Health Services and Outcomes Research Methodology (2019) 19:106–126
https://doi.org/10.1007/s10742-018-0187-2

Causal inference for multi-level treatments with machine-learned propensity scores

Lin Lin[1] · **Yeying Zhu**[1] · **Liang Chen**[2]

Received: 24 February 2017 / Revised: 11 August 2018 / Accepted: 23 August 2018 /
Published online: 30 August 2018

Abstract
Propensity score-based methods have been widely developed to adjust for confounders in observational studies to estimate causal treatment effect for binary treatments. We generalize these causal inference methods to the multi-level treatment case. We review the generalized causal inference framework and several propensity score estimation methods. We conduct a comprehensive simulation study to evaluate the performance of multinomial logistic regression, generalized boosted models, random forest and data adaptive matching score for estimating propensity scores based on inverse probability of treatment weighting. From our findings, multinomial logistic regression is susceptible to yielding extreme weights while a mis-specified model is assumed, which results in poor performance of the inverse probability weighted estimator. On the other hand, machine-learned propensity scores tend to have less biased and more stable performance, and the data adaptive matching score tends to perform the best overall. The above-mentioned propensity score based methods are applied to the Taobao dataset to evaluate the causal effect of reputation on sales.

Keywords Causal inference · Generalized propensity score · Multinomial logistic regression · Generalized boosted model · Random forest · Data adaptive matching score

1 Introduction

When investigating the causal effect of a treatment, one of the most common issues in observational studies is that the groups of subjects in the study are not randomly assigned. With the presence of confounders, which are the variables associated with both the treatment and

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s10742-018-0187-2) contains supplementary material, which is available to authorized users.

✉ Yeying Zhu
yeying.zhu@uwaterloo.ca

[1] Department of Statistics and Actuarial Science, University of Waterloo, Waterloo, ON N2L 3G1, Canada

[2] Economics and Management School, Wuhan University, Wuhan 430072, Hubei, People's Republic of China

the outcome, non-random group assignment could result in biased estimates of the treatment effect. Propensity score-based methods (Rosenbaum and Rubin 1983) are often used to adjust for confounding. In the case of a binary treatment, i.e. treatment versus control, the propensity score is defined to be the conditional probability of being assigned to the treatment group given the observed covariates. The propensity score acts as a summary statistic that incorporates the information on all the observed covariates so that instead of balancing each covariate individually, causal inference can be made by simply balancing on the propensity score. By doing so, the bias due to all observed confounders can be removed (Rosenbaum and Rubin 1983).

The initial work on propensity scores focused on the case of binary treatment, though there is more recent work generalizing the propensity score to treatments with more than two levels. This stream of work began with the definition of a generalized propensity score, which is the conditional probability of being assigned to a particular treatment group given the observed covariates (Imbens 2000). Regression adjustment for propensity scores as well as inverse probability of treatment weighting (IPTW) were generalized to the multi-level treatment case (Imbens 2000). Then propensity score matching was generalized to the multi-level treatment case (Lechner 2001), followed by a generalized version of propensity score subclassification (Imai and Dyk 2004). Further work generalized doubly robust estimation using propensity scores to the multi-level treatment case (Uysal 2014; Tu et al. 2013). In this paper, we focus on IPTW to estimate causal effects.

An important topic to address in this paper is the estimation of the generalized propensity score. A natural extension to the logistic regression model for the case of binary treatment would be the multinomial logistic regression model for multi-level treatments (e.g., Spreeuwenberg et al. (2010) & Feng et al. (2012)). For treatments or treatment levels that are correlated, Imbens (2000) recommended nested logistic models; Lechner (2002) and Imai and Dyk (2004) suggested that multinomial probit models can be used. A comparison of the above-mentioned methods can be found in (Tchernis et al. 2005). Recently, machine learning methods have also been considered for propensity score estimation. These methods include classification and regression trees (CART), bagged CART, random forests (RF), and generalized boosted models (GBM), which were shown to be less biased and have better 95% confidence interval coverage than logistic regression for the binary case (Lee et al. 2010). A model-averaged approach called data-adaptive matching score (DAMS) that balances proprensity score estimates from parametric (logistic regression) and non-parametric (random forest) models was also suggested (Zhu et al. 2014). Recently, GBM were generalized to the multi-level treatment case (McCaffrey et al. 2013). In this article, we generalize RF and DAMS to the multi-level treatment case and to evaluate their performance through comprehensive simulation studies.

The rest of the paper is organized as follows. In Sect. 2, we review the causal inference framework for multi-level treatments. In Sect. 3, we describe several existing propensity score estimation methods in the literature and extend some of them to the multi-level treatments. We then discuss the IPTW method for estimating causal treatment effects in Sect. 4. Section 5 describes a simulation study to evaluate the performance of several propensity score estimation methods for multi-level treatments. A data application to the Taobao dataset is presented in Secti. 6. Lastly, we conclude with a discussion in Sect. 7.

2 The causal inference framework for multi-level treatments

Suppose that we have a total of n subjects, indexed i=1,...,n. Let A_i be the observed treatment status for the ith subject, so $A_i = t$ if subject i was observed under treatment $t \in \{1,...,M\}$, where there are a total of M treatments. We denote indicator variables, indicating membership of a particular treatment group t, as $A_i(t) = I(A_i = t)$. Let Y_i denote the observed outcome for subject i. We further define a set of potential outcomes $\{Y_i(1), \ldots, Y_i(M)\}$ where $Y_i(t)$ is the hypothetical outcome value if the ith subject were assigned to treatment t (Imbens 2000; Imai and Dyk 2004). To be noted, we only observe one of the potential outcomes for each subject. Let X_i denote the $p \times 1$ vector of p observed pre-treatment covariates for subject i.

We are interested in two types of causal effects: the average treatment effect (ATE) and the average treatment effect among the treated (ATT). The ATE of treatment t relative to treatment s is the difference of mean outcomes had the entire population been observed under t versus had the entire population been observed under s. The ATE would be of interest if we were interested in evaluating the overall effect of the treatment(s) on the population.

$$\begin{aligned} ATE_{ts} &:= E[Y(t)] - E[Y(s)] \\ &= \mu_t - \mu_s. \end{aligned}$$

The average treatment effect of treatment s among those treated with t (ATT of t relative to s) is the difference in mean outcome among subjects who were treated with t versus the mean outcome they would have had if they received s (McCaffrey et al. 2013). The ATT would be of interest if the treatment(s) targets a particular group and we were interested in the effect of the treatment(s) on that group.

$$\begin{aligned} ATT_{t,ts} &:= E[Y(t)|A = t] - E[Y(s)|A = t] \\ &= \mu_{t,t} - \mu_{t,s}. \end{aligned}$$

The generalized propensity score is the conditional probability of subject i receiving treatment t given covariates x (Imbens 2000).

$$\begin{aligned} p_i(t|x) &:= P(A_i = t|X_i = x) \\ &= P(A_i(t) = 1|X_i = x) \end{aligned}$$

For generality going forward, we drop the index in the notation as needed. There are also some key assumptions on the generalized propensity score similar to those for the binary case, and these assumptions are necessary for valid causal inference.

Assumption 1 The stable unit treatment value assumption states that the distribution of potential outcomes for one subject/unit is independent of the treatment assignment of any other subject (Imai and Dyk 2004).

$$(Y_i(1), \ldots, Y_i(M)) \perp A_j \text{ for } i \neq j.$$

This assumption means that a subject's outcome does not depend on the treatment assignment of any other subject, which is a strong assumption that may not hold if, for example, we were investigating treatments for infectious diseases.

Assumption 2 The weak unconfoundedness assumption states that the treatment assignment indicator does not depend on the potential outcome given the observed covariates (Imbens 2000):

$$A(t) \perp Y(t)|X.$$

This is a weaker version of the Strong Unconfoundedness/Ignorability/No Unmeasured Confounders Assumption in the binary case, where the treatment assignment does not depend on the potential outcomes given the observed covariates (Rosenbaum and Rubin 1983). This assumption means that we can model the conditional distribution of the treatment assignment given the covariates without having to condition on the outcome. Furthermore, if this assumption holds, then the treatment assignment is weakly unconfounded given the generalized propensity score (Imbens 2000), i.e.

$$A(t) \perp Y(t)|p(t|X).$$

This assumption indicates that instead of conditioning on the covariates X, it is sufficient to condition on the generalized propensity score $p(t|X)$.

