)\right]$.[1] The second one is the Average Treatment effect on the Treated, defined as ATT $:= \mathbb{E}[Y_i(1) - Y_i(0) | Z_i = 1]$. The weighting strategies will vary depending on whether the goal is to estimate the ATE or ATT. We discuss this in more detail in the later sections.

2.3 | Assumptions

We have already made two fundamental assumptions in the formulation of the estimation problem, namely random sampling of units from a population and SUTVA. In addition to these, we need two more assumptions to ensure identifiability of the previous causal estimands.

(a) **Unconfoundedness/Ignorability** : $\{Y_i(0), Y_i(1)\} \perp\!\!\!\perp Z_i | X_i$ for all $i = 1, 2, ..., n$; that is, the potential outcomes are conditionally independent of the treatment assignment given the observed covariates.

(b) **Positivity/Overlap/Probabilistic Assignment**: $0 < P(Z_i = 1 | X_i = x) < 1$ for all $x \in \mathcal{X}$, where $\mathcal{X}$ is the support of X_i. This means that for any value of the (multidimensional) covariate in its support, a unit possessing that value of the covariate has strictly positive probability of receiving treatment or control.

Conditions (a) and (b) jointly form the assumption of strong ignorability.[6] The unconfoundedness assumption can be re-written as $\{Y_i(0), Y_i(1)\} | Z_i = 1, X_i \sim \{Y_i(0), Y_i(1)\} | Z_i = 0, X_i$. This implies that if the treatment and control groups have the same distribution of observed covariates (i.e., $X_i | Z_i = 1 \sim X_i | Z_i = 0$ or equivalently $X_i \perp\!\!\!\perp Z_i$), then they have the same distribution of potential outcomes, making the two groups comparable. This reinforces the need to adjust for observed covariates to yield balanced treatment and control groups.

[1] This estimand is often referred to as the Population Average Treatment Effect (PATE) to distinguish it from its sample counterpart SATE $:= \frac{1}{n} \sum_{i=1}^n \{Y_i(1) - Y_i(0)\}$. In this paper we restrict our attention to the population version of the treatment effects, since in this setting, the potential outcomes in the sample $\{Y_i(0), Y_i(1); i = 1, 2, ..., n\}$ are regarded as random variables. However, it is worthwhile to note that due to random sampling, any reasonable estimator of the SATE can be conceptualized as a reasonable estimator of the PATE and vice-versa.[25] For the remainder of the paper, the PATE will simply be referred to as the ATE.

Now, we can express the ATE as

$$\text{ATE} = \mathbb{E}[Y_i(1) - Y_i(0)] = \mathbb{E}[m_1(X_i)] - \mathbb{E}[m_0(X_i)] \tag{2.3.1}$$

where $m_z(x) = \mathbb{E}[Y_i(z)|X_i = x]$, $z \in \{0, 1\}$. Unconfoundedness implies, $m_z(x) = \mathbb{E}[Y_i(z)|X_i = x, Z_i = z] = \mathbb{E}[Y_i^{\text{obs}}|X_i = x, Z_i = z]$. Under positivity, one can identify $m_z(x)$ for all x in $\mathcal{X}$ and consequently, the ATE. Identification of the ATT can be ensured similarly. Given the above assumptions, we now provide a brief description of the weight-based estimators of the ATE and the ATT.

2.4 | Estimators

In this section, we consider four different types of estimators. Denote the propensity score for unit i as $e(X_i) = P(Z_i = 1|X_i)$. If the estimand of interest is the ATE, then the corresponding inverse probability weights are as follows

$$w_i = \frac{Z_i}{e(X_i)} + \frac{1 - Z_i}{1 - e(X_i)}. \tag{2.4.1}$$

The weight received by each unit is computed by inverting the probability of the treatment that the subject received (treatment or control) and hence the name IPW. In case of estimating the ATE, the *target population* is the population of treatment and control units from which the sample was drawn. Here, both the treated and control units are weighted to represent this target population. Similarly, when the estimand of interest is the ATT, the target population changes to the population of treated units and the weights now are

$$w_i = Z_i + \frac{(1 - Z_i)e(X_i)}{1 - e(X_i)}. \tag{2.4.2}$$

Here, we weight the control units to represent the treated units. That is why each treated unit receives a weight of one. When the propensity scores are unknown, we plug in their estimates $\{\hat{e}(X_i), i = 1, 2, ..., n\}$ in the expression of the weights.

(i) **Horvitz-Thompson estimator**: The Horvitz-Thompson estimator of the ATE is $T_{\text{HT}}^{\text{ATE}} = \frac{1}{n}\sum_{i=1}^{n} \frac{Z_i Y_i^{\text{obs}}}{e(X_i)} - \frac{1}{n}\sum_{i=1}^{n} \frac{(1-Z_i)Y_i^{\text{obs}}}{1-e(X_i)}$. Following much of the causal inference literature, we call this the Horvitz-Thompson estimator for its structural similarity with the Horvitz-Thompson (HT) estimator of the population mean in the survey sampling literature.[26] Similarly, the HT estimator for the ATT is $T_{\text{HT}}^{\text{ATT}} = \frac{1}{n_t}\sum_{i=1}^{n} Z_i Y_i^{\text{obs}} - \frac{1}{n_t}\sum_{i=1}^{n}(1 - Z_i)\frac{e(X_i)Y_i^{\text{obs}}}{1-e(X_i)}$.

(ii) **Linear weighted difference estimator** or **Hajek estimator**: The Hajek estimator is denoted by $T_{\text{lin}} = \frac{1}{\sum_{i=1}^{n} w_i Z_i}\sum_{i=1}^{n} w_i Z_i Y_i^{\text{obs}} - \frac{1}{\sum_{i=1}^{n} w_i(1-Z_i)}\sum_{i=1}^{n} w_i(1 - Z_i)Y_i^{\text{obs}}$. Here we use the same notation for the Hajek estimator of the ATE and ATT. The reason is that T_{lin} has the same form for any general set of weights $\{w_1, ..., w_n\}$ (including the forms in (2.4.1) and (2.4.2)) and it is computed by taking the difference of the weighted average of the outcome variable in the two groups.

The above two estimators do not require a model for the outcome. However, they do require a model for the propensity score if it is unknown. Estimators (i) and (ii) are consistent for their corresponding estimands provided the propensity score model is correctly specified. The following third and fourth types of estimators fall under a large class of doubly robust (henceforth, DR) estimators, which combine a propensity model with a model for the outcome variable such that the estimator is consistent if at least one of the two models is correctly specified.[27] There are many ways one can construct a DR estimator for the ATE and the ATT. In this paper, we consider a residual bias-corrected estimator,[11] which augments an outcome regression-based estimator of the treatment effect with a weighted average of the residuals. Specifically, suppose we fit two separate regression models in the two treatment groups (e.g., using ordinary least squares) and let $\hat{m}_i^{(1)}$ and $\hat{m}_i^{(0)}$ be the predicted values of $Y_i(1)$ and $Y_i(0)$, respectively, based on the two regression model fits. In addition, let $\hat{\epsilon}_i^{(1)}$ and $\hat{\epsilon}_i^{(0)}$ be the corresponding residuals. Denote $\hat{\mu}^{(z)} = \frac{1}{n}\sum_{i=1}^{n} \hat{m}_i^{(z)}$, $z \in \{0, 1\}$. The Horvitz-Thompson version of the bias-corrected doubly robust estimator of ATE is given as follows:

(iii) **DR Horvitz-Thompson (DR HT) estimator:** The DR HT estimator for the ATE is $T_{\text{DR HT}}^{\text{ATE}} = \left\{\hat{\mu}^{(1)} + \frac{1}{n}\sum_{i=1}^{n} \frac{Z_i}{\hat{e}(X_i)}\hat{\epsilon}_i^{(1)}\right\} - \left\{\hat{\mu}^{(0)} + \frac{1}{n}\sum_{i=1}^{n} \frac{1-Z_i}{1-\hat{e}(X_i)}\hat{\epsilon}_i^{(0)}\right\}$. The DR HT estimator for the ATT is $T_{\text{DR HT}}^{\text{ATT}} = \frac{1}{n_t}\sum_{i=1}^{n} Z_i Y_i^{\text{obs}} - \left\{\hat{\mu}^{(0)} + \frac{1}{n_t}\sum_{i=1}^{n} \frac{(1-Z_i)e(X_i)}{1-\hat{e}(X_i)}\hat{\epsilon}_i^{(0)}\right\}$.

Similarly, we can compute the Hajek version of the bias corrected doubly robust estimator for a general set of weights $\{w_1, w_2, ..., w_n\}$.

(iv) **DR Hajek estimator:** The DR Hajek estimator for both the ATE and the ATT has the form $T_{\text{DR lin}} = \left\{\hat{\mu}^{(1)} + \frac{1}{\sum_{i=1}^{n} Z_i w_i} \sum_{i=1}^{n} Z_i w_i \hat{\epsilon}_i^{(1)}\right\} - \left\{\hat{\mu}^{(0)} + \frac{1}{\sum_{i=1}^{n} (1-Z_i) w_i} \sum_{i=1}^{n} (1 - Z_i) w_i \hat{\epsilon}_i^{(0)}\right\}$.

3 | WEIGHTING WISH LIST: WHAT ARE WE WEIGHTING FOR?[2]

3.1 | Weighting for balance

One of the main motivations for weighting is to balance the distributions of the observed covariates. Usually, covariate balance is assessed in terms of the absolute standardized mean difference (ASMD) for each covariate,[29,4] defined as

$$\text{ASMD}(x) = \frac{|\bar{x}_{w,t} - \bar{x}_{w,c}|}{\sqrt{(s_t^2 + s_c^2)/2}}, \tag{3.1.1}$$

where $\bar{x}_{w,t}$ and $\bar{x}_{w,c}$ are the weighted means of covariate x in the treatment and control groups respectively, and s_t^2 and s_c^2 are their corresponding unweighted sample variances. Instead, we propose using the *target* absolute standardized mean difference (TASMD), defined as

$$\text{TASMD}(x_g) = \frac{|\bar{x}_{w,g} - \bar{x}^*|}{s_g}, \tag{3.1.2}$$

where $\bar{x}_{w,g}$ and s_g are, respectively, the weighted mean and unweighted sample standard deviation of covariate x in treatment group $g \in \{t, c\}$, and $\bar{x}^*$ is the mean in a target population. Alternatively, one can define TASMD(x_g) as the ratio of $|\bar{x}_{w,g} - \bar{x}^*|$ to s^*, where s^* is the standard deviation of covariate x in the target population.[3] While the ASMD measures the imbalance between the treatment and control groups, one relative to the other, the TASMD measures the imbalance of each of the groups relative to the target, so that balance between the two treatment groups is achieved as a by-product of balance relative to the target.

This is a subtle but important point in practice, since certain weighting schemes can produce weighted treatment groups that are well-balanced relative to each other but not relative to the target. As a result, these weighting schemes may fail to aim at the estimand of interest if there is effect modification by covariates (see Equation 3.2.1). Consider, for example, weighting schemes that require truncating the weights at arbitrary thresholds or the implied weights by traditional linear regression modeling of the outcome.[30] In both cases, covariate balance in terms of ASMD can appear to be adequate, but the covariate structure of the weighted groups can be distorted relative to the target population of interest. By evaluating covariate balance in terms of TASMD, we can tell whether this is the case in practice.

The TASMD gives us some versatility. In practice, we can use the TASMD to assess balance relative to different target populations (defined by different estimands of interest), such as the population of treated units when estimating the ATT, the overall population when estimating the ATE, or a specific population characterized by its observed covariates when estimating the conditional average treatment effect (CATE). In any given application, we may not have data from the target population itself, but from a random sample from that population. In that case, we will compute $\bar{x}^*$ in the random sample. For instance, when estimating the ATE, the full sample of treatment and control units can be regarded as a proxy of the target population under random sampling. Thus, the covariate mean in the full sample can be used while computing the TASMD (see Equation 3.2.1 for a formal argument). Procedurally, the TASMD can not only target a population but also the covariate profile of a single data observation, such as a single patient.

In principle, it is desirable to assess balance in terms of other functions of the joint distribution of the observed covariates beyond their marginal means. In general, we propose using

$$\text{TASMD}\{B_k(X)_g\} = \frac{|\overline{B_k(X)}_{w,g} - \overline{B_k(X^*)}|}{s_g\{B_k(X)\}}, \tag{3.1.3}$$

where $B_k(X)$ is a suitable transformation of the p-dimensional observed covariate X, indexed by $k = 1, ..., K$, $\overline{B_k(X)}_{w,g}$ is the weighted mean of $B_k(X)$ in treatment group $g \in \{t, c\}$, $s_g\{B_k(X)\}$ is the unweighted sample standard deviation of $B_k(X)$ in the same group, and $\overline{B_k(X^*)}$ is the mean of $B_k(X)$ in the target population. In particular, if $K = p$ and $B_k(X)$ retrieves the kth coordinate of X, then (3.1.3) reduces to (3.1.2) in order to assess balance of the kth original covariate. In more generality,

[2] We borrow the phrase "what are we weighting for" from Solon et al. (2015).[28]

[3] The latter definition is valid unless the target population corresponds to a target individual, in which case the former definition given by (3.1.2) will be preferable.

