Health Serv Outcomes Res Method (2017) 17:175–197
DOI 10.1007/s10742-016-0168-2

Optimizing variance-bias trade-off in the TWANG package for estimation of propensity scores

Layla Parast[1] · Daniel F. McCaffrey[3] · Lane F. Burgette[2] · Fernando Hoces de la Guardia[1] · Daniela Golinelli[4] · Jeremy N. V. Miles[1] · Beth Ann Griffin[2]

Received: 4 April 2016 / Revised: 17 November 2016 / Accepted: 9 December 2016 /
Published online: 26 December 2016

Abstract While propensity score weighting has been shown to reduce bias in treatment effect estimation when selection bias is present, it has also been shown that such weighting can perform poorly if the estimated propensity score weights are highly variable. Various approaches have been proposed which can reduce the variability of the weights and the risk of poor performance, particularly those based on machine learning methods. In this study, we closely examine approaches to fine-tune one machine learning technique [generalized boosted models (GBM)] to select propensity scores that seek to optimize the variance-bias trade-off that is inherent in most propensity score analyses. Specifically, we propose and evaluate three approaches for selecting the optimal number of trees for the GBM in the `twang` package in R. Normally, the `twang` package in R iteratively selects the optimal number of trees as that which maximizes balance between the treatment groups being considered. Because the selected number of trees may lead to highly variable propensity score weights, we examine alternative ways to tune the number of trees used in the estimation of propensity score weights such that we sacrifice some balance on the pre-treatment covariates in exchange for less variable weights. We use simulation studies to illustrate these methods and to describe the potential advantages and disadvantages of each method. We apply these methods to two case studies: one examining the effect of dog ownership on the owner's general health using data from a large, population-based survey in California, and a second investigating the relationship between abstinence and a long-term economic outcome among a sample of high-risk youth.

Keywords Causal inference · Propensity score · Machine learning

✉ Layla Parast
parast@rand.org

[1] RAND Corporation, 1776 Main Street, Santa Monica CA 90403, USA

[2] RAND Corporation, 1200 South Hayes Street, Arlington VA 22202, USA

[3] Educational Testing Service, 660 Rosedale Road, Princeton NJ 08541, USA

[4] Mathematica Policy Research, 1100 1st Street, NE, Washington DC 20002, USA

1 Introduction

In studies aimed at evaluating treatments or interventions in health settings, it is often infeasible or impractical to consider random treatment assignment. In such observational (non-experimental) studies, the pre-treatment characteristics of individuals who receive treatment may be very different from those who do not receive treatment. If these pre-treatment characteristics are also associated with the outcome of interest, estimates of the treatment effect may be incorrect due to selection bias. A number of statistical methods have been developed to account and adjust for selection bias and obtain unbiased treatment effect estimates. Several of these methods, including regression adjustment, matching methods and propensity score methods, rely on adjusting for group differences by using the observed pre-treatment covariates available to the researchers. For example, propensity score methods involve the estimation of the propensity score, an individual's probability of assignment to (or selection into) the treatment group, which is then used to balance the treatment and control groups with respect to the observed pre-treatment characteristics. The propensity score can be used to create balance by matching, weighting, or subclassifying on the estimated propensity scores in the two groups which in turn allows for a valid comparison and a more robust estimate of the treatment effect of interest. Here, we focus on methods for fine-tuning the estimated weights which are equal to the inverse of the estimated propensity scores when interest lies in average treatment effects across the population and are commonly referred to as inverse-probability or propensity score weights.

Though propensity score weighting can be used to reduce or eliminate bias when estimating a treatment effect, it often comes at a price whereby the variance of the treatment effect estimates increases due to a reduction in the effective sample size. This is commonly referred to as "the variance-bias trade-off" that is at the core of many statistical methodological problems. When using the inverse of the propensity scores as weights, highly variable weights can lead to a few observations greatly influencing the estimated treatment effect, which can result in low precision of the estimated treatment effects. The primary cause of highly variable weights is poor overlap in pre-treatment characteristic values between the two groups, so that only a few members of either group are representative of the other group and thus, receive large weights. Large weights are more common when there are many predictors as they can create more separation between groups. Outliers values in the covariates and model mis-specification and extrapolation of linear models into regions with sparse data can also lead to large weights. While this variability issue is a concern regardless of the estimation method used to estimate the propensity scores, in this paper we are particularly interested in the context of propensity scores that are estimated using generalized boosted models (GBM) (McCaffrey et al. 2004). Numerous propensity score estimation approaches have been proposed ranging from simple logistic regression modeling to machine learning approaches, including GBM. These machine learning approaches provide alternatives to parametric estimation of propensity scores as a way to minimize bias from incorrect assumptions about the form of the model used (McCaffrey et al. 2004; van der Laan 2014; Imbens 2000; Robins et al. 2000). These methods eliminate reliance on a simple parametric logistic regression model and do not require the researcher to determine which pre-treatment covariates and their respective interactions should be included in the model. It has been shown that the resulting weights from these approaches yield more precise treatment effect estimates and

lower mean squared error than traditional logistic regression methods (Harder et al. 2010; Lee et al. 2010; Pirracchio et al. 2015).

In this study, we examine ways to fine-tune GBM to select propensity scores that seek to optimize the variance-bias trade-off that is inherent in most propensity score analyses. We build from previous work in Golinelli et al. (2012) who investigated whether it is best to optimize the balance or to settle for a less than optimal balance in hopes of reducing variability. Those authors found that "every step toward better balance usually means an increase in variance and at some point a marginal decrease in bias may not be worth the associated increase in variance". Currently, the R package `twang` (Ridgeway 2016), which uses GBM for propensity score estimation, selects the optimal number of trees (described below in Sect. 2.2) as that which maximizes balance (as measured by a particular balance metric) between the treatment groups being considered. However, the selected number of trees may lead to highly variable propensity score weights, which would lead to a large design effect, and thus lower power to detect a true treatment effect. In this paper, we propose and examine alternative ways to tune the number of trees for GBM for estimation of propensity score weights such that we sacrifice some balance on the pre-treatment covariates in exchange for less variable weights. Using simulation studies, we illustrate these methods and describe the potential advantages and disadvantages of each method. We apply these procedures to two case studies: one examining the effect of dog ownership on the owner's general health using data from a large, population-based survey in California, and a second investigating the relationship between abstinence and a long-term economic outcome among a sample of high-risk youth.

2 Propensity scores: use and estimation

2.1 Using propensity scores to obtain a treatment effect estimate

Let Y_i denote the outcome of interest for individual i, T_i denote the treatment or intervention, where $T_i = 0$ or 1, and $\mathbf{X}_i$ denote the vector of available baseline/pre-treatment covariates. Each individual has two potential outcomes: the Y_i that would be observed if the individual was assigned to treatment group 1 i.e. $T_i = 1$ and the Y_i that would be observed if the individual was assigned to treatment group 0 i.e. $T_i = 0$. However, only one of these potential outcomes is observable for each individual. To rigorously define our estimate of interest we define Y_{1i} and Y_{0i} to denote the potential outcomes when $T_i = 1$ and $T_i = 0$, respectively. Using this notation, a common treatment effect of interest might be the average treatment effect on the population (ATE):

$$ATE = E(Y_{1i} - Y_{0i}) = E(Y_{1i}) - E(Y_{0i}) \equiv \tau. \quad (1)$$

If Y_{1i} and Y_{0i} were observed for every individual, then $E(Y_{1i})$ and $E(Y_{0i})$ could simply be estimated using $n^{-1}\sum_{i=1}^{n} Y_{1i}$ and $n^{-1}\sum_{i=1}^{n} Y_{0i}$, respectively, where n is the number of individuals. However, these are never both observed for the same individual. If one were to instead use $n_j^{-1}\sum_{i=1}^{n} Y_{ji} I(T_i = j)$ for $j = 0, 1$, where n_j is the number of individuals in group j, to estimate these quantities, the obtained treatment effect estimate will be biased unless treatment assignment and the potential outcomes are independent i.e. $T_i \perp Y_{1i}, Y_{0i}$. In an observational study, it is generally not appropriate to assume such independence. Often, there are individual characteristics that may be associated with both treatment assignment and the potential outcomes. For example, in a study examining the effect of

dog ownership on general health, individual factors such as gender, age, marital status, and socio-economic factors are likely associated with both the likelihood of owning a dog and general health. While such factors cannot be ignored, if it is possible to identify this set of factors, denoted by $\mathbf{Z}_i$, a subset of $\mathbf{X}_i$, then it may be reasonable to make the assumption:

$$T_i \perp Y_{1i}, Y_{0i} \mid \mathbf{Z}_i. \tag{2}$$

Under this assumption, methods that appropriately account for the differential distribution of $\mathbf{Z}_i$ within each treatment group will lead to valid estimation of the treatment effect estimate (Rosenbaum and Rubin 1983b). This assumption is often referred to as the assumption of no unmeasured confounders or the assumption of strong ignorability (Robins et al. 2000).