Assumption 3 The sufficient overlap or positivity assumption states that there is a non-zero probability of being assigned to each treatment (Rosenbaum and Rubin 1983; McCaffrey et al. 2013), i.e.,

$$0 < p(t|X) < 1 \quad \text{for all} \quad t, X.$$

This assumption means that for each subject, it is possible that there is at least one comparable person in the population in each of the treatment groups. With this assumption we can estimate the ATE without relying on extrapolation. However, Assumption 3 is harder to satisfy compared to a binary treatment setting, especially as the number of treatment gets large and the dimension of X is large. If this assumption does not hold, one can modify the study population of interest to ensure sufficient overlap (e.g., Crump et al. (2009)) or find a linear combination of X, $X'\beta$, which often has a much lower dimension than X, such that $0 < p(t|X'\beta) < 1$ (Luo et al. 2017).

Given this framework and a correct generalized propensity score model, we can obtain a theoretically unbiased estimate of the average potential outcome for each treatment group by adjusting for the propensity score.

3 Generalized propensity score estimation

In this paper we will focus on four methods for modeling the generalized propensity score $p(t|X)$: multinomial logistic regression (MLR), generalized boosted models (GBM), random forests (RF) and data-adaptive matching score (DAMS). A brief review of these methods can be found in Zhu and Lin (2016). We wish to model the probability of assignment to each of the M treatment levels, conditional on the covariates X. The methods of propensity score estimation differ slightly when estimating the ATE versus the ATT. This is because when estimating the ATE, we need to consider the whole population, whereas when estimating the ATT, we only need to consider the subpopulation consisting of the relevant treatment groups.

3.1 Generalized propensity scores for ATE estimation

In this subsection we will discuss the methods listed above with a focus on ATE estimation. Multinomial logistic regression (MLR) is an extension of logistic regression to cases where the number of classes is larger than two (McFadden 1973) . Instead of assuming an underlying binomial distribution for the treatment conditional on the covariates as in logistic regression, we now assume an underlying multinomial distribution. We assume the generalized propensity score for each treatment level follows:

$$p(t|x)_{MLR} = \frac{1}{1 + \sum_{s=2}^{M} e^{\beta_s' x}} \text{ for } t = 1,$$
$$p(t|x)_{MLR} = \frac{e^{\beta_t' x}}{1 + \sum_{s=2}^{M} e^{\beta_s' x}} \text{ for } t = 2, \ldots, M,$$

where $\beta_s = (1, \beta_{s1}, \ldots, \beta_{sp})$ for $s = 2, \ldots, M$ and p is the number of covariates. Then β's are estimated by maximizing the likelihood function:

$$L(\beta) = \prod_{i=1}^{n} \prod_{t=1}^{M} p_i(t|x)^{A_i(t)}.$$

In MLR, it assumes a parametric model for the propensity score and therefore, the consistency of the subsequent causal effect estimator relies on the correct specification of the model. However, when the dimension of the covariates is large, it is a challenging task and one also needs to consider the possible nonlinear and interaction terms among the covariates.

As an alternative to parametric regression, we now consider two machine learning techniques for propensity score estimation: GBM and RF. The GBM algorithm for the binary treatment case have been outlined in McCaffrey et al. (2004). GBM is built on an ensemble of regression trees and each regression tree iteratively fits the residuals from the previous tree to approximate the propensity score function. The number of trees to be generated is determined by achieving the maximum balance in covariate distribution among different treatment groups. The algorithm automatically includes interaction and nonlinear terms as the regression tree allows for multi-level splits. Zhu et al. (2015) discusses the variable selection issue in GBM.

In the multi-level treatment case, since the ATE defines on the entire population, we are interested in the probability that each subject is assigned to a particular treatment t as opposed to any other treatment (McCaffrey et al. 2013). It is suggested we should fit a GBM that balance the covariates between the treatment t group and the entire sample as in the binary case (McCaffrey et al. 2013). One can repeat the same procedure for each of the M treatments to obtain the generalized propensity scores $\hat{p}(t|x)$ for $t = 1, \ldots, M$.

The third method we will consider is random forests, which is a very popular regression/classification method in machine learning (Breiman 2001). A random forest (RF) is the aggregation of a collection of regression or classification trees fitted on bootstrap samples of the original dataset with the original sample size. Different from GBM in estimating the generalized propensity scores, RF uses a collection of classification trees as opposed to

regression trees. A key feature of RF lies in that each tree in the tree ensemble is built on a random subset of the original covariates on a bootstrap sample of the original dataset to avoid overfitting. Given a vector of covariates, each tree votes for one class label. Then the generalized propensity score for treatment t can be estimated as the fraction of trees that classify/predict the given subject into treatment group t out of the entire collection of trees.

In RF, if none of the trees predict a particular treatment level, then the RF estimator of the propensity score for that treatment will be zero. This is in violation of the positivity assumption and may lead to infinite weight while employing inverse probability weighting method to estimate causal effects. A common way to deal with this issue is to trim extreme or infinite weights to smaller values. However, such a trimming strategy is somewhat arbitrary and according to a simulation study in Lee et al. (2011), trimming RF propensity scores does not significantly reduce bias and standard error of the causal estimates.

Another data-adaptive approach to deal with extreme values without *ad-hoc* adjustment is called the data-adaptive matching score (DAMS), which is a weighted average of the propensity score estimates from a parametric model (such as a MLR model) and a nonparametic model (such as a RF model) (Zhu et al. 2014). This DAMS estimator can be denoted as

$$\hat{p}(t|x)_{DAMS} = \lambda \hat{p}(t|x)_{MLR} + (1-\lambda)\hat{p}(t|x)_{RF}, \tag{1}$$

$$\text{where } \lambda = \frac{\hat{p}(t|x)_{MLR}{}^{A(t)}[1-\hat{p}(t|x)_{MLR}]^{1-A(t)}}{\hat{p}(t|x)_{MLR}{}^{A(t)}[1-\hat{p}(t|x)_{MLR}]^{1-A(t)} + \hat{p}(t|x)_{RF}{}^{A(t)}[1-\hat{p}(t|x)_{RF}]^{1-A(t)}}.$$

According to Zhu et al. (2014), the form of λ has a Bayesian interpretation. If MLR is employed as the parametric model, the combined estimator (1) will never produce a value of 0 or 1. By a series of simulation studies performed in Zhu et al. (2014), the authors in that paper found that by model averaging, the bias and the variance of the treatment effect estimates can always be reduced, compared to LR or RF models alone.

It is obvious that MLR and RF model would produce estimated propensity scores that sum to 1 across all treatments. However, for GBM and DAMS, this is not necessarily the case. In McCaffrey et al. (2013), the authors argue that it is not a problem as long as the weights produced by the estimates can achieve balance between the treatment group of interest and the entire sample. In fact, the important tuning parameter, the number of trees generated by GBM, is suggested to be selected by achieving the best balance between the treatment group of interest and the entire sample.

3.2 Generalized propensity scores for ATT estimation

Recall that when estimating the ATT, we only need to consider the subpopulation consisting of the relevant treatment groups. For example, if we were interested in the ATT of t relative to s, we would only use the subpopulation of treatment group t and treatment group s. Hence, estimating propensity scores in this case is equivalent to the binary treatment case.

Instead of MLR, we can simply use logistic regression on the subpopulation to model $p(t|x)$ by considering treatment t as the treatment and treatment s as the control in the binary case. Similarly, the GBM algorithm for the binary treatment case can be used to estimate $p(t|x)$. Lastly, the RF algorithm is also similar, except now we only classify the

subpopulation into treatment group t and treatment group s. Once we obtain the estimates from LR and RF, we can calculate the DAMS estimator of the propensity score as in (1).

4 Treatment effect estimation

There are various propensity score-based methods to estimate an ATE or an ATT including matching, stratification, IPTW, regression adjustment, and doubly robust estimation (Rosenbaum and Rubin 1983; Rubin 1974). In this paper, we focus on IPTW method.