$B(X) = \big(B_1(X), ..., B_K(X)\big)^\top$ can be modified to assess balance of higher-order univariate and multivariate moments of X,[31] basis functions for general function-spaces including sieves,[21] and representers of Reproducing Kernel Hilbert Spaces.[32]

In Section 4.2, we will describe the balancing approach to weighting and see that this approach directly constrains or minimizes the TASMD in Equation (3.1.3). Following much of the applied causal inference literature, after weighting, we recommend plotting the TASMDs for each treatment group in order to visualize targeted balance and assess the comparability of each of the treatment groups relative to the target. For other graphical diagnostics, we can also plot the weighted empirical cumulative distribution function of the covariates in the treatment groups and compare them to the cumulative distribution function of the covariates in the target. We note that as a measure of targeted covariate balance, the TASMD naturally extends to multi-valued (non-binary) treatments.

3.2 | Weighting for stability

As already discussed, we balance the covariates in order to remove biases. However, we also want to produce an estimator that has low variance; i.e., one that is stable. In an influential paper, Kang and Schafer (2007)[11] exemplified how weights obtained through a misspecified propensity score model can be highly variable and produce weighted estimators that are highly unstable. As alluded to in the introduction, the stability of the estimator is sometimes overlooked in practice. This motivates conducting diagnostics for the variability of the weights.

In order to explain the problem of the variability of the weights, let us consider the problem of estimating the ATE using the Hajek estimator T_{lin}. We compute the mean squared error of T_{lin} as

$$MSE[T_{\text{lin}}] = \big(Bias[T_{\text{lin}}]\big)^2 + Var[T_{\text{lin}}].$$

Without loss of generality assume that the weights in each group have been normalized so that they add up to one. We can now express the bias and variance of T_{lin} as follows

$$\begin{aligned}
Bias[T_{\text{lin}}] &= \mathbb{E}\big[\mathbb{E}[T_{\text{lin}}|\boldsymbol{X}, \boldsymbol{Z}]\big] - \mathbb{E}\Big[\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[Y_i(1) - Y_i(0)|\boldsymbol{X}, \boldsymbol{Z}]\Big] \\
&= \mathbb{E}\Big[\frac{1}{\sum_{i=1}^{n} Z_i w_i}\sum_{i=1}^{n} Z_i w_i \mathbb{E}[Y_i(1)|X_i] - \frac{1}{\sum_{i=1}^{n}(1 - Z_i)w_i}\sum_{i=1}^{n}(1 - Z_i)w_i\mathbb{E}[Y_i(0)|X_i]\Big] \\
&\quad - \mathbb{E}\Big[\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[Y_i(1) - Y_i(0)|X_i, Z_i]\Big] \\
&= \mathbb{E}\Big[\sum_{i=1}^{n} Z_i w_i m_1(X_i) - \frac{1}{n}\sum_{i=1}^{n} m_1(X_i)\Big] - \mathbb{E}\Big[\sum_{i=1}^{n}(1 - Z_i)w_i m_0(X_i) - \frac{1}{n}\sum_{i=1}^{n} m_0(X_i)\Big]. \qquad (3.2.1)
\end{aligned}$$

The terms inside the two expectations state that the weights that balance the weighted mean of $m_1(X)$ in the treatment group relative to the full sample and and the weighted mean of $m_0(X)$ in the control group relative to the full sample, would completely remove the bias of T_{lin}. Turning our attention to the variance of T_{lin}, we use the following decomposition

$$Var[T_{\text{lin}}] = \mathbb{E}\big[Var[T_{\text{lin}}|\boldsymbol{X}, \boldsymbol{Z}]\big] + Var\big[\mathbb{E}[T_{\text{lin}}|\boldsymbol{X}, \boldsymbol{Z}]\big].$$

Now, suppose the weights exactly balance their respective conditional mean functions, relative to the target, i.e., $\sum_{i=1}^{n} Z_i w_i m_1(X_i) = \frac{1}{n}\sum_{i=1}^{n} m_1(X_i)$ and $\sum_{i=1}^{n} Z_i w_i m_0(X_i) = \frac{1}{n}\sum_{i=1}^{n} m_0(X_i)$. In that case, the second term in the above variance decomposition would be functionally independent of the weights. Let us now assume that the potential outcome models are homoscedastic: $Var[Y_i(1)|X_i] = \sigma_1^2$ and $Var[Y_i(0)|X_i] = \sigma_0^2$ for $i = 1, 2, ..., n$. Expanding the conditional variance in the first term we get

$$\begin{aligned}
Var[T_{\text{lin}}|\boldsymbol{X}, \boldsymbol{Z}] &= \sum_{i=1}^{n} Z_i w_i^2 Var[Y_i(1)|X_i] - \sum_{i=1}^{n}(1 - Z_i)w_i^2 Var[Y_i(0)|X_i] \\
&= \sigma_1^2\sum_{i=1}^{n} Z_i w_i^2 + \sigma_0^2\sum_{i=1}^{n}(1 - Z_i)w_i^2. \qquad (3.2.2)
\end{aligned}$$

Thus, conditional on the observed covariates and treatment assignment, the variance of the Hajek estimator is a weighted sum of the sum of squares of the weights in the two groups. Since the weights are normalized in both the treatment and control groups, controlling the sum of squares of the weights in the two groups is equivalent to controlling the variance of the weights in the two groups. It is therefore desirable to ensure that the variability of the weights in the two groups is not too large.

High variability in the weights often stems from a few observations that receive very large weight values in order for their corresponding weighted treatment group to represent the relevant target population. With modeling approaches to weighting, the presence of a few very large weights often relates to violations of the positivity assumption. The literature distinguishes between two types of violations of the positivity assumption: theoretical (or deterministic) violations, and practical (or random) violations.[33,34] A theoretical violation concerns the data generating mechanism of the treatment assignment, and implies that for certain values in the covariate space, the probability of receiving treatment (or control) is actually zero. Conversely, a practical violation concerns the data at hand and occurs when for certain values of the covariates in the sample, we only observe treated or control units, but not both, in spite both having a non-zero probability of being assigned to both treatments. In both cases, the estimated propensity scores can be very close to zero or one for some units, and the resulting weights can be very large, leading to variable weights and unstable effect estimates.

As diagnostics for stability of the weights, the following measures can be used.

1. *Moment-based measures of variability:* coefficient of variation of the weights, which is given by the standard deviation of the weights divided by the mean of the weights. Equivalently, one can also look at the standard deviation of the normalized weights.

2. *Effective sample size:* the variance of the weights has a direct relationship with the *effective sample size of a weighted sample*. For a general weighted sample with weights $\{w_1, ..., w_n\}$, Kish's approximate formula for the effective sample size is $n_{\text{eff}} = \frac{(\sum_{i=1}^n w_i)^2}{\sum_{i=1}^n w_i^2}$.[35] An unusually low value of the effective sample size would indicate the possible presence of a few influential observations (with extreme weights).

3. *Measures of extremity:* maximum of the weights, 95th and 99th percentile of weights.

4. *Graphical diagnostics:* boxplot of the weights.

All of the aforementioned measures of stability could be used for comparative diagnostics, i.e., diagnostics to compare two or more weighting methods. The effective sample size, however, can be conveniently used as a standalone diagnostic, since one can always quantify the proportion of the total sample size that is effectively being used in the weighted sample by simply taking the ratio of the effective sample size and the total sample size. In the same sense, the effective sample size is simple and intuitive for practice.

3.3 | Weighting to 'be positive'

We can think of weighting as creating copies of each unit in the sample. If the weights are positive — or strictly speaking, if the weights are non-negative — it is intuitive to think of the number of copies being proportional to the value of the weights. Negative weights can produce a very well-balanced sample along with stability; however, it is less clear what the weighted sample represents in this case. The interpretability of the weights and the weighted sample is one motivation to consider weights that are non-negative.

There is a more formal motivation to advocating the use of non-negative weights, which stems from the notion of "sample boundedness" of an estimator. An estimator is sample bounded if its value lies in the parameter space of the targeted estimand with probability 1 (when the sample space of the estimator is finite).[20]

Now, let us again consider the Hajek estimator $T_{\text{lin}} = \frac{1}{\sum_{i=1}^n w_i Z_i} \sum_{i=1}^n w_i Z_i Y_i^{\text{obs}} - \frac{1}{\sum_{i=1}^n w_i(1-Z_i)} \sum_{i=1}^n w_i(1 - Z_i) Y_i^{\text{obs}}$. When the estimand is the ATE, the first term in T_{lin} given by $\frac{1}{\sum_{i=1}^n w_i Z_i} \sum_{i=1}^n w_i Z_i Y_i^{\text{obs}}$ estimates $\mathbb{E}[Y_i(1)]$ (and similarly, the second term in T_{lin} estimates $\mathbb{E}[Y_i(0)]$). If the w_i's for the treated individuals are positive, the first term of T_{lin} lies in the interval $[Y_{1,min}, Y_{1,max}]$, where $Y_{1,min}$ and $Y_{1,max}$ are the minimum and maximum values of the observed outcome in the treatment group. Since the observed outcomes for the treated units are identical to their respective potential outcomes under treatment condition, the interval $[Y_{1,min}, Y_{1,max}]$ is a subset of the parameter space of $\mathbb{E}[Y_i(1)]$. Thus, $\frac{1}{\sum_{i=1}^n w_i Z_i} \sum_{i=1}^n w_i Z_i Y_i^{\text{obs}}$ satisfies the sample boundedness property as an estimator of $\mathbb{E}[Y_i(1)]$. This combined with a similar observation for the second term of T_{lin} implies that if $w_i \geq 0$ for all $i \in \{1, 2, ..., n\}$, the Hajek estimator for the ATE satisfies the sample boundedness property. On the other hand, if at least one of the weights is negative, the Hajek estimator may lie outside the corresponding parameter space with positive probability, thereby failing to be sample bounded. With limited dependent variables, such as binary outcomes,? negative weights can produce estimates outside the feasible range of the outcome variable, that are not easy to interpret.

The idea of sample boundedness in the context of weighting is closely connected to the notions of extrapolation away from — and interpolation based on — the observed data. A common criticism of regression based adjustments in observational studies is that such adjustments often extrapolate and, therefore, may depend heavily on model specification.[36] In contrast, by restricting the weights to be positive, we can force all weight-based adjustments to be an interpolation as opposed to an extrapolation of the observed data. However, when there is a lack of *interpolation overlap* in the sample (i.e., the covariates for some treated unit fall outside the convex hull of the covariates for the control units and vice versa),[37] it may not be possible to estimate the average treatment effect through an interpolation of the observed outcomes. In such cases, one can either change the target estimand (e.g., the average effect in the region of overlap; see, e.g., Crump et al. 2009[38] and Li et al. 2018[39]), or rely on outcome model based adjustments to estimate the treatment effect in the region of non-overlap. In our minds, the feasibility of covariate adjustments with weights that are non-negative (and thus, do not extrapolate) is a fundamental diagnostic on violations to the positivity assumption in practice.

3.4 | Weighting for external validity

Even if we obtain weights that are non-negative, induce satisfactory balance between the treatment and control group, and are reasonably stable, we should ask ourselves the question, to whom does our estimated treatment effect apply to? In this paper, we have assumed the sample is randomly drawn from a certain population, and this random sampling along with the assumptions of unconfoundedness and positivity enables us to identify the ATE for that population. As previously mentioned, these three assumptions ensure that any reasonable estimator of the Sample Average Treatment Effect (SATE) would be reasonable as an estimator of the Population Average Treatment Effect (PATE) and vice-versa. Therefore, under these three assumptions, weighting for an internally valid estimate would also enjoy external validity.

In practice, before the study is conducted, the investigator may want to fix a target population to generalize the in-sample estimates. The study can then be designed either by using a sample of units belonging to the target population, or by using a sample which is disjoint from the target population. In the former setup, the associated estimand is the PATE, which has been our quantity of interest throughout the paper. However, the working sample is not always a random sample from that target population. Thus, inferences concerning the target population under the assumption of random sampling can be misleading. The second scenario involves inference to a population that consists of a disjoint set of units to those in the sample. The corresponding average treatment effect in that population is often referred to as the Target Average Treatment Effect (TATE).[40,41] One can also think of the TATE as the average effect of the treatment on the 'non-participants' under the former setup; i.e., the average effect on the units which are not in the sample. More formally, if S is the indicator of sample membership of a unit in the population, i.e., $S = 1$ if the unit is included in the sample and zero otherwise, then TATE $:= \mathbb{E}[Y(1) - Y(0) | S = 0]$. More generally, we can consider the problem of estimating the average treatment effect of any target population (irrespective of whether it consists of a disjoint set of units or not). In all cases, the validity of the inference depends on the degree of representativeness of the sample to the relevant target population.

