International Journal of Statistics and Probability; Vol. 10, No. 1; January 2021
ISSN 1927-7032 E-ISSN 1927-7040
Published by Canadian Center of Science and Education

Causal Subclassification Tree Algorithm and Robust Causal Effect Estimation via Subclassification

Tomoshige Nakamura[1], Mihoko Minami[2]

[1] School of Science and Technology, Keio University, Kanagawa, Japan

[2] Department of Mathematics, Keio University, Kanagawa, Japan

Correspondence: Tomoshige Nakamura, School of Science and Technology, Keio University, 3-14-1 Hiyoshi, Kohoku-ku, Yokohama, Kanagawa, Japan. E-mail: tomoshige.nakamura@gmail.com

Received: September 29, 2020 Accepted: November 9, 2020 Online Published: November 30, 2020

doi:10.5539/ijsp.v10n1p40 URL: https://doi.org/10.5539/ijsp.v10n1p40

Abstract

In observational studies, the existence of confounding variables should be attended to, and propensity score weighting methods are often used to eliminate their effects. Although many causal estimators have been proposed based on propensity scores, these estimators generally assume that the propensity scores are properly estimated. However, researchers have found that even a slight misspecification of the propensity score model can result in a bias of estimated treatment effects. Model misspecification problems may occur in practice, and hence, using a robust estimator for causal effect is recommended. One such estimator is a subclassification estimator. Wang, Zhang, Richardson, & Zhou (2020) presented the conditions necessary for subclassification estimators to have $\sqrt{N}$-consistency and to be asymptotically well-defined and suggested an idea how to construct subclasses.

In this paper, we propose a subclass construction algorithm, a causal subclassification tree, that satisfies the conditions of Wang et al. (2020). We also explain the connection between our algorithm and Wang et al. (2020)'s suggested method and show that their method does not satisfy their conditions. We performed simulation experiments in a similar setting as in Kang and Schafer (2007). The simulation results show that the causal subclassification algorithm improves the performance of weighting estimators for causal effects, especially when the propensity model and/or outcome model is misspecified. It should also be noted that when machine learning methods such as random forest and XgBoost are used to estimate propensity scores, the doubly robust estimator with the causal subclassification tree shows the best performance when both models are misspecified. We also compared our method to Wang et al. (2020)'s suggested method through the simulation and found that the causal subclassification tree has almost the same bias but a smaller variance compared to Wang et al. (2020)'s suggested method.

Keywords: causal inference, subclassification estimator, decision tree, robust inference

1. Introduction

One of the main purposes of observational studies is to estimate the causal effect of an intervention or a treatment on the outcome. The ideal setting to estimate causal effect is a randomized control trial (RCT). However, in practice, treatments or interventions cannot be assigned randomly. Typical examples are advertisement placement in marketing and smoking status in epidemiology. There are various reasons that make random assignment difficult, ranging from practical to ethical. Particularly in observational studies, treatments or interventions often depend on covariates because controlling confounding variables is difficult. In cases where the outcome is also affected by covariates, if we do not pay enough attention to assignment mechanisms, the estimated effect of the treatment may be biased.

The most commonly used technique for removing the effects of confoundings is the propensity score, proposed by Rosenbaum and Rubin (1983). Causal effect estimation methods using the propensity score include matching (e.g., Rosenbaum & Rubin, 1985; Rosenbaum, 1989; Abadie & Imbens, 2006), subclassification (e.g., Rosenbaum & Rubin,1984; Rosenbaum, 1991; Hansen, 2004), weighting (e.g., Rosenbaum, 1987; Robins, Hernán, & Brumback, 2000; Hirano, Imbens, & Ridder, 2003), and regression (e.g., Heckman, Ichimura, & Todd, 1997;). In recent years, there has been growing interest in individual causal effects. BART(Hill, 2011), Causal Forest (Wager & Athey, 2018), Generalized Random Forest (Athey, Tibshirani, & Wager, 2019) and R-Learner (Nie & Wager, 2020) have been proposed as methods for estimating heterogeneous causal effects.

When we estimate causal effects using the propensity score model, we need to be careful to avoid model misspecification of the propensity score. For example, Kang and Schafer (2007) pointed out the instability of the IPW estimator (Hirano et al., 2003) and the Doubly Robust estimator (Robins, Rotnitzky, & Zhao, 1994) caused by misspecification of the propensity score model. In practice, it may be difficult to get an exact specification of the true propensity score model; however, we need to consider this aspect when using the model.

One popular approach to address the model misspecification problem is subclassification, which involves grouping units into K subclasses based on their estimated propensity scores. For example, Rosenbaum and Rubin (1984) recommended using $K = 5$ as the number of subclasses. Unlike the IPW estimator, the ordinal subclassification estimator (cf. Rosenbaum & Rubin, 1983) does not depend on the estimated propensity score itself, but uses the rank information of the estimated propensity score. Therefore, the subclassification estimator is robust with respect to propensity score outliers or slight model misspecification for propensity score.

However, to use a subclassification estimator in practice, it is necessary to determine a method for subclass construction (i.e., we need to choose the lower and upper bounds of the propensity score in each subclass). A commonly used approach to determine the propensity score cut-off points of subclasses is to use quantiles of estimated propensity scores with a fixed number of subclasses, 5 to 10. A problem with this approach is that the length of each subclass does not converge to 0 even if the sample size N goes to infinity, and hence, the subclassification estimator is asymptotically biased. To avoid this problem, Wang et al. (2020) proposed conditions between sample size and number of subclasses at which the subclassification estimator becomes $\sqrt{N}$-consistent and asymptotically well defined. They also proposed a theoretical guideline to construct subclasses and suggested constructing subclasses such that each subclass contains at least one observation from treatment and control groups. However there are two concerns. One is about the variance of the subclassification estimator obtained by their suggested method. When the number of samples in each subclass becomes small, then the variance of causal estimators in each subclass may become large, so the subclassification estimator may have large variance. The other concern is whether the suggested method satisfies their theoretical guideline.

In this paper, we propose a new subclass construction algorithm that satisfies Wang et al. (2020)'s conditions for a subclassification estimator to become $\sqrt{N}$-consistent and asymptotically well-defined. Additionally, we show that our algorithm includes Wang et al. (2020)'s suggested method as a special case, and prove that their method does not satisfy their conditions that the subclassification estimator becomes asymptotically well-defined.

We call this algorithm the causal subclassification tree. In our algorithm, we use the decision tree by Breiman, Friedman, Stone, and Olshen (1984) with the sum of variance of propensity scores or sum of variance of potential outcomes in child nodes as the impurity measure. To achieve Wang et al. (2020)'s conditions, we impose constraints on splitting rules controlled by two parameters: on each split leaves at least a fraction α of the available training examples on each side of the split and, the trees fully grown to depth ℓ for some $\ell \in \mathbf{N}$, i.e., between ℓ and $2\ell - 1$ observations of treatment and control groups in each terminal node of the tree. Next, terminal nodes are used as strata for subclassification. In section 3, we show that the number of leaves generated by our algorithms with $\ell = O(\sqrt{N}\log(N))$ satisfies the conditions by Wang et al. (2020). We also proposed a method for selecting parameters (ℓ, α) and extended our subclassification algorithm to multi-class treatment.

Some researchers claim that when propensity scores are estimated by machine learning (ML) methods, such as random forest, boosting, or neural networks, the resulting propensity score weighting estimator has a larger bias or RMSE in some cases than when a logistic regression model is used for propensity score estimation (i.e., Cannas & Arpino, 2019). We also confirm this claim, as we show in section 4. However since combining a causal subclassification tree algorithm and weighting estimator with the ML method shows a relatively small bias and variance compared to the original weighting estimator with ML methods, when the propensity score model is misspecified, the weighting estimator with the ML method combined with a causal subclassification tree shows a better performance than a weighting estimator with a logistic regression model and CBPS.

This paper is organized as follows. The rest of Section 1 describes the symbols and assumptions used in this paper. Section 2 introduces a general subclassification estimator for causal effect and the results of Wang et al. (2020) on $\sqrt{N}$-consistency of the subclassification estimator. In section 3, we propose the subclassification algorithm and show its asymptotic properties. We then extend our algorithm to the case of multi-class treatment and estimation for weighted average treatment effects. In section 4, we confirm that the weighting estimators for causal effects become robust by applying our methods through simulation using Kang and Schafer (2007)'s settings. We end with a discussion of the findings in section 5.

