Journal of Machine Learning Research 13 (2012) 2955-2994 Submitted 7/11; Revised 6/12; Published 10/12

# Facilitating Score and Causal Inference Trees for Large Observational Studies

**Xiaogang Su** XGSU@UAB.EDU
*School of Nursing*
*University of Alabama at Birmingham*
*1720 2nd Ave S*
*Birmingham, AL 35294, USA*

**Joseph Kang** JOSEPH-KANG@NORTHWESTERN.EDU
*Department of Preventive Medicine*
*Northwestern University*
*680 N. Lake Shore, Suite 1410*
*Chicago, IL 60611, USA*

**Juanjuan Fan** JJFAN@SCIENCES.SDSU.EDU
**Richard A. Levine** RALEVINE@SCIENCES.SDSU.EDU
*Department of Mathematics and Statistics*
*San Diego State University*
*5500 Campanile Drive, GMCS 415*
*San Diego, CA 92182, USA*

**Xin Yan** XIN.YAN@UCF.EDU
*Department of Statistics*
*University of Central Florida*
*4000 Central Florida Blvd*
*Orlando, FL 32816, USA*

**Editor:** Peter Spirtes

## Abstract

Assessing treatment effects in observational studies is a multifaceted problem that not only involves heterogeneous mechanisms of how the treatment or cause is exposed to subjects, known as propensity, but also differential causal effects across sub-populations. We introduce a concept termed the facilitating score to account for both the confounding and interacting impacts of covariates on the treatment effect. Several approaches for estimating the facilitating score are discussed. In particular, we put forward a machine learning method, called causal inference tree (CIT), to provide a piecewise constant approximation of the facilitating score. With interpretable rules, CIT splits data in such a way that both the propensity and the treatment effect become more homogeneous within each resultant partition. Causal inference at different levels can be made on the basis of CIT. Together with an aggregated grouping procedure, CIT stratifies data into strata where causal effects can be conveniently assessed within each. Besides, a feasible way of predicting individual causal effects (ICE) is made available by aggregating ensemble CIT models. Both the stratified results and the estimated ICE provide an assessment of heterogeneity of causal effects and can be integrated for estimating the average causal effect (ACE). Mean square consistency of CIT is also established. We evaluate the performance of proposed methods with simulations and illustrate their use with the NSW data in Dehejia and Wahba (1999) where the objective is to assess the impact of

a labor training program, the National Supported Work (NSW) demonstration, on post-intervention earnings.

**Keywords:** CART, causal inference, confounding, interaction, observational study, personalized medicine, recursive partitioning

## 1. Introduction

Comparative studies that involve evaluation of the effect of an investigational treatment or a putative cause on an outcome variable are fundamental in many application fields. The data may come from either a designed experiment or an observational study. Regardless of the data sources, two major issues exist when assessing the treatment effect: confounding and interaction effects of covariates.

A confounding variable or confounder is an extraneous covariate that relates to both the treatment and the response and hence influences the treatment effect estimation. Controlling or adjusting for confounders can be done in either design or analysis. In designed experiments, randomization, matching, cohort restriction, and stratification are commonly-used ways to effectively control for confounding variables. However, observational studies are often the only available choice due to ethical or practical considerations. Causal inference with observational data is particularly challenging. The main obstacle is the nonrandom treatment assignment mechanism, in which the subjects select a treatment that they believe best serve their interests or are exposed to a treatment according to individual traits. As a result, systematic imbalance or heterogeneity may exist between individuals in the treated group and those in the control group. Thus it is crucial to control for confounders in the analysis stage of such data. Common approaches include analysis of covariance (ANCOVA), propensity score methods (Rosenbaum and Rubin, 1983), and directed acyclic graphs (DAGs; Pearl 2000 and Spirtes, Glymour, and Scheines 2001). Even with randomized experimental data, covariate imbalance can also be revealed when examining data in a multivariate manner. Consider a hypothetical example where $m$ older women and $m$ younger men are assigned to the treated group while $m$ older men and $m$ younger women are assigned to the control group. The data appear to be perfectly balanced in terms of either age or gender, despite the perfect imbalance at their combination levels. When the dimension of covariates gets high, each experimental unit essentially represents an unique individual that is not replicable, which makes randomization less relevant. This partially explains why covariate adjustment is practiced even with randomized experimental data. Associated with variable selection issues, additional challenges present themselves in the form of over- or under- adjustment when confounders are incorrectly identified. For example, under-adjustment occurs when an important confounder is uncollected in the data or excluded from the model. On the other hand, some intermediary outcome variables, often referred to effect-mediators, are important in understanding the mechanism how and why the treatment becomes effective. As an example of over-adjustment, the treatment effect would be under-estimated when a mediator is mistakenly considered as a confounder and included in the model for adjustment. Over-adjustment also may occur when controlling for a collider that correlates with both the treatment and the outcome via an '$M$-diagram' (Greenland, 2003).