This is the well-known problem of generalizability (or external validity), which is highly relevant for both randomized experiments and observational studies. There exists a growing literature on statistical techniques that use weighting to assess generalizability to a target population.[42,43,44] In Section 3.1, we have discussed a balance diagnostic for this goal. Under certain assumptions regarding the sample selection mechanisms,[4] the PATE (or TATE) can be identified based on the sample observations. One can then weight the sample appropriately towards the covariate distributions of the target population and compute weighted estimators (both linear and doubly-robust) of the PATE.[45]

4 | TWO APPROACHES TO WEIGHTING

4.1 | The modeling approach

The modeling approach essentially focuses on maximizing the fit of a model for the unknown propensity score as a function of the observed covariates. In practice, the most common way of calculating the weights in the modeling approach is to consider

[4] These are analogous to the positivity and unconfoundedness assumptions in an observational study.

the following logistic regression model

$$\text{logit}\{e(X_i)\} = \beta_0 + \beta_1 X_{i1} + ... + \beta_p X_{ip}$$

The coefficients $(\beta_0, ..., \beta_p)$ are estimated using a maximum likelihood procedure. Denote the estimated coefficients by $(\hat{\beta}_0, ..., \hat{\beta}_p)$. The estimate of the ith propensity score is $\hat{e}(X_i) = \frac{\exp(\hat{\beta}_0 + \hat{\beta}_1 X_{i1} + ... + \hat{\beta}_p X_{ip})}{1 + \exp(\hat{\beta}_0 + \hat{\beta}_1 X_{i1} + ... + \hat{\beta}_p X_{ip})}$. The estimated propensity scores are then utilized to get the weights and relevant estimators of the treatment effects. One can also consider a logistic regression model on some transformations of the covariates, instead of the original covariates. In certain situations, subject-matter knowledge about the relationship between covariates and treatment can guide a suitable specification for the propensity score. Alternative methods in the modeling approach use machine learning methods to fit more flexible models for the propensity score.[46] However, in general, it is very hard for an investigator to come up with the right propensity score model in an observational study.[5]

Now, it is worth repeating that the weights obtained under the modeling approach are not guaranteed to satisfy the desiderata for weighting mentioned in the previous section, especially the properties of balance and stability. If the propensity scores are known, the IPW will create a weighted sample that is stochastically balanced in terms of the distribution of covariates. This is because, for any (measurable) function $h : \mathbb{R}^k \to \mathbb{R}$, we have

$$\mathbb{E}\Big[\frac{h(X_i)Z_i}{e(X_i)}\Big] = \mathbb{E}\Big[\frac{h(X_i)}{e(X_i)}\mathbb{E}[Z_i|X_i]\Big] = \mathbb{E}[h(X_i)] = \mathbb{E}\Big[\frac{h(X_i)(1-Z_i)}{1-e(X_i)}\Big] \tag{4.1.1}$$

However, (4.1.1) may not hold, even asymptotically, if the working model for the propensity score is misspecified. Moreover, (4.1.1) does not imply balance in finite samples of any transformation of $h(\cdot)$ of the covariates. Therefore, given a particular specification of a statistical model for the propensity score, even if one obtains the weights by optimizing the fit of that model to the observed data, the job is far from over. In order for the resulting weighted data to be taken to the subsequent stage of outcome analysis, it should pass several diagnostic tests for balance and stability, such as the ones discussed in the previous section. If the estimated scores $\hat{e}(X)$ are close to zero for some treated units (or close to one for some control units), the corresponding estimated inverse probability weights would be so high that when normalized, a few units would dominate the analyses (leading to a markedly small effective sample size). Such extremeness and high dispersion in the weights can lead to unstable treatment effect estimators. It is sometimes recommended to truncate the extreme weights. In that case, one needs to decide on a threshold for truncating (e.g., the 95th percentile of the weights). One issue with truncating is that, once the weights have been truncated, the investigator has to re-evaluate balance in the weighted sample. If the resulting covariate balance is unsatisfactory after truncating, either the truncating threshold or the entire propensity model has to be modified and weights need to be re-computed. Besides, truncating the weights would mean that the investigator does not believe in the propensity model that he/she posited in the first place. See Figure 2 for a flowchart of the modeling and balancing approaches to weighting.

4.2 | The balancing approach

We now turn our attention to a more recent approach to weighting that differs from the modeling approach in terms of its motivation and implementation. Instead of optimizing the fit of the propensity score model, the balancing approach directly optimizes towards covariate balance in the weighted sample. However, as discussed in Section 3, the stability of the weights is another important aspect to consider in addition to balance and positiveness. Here, we illustrate a broad class of weighting methods which fall under the balancing approach termed minimal dispersion approximately balancing weights[21] or simply minimal weights. Under this approach, the investigator proposes a set of covariate balance requirements and solves for the weights that minimize a measure of dispersion of the weights subject to these requirements. Examples of this approach include the methods by Hainmueller (2012),[17] Zubizarreta (2015),[10], and Chan et al. (2016)[18]. Related approaches are Imai and Ratkovic (2014),[14] Fan et al. (2016),[16] Tan (2020),[47] Zhao and Percival (2017),[48] Li et al. (2018),[39] Wong and Chan (2018),[32] Yiu and Su (2018),[49] Zhao (2019),[50] and Ning et al. (2018).[51] Unless stated otherwise, in the rest of the paper, we will use the terms minimal weights and the balancing approach to weighting interchangeably, although the latter class is more general.

Suppose we are interested in estimating the ATT. In that case the strategy should be to weight the individuals in the control group such that the resulting weighted control units have similar covariate distributions as the treated units. Denoting $\boldsymbol{w} =$

[5] As Kang and Schafer (2007)[11] point out, "Except by divine revelation, it is unlikely that an analyst who sees only x_i would ever formulate a correct π-or y-model. Rather, he or she would naturally be drawn to models that are linear and logistic in the x_{ij}'s and those incorrect models look trustworthy."

$(w_1, ..., w_n)$ as the vector of weights for n units, the minimal weighting method chooses an optimal $\boldsymbol{w}$ by solving the following convex optimization problem

$$\begin{aligned} \underset{\boldsymbol{w}}{\text{minimize}} \quad & \sum_{i:Z_i=0} \psi(w_i) \\ \text{subject to} \quad & \Big| \sum_{i:Z_i=0} w_i B_k(X_i) - \frac{1}{n_t} \sum_{i:Z_i=1} B_k(X_i) \Big| \le \delta_k,\ k = 1, 2, ..., K \end{aligned} \tag{4.2.1}$$

Let us now scrutinize each component of the above optimization problem. First, $\psi : \mathbb{R} \to \mathbb{R}$ is a convex function of the weights. $\boldsymbol{B}(X) = \left(B_1(X), B_2(X), ..., B_K(X)\right)^T$ is a vector of K many real-valued functions or transformations of the covariate vector X. The kth constraint restricts the absolute value of the difference of the weighted mean of the $B_k(X_i)$s in the control group and the mean of that in the treatment group to be no larger than a pre-specified tolerance level denoted by $\delta_k (\ge 0)$. For instance, if B_k maps the entire covariate vector to its jth component, then the kth constraint directly controls the absolute mean difference of the jth covariate in the weighted control and treatment group to a desired level set by the investigator. Setting δ_k to zero would make the kth constraint exact as opposed to approximate. If the targeted estimand is the ATE, then we solve this optimization problem twice. In the first step, we solve for the weights in the control group by restraining the difference between the weighted mean of the $B_k(X_i)$s among the control units and the unweighted mean of the $B_k(X_i)$s among the full sample of n units, for $k = 1, ..., K$. We then weight the treated units in a similar manner to balance their covariate distributions relative to the full sample.

A special case of minimal weights are the Stable Balancing Weights (SBW).[10] SBW corresponds to the case where $\psi(w) = (w - 1/n_c)^2$. Apart from the K balancing constraints, the weights for the control units are constrained to be non-negative. Thus, SBW forces all weight based adjustments to be an interpolation of the available data, thereby guaranteeing sample boundedness of all Hajek type estimators. Furthermore, the weights generated by SBW are usually normalized, so that they add up to one. Note that, normalization can be imposed on the weights in the optimization problem itself by setting B_k to be the constant function 1 and δ_k to be zero. However, we keep this normalization constraint separate from the balancing constraints, as they serve different purposes. Therefore, under SBW, the optimization problem in (4.2.1) takes the following form:

$$\begin{aligned} \underset{\boldsymbol{w}}{\text{minimize}} \quad & \sum_{i:Z_i=0} (w_i - \bar{w}_c)^2 \\ \text{subject to} \quad & \Big| \sum_{i:Z_i=0} w_i B_k(X_i) - \frac{1}{n_t} \sum_{i:Z_i=1} B_k(X_i) \Big| \le \delta_k,\ k = 1, 2, ..., K-2 \\ & \sum_{i:Z_i=0} w_i = 1 \\ & w_i \ge 0,\ i : Z_i = 0 \end{aligned} \tag{4.2.2}$$

where $\bar{w}_c$ is the average of the control weights. The SBW method thus directly attempts to minimize the variance of the control weights subject to the given positivity and balancing constraints. Note that the objective function in (4.2.2) can be written as $\sum_{i:Z_i=0} w_i^2 - \frac{1}{n_c}$ subject to the given constraints. Thus, SBW attempts to minimize $\sum_{i:Z_i=0} w_i^2 = \sum_{i=1}^n (1 - Z_i) w_i^2$ subject to the given constraints. Hence, (4.2.2) relates to an analog of the variance formula in (3.2.2) (where the estimand of interest is the ATT). Also, the effective sample size of the weighted control group with normalized weights is $\frac{1}{\sum_{i:Z_i=0} w_i^2}$ (by Kish's approximate formula).[35] Minimizing the variance of the weights in the control group is, therefore, equivalent to maximizing the effective sample size of the weighted control group. In this spirit, SBW is similar to cardinality matching[52] where one aims to find the pair-matched sample of highest size that satisfies certain pre-specified approximate balancing constraints.

In order to estimate the ATE through SBW, we need to solve the above optimization problem twice — first to find the weights for the control group that match its covariate distribution to the full sample of treated and control units, followed by the finding the weights for the treatment group with the same property. From the above optimization problem, it is evident that by suitably selecting the functions B_k, one can balance not only the means of the original covariates, but also higher order moments of a covariate, interactions of two or more covariates, and even quantiles of the covariates. Thus, in the balancing approach, instead of directly specifying a probability model for treatment propensities, the investigator only has to specify the desired balancing criterion to obtain a set of weights that satisfy the given balance requirements by design. Moreover, the convex optimization in (4.2.2) can be solved in polynomial time and thus is suitable for dealing with large data sets.

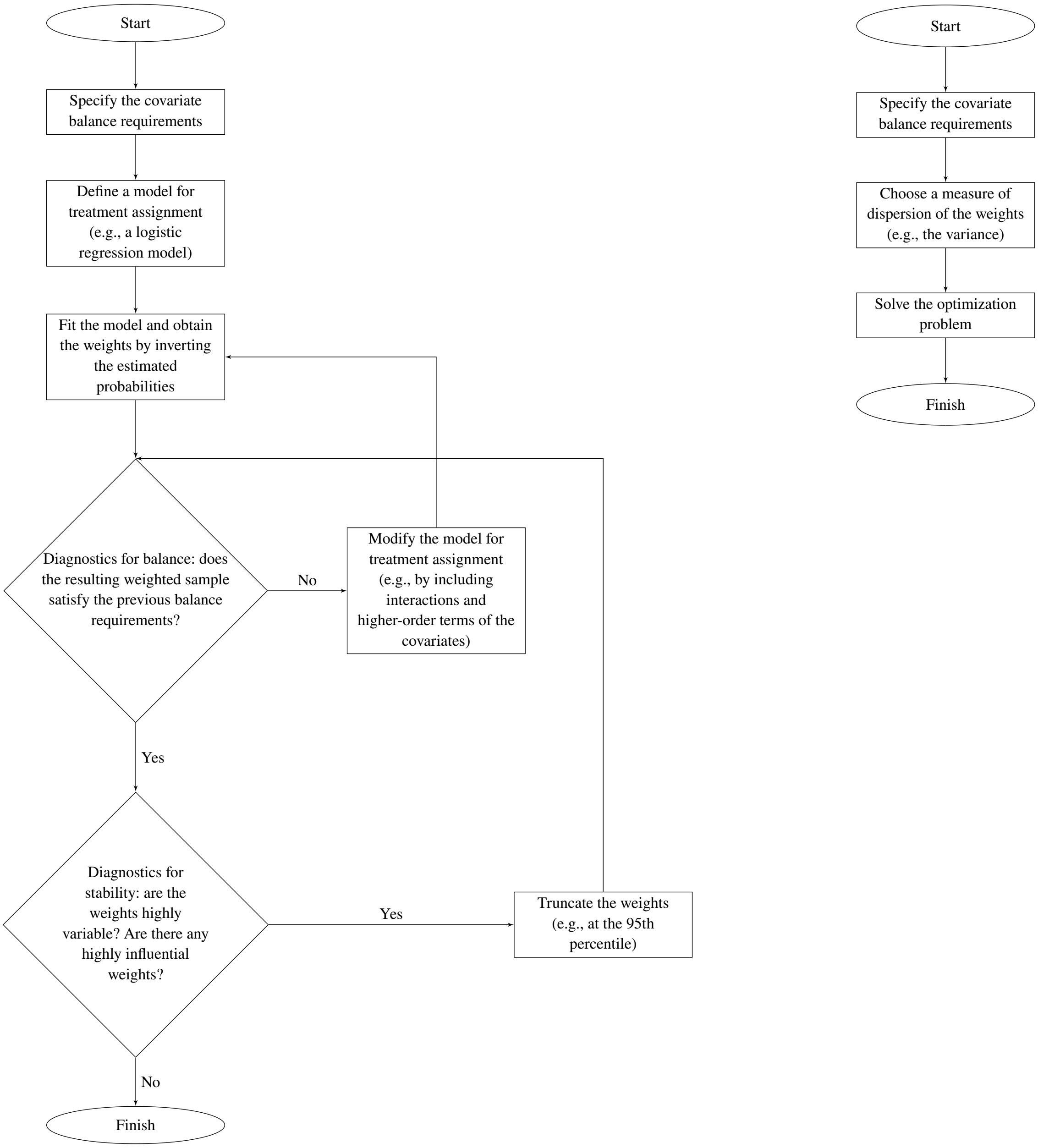

Start
Specify the covariate balance requirements
Define a model for treatment assignment (e.g., a logistic regression model)
Fit the model and obtain the weights by inverting the estimated probabilities
Diagnostics for balance: does the resulting weighted sample satisfy the previous balance requirements?
No
Modify the model for treatment assignment (e.g., by including interactions and higher-order terms of the covariates)
Yes
Diagnostics for stability: are the weights highly variable? Are there any highly influential weights?
Yes
Truncate the weights (e.g., at the 95th percentile)
No
Finish
Start
Specify the covariate balance requirements
Choose a measure of dispersion of the weights (e.g., the variance)
Solve the optimization problem
Finish

FIGURE 2 Flowcharts for the modeling (left) and balancing (right) approaches to weighting.