1.1 Notation for Binary Treatment

Let us consider a random sample of N observations from a population $\mathcal{P}$. For each unit i, we observe a binary treatment variable $T_i \in \{0, 1\}$, a p-dimensional column vector of observed pre-treatment covariates X_i, whose support is denoted by $\mathcal{X}$, and outcome $Y_i \in \mathbf{R}$. The propensity score is defined as the conditional probability of receiving the treatment given the covariates X_i.

Following Rosenbaum and Rubin (1983), we assume that the true propensity score is bounded away from 0 and 1:

$$0 < \Pr[T_i = 1|X_i = x] < 1, \qquad \text{for all} \;\; x \in \mathcal{X} \tag{1}$$

This assumption means that all individuals in the data have a non-zero probability of receiving or not receiving treatment.

To estimate causal effects from the data, additional assumptions must be made to the data generation process. We assume the unconfoundedness assumption,

$$\{Y_i(1), Y_i(0)\} \perp\!\!\!\perp T_i \big| X_i = x$$

where $Y_i(t)$ are potential outcomes corresponding to treatment assignment $t \in \{0, 1\}$.

Under the unconfoundedness assumption, Rosenbaum and Rubin (1983) showed that treatment assignment is ignorable given the (true) propensity score $\pi(X_i)$.

$$\{Y_i(1), Y_i(0)\} \perp\!\!\!\perp T_i \big| \pi(X_i)$$

From this result, since expectation of potential outcomes is $\mathbb{E}[Y_i(t)|\pi(X_i), T_i = t] = \mathbb{E}[Y_i(t)|\pi(X_i)]$, the average treatment effect can be written as the difference between the expected outcomes of the treatment group and the control group conditions on propensity score,

$$\begin{aligned}\mathbb{E}[Y_i(1) - Y_i(0)] &= \mathbb{E}[\mathbb{E}[Y_i(1)|T_i = 1, \pi(X_i)] - \mathbb{E}[Y_i(0)|T_i = 0, \pi(X_i)]] \\ &= \mathbb{E}[\mathbb{E}[Y_i|T_i = 1, \pi(X_i)] - \mathbb{E}[Y_i|T_i = 0, \pi(X_i)]]\end{aligned}$$

In fact, the propensity score is a continuous quantity on $(0, 1)$, and there is no guarantee that there will be samples with the same propensity score. In general, the following three estimators are often used. The first is a matching method that measures the closeness of the propensity score between samples using an appropriate distance function or kernel function, and compares those with close distances. The second is an Inverse Probability Weighting estimator (IPW estimator, Hirano et al., 2003) that weights samples by the reciprocal of the propensity score. The last is a stratification method that divides the data into subclasses of 5 to 10 based on the propensity score and then compares them by subclass.

2. Subclassification Estimator

Subclassification estimators are constructed in two steps. First, the observed data are divided into subclasses based on propensity scores, and then the averages of outcomes of treatment and control groups belonging to the same subclass are compared (Rosenbaum & Rubin, 1984).

Let $\hat{\pi}_{\min}$ and $\hat{\pi}_{\max}$ be the minimum and the maximum of estimated propensity scores, respectively, and $\hat{q}_k, k = 0, 1, \ldots, K$ be the values such that $\hat{\pi}_{\min} = \hat{q}_0 < \hat{q}_1 < \cdots < \hat{q}_{K-1} < \hat{q}_K = \hat{\pi}_{\max}$ where K is the number of subclasses. The set of $\hat{C}_k = [\hat{q}_{k-1}, \hat{q}_k)$, $(k = 1, 2, ..., K - 1)$ and $\hat{C}_K = [\hat{q}_{K-1}, \hat{q}_K]$ forms a partition of the interval $[\hat{\pi}_{\min}, \hat{\pi}_{\max}]$. Let N_k be the number of units in subclass k and N_{tk} be the number of subjects in treatment group t in subclass k for $t = 0, 1$ and $k = 1, \cdots, K$. That is, $N_k = \sum_{i=1}^{N} I(\hat{\pi}(X_i) \in \hat{C}_k)$, and $N_{tk} = \sum_{i=1}^{N} I(\hat{\pi}(X_i) \in \hat{C}_k) I(T_i = t)$. Then the subclassification estimator for average treatment effect $\hat{\tau}_S$ is given by

$$\hat{\tau}_S = \sum_{k=1}^{K} \frac{N_k}{N} \left\{ \frac{1}{N_{1k}} \sum_{i=1}^{N} T_i Y_i I(\hat{\pi}_i \in \hat{C}_k) - \frac{1}{N_{0k}} \sum_{i=1}^{N} (1 - T_i) Y_i I(\hat{\pi}_i \in \hat{C}_k) \right\}.$$

For subclassification estimators, setting $K = 5$ is recommended in Rosenbaum and Rubin (1984), and the quantiles of estimated propensity score are used for K cut-off points. However, as is well known, if K is fixed, $\hat{\tau}_S$ may become a biased estimator for τ (e.g., Lunceford & Davidian, 2004).

Wang et al. (2020) proposed increasing K according to the sample size N to avoid bias that may occur when K is fixed. That is, the number of subclasses $K = K(N)$ is set to satisfy the following properties:

$$K(N) \to \infty (N \to \infty) \qquad \text{and} \qquad K(N)/\sqrt{N} \to \infty (N \to \infty) \tag{2}$$

They also proposed the Hybrid Estimator $\hat{\tau}_H$,

$$\hat{\tau}_H = \sum_{k=1}^{K(N)} \frac{N_k}{N} \left\{ \frac{1}{N_{1k}} \sum_{i=1}^{N} Z_i Y_i I(\hat{\pi}_i \in \hat{C}_k) - \frac{1}{N_{0k}} \sum_{i=1}^{N} (1 - Z_i) Y_i I(\hat{\pi}_i \in \hat{C}_k) \right\} \tag{3}$$

and showed that $\hat{\tau}_H$ is $\sqrt{N}$-consistent estimator for τ under conditions (2). Moreover, they showed that if the condition

$$(K(N))\log(K(N))/N \to 0. \tag{4}$$

is satisfied, then $\hat{\tau}_H$ becomes asymptotically well defined :

$$\Pr(N_{tk} > 0) \to 1 \quad \text{for all} \quad t, k.$$

So, to $\hat{\tau}_H$ be asymptotically well-defined, the condition is that the number of subclasses should grow slowly enough so that all subclasses contain at least one observations from treatment and control group.

In Wang et al. (2020), they proposed using the maximal number of subclasses such that the hybrid estimator is well defined:

$$K_{\max} = \max\{K : \hat{\tau}_H \text{ is well defined}\},$$

and called the hybrid estimator with $K_{\max}$ as the full subclassification estimator $\hat{\tau}_{FS}$ given by

$$\hat{\tau}_{FS} = \sum_{k=1}^{K_{\max}} \frac{N_k}{N} \left\{ \frac{1}{N_{1k}} \sum_{i=1}^{N} Z_i Y_i I(\hat{\pi}_i \in \hat{C}_k) - \frac{1}{N_{0k}} \sum_{i=1}^{N} (1 - Z_i) Y_i I(\hat{\pi}_i \in \hat{C}_k) \right\}. \tag{5}$$

Although Wang et al. (2020) provided a theoretical guideline for choosing K, they did not give a practical procedure for finding $K_{\max}$, but suggested using the largest K such that all subclasses $\hat{C}_1, ...\hat{C}_K$ contain at least one observation from each treatment group. However, does their suggested procedure satisfy conditions (2) and (4) ?

In the following sections, we first propose a new algorithm controlled by two parameters to generate subclasses and show the conditions for its parameters so that the number of subclasses $K(N)$ generated by the proposed algorithm satisfies conditions (2) and (4). We then point out that Wang et al. (2020)'s suggested procedure is a special case of our algorithm and show that the resulting subclassification estimator with their suggested procedure does not satisfy condition (4).

3. Causal Subclassification Tree

In this section, we propose an algorithm named causal subclassification tree, which generates subclasses $C_1, ..., C_K$ automatically from estimated propensity scores and potential outcomes. We evaluate upper and lower bounds of the number of subclasses $K(N)$ generated by our algorithm, and show such $K(N)$ satisfies conditions (2) and (4). Then, we discuss the robustness of the subclassification estimator with our algorithm in section 3.3, and explain the advantages of our algorithms in practical situations. The parameter tuning procedures and the extension of our algorithm to multi-class treatment are described at the end of this section.