IPTW works on the principle that observations in any treatment group can be reweighted so that the covariate distribution in that treatment group matches those of any other treatment group provided that the causal inference assumptions hold. This reweighting procedure makes the two groups comparable so that the difference in outcomes between the two groups is not attributed to the differences in covariates between the two groups. Specifically, if we use the inverse of the probability of being in a particular treatment group t (i.e. the inverse of the propensity score) as the weight for an observation in treatment group t, then we can balance treatment group t with other treatment groups (Frölich 2004). The weighting is slightly different when estimating ATEs versus when estimating ATTs.

When estimating ATEs, we are interested in the population average outcome, so the weights need to balance treatment groups of interest with the entire population. Since the propensity score $p(t|x)$ for estimating ATE is the probability of being assigned to treatment t out of all the possible treatments, the propensity score is the inclusion probability of the treatment t group. Then each subject i in the treatment t group represents $\frac{1}{p_i(t|x)}$ people in the population, so the weight for observation i is

$$w_i(t) = \frac{1}{\hat{p}_i(t|x)}.$$

Hence, the estimate for the population average outcome for treatment t is the weighted average of the outcomes in the treatment group t:

$$\hat{\mu}_t = \frac{\sum_{i=1}^{n} A_i(t) w_i(t) Y_i}{\sum_{i=1}^{n} A_i(t) w_i(t)}.$$

The estimated ATE is then the difference in the estimated population average outcomes:

$$\widehat{ATE}_{ts} = \hat{\mu}_t - \hat{\mu}_s.$$

On the other hand, estimating the ATT of treatment t relative to treatment s focuses on the subpopulation of the treatment t group. To estimate the ATT, we need to estimate the average outcome for treatment t for those actually treated with t as well as the average outcome for treatment s for those actually treated with t. To estimate the former, we simply use the unweighted average of the outcomes from the sample treatment t group:

$$\hat{\mu}_{t,t} = \frac{\sum_{i=1}^{n} A_i(t) Y_i}{\sum_{i=1}^{n} A_i(t)},$$

since our sample treatment t group is assumed to be representative of the population treatment t group.

To estimate the latter, we use a weighted average of the outcomes from the sample treatment s group, weighted so that the covariate distribution in the s group matches that of the t group. Each subject i in the treatment s group represents $\frac{1}{p_i(s|x)}$ people in the population, but each similar person in the population has a $p_i(t|x)$ probability of being in treatment group t, so the weight of observation i in group s if he/she were to be in group t is

$$w_i(t,s) = \frac{\hat{p}_i(t|x)}{\hat{p}_i(s|x)}. \tag{2}$$

Note that since estimating the propensity scores for estimating the ATT is equivalent to the binary case, then $\hat{p}_i(s|x) = 1 - \hat{p}_i(t|x)$, so the weight in (2) can be rewritten as

$$w_i(t,s) = \frac{\hat{p}_i(t|x)}{1 - \hat{p}_i(t|x)}.$$

Another explanation for this weighting scheme is that subjects in treatment group s with covariate values that are much more common in their own treatment group than in treatment group t (i.e. $p_i(s|x)$ is very large relative to $p_i(t|x)$ so that $\frac{p_i(t|x)}{p_i(s|x)}$ is small) will have small weights since they are relatively common in group s but not in group t (McCaffrey et al. 2013). On the other hand, subjects with covariate values that are much more common in group t than in group s (i.e. $p_i(s|x)$ is very small relative to $p_i(t|x)$ so that $\frac{p_i(t|x)}{p_i(s|x)}$ is large) will have large weights since they represent group t better and there are few of them in group s (McCaffrey et al. 2013).

Then, the estimated average outcome for treatment s for those actually treated with t is

$$\hat{\mu}_{t,s} = \frac{\sum_{i=1}^{n} A_i(s) w_i(t,s) Y_i}{\sum_{i=1}^{n} A_i(s) w_i(t,s)}.$$

The estimate of the ATT of treatment t relative to treatment s is then the difference between two average outcomes:

$$\widehat{ATT}_{t,ts} = \hat{\mu}_{t,t} - \hat{\mu}_{t,s}.$$

5 Simulation studies

In our simulation study, we investigate the performance of the various propensity score estimation methods described in Sect. 3 along with the IPTW estimator in estimating ATEs and ATTs. The propensity score estimation methods are MLR, GBM and RF. In addition, we compare the performance of the model-averaged estimator, DAMS, with MLR and RF as the parametric and nonparametric components.

5.1 Setup

The simulation structure is adapted from the binary treatment case in Lee et al. (2010). For each simulated dataset, we have the following variables:

(1) Four confounding variables associated with both the treatment assignment and the outcome, X_1, X_2, X_3, and X_4. Here, X_1 and X_3 are binary, while X_2 and X_4 are continuous.
(2) Three covariates associated with the treatment assignment only, X_5, X_6, and X_7. Here, X_5 and X_6 are binary, and X_7 is continuous.
(3) Three covariates associated with the outcome only, X_8, X_9, and X_{10}. Here, X_8 and X_9 are binary, while X_{10} is continuous. Let $X = (X_1, \ldots, X_{10})$ be the vector of all ten covariates.
(4) The three-level treatment variable A with levels 0, 1, and 2. The true generalized propensity scores (probability of assignment into each treatment level) are:

$$P(A = 0|X) = \frac{1}{1 + e^{f_1(X)} + e^{f_2(X)}},$$
$$P(A = 1|X) = \frac{e^{f_1(X)}}{1 + e^{f_1(X)} + e^{f_2(X)}},$$
$$P(A = 2|X) = \frac{e^{f_2(X)}}{1 + e^{f_1(X)} + e^{f_2(X)}},$$

where X is the vector of the covariates and $f_1(X)$ and $f_2(X)$ are functions of X.
(5) The continuous outcome variable Y with

$$E[Y|X] = -3.85 + 0.3X_1 - 0.36X_2 - 0.73X_3 - 0.2X_4 + 0.71X_8 - 0.19X_9 + 0.26X_{10} - 0.4I(A = 1) - 0.7I(A = 2)$$

so we have the following true ATEs: $ATE_{10} = -0.4$, $ATE_{20} = -0.7$, and $ATE_{21} = -0.3$.

The covariates X_1, X_2, X_3, X_4, X_7, and X_{10} are generated through independent normal distributions with mean 0 and standard deviation 1. The rest of the covariates, X_5, X_6, X_8, X_9 are generated from normal distributions and have a correlation structure as follows:

$$corr(X_1, X_5) = 0.2,\ corr(X_2, X_6) = 0.9,\ corr(X_3, X_8) = 0.2,\ corr(X_4, X_9) = 0.9.$$

Then, X_1, X_3, X_5, X_6, X_8 and X_9 are dichtomized at the sample average (1 if the observation is greater than the sample average, 0 otherwise).

We consider several scenarios (Lee et al. 2010) where the treatment assignment is related to the covariates with various degrees of non-linearity. The complexity of the relationships range from a simple scenario such as having additivity and linearity with the main effects only in the functions $f_1(X)$ and $f_2(X)$ (scenario A), to a complicated scenario such as having moderate non-additivity and non-linearity with ten two-way interaction terms and three quadratic terms in $f_1(X)$ and $f_2(X)$ (scenario G). The exact forms of $f_1(X)$ and $f_2(X)$ in each scenario are listed below and 1000 datasets were generated under each scenario with each dataset containing $n = 1000$ observations.