4.3 | Connection between the two approaches

The desirable properties of a weighting method outlined in Section 3 provide a motivation for the balancing approach to weighting. This is evidenced by the primal optimization problem for the minimal weights, given in (4.2.1). By contrast, owing to the balancing properties of the propensity score, estimation of the propensity scores has been the main focus of the modeling approach to weighting in practice. However, even though minimal weighting is designed from the perspective of diagnostics, Wang and Zubizarreta (2020)[21] showed that, the optimization problem (4.2.1) implicitly fits a model for the propensity score. The resulting minimal weights can thus be viewed as estimates of the true (normalized) inverse probability weights, where the true inverse probability weights are estimated through the implied propensity score model. This connects the optimization problem under the balancing approach to the seemingly disjoint estimation problem under the modeling approach.

In particular, investigating the dual formulation of (4.2.1) enables us to identify this connection between the two approaches. Let $\boldsymbol{w}^* = (w_1^*, w_2^*, ..., w_n^*)$ be the optimal set of weights obtained by solving the primal problem in (4.2.1). We can then express

$\boldsymbol{w}^*$ as

$$w_i^* = \rho'\{\boldsymbol{B}(X_i)^T\lambda^*\} - \frac{1}{n_t} = \rho'\{\lambda_1^* B_1(X_i) + ... + \lambda_K^* B_K(X_i)\} - \frac{1}{n_t}, \quad i = 1, 2, ..., n \tag{4.3.1}$$

where $\lambda^* = (\lambda_1^*, ..., \lambda_K^*)$ is the solution to the corresponding dual problem of (4.2.1) and ρ is a real valued differentiable function which depends on the objective function ψ in (4.2.1). The dual problem can equivalently be written in the following form

$$\underset{\lambda}{\text{minimize}} \quad \sum_{i=1}^{n} \Big[-(1 - Z_i)\rho\{\boldsymbol{B}(X_i)^T\lambda\} + \frac{\boldsymbol{B}(X_i)^T\lambda}{n_t} \Big] + |\lambda|^T\boldsymbol{\delta} \tag{4.3.2}$$

Here λ is a $K \times 1$ vector of dual variables and $\boldsymbol{\delta} = (\delta_1, ..., \delta_K)^\top$ is the vector of balance tolerances used in (4.2.1). The derivation is provided in the Appendix.

Now, suppose we decide to estimate the propensity score using a generalized linear model of Z_i on $\boldsymbol{B}(X_i)$ with the link function $g\{e(x)\} = (\rho')^{-1}\big[\frac{1}{n_t\{1-e(x)\}}\big]$. More precisely, we model the propensity scores as $e(X_i) = 1 - \frac{1}{n_t\rho'\{\boldsymbol{B}(X_i)^T\lambda\}}$. Moreover, suppose that instead of estimating λ through standard maximum likelihood estimation, we estimate it by minimizing the objective function given in (4.3.2). Note that (4.3.2) is nothing but a regularized version of the loss function $L(\lambda) = \sum_{i=1}^{n}\big[-(1-Z_i)\rho\{\boldsymbol{B}(X_i)^T\lambda\} + \frac{\boldsymbol{B}(X_i)^T\lambda}{n_t}\big]$, regularized by the L_1 penalty on λ. In this case, the resulting estimated inverse probability weights are proportional to the primal solutions $\boldsymbol{w}^*$. Thus, computation of optimal weights through minimal weighting is equivalent to IPW based on a shrinkage estimation of propensity score.

A similar connection between the modeling and balancing approaches has been explored by Tan (2020).[47] Leveraging the idea of calibration in survey sampling, Tan (2020)[47] proposed to estimate the propensity score using a logistic link function and an L_1 regularized version (with tuning parameter κ) of the calibration loss, which is the loss function implied by the estimation of propensity scores subject to exact balancing constraints. The regularized calibration loss is given as follows.

$$\underset{\lambda}{\text{minimize}} \quad \sum_{i=1}^{n} \big[(1 - Z_i)\exp\{-\lambda^T B(X_i)\} + Z_i\lambda^T B(X_i)\big] + \kappa||\lambda||_1 \tag{4.3.3}$$

The estimated propensity scores are given by $\hat{e}(X_i) = \frac{\exp\{\hat{\lambda}^T B(X_i)\}}{1+\exp\{\hat{\lambda}^T B(X_i)\}}$, where $\hat{\lambda}$ is the solution of (4.3.3). Using both theory and empirical evaluations, Tan (2020)[47] showed that these estimated propensity scores estimate the true propensity scores more accurately than the model-based estimates of the propensity scores in terms of mean squared relative errors. Controlling the relative error of the propensity score provides more control over the errors of the resulting IPW estimators. Tan (2020)[47] also showed that the dual problem of (4.3.3) has the same form as the optimization problem in entropy balancing,[17] but with approximate balancing constraints (with uniform tolerance κ) as opposed to exact constraints. In this sense, the regularized calibration weights can be regarded as a special case of minimal weights.

From a finite sample perspective, the balancing approach emphasizes estimating the propensity score in such a way that the induced weighted sample is balanced in terms of the empirical distribution of the observed covariates. Mathematically, the balancing and modeling approaches to weighting are equivalent, as the balancing approach is actually modeling the propensity scores with a loss function that directly targets covariates and restrains the dispersion of the weights.

4.4 | More on the balancing approach

4.4.1 | Asymptotic properties

There is a rigorous literature on the asymptotic properties of propensity score methods (and IPW with estimated propensity scores, in particular) in the context of average treatment effect estimation.[53,7,54] The connection between minimal weights and shrinkage estimation of the propensity scores allows several attractive large sample properties to be established for the balancing weights and the corresponding weighted estimators. Wang and Zubizarreta (2020)[21] have established these properties in the framework of estimation with incomplete outcome data, but they carry over to the setup of estimating average treatment effects in causal inference. The properties rely on smoothness conditions on the propensity score function and the mean function of the response given covariates. The following briefly summarizes these results.

First, under certain assumptions, the minimal weights are consistent estimators of the true inverse probability weights. At a high level, the underlying assumptions mainly include standard regularity conditions required for consistency of minimum risk estimators, mild restrictions on the structure of the objective function ψ and the vector of basis functions $\boldsymbol{B}(X)$, in addition to a uniform approximability condition that essentially means that the propensity score function can be well-approximated by the basis functions (through the inverse link) uniformly over the covariate space. Second, Wang and Zubizarreta (2020)[21] provide asymptotic results of the standard linear weighted estimator, which has the same form as the Hajek estimator in our case. More specifically, they show that under additional assumptions on the conditional mean function of the outcome given covariates,[6] the weighting estimator based on the minimal weights is consistent for the population mean of the outcome variable (which translates to the average treatment effect in the causal inference setting) and asymptotically Normally distributed. Furthermore, the asymptotic variance of this estimator attains the semiparametric efficiency bound. Hence, similar to the weighting estimators in Hirano et al. (2003),[7] Chan et al. (2016),[18] and Fan et al. (2016),[16] minimal weights lead to a semiparametrically efficient estimator of the average treatment effect.

4.4.2 | An algorithm for selecting the degree of approximate balance

An important question in practice is how to select the tuning parameters δ_k, $k = 1, ..., K$, which constrain the covariate imbalances in the functions $\boldsymbol{B}_k$ of the covariates and regulate the trade-off between the degree of approximate balance and the dispersion of the weights. Small values of these parameters force the weighted sample to be tightly balanced, but potentially come at the cost of highly dispersed weights or a lower effective sample size after weighting. Similarly, high values of the parameters relax the feasible set of the weights and lead to more uniform weights, and thus, to a higher effective sample size, but at the likely cost of higher covariate imbalances.

Suppose that the investigator has substantive knowledge of which functions of the covariates are relevant. Alternatively, the investigator can learn these functions, e.g., by an outcome-adaptive Lasso[55] in split samples, where a separate, small planning sample is used to guide the study design.[56] Unless the investigator has strong a priori knowledge of which covariate functions are more prognostically important than others, he/she may want to choose the parameters δ_k through a data-driven procedure, as opposed to setting them manually. Wang and Zubizarreta (2020)[21] proposed a bootstrap-based algorithm to select a common tuning parameter for all balancing constraints in the minimal weighting optimization problem (see also Zhao 2019).[50] Here we present a slight variant of that algorithm.

Algorithm 1 Selection of uniform tuning parameter δ for minimal weights

```
Fix $\mathcal{D}$, the grid of covariate imbalances (in units of standard deviation).
for $\delta \in \mathcal{D}$, do
    Compute the weights $w_i$, $i = 1, ..., n$, for the original sample by solving (4.2.1) with tolerance $\delta$ (in standard deviations) for all the balancing constraints.
    for $b \in \{1, ..., N_{boot}\}$, where $N_{boot}$ is the number of bootstrap samples, do
        Draw a bootstrap sample $\mathcal{S}_b$ from the original sample.
        for $k \in \{1, ..., K\}$, where $K$ is the number of balancing constraints, do
            Calculate the covariate imbalance measure $C_{k,b}(\delta)$ corresponding to the $k$th balancing constraint on $\mathcal{S}_b$.
        end
        Compute the mean imbalance for $b$th bootstrap sample, i.e., $\xi_b(\delta) = \frac{1}{K}\sum_{k=1}^{K} C_{k,b}(\delta)$.
    end
    Compute the average imbalance over all bootstrap samples, i.e., $\Xi(\delta) = \frac{1}{N_{boot}}\sum_{b=1}^{N_{boot}} \xi_b(\delta)$ .
end
Choose $\delta^* = \arg\min_{\delta \in \mathcal{D}} \Xi(\delta)$.
```

When the estimand of interest is the ATT, the covariate imbalance measure $C_{k,b}(\delta)$ is given by

$$C_{k,b}(\delta) = \Big| \sum_{i \in \mathcal{S}_b} (1 - Z_i) w_i B_k(X_i) - \frac{1}{\sum_{i \in \mathcal{S}_b} Z_i} \sum_{i \in \mathcal{S}_b} Z_i B_k(X_i) \Big|$$

Wang and Zubizarreta (2020)[21] considered drawing K independent bootstrap samples corresponding to each of the K balancing constraints and calculating the average of the covariate imbalance measures across all the bootstrap samples. In Algorithm

[6] Notably, they consider a similar uniform approximability assumption of the mean function by the chosen basis functions $\boldsymbol{B}(x)$.

1, instead of computing imbalance in the bootstrap sample for one constraint, we average the imbalances over all the constraints. Aside from computing a simple average of the imbalances in line 9 of Algorithm 1, one can also consider calculating the maximum imbalance for the bth bootstrap sample, i.e., $\psi_b(\delta) = \max_{k=1,...,K} C_{k,b}(\delta)$ for $b \in \{1, 2, ..., N_{boot}\}$.

Now, since the sample is assumed to be drawn randomly from a population, and by (4.1.1), the true IPW would balance, on an average, any transformation of the covariates over different random draws from that population. Leveraging this idea and the fact that the minimal weights are implicitly estimating the IPW, we evaluate the covariate imbalances over different bootstrap samples. Here the bootstrap samples work as proxy random samples from the target population.

How should one choose the grid of covariate imbalances $\mathcal{D}$ in practice? From an asymptotic perspective, the imbalance δ should be smaller than $K^{-\frac{1}{2}}$ in order for the consistency results to hold.[21] In practice, a standardized mean difference no larger than 0.1 (or 0.2) is commonly considered as satisfactory covariate balance.[57] However, this threshold is not formally justified, and covariate balance diagnostics, not unlike regression diagnostics, tend to be informal in nature (see Chapter 9 of Rosenbaum 2010 for a related discussion).[58] Based on these considerations, we recommend setting the grid values in $\mathcal{D}$ to be smaller than $\min(0.1, K^{-\frac{1}{2}})$.