3.1 Algorithm of Causal Subclassification Tree

Let the parent node be P and the child nodes C, C', which are generated from P. Further, let (N_{0P}, N_{1P}), (N_{0C}, N_{1C}), $(N_{0C'}, N_{1C'})$ be the number of samples of treatment and control groups in parent node P, child node C, and child node C', respectively. Then, splitting the node is controlled by parameters (ℓ, α) as follows. For a given $\ell \in \mathbf{N}$ and $\alpha \in (0, 0.5]$, the causal subclassification tree splits the parent node P into C, C' only when the conditions

$$N_{1C}, N_{1C'}, N_{0C}, N_{0C'} \geq \ell, \tag{6}$$

$$N_{0C}, N_{0C'} \geq \alpha N_{0P}, \qquad N_{1C}, N_{1C'} \geq \alpha N_{1P}. \tag{7}$$

are satisfied. Condition (6) controls the minimum leaf size of a tree so that each leaf contains at least ℓ units of each group, and condition (7) controls the ratio of units that child nodes C and C' contain from the parent node P. A moderate value of α prevents extremely imbalanced child nodes

Next, we describe our splitting rules. Let the estimated propensity score for sample i be $\hat{\pi}_i$. For a split of node P to C and C', we define the splitting criteria $\Delta_1(C, C')$ as follows.

$$\Delta_1(C, C') = \frac{1}{|\{i : \hat{\pi}_i \in C\}|} \sum_{\{i:\hat{\pi}_i \in C\}} (\hat{\pi}_i - \bar{\pi}_C)^2 + \frac{1}{|\{i : \hat{\pi}_i \in C'\}|} \sum_{\{i:\hat{\pi}_i \in C'\}} (\hat{\pi}_i - \bar{\pi}_{C'})^2 \tag{8}$$

where, $\bar{\pi}_C$ and $\bar{\pi}_{C'}$ are the averages of the estimated propensity scores in the child nodes C and C'. At node P, we choose the split that minimizes $\Delta(C, C')$ among all the splits that satisfy condition (6) and (7). Nodes are split recursively until no split satisfies conditions (6) and (7), i.e., each leaf has more than ℓ but less than $2\ell - 1$ samples of either classes. The procedure for the causal subclassification tree is described as follows.

Procedure.1 Causal Subclassification Tree

Input: estimated propensity score$\{\hat{\pi}_i\}_{i=1}^N$*, treatment variable* $\{T_i\}_{i=1}^N$*, minimum leaf size* ℓ*, sample fraction* α*.*

1. *find the split of the node that minimizes* $\Delta_1(C, C')$ *among the splits satisfying conditions (6) and (7), and apply this split to the node. Repeat this step recursively until no split satisfies (6) and (7).*
2. *Regard terminal nodes generated by the tree as subclasses* $C_1, C_2, ..., C_K$*.*
3. *Compute subclassification estimator* $\hat{\tau}_H$*, with equation (3)*

This criterion (8) is the sum of the variances of the estimated propensity scores in the child nodes C and C'. Instead of this criterion, we can also use the sum of the squared prediction error, a Gini coefficient, or information gain as the criterion.

We can use the following splitting criteria $\Delta_2(C, C')$ instead of $\Delta_1(C, C')$, which uses potential outcomes $Y(1)$ and $Y(0)$:

$$\Delta_2(C, C') = \frac{1}{|\{i : \hat{\pi}_i \in C\}|} \sum_{\{i:\hat{\pi}_i \in C\}} \left\{ T_i \left(Y_i - \bar{Y}_C(1)\right)^2 + (1 - T_i)\left(Y_i - \bar{Y}_C(0)\right)^2 \right\}$$
$$+ \frac{1}{|\{i : \hat{\pi}_i \in C'\}|} \sum_{\{i:\hat{\pi}_i \in C'\}} \left\{ T_i \left(Y_i - \bar{Y}_{C'}(1)\right)^2 + (1 - T_i)\left(Y_i - \bar{Y}_{C'}(0)\right)^2 \right\}$$

where $\bar{Y}_C(t)$ and $\bar{Y}_{C'}(t)$ are defined as

$$\bar{Y}_C(1) = \frac{1}{|\{i : \hat{\pi}_i \in C, T_i = t\}|} \sum_{\{i:\hat{\pi}_i \in C\}} T_i Y_i,$$
$$\bar{Y}_C(0) = \frac{1}{|\{i : \hat{\pi}_i \in C, T_i = 0\}|} \sum_{\{i:\hat{\pi}_i \in C\}} (1 - T_i) Y_i$$

and

$$\bar{Y}_{C'}(1) = \frac{1}{|\{i : \hat{\pi}_i \in C', T_i = t\}|} \sum_{\{i:\hat{\pi}_i \in C'\}} T_i Y_i,$$
$$\bar{Y}_{C'}(0) = \frac{1}{|\{i : \hat{\pi}_i \in C', T_i = 0\}|} \sum_{\{i:\hat{\pi}_i \in C'\}} (1 - T_i) Y_i,$$

which are the averages of the observed potential outcomes in child nodes C and C', respectively. Athey and Imbens (2016) pointed out that, when splits and outcomes are not independent, the subclassification estimator may have a bias. This problem can be avoided by applying the Double-Sample Tree (Wager & Athey, 2018) to the procedure. 1. Double-sample tree is a technique to achieve honesty (Wager & Athey, 2018) by dividing its training subsample into two halves $\mathcal{I}$ and $\mathcal{J}$. It uses the $\mathcal{J}-$sample to place the splits, while holding out the $\mathcal{I}$-sample to make the within-leaf estimation.

Procedure.2 Causal Subclassification Tree (Double-Sample)

Input: estimated propensity score $\{\hat{\pi}_i\}_{i=1}^N$*, outcomes* $\{Y_i\}_{i=1}^N$*, treatment variable* $\{T_i\}_{i=1}^N$*, minimum leaf size k¢sample fraction* α*.*

1. *Divide data randomly into disjoint sets of size* $|\mathcal{I}| = \lfloor N/2 \rfloor$ *and* $|\mathcal{J}| = \lceil N/2 \rceil$.
2. *Find the split of the node that minimizes* $\Delta_2(C, C')$ *among satisfying conditions (6) and (7), and apply this split to the nodes with all observations from* $\mathcal{I}$*-data, and* X*- or* T*-observations from* $\mathcal{J}$*-data, but without outcomes* Y*-observations from* $\mathcal{J}$*-data. Repeat this step recursively until no split satisfies (6) and (7) for* $\mathcal{I}$*-data.*
3. *Regard terminal nodes generated by the tree as subclasses* $C_1, C_2, ..., C_K$*.*
4. *Compute subclassification estimator* $\hat{\tau}_H$*, with equation (3), using* $\mathcal{I}$*-data.*

Remark 1. *Here, ℓ and α are parameters for controlling the tree so that extreme division is not generated. These parameters are also used for Wager and Walther (2015) and Causal Forest (Wager & Athey, 2018) for generating α-regular tree.*

3.2 Properties of the Numbers of Subclasses Generated by Causal Subclassification Tree

In this section, we discuss the asymptotic behavior of the number of subclasses $K(N)$ generated by the causal subclassification tree. The following theorem shows its upper and lower bounds.

Theorem 1 *The number of subclasses $K(N)$ generated by the causal subclassification tree with parameter (ℓ, α) satisfies the following inequality.*

$$2^{\log \frac{\min\{N_1, N_0\}}{2\ell-1} / \log \alpha^{-1}} \leq K(N) \leq 2^{\log \frac{\max\{N_1, N_0\}}{\ell} / \log(1-\alpha)^{-1}} \tag{9}$$

where N is the number of observations and N_1 and N_0 are the number of observations in the treatment and control groups, respectively. Thus, $N_1 + N_0 = N$.

Proof.

By assumption (1), the propensity score is $0 < p(x) < 1$ for any $x \in \mathcal{X}$, so that as $N \to \infty$, it holds $N_0, N_1 \to \infty$.

Let $d_{\min}$ and $d_{\max}$ be the minimum and the maximum depth of the causal subclassification tree with sample size (N_1, N_0) and parameters (ℓ, α). Because a child node contains at least $100\alpha\%$ units of its parent node by condition (7), the following inequality holds,

$$\min\{N_1, N_0\}\alpha^{d_{\min}} \leq 2\ell - 1$$

Additionally, because the child node contains at the most $100(1-\alpha)\%$ units of its parent node, it holds

$$\ell \leq \max\{N_1, N_0\}\,(1-\alpha)^{d_{\max}}\,.$$

Thus, we have the following inequality:

$$\log \frac{\min\{N_1, N_0\}}{2\ell - 1} \Big/ \log \alpha^{-1} \leq d_{\min}$$

$$d_{\max} \leq \log \frac{\max\{N_1, N_0\}}{\ell} \Big/ \log(1-\alpha)^{-1}$$

where we use logarithm base as 2, because two leaves are generated from one split. The number of terminal nodes in a tree is larger than $2^{d_{\min}}$ and smaller than $2^{d_{\max}}$. Therefore, the inequality (9) holds.