Scenario A

$$f_1(X) = 0.8X_1 - 0.25X_2 + 0.6X_3 - 0.4X_4 - 0.8X_5 - 0.5X_6 + 0.7X_7$$
$$f_2(X) = -0.4X_1 - 0.1X_2 + 0.45X_3 + 0.7X_4 + 0.2X_5 - 0.9X_6 - 0.35X_7$$

Scenario B

$$f_1(X) = 0.8X_1 - 0.25X_2 + 0.6X_3 - 0.4X_4 - 0.8X_5 - 0.5X_6 + 0.7X_7 - 0.25X_2^2$$
$$f_2(X) = -0.4X_1 - 0.1X_2 + 0.45X_3 + 0.7X_4 + 0.2X_5 - 0.9X_6 - 0.35X_7 - 0.1X_2^2$$

Scenario C

$$f_1(X) = 0.8X_1 - 0.25X_2 + 0.6X_3 - 0.4X_4 - 0.8X_5 - 0.5X_6 + 0.7X_7 - 0.25X_2^2 - 0.4X_4^2 + 0.7X_7^2$$
$$f_2(X) = -0.4X_1 - 0.1X_2 + 0.45X_3 + 0.7X_4 + 0.2X_5 - 0.9X_6 - 0.35X_7 - 0.1X_2^2 + 0.7X_4^2 - 0.35X_7^2$$

Scenario D

$$f_1(X) = 0.8X_1 - 0.25X_2 + 0.6X_3 - 0.4X_4 - 0.8X_5 - 0.5X_6 + 0.7X_7 + 0.4X_1 \times X_3 - 0.175X_2 \times X_4 - 0.2X_4 \times X_5 - 0.4X_5 \times X_6$$
$$f_2(X) = -0.4X_1 - 0.1X_2 + 0.45X_3 + 0.7X_4 + 0.2X_5 - 0.9X_6 - 0.35X_7 - 0.2X_1 \times X_3 - 0.07X_2 \times X_4 + 0.35X_4 \times X_5 - 0.1X_5 \times X_6$$

Scenario E

$$f_1(X) = 0.8X_1 - 0.25X_2 + 0.6X_3 - 0.4X_4 - 0.8X_5 - 0.5X_6 + 0.7X_7 - 0.25X_2^2 + 0.4X_1 \times X_3 - 0.175X_2 \times X_4 - 0.2X_4 \times X_5 - 0.4X_5 \times X_6$$
$$f_2(X) = -0.4X_1 - 0.1X_2 + 0.45X_3 + 0.7X_4 + 0.2X_5 - 0.9X_6 - 0.35X_7 - 0.1X_2^2 - 0.2X_1 \times X_3 - 0.07X_2 \times X_4 + 0.35X_4 \times X_5 - 0.1X_5 \times X_6$$

Scenario F

$$f_1(X) = 0.8X_1 - 0.25X_2 + 0.6X_3 - 0.4X_4 - 0.8X_5 - 0.5X_6 + 0.7X_7 + 0.4X_1 \times X_3 + 0.4X_1 \times X_6 - 0.175X_2 \times X_4 + 0.175X_2 \times X_3 + 0.3X_3 \times X_4 + 0.3X_3 \times X_5 - 0.2X_4 \times X_5 - 0.28X_4 \times X_6 - 0.4X_5 \times X_6 - 0.4X_5 \times X_7$$
$$f_2(X) = -0.4X_1 - 0.1X_2 + 0.45X_3 + 0.7X_4 + 0.2X_5 - 0.9X_6 - 0.35X_7 - 0.2X_1 \times X_3 - 0.2X_1 \times X_6 - 0.07X_2 \times X_4 - 0.07X_2 \times X_3 + 0.225X_3 \times X_4 + 0.225X_3 \times X_5 + 0.49X_4 \times X_5 + 0.49X_4 \times X_6 + 0.1X_5 \times X_6 + 0.1X_5 \times X_7$$

Scenario G

$$f_1(X) = 0.8X_1 - 0.25X_2 + 0.6X_3 - 0.4X_4 - 0.8X_5 - 0.5X_6 + 0.7X_7 - 0.25X_2^2 - 0.4X_4^2 + 0.7X_7^2 + 0.4X_1 \times X_3 + 0.4X_1 \times X_6 - 0.175X_2 \times X_4 + 0.175X_2 \times X_3 + 0.3X_3 \times X_4 + 0.3X_3 \times X_5 - 0.2X_4 \times X_5 - 0.28X_4 \times X_6 - 0.4X_5 \times X_6 - 0.4X_5 \times X_7$$
$$f_2(X) = -0.4X_1 - 0.1X_2 + 0.45X_3 + 0.7X_4 + 0.2X_5 - 0.9X_6 - 0.35X_7 - 0.1X_2^2 + 0.7X_4^2 - 0.35X_7^2 - 0.2X_1 \times X_3 - 0.2X_1 \times X_6 - 0.07X_2 \times X_4 - 0.07X_2 \times X_3 + 0.225X_3 \times X_4 + 0.225X_3 \times X_5 + 0.49X_4 \times X_5 + 0.49X_4 \times X_6 + 0.1X_5 \times X_6 + 0.1X_5 \times X_7$$

5.2 Simulation comparison

We used R version 3.1.2 for the simulation and used the following packages/methods to analyze each dataset. MLR propensity scores were fitted using the *nnet* package (Venables and Ripley 2002). While we fit the MLR model, we include only main effects with all the available covariates ($X_1 - X_{10}$). RF propensity scores were fitted using the *randomForest* package (Liaw and Wiener 2002) with default setting. That is, we generate 500 trees (*ntree* = 500) and at each node, a best split will be selected from 3 randomly selected covariates. No trimming will be performed for the estimated propensity scores. GBM for propensity score estimation was fitted using the *gbm* package (Ridgeway 2012) with the shrinkage parameter set to 0.01 (*shrinkage* = 0.01), allowing for two-way interactions (*interaction.depth* = 2), and to fit a maximum of 3000 trees (*n.trees* = 3000). The optimal number of trees by GBM is found by optimizing the average standardized absolute mean difference (ASAM) (McCaffrey et al. 2013). The DAMS propensity scores were then calculated by the method described in Sect. 3.1 using the MLR propensity scores and the RF propensity scores. For reference, we also include an oracle estimator which builds on the true MLR model, although the true model is unknown in reality. IPTW for estimating ATEs was implemented as described in Section 4.1. Lastly, the

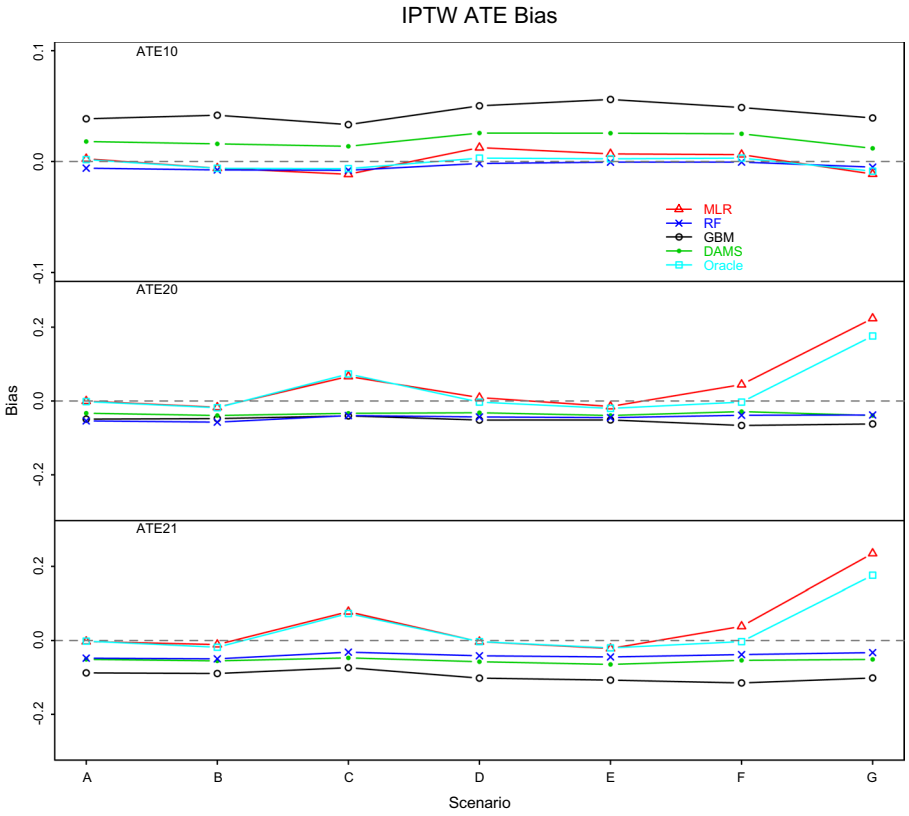

Fig. 1 Performance plots of bias for estimating ATEs

theoretical standard error estimates of the IPTW ATEs for each dataset are calculated using the *survey* package in R.