In terms of influence of covariates on treatment effect assessment, another equally important issue is interaction, also known as effect modification or effect moderation (see, e.g., VanderWeele and Robins 2007 and VanderWeele 2009), which is concerned with differential treatment effects at different levels or values of covariates. An effect modifier is a covariate that interacts with the treatment and changes the direction and/or degree of its causal effect on the outcome. Existence of

interaction complicates model interpretation. Detection of interaction is challenging. While interactions are mostly formulated via cross-product terms in a linear model and restricted to be of the first- or second-order, complex nonlinear or higher-order interactions may exist in reality. It is also important to distinguish between qualitative interactions and quantitative interactions. Qualitative interaction (Gail and Simon, 1985) occurs when there is a directional change in terms of treatment preference, a cause of greater concern to practitioners. Closely related to treatment-by-covariate interactions, subgroup analysis (see, e.g., Lagakos 2006) is an integral part in the analysis of clinical trials. Practitioners and regulatory agencies are keen to know if there are subgroups of trial participants who are more or less likely to be helped or harmed by the intervention under investigation. Subgroup analysis helps explore the heterogeneity of the treatment effect across sub-populations and extract the maximum amount of information from the available data. On the other hand, subgroup analysis is subject to malpractice owing to difficulties in subgroup determination, multiple testings, and lack of power. The new stimulating concept of personalized medicine or personalized treatments (see, e.g., Jain 2009) is intended to refine the traditional medical decisions by capitalizing on results of subgroup analysis or the knowledge of individualized treatment effects. Nevertheless, sorting out differential causal effects often entails large data that are collected at post-trial periods, for example, the Medicare or Medicaid databases.

Assessments of confounding and interaction intervene with each other. First of all, confounding emerges as one primary issue in the assessment of the main effect of treatment, also known as the average causal effect (ACE). However, ACE implicitly assumes homogeneity or unimportant heterogeneity of causal effects. When strong treatment-by-covariate interaction exists, ACE may become less practically useful. This is the case especially when the interaction is qualitative. Suppose, for example, that the treatment effect is $\delta$ for half of the data (say, males) and $-\delta$ for the other half (say, females), both having important scientific implications. The ACE in this case is null. When solely based on ACE, one would arrive at the misleading conclusion that the treatment does not have an effect. On the other hand, when the estimation bias caused by inadequately handled confounders gets overwhelming, it may be disguised as differential treatment effects. We shall illustrate more on this point later with simulation in Section 4. Therefore, it is crucial to have both confounding and interaction well addressed in comparative analysis.

Rubin's causal model (Rubin, 1974, 1977, 1978, 2005) provides a general framework for making these assessments, within which the treatment effect is finely calibrated at three different hierarchical levels (i.e., unit, subpopulation, and population) using a counterfactual model and the concept of potential outcomes. In this article, causal inference is explicitly reformulated as a predictive modeling problem within the framework of Rubin's causal model. To approach, we introduce a concept, termed facilitating score, to address both the confounding and interacting impact of extraneous variables on causal inference. Conditional on the facilitating score, homogeneity can be achieved in both the assignment mechanism and and the effect of the treatment. Then we put forward a causal inference tree (CIT) procedure, to approximate the facilitating score with a piecewise constant function. CIT recursively splits data into disjoint groups in such a way that both treatment assignment mechanisms and the treatment effects become more homogeneous within each group. On the basis of CIT, a group of recursive partitioning methods are devised to make causal inference at different levels.

The remainder of this paper is arranged in the following manner. In Section 2, following an outline of Rubin's causal inference framework, the concept of facilitating score is introduced and methods for estimating the facilitating score are discussed. Section 3 presents the CIT methodology

in detail. Section 4 contains simulation studies that are designed to investigate the performance of CIT. An illustration is provided via a real data example in Section 5. In Section 6, we extend the results to situations where the treatment variable is ordinal or nominal. Section 7 ends the article with a brief discussion.

## 2. Facilitating Scores

We first review Rubin's causal models, then we introduce the facilitating score concept and discuss methods for estimating the facilitating score.

### 2.1 Causal Inference

In Rubin's causal model (Rubin, 1974, 1977, 1978, 2005), a fine calibration of treatment effect is facilitated by a comparison between the observed outcome on an individual or unit and the potential outcome if the individual had been assigned to the counterfactual treatment group. Adopting his notations, let $\Omega = \{\omega\}$ be a finite population with $N$ units, endowed with a probability measure $P$ that places uniform mass $1/N$ on each unit. Let $T = T(\omega)$ be a binary treatment assignment variable with value 1 if unit $\omega$ receives the putative treatment and 0 otherwise. While the term 'treatment assignment' or 'selection' is best suitable for designed experiments, we shall use it throughout this article. In addition, let $\mathbf{X} = \mathbf{X}(\omega)$ be a $p$-dimensional vector of measured covariates for unit $\omega$.

Let $Y_0 = Y_0(\omega)$ be the response that would have been observed if unit $\omega$ were assigned to the control group and let $Y_1 = Y_1(\omega)$ be the response that would have been observed if unit $\omega$ received the treatment. These two variables are called *potential outcomes* (Neyman, 1923). In reality, either $Y_0(\omega)$ or $Y_1(\omega)$, but not both, can actually be observed depending on the value of $T(\omega)$, an inherent fact called the *fundamental problem of causal inference* (Holland, 1986). Thus the observed outcome is

$$Y(\omega) = \{1 - T(\omega)\}Y_0(\omega) + T(\omega)Y_1(\omega).$$

Throughout this paper, we consider random sampling from $\Omega$ so that $\{\omega_1, ..., \omega_n\}$ forms an independent and identically distributed (iid) sample of size $n$. The available data $\{(y_i, t_i, \mathbf{x}_i) = (y(\omega_i), t(\omega_i), \mathbf{x}(\omega_i)) : i = 1, \ldots, n\}$ consist of $n$ realizations of $Y$, $T$, and $\mathbf{X}$. For the sake of simplicity, we sometimes omit unit $\omega$ from the notations.