Theorem 2 *For $\alpha \in (0, 0.5]$, and $\ell = \sqrt{N}/\log(N)$, the number of subclasses $K(N)$ generated by causal subclassification tree satisfies condition (2) and (4);*

$$K(N) \to \infty, \quad K(N)/\sqrt{N} \to \infty \quad \textit{and} \quad K(N)\log(K(N))/N \to 0$$

as $N \to \infty$.

Proof. By assumption (1), $N_0/N \to P(W = 0) \neq 0$, $N_1/N \to P(W = 1) \neq 0$. So, $N_j = O(N)$. The lower bound in (9) is shown to be

$$\begin{aligned} 2^{\log \frac{\min\{N_1, N_0\}}{2\ell-1} / \log \alpha^{-1}} &= 2^{\log \frac{\min\{N_1, N_0\}}{2\sqrt{N}/\log(N) - 1} / \log \alpha^{-1}} \\ &= 2^{\log \frac{\min\{N_1, N_0\}\log(N)}{2\sqrt{N} - \log(N)} / \log \alpha^{-1}} \\ &= O(\sqrt{N}\log(N)), \end{aligned}$$

and the upper bound is also shown to be

$$\begin{aligned} 2^{\log \frac{\max\{N_1, N_0\}}{\ell} / \log(1-\alpha)^{-1}} &= 2^{\log \frac{\max\{N_1, N_0\}}{\sqrt{N}/\log(N)} / \log(1-\alpha)^{-1}} \\ &= 2^{\log \frac{\max\{N_1, N_0\}\log(N)}{\sqrt{N}} / \log(1-\alpha)^{-1}} \\ &= O(\sqrt{N}\log(N)). \end{aligned}$$

Thus, by inequality (9),

$$K(N) = O(\sqrt{N}\log(N))$$

and conditions (2) and (4) hold.

According to theorem 2, if we set $\ell = \sqrt{N}/\log(N)$, then $K(N)$ satisfies conditions (2) and (4). Therefore, the subclassification estimator obtained by the causal stratification tree is an $\sqrt{N}$-consistent estimator for τ and is asymptotically well defined.

3.3 Properties of Causal Subclassification Tree

In this section, we first explain the connection between the causal subclassification algorithm and Wang et al. (2020)'s strategy : using the largest K such that all subclasses $\hat{C}_1, ...\hat{C}_k$ contain at least one observation from each treatment group. Their strategy to generate subclasses can be considered as a special case of our algorithm with a small enough parameter α and parameter $\ell = 1$. When we set $\ell = 1$, the number of units from both/either of the treatment or control groups in each subclass is 1, so this is one of the possible implementations of their strategy. However, if we set $\ell = 1$ for any sample size N, because the lower and upper limits of the number of subclasses $K(N)$ are given by inequality (9),

$$2^{\log\min\{N_1,N_0\}/\log\alpha^{-1}} \le K(N) \le 2^{\log\max\{N_1,N_0\}/\log(1-\alpha)^{-1}},$$

the order of $K(N)$ becomes $O(N)$. If $K(N) = O(N)$, the condition (2) for $\sqrt{N}-$consistency is satisfied, but the condition (4) is not satisfied, because $K(N)\log(N)/N \to \infty$. This result holds for any criterion function of the causal subclassification tree, so the subclassification estimator with Wang et al. (2020)'s strategy does not satisfy condition (4) and it is not guaranteed to be asymptotically well-defined.

Although the subclassification estimator with fixed ℓ does not satisfy the condition (4), if we increase ℓ in the order of $\sqrt{N}/\log(N)$, the resulting subclassification estimator becomes asymptotically well-defined. We choose ℓ, which minimizes the loss function introduced in section 3.5 by cross-validation from the range

$$\left[\frac{\sqrt{\min\{N_1,N_0\}}}{2\log(\min\{N_1,N_0\})}, \frac{3\sqrt{\max\{N_1,N_0\}}}{2\log(\max\{N_1,N_0\})}\right].$$

By theorem 1 and theorem 2, the subclassification estimator with causal subclassification tree algorithm with our strategy satisfies conditions (2) and (4), therefore it is $\sqrt{N}-$consistent and asymptotically well defined.

We would like to mention/emphasize here that, unlike conventional methods that automatically choose cut-off points of propensity score estimates based on quantiles, the causal subclassification tree chooses cut-off points that minimize criteria such as Δ_1 or Δ_2. This is because the choice of cut-off points affects the performance of the resulting subclassification estimator, and we would like to control it as much as we can by defining criteria that reflect our thought of "good performance" and choosing cut-off points based on such a performance. For example, if we use quantiles of the estimated propensity score, it is difficult to explain why we use such cut-off points or answer whether there are any better cut-off points. For these practical reasons, data-driven cut-off point determination is important in data analysis.

If we use a causal subclassification tree, because the criterion function Δ is needed, then the resulting subclasses satisfy the properties that are led by Δ. For example, if we use Δ_1 as a criterion function in the causal subclassification tree, we can obtain subclasses that have a small variance of estimated propensity scores in each subclass, or if we use Δ_2 as a criterion function, subclasseses such that the variance of observed outcomes in each subclass is small are generated. In addition, if we use Δ_3 as a criterion function,

$$\Delta_3(C, C') = \frac{1}{|\{i : \hat{\pi}_i \in C\}|}\sum_{\{i:\hat{\pi}_i\in C\}} T_i \log\frac{N_{1C}}{N_C} + (1 - T_i)\log\frac{N_{0C}}{N_C} \tag{10}$$

$$+ \frac{1}{|\{i : \hat{\pi}_i \in C'\}|}\sum_{\{i:\hat{\pi}_i\in C'\}} T_i \log\frac{N_{1C'}}{N_{C'}} + (1 - T_i)\log\frac{N_{0C'}}{N_{C'}} \tag{11}$$

we can generate subclasses that minimize the likelihood of distribution of treatment variables T. Therefore, from these discussions, one of the attractive properties of using the causal subclassification tree is that researchers can change the arbitrary criterion function with intention using properties so that subclasses may be satisfied.

3.4 Subclassification Estimator for Weighted Average Treatment Effect

In practical situations, we may want to estimate not only the average treatment effect, but also the weighted average treatment effects. For example, we are often interested in the average treatment effect on the treated (ATT), average treatment effect on the untreated (ATU), or the treatment effect on the overlap (ATO, Li, Morgan, & Zaslavsky, 2018). In this section, we introduce subclassification estimators for the weighted average treatment effect. The IPW estimator for average treatment effect on the treated(ATT), $\mathbb{E}[Y_i(1) - Y_i(0)|T_i = 1]$, is given by

$$\tau_{ATT} = \frac{1}{N_1}\sum_{i=1}^{N} T_i Y_i - \sum_{i=1}^{N} \frac{\hat{\pi}_i}{1-\hat{\pi}_i}(1-T_i)Y_i \Big/ \sum_{i=1}^{N} \frac{\hat{\pi}_i}{1-\hat{\pi}_i}(1-T_i)$$

The subclassification estimator for ATT can be obtained by replacing $1/\hat{\pi}_i$ with subclassification weight $1/\hat{p}_i$ as follows:

$$\hat{\tau}^S_{ATT} = \frac{1}{N_1}\sum_{k=1}^{K}\left\{N_{1k}\left(\frac{1}{N_{1k}}\sum_{\{i\in C_k\}} T_i Y_i - \frac{1}{N_{0k}}\sum_{\{i\in C_k\}}(1-T_i)Y_i\right)\right\}$$

Here, $\left(\frac{1}{N_{1k}}\sum_{\{i\in C_k\}} T_i Y_i - \frac{1}{N_{0k}}\sum_{\{i\in C_k\}}(1-T_i)Y_i\right)$ is the difference in the average of the potential outcomes in the treatment and control groups on subclass C_k. Therefore, the subclassification estimator for ATT can be considered as a weighted mean of causal effects of subclasses with sample sizes of the treatment group. This shows that a subclass with a high probability of receiving treatment is assigned a large weight.