To evaluate the performance of each of the propensity score estimation methods for IPTW estimator, we consider several metrics. First, we consider the bias, which is the difference between the average of the ATE estimates from the 1000 datasets and the respective true ATE. i.e.,

$$Bias = \overline{\widehat{ATE}} - ATE$$

where $\overline{\widehat{ATE}} = \frac{1}{1000}\sum_{j=1}^{1000}\widehat{ATE}_j$.

We also consider the empirical standard error, which is the sample standard deviation of the 1000 ATE estimates. i.e.,

$$empirical\ SE = \sqrt{\frac{1}{999}\sum_{j=1}^{1000}(\widehat{ATE}_j - \overline{\widehat{ATE}})^2}.$$

Another measurement of the standard error we consider is the average standard error, which is the average of the 1000 theoretical standard errors of a particular ATE. i.e.,

$$averageSE = \frac{1}{1000}\sum_{j=1}^{1000}\widehat{SE}_j,$$

where $\widehat{SE}_j$ is the theoretical standard error from the jth dataset calculated by the sandwich standard error formula using the *survey* package.

Another metric we consider is the 95% confidence interval coverage. We calculate 95% confidence intervals based on theoretical standard error estimates:

$$95\%\ CI_j = (\widehat{ATE}_j - 1.96 \times \widehat{SE}_j, \widehat{ATE}_j + 1.96 \times \widehat{SE}_j).$$

The coverage is calculated as the percentage of the 95% confidence intervals that contains the true value of the ATE.

Lastly, we consider the root mean squared error (RMSE), which is the average of the squared difference between the ATE estimates and the true ATE. i.e.,

$$RMSE = \sqrt{\frac{1}{1000}\sum_{j=1}^{1000}(\widehat{ATE}_j - ATE)^2}.$$

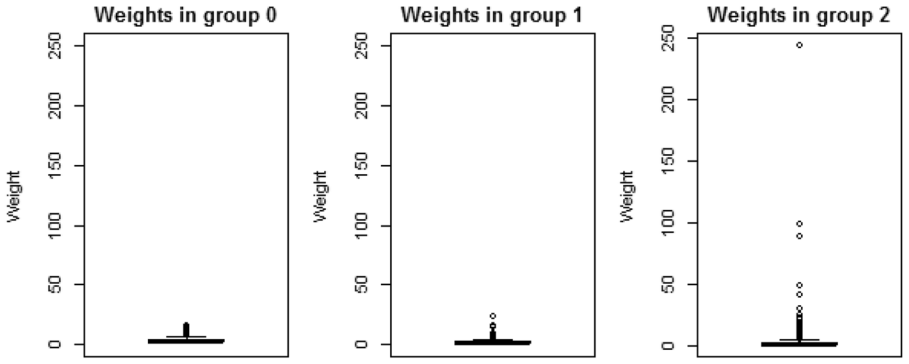

Fig. 2 Box plots of MLR weights in each group

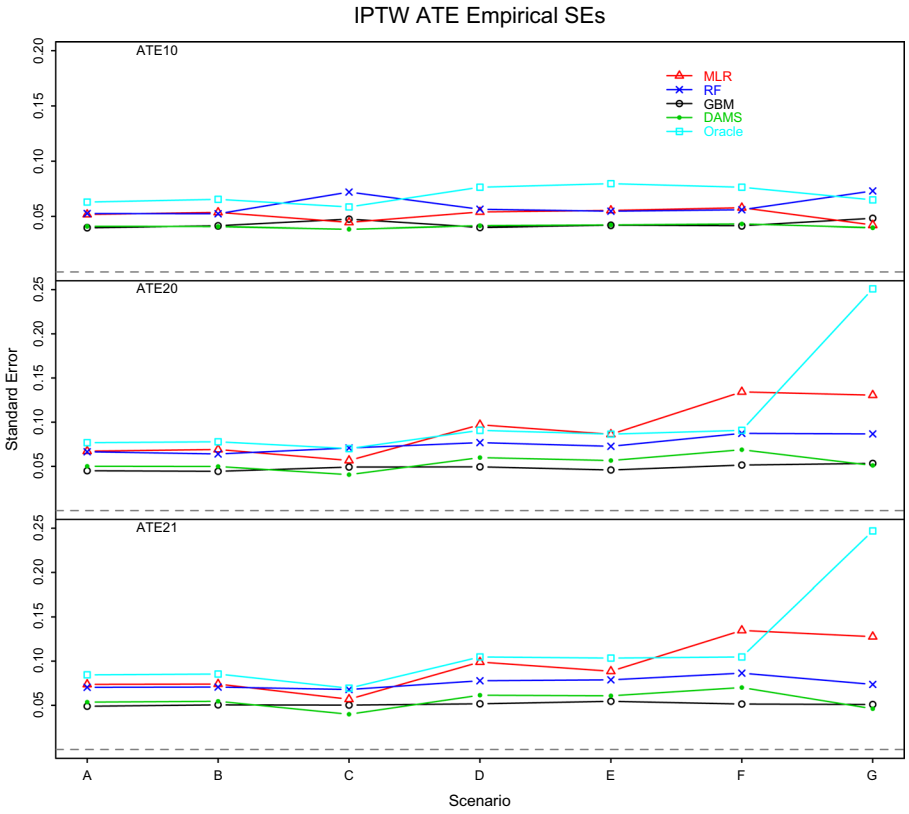

Fig. 3 Performance plots of empirical standard error for estimating ATEs

Table 1 Comparison of empirical standard error and average standard error of $\widehat{ATE}_{10,IPTW}$ in scenario A

Method	Empirical standard error	Average standard error
MLR	0.052	0.078
RF	0.053	0.076
GBM	0.040	0.068
DAMS	0.041	0.071

5.3 Simulation results for estimating ATEs

We present the results from the simulations for each of the seven scenarios. Plots for each metric are given for each of the three ATE estimates: $\widehat{ATE}_{10}$, $\widehat{ATE}_{20}$, and $\widehat{ATE}_{21}$.

Bias Looking at Fig. 1, estimating generalized propensity scores by GBM seems to result in the most biased $\widehat{ATE}_{10,IPTW}$ in all seven scenarios, and this method seems to consistently overestimate the ATE. On the other hand, MLR and RF seem to result in lower bias in $\widehat{ATE}_{10,IPTW}$. The bias of $\widehat{ATE}_{10,IPTW}$ seems to be small and similar across different propensity score estimation methods.