Causal inference is concerned with the comparison of the two potential outcomes via the observed data. Holland and Rubin (1988) distinguished three levels of causal inferences: unit level, subpopulation level, and population level. The lowest level of causal inference is a comparison of $Y_0(\omega)$ and $Y_1(\omega)$, typically the difference $Y_1(\omega) - Y_0(\omega)$, for each unit $\omega$. Subpopulations can be formed by restricting the values of covariates to a partition of $\Omega$. The causal effect in a subpopulation $\{\omega : \mathbf{X}(\omega) \in B\}$ is $E(Y_1|\mathbf{X} \in B) - E(Y_0|\mathbf{X} \in B)$ for some Borel set $B$ in the predictor space $\mathbb{X}$. The average causal effect (ACE) over the entire population $\Omega$ is $E(Y_1) - E(Y_0)$. These three levels form a hierarchy of causal inference in decreasing order of strength, in the sense that knowledge of upper-level causal inferences can be inferred from that of lowered-level causal inferences, but not vice versa. A preponderance of the literature in causal inference is centered on schemes for making the population-level inference or estimating ACE under various scenarios.

Rosenbaum and Rubin (1983) introduced the concept of balancing score to tackle the confounding issue in causal inference. A balancing score $b(\mathbf{x})$ accounts for the dependence between $\mathbf{X}$ and

treatment assignment or selection $T$; that is

$$\mathbf{X} \perp\!\!\!\perp T \,|\, b(\mathbf{X}).$$

Treated and untreated subjects sharing the same balancing score tend to have the same distribution of covariates. Various covariate adjustment techniques implicitly adjust for an estimated scalar balancing score. They showed that the propensity score

$$e(\mathbf{x}) = P(T = 1|\mathbf{X} = \mathbf{x}),$$

which is defined as the conditional probability of assignment to the treated group given the measured covariates $\mathbf{X}$, is the coarsest balancing score. Namely, $b(\mathbf{x})$ is a balancing score if and only if $b(\mathbf{x})$ is finer than $e(\mathbf{x})$, that is, $e(\mathbf{x})$ is a function of $b(\mathbf{x})$.

Propensity score based matching, stratification (or subclassification), and adjustment have been extensively used to balance the discrepancy in covariates between the treatment groups in the assessment of ACE. In propensity score analysis, the assumption of *strong ignorability* plays a pivotal role. Similar to that of missing at random (MAR) in the missing data literature (Rubin, 1976), this assumption states that $P(T|\mathbf{X}, Y_0, Y_1) = P(T|\mathbf{X})$ or,

$$T \perp\!\!\!\perp (Y_0, Y_1)|\mathbf{X}.$$

It is possible that strong ignorability is violated even there are no unmeasured variables that are direct causes of any pair of measured variables. See, for example, Greenland (2003) for more discussions. It is worth noting that this assumption does not imply that $T \perp\!\!\!\perp Y \,|\, \mathbf{X}$. To illustrate, consider a simple example where the causal effect at the unit level is constant, namely, $Y_1(\omega) - Y_0(\omega) = \delta$ for any $\omega$. Suppose that $Y_0 = f(\mathbf{X}) + \varepsilon$ and $Y_1 = f(\mathbf{X}) + \delta + \varepsilon$, where $\varepsilon \perp\!\!\!\perp \mathbf{X}$ is the error term. It follows that $Y = \delta T + f(\mathbf{X}) + \varepsilon$. The ignorability assumption amounts to $\varepsilon \perp\!\!\!\perp T \,|\, \mathbf{X}$, which, by no means, implies $Y \perp\!\!\!\perp T \,|\, \mathbf{X}$.

Under this assumption of strong ignorability, Rosenbaum and Rubin (1983) established that $(Y_1, Y_0) \perp\!\!\!\perp T|b(\mathbf{X})$ when $0 < e(\mathbf{X}) < 1$.It follows that

$$E(Y_1|b(\mathbf{X}), T = 1) - E(Y_0|b(\mathbf{X}), T = 0) = E(Y_1|b(\mathbf{X})) - E(Y_0|b(\mathbf{X})). \tag{1}$$

Therefore, the population-level causal interpretation may be achieved by averaging over the distribution of $b(\mathbf{X})$,

$$E(Y_1 - Y_0) = E_{b(\mathbf{X})}\{E(Y_1|b(\mathbf{X})) - E(Y_0|b(\mathbf{X}))\}. \tag{2}$$

Equations (1) and (2) provide the basis for propensity score based methods.