Similarly, the subclassification estimator for the Average Treatment effect on the Untreated (ATU), $\mathbb{E}[Y_i(1)-Y_i(0)|T_i = 0]$, is given by

$$\hat{\tau}^S_{ATU} = \frac{1}{N_0}\sum_{k=1}^{K}\left\{N_{0k}\left(\frac{1}{N_{0k}}\sum_{\{i\in C_k\}} T_i Y_i - \frac{1}{N_{0k}}\sum_{\{i\in C_k\}}(1-T_i)Y_i\right)\right\},$$

and the subclassification estimator for the Average Treatment effect on the Overlap (ATO, Li et al. (2018)) is given by

$$\hat{\tau}^S_{ATO} = \sum_{k=1}^{K}\left\{\frac{N_{0k}N_{1k}}{N_k}\left(\frac{1}{N_{1k}}\sum_{\{i\in C_k\}} T_i Y_i - \frac{1}{N_{0k}}\sum_{\{i\in C_k\}}(1-T_i)Y_i\right)\right\} \Big/ \sum_{k=1}^{K}\frac{N_{0k}N_{1k}}{N_k}$$

3.5 A Tuning Algorithm for Parameter (ℓ, α)

We discuss three parameter tuning methods for (ℓ, α). All of them use a cross-validation method for choosing parameters, and to choose ℓ and α such that $K(N)$ satisfies condition (2) and (4), we explore a range

$$\left[\frac{\sqrt{\min\{N_1, N_0\}}}{2\log(\min\{N_1, N_0\})}, \frac{3\sqrt{\max\{N_1, N_0\}}}{2\log(\max\{N_1, N_0\})}\right]$$

for the parameter ℓ and $(0, 0.5]$ for the parameter α, where N_1 and N_0 are the sample sizes of treatment and control groups. For example, when the sample size $N = 1000$ with $(N_0, N_1) = (200, 800)$, we explore [2,6] for the parameter ℓ as candidates. The first method involves finding the minimizer of the cross-validation prediction error of the causal stratification tree. The second method involves finding the minimizer of the measure of covariates imbalance, for example, Imai and Ratkovic (2014)'s imbalance measure:

$$Imbalance = \left(\left(\frac{1}{N}\sum_{i=1}^{N} w_i X_i\right)^T \left(\frac{1}{N}\sum_{i=1}^{N} X_i X_i^T\right)^{-1} \left(\frac{1}{N}\sum_{i=1}^{N} w_i X_i\right)^T\right)^{1/2},$$

or the one by Wang and Zubizarreta (2019):

$$Imbalance = \left\|\frac{\sum_{i=1}^{N} w_i T_i X_i}{\sum_{i=1}^{N} w_i T_i} - \frac{1}{N}\sum_{i=1}^{N} X_i\right\| / \mathrm{sd}(X)$$

where w_i are the inverse of estimated propensity scores. These two approaches are intuitive. However, it is well known that if the model for propensity score contains all variables that affect potential outcomes including confoundings, then the mean square error of IPW estimator becomes smaller than in the case when the propensity score model contains only confounding variables (Brookhart et al., 2006).. Hence, we consider a measure Γ that uses a variance of potential outcomes, defined as follows:

$$\Gamma = \frac{1}{N}\sum_{k=1}^{K} N_k \Gamma_k$$

where Γ_k is

$$\begin{aligned}\Gamma_k &= \frac{1}{N_{1k}}\sum_{i=1}^{N} T_i(Y_i(1)-\bar{Y}^k(1))^2 I(\hat{\pi}_i \in C_k)\\ &\quad + \frac{1}{N_{0k}}\sum_{i=1}^{N}(1-T_i)(Y_i(0)-\bar{Y}^k(0))^2 I(\hat{\pi}_i \in C_k), \quad \text{and}\\ \bar{Y}^k(j) &= \sum_{i=1}^{N} T_i Y_i(j) I(\hat{\pi}_i \in C_k)/N_j, \quad j=0,1.\end{aligned}$$

Here, Γ_k is the sum of the variances of observed outcome of each group in the subclass C_k. We find a parameter (ℓ, α) that minimizes Γ by cross validation.

3.6 Extension to Multi-class Treatment Regimes

We extend the causal subclassification tree to causal inference with multi-class treatments. Let T_i be the treatment that takes one of the K integer values, i.e., $T_i \in \mathcal{T} = \{0,1,2,...,K-1\}$, where $K \geq 2$, and $Y_i(t), t = 0,1,...,K-1$ are potential outcomes corresponding to treatment t. Following Imbens (2000), we define the generalized propensity score as

$$\pi_i^k = \pi^k(X_i) = \Pr(T_i = k|X_i)$$

where all conditional probabilities sum to 1, i.e., $\sum_{k=0}^{K-1}\pi^k(X_i) = 1$. We may use a multinomial logistic regression model to estimate the generalized propensity scores.

In causal inference with multi-class treatment regimes, we are interested in difference of effects between two different treatments t and t', $\tau(t,t') = \mathbb{E}[Y_i(t) - Y_i(t')]$. An IPW estimator for $\tau(t,t')$, proposed by Feng, Zhou, Zou, Fan, and Li (2012), is defined as

$$\hat{\tau}(t,t')^{IPW} = \frac{\sum_{i=1}^{N} I(T_i = t)Y_i/\pi_i^t}{\sum_{i=1}^{N} I(T_i = t)/\pi_i^t} - \frac{\sum_{i=1}^{N} I(T_i = t')Y_i/\pi_i^{t'}}{\sum_{i=1}^{N} I(T_i = t')/\pi_i^{t'}}.$$

However, this estimator becomes unstable when there are some misspecification in the propensity score model. Thus, we propose a new subclassification method that extends the causal subclassification tree algorithms to multi-class treatment regimes.

Let $\hat{\pi}_i = (\hat{\pi}_i^0, \hat{\pi}_i^1, ..., \pi_i^{K-1})^T$ be the estimated generalized propensity scores. We can extend the causal subclassification tree to multi-class treatment regimes only by replacing Δ_1 in (8) with the following Δ_3:

$$\Delta_3(C_1, C_2) = \sum_{j=1}^{2}\frac{1}{|\{i : \hat{e}_i \in C_j\}|}\sum_{\{i:\hat{\pi}_i \in C_j\}}\left(\hat{\pi}_i - \bar{\pi}_{C_j}\right)^T\left(\hat{\pi}_i - \bar{\pi}_{C_j}\right).$$

Then, just as in the binary treatment case, we define the weights for each treatment t as follows.

$$w_{it} = \sum_{k=1}^{K(N)}\frac{N_k}{N_{tk}} I(\hat{\pi}_i \in \hat{C}_k) = \sum_{k=1}^{K(N)}\frac{1}{\hat{p}_{tk}} I(\hat{\pi}_i \in \hat{C}_k)$$

where N_k is the number of units in subclass C_k, and N_{tk} is the number of units that received treatment t in subclass C_k. A subclassification estimator for $\tau(t,t')$ is defined by replacing $1/\pi_i^t$ and $1/\pi_i^{t'}$ with w_{it} and $w_{it'}$, respectively, as

$$\hat{\tau}_H(t,t') = \sum_{i=1}^{N} w_{it}T_iY_i \Big/ \sum_{i=1}^{N} w_{it} - \sum_{i=1}^{N} w_{i'}(1-T_i)Y_i \Big/ \sum_{i=1}^{N} w_{it'}.$$

4. Simulation Studies

We apply the causal subclassification tree to the simulation data in almost the same settings as Kang and Schafer (2007), where propensity score models were slightly misspecified. In the controversial paper, Kang and Schafer (2007) conducted a set of simulation experiments to study the performance of propensity score weighting methods. They found that the misspecification of a propensity score model can affect the performance of various weighting methods. In particular, they showed that although the doubly robust estimator of Robins et al. (1994) provides a consistent estimate for the treatment effect if either the outcome model or the propensity score model is correct, the performance of the doubly robust estimator can deteriorate when both models are slightly misspecified. In this section, we describe the simulation experiment performed with the data generated in Kang and Schafer (2007)'s model. We then examine whether the subclassification estimator with the causal subclassification tree can improve the empirical performance of propensity score weighting estimators and full subclassification estimator constructed by Wang et al. (2020)'s procedure.