The use of propensity scores is one method that allows for such valid estimation. The propensity score is the probability of being in treatment group 1 given individual characteristics, $\mathbf{Z}_i$: $p_i = P(T_i = 1|\mathbf{Z}_i)$. Previous work in this area has shown that when Assumption (2) holds and p_i is known or can be consistently estimated, then

$$T_i \perp Y_{1i}, Y_{0i} \mid p_i$$

(Rosenbaum and Rubin 1984, 1983b; Hernán et al. 2000). A valid ATE estimate can then be obtained by weighting using the propensity score:

$$\widehat{ATE} = \frac{\sum_{i:T_i=1} Y_{1i} W_i}{\sum_{i:T_i=1} W_i} - \frac{\sum_{i:T_i=0} Y_{0i} W_i}{\sum_{i:T_i=0} W_i} \tag{3}$$

where

$$W_i = \begin{cases} 1/p_i & \text{if} \quad T_i = 1 \\ 1/(1-p_i) & \text{if} \quad T_i = 0. \end{cases} \tag{4}$$

Similar logic can be used to develop a valid estimator for the average treatment effect on the treated population (ATT), another commonly used estimate of the effect of the treatment. As mentioned previously, although propensity score weights can reduce the potential for bias by balancing the covariate distribution between the treatment and control groups, that reduction comes at the cost of unequal weighting of observations. In general, a weighted sample mean has greater variance than an unweighted sample mean of the same size sample. If the weights are independent of the outcomes, the ratio of the variances of weighted and unweighted means for group $j = 0, 1$, for control and treatment, equals $D^j = \frac{n_j \sum_{i=1}^{n_j} W_i^2}{(\sum_{i=1}^{n_j} W_i)^2}$, which is typically referred to as the design effect (DEFF). Even though propensity score weights are unlikely to be independent of the outcomes, the DEFF is commonly used to assess the variability in the weights and potential impacts of weighting on the precision of the treatment effect estimate. In addition to group specific design effects, we also consider the weighted average DEFF,

$$D = D^1(1-q) + D^0 q = \frac{n_1 \sum_{i=1}^{n_1} W_i^2}{(\sum_{i=1}^{n_1} W_i)^2}(1-q) + \frac{n_0 \sum_{i=0}^{n_0} W_i^2}{(\sum_{i=1}^{n_0} W_i)^2} q$$

where $q = n_1/n$, and we refer to $(\sum_{i=1}^{n_j} W_i)^2/(\sum_{i=1}^{n_j} W_i^2)$ as the effective sample size in treatment group j. In our numerical examples, we will examine both the DEFF and the variance of the treatment effect estimate.

2.2 Propensity score estimation

Because the propensity scores are unknown they must be estimated. There has been an immense amount of work focused on developing methods to estimate propensity scores (e.g., McCaffrey et al. 2004; van der Laan 2014; Breiman et al. 1984; Hill 2011; Imai and Ratkovic 2014; Liaw and Wiener 2002). In practice, propensity scores are most commonly estimated using parametric methods such as logistic regression. Standard implementation using logistic regression involves beginning with a model that only includes main effects for the observed pre-treatment characteristics and adding squared terms and interactions of covariates to the model to improve the propensity estimates when sufficient balance is not obtained. The resulting fits have been criticized for yielding highly variable weights and unstable weighted means (Kang and Schafer 2007). Moreover, the model fitting process can be time consuming especially with many covariates and highly disparate groups. Consequently, authors have proposed generalized boosted models (GBM) and other machine learning methods as promising alternatives to logistic regression for propensity score estimation.

GBM is a nonparametric approach to model outcomes (binary, discrete, or continuous) that allows for interactions among covariates and flexible functional forms for the regression surface. It is also invariant to monotonic transformations of covariates. GBM approximates the regression surface through a piecewise constant model, in which the regression surface is constant over regions of the covariate space. The fitting algorithm involves partitioning the covariate space and assigning values to constant functions in the selected regions. Model building is automated through an iterative algorithm that adds terms to maximize the likelihood conditional on the model chosen through the previous iterations. Heuristically, GBM models an outcome as a sum of simple regression tree fits and each iteration of the fitting algorithm adds an additional tree fit to the residuals of the model from the previous iteration. Given the sum-of-trees formulation, the number of iterations used in the fitting algorithm is commonly referred to as the "number of trees" in the model. The GBM algorithm improves the fit to the data with each additional tree, requiring an external criterion to select the number of trees that is optimal in a given situation and controls between overfitting to the data and underspecification of the model. In prediction applications, an external measure of fit estimated through cross-validation or a holdout sample is used to select the number of trees. However, for propensity score estimation, the balance of the covariates across treatment and control groups, that is, the similarity in the weighted distributions of covariates from the two groups, is used to select the optimal number of trees in the GBM. Intuitively, since weights derived from the true propensity scores would achieve balance, a well-fitting estimate of the propensity score should also achieve balance, given sufficient propensity score overlap between groups. For more details on GBM for propensity score estimation, see McCaffrey et al. (2004) and Burgette et al. (2015).

In this paper, we focus specifically on propensity score estimation using GBM. The R package `twang`, an acronym of Toolkit for Weighting and Analysis of Nonequivalent Groups, includes functions to estimate propensity scores using GBM. This package relies on another R package, `gbm` (Ridgeway 2015), to fit the GBM and then uses user-selected criteria to select an optimal number of trees.

3 Methods to select the optimal number of trees in `twang`

3.1 Current method

The `twang` package uses covariate balance to select the number of trees for the GBM models of propensity scores. Specifically, as implemented for propensity score estimation, the GBM function selects the number of trees which minimizes the differences between the two treatment groups as measured by one of four measures: the mean of the absolute standardized bias (ASB), the maximum of the ASB, the mean of the Kolmogorov–Smirnov (KS) statistic, and the maximum of the KS statistic. By definition, the ASB for a particular covariate equals the absolute value of the covariate weighted mean in treatment group 1 minus the covariate weighted mean in treatment group 0 divided by the pooled sample standard deviation. This measure is calculated separately for each covariate and the overall balance measure used for tuning the GBM equals either the mean or the maximum of the covariate-specific ASB values. For each covariate, the KS statistic equals the maximum absolute difference in the propensity score weighted empirical cumulative distribution functions of the treatment and control groups. Again, the statistic is calculated separately for each covariate and aggregated across covariates by either the mean or maximum to create the balance measure used in selecting the number of trees for the GBM model. In this paper, we choose the number of trees which minimizes the maximum of the absolute standardized bias as the one to provide information for the current `twang` method in our simulations and case studies.

3.2 *p* value based method

One straightforward and relatively simplistic alternative to using one of the current `twang` criteria to select the number of trees is to select the smallest number of trees such that there are no statistically significant imbalances in the pre-treatment covariates after weighting, provided that number is less than the number of trees selected by `twang`. We denote this method by letting M equal the number of covariates, $m = 1, \ldots, M$ and t^* equal the number of trees selected by the current `twang` method, and define $\mathcal{L}$ as a grid of L equally spaced points between 1 and t^*. We define $t^p(c_t)$ as:

$$t^p(c_t) = \min\{l \in \mathcal{L} : p_{lm} > c_t \quad \forall m\}$$

where p_{lm} denotes the p value for testing the significance in the weighted mean difference for covariate m using weights obtained when the number of trees is equal to l. That is, $t^p(c_t)$ is the smallest l such that the p values assessing balance associated with each covariate are greater than c_t. In our numerical examples, we consider two values for c_t, 0.05 and 0.10 given these are commonly used thresholds for determining statistically significant and moderately statistically significant findings. For a given sample, smaller imbalances in covariates across groups will lead to $p < 0.10$ more often than $p < 0.05$, so that larger values for c_t put a greater price on imbalance and will tend to lead to GBM fits using more trees.

3.3 Absolute standardized bias based method

Another simple alternative for fine-tuning GBM is to select the smallest number of trees such that the ASB for all covariates after weighting is below a prespecified threshold. Therefore, we define $t^a(c_{\text{ASB}})$ as:

$$t^a(c_{\text{ASB}}) = \min\{l \in \mathcal{L} : \text{ASB}_{lm} < c_{\text{ASB}} \quad \forall m\}$$

where ASB_{lm} denotes the ASB for covariate m using weights obtained when the number of trees is l. That is, $t^a(c_{\text{ASB}})$ is the smallest l such that the ASB associated with every covariate is less than c_{ASB}. In our numerical examples, we consider two values for c_{ASB}, 0.10 and 0.20 which are commonly used thresholds for ASB in the literature.The 0.20 threshold is more liberal than the 0.10 threshold advocated by some authors as indicative of good balance (Austin 2007, 2009; Austin and Stuart 2015; Normand et al. 2001; Hankey and Myers 1971). Smaller values of c_{ASB} put a higher cost on imbalance and will tend to yield GBM fits with more trees.

We note that we have chosen both the ASB and p value methods for consideration in our evaluation because of their simplicity and intuitiveness for most applied practitioners and policy makers. These are simple criteria that might naturally be considered by any analyst or researcher who is trying to fine-tune `twang` when selecting the optimal iteration to yield more stable weights at the cost of slight reductions in balance. Our goal is to study the relative performance of these simpler methods versus a more complex method motivated by mean-squared error (MSE) considerations, described in the next section, in order to illustrate both their advantages and their disadvantages.