### 2.2 Facilitating Score

Parallel to confounding, interaction is concerned with differential causal effects among units or subpopulations. It is important to note that both Equation (1) and (2) involve a reduction of hierarchy in causal inference, where individual-level inferences are integrated to make subpopulation-level inferences on $\Omega_b = \{\omega : b(\mathbf{X}(\omega)) = b\}$ or sub-population level inferences are reduced to the population-level inference on $\Omega$. Such a reduction may not be taken for granted, because it implicitly assumes homogenous lower-level causal effects. Specifically, if substantial differences in causal effects are present at a lower level of inference, then transition to an upper-level inference may not be plausible

and conclusions based on upper-level causal effects can be misleading. This can be particularly problematic when qualitative interactions exist.

To gain insight, note that with balancing score $b(\mathbf{X})$,

$$\mathbf{X} \not\perp\!\!\!\perp (Y_0, Y_1) \,|\, b(\mathbf{X}).$$

As a result, $\delta_b(\mathbf{X}) = E(Y_1|b(\mathbf{X}) = b) - E(Y_0|b(\mathbf{X}) = b)$ in (2) is not a constant, but a function of $\mathbf{X}$ within the subpopulation $\Omega_b$. If $\delta_b(\mathbf{X})$ varies substantially with $\mathbf{X}$, we say that a treatment-by-covariate interaction exists. In this case, the overall causal effect $\delta_b$ in $\Omega_b$ becomes less pertinent as it implicitly assumes that $\delta_b(\mathbf{X})$ can be reduced to a constant $\delta_b$. A fine delineation of treatment effect $\delta_b(\mathbf{X})$ at the individual level is desirable in the efforts of advancing personalized medicines. Even if estimating $\delta_b$ is of interest, it cannot be summarized by direct comparison of treatment means. Instead, it should be obtained by integrating over the distribution of $\mathbf{X}$ in $\Omega_b$, that is, $\delta_b = \int_{\Omega_b} \delta_b(\mathbf{x})\, d\mu(\mathbf{x})$. Direct comparison of treatment means in $\Omega_b$ makes another implicit assumption that, within $\Omega_b$, $\mathbf{X}$ follows a uniform distribution. The same problem remains when using (2) for ACE estimation.

It is therefore critical to take both heterogeneous treatment assignment mechanisms and differential treatment effects into consideration when assessing the treatment effects. We introduce a concept termed facilitating score to address these two issues simultaneously.

**Definition 1** *A facilitating score* $\mathbf{a}_0(\mathbf{X})$ *is a* $q_0$*-dimensional (*$0 < q_0 \leq p$*) function of* $\mathbf{X}$ *such that* $\mathbf{X} \perp\!\!\!\perp (Y_0, Y_1, T) \,|\, \mathbf{a}_0(\mathbf{X})$.

In this definition, the joint independence between $\mathbf{X}$ and $(Y_0, Y_1, T)$ given $\mathbf{a}_0(\mathbf{X})$ can be relaxed as two marginal independence conditions: $\mathbf{X} \perp\!\!\!\perp T \,|\, \mathbf{a}_0(\mathbf{X})$ and $\mathbf{X} \perp\!\!\!\perp (Y_0, Y_1) \,|\, \mathbf{a}_0(\mathbf{X})$, which separately address the confounding effect and the interacting effect of $\mathbf{X}$. But, if strong ignorability, that is, $T \perp\!\!\!\perp (Y_0, Y_1) \,|\, \mathbf{X}$, is further assumed, it follows that $T \perp\!\!\!\perp (Y_0, Y_1) \,|\, \mathbf{a}_0(\mathbf{X})$ and hence the marginal independence implies the joint independence as well. Existence of $\mathbf{a}_0(\mathbf{X})$ is guaranteed, since $\mathbf{X}$ itself can be regarded as a facilitating score.

Nevertheless, Definition 1 places strong requirements on $\mathbf{a_0}(\mathbf{X})$. Estimating the facilitating score essentially involves jointly modeling $\{Y_0, Y_1, T\}$ conditional on $\mathbf{X}$, which is unworkable since $(Y_0, Y_1)$ can not be observed at the same time. To get around this difficulty, we next consider a weaker definition of facilitating score that is more practically useful.

**Definition 2** *A weak facilitating score* $\mathbf{a}(\mathbf{X})$ *is a q-dimensional (*$0 < q \leq p$*) function of* $\mathbf{X}$ *such that (i)* $\mathbf{X} \perp\!\!\!\perp T \,|\, \mathbf{a}(\mathbf{X})$ *and (ii)* $E(Y_1 - Y_0|\mathbf{X}) = E(Y_1 - Y_0|\mathbf{a}(\mathbf{X}))$.

By condition (i), a weak facilitating score $\mathbf{a}(\mathbf{X})$ must be a balancing score; by condition (ii), any effect moderation owing to $\mathbf{X}$ can be fully represented by $\mathbf{a}(\mathbf{X})$. Condition (ii) is equivalent to saying that $E(Y_1 - Y_0|\mathbf{a}(\mathbf{X}) = \mathbf{a})$ is independent of $\mathbf{X}$. However, this does not necessarily imply that

$$E(Y_1|\mathbf{X}) = E(Y_1|\mathbf{a}(\mathbf{X})) \;\text{ and }\; E(Y_0|\mathbf{X}) = E(Y_0|\mathbf{a}(\mathbf{X})). \tag{3}$$