Now, if there is limited overlap in covariate distributions, it may not be possible to restrain the imbalances of all the functions B_k to be less than any $\delta \in \mathcal{D}$. In that case, the balancing approach will not produce a solution (i.e., there are no non-negative weights that approximately balance the functions of the covariates as permitted by $\mathcal{D}$). Then, one might resort to other modes of analysis, such as specifying a parametric model for the potential outcomes. However, such analysis would require extrapolating across non-overlapping portions of the covariate space and producing results that are highly sensitive to model misspecification. As we know, estimates based on a misspecified outcome model can be severely biased. In light of these issues, we view the infeasibility of the balancing approach as a helpful diagnostic for practical violations to the positivity assumption. An infeasible weighting solution for any $\delta \in \mathcal{D}$ helps the investigator to realize that there are no weights that can yield satisfactory balance by interpolating for the data at hand.

5 | A SIMULATION STUDY

In this section, we conduct a simulation study in order to evaluate the finite sample performances of the estimators of the ATE defined in Section 2.4 computed under different weighting methods. Here we use a simulation design originally explored in Hainmueller (2012).[17]

5.1 | Study design

There are six observed covariates $X_1, X_2, ..., X_6$ where

$$\begin{aligned} \begin{pmatrix} X_1 \\ X_2 \\ X_3 \end{pmatrix} &\sim \mathcal{N}_3\left[\begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 2 & 1 & -1 \\ 1 & 1 & -0.5 \\ -1 & -0.5 & 1 \end{pmatrix}\right] \\ X_4 &\sim Unif[-3, 3] \\ X_5 &\sim \chi^2_1 \\ X_6 &\sim Bernoulli(0.5) \end{aligned}$$

Here X_4, X_5 and X_6 are mutually independent and are independent of $(X_1, X_2, X_3)^\top$. The set of observed covariates consists of both continuous and binary variables. Given the six covariates, the treatment indicator is given by $Z = \mathbb{1}(X_1 + 2X_2 - 2X_3 - X_4 - 0.5X_5 + X_6 + \epsilon > 0)$. The error term ϵ is a Normal random variable with mean zero and variance σ^2_ϵ. One can tune the degree of overlap between the treatment and control group by appropriately regulating the value of σ^2_ϵ. Following Hainmueller (2012),[17] we consider two designs corresponding to two distinct values of the error variance: $\sigma^2_\epsilon = 100$ (we call this the strong overlap case) and $\sigma^2_\epsilon = 30$ (we call this the weak overlap case). Finally, given a data generating mechanism for $\boldsymbol{X} = (X_1, ..., X_6)^\top$ and Z, the potential outcomes $(Y(0), Y(1))$ are generated using two different outcome designs:

(A) $Y(0) = X_1 + X_2 + X_3 - X_4 + X_5 + X_6 + \eta, Y(1) = Y(0)$

(B) $Y(0) = (X_1 + X_2 + X_5)^2 + \eta, Y(1) = Y(0)$

In each case, η follows a standard Normal distribution and is independent of the covariates. A random sample of size $n = 600$ is drawn from each of the above data generating designs. A few observations are in order. First, in all the cases, the unit-level causal effects are zero for all units, leading to an ATE of zero. Second, it follows from the definition of the propensity score that a probit model regressing the treatment indicator on the original covariates is correctly specified. Finally, for the outcome designs, in Design (A), the potential outcomes are linearly related to the original covariates. In Design (B), on the other hand, the potential outcomes are linearly related to transformations of the original covariates. Moreover, all of the covariates are prognostically important in Design (A), but not in Design (B).

5.2 | Simulation results

We compute 51 estimators in total, 48 of them based on weights (some, in combination with some form of outcome regression) and 3 of them solely based on outcome regression model fits. The full set of weight-based estimators can be cross-classified according to the weighting method used and the way the weights are incorporated in computing the estimator. Five different types of weighting methods are compared, which are labelled as Simple Logistic, Complex Logistic, SBW-1, SBW-2 and SBW-3. Detailed description of the labels are given in Table 1.

TABLE 1 Description of weighting methods used

Label	Weighting Method
Simple Logistic	Logistic regression model with all original covariates considered in the linear predictor.
Complex Logistic	Extension of Simple Logistic with the addition of squares of the covariates and all possible pairwise products of the covariates.
SBW-1	SBW with approximate mean balancing constraints on all original covariates at a common tolerance level chosen using the tuning algorithm.
SBW-2	Extension of SBW-1 by imposing additional approximate balancing constraints for decile indicators for all continuous covariates with a common tolerance level chosen using the tuning algorithm.
SBW-3	Extension of SBW-1 by imposing additional approximate balancing constraints for second order moments of all original covariates and first order moments of all pairwise products of the covariates. The common tolerance level of the approximate balancing constraints is chosen using the tuning algorithm.

Using these weighting methods, we compute the Horvitz-Thompson, Hajek, and the residual bias corrected doubly robust (DR) estimator of the ATE (defined in Section 2.4).[7] We use three types of regression model specifications for the outcome variable and use it to estimate the ATE either through g-computation (without taking the weights into account) or through a doubly robust estimator (which takes the weights into account). Specifically, we consider a linear regression on the original covariates, Bayesian Additive Regression Trees (BART),[59,60] and a random forest based on the original covariates. In order to allow for treatment-covariate interactions, we fit two separate outcome regression models: one with the observations in the treatment group and the other in the control group. We have kept the same specification for both the outcome regression models (e.g., if we fit a linear regression model for Y^{obs} on $\boldsymbol{X}$ in the treatment group, then we also fit a linear regression model for Y^{obs} on $\boldsymbol{X}$ in the control group).

[7] The results in the simulation for the 'DR Horvitz-Thompson' and corresponding 'DR Hajek' estimators turn out to be almost identical. We only report the results for 'DR Hajek' in this paper.

The following two subsections contain the results of the simulation study under the strong overlap and the weak overlap case. In each case, we compare the performance of the estimators described earlier for outcome designs (A) and (B). The R code for this simulation study is available on the corresponding author's website.[8]

5.2.1 | Strong overlap

The strong overlap case corresponds to the scenario where the variance of ϵ is $\sigma_\epsilon^2 = 100$. Note that the true propensity score under the simulation design outlined in Section 4.1 is given by $e(\boldsymbol{X}) = 1 - \Phi\left\{\frac{-g(\boldsymbol{X})}{\sigma_\epsilon}\right\}$ where $g(\boldsymbol{X}) = X_1 + 2X_2 - 2X_3 - X_4 - 0.5X_5 + X_6$ and Φ denotes the CDF of a standard Normal distribution. A high value of σ_ϵ^2 tends to wash out the effect of $g(\boldsymbol{X})$ in the treatment selection process, leading to a high degree of overlap between the treatment and control groups.

Table 2 contains the Monte Carlo estimates of the bias and RMSE (multiplied by 100) of all the estimators defined earlier based on $M = 800$ simulations, for both outcome designs (A) and (B).

TABLE 2 Monte Carlo estimates of Bias and RMSE for different estimators for outcome designs A and B in the *strong overlap* case. The tabulated values have been multiplied by 100.

Estimators	Weights	Outcome design A		Outcome design B	
		Bias	RMSE	Bias	RMSE
Horvitz Thompson	Simple Logistic	1.31	12.43	2.04	112.80
	Complex Logistic	1.31	15.50	-3.56	73.48
Hajek	Simple Logistic	1.36	12.09	1.89	103.99
	Complex Logistic	1.79	14.62	-1.50	62.45
	SBW-1	1.76	9.41	-2.67	87.59
	SBW-2	1.22	10.06	-2.37	80.60
	SBW-3	3.15	10.44	-1.74	15.26
DR Hajek (Linear)	Simple Logistic	-0.13	9.40	0.80	117.21
	Complex Logistic	-0.05	9.75	-0.38	61.47
	SBW-1	-0.21	9.17	-0.54	88.07
	SBW-2	-0.19	9.74	-1.07	79.69
	SBW-3	-0.16	9.61	-3.20	19.38
DR Hajek (BART)	Simple Logistic	1.01	10.14	20.10	44.83
	Complex Logistic	1.05	10.26	19.91	44.10
	SBW-1	1.23	10.13	20.18	44.81
	SBW-2	1.06	10.17	20.03	44.75
	SBW-3	1.31	10.19	19.62	43.73
DR Hajek (RF)	Simple Logistic	10.32	15.13	7.15	69.10
	Complex Logistic	10.13	15.14	5.67	54.37
	SBW-1	11.12	15.58	4.88	68.26
	SBW-2	10.06	14.80	5.12	67.20
	SBW-3	11.27	15.62	3.93	49.43
g-computation (Linear)		-0.2	9.17	-0.79	89.20
g-computation (BART)		1.70	10.20	20.20	44.83
g-computation (RF)		18.00	22.04	10.24	72.04

For outcome design (A), we observe that regardless of the underlying weighting method used, the bias and RMSE values are considerably small. This should come as no surprise since the outcome model is linear in the observed covariates and all the propensity score specifications as well as the balancing constraints in SBW include the observed covariates. However, it is evident from Table 2 that among the linear weighted estimators (i.e., Hajek estimators), those incorporating SBW-based weights have the lowest RMSEs. As we increase the number of constraints in SBW (i.e., SBW-2 and SBW-3), the Hajek estimators

[8] We used the R package `sbw` (version 1.0) for computing the SBW weights.

show an increase in RMSE, although not exceeding the RMSE of the modeling counterparts. Here SBW-2 and SBW-3 attempts to restrict the difference in means of a multitude of functions of the covariates in addition to the correct functions. It is thus apparent that overbalancing in SBW comes with a cost of increased RMSE, however, not to the extent of the weights based on the modeling approach.

For outcome design (A), two linear regression models (one in the treatment group and another in the control group) of the response variable on the observed covariates are correctly specified. Thus, a linear regression estimator or a DR estimator where a linear regression model (on the observed covariates) is used is consistent for the ATE, and Table 2 supports this proposition. The DR estimators based on a linear response model have similar performance in terms of bias and RMSE, regardless of the weighting method used. A similar pattern is noted for DR estimators involving other types of response models (e.g., BART and random forests), although the bias and RMSE for these estimators are higher than those for the DR Hajek Linear estimator. The important feature of this simulation design is that the degree of separation between the treatment and the control groups in terms of the covariate distribution is small; in other words, the two groups are not too imbalanced to begin with. Thus, in this case, properly weighted treatment and control groups can enhance the performance of a DR estimator only to a very small extent. The DR estimators with BART are nearly as accurate as those based on linear outcome models. However, estimators involving random forests perform poorly in terms of bias and RMSE in this setting. The poor performance of random forest may be attributed to the moderate sample size of 600. Since we fit two random forests in the treatment and control group separately, each of the models ends up utilizing a small number of observations.

The right half of Table 2 contains the bias and variance of all the estimators for outcome design (B). Unlike outcome model (A), the conditional mean function of the potential outcomes is linear in squares and products of a subset of covariates. Now, SBW-3 attempts to choose optimal weights by balancing all prognostically important transformations of the covariates.[9] Despite the existence of strong overlap between the treatment and the control group, the bias and RMSE of the resulting Hajek estimator based on SBW-3 is substantially lower than the Hajek estimators (as well as Horvitz-Thompson estimators) based on any other weighting method used. Complex Logistic, which also incorporates the squares and interaction terms along with the linear terms in the specification of the propensity score model, yields the smallest bias and RMSE among all the linear estimators based on the modeling approach.

Here, the DR estimators behave differently from the previous scenario in the sense that there appears to be a role reversal of the outcome model and choice of weighting method in determining the accuracy of the corresponding estimators. Note that a linear model for the potential outcomes on the covariates is misspecified. The performance of the linear regression estimator of ATE is strikingly improved by augmenting it with the weights given by SBW-3, as seen in Table 2. The use of weights under the balancing approach also helps improve the performance of random forest. However, quite surprisingly, BART appears to be relatively insensitive to the choice of weights augmented to form a DR estimator. Despite having high biases, DR estimators using BART have considerably smaller RMSEs than DR estimators using linear regression or random forests.

5.2.2 | Weak overlap

In this section, we consider the case where $\sigma_\epsilon^2 = 30$, denoting weak overlap. Under this scenario, the Monte Carlo estimates of bias and RMSE for outcome designs (A) and (B) are shown in Table 3.

For outcome design (A), the performance of the modeling approaches is worse (as compared to the strong overlap case), as evidenced by the bias and RMSE in the left half of Table 3. The deterioration in accuracy is more apparent among the Hajek estimators. This indicates that even though the propensity models include the right functions of the covariates, it is comparatively more difficult for modeling approaches to balance covariate distributions under weak overlap as compared to strong overlap. However, the accuracy measures of the Hajek estimators driven by balancing weights are of similar order to the former case. The balancing weights successfully remove covariate imbalance between the treatment and control groups relative to the full sample while simultaneously optimizing towards a stable set of weights, justifying the small bias and RMSE. The general accuracy patterns of the estimators under the strong overlap case carry forward to the weak overlap case. For instance, the overall performance of the DR estimators with a linear response model dominates the other forms of DR estimators in the study. Besides, the part of a DR estimator stemming from the outcome regression seems to overshadow the weighted adjustment, leading to similar order of bias (and RMSE) across estimators with different weighting methods and a fixed response model. The results for outcome design (B) are also similar to its strong overlap counterpart. More concretely, estimators built on SBW-3 weights have uniformly smaller RMSE than their competitors in the respective groups. Although SBW-1 does not explicitly force balance on

[9] SBW-3 also contains additional balance conditions for other functions of the covariates that are unrelated to the potential outcomes, e.g., X_3, X_4^2, X_3X_6, etc.