3.4 Mean squared error based method

The final criterion we consider for selecting the optimal number of trees in `twang` directly considers both the possible bias and variance from a model with a given number of trees by tuning the GBM to minimize an approximation to the mean squared error (MSE) of the estimated treatment effect. We describe our approach first assuming a single covariate X_i is available.

Consider the linear model for the potential control outcomes:

$$Y_{0i} = \beta_0 + X_i\alpha + e_i \tag{5}$$

where $E(e_i) = 0$, $e_i \perp X_i$ and e_i are independent and identically distributed with variance σ_e^2. In addition, let the treatment effect be constant τ, so that $Y_{1i} = Y_{0i} + \tau$, and the observed data equal $Y_i = \beta_0 + X_i\alpha + T_i\tau + e_i$. Let

$$\tilde{Y}_j = \frac{\sum_{i:T_i=j} Y_{1i}W_i}{\sum_{i:T_i=j} W_i},$$

for $j = 0, 1$, for the weights defined by (4) and $\tilde{X}_1$, $\tilde{X}_0$, $\tilde{e}_1$ and $\tilde{e}_0$ equal the corresponding values for X and e. Let D_t, D_t^0, and D_t^1 be the design effect defined earlier overall (see Sect. 2.1), in treatment group 0, and in treatment group 1, respectively, using weights obtained when the number of trees is t and note that

$$E[D_t^1/n_1] + E[D_t^0/n_0] = \frac{E[D_t]}{nq(1-q)}.$$

The estimated additive treatment effect estimate, $\widehat{ATE} = \tilde{Y}_1 - \tilde{Y}_0$ and the MSE for this estimator is

$$\begin{aligned} E[(\tilde{Y}_1 - \tilde{Y}_0 - \tau)^2] &= E[(\alpha\tilde{X}_1 + \tau + \tilde{e}_1 - (\alpha\tilde{X}_0 + \tilde{e}_0) - \tau)^2] \\ &= \alpha^2\sigma_X^2 \frac{E[(\tilde{X}_1 - \tilde{X}_0)^2]}{\sigma_X^2} + E[(\tilde{e}_1 - \tilde{e}_0)^2]. \end{aligned}$$

Because e_i is independent of $X_{i'}$ for all i and i' and $E[e] = 0$, $E[(\tilde{e}_1 - \tilde{e}_0)]^2 = \sigma_e^2(E[D_t^1/n_1] + E[D_t^0/n_0])$. For this simple linear model, $R^2 = \alpha^2\sigma_X^2/(\alpha^2\sigma_X^2 + \sigma_e^2)$ since $\text{var}\ (Y_0) = \alpha^2\sigma_X^2 + \sigma_e^2$ is the large sample coefficient of variation for a regression of Y_0 on X. For a given number of trees, we let $\Delta_t = E[(\tilde{X}_{t1} - \tilde{X}_{t0})^2]/\sigma_X^2$, then the MSE for a given number of trees can be written as:

$$MSE_t = \ \text{var}\ (Y_0)\left\{R^2\Delta_t + (1-R^2)\frac{E[D_t]}{nq(1-q)}\right\}. \tag{6}$$

Given any two tree selections, t_1 and t_2, one would select t_2 over t_1 if $MSE_{t_1} > MSE_{t_2}$ or

$$\text{var}\ (Y_0)\left\{R^2\Delta_{t_1} + (1-R^2)\frac{E[D_{t_1}]}{nq(1-q)}\right\} > \ \text{var}\ (Y_0)\left\{R^2\Delta_{t_2} + (1-R^2)\frac{E[D_{t_2}]}{nq(1-q)}\right\}$$

Rearranging yields that t_2 should be selected if

$$1 < \left[\frac{(1-R^2)}{R^2}\right]\frac{\frac{1}{nq(1-q)}(E[D_{t_1}] - E[D_{t_2}])}{\Delta_{t_2} - \Delta_{t_1}}.$$

To use this formula to select the number of trees, estimates of R^2, Δ_t, and $E[D_t]$ for $t = t_1$ and t_2 are needed. To obtain an estimate of $E[D_t]$, one can use the observed design effects for the weights from each group for each model. To obtain estimates of Δ_t for $t = t_1, t_2$, one can use the ASB for covariate X for each model, which we denote, $\text{ASB}_t(X)$. For R^2, it is not possible to directly estimate R^2 from a regression of Y_0 on X because Y_0 is not observed for the entire sample. However, one can estimate α and σ_e^2 from a regression of Y on X using either the entire sample or the control group only, and σ_X^2 as the pooled sample variance of X, and then estimate R^2 from these two values: $\widehat{R}^2 = \sigma_x^2/(\sigma_x^2 + \sigma_e^2)$. The number of trees can then be selected as $t^m(Y, X)$ where:

$$t^m(Y,X) = \min\left\{l \in \mathcal{L} : 1 < \left[\frac{(1-\widehat{R}^2)}{\widehat{R}^2}\right]\frac{\frac{1}{nq(1-q)}(D_{t^*} - D_l))}{(\text{ASB}_l(X))^2 - \text{ASB}_{t^*}(X)^2)}\right\}. \tag{7}$$

We have explicitly included Y in our notation of $t^m(Y, X)$ to emphasize the use of outcome information in this approach. Alternatively, if the outcomes are unavailable or if the analyst desires to keep the design phase in which weights are estimated separate from the outcome analysis phase to avoid any data snooping, then plausible values for R^2 could be chosen and used in this approach. For example, let R_c^2 denote a plausible R^2 value chosen based on substantive knowledge. Then, the number of trees can be selected using this same

approach, but without using outcome Y information, as $t^m(X)$ defined parallel to (7) but with $\widehat{R}^2$ replaced by R_c^2.

This approach can be extended to multiple covariates. In this case, $Y_0 = \beta_0 + \boldsymbol{\beta}'X + e$,

$$R^2 = \sigma^2_{E[Y|X]}/(\sigma^2_{E[Y|X]} + \sigma^2_e), \text{and}$$

$$\Delta_t = E\left[\left(\sum_{i:T_i=1} W_i\boldsymbol{\beta}'X_i/\sum_{i:T_i=1} W_i - \sum_{i:T_i=0} W_i\boldsymbol{\beta}'X_i/\sum_{i:T_i=0} W_i\right)^2\right]/\sigma^2_{E[Y|X]},$$

where $\sigma^2_{E[Y|X]} = \boldsymbol{\beta}'\Sigma_X\boldsymbol{\beta}$. We can estimate $\boldsymbol{\beta}$ as $\widehat{\boldsymbol{\beta}}$ and σ^2_e from a regression of Y on the covariates and a treatment indicator or from a regression of Y on the covariates in the control group. Let $\widehat{\Sigma}_X$ equal the pooled sample estimate of the variance-covariance matrix of the covariates and $\widehat{\sigma}^2_{E[Y|X]} = \widehat{\boldsymbol{\beta}}'\widehat{\Sigma}_X\widehat{\boldsymbol{\beta}}$, we can estimate R^2 by

$$\widehat{R}^2 = \widehat{\sigma}^2_{E[Y|X]}/(\widehat{\sigma}^2_{E[Y|X]} + \widehat{\sigma}^2_e) \tag{8}$$

and $\text{ASB}_t(Y \mid \mathbf{X}) = (\sum_{i:T_i=1} W_i\widehat{\boldsymbol{\beta}}'X_i/\sum_{i:T_i=1} W_i - \sum_{i:T_i=0} W_i\widehat{\boldsymbol{\beta}}'X_i/\sum_{i:T_i=0} W_i)/\widehat{\sigma}_{E[Y|X]}$. The number of trees can then be selected as $t^m(Y,X)$ where:

$$t^m(Y,X) = \min\left\{l \in \mathcal{L} : 1 < \left[\frac{(1-\widehat{R}^2)}{\widehat{R}^2}\right]\frac{\frac{1}{nq(1-q)}(D_{t^*} - D_l))}{(\text{ASB}_l(Y|\mathbf{X}))^2 - \text{ASB}_{t^*}(Y|\mathbf{X})^2)}\right\}.$$

Alternatively, as in the single covariate setting, if analysts do not wish to use the outcomes in the estimation of the propensity scores or the outcomes are unavailable, the analyst can pose plausible values for R^2. Also, since $\boldsymbol{\beta}$ cannot be estimated without an outcome, analysts cannot estimate $\text{ASB}_t(Y \mid \mathbf{X})$. As an alternative one could use $\text{ASB}(X^*)$ and $\text{ASB}(X^l)$ where X^* and X^l are the baseline covariates with the most imbalance when the number of trees is t^* or l respectively, in the inequality for selecting t^*. Similar to the single covariate setting, we denote the resulting number of trees selected using this approach as $t^m(X)$. In our numerical examples, we examine the performance of both $t^m(Y,X)$ and $t^m(X)$.