There could exist a common function $g(\mathbf{X})$ that has been cancelled out in Condition (ii). Namely,

$$g(\mathbf{X}) = E(Y_1|\mathbf{X}) - E(Y_1|\mathbf{a}(\mathbf{X})) = E(Y_0|\mathbf{X}) - E(Y_0|\mathbf{a}(\mathbf{X})).$$

A facility score must also be a weak facilitating score, but not vice versa. We use the term 'facilitating' because conditioning on $\mathbf{a}(\mathbf{X})$ helps facilitate causal inference, in the sense that causal inference within the sub-population $\Omega_{\mathbf{a}} = \{\omega : \mathbf{a}(\mathbf{X}(\omega)) = \mathbf{a}\}$ can be conveniently obtained via direct comparison of sample mean responses. This is because both propensity and the treatment effect $\delta_{\mathbf{a}}$ become constant within $\Omega_{\mathbf{a}}$.

Since the propensity $e(\mathbf{X})$ is the coarsest balancing score, it follows that $e(\mathbf{X}) = e$ in $\Omega_{\mathbf{a}}$. In some scenarios, $e(\mathbf{X})$ is explicitly a separate component of $\mathbf{a}(\mathbf{X})$, as exemplified by the parametric approach outlined in Section 2.3; but this is not necessarily true in general, as exemplified by the semi-parametric approach outlined in the same section. In terms of stratification, $\Omega_{\mathbf{a}}$ provides additional refinements of $\Omega_e = \{\omega : e(\mathbf{X}(\omega)) = e\}$ in order to achieve homogeneous within-stratum treatment effects.

**Theorem 3** *Suppose that the conditional joint density of $(Y,T)$ given $\mathbf{X}$, $f_{Y,T|\mathbf{X}}(Y,T|\mathbf{X})$, can be written as $f_{Y,T|\mathbf{X}}(Y,T|\mathbf{X}) = g(Y,T,\mathbf{h}(\mathbf{X}))$ for some function $g(\cdot)$. In other words, $(Y,T) \perp\!\!\!\perp \mathbf{X}\,|\,\mathbf{h}(\mathbf{X})$. Assuming that treatment assignment is strongly ignorable, $\mathbf{h}(\mathbf{X})$ is a weak facilitating score when $0 < e(\mathbf{X}) < 1$.*

We defer the proof of Theorem 3 to Appendix A, where it is established as a special case of a more general result in Theorem 7. Theorem 3 basically states that both confounding and interacting effect of $\mathbf{X}$ on causal inference with the potential outcomes $(Y_1, Y_0)$ can be handled by working with the observed data $(Y,T,\mathbf{X})$. More specifically, if the joint density of $(Y,T)$ given $\mathbf{X}$ can be accounted for by a vector-valued function $\mathbf{h}(\mathbf{X})$, that is, $(Y,T) \perp\!\!\!\perp \mathbf{X}\,|\,\mathbf{h}(\mathbf{X})$, then $\mathbf{h}(\mathbf{X})$ is a weak facilitating score. Besides, it can be shown that Equation (3) holds for $\mathbf{h}(\mathbf{X})$, that is, $E(Y_1|\mathbf{X}) = E(Y_1|\mathbf{h}(\mathbf{X}))$ and $E(Y_0|\mathbf{X}) = E(Y_0|\mathbf{h}(\mathbf{X}))$. This condition will be relaxed in Section 2.3.

Estimation of $\mathbf{h}(\mathbf{X})$ involves modeling the joint distribution of $(Y,T)$ given $\mathbf{X}$. Searching for a satisfactory $\mathbf{h}(\mathbf{X})$ is not an easy task; we have to look for approximate solutions. On the other hand, it is no longer unattainable as the involved elements $(Y,T,\mathbf{X})$ are all observed. Although $\mathbf{h}(\mathbf{X})$ is generally set as vector-valued, its dimension should be small in order to be practically useful.

### 2.3 Estimating the Facilitating Score

We shall discuss three proposals for finding useful approximations of $\mathbf{h}(\mathbf{X})$, which are parametric, semiparametric, and nonparametric in nature, respectively. While they are all methodologically interesting, we deem the nonparametric approach most practically useful.

The first method is parametric. Consider

$$\begin{aligned} f(Y,T|\mathbf{X}) &= f(Y|T,\mathbf{X}) \cdot f(T|\mathbf{X}) \\ &= \{f(Y|T=1,\mathbf{X})\}^T \cdot \{f(Y|T=0,\mathbf{X})\}^{1-T} \cdot f(T|\mathbf{X}) \end{aligned} \tag{4}$$

by Bayes' rule. With a parametric approach, we assume a model for each of the terms in (4): propensity score model for $f(T|\mathbf{X})$ and outcome regression models for $f(Y|T=0,\mathbf{X})$ and $f(Y|T=1,\mathbf{X})$. It is convenient to model $T|\mathbf{X}$ with logistic regression and model $Y|(T,\mathbf{X})$ with Gaussian linear regression so that

$$f(Y,T|\mathbf{X}) = \frac{1}{\sigma}\phi\left(\frac{Y-\mu}{\sigma}\right) \cdot \{\pi(h_3(\mathbf{X}))\}^T \{1-\pi(h_3(\mathbf{X}))\}^{1-T}, \tag{5}$$

where $\sigma$ is the constant error variance;

$$\mu = E(Y|T,\mathbf{X}) = \gamma_0 + \gamma_1 T + h_1(\mathbf{X}) + T \cdot h_2(\mathbf{X}); \tag{6}$$

$\phi(\cdot)$ is the density function of the standard normal $\mathcal{N}(0,1)$ distribution; and $\pi(x) = \exp(x)/(1+\exp(x))$ is the logistic or expit function.