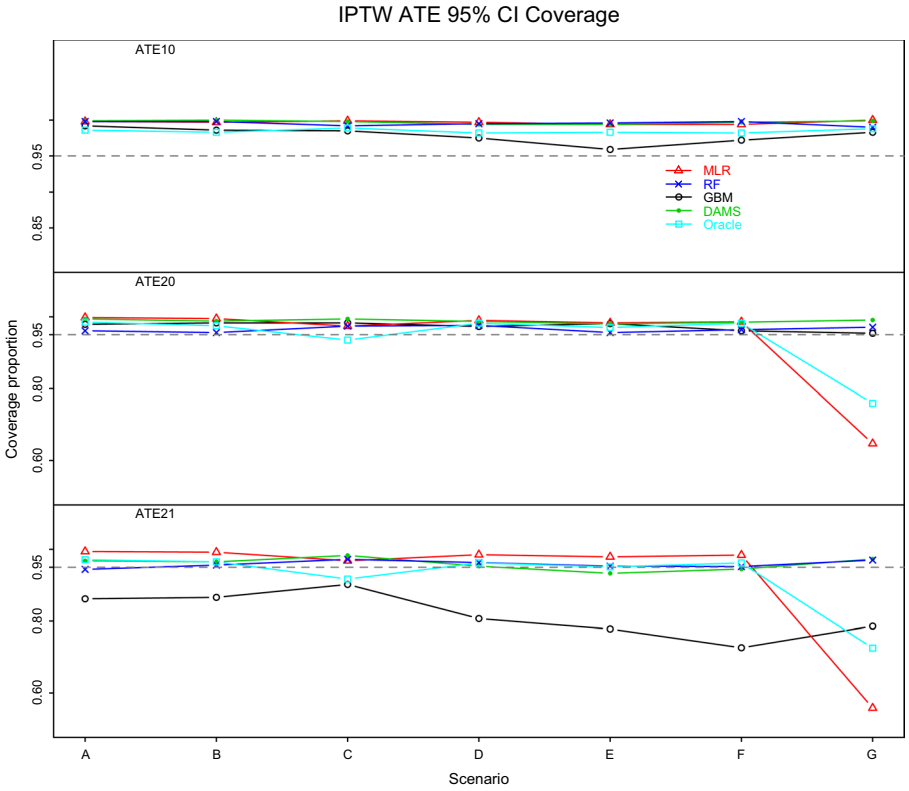

Fig. 4 Performance plots of 95% confidence interval coverage for estimating ATEs

The bias of $\widehat{ATE}_{20,IPTW}$ and $\widehat{ATE}_{21,IPTW}$ seem to be similar for each propensity score estimation method across all scenarios. In contrast to $\widehat{ATE}_{10,IPTW}$, GBM now results in consistently negatively biased estimates. In scenarios C and G where there is moderate non-linearity, MLR seems to result in a higher bias whereas the bias of the other three methods remain consistent across the different scenarios. In fact, all methods except for MLR tend to under-estimate the ATE. Additionally, DAMS and RF perform very similarly and tend to result in low bias.

Note that MLR results in lower bias for $\widehat{ATE}_{10,IPTW}$ in scenario G compared to the other two ATEs. If we look at the distribution of MLR weights in scenario G for each of the treatment groups in a random dataset, there is a subject with an extreme weight of around 250 in group 2, whereas the weights in groups 0 and 1 do not exceed 50 (Fig. 2). Since extreme weights in IPTW can lead to large bias and variance of the estimated causal effects, this is likely the reason for the higher bias in $\widehat{ATE}_{20}$ and $\widehat{ATE}_{21}$ as these two ATEs involve treatment group 2.

Since there is considerable non-linearity and non-additivity in scenario G, the simple MLR model with linear terms only did not do an adequate job in estimating the propensity scores, hence resulting in extreme weights. On the other hand, the machine learning methods were able to better account for complexity in the propensity score model.

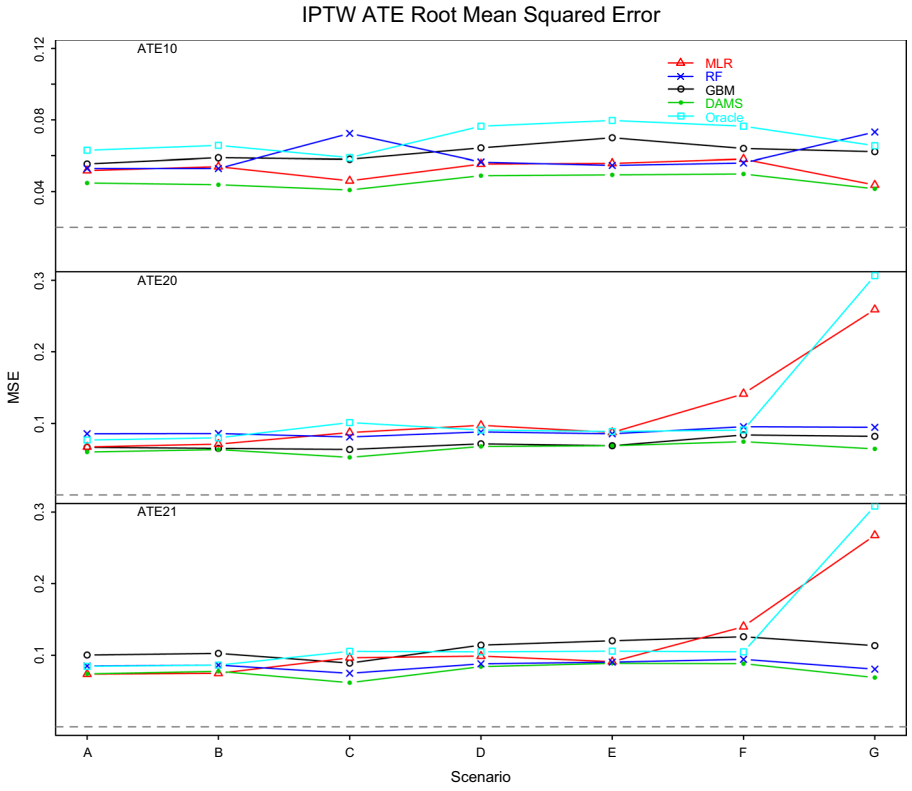

Fig. 5 Performance plots of root mean squared error for estimating ATEs

Empirical standard error From Fig. 3, the empirical SEs for $\widehat{ATE}_{10,IPTW}$ are similar between all methods and generally are constant across all scenarios. In scenarios C and G with non-linearity, RF seems to yield slightly higher empirical SEs.

As was the case for bias, the empirical SE performance patterns for $\widehat{ATE}_{20,IPTW}$ and $\widehat{ATE}_{21,IPTW}$ are very similar. As the complexity of the scenarios increases, MLR yields higher empirical SEs compared to the other three methods. GBM and DAMS tend to result in similar levels of empirical SE that are lower compared to the other two methods.

95% confidence interval coverage In Fig. 4 we see that the 95% confidence interval coverage for $\widehat{ATE}_{10,IPTW}$ is consistently above 95% for all four methods. This is because the theoretical standard errors were overestimated, so the average SE was higher than the empirical SE, as shown in Table 1. The reason is that the sandwich formula does not account for the variability in estimating the propensity scores, which leads to a smaller number of estimating equations while constructing the variance estimator and therefore enlarge the variance. This is theoretically proved by Lunceford and Davidian (2004) and widely noted in some other simulation studies, such as Lee et al. (2010) and Xie et al. (2017).

The 95% confidence interval coverage for $\widehat{ATE}_{20,IPTW}$ is similarly high for GBM, DAMS, and RF, but for MLR the coverage is very low in scenario G with moderate non-additivity and non-linearity. This is because MLR yields a much higher bias in scenario G.

The 95% confidence interval coverage for $\widehat{ATE}_{21,IPTW}$ for MLR, DAMS, and RF follow the same pattern as for $\widehat{ATE}_{20,IPTW}$. However, for GBM, we see that the 95% confidence interval coverage is consistently below 95% due to the relatively large bias in estimating ATE_{21}. Overall, DAMS and RF yield the most consistent 95% confidence interval coverage above 95% for all three ATEs.

Root mean squared error Looking at Fig. 5, the RMSE for $\widehat{ATE}_{10,IPTW}$ is consistently low for all four methods and all seven scenarios. For $\widehat{ATE}_{20,IPTW}$ and $\widehat{ATE}_{21,IPTW}$, GBM, DAMS, and RF yield similarly low MSEs in all scenarios, but the MSE for MLR is higher in scenario G. This is likely related to the high bias of $\widehat{ATE}_{20,IPTW}$ and $\widehat{ATE}_{21,IPTW}$ in scenario G when using MLR.

Overall, DAMS and RF result in the lowest bias in all scenarios and for each ATE. GBM and DAMS result in the lowest empirical SEs. DAMS and RF result in the most consistent 95% confidence interval coverage above 95%. Lastly, all methods except for MLR perform similarly well in terms of having lower and consistent MSEs across all seven scenarios.

Interestingly, the oracle model, which builds on the true propensity score model, does not provide the best estimation of the causal estimator in most of the scenarios. Especially, the oracle model yields the largest standard errors compared to the rest of the methods in almost all the scenarios. The reason is in the oracle model, we only include the covariates related to the treatment (X_1–X_7). We ignore the covariates that are only related to the outcome (X_8-X_{10}), which results in the larger standard error of the oracle model. This phenomenon is also noted in Brookhart and Laan (2006) and Zhu et al. (2015).

5.4 Simulation results for estimating ATEs when sample size is small

The simulation results with a sample size of 200 are displayed in Figures 1–4 in the Electronic Supplementary Material. The relative performance of different propensity score methods are similar to the results with $n = 1000$. The small bias shown in Fig. 1 of the Electronic Supplementary Material indicates that all four methods generally perform well with a small sample size.