TABLE 3 Monte Carlo estimates of Bias and RMSE for different estimators for outcome designs A and B in the in the *weak overlap* case. The tabulated values have been multiplied by 100.

Estimators	Weights	Outcome design A		Outcome design B	
		Bias	RMSE	Bias	RMSE
Horvitz Thompson	Simple Logistic	7.95	30.17	22.48	189.41
	Complex Logistic	11.30	36.45	21.95	146.99
Hajek	Simple Logistic	9.24	28.14	25.28	155.52
	Complex Logistic	12.53	32.49	24.23	107.76
	SBW-1	2.06	9.54	9.54	93.05
	SBW-2	2.64	12.24	-1.11	118.84
	SBW-3	9.72	16.73	-12.55	38.19
DR Hajek (Linear)	Simple Logistic	-0.08	10.76	13.87	199.12
	Complex Logistic	0.06	11.59	11.28	134.13
	SBW-1	0.12	9.19	12.97	93.76
	SBW-2	0.02	11.34	1.16	117.10
	SBW-3	-0.05	10.83	-5.56	38.65
DR Hajek (BART)	Simple Logistic	4.41	12.83	39.11	70.51
	Complex Logistic	4.56	13.04	38.91	69.74
	SBW-1	5.42	12.83	39.29	70.34
	SBW-2	4.39	12.97	38.96	70.14
	SBW-3	5.28	13.04	38.30	69.06
DR Hajek (RF)	Simple Logistic	22.76	27.05	16.52	102.39
	Complex Logistic	22.84	27.33	13.50	83.69
	SBW-1	25.62	28.74	8.54	98.15
	SBW-2	22.12	25.66	9.36	98.72
	SBW-3	25.64	28.78	1.68	77.31
g-computation (Linear)		0.16	9.18	10.76	100.376
g-computation (BART)		6.10	13.13	39.34	70.44
g-computation (RF)		37.30	40.21	14.15	100.88

the squares and products of covariates, in this case, the resulting Hajek estimator has smaller bias and RMSE than Complex Logistic.

5.2.3 Summary of results and discussion

First, under both strong and weak overlap of the covariate distributions, the weighted estimators under the balancing approach tend to be at least as good as those based on the modeling approach. Second, when the correct transformations of the covariates are balanced through the balancing approach, the resulting estimators show markedly better performance than the estimators based on the modeling approach. On the other hand, when the balancing approach does not include the correct transformations in its balancing constraints (e.g., SBW-1 under outcome design (B)), we still obtain satisfactory weighted estimators, and in several cases it outperforms the corresponding weighted estimators incorporating modeling weights. This is an important observation from the perspective of Section 4.3, where we established that solving for the minimal weights is implicitly connected to fitting a model for the propensity score (using a certain loss function) and obtaining the inverse probability weights. In particular, the transformations of the covariates included in the balancing constraints are the same as the transformations that are combined linearly and linked to the propensity score through the function $g(\cdot)$ (in the spirit of generalized linear models). Therefore, in the context of this simulation study, SBW-1 and Simple Logistic essentially use the same linear predictor for their respective propensity score specifications, albeit with different link functions. The simulation results indicate that the weighted estimators with SBW-1 weights outperform the respective estimators with Simple Logistic weights almost uniformly in terms of RMSE. Third, in regards to the optimization problem under the balancing approach, we observe that overbalancing can hurt in terms of loss of accuracy of the weighted estimators. However, despite the overall increase in bias and RMSE, the estimators with minimal weights still perform as well as the estimators based on the modeling approach. These observations are congruent with the results of Tan (2020)[47] discussed in Section 4.3, which explain why the balancing approach is expected to perform better than the modeling approach, especially when the propensity score model is misspecified.[10]

We also evaluated the performance of SBW with exact as opposed to approximate balancing constraints. In particular, we considered SBW-1, SBW-2, and SBW-3 with $\delta_k = 0$. For some simulated data sets, these weighting optimization problems were infeasible, stating that for those given data sets, there is no set of non-negative weights that exactly balance the means of the required transformations of the covariates. In other words, for those data sets there is no sample bounded estimator that uses non-negative weights that perfectly balances the covariates. To the investigator, it may sound discouraging not to find a weighting solution, but we view it as helpful information about the data at hand, as it warns the investigator about practical violations of the positivity assumption. With this information, the investigator can choose to estimate the effect under stronger parametric assumptions or change the question asked of the data by targeting a more local estimand.[38]

If we restrict the comparisons to those simulated data sets under outcome design (A) and the strong overlap case ($\sigma^2_\epsilon = 100$) where the exact balancing version of SBW-1 is feasible, we find that the resulting Hajek estimator has bias and RMSE of 0.03 and 8.63 respectively, outperforming all the other Hajek estimators. On the other hand, under outcome design (B), the Hajek estimator has bias and RMSE of 1.51 and 92.11 respectively. Despite having a very small bias, the RMSE of this estimator is the highest among all the SBW-based Hajek estimators, which occurs due to the increased variability of the weights. Thus, exact balancing may be problematic if the functions of the covariates that are balanced are different from functions that actually intervene in the conditional means of the two potential outcomes. SBW with approximate balancing (the level of approximation chosen by the tuning algorithm) seem to exhibit desirable performances in both the cases.

6 | AN OBSERVATIONAL STUDY OF AIR POLLUTION

6.1 | The air quality data set

In this section, we analyze data from an observational study by Zigler et al. (2018).[61] The study evaluated the impact of ambient fine particulate matter $PM_{2.5}$ on health outcomes of Medicare beneficiaries. Counties were designated as 'attainment' areas and 'non-attainment' areas depending on whether they met the 1997 National Ambient Air Quality Standard (NAAQS) for $PM_{2.5}$. The non-attainment areas constituted the treatment group and the attainment areas constituted the control group. The treatment and control group sizes were $n_t = 292$ and $n_c = 537$, respectively, for a total of $n = 829$ counties in the study. For simplicity, in this section we restrict our attention to a subset of this data set, which includes the treatment indicator and a set of 16 baseline covariates (15 continuous and 1 binary).

We consider two weighting methods under the modeling and balancing approaches: Simple Logistic and Complex Logistic regression, and SBW-1 and SBW-3, respectively. The definitions of the weighting methods are the same as those given in Table 1, except for Complex Logistic and SBW-3. Here, Complex Logistic is based on a logistic regression of the treatment indicator on the original covariates and their squares. In the same spirit, SBW-3 includes approximate balancing constraints for the original covariates and their squares. In the following section, we will perform some of the diagnostics (outlined in Section 3) for each of the weighting methods considered. Specifically, we shall first check whether the weights induce satisfactory balance in the treatment and the weighted control group relative to the target, followed by an inspection of the degree of variability of the weights and the presence of extreme or highly influential weights. In the online supplementary materials, we perform similar diagnostics based on the widely used Lalonde study on the effectiveness of a labor training program.[62,63] We also provide estimates of the ATT and corresponding standard error estimates under different weighting methods. The code for replication for both studies is available on the corresponding author's website.[11]

6.2 | Diagnostics for balance

In order to perform targeted balance diagnostics, we compute the TASMD of the observed covariates and their squares for the weighted treatment and the weighted control group, relative to the target. Since the estimand of interest is the ATE, for any

[10] Although Simple Logistic and Complex Logistic included the right transformations of the covariates in their propensity score specifications, the propensity score models were still misspecified, as the true propensity score involved a probit link function. However, it is well-known that usually logit and probit functions lead to similar estimates of the response probabilities.

[11] We used the R package `sbw` (version 1.0) and the *quadprog* optimization solver for computing the SBW weights for both studies.

covariate (or any transformation of the covariate vector), the target value is the average of that covariate in the full sample. Figure 3 depicts the TASMDs in both treatment and control group relative to the full sample, for the 16 observed covariates in the data.

FIGURE 3 Plot of TASMD of the original covariates in the weighted treatment group (left) and the weighted control group (right) relative to the full sample, for different weighting methods. The dotted and dashed vertical lines are drawn at 0.1 and 0.2 TASMD, respectively.

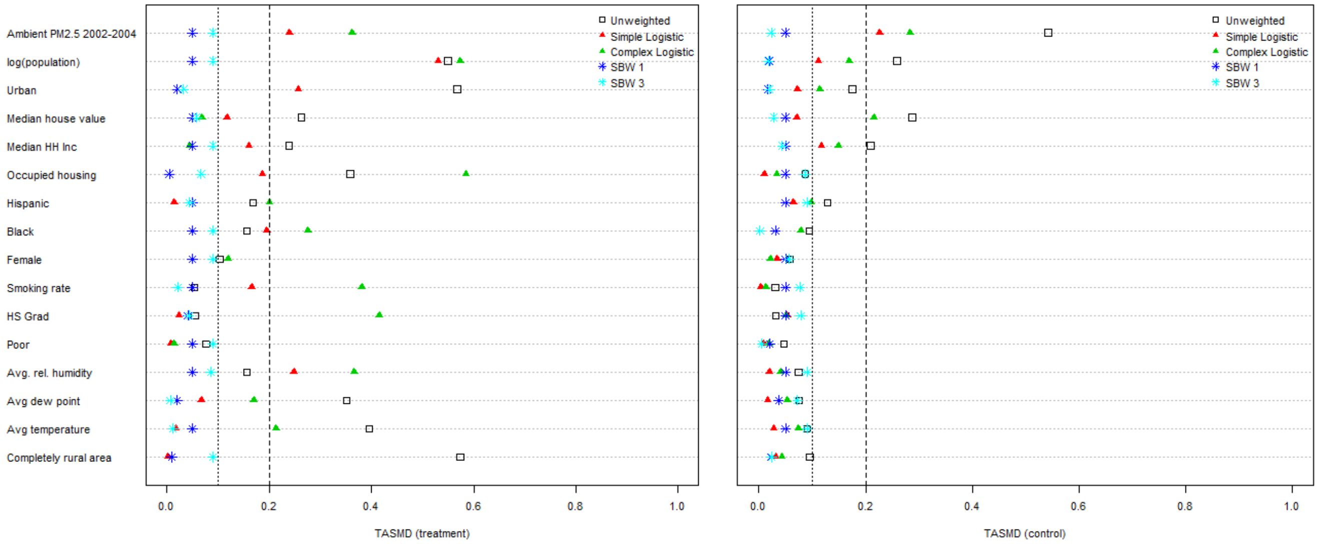
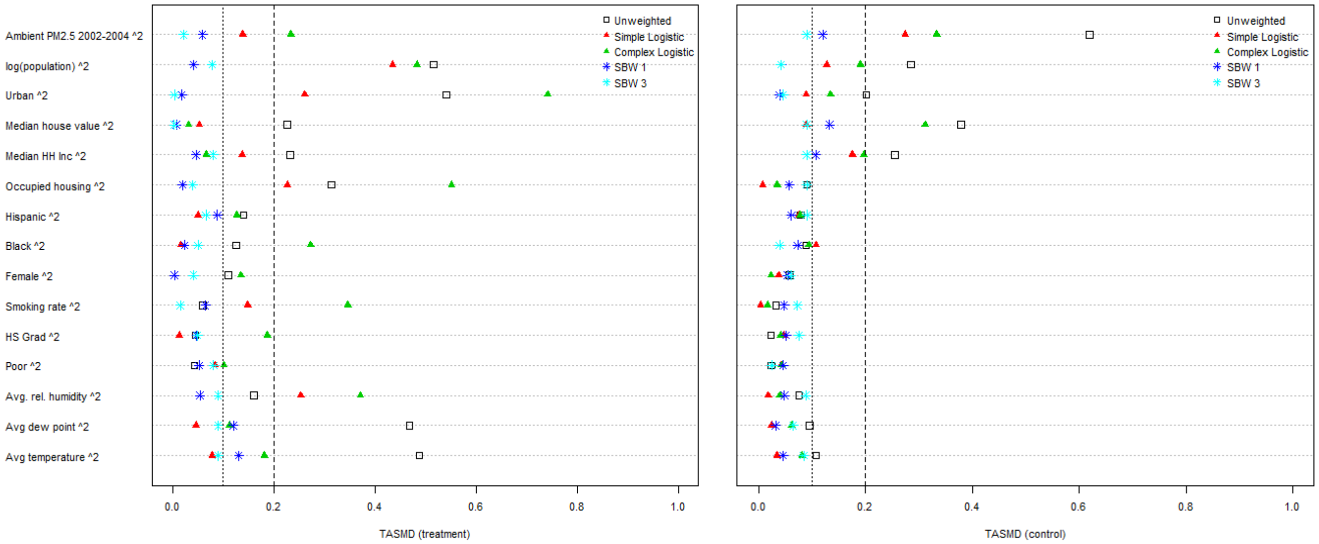


Figure 3 indicates that for most of the covariates, there are substantial mean imbalances in both treatment and control groups (greater than 0.2 TASMD) relative to the full sample before weighting, with the treatment group having a substantially higher number of cases of poorly balanced covariates. Overall, the modeling approach as implemented with logistic regression does a poor job in reducing the degree of imbalances below 0.1 TASMD, or even 0.2 TASMD. While Simple Logistic manages to adequately balance the means of covariates such as 'Avg dew point' and 'Completely rural area' relative to the target, it fails to sufficiently remove imbalances on covariate-means such as 'Ambient PM2.5 2002-2004,' 'log(population),' and 'Median HH Inc.' A considerable number of covariates remain highly imbalanced after weighting by Complex Logistic. Moreover, the degree of imbalances relative to the target for some covariates (e.g., 'HS Grad') become substantially exacerbated as compared to the unweighted sample. SBW-1, which implicitly fits the same propensity score model as Simple Logistic, almost always outperforms the modeling approaches in terms of having small TASMDs in the treatment group, while obtaining a better balanced sample overall in the control group with all TASMDs less than 0.1. In fact, the TASMDs yielded by both SBW-1 and SBW-3 are less than 0.1 for all the covariates and for both the treatment groups. On the other hand, in the treatment group, 13 covariates have TASMDs greater than 0.1 (11 greater than 0.2) under Complex Logistic and 10 covariates have TASMDs greater than 0.1 (5 greater than 0.2) under Simple Logistic.