As noted above, if the regression coefficients and residual variance are estimated using the data, then outcome information, Y, is used to select the number of trees, and will contribute to the estimation of the propensity score weights. Intuitively, this should help improve the accuracy of the results as additional information contributes to the estimation process. However, it creates the potential for the results of the treatment effect estimation to influence the weight selection. This can be avoided by using plausible values for R^2 rather than estimating it or by fitting the outcome regression model only to the control cases, and using bounds for ASB for the unknown Δ_t values. In addition, the linear model (5) with equal variances of e in the two groups is used as an approximation to construct the inequality above. In general, this assumption might not be true. In fact, if (5) was the true model, then propensity score weighting would not be necessary, as one could instead use simple regression adjustment to obtain an unbiased estimate of the treatment effect. The linear model is used to motivate the inequality, which will hopefully be sufficiently accurate to provide a way of tuning the propensity score model to yield estimated treatment effects with lower MSE than simply picking the model that yields the best covariate balance. It is important to note that all of the alternative methods described here focus on

potentially reducing the number of trees selected by `twang`; we do not focus on increasing the number of trees.

4 Simulation study

We examined the performance of the proposed alternatives and compared each to the existing approach used in `twang` in two simulation settings with $n = 300$ and $n = 2000$. In setting (i), data were generated as:

$$\begin{aligned} X &\sim N(0, .64) \\ Z &\sim \text{Exp}(1.5) \\ p &= 1/[1 + \exp\{-1.1(X + Z - 2/3)\}] \\ T &= 1(U < p), \quad U \sim Unif(0, 1) \\ Y &= 0.3X + 0.3Z + \tau T + N(0, \sigma_1^2) \end{aligned}$$

where X and Z denote the pre-treatment baseline covariates, Y denotes the outcome, $\tau = 0$ and $\sigma_1 = 2.5$. In setting (ii), data were generated similarly with the exception that $\sigma_1 = 0.7$. The treatment effect of interest in our estimation procedures is τ. In setting (i) the R^2 was 0.015 and in setting (ii) it was 0.166. Recall that an estimate of the R^2 within each simulation replication is what is used in the MSE-based approach to select $t^m(Y, X)$; this value is set to 0.20 when implementing the $t^m(X)$ method. We specifically chose these two simulation settings and sample sizes to illustrate the differences in these methods, describe the advantages and disadvantages of each method, and demonstrate expected results under certain conditions. For example, when R^2 is very small and the ASB for all covariates after using `twang` is low, and the sample size is relatively small, we would expect the MSE-based method to substantially reduce the number of trees selected as optimal. When R^2 is moderate or large, we would *not* expect the MSE-based method to reduce trees compared to the current `twang` methodology. When the sample size is large, we would generally *not* expect the p value based method or the MSE-based method to substantially reduce trees (unless R^2 is very small). When the sample size is small, we would generally expect the p value based method and the MSE-based method to reduce trees (unless R^2 is moderate or large). Since the ASB-based method depends only on the ASB, our expectations depend on whether `twang` is initially able to achieve balance within the pre-determined threshold. Within a particular setting, we would expect the ASB method to reduce trees more dramatically as the sample size increases.

Results are summarized in Tables 1 ($n = 300$) and 2 ($n = 2000$) across 1000 replications of each setting. Recall that $t^m(Y, X)$ indicates the MSE-based method that uses outcome information, $t^m(X)$ indicates the MSE-based method that does not use outcome information, $t^a(0.20)$ and $t^a(0.10)$ indicate the ASB-based methods that use a 0.20 and 0.10 threshold, respectively, $t^p(0.05)$ and $t^p(0.10)$ indicate the p value based methods that use a 0.05 and 0.10 threshold, respectively. These tables show the average number of trees selected, the average maximum ASB, the standardized bias of the treatment effect estimate (standardized using the standard deviation of the of Y in the control group), the standardized variance of the treatment effect estimate, the standardized MSE of the treatment effect estimate, and the average design effect induced by the estimated propensity score weights (design effect is 1 in the unweighted approach).

Table 1 Simulation study results with $n = 2000$; maximum ASB, standardized bias, standardized var, and standardized MSE all multiplied by 100, $t^m(Y,X)$ indicates the MSE-based method that uses outcome information, $t^m(X)$ indicates the MSE-based method that does not use outcome information, $t^a(0.20)$ and $t^a(0.10)$ indicate the ASB-based methods that use a 0.20 and 0.10 threshold, respectively, $t^p(0.05)$ and $t^p(0.10)$ indicate the p value based methods that use a 0.05 and 0.10 threshold, respectively

Method	n=2000			
	Setting (i) # of Trees	Setting (ii)	Setting (i) Maximum ASB	Setting (ii)
Unweighted	–	–	70.45	70.45
TWANG	926	924	12.57	12.68
$t^m(Y,X)$	695	892	13.15	12.66
$t^m(X)$	868	864	12.56	12.68
$t^a(0.20)$	317	319	18.89	18.90
$t^a(0.10)$	907	913	12.81	12.82
$t^p(0.05)$	913	916	12.59	12.69
$t^p(0.10)$	926	924	12.57	12.68
	Standardized bias		Standardized var	
Unweighted	11.02	36.79	0.22	0.22
TWANG	2.03	6.98	0.28	0.23
$t^m(Y,X)$	2.14	6.98	0.28	0.23
$t^m(X)$	2.04	6.99	0.28	0.23
$t^a(0.20)$	3.15	10.45	0.26	0.21
$t^a(0.10)$	2.07	7.05	0.28	0.23
$t^p(0.05)$	2.03	6.98	0.28	0.23
$t^p(0.10)$	2.03	6.98	0.28	0.23
	Standardized MSE		Design effect	
Unweighted	1.43	13.75	1	1
TWANG	0.32	0.72	1.26	1.26
$t^m(Y,X)$	0.32	0.72	1.25	1.26
$t^m(X)$	0.32	0.72	1.26	1.26
$t^a(0.20)$	0.36	1.30	1.18	1.18
$t^a(0.10)$	0.32	0.73	1.26	1.26
$t^p(0.05)$	0.32	0.72	1.26	1.26
$t^p(0.10)$	0.32	0.72	1.26	1.26

We first describe the results when $n = 300$ in Table 2. In setting (i), where R^2 is very small, the MSE-based approach, $t^m(Y,X)$, greatly reduced the number of trees by 65%, from 502 to 176, as expected. This leads to a design effect reduction from 1.22 to 1.10, though the observed difference in variance of the treatment effect estimate is negligible. Also as expected, the bias of the treatment effect estimate increases slightly using this approach. Unfortunately, the MSE increases more than we would have expected as a result of this increase in bias. In this setting, the other alternative methods reduce the number of trees somewhat, but not substantially except for the p value based method with the 0.05

threshold which reduces the trees by 25% from 502 to 378. This is not surprising given the large ASB still present even after `twang` is used with 502 trees. In setting (ii), where R^2 is moderate, as expected, both MSE-based approaches do not dramatically reduce the number of trees. The p value based approach with a 0.05 threshold reduces the number of trees the most, but at a larger cost to bias.

We now describe the results when $n = 2000$ in Table 1. Regardless of setting, the MSE-based approaches do not dramatically reduce the number of trees, although as expected the reduction is greater in setting (i) with the small R^2. Similarly, both t^p methods do not reduce trees because with the large samples even small differences are significant, which is one of the concerns with a tuning method that uses p values to evaluate bias. The only method that results in a large reduction in trees is $t^a(0.20)$ because an ABS of 0.20 is much greater than the ABS achieved with `twang`. However, because the variance of the estimated treatment effects is small, bias dominates and the MSE for $t^a(0.20)$ is larger than for any of the other methods. As might be expected with a large sample, even small increases in bias due to imbalance dominate any reduction in variance associated with a reduction in DEFF. Consequently, there is no benefit to reducing trees by a large amount and the one method that does greatly reduce trees, results in greater MSE than other methods.

5 Case studies

5.1 The effect of dog ownership on general health

We applied these proposed methods to investigate the effect of dog ownership on general health. Several studies have found that owning and/or interacting with a pet (mostly a dog) has benefits for the individual including mental health outcomes such as decreased anxiety and physical health outcomes such as improved immune response (Wells 2009a, b; McConnell et al. 2011). Our analysis used survey response data from the 2003 California Health Interview Survey (CHIS 2003), a population-based, random-digit dial telephone survey of California households. CHIS is the largest state-level health survey and is designed to provide population-based estimates for the state of California, California counties, and major ethnic groups. CHIS collected extensive information on health status, health conditions, health-related behaviors, health insurance coverage and access to health care services as well as demographic and socioeconomic information. Within each household, an interview was conducted with a randomly selected adult (age 18 and over). CHIS 2003 was conducted between August 2003 and February 2004. Interviews were conducted in English, Spanish, Chinese, Vietnamese, and Korean. The demographic characteristics of the CHIS sample (such as race, ethnicity, and income) are very similar to those obtained from Census data, and additional research suggests that CHIS data are representative of the California population (Lee et al. 2009; CHIS 2003). Detailed information about the CHIS methodology is available elsewhere (Survey 2005; Ponce et al. 2004).