Our data-generating process is as follows. There are four pretreatment covariates $X_i = (X_{i1}, X_{i2}, X_{i3}, X_{i4})$ each of which is independently, identically distributed as $N(0, 1)$. The true outcome model is a linear regression with these covariates and the error term is an independently and identically distributed standard normal random variate:

$$y_i = 210 + 5T_i + 27.4X_{i1} + 13.7X_{i2} + 13.7X_{i3} + 13.7X_{i4} + \varepsilon_i$$

where $\varepsilon_i \sim N(0, 5)$. The mean outcome of the treated observations is 210, which is the quantity of interest to be estimated. The true propensity score model is a logistic regression, with X_i being the linear predictor such that the mean probability of receiving the treatment equals 0.5,

$$\pi_i = \text{expit}(-X_{i1} + 0.5X_{i2} - 0.25X_{i3} - 0.1X_{i4}).$$

Finally, only the non-linear transforms of covariates are observed, and they are given by $Z_i = (Z_{i1}, Z_{i2}, Z_{i3}, Z_{i4})$ where

$$\begin{aligned} Z_{i1} &= \exp(X_{i1}/2), \\ Z_{i2} &= X_{i2}/(1 + \exp(X_{i1})) + 10, \\ Z_{i3} &= (X_{i1}X_{i3}/25 + 0.6)^3, \\ Z_{i4} &= (X_{i2} + X_{i4} + 20)^2. \end{aligned}$$

In this simulation, three propensity score weighting estimators are investigated: the Horvitz-Thompson estimator (HT, Horvitz & Thompson, 1952), the inverse propensity score weighting estimator (IPW, Hirano et al., 2003), and the doubly robust estimator (DR, Robins et al., 1994) given by

$$\hat{\tau}_{\text{HT}} = \frac{1}{N}\sum_{i=1}^{N}\frac{T_iY_i}{\hat{\pi}_i} - \frac{1}{N}\sum_{i=1}^{N}\frac{(1-T_i)Y_i}{1-\hat{\pi}_i},$$

$$\hat{\tau}_{\text{IPW}} = \sum_{i=1}^{N}\frac{T_iY_i}{\hat{\pi}_i}\Big/\sum_{i=1}^{N}\frac{T_i}{\hat{\pi}_i} - \sum_{i=1}^{N}\frac{(1-T_i)Y_i}{1-\hat{\pi}_i}\Big/\sum_{i=1}^{N}\frac{1-T_i}{1-\hat{\pi}_i} \qquad \text{and}$$

$$\begin{aligned} \hat{\tau}_{\text{DR}} &= \frac{1}{N}\sum_{i=1}^{N}\left\{X_i^T\hat{\gamma}_{\text{OLS1}} + \frac{T_i(Y_i - X_i^T\hat{\gamma}_{\text{OLS}})}{\hat{\pi}_i}\right\} \\ &\quad - \frac{1}{N}\sum_{i=1}^{N}\left\{X_i^T\hat{\gamma}_{\text{OLS0}} + \frac{(1-T_i)(Y_i - X_i^T\hat{\gamma}_{\text{OLS}})}{1-\hat{\pi}_i}\right\}, \\ \text{where} \qquad \hat{\gamma}_{\text{OLS1}} &= \left\{\sum_{i=1}^{N}T_iX_iX_i^T\right\}^{-1}\sum_{i=1}^{N}T_iX_iY_i\text{and} \\ \hat{\gamma}_{\text{OLS0}} &= \left\{\sum_{i=1}^{N}(1-T_i)X_iX_i^T\right\}^{-1}\sum_{i=1}^{N}(1-T_i)X_iY_i. \end{aligned}$$

The true model for propensity scores is a logistic regression with X_i as covariates. Thus, the models with Z_i as covariates are misspecified for these data. Likewise, the true outcome model is a linear regression with X_i as covariates and the models with Z_i as covariates are misspecified. However, only the DR estimator uses an outcome model, because outcome models need not be considered for HT and IPW estimators. When we use causal subclassification weights(CSW) w_{i1} instead of estimated propensity scores π_i in each estimator, we denote them as HT-CSW, IPW-CSW, and DR-CSW. In addition, when we use full subclassification weights (FSW, Wang et al., 2020), we denote them as HT-FSW, IPW-FSW, and DR-FSW. To obtain full subclassification weights, following the discussion in section 3.3, we use a causal subclassification tree with the parameter $\ell = 1$.

To estimate the propensity score, Kang and Schafer (2007) used logistic regression with Z_i as predictors, i.e., $\pi(X_i) = \text{logit}^{-1}(Z_i^T\beta)$, that is, the model was misspecified because the true propensity score model is a logistic regression with

X_i as predictors. In our simulation, in addition to logistic regression, we use Covariate Balancing Propensity Score (CBPS, Imai & Ratkovic, 2014), Random Forest (Breiman, 2001), and XgBoost (Chen & Guestrin, 2016) to estimate the propensity score. We use the same propensity and outcome model specifications as Kang and Schafers and investigate whether the use of causal subclassification weights $\hat{p}_i$ instead of the estimates $\hat{\pi}_i$ by LR, CBPS, RF, and XgBoost improves the empirical performance of these estimators. For each estimated propensity score with several estimation methods, we computed causal subclassification weights using the causal subclassification tree (Procedure.2). To train each tree, we used the parameter-tuning algorithm in subsection 3.5, which uses potential outcomes. In other words, our simulation study examines how replacing the reciprocal of estimated propensity score with causal subclassification weights will improve the empirical performance of the three commonly used weighting estimators. As in Kang and Schafer's study, we conducted simulations under the following four scenarios:

(a) both propensity score and outcome models are correctly specified,

(b) only the propensity score model is correct,

(c) only the outcome model is correct, and

(d) both the propensity score and the outcome models are misspecified.

For each scenario, we conducted 1000 Monte Carlo simulations with a sample size of 1000 and computed the bias, variance, and root-mean-squared error (RMSE) for each estimator. The results are presented in Table 1, 2, 3. HT and IPW estimators do not depend on outcome models, and hence, we do not fill the values into corresponding cells in each table. As discussed in section 3.3, HT-CSW and IPW-CSW are the same estimators, and hence, we do not write the result of HT-CSW in each table. For each scenario, we computed the bias, variance, and RMSE for each weighting estimator on the basis of five different estimation methods for the propensity score:

(i) the standard logistic regression with X_i being the linear predictor as in the original simulation study (LR),

(ii) the just-identified CBPS (Imai & Ratkovic, 2014) estimation with the covariate balancing moment conditions with respect to X_i and without the score condition (CBPS(exact)),

(iii) the overidentified CBPS estimation (Imai & Ratkovic, 2014) with both covariate balancing and score conditions (CBPS(over)),

(iv) the Random Forest with X_i as predictors (RF), and

(v) the XgBoost with X_i as predictors (XgBoost).

In the first scenario (a), the case where both models are correct, HT, IPW, and DR estimators have relatively low bias when LR and CBPSs are used as models for propensity score regardless of whether we use the causal subclassification tree. However, HT and IPW estimators have a larger absolute bias compared to those with CSW and FSW when RF and XgBoost are used. Each subclassification estimator has about the same or lower variances compared to those with estimated propensity scores themselves. Comparing CSW and FSW, both have almost the same absolute bias; however, the variances of FSW estimators are twice as large as those of CSW. As a result, the RMSE of estimators with causal subclassification weights are relatively smaller than those with original weights except CBPS(exact), especially when RF and XgBoost are used, and also smaller than those of full subclassification weights. DR is not sensitive to the choice of propensity score estimation methods because of its property.

The second simulation scenario, (b), shows the performance of various estimators when the propensity score model is correct but the outcome model is misspecified. The results for HT and IPW are same as those of the first scenario (a) because these estimators only depend on the model for propensity scores. For the DR estimator, DR-CSW and FSW have relatively similar biases and smaller variances compared to DR when LR and CBPS are used for the propensity score model. In contrast, when RF and XgBoost are used for the propensity score model, DR-CSW has a smaller bias, but a relatively large variance. As a consequence, RMSE for all estimators with causal subclassification weights are smaller than those with the estimated propensity score themselves. Comparing DR-CSW and DR-FSW, both have almost the same absolute bias; however, the variances of FSW estimators are twice as large as those of CSW. As a result, all RMSE of the DR-CSW estimator are smaller than those of DR-FSW.

The third scenario, (c), is the situation where the propensity score model is misspecified whereas the outcome models for estimators DR is correct. Because HT and IPW estimators rely only on the propensity scores, these estimators may have large bias and/or variance. For LR, IPW-CSW and IPW-FSW have larger biases but smaller variances compared to the IPW estimator (for HT estimator, HT-CSW and HT-FSW have smaller biases and variances). RMSE for IPW-CSW and IPW-FSW are smaller than those for HT and IPW. For IPW-CSW with both, CBPS have larger bias and variance compared to IPW. For HT and IPW estimators with RF and XgBoost, IPW-CSW have about the same or smaller bias and variance than HT and IPW, respectively. When LR, RF, and XgBoost are used, HT-CSW and IPW-CSW have relatively smaller RMSE, but when both CBPS are used, HT-CSW and IPW-CSW have larger RMSE. This is an interesting phenomenon. This is because the subclassification estimator with our weights has a consistency when the order of propensity score is correctly specified (by the properties of regression tree, i.e., Wager & Walther, 2015). LR, RF, and XgBoost are non-linear prediction models for W, but CBPS is not. CBPS finds the solution for balancing equations for covariates of treatment and control groups, and hence, the performances of HT and IPW estimators with LR, RF, and XgBoost are improved with CSW but the estimators with CBPS are not. Comparing the subclassification estimators CSW with FSW without DR in scenario (c), in short, a similar result is confirmed for scenario (a) and (b). Both of these estimators have similar absolute biases, and FSW has larger variances than CSW.