In Figure 4, we display the TASMDs for the squares of the 15 continuous covariates using different weighting methods. Similar to the former case, the unweighted sample is highly imbalanced in terms of the second order moments of the covariates and weighting by the modeling approaches fails to yield satisfactory balance relative to the full sample for these functions. For the modeling approaches, the residual imbalance towards the target is present both in higher numbers and higher magnitude in the treatment group than in the control group. In the treatment group, among the squares of 15 covariates, 8 have TASMDs greater than 0.1 (5 greater than 0.2) under Simple Logistic and 13 covariates have TASMDs greater than 0.1 (7 greater than 0.2) under Complex Logistic. On the other hand, only 2 covariate-squares have TASMDs (in the treatment group) strictly greater than 0.1 under SBW-1; no covariate-square has TASMD greater than 0.2, and none have TASMDs greater than 0.1 under SBW-3. Interestingly, despite not restraining the mean imbalance of the squares of the covariates explicitly, SBW-1 successfully induces a weighted sample which is reasonably balanced towards the target with respect to the said transformations.

FIGURE 4 Plot of TASMD of the squares of the original covariates for the weighted treatment group (left) and the weighted control group (right) relative to the full sample, for different weighting methods. The dotted and dashed vertical lines are drawn at 0.1 and 0.2 TASMD, respectively.



Note that here we have chosen the degree of approximate covariate balance using Algorithm 1; we could have set a δ manually such that if feasible, all the TASMDs are near zero by construction. However, doing so could have resulted in a considerably larger variance of the weights that does not generalize well according to the algorithm.

6.3 | Diagnostics for stability

The next stage of weight diagnostics comprises checks for variability and extremeness of the weights. Following Section 3, we compute different measures of stability of the (normalized) weights; namely, the standard deviation, 99th percentile, and maximum value. We also tabulate the total effective sample size for each weighting method by adding the effective sample sizes of the corresponding weighted treatment and control groups. Table 4 reports the aforementioned measures separately for the treatment weights and control weights.

TABLE 4 Stability measures of the normalized weights in the treatment and control groups corresponding to different weighting methods. The measures include standard deviation (SD), 99th percentile (99th Perc.) and maximum (Max). We also tabulate the total effective sample size (Total ESS) by adding the ESS of treatment and control groups.

Weighting Method	Treatment Group			Control Group			Total ESS
	SD	99th Perc.	Max.	SD	99th Perc.	Max.	
Simple Logistic	0.0204	0.0393	0.3389	0.0019	0.0091	0.0362	268.3949
Complex Logistic	0.0214	0.0219	0.2552	0.0014	0.0077	0.0244	344.5162
SBW-1	0.0108	0.0555	0.0622	0.0012	0.0048	0.0069	398.8791
SBW-3	0.0091	0.0393	0.0523	0.0013	0.0051	0.0082	400.7916

We observe that both the treated and control weights have smaller dispersion under the balancing approach than the modeling approach, as evidenced by the standard deviation of the weights. Furthermore, the modeling approach exhibits more extreme weights. For instance, in the treatment group, the maximum weight under Simple Logistic is 0.3389, which is more than 5 times the maximum weight under SBW-1 and almost 99 times more than $1/n_t = 0.0034$, the uniform weights. Observations with such extreme weights are highly influential in the sense that they can dominate the contribution of other observations in any weighted estimator of the treatment effect. A useful measure in this regard is the effective sample size (ESS). We see that the total ESS based on the Complex Logistic approach is approximately 268, whereas the total ESS for SBW-3 is approximately 400. Roughly speaking, this means that effectively, SBW-3 is using almost 50 percent more observations than used by Complex Logistic.

To summarize, the studied balancing approach to weighting substantially reduces the strong covariate imbalance that was present in the data. This approach also provides weights that are less dispersed and extreme than those under the modeling approach, leading to a higher effective sample size of the weighted sample.

7 | SUMMARY AND REMARKS

Weighting methods are broadly applicable to causal inference problems and problems of estimation with incomplete outcome data. Weighting is modular, in the sense that it is part of the design stage of the study, and can be used in conjunction with other methods in the outcome analysis stage, such as statistical machine learning methods. This generality of weighting has led to a rapid increase in its usage in practice. Traditionally, the weights are estimated from the data through a modeling approach where one explicitly models the probability of receiving treatment (or control) as a function of the observed covariates and subsequently inverts these probabilities appropriately to get the desired weights. However, there are two practical drawbacks associated with this approach. First, the estimated inverse probability weights obtained through a certain specification of the propensity score model may not produce satisfactory covariate balance in the weighted sample, especially if the propensity model is misspecified. Second, the resulting weights can be extremely large, leading to instability of the treatment effect estimates.

In this paper, we have delineated desirable properties of a weighting scheme, and building on the work by Austin and Stuart (2015),[5] we have presented a set of weight diagnostics directed towards assessing these properties. These properties essentially advocate the usage of weights which are non-negative, are stable, and induce adequate balance relative to the target population of interest. This motivated us to discuss a recent approach to weighting, which we term as the balancing approach (see also Hirshberg and Zubizarreta 2017).[8] The balancing approach proposes a solution to standard problems associated with the modeling approach by building weights which are 'small' by construction and which flexibly balance pre-specified functions of the covariates across treatment and control groups. These functions can be many and span a general function space. Also, the balancing approach can ensure that the weights do not extrapolate.

We have provided a review of the recent progress made in this realm. In particular, we have considered the minimal dispersion approximately balancing class of weights, or minimal weights.[21] We have shown that in the framework of ATT estimation, minimal weights estimate a normalized version of the inverse probability weights under a form of shrinkage estimation of the propensity score. In order to choose the tuning parameter for the optimization problem for the minimal weights, we have proposed a slight variant to the algorithm given in the paper by Wang and Zubizarreta (2020).[21] Finally, using the simulation setup in Hainmueller (2012)[17] and an empirical study on air quality,[61] we have evaluated the comparative performances of these two approaches to weighting. In view of the simulation and empirical study results, we recommend using the balancing approach to weighting as it systematically results in better covariate balance with weights that are minimally dispersed. As a result, treatment effect estimates tend to be more accurate and stable.

There are several open questions about the balancing approach which are intriguing from both theoretical and applied perspectives. First, in an observational study the investigator may not have enough substantive knowledge on which functions of the covariates to balance. As mentioned in Section 3.1, ideally one would like to balance the conditional means of the potential outcomes, i.e., $m_0(X_i)$ and $m_1(X_i)$ (when the estimand of interest is the ATE). Assuming a particular parametric structure on $m_0(\cdot)$ and $m_1(\cdot)$ allows us to target balance on these functions through the balancing approach. Since the true structure of the conditional means is generally unknown, there has been a growing interest in weighting methods that address non-parametric forms of covariate balance. In some recent papers, investigators have tried to balance transformations of the covariates belonging to certain infinite dimensional classes of smooth functionals, e.g., Reproducing Kernel Hilbert Spaces.[32,64] Assuming that the true conditional mean functions belong to that class of functions and some other technical conditions are satisfied, the resulting weighted estimators are shown to have appealing asymptotic properties. There is a broad scope for research on developing weights in the balancing approach which targets balance for a sufficiently general class of functions (see, e.g., Athey et al. 2018,[65] Hirshberg and Wager 2018,[66] Hirshberg et al. 2019,[67] Kallus 2019,[64] and Wang and Zubizarreta 2020[21]).

Second, another important aspect that specifically concerns minimal weights is the choice of the tuning parameter for approximate covariate balance. The tuning parameter plays a crucial role in the optimization problem of the balancing approach since the tuning parameter trades covariate balance for weight dispersion. In this paper, we have described a bootstrap-based algorithm which selects a uniform tuning parameter for all the balancing constraints. A problem with using a common value for the tuning parameter is that the algorithm treats all the constraints equally and hence tries to force equal degrees of balance for all the associated covariate functions simultaneously. Thus, with a very large number of balancing constraints, the chosen tuning parameter might be quite large, resulting in weaker balance in the weighted sample for some important covariates which are a priori known to be prognostically important. More work needs to be done on extending this algorithm to adaptively select the tuning parameters based on the relative importance of each balancing constraint.

References

1. Rubin DB. For objective causal inference, design trumps analysis. *Annals of Applied Statistics* 2008; 2(3): 808–840.
2. Rubin DB. Matching to remove bias in observational studies. *Biometrics* 1973; 29: 159–183.
3. Rosenbaum PR, Rubin DB. Reducing bias in observational studies using subclassification on the propensity score. *Journal of the American Statistical Association* 1984; 79(387): 516–524.
4. Stuart EA. Matching methods for causal inference: a review and a look forward. *Statistical Science* 2010; 25(1): 1–21.
5. Austin PC, Stuart EA. Moving towards best practice when using inverse probability of treatment weighting (IPTW) using the propensity score to estimate causal treatment effects in observational studies. *Statistics in medicine* 2015; 34(28): 3661–3679.
6. Rosenbaum PR, Rubin DB. The central role of the propensity score in observational studies for causal effects. *Biometrika* 1983; 70(1): 41–55.
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. Hirshberg DA, Zubizarreta JR. On Two Approaches to Weighting in Causal Inference. *Epidemiology* 2017; 28(6): 812–816.
9. Imai K, King G, Stuart E. Misunderstandings among experimentalists and observationalists about causal inference. *Journal of the Royal Statistical Society, Series A* 2008; 171(2): 1–22.
10. Zubizarreta JR. Stable weights that balance covariates for estimation with incomplete outcome data. *Journal of the American Statistical Association* 2015; 110(511): 910–922.
11. Kang JDY, Schafer JL. Demystifying double robustness: a comparison of alternative strategies for estimating a population mean from incomplete data (with discussion). *Statistical Science* 2007; 22(4): 523–539.
12. Cole SR, Hernán MA. Constructing inverse probability weights for marginal structural models. *American Journal of Epidemiology* 2008; 168(6): 656–664.
13. Graham BS, Pinto CCDX, Egel D. Inverse probability tilting for moment condition models with missing data. *The Review of Economic Studies* 2012; 79(3): 1053–1079.
14. Imai K, Ratkovic M. Covariate balancing propensity score. *Journal of the Royal Statistical Society: Series B* 2014; 76(1): 243–263.
15. Tan Z. Regularized calibrated estimation of propensity scores with model misspecification and high-dimensional data. *arXiv preprint arXiv:1710.08074* 2017.
16. Fan J, Imai K, Liu H, Ning Y, Yang X. Improving covariate balancing propensity score: A doubly robust and efficient approach. 2016.
17. Hainmueller J. Entropy balancing for causal effects: a multivariate reweighting method to produce balanced samples in observational studies. *Political Analysis* 2012; 20(1): 25-46.

18. Chan KCG, Yam SCP, Zhang Z. Globally efficient nonparametric inference of average treatment effects by empirical balancing calibration weighting. *Journal of the Royal Statistical Society: Series B* 2016; 78(3): 673–700.

19. Santacatterina M, Bottai M. Optimal probability weights for inference with constrained precision. *Journal of the American Statistical Association* 2018; 113(523): 983–991.

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

21. Wang Y, Zubizarreta JR. Minimal dispersion approximately balancing weights: asymptotic properties and practical considerations. *Biometrika* 2020; 107(1): 93–105.

22. Neyman J. On the application of probability theory to agricultural experiments. *Statistical Science* 1923, 1990; 5(5): 463–480.

23. Rubin DB. Estimating causal effects of treatments in randomized and nonrandomized studies.. *Journal of Educational Psychology* 1974; 66(5): 688.

24. Rubin DB. Randomization analysis of experimental data: the Fisher randomization test comment. *Journal of the American Statistical Association* 1980; 75(371): 591–593.

25. Imbens GW. Nonparametric estimation of average treatment effects under exogeneity: A review. *The Review of Economics and Statistics* 2004; 86(1): 4–29.

26. Horvitz DG, Thompson DJ. A generalization of sampling without replacement from a finite universe. *Journal of the American statistical Association* 1952; 47(260): 663–685.

27. Robins JM, Rotnitzky A, Zhao LP. Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association* 1994; 89(427): 846–866.

28. Solon G, Haider SJ, Wooldridge JM. What are we weighting for?. *Journal of Human Resources* 2015; 50(2): 301–316.

29. Rosenbaum PR, Rubin DB. Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. *The American Statistician* 1985; 39(1): 33–38.