Our sample for analysis consisted of the 8526 adults who had a child in the home; 27.0% of these respondents owned a dog. Available individual characteristics included age, gender, race/ethnicity, household size, marriage status, whether the individual received TANF (Temporary Assistance for Needy Families), household annual income, whether the individual worked full time, whether the individual had a spouse that worked full time,

Table 2 Simulation study results with $n = 300$; maximum ASB, standardized bias, standardized var, and standardized MSE all multiplied by 100, $t^m(Y,X)$ indicates the MSE-based method that uses outcome information, $t^m(X)$ indicates the MSE-based method that does not use outcome information, $t^a(0.20)$ and $t^a(0.10)$ indicate the ASB-based methods that use a 0.20 and 0.10 threshold, respectively, $t^p(0.05)$ and $t^p(0.10)$ indicate the p value based methods that use a 0.05 and 0.10 threshold, respectively

Method	n = 300			
	Setting (i) # of Trees	Setting (ii)	Setting (i) Maximum ASB	Setting (ii)
Unweighted	–	–	70.02	70.01
TWANG	502	494	23.49	23.38
$t^m(Y,X)$	176	446	40.55	23.44
$t^m(X)$	452	445	23.62	23.52
$t^a(0.20)$	461	453	23.97	23.86
$t^a(0.10)$	501	493	23.50	23.39
$t^p(0.05)$	378	369	25.38	25.35
$t^p(0.10)$	445	436	24.17	24.08
	Standardized bias		Standardized var	
Unweighted	11.56	36.49	1.36	1.46
TWANG	4.27	12.36	1.61	1.62
$t^m(Y,X)$	5.78	12.39	1.60	1.60
$t^m(X)$	4.31	12.46	1.60	1.60
$t^a(0.20)$	4.37	12.62	1.60	1.60
$t^a(0.10)$	4.27	12.36	1.61	1.62
$t^p(0.05)$	4.62	13.42	1.56	1.56
$t^p(0.10)$	4.40	12.74	1.59	1.59
	Standardized MSE		Design effect	
Unweighted	2.69	14.77	1	1
TWANG	1.79	3.14	1.22	1.21
$t^m(Y,X)$	1.93	3.13	1.10	1.21
$t^m(X)$	1.78	3.15	1.21	1.21
$t^a(0.20)$	1.79	3.19	1.21	1.21
$t^a(0.10)$	1.79	3.14	1.22	1.21
$t^p(0.05)$	1.78	3.36	1.19	1.18
$t^p(0.10)$	1.79	3.21	1.20	1.20

whether the individual lived in a house, and a rural/urban measure (1 = urban; 2 = 2nd city; 3 = suburban; 4 = town and rural) for the individual's address. General health status of the individual was measured as the self-reported response to the question "Would you say that in general your health is excellent, very good, good, fair or poor?" Responses were coded from 1 to 5 with 5 indicating "Excellent." Dog ownership was assessed with the question "Do you have any dogs that you allow inside your home?"

Our goal was to examine the effect of dog ownership on general health. However, since individuals who own a dog are different from those who do not own in a dog in ways that

Table 3 Distribution of baseline covariates among those who do own a dog versus do not own a dog

	Own a dog Mean (SD) or % n = 2306	Do not own a dog Mean (SD) or % n = 6220	ASB	*p* value
Age	38.9 (8.05)	36.51 (8.34)	0.29	<0.001
Male	37.3%	37.8%	0.01	0.660
Race				<0.001
Latino	12.7%	36.6%	0.72	<0.001
Pacific Islander	0.3%	0.3%	0.01	–
American Indian or Alaska native	1.5%	1.4%	0.01	–
Asian	3.2%	11.8%	0.49	–
African American	3.2%	6.7%	0.20	–
White	76.6%	40.8%	0.85	–
Other	2.5%	2.4%	0.01	–
Household size	4.21 (1.26)	4.28 (1.38)	0.05	0.025
Married	78.9%	72.3%	0.15	<0.001
On TANF	2.9%	5.6%	0.13	<0.001
Household annual income	11.03 (1.03)	10.52 (1.39)	0.38	<0.001
Works full time	64%	61.6%	0.05	0.042
Spouse works full time	57.1%	49.2%	0.16	<0.001
Lives in a house	89.5%	64.4%	0.55	<0.001
Rural/urban	2.34 (1.08)	2 (1.07)	0.32	<0.001

Rural/urban: 1 = urban; 2 = 2nd city; 3 = suburban; 4 = town and rural i.e. higher is "more rural",
ASB absolute standardized bias

may also be associated with general health, differences in such individual characteristic must be accounted for when examining the effect of dog ownership. Table 3 shows the distribution of individual characteristics by ownership group; the groups differ significantly on almost all covariates. While dog owners and non-owners do not appear to differ in gender, younger individuals and non-Whites are less likely to own a dog. Dog owners tend to have higher incomes and lower likelihood of receiving TANF. Dog owners are more likely to be married, to work full time, to have a spouse that works full time, to live in a house, and to live in a more rural area.

Table 4 shows the results of our analysis. Our treatment effect of interest was the difference in general health between dog owners and non-dog owners. The unweighted mean general health among dog owners was 3.84, while the unweighted mean general health among non-dog owners was 3.57. The difference, 0.27, has a standard error of 0.02 and is statistically significant. After accounting for selection bias using `twang`, the estimated treatment effect reduced to 0.04 with a standard error of 0.03 and was no longer significant. Using `twang` removed differences in the means by balancing the groups on the covariates (e.g. the maximum ASB was just 0.09 for `twang` and it was 0.85 unweighted), but this reduction came at a cost of over a 50% increase in the standard error. This is the type of situation where alternative methods for tuning the number of trees could potentially be of value.

Interestingly, in this example, the MSE-based $t^m(X, Y)$ method and both p value methods $t^p(0.05)$ and $t^p(0.10)$, selected the exact same number of trees as `twang`. For the

Table 4 Treatment effect estimates using each approach with corresponding standard error, p value, design effect, maximum ASB and number of trees for dog ownership study; $t^m(Y,X)$ indicates the MSE-based method that uses outcome information, $t^m(X)$ indicates the MSE-based method that does not use outcome information, $t^a(0.20)$ and $t^a(0.10)$ indicate the ASB-based methods that use a 0.20 and 0.10 threshold, respectively, $t^p(0.05)$ and $t^p(0.10)$ indicate the p value based methods that use a 0.05 and 0.10 threshold, respectively

	Unweighted	`twang`	$t^a(.20)$	$t^a(.10)$	$t^m(Y,X)$	$t^m(X)$	$t^p(0.05)$	$t^p(0.10)$
Case study: dog ownership								
Estimate	0.27	0.04	0.06	0.03	0.04	0.03	0.04	0.04
SE	0.02	0.03	0.03	0.03	0.03	0.03	0.03	0.03
p value	<0.01	0.29	0.04	0.31	0.29	0.29	0.29	0.29
Design effect	1	1.57	1.28	1.53	1.57	1.56	1.57	1.57
Maximum ASB	0.85	0.09	0.19	0.10	0.09	0.09	0.09	0.09
Number of trees	–	2684	274	1698	2684	2465	2684	2684

p value methods, our large sample resulted in significant differences for at least one covariate at the 0.10 and even the 0.05 level. Thus, these methods did not allow for a reduction in trees. For the MSE-based method, even though we estimate the R^2 used in the $t^m(Y,X)$ method at just 0.18, because of the large sample, even a small increase in imbalance in the covariates coming from fewer trees was too costly, in terms of increasing our approximated MSE. Not surprisingly, the $t^m(X)$ approach performs similarly to $t^m(Y,X)$; this method assumes that $R^2 = 0.20$ (close to our estimated 0.18) and the maximum ASB among all covariates was close to our estimate of $E[Y \mid X]$. In contrast, the ASB-based methods $t^a(0.20)$ and $t^a(0.10)$, which rely only on ASB of the covariates and are invariant to sample size, yielded somewhat different results. Since the maximum ASB using `twang` was 0.09, letting the maximum increase to 0.10, with $t^a(0.10)$ had limited effects on the estimated treatment effect or its standard error, even though the number of trees fell by almost 1000. In our experience, it is common for the balance and treatment effects to be similar across relatively large ranges for the number of trees. Allowing for greater imbalance by using the $t^a(0.20)$ rule had a greater impact on the estimated treatment effect. However, even though the design effect decreased from 1.57 to 1.28, the change in the standard error was negligible.

In summary, the similarity in performance for the MSE and p value based methods compared to `twang` is expected given (a) the large sample size and (b) the degree of imbalance in the baseline covariates. Both the R^2 and ASB values result in the $t^m(Y,X)$ and $t^m(X)$ methods allowing for no or little reduction in trees because the price in terms of potential bias was too great.

Figure 1 displays the standardized bias for a subset of the methods for each covariate and demonstrates that the weighted methods, except the ASB-based method with the 0.20 threshold, control the standardized bias and result in well-balanced groups. Here it can be seen quite clearly that the ASB-based tree method allows for greater imbalance between the treatment and control groups when compared to the other tree based methods. In doing so, it provides a more dramatic reduction in variability of the propensity score weights and therefore a smaller design effect but yet clearly at a potentially large cost to bias. With the large sample size in this analysis, power may not be an issue and thus, it might not be necessary to tolerate an increase in bias in exchange for more precise treatment effect estimates.