5.5 Simulation results for estimating ATTs

We also examine the performance of IPTW for estimating ATT using different propensity score estimation methods. Figures 5, Figure 6 and Figure 7 in the Electronic Supplementary Material show the bias, standard error and the root mean squared error of the estimated treatment effects, respectively. The sample size is 1000. Again, the nonparametric algorithms, RF, DAMS and GBM lead to less biased estimates compared to MLR, when the MLR model is misspecified. In addition, MLR and RF leads to larger variance due to the fact that they are more likely to produce extreme weights (plots not shown). Overall, DAMS and GBM perform the best.

6 Data application

For an illustration, we apply the proposed methodology to the Taobao dataset collected from Taobao.com, China's largest e-commerce platform. Taobao commands 96.5% of the market share of C2C e-commerce in China. Consumers can buy almost everything they

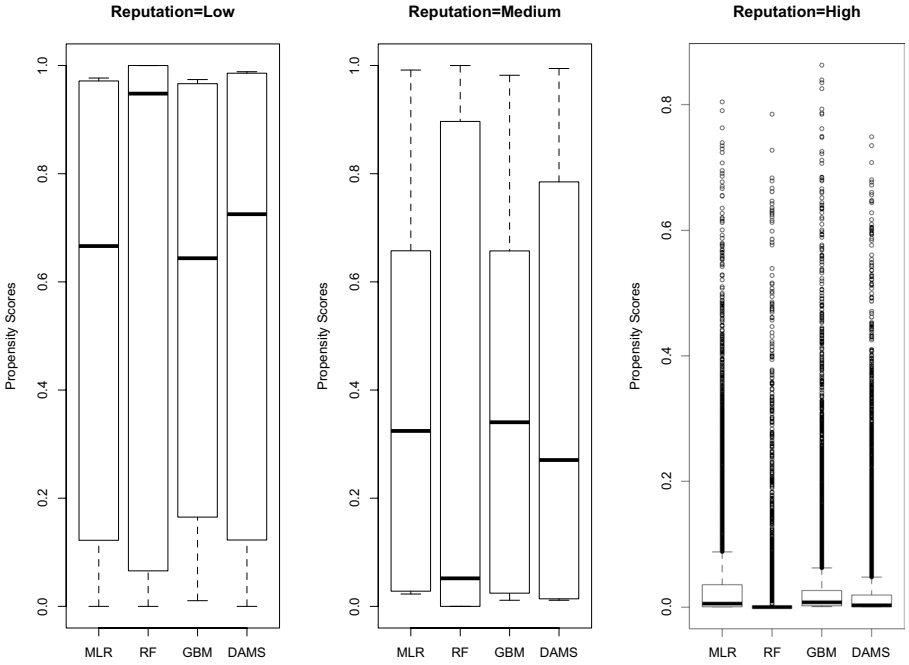

Fig. 6 The boxplot of generalized propensity scores estimated by each method

Table 2 Seller's reputation and average sales

Score	Grade	Reputation	Number of sellers	Mean sales
<=250	0–5	Low (0)	5587	2.64×10^4
251–10,000	6–10	Medium (1)	3881	1.47×10^5
>= 10,001	11–20	High (2)	532	4.75×10^5

Table 3 Causal effect estimates of reputation on sales

GPS method	MLR	RF	GBM	DAMS
$\widehat{ATE}_{10,IPTW}$ ("medium" vs. "low")	9.56×10^4	3.76×10^4	8.58×10^4	8.68×10^4
$\widehat{ATE}_{20,IPTW}$ ("high" vs. "low")	3.95×10^5	2.90×10^5	3.41×10^5	3.93×10^5
$\widehat{ATE}_{21,IPTW}$ ("high" vs. "medium")	3.00×10^5	2.52×10^5	2.55×10^5	3.06×10^5

need online, often times at prices much lower than what traditional retailers offer. Despite its success, Taobao has been criticized for contributing to trade in counterfeit goods. Lack of trust has been one of the most formidable barriers to people for engaging in e-commerce. Unlike in the offline world, buyers do not have the opportunity to verify the legitimacy of the product. Taobao introduced the "Alipay" online payment service in 2004 and offers a feedback system to reduce the likelihood of fraud and encourage trust based upon reputation.

In this analysis, we are interested in examining the causal effect of a seller's rating (reputation) on sales. The data are at the seller level in the electronic sector aggregated from millions of transactions on Taobao.com from June 2011 to December 2011. The dataset we get is a simple random sample from the original database with a sample size of ten thousand. In this analysis, the outcome variable is *sales*, which is the gross revenue (unit: CN yuan; 1 CN yuan≈ 0.16 US dollar) for each seller from June 2011 to December 2011. The treatment we are interested in is *reputation*: the rating of the seller. There are 13 potential confounders related to seller's characteristics: seller's age, seller's gender, whether the seller provides free return within 7 days of purchase, whether the seller provides fast shipping, whether the seller offers bundle discount, whether the seller offers "second kill" (a promotion program), whether the seller participates in Taobao's ad auction, whether the seller has a premium store, whether the seller offers the social referral promotion, whether the seller allows payment upon delivery, whether the seller agrees to pay 10 times more if the product is found to be fake, whether the seller provides 30 day free repair and whether the products provided by the seller are exactly as described on the website.

Regarding the treatment variable, the website adopts a similar reputation rating system as on eBay. For each transaction, the buyer can choose to rate the seller by leaving feedback. The seller earns +1 point for each positive rating, no points for each neutral rating, and -1 point for each negative rating. If the buyer does not submit feedback in 15 days upon completion of the transaction, the seller obtains a positive rating. The cumulative rating score is then categorized into twenty grades from 0 to 20 (Fan et al. 2016). The variable

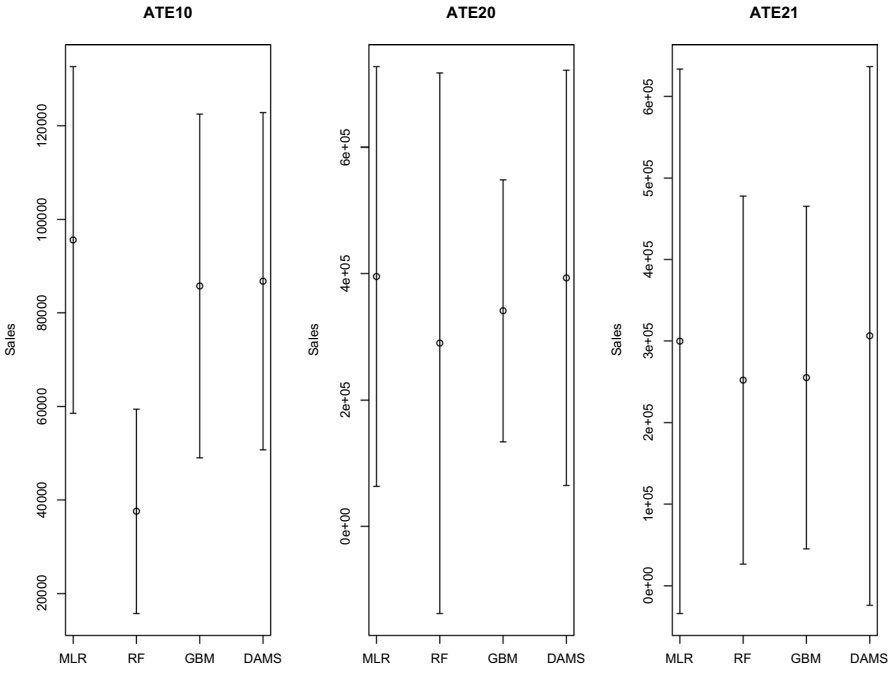

Fig. 7 The 95% confidence interval of the causal effects by each method. ATE_{10} is the average treatment effect of "medium" versus "low" reputation; ATE_{20} is the average treatment effect of "high" versus "low" reputation; and ATE_{21} is the average treatment effect of "high" versus "medium" reputation

reputation represents the grade of each seller at the end of December 2011. Our data shows that 94.68% of all the sellers have a reputation grade between 0 and 10. Taobao.com divide reputation grade into 4 categories and each is characterized by a system of symbols: 0-5 (hearts), 6-10 (diamonds), 11-15 (crowns), 16-20 (golden crowns). Sellers symbolized by hearts and diamonds are usually new online startups and small business sellers. On the other hand, sellers symbolized by crowns and golden crowns are usually established sellers and run by companies. Based on this, we redefine the reputation variable as "low (0)" with a reputation grade between 0 and 5, "medium (1)" with a reputation grade between 6 and 10, and "high (2)" with a reputation grade between 11 and 20. Table 2 shows the rating score, the corresponding grade and the reputation level. It also reports the frequency of each reputation level in our dataset. The last column shows the average value of sales for each reputation level. Overall, there is an increasing trend in the sales when reputation increases.