30. Chattopadhyay A, Zubizarreta JR. On the implied weights of linear regression in causal inference. *working paper* 2020.

31. Zubizarreta JR. Using mixed integer programming for matching in an observational study of kidney failure after surgery. *Journal of the American Statistical Association* 2012; 107(500): 1360–1371.

32. Wong RK, Chan KCG. Kernel-based covariate functional balancing for observational studies. *Biometrika* 2018; 105(1): 199–213.

33. Petersen ML, Porter KE, Gruber S, Wang Y, Laan v. dMJ. Diagnosing and responding to violations in the positivity assumption. *Statistical methods in medical research* 2012; 21(1): 31–54.

34. Westreich D, Cole SR. Invited commentary: positivity in practice. *American journal of epidemiology* 2010; 171(6): 674–677.

35. Kish L. Survey Sampling New York John Wiloy y Sons. *New York, EUA* 1965; 26.

36. Imbens GW. Matching methods in practice: three examples. *Journal of Human Resources* 2015; 50(2): 373–419.

37. Fogarty C, Mikkelsen M, Gaieski D, Small D. Discrete optimization for interpretable study populations and randomization inference in an observational study of severe sepsis mortality. *Journal of the American Statistical Association* 2015: forthcoming.

38. Crump RK, Hotz VJ, Imbens GW, Mitnik OA. Dealing with limited overlap in estimation of average treatment effects. *Biometrika* 2009; 96(1): 187–199.

39. Li F, Morgan KL, Zaslavsky AM. Balancing covariates via propensity score weighting. *Journal of the American Statistical Association* 2018; 113(521): 390–400.

40. Kern HL, Stuart EA, Hill J, Green DP. Assessing methods for generalizing experimental impact estimates to target populations. *Journal of research on educational effectiveness* 2016; 9(1): 103–127.

41. Nguyen TQ, Ebnesajjad C, Cole SR, Stuart EA, others . Sensitivity analysis for an unobserved moderator in RCT-to-target-population generalization of treatment effects. *The Annals of Applied Statistics* 2017; 11(1): 225–247.

42. Stuart EA, Cole SR, Bradshaw CP, Leaf PJ. The use of propensity scores to assess the generalizability of results from randomized trials. *Journal of the Royal Statistical Society: Series A* 2011; 174(2): 369–386.

43. Hartman E, Grieve R, Ramsahai R, Sekhon JS. From sample average treatment effect to population average treatment effect on the treated: combining experimental with observational studies to estimate population treatment effects. *Journal of the Royal Statistical Society: Series A* 2015; 178(3): 757–778.

44. Andrews I, Oster E. Weighting for external validity. tech. rep., National Bureau of Economic Research; 2017.

45. Dahabreh IJ, Robertson SE, Stuart EA, Hernan MA. Transporting inferences from a randomized trial to a new target population. *arXiv preprint arXiv:1805.00550* 2018.

46. Lee BK, Lessler J, Stuart EA. Improving propensity score weighting using machine learning. *Statistics in medicine* 2010; 29(3): 337–346.

47. Tan Z. Regularized calibrated estimation of propensity scores with model misspecification and high-dimensional data. *Biometrika* 2020; 107(1): 137–158.

48. Zhao Q, Percival D. Entropy balancing is doubly robust. *Journal of Causal Inference* 2017; 5(1).

49. Yiu S, Su L. Covariate association eliminating weights: a unified weighting framework for causal effect estimation. *Biometrika* 2018.

50. Zhao Q. Covariate balancing propensity score by tailored loss functions. *The Annals of Statistics* 2019; 47(2): 965–993.

51. Ning Y, Peng S, Imai K. Robust Estimation of Causal Effects via High-Dimensional Covariate Balancing Propensity Score. *arXiv preprint arXiv:1812.08683* 2018.

52. Visconti G, Zubizarreta JR. Handling limited overlap in observational studies with cardinality matching. *Observational Studies* 2018; 4: 217–249.

53. Hahn J. On the role of the propensity score in efficient semiparametric estimation of average treatment effects. *Econometrica* 1998: 315–331.

54. Ichimura H, Linton O. Asymptotic expansions for some semiparametric program evaluation estimators. *Identification and Inference for Econometric Models* 2005: 149–170.

55. Shortreed SM, Ertefaie A. Outcome-adaptive lasso: Variable selection for causal inference. *Biometrics* 2017; 73(4): 1111–1122.

56. Heller R, Rosenbaum PR, Small DS. Split samples and design sensitivity in observational studies. *Journal of the American Statistical Association* 2009; 104(487): 1090–1101.

57. Normand SLT, Landrum MB, Guadagnoli E, et al. Validating recommendations for coronary angiography following acute myocardial infarction in the elderly: a matched analysis using propensity scores. *Journal of clinical epidemiology* 2001; 54(4): 387–398.

58. Rosenbaum PR. *Design of observational studies*. Springer . 2010.

59. Chipman HA, George EI, McCulloch RE, others . BART: Bayesian additive regression trees. *The Annals of Applied Statistics* 2010; 4(1): 266–298.

60. Hill JL. Bayesian nonparametric modeling for causal inference. *Journal of Computational and Graphical Statistics* 2011(1): 217–240.

61. Zigler CM, Choirat C, Dominici F. Impact of National Ambient Air Quality Standards Nonattainment Designations on Particulate Pollution and Health.. *Epidemiology (Cambridge, Mass.)* 2018; 29(2): 165–174.

62. LaLonde RJ. Evaluating the econometric evaluations of training programs with experimental data. *The American Economic Review* 1986: 604–620.

63. Dehejia R, Wahba S. Causal effects in nonexperimental studies: Reevaluating the evaluation of training programs. *Journal of the American Statistical Association* 1999; 94(443): 1053–1062.

64. Kallus N. Generalized optimal matching methods for causal inference. *Journal of Machine Learning Research* 2019: forthcoming.

65. Athey S, Imbens GW, Wager S. Approximate residual balancing: debiased inference of average treatment effects in high dimensions. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 2018.

66. Hirshberg DA, Wager S. Augmented minimax linear estimation. *arXiv preprint arXiv:1712.00038* 2018.

67. Hirshberg DA, Maleki A, Zubizarreta JR. Minimax linear estimation of the retargeted mean. *arXiv preprint arXiv:1901.10296* 2019.

APPENDIX

Minimal weights and inverse probability weights

Here we state and prove a version of Theorem 1 of Wang and Zubizarreta (2020),[21] in the context of estimation of the Average Treatment Effect for the Treated (ATT) using minimal weights.

Theorem 1. The dual optimization problem of (4.2.1) has the following form

$$\underset{\lambda}{\text{minimize}} \sum_{i=1}^{n} \Big[-(1-Z_i)\rho\{\boldsymbol{B}(X_i)^T\lambda\} + \frac{\boldsymbol{B}(X_i)^T\lambda}{n_t}\Big] + |\lambda|^T\boldsymbol{\delta}$$

Here λ is a $K \times 1$ vector of dual variables corresponding to the K balancing constraints in (4.2.1) and $\rho(t) = \frac{t}{n_t} + t(\psi')^{-1}(-t) + \psi\{(\psi')^{-1}(-t)\}$. Moreover, for primal solution $\boldsymbol{w}^*$ and dual solution λ^*, we have $w_i^* = \rho'\{\boldsymbol{B}(X_i)^T\lambda^*\} - \frac{1}{n_t}$

Proof. We follow the proof technique used by Wang and Zubizarreta.[21] Firstly, we can write kth balancing constraint of (4.2.1) as follows:

$$\big(B_k(X_1), ..., B_k(X_n)\big)\Big(\frac{Z_1}{n_t} - (1-Z_1)w_1, ..., \frac{Z_1}{n_t} - (1-Z_1)w_n\Big)^T \le \delta_k$$

where $k = 1, 2, ..., K$. Let A be a $K \times n$ matrix whose (i,j)th element is $B_i(X_j)$. Let us further denote $Q = \begin{pmatrix} A \\ -A \end{pmatrix}$ and $\boldsymbol{d} = \begin{pmatrix} \boldsymbol{\delta} \\ -\boldsymbol{\delta} \end{pmatrix}$. Denoting $s_i = \frac{Z_i}{n_t} - (1-Z_i)w_i$ and $\boldsymbol{s} = (s_1, s_2, ..., s_n)^T$, we can write the primal problem as

$$\begin{aligned} &\underset{\boldsymbol{w}}{\text{minimize}} \quad \sum_{i=1}^{n}(1-Z_i)\psi(-s_i) \\ &\text{subject to} \quad \boldsymbol{Qs} \le \boldsymbol{d} \end{aligned} \tag{A1}$$

For convenience, let us denote $h(t) = \psi(-t)$. (A1) gives us a convex optimization problem in $\boldsymbol{s}$ with linear constraints. Let $\boldsymbol{Q}_i$ be the ith column of $\boldsymbol{Q}$. The dual of (A1) is as follows,

$$\begin{aligned} &\underset{\lambda}{\text{maximize}} \quad \sum_{i=1}^{n}\{-h_i^*(Q_i^T\lambda)\} - \lambda^T\boldsymbol{d} \\ &\text{subject to} \quad \lambda \ge \boldsymbol{0} \end{aligned} \tag{A2}$$

where

$$\begin{aligned} h_i^*(t) &= \sup_{s_i}\{ts_i - (1-Z_i)h(s_i)\} \\ &= \sup_{w_i}\Big[\Big\{\frac{Z_i}{n_t} - (1-Z_i)w_i\Big\}t - (1-Z_i)h(-w_j)\Big] \\ &= \Big\{\frac{Z_i}{n_t} - (1-Z_i)w_i^*\Big\}t - (1-Z_i)h(-w_j^*) \end{aligned}$$

where w_j^* satisfies $\frac{\partial}{\partial w_i}\Big[\big\{\frac{Z_i}{n_t} - (1-Z_i)w_i\big\}t - (1-Z_i)h(-w_j)\Big] = 0$. Solving for w_i, we get $w_i^* = -(h^{'})^{-1}(t)$.

Substituting w_i^* in the expression of $h_i^*(t)$, we get $h_i^*(t) = -(1-Z_i)\rho(t) + \frac{t}{n_t}$, where $\rho(t) = \frac{t}{n_t} - t(h^{'})^{-1}(t) + h\{(h^{'})^{-1}(t)\} = \frac{t}{n_t} + t(\psi^{'})^{-1}(-t) + \psi\{(\psi^{'})^{-1}(-t)\}$. Taking derivative of $\rho(t)$ gives us $\rho'(t) = \frac{1}{n_t} - (h^{'})^{-1}(t)$. Hence

$$w_i^* = \rho'(t) - \frac{1}{n_t} \tag{A3}$$

The dual objective function in (A2) can be written as $\sum_{i=1}^n \Big\{(1-Z_i)\rho(Q_i^T\lambda) - \frac{Q_i^T\lambda}{n_t}\Big\} - \lambda^T\boldsymbol{d}$ and hence we can alternatively write the dual optimization problem as follows

$$\begin{aligned} &\underset{\lambda}{\text{minimize}} \quad \sum_{i=1}^n \Big\{-(1-Z_i)\rho(Q_i^T\lambda) + \frac{Q_i^T\lambda}{n_t}\Big\} + \lambda^T\boldsymbol{d} \\ &\text{subject to} \quad \lambda \geq \boldsymbol{0} \end{aligned} \tag{A4}$$

We see that the above form of the dual has similar structure to the dual form in Lemma 1 of Wang and Zubizarreta.[21] It is now easy to see (following Proof of Theorem 1 of the same paper) that (A4) is equivalent to

$$\underset{\lambda}{\text{minimize}} \sum_{i=1}^n \Big[-(1-Z_i)\rho\{\boldsymbol{B}(X_i)^T\lambda\} + \frac{\boldsymbol{B}(X_i)^T\lambda}{n_t}\Big] + |\lambda|^T\boldsymbol{\delta}$$

The second part of the theorem follows from (A3).

□

The dual optimization problem can be viewed as a penalized version of an empirical loss function given by

$$\frac{1}{n}\sum_{i=1}^n \Big[-(1-Z_i)n\rho\{\boldsymbol{B}(X_i)^T\lambda\} + \frac{n\boldsymbol{B}(X_i)^T\lambda}{n_t}\Big]$$

where the loss function is regularized by the L_1 penalty. Now, we can write the expected loss (conditional on the covariates) as

$$\mathbb{E}\Big[-(1-Z_i)n\rho\{\boldsymbol{B}(X_i)^T\lambda\} + \frac{n\boldsymbol{B}(X_i)^T\lambda}{n_t}\Big|X_i\Big] = -n\{1-e(X_i)\}\rho\{\boldsymbol{B}(X_i)^T\lambda\} + \frac{n\boldsymbol{B}(X_i)^T\lambda}{n_t}$$

Minimizes this expected loss over $\lambda \geq 0$ gives us the following

$$\begin{aligned} \{1-e(X_i)\}\rho'\{\boldsymbol{B}(X_i)^T\lambda\} &= \frac{1}{n_t} \\ \rho'\{\boldsymbol{B}(X_i)^T\lambda\} - \frac{1}{n_t} &= \frac{1}{n_t}\frac{e(X_i)}{1-e(X_i)} \end{aligned} \tag{A5}$$

Equations (A3) and (A5) establishes the connection between inverse probability weights and minimal weights under the framework of ATT estimation.

□