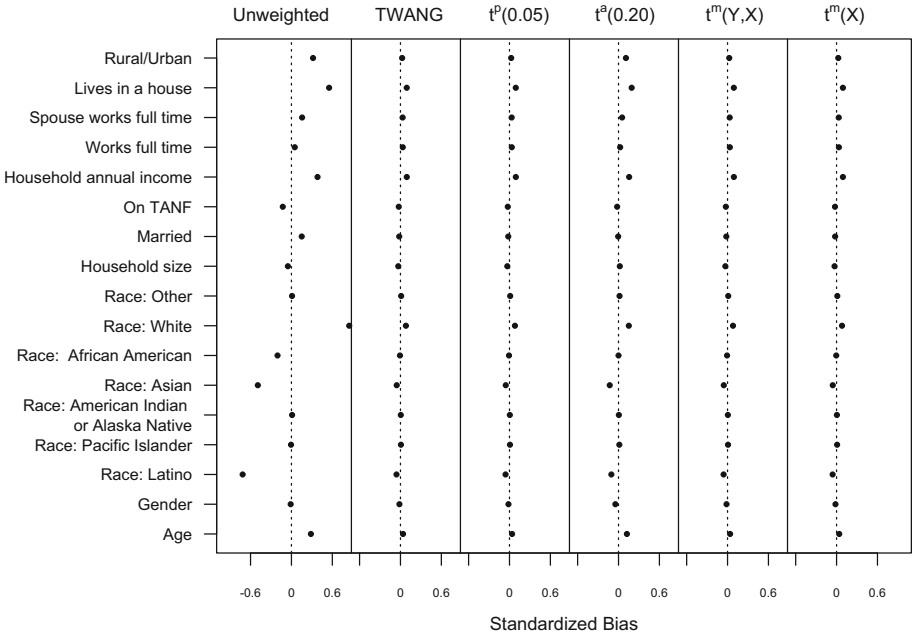

Fig. 1 Standardized bias for each covariate in the dog ownership case study using `twang`, $t^p(0.05)$, $t^a(0.20)$, $t^m(Y,X)$, and $t^m(X)$

5.2 Abstinence and long-term economic outcomes

In our second case study, we applied our proposed methods to investigate the relationship between abstinence and a long-term economic outcome among a sample of high-risk youth (Griffin et al. 2011). Here, it is of interest to understand whether youth who demonstrate short-term successes after substance abuse treatment are less likely to experience adverse long-term economic outcomes compared to those who do not demonstrate such short-term successes (Kaestner 1990, 1994; Ringel et al. 2006, 2007; Register and Williams 1992). Detailed information on this study may be found in Griffin et al. (2011). As might be expected, one of the most challenging aspects of their analysis was controlling for the observed differences in pre-treatment covariates between the youth who abstained from using drugs versus those who used drugs. We utilize data from this case study to illustrate the relative performance of our proposed methods on case study data where selection is very strong and sample size is much smaller than in the dog ownership study.

The dataset for this case study includes 353 adolescent offenders in Los Angeles who were adjudicated as delinquent and sent to one of seven residential group homes from February 1999 to May 2000 (Morral et al. 2004). Youth data come from the baseline, 3, 6, 12, and 87-month follow-ups where each youth was interviewed using the Global Appraisal of Individual Needs (GAIN), a structured clinical interview that collects information on eight main topic domains (background, substance use, physical health, risk behaviors, mental health, and environment, legal, as well as vocational factors) (Dennis 1999). Among the 353 youth in the dataset, 22% abstained from using drugs between baseline and the 12-month follow-up. In this analysis, we aimed to balance the abstainers and the drug users on four individual level characteristics: internal mental distress scale (a

count of past-year symptoms related to internalizing disorders including somatic, anxiety, depression, traumatic stress and suicide/homicide thoughts), the social risk scale (a sum of items indicating how many people (none, a few, some, most, all) the respondent hangs out with socially are involved in drug use, getting drunk, fighting, illegal activities, school or work, treatment, or are in recovery), the substance frequency scale (7-item scale that sums days of use during the past 90 days for alcohol, marijuana, and other illicit drugs), and number of days in the past 90 the youth was drunk or high for most of the day. As shown in Table 5, youth who abstained from using drugs were significantly better on all four pre-treatment measures. They had lower means on all four variables prior to weighting and the differences were quite large for substance use variables with ABS of 1.0 and for social risk with an ABS of greater than 0.6.

Our aim was to examine the effect of abstinence on an economic outcome at the 87-month follow-up. Thus, our outcome measure was total legitimate income which was measured using responses to the question: "During the past 90 days, about how much money did you receive from wages or salary from a legitimate job or business?". Table 6 shows the results of our analysis; here, we estimated the ATT (average treatment effect in the treated) where our treatment effect of interest was the difference in total legitimate income during the past 90 days between abstainers and drug users for youth like those who abstained. The unweighted mean 90-day income among abstainers was $2992.30, while the unweighted mean 90-day income among drug users was $1815.82. The unweighted analysis results show that this difference was significant; that is, youth like those who abstained earned significantly higher 90-day income in young adulthood than youth who did not abstain. However, given the large differences in pre-treatment measures between groups, concern existed that the estimated effect may be biased by selection. These results show that `twang` was very successful at balancing the group, reducing the maximum ABS form 1.0 to 0.05; however, achieving this balance required relatively large variability in the weights with a DEFF of 1.74 and a standard error that is nearly 25% larger than that for the unweighted analysis. Given the small ABS for `twang` and large DEFF there seems to be potential for an alternative tuning method to yield a more accurate estimate.

All the alternative methods result in large reductions in the number of trees. Because `twang` yields a maximum ABS of 0.05, we can reduce the number of trees by large numbers and still have the maximum ABS below 0.20 or even 0.10 and the $t^a(0.10)$ and $t^a(0.20)$ approaches reduce the number of trees by 73 and almost 90%, with corresponding reductions in DEFF of 18 and 28% and standard errors of 10 and 14%, respectively. Similarly because the sample size is so small even relatively large differences in covariate means across groups are not significant, so both the $t^p(0.05)$ and $t^p(0.10)$ result in very reduced numbers of trees (229 and 286, respectively, because large differences are required for rejection at $p = 0.05$) and substantially reduced DEFFs and standard errors.

Table 5 Distribution of baseline covariates among abstainers versus drug users

	Abstainers Mean (SD) n = 84	Drug users Mean (SD) n = 269	ASB	*p* value
Internal mental distress scale	5.80 (4.7)	6.62 (5.6)	0.18	0.1790
Substance frequency scale	0.08 (0.1)	0.22 (0.2)	1.00	<0.001
Social risk scale	7.49 (4.6)	10.43 (4.9)	0.64	<0.001
Number of days drunk/high	8.15 (19.1)	27.41 (32.7)	1.01	<0.001

Table 6 Treatment effect estimates using each approach with corresponding standard error, p value, design effect, maximum ASB and number of trees for abstinence study; $t^m(Y,X)$ indicates the MSE-based method that uses outcome information, $t^m(X)$ indicates the MSE-based method that does not use outcome information, $t^a(0.20)$ and $t^a(0.10)$ indicate the ASB-based methods that use a 0.20 and 0.10 threshold, respectively, $t^p(0.05)$ and $t^p(0.10)$ indicate the p value based methods that use a 0.05 and 0.10 threshold, respectively

	Unweighted	`twang`	$t^a(0.20)$	$t^a(0.10)$	$t^m(Y,X)$	$t^m(X)$	$t^p(0.05)$	$t^p(0.10)$
Case study: abstinence								
Estimate	1176.48	1151.26	1357.08	1307.31	1177.8	1357.08	1361.71	1357.08
SE	417.93	516.51	445.36	465.55	417.98	445.36	439.78	445.36
p value	0.01	0.03	<0.01	0.01	0.01	<0.01	<0.01	<0.01
Design effect	1.00	1.74	1.25	1.42	1.00	1.25	1.21	1.25
Maximum ASB	1.01	0.05	0.18	0.10	1.00	0.18	0.24	0.18
Number of trees	–	2797	286	743	1	286	229	286

For $t^m(X)$, our assumption that $R^2 = 0.20$ turned out to be substantially larger than the estimate which equaled 0.01; this method produced results similar to the p value based methods. However, the $t^m(Y,X)$ approach used the estimated 0.01 value, which explains why this method allows for a dramatic reduction in trees. In essence, this method is willing to allow for more imbalance because the very low R^2 indicates that these variables are not strongly associated with the outcome, and are thus not likely to be confounders. Given this low R^2 and small sample size, the price in terms of design effect that is required in order to obtain balance, as shown by `twang`, is not deemed worthwhile using the MSE approach. The $t^m(Y,X)$ approach essentially results in an estimate equal to the unweighted estimate. The selection of a single tree by the $t^m(Y,X)$ method is simply an artifact of our grid approach which defined $\mathcal{L}$ as a grid of L equally spaced points between 1 and t^*.