**Proposition 4** *Suppose that the propensity model can be specified by $e(\mathbf{X}) = e(h_3(\mathbf{X}))$ as in (5) and the conditional mean response given $(T,\mathbf{X})$ is formulated by (6). Under the assumption of strong ignorability, $\mathbf{h}(\mathbf{X}) = (h_2(\mathbf{X}), h_3(\mathbf{X}))^t$ is a weak facilitating score.*

The proof is provided in Appendix B. Proposition 4 says that $h_1(\mathbf{X})$ is not a necessary component of a weak facilitating score. It holds as long as the conditional mean outcome is specified by (6); in other words, normality is not needed either. Besides, note that Equation (3) is not required with this definition of $\mathbf{h}(\mathbf{X})$.

To continue with the parametric approach, linearity is further enforced so that $h_j(\mathbf{X}) = \beta_j^t \mathbf{X}_j$ for $j = 1,2,3$, where $\mathbf{X}_j$ contains selected components of $\mathbf{X}$. The involved parameters $\theta = \{\beta, \gamma, \sigma\}$ can be estimated via maximum likelihood in a straightforward manner. The likelihood function is

$$L(\theta) = \prod_{i=1}^{n} \frac{1}{\sigma} \phi\left(\frac{y_i - \mu_i}{\sigma}\right) \cdot \prod_{i=1}^{n} \left\{\pi(\beta_3^t \mathbf{x}_i)\right\}^{t_i} \left\{1 - \pi(\beta_3^t \mathbf{x}_i)\right\}^{1-t_i} = L_1 \cdot L_2. \tag{7}$$

Clearly there is a variable selection issue involved. Note that $(\beta_1, \beta_2)$ are involved only in $L_1$ for the outcome regression model while $\beta_3$ is involved only in $L_2$ for the propensity score model. This property not only simplifies the likelihood optimization, but also allows for variable selection to be performed separately for the propensity model and outcome regression models.

With an estimated $\widehat{\mathbf{h}}(\mathbf{x}) = (\hat{\beta}_2^t \mathbf{x}, \hat{\beta}_3^t \mathbf{x})^t$, data can be stratified via combined use of the medians or terciles of its components, similar to propensity score subclassification. While this parametric method provides a feasible approach for stratification, there are several difficulties in practice. First of all, it is a two-step approach. The final results rely on correct model specifications. Moreover, the number of strata has to be rather arbitrarily determined. The fact that $\widehat{\mathbf{h}}(\mathbf{x})$ is vector-valued contributes added difficulties to execution. In particular, as the dimension of $\widehat{\mathbf{h}}(\mathbf{x})$ increases, the number of strata grows precipitously. Even with only two categories induced by each component, there are $2^q$ subclasses for a $q$-dimensional $\widehat{\mathbf{h}}(\mathbf{x})$.

Another intuitive semi-parametric approach to estimate $\mathbf{h}(\mathbf{X})$ is via dimension reduction techniques. In view of $(Y,T) \perp\!\!\!\perp \mathbf{X} | \mathbf{h}(\mathbf{X})$, if it is further assumed that $\mathbf{h}(\mathbf{X})$ is linear in $\mathbf{X}$ so that $\mathbf{h}(\mathbf{X}) = \mathbf{B}\mathbf{X}$, then the subspace spanned by columns of $\mathbf{B}$, $\mathbb{S}(\mathbf{B})$, is called the dimension-reduction subspace that accounts for the conditional distribution of $(Y,T)$ given $\mathbf{X}$. Let $\mathbb{S}_{(Y,T)|\mathbf{X}}$ denote the intersection of all dimension-reduction subspaces. Under some regular assumptions, $\mathbb{S}_{(Y,T)|\mathbf{X}}$ is also a subspace, termed the central dimension-reduction subspace or central space. Sliced inverse regression (SIR; Li 1991) and its variants can be used to estimate $\mathbb{S}_{(Y,T)|\mathbf{X}}$. While further research efforts are needed in handling the bivariate response $(Y,T)$, there is no additional conceptual complication involved. For example, one convenient approach is to first introduce $(2S)$ slice indicator variables

$$Z_{st} = I\{(y'_{k-1} < Y \le y'_k) \cap (T = t)\},$$

where $s = 0,1,\ldots,S$; $t = 0,1$; and $\{-\infty = y'_0 < y'_1 < \cdots < y'_S = +\infty\}$ are pre-specified grid points that define $S$ slices for $Y$. Then the sliced regression method (Wang and Xia, 2008) can be applied

to estimate the central mean space of $\mathbf{Z} = (Z_{st})$, which approximates the central space $\mathbb{S}_{(Y,T)|\mathbf{X}}$. Nevertheless, the same above-mentioned difficulties as in the parametric approach remain when it comes to stratification on the estimated linear facilitating scores.