To estimate the generalized propensity score, we employ MLR, RF, GBM and DAMS with the same parameter setting as described in the simulation study. Figure 6 shows the boxplot of the estimated generalized propensity scores for three different treatment groups. From the figure, we may doubt that RF is not very reliable in this analysis since its estimates are quite different from the other three methods. We observe that the propensity of having a "high" reputation is noticeably lower than the other two groups, which is reasonable.

We then apply IPTW to estimate ATE. The results based on different propensity score models are displayed in Table 3. Again, RF seems to work quite differently and the rest of the three methods yield similar results. For example, we can conclude that sellers with "medium" reputation has an increase of around 9×10^4 CN Yuan in half-year sales, compared to "low" reputation.

We also use the sandwich formula to obtain the standard error and the 95% confidence interval for each causal effect. The results are displayed in Fig. 7. Again, the confidence interval based on different algorithms are quite similar with an exception of RF. Based on DAMS, at $\alpha = 0.05$, the casual effect of "medium" reputation versus "low" reputation as well as "high" reputation versus "low" reputation are significantly larger than 0, while the causal effect of "high" reputation versus "medium" reputation is not significantly different from 0.

7 Discussion

From our simulation studies, we found that using MLR to estimate the generalized propensity scores can result in extreme weights that in turn result in more bias, higher empirical standard errors, poor 95% confidence interval coverage, and higher mean squared errors when the relationship between the treatment assignment and the covariates is non-linear and non-additive. On the other hand, GBM, DAMS, and RF tend to be more stable across different levels of complexity in the relationship between the treatment assignment and the covariates. Considering all metrics listed in Sect. 5.2, DAMS performed the best out of the four propensity score estimation methods in combination with IPTW. In conclusion, we recommend machine learning methods for propensity score estimation in multi-level treatment settings.

In real applications, the true propensity score model is never known for observational studies. Researchers may apply different algorithms and see how different the results are. If

one algorithm produces results that are quite different from the rest, it is probably not very reliable. Boxplot of the propensity score estimates or inverse weights should be checked to detect outliers. Balance in the covariates should be evaluated before and after the propensity score based adjustment. In general, we believe a model-averaging estimator, such as DAMS or super learner (Laan et al. 2007) is usually superior compared to a single model.

Funding This study was funded by Social Sciences and Humanities Research Council Insight Development Grant (Grant number: 430-2016-00163) and by National Sciences and Engineering Research Council of Canada (Grant number: RGPIN-2017-04064)

Compliance with ethical standards

Conflict of interest All authors declares that they have no conflict of interest.

Ethical approval This article does not contain any studies with human participants or animals performed by any of the authors.

References

Breiman, L.: Random forests. Mach. Learn. **45**(1), 5–32 (2001)
Brookhart, M.A., van der Laan, M.J.: A semiparametric model selection criterion with applications to the marginal structural model. Comput. Stat. Data Anal. **50**(2), 475–498 (2006)
Crump, R.K., Hotz, V.J., Imbens, G.W., Mitnik, O.A.: Dealing with limited overlap in estimation of average treatment effects. Biometrika **96**(1), 187–199 (2009)
Fan, Y., Ju, J., Xiao, M.: Reputation premium and reputation management: evidence from the largest e-commerce platform in china. Int. J. Ind. Organ. **46**, 63–76 (2016)
Feng, P., Zhou, X.H., Zou, Q.M., Fan, M.Y., Li, X.S.: Generalized propensity score for estimating the average treatment effect of multiple treatments. Stat. Med. **31**(7), 681–697 (2012)
Frölich, M.: Programme evaluation with multiple treatments. J. Econ. Surv. **18**(2), 181–224 (2004)
Imai, K., Van Dyk, D.: Causal inference with general treatment regimes. J. Am. Stat. Assoc. **99**(467), 854–866 (2004)
Imbens, G.W.: The role of the propensity score in estimating dose-response functions. Biometrika **87**(3), 706–710 (2000)
Lechner, M.: Identification and estimation of causal effects of multiple treatments under the conditional independence assumption. In: Econometric Evaluation of Labour Market Policies pp. 43–58 (2001)
Lechner, M.: Program heterogeneity and propensity score matching: an application to the evaluation of active labor market policies. Rev. Econ. Stat. **84**(2), 205–220 (2002)
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)
Liaw, A., Wiener, M.: Classification and regression by randomforest. R News **2**(3), 18–22 (2002). http://CRAN.R-project.org/doc/Rnews/
Lunceford, J.K., Davidian, M.: Stratification and weighting via the propensity score in estimation of causal treatment effects: a comparative study. Stat. Med. **23**(19), 2937–2960 (2004)
Luo, W., Zhu, Y., Ghosh, D.: On estimating regression-based causal effects using sufficient dimension reduction. Biometrika **104**(1), 51–65 (2017)
McCaffrey, D.F., Griffin, B.A., Almirall, D., Slaughter, M.E., Ramchand, R., Burgette, L.F.: A tutorial on propensity score estimation for multiple treatments using generalized boosted models. Stat. Med. **32**(19), 3388–3414 (2013)
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), 403–425 (2004)
McFadden, D.: Conditional logit analysis of qualitative choice behavior. In: Zarembka, P. (ed.) Frontiers in Econometrics, pp. 105–142. Academic Press, New York (1973)

Ridgeway, G.: gbm: Generalized Boosted Regression Models (2012). R package version 1.6-3.2. http://CRAN.R-project.org/package=gbm. Accessed 8 Nov 2016

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

Rubin, D.B.: Estimating causal effects of treatments in randomized and nonrandomized studies. J. Educ. Psychol. **66**(5), 688–701 (1974)

Spreeuwenberg, M.D., Bartak, A., Croon, M.A., Hagenaars, J.A., Busschbach, J.J., Andrea, H., Twisk, J., Stijnen, T.: The multiple propensity score as control for bias in the comparison of more than two treatment arms: an introduction from a case study in mental health. Med. Care **48**(2), 166–174 (2010)

Tchernis, R., Horvitz-Lennon, M., Normand, S.L.T.: On the use of discrete choice models for causal inference. Stat. Med. **24**(14), 2197–2212 (2005)

Tu, C., Koh, W.Y., Jiao, S.: Using generalized doubly robust estimator to estimate average treatment effects of multiple treatments in observational studies. J. Stat. Comput. Simul. **83**(8), 1518–1526 (2013)

Uysal, S.D.: Doubly robust estimation of causal effects with multivalued treatments: an application to the returns to schooling. J. Appl. Econom. **30**(5), 763–786 (2014)

van der Laan, M.J., Polley, E.C., Hubbard, A.E.: Super learner. Stat. Appl. Genet. Mol. Biol. **6**(1), 1–21 (2007)

Venables, W.N., Ripley, B.D.: Modern Applied Statistics with S, 4th edn. Springer, New York (2002)

Xie, Y., Zhu, Y., Cotton, C.A., Wu, P.: A model averaging approach for estimating propensity scores by optimizing balance. Stat. Methods Med. Res. (2017). https://doi.org/10.1177/0962280217715487

Zhu, Y., Ghosh, D., Mitra, N., Mukherjee, B.: A data-adaptive strategy for inverse weighted estimation of causal effects. Health Serv. Outcomes Res. Methodol. **14**(3), 69–91 (2014)

Zhu, Y., Lin, L.L.: Propensity score modeling and evaluation. In: Statistical Causal Inferences and Their Applications in Public Health Research, pp. 111–124. Springer (2016)

Zhu, Y., Schonbach, M., Coffman, D.L., Williams, J.S.: Variable selection for propensity score estimation via balancing covariates. Epidemiology **26**(2), e14–e15 (2015)