This case study demonstrates one of the potential risks of developing the adjustment for selection without use of the outcome information. The four selected pre-treatment variables were identified as clear risk factors for later substance use and negative outcomes by experts and the literature, but they turned out to be unrelated to outcomes measured 87 months later for the adolescents in this study. By relying on outcome information, $t^m(Y,X)$ chooses to ignore the covariate imbalance and essentially return unweighted results. The other approaches ignore the outcomes and choose a GBM model which yields variable weights so as to balance the covariates and potentially unnecessarily degrade the precision of the estimated effects.

6 Discussion

In this paper we consider the problem of the inherent variance-bias tradeoff in using weighting methods to control for selection in observational studies: balancing covariates to remove bias results in variable weights which may potentially reduce the precision of estimated treatment effects. In the context of selecting the complexity of GBMs for the probability of treatment, we examined several alternative approaches to select the number of trees used in the GBM. One of our approaches aimed to explicitly consider both balance in the covariates and the variability in the resulting weights and attempted to select a model that would minimize the MSE of the estimated treatment effect. We also examined two

other approaches that focused specifically on the covariate balance by selecting the smallest number of trees such that some specified threshold, in terms of its standardized absolute value or statistical significance tests p values, was achieved. Although these latter two alternatives do not explicitly account for variance when tuning the GBM, they are straightforward and easy to implement without complex derivations. They are unlike the approach which attempts to minimize MSE, which requires additional computations and most importantly, requires either use of (a) the outcome data to estimate the strength of the covariates in terms of outcome prediction or (b) an educated guess about the predictive strength. All of the examined methods aim to reduce the variance of the propensity score weights at the expense of a (hopefully) small amount of bias.

The results from the simulation study and two case studies show that the methods generally work as expected in terms of reducing the number of trees, DEFF, MSE and increasing bias. Surprisingly, it does not appear than any one method is superior for multiple settings, when comparing to standard `twang`. When the sample size is large, a change in the DEFF has less of an impact on the MSE of the treatment effect and thus, increasing imbalance is generally more costly, and tuning GBM by any method other than minimizing imbalance (the standard method of `twang`) is generally suboptimal. The MSE-based method results essentially confirm this by consistently selecting a similar number of trees as `twang`. Also with a large sample size, even small differences in group means are significant and thus, the p value-based methods also perform similarly to `twang`. In contrast, the ABS threshold method is completely insensitive to sample size and can perform poorly with large samples as demonstrated by increased bias in setting (ii) for $t^a(0.20)$ in our simulation study with $n = 2000$ and for the case study on pet ownership. On the other hand, with small samples, the p value based methods will tend to underweight the costs of imbalance because only large differences are significant and thus, for small samples this method has a great risk of poor performance in terms of increases in bias. With smaller samples, the MSE-based method allows both the imbalance and the predictive strength (R^2) to drive the tree selection. When the predictive strength is very weak, this method will likely indicate that an unweighted approach is optimal which is reasonable given that this likely implies that the covariates being used in the propensity score model are not confounders. However, as shown in the simulation results, this may be dangerous if there is still some association between the covariates and the outcome, and a selection bias, and may lead to an increase in bias that is not compensated by a substantial enough reduction in variance.

The theory behind the MSE-based approach attempts to account for all the relevant factors, adjusting the relative cost of DEFF and imbalance depending on the strength of the covariate-outcome relationship and the sample size, and this is reflected in the number of trees, the DEFF and the imbalance of the covariates in our simulation studies. However, the MSE of $t^m(Y, X)$ is generally not smaller than the MSE of the standard `twang` approach. We expect that the primary reason for the surprising result is that the variance in the estimated treatment effect is not sufficiently reduced by the method. We expect several factors are at play. First, the DEFF is not fixed but varies, and hence the variance of the weighted mean of the residuals is greater than $\{(1 - R^2)D_t\}/\{nq(1 - q)\}$. Second, we estimate R^2 and the regression coefficients, and this adds to variability in the amount of estimated imbalance, which increases the contribution of the imbalance to the variance of the estimated treatment effect. Third, we use squared difference in weighted covariate means from one sample to estimate Δ^2, the expected value of this squared difference, and this too increases the contribution of imbalance to the variance of the estimated treatment

effects. Because of these factors the MSE-based method tends to not achieve the expected reduction in variance but has closer to the expected increase in bias, and consequently the MSE in the treatment effect is larger than expected. Our results suggest that improving on the standard `twang` approach of tuning the model to minimize imbalance may be difficult in practice, in part because increasing imbalance is costly in terms of both bias and variance.

The implications from our findings are important to consider. While propensity score weighting has been shown to reduce bias when estimating treatment effects, it also often reduces power at the same time. In all cases, a certain reduction in power must be expected due to the variability in the estimated weights. Our work highlights that there are potentially meaningful ways to optimize propensity score machine learning methods to allow for minimal bias and less variability. However, caution should be used, particularly with small sample sizes. As discussed above, substantial improvement on the standard `twang` approach of tuning the model to minimize imbalance may be difficult in practice. It is important to note that the fine-tuning required by GBM is similar to the fine-tuning inherent in almost all machine learning methods. We expect that the alternative approaches we have explored in this paper may lend themselves nicely to other machine learning methods such as LASSO, splines, the superlearner, or random forests (Tibshirani 1996; Breiman 2001; van der Laan et al. 2007). Another alternative one might consider would be weight trimming. However, Lee et al. (2011) demonstrated that weight trimming after use of `twang` to obtain propensity score weights does not improve performance compared to no trimming of `twang` weights. In fact, they found that in some cases weight trimming can induce bias. In general, weight trimming cannot be optimal because trimming weights will increase imbalance in the covariates to reduce variability in the weights and weight trimming approaches generally do not account for the fact that the relative costs of imbalance in the covariate and variability in the weights depends on R^2 and the sample size. Hence, like the ABS method, when the sample size or R^2 is large weight trimming is likely to perform poorly.

Our work should be considered along with its limitations. First, the use of outcome information in one of the proposed MSE-based approaches should be carefully considered. Some recent research has strongly cautioned against the use of outcome information in the estimation of propensity score weights (Stuart et al. 2013; Rubin 2004; Rosenbaum 2010; Hansen 2008). As described by Stuart et al. (2013), propensity score methods tend to be conducted without use of the outcome variable in an effort to separate the design and analysis stages of a study and allow for use of a single set of propensity scores for multiple outcomes. However, others have argued that without the use of outcome information, instrumental variables which are related to the treatment but not related to the outcome may be included in the propensity score model and result in decreased precision (Brookhart et al. 2006; Westreich et al. 2011). Second, our simulation study includes only a small number of conditions which, although chosen to control features of the data which would affect the performance of alternative approaches to balance bias and variance when tuning the propensity score model, did not produce data with large design effects. The relative performance of alternative methods may differ in settings with more variable weights, like the case studies. However, by including the case studies we are able to demonstrate the impact of the proposed alternatives in cases with more variable weights and in particular, show that the MSE-based approach can guard against modeling with a large number of instruments (abstinence example) which, as noted above, can degrade the accuracy of estimated treatment effects. A third limitation of our work is our focus on the balance

metric ASB; however, there are a number of other available metrics to assess balance (Imbens and Rubin 2015). Additionally, while the data from our case studies produce highly variable weights as demonstrated by the large design effects, our simulation study does not appear to produce situations with highly variable weights. We expect that these alternatives may perform differently in more extreme simulated data structures. Another limitation of `twang` and our proposed approaches is that they can be computationally intensive. Our alternatives rely on a grid search algorithm and thus can require long amounts of processing time. Lastly, as with any propensity score approach, we require the strong assumption that there are no unmeasured confounders. This assumption is impossible to test in practice but one could (and should) consider sensitivity analyses to examine how sensitive the observed findings might be to violations of this assumption (Griffin et al. 2013; Rosenbaum and Rubin 1983a; Higashi et al. 2005).

Funding This study was funded by National Institutes of Health grant 1R01DA034065-01A1 and National Institute of Child Health and Human Development grant R01HD066591.

Compliance with ethical standards

Conflict of interest The authors declare that they have no conflict of interest.

Ethical approval and informed consent This study used only secondary de-identified datasets.