In the next section, we consider yet another recursive partitioning based nonparametric alternative, which seems to provide a more satisfactory solution to the problem. Hereafter, we refer to this method as CIT for causal inference tree. CIT combines estimation of $\mathbf{h}(\mathbf{x})$ and data stratification into one step. On the basis of CIT, we devise methods for making causal inference at different levels.

## 3. Causal Inference via Recursive Partitioning

Tree-based methods (Morgan and Sonquist 1963 and Breiman et al. 1984) approximate the underlying function of interest with piecewise constants by recursively partitioning the predictor space. At the same time, a tree structure offers natural grouping of data with easily interpretable splitting rules. With an automated algorithmic approach, CIT seeks disjoint groups that have homogeneous joint density of $(Y,T)$ within each. The resultant grouping rules, which are induced by binary splits on the covariates $\mathbf{X}$, are meaningfully interpretable, implying a nonparametric estimation of the facilitating score.

In this section, we first follow the CART (Breiman et al., 1984) convention to construct one single CIT, which consists of three steps: growing a large tree and selecting the optimal subtree via pruning and cross validation. On the basis of CIT, methods for causal inference at different levels are then developed. CIT itself provides a natural stratification of data for subpopulation inference. An aggregated grouping method is introduced in order to enhance its performance. Conditional inference at the individual unit level can also be obtained by combining results from ensemble CIT models. Both stratified and individualized causal effect estimates can help depict variations in propensity and treatment effects and make available a natural evaluation of the plausibility of treatment comparability and ACE assessment. These results can also be integrated for estimating ACE estimates. Finally, we establish the mean square risk consistency of CIT under conditions similar to those in CART (Breiman et al., 1984).

### 3.1 Causal Inference Trees (CIT)

A tree model can be expressed as a graph with connected nodes, each node corresponding to a subset of the data. We use $\mathcal{T}$ as a generic notation for a tree structure and $\tau$ for a node. In tree modeling, the effects of $\mathbf{X}$ are exclusively explained by the splitting rules. To start the tree construction, we consider one single split of data. When restricted to a node $\tau$, the distribution of $(Y,T)$ no longer depends on $\mathbf{X}$, implying a constant propensity and a constant treatment effect. Following decomposition of the joint density $f_\tau(Y,T) = f_\tau(Y|T)\, f_\tau(T)$ within node $\tau$, it is convenient to impose that

$$T \sim \text{Bernoulli}\,(\pi_\tau) \quad \text{and} \quad Y|T \sim \mathcal{N}\left\{\mu = (1-T)\mu_{\tau 0} + T\mu_{\tau 1},\, \sigma_\tau^2\right\}.$$

We would like to comment that recursive partitioning can be viewed as a localized approach with local optimality achieved at each split. In local areas, the model needs not to be complicated and often employs a parametric form. The procedure starts with splits that are built upon something that is relatively simple and then evolves into a comprehensive model by recursively bisecting.

The resultant tree model is nonparametric in nature and relatively robust to local distributional assumptions.

The associated log-likelihood function becomes

$$l_\tau = -\frac{n_\tau}{2}\ln(2\pi\sigma^2) - \frac{\sum_{i\in\tau}(y_i-\mu_i)^2}{2\sigma^2} + n_{\tau 1}\ln\pi_\tau + n_{\tau 0}\ln(1-\pi_\tau)$$

where $\{n_\tau, n_{\tau 0}, n_{\tau 1}\}$ are the total number of observations in node $\tau$, the number of observations in node $\tau$ that are assigned to the control group, and the number of observations in node $\tau$ that are assigned to the treatment group, respectively. Maximum likelihood estimates of the involved parameters are explicitly available: $\hat{\pi}_\tau = n_{\tau 1}/n_\tau$, $\hat{\mu}_{\tau 0} = \bar{y}_{\tau 0}$, $\hat{\mu}_{\tau 1} = \bar{y}_{\tau 1}$, and $\hat{\sigma}^2 = SSE_\tau/n_\tau$, where

$$SSE_\tau = \sum_{\{i\in\tau:\ t_i=1\}}(y_i-\bar{y}_{\tau 1})^2 + \sum_{\{i\in\tau:\ t_i=0\}}(y_i-\bar{y}_{\tau 0})^2,$$

and $\{\bar{y}_{\tau 0}, \bar{y}_{\tau 1}\}$ are the sample average responses of the control and treatment groups in node $\tau$, respectively. Up to a constant, the maximized log-likelihood function in node $\tau$ is

$$\hat{l}_\tau \propto -\frac{n_\tau}{2}\ln(n_\tau \cdot SSE_\tau) + n_{\tau 1}\ln n_{\tau 1} + n_{\tau 0}\ln n_{\tau 0}. \tag{8}$$

Note that we have assumed a mean-shift Gaussian model with the same constant variance for the causal effect. If different variances are considered, the final form of $\hat{l}_\tau$ would be slightly different.