In the final simulation scenario (d), the performance of estimator DR deteriorated because both the propensity score and the outcome models are misspecified. The bias and RMSE for DR are the largest in all scenarios. However, for all the propensity score estimation methods, DR-CSW and DR-FSW have relatively smaller biases and RMSE compared to DR. Both DR and DR-CSW with RF or XgBoost have small RMSE compared to other cases; especially, DR-CSW with ML methods achieves the smallest RMSE among the propensity score models. As in the previous scenarios, let us compare the biases and variances of the subclassification estimators of CSW and FSW; the same result is seen, where both of these estimators have a similar absolute bias, and FSW has a larger variance than CSW.

Summarizing the above discussions, weights that we proposed tend to improve the performance of all estimators with original estimated propensity scores and show a robustness for model misspecification. Especially regarding RF and XgBoost, all estimators with causal subclassification weights improve bias, variance, and RMSE dramatically than those without CSW. If the propensity score estimates are directly used, machine learning methods, such as RF and Xgboost, are not recommended. However, by combining them with our proposal algorithm, causal subclassification tree, these machine learning methods overperform LR and CBPSs when models are misspecified. Furthermore, by comparing the result of the subclassification estimator for CSW and FSW, for all scenarios, both have the same bias, but FSW has twice as large of a variance as CSW, so for all scenarios and estimators, the subclassification estimators with CSW have smaller RMSE than those of FSW.

4. Concluding Remarks

In observational studies, the existence of confounding variables should be attended to and propensity score weighting methods are often used to eliminate their effects. The propensity score methods are extended for the various settings including multiple treatment, longitudinal, or time-varying treatments. Several papers discuss the risk of misspecification of propensity score models (e.g., Drake, 1993); however, in practice, insufficient attention has been paid for the propensity score estimation or the instability of weighting estimators.

In this paper, we proposed a subclass construction algorithm, a causal subclassification tree, that satisfies the conditions given by Wang et al. (2020) for the subclassification estimator to have $\sqrt{N}$-consistency and to be asymptotically well defined. We also showed that Wang et al. (2020)'s suggested procedure does not satisfy condition (4). The advantage of the causal subclassification tree is that, unlike conventional methods that do attend to the choice of cut-off points of propensity score estimates, the causal subclassification tree chooses cut-off points that minimize criteria. By defining criteria that reflect our thoughts about "good performance," we can control the performance of the causal estimator.

The simulation study showed that the proposed method improves the performance of propensity score weighting estimators, especially when the propensity score model is misspecified. When both models are misspecified, doubly robust estimators with our weights performed better than the original one, and doubly robust estimators with machine learning methods achieve smaller RMSE than those using ordinal propensity score estimation methods. That is, doubly robust estimators with machine learning methods combined with our algorithm have the smallest RMSE when both models are misspecified.

We also compared the empirical performance of subclassification estimator using our algorithm and the subclassification estimator constructed by Wang et al. (2020)s suggested procedure. The simulation result shows that from the view of estimation bias, there is little difference in the two estimators, but from the view of the variance of estimators, the full subclassification estimator has twice as much variance as that with a causal subclassification tree.

Lastly, we mention future work. If we use subclassification methods, the following two remaining effects for causal estimators in each subclass should be considered. One is the effect of residual within-subclass confounding (e.g., Lunceford

& Davidian, 2004), and the other is the effect of the variation of outcomes within-subclass that are affected by covariates regarding outcomes and error terms. If ℓ is set small, the residual within-subclass confounding becomes sufficiently small; however, the effects of the outcome variations may not be adjusted. We will now try to overcome such a problem by combining regression methods with a causal subclassification tree to estimate leaf-wise causal effect estimation for stabilizing the resulting estimator.

Acknowledgements

This work was partially supported by the Japan Society for the Promotion of Science KAKENHI Grant Number JP 17K00055.

References

Abadie, A., & Imbens, G. W. (2006). Large Sample Properties of Matching Estimators for Average Treatment Effects. *Econometrica, 74*(1), 235-267. https://doi.org/10.1111/j.1468-0262.2006.00655.x

Athey, S., & Imbens, G. W. (2016). Recursive partitioning for heterogeneous causal effects. *Proceedings of the National Academy of Sciences, 113*(27), 7353-7360. https://doi.org/10.1073/pnas.1510489113

Athey, S., Tibshirani, J., & Wager, S. (2019). Generalized random forests. *Annals of Statistics, 47*(2), 1148-1178. https://doi.org/10.1214/18-AOS1709

Breiman, L. (2001). Random forests. *Machine Learning, 45*, 5-32. https://doi.org/10.1023/A:1010933404324

Breiman, L., Friedman, J., Stone, C. J., & Olshen, R. (1984). *Classification and Regression Trees*. Chapman and Hall/CRC, first edition.

Brookhart, M. A., Schneeweiss, S., Rothman, K. J., Glynn, R. J., Avorn, J., & Sturmer, T. (2006). Variable Selection for Propensity Score Models. *American Journal of Epidemiology, 163*(12), 1149-1156. https://doi.org/10.1093/aje/kwj149

Cannas, M., & Arpino, B. (2019). A comparison of machine learning algorithms and covariate balance measures for propensity score matching and weighting. *Biometrical Journal, 61(4)*, 1049-1072. https://doi.org/10.1002/bimj.201800132

Chen, T., & Guestrin, C. (2016). Xgboost: A scalable tree boosting system. *In Proceedings of the 22Nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Association for Computing Machinery*, 785-794. https://doi.org/10.1145/2939672.2939785

Drake, C. (1993). Effects of Misspecification of the Propensity Score on Estimators of Treatment Effect. *Biometrics, 49*(4), 1231-1236. https://doi.org/10.2307/2532266

Feng, P., Zhou, X. H., Zou, Q. M., Fan, M. Y., & Li ,X. S. (2012). Generalized propensity score for estimating the average treatment effect of multiple treatments. *Statistics in Medicine, 31*(7), 681-697. https://doi.org/10.1002/sim.4168

Hansen, B. B. (2004). Full Matching in an Observational Study of Coaching for the SAT. *Journal of the American Statistical Association, 99*, 609-618. https://doi.org/10.1198/016214504000000647

Heckman, J. J., Ichimura, H., & Todd, P. E. (1997). Matching As An Econometric Evaluation Estimator: Evidence from Evaluating a Job Training Programme. *The Review of Economic Studies, 64*(4), 605-654. https://doi.org/10.2307/2971733

Hill, J. L. (2011). Bayesian nonparametric modeling for causal inference. *Journal of Computational and Graphical Statistics, 20*(1), 217-240. https://doi.org/10.1198/jcgs.2010.08162

Hirano, K., Imbens, G. W., & Ridder, G. (2003). Efficient estimation of average treatment effects using the estimated propensity score. *Econometrica, 71*(4), 1161-1189. https://doi.org/10.1111/1468-0262.00442

Horvitz, D. G., & Thompson, D. J. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association, 47*, 663-685. https://doi.org/10.1080/01621459.1952.10483446

Imai, K., & Ratkovic, M. (2014). Covariate balancing propensity score. *Journal of the Royal Statistical Society: Series B, 76*, 243-263. https://doi.org/10.1111/rssb.12027

Kang, J. D. Y., & Schafer, J. L. (2007). Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data. *Statistical Science, 22*, 523-539. https://doi.org/10.1214/07-STS227

Li, F., Morgan, K. L., & Zaslavsky, A. M. (2018). Balancing covariates via propensity score weighting. *Journal of the*

American Statistical Association, 113, 390-400. https://doi.org/10.1080/01621459.2016.1260466

Lunceford, J. K., & Davidian, M. (2004). Stratification and weighting via the propensity score in estimation of causal treatment effects: a comparative study. *Statistics in Medicine, 23*, 2937-2960. https://doi.org/10.1002/sim.1903

Nie, X., & Wager, S. (2020). *Quasi-Oracle Estimation of Heterogeneous Treatment Effects.* Biometrika, forthcoming. https://doi.org/10.1093/biomet/asaa076