References

Austin, P.C.: The performance of different propensity score methods for estimating marginal odds ratios. Stat. Med. **26**(16), 3078–3094 (2007)

Austin, P.C.: Balance diagnostics for comparing the distribution of baseline covariates between treatment groups in propensity-score matched samples. Stat. Med. **28**, 3083–3107 (2009)

Austin, P.C., Stuart, E.A.: Moving towards best practice when using inverse probability of treatment weighting (IPTW) using the propensity score to estimate causal treatment effects in observational studies. Stat. Med. **34**(28), 3661–3679 (2015)

Breiman, L.: Random forests. Mach. Learn. **45**(1), 5–32 (2001)

Breiman, L., Friedman, J., Stone, C.J., Olshen, R.A.: Classification and regression trees. CRC Press, New York (1984)

Brookhart, M.A., Schneeweiss, S., Rothman, K.J., Glynn, R.J., Avorn, J., Stürmer, T.: Variable selection for propensity score models. Am. J. Epidemiol. **163**(12), 1149–1156 (2006)

Burgette, L., McCaffrey, D.F., Griffin, B.A.: Propensity score estimation with boosted regression. In: Pan, W. (ed.) Propensity Score Analysis: Fundamentals and Developments. Guilford Publications, New York (2015)

California Health Interview Survey (CHIS): CHIS 2003 Methodology Report Series. UCLA Center for Health Policy Research, Los Angeles, CA (2005)

Dennis, M.L.: Overview of the Global Appraisal of Individual Needs (Gain): Summary. Chestnut Health Systems, Bloomington, IL (1999)

Golinelli, D., Ridgeway, G., Rhoades, H., Tucker, J., Wenzel, S.: Bias and variance trade-offs when combining propensity score weighting and regression: with an application to hiv status and homeless men. Health Serv. Outcomes Res. Method. **12**(2–3), 104–118 (2012)

Griffin, B.A., Ramchand, R., Edelen, M.O., McCaffrey, D.F., Morral, A.R.: Associations between abstinence in adolescence and economic and educational outcomes seven years later among high-risk youth. Drug Alcohol Depend. **113**(2), 118–124 (2011)

Griffin, B.A., Eibner, C., Bird, C.E., Jewell, A., Margolis, K., Shih, R., Slaughter, M.E., Whitsel, E.A., Allison, M., Escarce, J.J.: The relationship between urban sprawl and coronary heart disease in women. Health Place **20**, 51–61 (2013)

Hankey, B.F., Myers, M.H.: Evaluating differences in survival between two groups of patients. J. Chronic Dis. **24**(9), 523–531 (1971)

Hansen, B.B.: The prognostic analogue of the propensity score. Biometrika **95**(2), 481–488 (2008)

Harder, V.S., Stuart, E.A., Anthony, J.C.: Propensity score techniques and the assessment of measured covariate balance to test causal associations in psychological research. Psychol. Methods **15**(3), 234 (2010)
Hernán, M.Á., Brumback, B., Robins, J.M.: Marginal structural models to estimate the causal effect of zidovudine on the survival of HIV-positive men. Epidemiology **11**(5), 561–570 (2000)
Higashi, T., Shekelle, P.G., Adams, J.L., Kamberg, C.J., Roth, C.P., Solomon, D.H., Reuben, D.B., Chiang, L., MacLean, C.H., Chang, J.T., et al.: Quality of care is associated with survival in vulnerable older patients. Ann. Intern. Med. **143**(4), 274–281 (2005)
Hill, J.L.: Bayesian nonparametric modeling for causal inference. J. Comput. Graph. Stat. **20**(1), 217–240 (2011)
Imai, K., Ratkovic, M.: Covariate balancing propensity score. J. R. Stat. Soc. Ser. B (Stat. Method.) **76**(1), 243–263 (2014)
Imbens, G.W.: The role of the propensity score in estimating dose-response functions. Biometrika **87**(3), 706–710 (2000)
Imbens, G.W., Rubin, D.B.: Causal Inference in Statistics, Social, and Biomedical Sciences. Cambridge University Press, Cambridge (2015)
Kaestner, R.: The effect of illicit drug use on the wages of young adults. Tech. rep., National Bureau of Economic Research (1990)
Kaestner, R.: New estimates of the effect of marijuana and cocaine use on wages. Ind. Labor Relat. Rev. **47**(3), 454–470 (1994)
Kang, J.D., Schafer, J.L.: Demystifying double robustness: a comparison of alternative strategies for estimating a population mean from incomplete data. Statistical Science, pp. 523–539 (2007)
Lee, B.K., Lessler, J., Stuart, E.A.: Improving propensity score weighting using machine learning. Stat. Med. **29**(3), 337–346 (2010)
Lee, B.K., Lessler, J., Stuart, E.A.: Weight trimming and propensity score weighting. PloS One **6**(3), e18,174 (2011)
Lee, S., Brown, E.R., Grant, D., Belin, T.R., Brick, J.M.: Exploring nonresponse bias in a health survey using neighborhood characteristics. Am. J. Public Health **99**(10), 1811 (2009)
Liaw, A., Wiener, M.: Classification and regression by randomforest. R News **2**(3), 18–22 (2002)
McCaffrey, D.F., Ridgeway, G., Morral, A.R.: Propensity score estimation with boosted regression for evaluating causal effects in observational studies. Psychol. Methods **9**(4), 403 (2004)
McConnell, A.R., Brown, C.M., Shoda, T.M., Stayton, L.E., Martin, C.E.: Friends with benefits: on the positive consequences of pet ownership. J. Personal. Soc. Psychol. **101**(6), 1239 (2011)
Morral, A.R., McCaffrey, D.F., Ridgeway, G.: Effectiveness of community-based treatment for substance-abusing adolescents: 12-month outcomes of youths entering phoenix academy or alternative probation dispositions. Psychol. Addict. Behav. **18**(3), 257 (2004)
Normand, S.L.T., Landrum, M.B., Guadagnoli, E., Ayanian, J.Z., Ryan, T.J., Cleary, P.D., McNeil, B.J.: Validating recommendations for coronary angiography following acute myocardial infarction in the elderly: a matched analysis using propensity scores. J. Clin. Epidemiol. **54**(4), 387–398 (2001)
Pirracchio, R., Petersen, M.L., van der Laan, M.: Improving propensity score estimators' robustness to model misspecification using super learner. Am. J. Epidemiol. **181**(2), 108–119 (2015)
Ponce, N.A., Lavarreda, S.A., Yen, W., Brown, E.R., DiSogra, C., Satter, D.E.: The california health interview survey 2001: translation of a major survey for california's multiethnic population. Public Health Rep. **119**(4), 388 (2004)
Register, C.A., Williams, D.R.: Labor market effects of marijuana and cocaine use among young men. Ind. Labor Relat. Rev. **45**(3), 435–448 (1992)
Ridgeway, G.: gbm: Generalized Boosted Regression Models. R package version 2.1.1. Retrieved from cran.r-project.org (2015)
Ridgeway, G., McCaffrey, D., Morral, A., Griffin, B.A., Burgette, L.: Twang: Toolkit for Weighting and Analysis of Nonequivalent Groups. R package version 9.5. Retrieved from cran.r-project.org (2016)
Ringel, J.S., Collins, R.L., Ellickson, P.L.: Time trends and demographic differences in youth exposure to alcohol advertising on television. J. Adolesc. Health **39**(4), 473–480 (2006)
Ringel, J.S., Ellickson, P.L., Collins, R.L.: High school drug use predicts job-related outcomes at age 29. Addict. Behav. **32**(3), 576–589 (2007)
Robins, J.M., Hernán, M.Á., Brumback, B.: Marginal structural models and causal inference in epidemiology. Epidemiology **11**(5), 550–560 (2000)
Rosenbaum, P.R.: Various practical issues in matching. In: Design of Observational Studies, pp. 187–195. Springer, New York (2010)
Rosenbaum, P.R., Rubin, D.B.: Assessing sensitivity to an unobserved binary covariate in an observational study with binary outcome. J. R. Stat. Soc. Ser. B (Methodol.) **45**(2), 212–218 (1983a)

Rosenbaum, P.R., Rubin, D.B.: The central role of the propensity score in observational studies for causal effects. Biometrika **70**(1), 41–55 (1983b)
Rosenbaum, P.R., Rubin, D.B.: Reducing bias in observational studies using subclassification on the propensity score. J. Am. Stat. Assoc. **79**(387), 516–524 (1984)
Rubin, D.B.: On principles for modeling propensity scores in medical research. Pharmacoepidemiol. Drug Saf. **13**(12), 855–857 (2004)
Stuart, E.A., Lee, B.K., Leacy, F.P.: Prognostic score-based balance measures can be a useful diagnostic for propensity score methods in comparative effectiveness research. J. Clin. Epidemiol. **66**(8), S84–S90 (2013)
Survey, C.H.I.: Technical Paper No. 1: The chis 2001 Sample: Response Rate and Representativeness. Ucla Center for Health Policy Research, Los Angeles, CA (2003)
Tibshirani, R.: Regression shrinkage and selection via the lasso. J. R. Stat. Soc. Ser. B (Methodol.) **58**(1), 267–288 (1996)
van der Laan, M.J.: Targeted estimation of nuisance parameters to obtain valid statistical inference. Int. J. Biostat. **10**(1), 29–57 (2014)
van der Laan, M.J., Polley, E.C., Hubbard, A.E.: Super learner. Stat. Appl. Genet. Mol. Biol. (2007). doi:10.2202/1544-6115.1309
Wells, D.L.: Associations between pet ownership and self-reported health status in people suffering from chronic fatigue syndrome. J. Altern. Complement. Med. **15**(4), 407–413 (2009a)
Wells, D.L.: The effects of animals on human health and well-being. J. Soc. Issues **65**(3), 523–543 (2009b)
Westreich, D., Cole, S.R., Funk, M.J., Brookhart, M.A., Stürmer, T.: The role of the c-statistic in variable selection for propensity score models. Pharmacoepidemiol. Drug Saf. **20**(3), 317–320 (2011)