Robins, J. M., Hernán, M. A., & Brumback, B. (2000): Marginal structural models and causal inference in epidemiology. *Epidemiology, 11*, 550-560. https://doi.org/10.1097/00001648-200009000-00011

Robins, J. M., Rotnitzky, A., & Zhao, L. P. (1994). Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association, 89*, 846-866. https://doi.org/10.1080/01621459.1994.10476818

Rosenbaum, P. R. (1987). Model-based direct adjustment. *Journal of the American Statistical Association, 82*, 387-394. https://doi.org/10.1080/01621459.1987.10478441

Rosenbaum, P. R. (1989). Optimal matching for observational studies. *Journal of the American Statistical Association, 84*, 1024-1032. https://doi.org/10.1080/01621459.1989.10478868

Rosenbaum, P. R. (1991). A characterization of optimal designs for observational studies. *Journal of the Royal Statistical Society: Series B, 53*, 597-610. https://doi.org/10.1111/j.2517-6161.1991.tb01848.x

Rosenbaum, P. R., & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika, 70*, 41-55. https://doi.org/10.1093/biomet/70.1.41

Rosenbaum, P. R., & Rubin, D. B. (1984). Reducing bias in observational studies using subclassification on the propensity score. *Journal of the American Statistical Association, 79*, 516-524. https://doi.org/10.1080/01621459.1984.10478078

Rosenbaum, P. R., & Rubin, D. B. (1985). Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. *The American Statistician, 39*, 33-38. https://doi.org/10.1080/00031305.1985.10479383

Wager, S., & Athey, S. (2018). Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association, 113*, 1228-1242. https://doi.org/10.1080/01621459.2017.1319839

Wager, S., & Walther, G. (2016). Adaptive concentration of regression trees, with application to random forests. https://arxiv.org/abs/1503.06388v3

Wang, L., Zhang, Y., Richardson, T. S., & Zhou, X. (2020). Robust estimation of propensity score weights via subclassification. https://arxiv.org/abs/1602.06366v2

Wang, Y., & Zubizarreta, J. R. (2020). Minimal dispersion approximately balancing weights: asymptotic properties and practical considerations. *Biometrika, 107*, 93-105. https://doi.org/10.1093/biomet/asz050

Table 1. Absolute Bias : N = 1000, iteration 1000

Model for Propensity Score	HT	HT-CSW	HT-FSW	IPW	IPW-CSW	IPW-FSW	DR	DR-CSW	DR-FSW
Both models correct									
LR	0.109	0.413	0.281	0.116	0.413	0.281	0.011	0.013	0.012
CBPS(over)	1.317	0.458	0.344	1.517	0.458	0.344	0.012	0.012	0.008
CBPS(exact)	0.002	0.440	0.309	0.012	0.440	0.309	0.012	0.013	0.016
RF	4.531	2.116	2.105	4.765	2.116	2.105	0.015	0.015	0.010
Xgboost	9.524	2.128	2.523	9.906	2.128	2.523	0.014	0.025	0.029
Only propensity score model correct									
LR	-	-	-	-	-	-	0.046	0.046	0.032
CBPS(over)	-	-	-	-	-	-	0.532	0.057	0.067
CBPS(exact)	-	-	-	-	-	-	0.017	0.068	0.058
RF	-	-	-	-	-	-	2.196	1.644	1.597
Xgboost	-	-	-	-	-	-	3.350	0.748	1.044
Only outcome model correct									
LR	39.773	5.411	5.340	1.184	5.411	5.340	0.021	0.001	0.002
CBPS(over)	3.054	6.574	6.432	5.968	6.574	6.432	0.013	0.003	0.002
CBPS(exact)	5.870	7.128	7.040	5.881	7.128	7.040	0.012	0.003	0.007
RF	6.532	4.198	4.202	5.833	4.198	4.202	0.012	0.006	0.013
Xgboost	9.973	0.773	0.499	11.454	0.773	0.499	0.012	0.004	0.009
Both models incorrect									
LR	-	-	-	-	-	-	18.186	4.761	4.765
CBPS(over)	-	-	-	-	-	-	7.106	5.016	5.046
CBPS(exact)	-	-	-	-	-	-	6.868	5.104	5.101
RF	-	-	-	-	-	-	2.780	2.592	2.584
Xgboost	-	-	-	-	-	-	4.170	1.217	0.888

Table 2. Variance : N = 1000, iteration 1000

Model for Propensity Score	HT	HT-CSW	HT-FSW	IPW	IPW-CSW	IPW-FSW	DR	DR-CSW	DR-FSW
Both models correct									
LR	37.434	1.537	3.077	2.224	1.537	3.077	0.157	0.164	0.220
CBPS(over)	32.059	1.838	3.699	1.658	1.838	3.699	0.150	0.166	0.206
CBPS(exact)	0.157	1.099	2.568	0.156	1.099	2.568	0.156	0.168	0.223
RF	5.954	1.750	3.379	1.235	1.750	3.379	0.139	0.166	0.199
Xgboost	44.105	9.527	11.991	13.202	9.527	11.991	0.136	0.212	0.279
Propensity score model correct									
LR	-	-	-	-	-	-	2.820	1.372	2.227
CBPS(over)	-	-	-	-	-	-	2.080	1.492	2.260
CBPS(exact)	-	-	-	-	-	-	1.824	1.076	1.857
RF	-	-	-	-	-	-	1.003	1.045	1.809
Xgboost	-	-	-	-	-	-	1.853	2.951	4.520
Outcome model correct									
LR	23414.292	2.455	3.720	118.030	2.455	3.720	4.513	0.173	0.212
CBPS(over)	52.207	3.951	5.823	3.523	3.951	5.823	0.176	0.172	0.214
CBPS(exact)	3.178	5.080	6.925	3.223	5.080	6.925	0.172	0.176	0.213
RF	6.097	1.871	3.249	1.488	1.871	3.249	0.139	0.165	0.201
Xgboost	48.657	9.354	11.383	11.481	9.354	11.383	0.136	0.204	0.271
Both models incorrect									
LR	-	-	-	-	-	-	4985.985	2.249	2.757
CBPS(over)	-	-	-	-	-	-	6.497	2.608	3.193
CBPS(exact)	-	-	-	-	-	-	4.556	2.713	3.254
RF	-	-	-	-	-	-	1.270	1.367	2.145
Xgboost	-	-	-	-	-	-	1.868	2.847	4.059

Table 3. Root of Mean Squared Error : $N = 1000$, iteration 1000

Model for Propensity Score	HT	HT-CSW	HT-FSW	IPW	IPW-CSW	IPW-FSW	DR	DR-CSW	DR-FSW
Both models correct									
LR	6.119	1.307	1.777	1.496	1.307	1.777	0.396	0.405	0.469
CBPS(over)	5.813	1.431	1.954	1.990	1.431	1.954	0.387	0.408	0.453
CBPS(exact)	0.396	1.137	1.632	0.395	1.137	1.632	0.395	0.410	0.473
RF	5.146	2.496	2.795	4.893	2.496	2.795	0.373	0.408	0.446
Xgboost	11.611	3.749	4.284	10.552	3.749	4.284	0.369	0.461	0.529
Propensity score model correct									
LR	-	-	-	-	-	-	1.680	1.172	1.493
CBPS(over)	-	-	-	-	-	-	1.537	1.223	1.505
CBPS(exact)	-	-	-	-	-	-	1.351	1.040	1.364
RF	-	-	-	-	-	-	2.414	1.936	2.088
Xgboost	-	-	-	-	-	-	3.616	1.874	2.368
Outcome model correct									
LR	158.102	5.633	5.677	10.928	5.633	5.677	2.124	0.416	0.460
CBPS(over)	7.844	6.868	6.870	6.256	6.868	6.870	0.420	0.414	0.463
CBPS(exact)	6.135	7.476	7.516	6.148	7.476	7.516	0.415	0.420	0.461
RF	6.983	4.415	4.573	5.960	4.415	4.573	0.373	0.407	0.448
Xgboost	12.170	3.155	3.411	11.945	3.155	3.411	0.368	0.452	0.520
Both models incorrect									
LR	-	-	-	-	-	-	72.916	4.992	5.046
CBPS(over)	-	-	-	-	-	-	7.549	5.270	5.353
CBPS(exact)	-	-	-	-	-	-	7.193	5.363	5.411
RF	-	-	-	-	-	-	3.000	2.843	2.970
Xgboost	-	-	-	-	-	-	4.388	2.080	2.202

Copyrights