Without loss of generality, we consider binary splits only. When a split $s$ bisects node $\tau$ into the left child node $\tau_L$ and the right child node $\tau_R$, the associated likelihood ratio test statistic is

$$LRT(s) = 2\cdot(\hat{l}_{\tau_L} + \hat{l}_{\tau_R} - \hat{l}_\tau), \tag{9}$$

where the maximized log-likelihood score for nodes $\tau_L$ and $\tau_R$, $\hat{l}_{\tau_L}$ and $\hat{l}_{\tau_R}$, can be obtained in the same manner as $\hat{l}_\tau$ in (8). The $LRT_s$ can be used as the splitting statistic to select the best split. After removing irrelevant components, we have

$$\begin{aligned} LRT(s) \quad \propto \quad & -n_{\tau_L}/2\cdot\ln(n_{\tau_L}SSE_{\tau_L}) - n_{\tau_R}/2\cdot\ln(n_{\tau_R}SSE_{\tau_R}) + \\ & n_{\tau_L 1}\ln n_{\tau_L 1} + n_{\tau_L 0}\ln n_{\tau_L 0} + n_{\tau_R 1}\ln n_{\tau_R 1} + n_{\tau_R 0}\ln n_{\tau_R 0}. \end{aligned}$$

The best split $s^\star$ is the one that yields the maximum $LRT(s)$ among all allowable splits. Accordingly node $\tau$ is split into $\tau_L$ and $\tau_R$ using $s^\star$. Subsequently, a similar procedure is applied to split either of $\tau_L$ and $\tau_R$. We repeat the procedure until some mild stopping rules are satisfied. This process results in a large initial tree, denoted as $\mathcal{T}_0$.

The final tree model is a subtree of $\mathcal{T}_0$. Nevertheless, it is practically infeasible to examine every subtree because the number of subtrees increases rapidly as the number of terminal nodes in $\mathcal{T}_0$ increases. The idea of pruning is to provide a subset of candidate subtrees by iteratively truncating off the 'weakest link' of $\mathcal{T}_0$. There are several pruning algorithms available, including the cost-complexity pruning of CART (Breiman et al., 1984) for trees that are built upon minimizing within-node impurity, the split-complexity pruning of LeBlanc and Crowley (1993) for trees that are built upon maximizing between-node differences, and the AIC pruning of Su, Wang, and Fan (2004) for trees that are built within the maximum likelihood framework. Since CIT is essentially likelihood based, the AIC pruning is adopted for direct use. We shall keep our descriptions concise by referring the reader to appropriate references for greater details.

In the AIC pruning algorithm, the performance of a given tree $\mathcal{T}$ is measured by the Akaike (1974) information criterion:

$$\text{AIC}_{\mathcal{T}} = -2 \cdot \hat{l}_{\mathcal{T}} + \lambda \times (4 \cdot |\widetilde{\mathcal{T}}|)$$

where the associated maximized log-likelihood of $\mathcal{T}$ is

$$\hat{l}_{\mathcal{T}} = \sum_{\tau \in \widetilde{\mathcal{T}}} \hat{l}_{\tau}; \tag{10}$$

$\lambda = 2$ is the penalty parameter for tree complexity; and $|\widetilde{\mathcal{T}}|$ denotes the number of terminal nodes in $\mathcal{T}$, with $\widetilde{\mathcal{T}}$ being the set of all terminal nodes in $\mathcal{T}$ and $|\cdot|$ for cardinality when the argument is a set. Note that each added terminal introduces four more new parameters $\{\pi_\tau, \mu_{\tau 0}, \mu_{\tau 1}, \sigma_\tau\}$. Thus the total number of parameters in tree $\mathcal{T}$ is $4 \cdot |\widetilde{\mathcal{T}}|$. A tree with a smaller AIC is preferable. Alternatively, the Bayesian information criterion (BIC; Schwarz 1978) with $\lambda = \ln(n)$ is another choice in common use. At each step, the algorithm examines all available internal nodes or links in the present tree and truncates the link that results in the subtree with the smallest AIC. The pruning procedure yields a nested sequence of subtrees $\mathcal{T}_0 \succ \mathcal{T}_1 \succ \cdots \mathcal{T}_M$, where $\mathcal{T}_M$ is the null tree structure with root node only and "$\succ$" is read as "has subtree".

The final step of tree size selection entails identifying the optimal tree $\mathcal{T}_\star$ from the subtree sequence. The same AIC or BIC measure can be used for this purpose. However, cross validation is needed to validate $\hat{l}_{\mathcal{T}}$ in Equation (10), which can be done via either the test sample method or resampling methods ($V$-fold cross-validation or bootstrapping), depending on the available sample size. Again, we refer readers to Su, Wang, and Fan (2004) for details.

**Remark** Using the parametric approach in Section 2.3, an alternative splitting statistic can be obtained by maximizing the between-node difference. To split node $\tau$, let $I_s$ denote the indicator function corresponding to a permissible split $s$ of $\tau$. Consider model

$$\log \frac{\Pr(T=1|\mathbf{x})}{\Pr(T=0|\mathbf{x})} = \beta_0 + \beta_1 I_s \text{ and}$$
$$y = \gamma_0 + \gamma_1 T + \gamma_2 I_s + \gamma_3 T \cdot I_s + \sigma\varepsilon \text{ with } \varepsilon \sim \mathcal{N}(0,1). \tag{11}$$

In view of Proposition 4, the Wald test statistic for testing $H_0 : \beta_1 = \gamma_3 = 0$ can be used as the splitting statistic. Since the log-likelihood function is separable for $\beta$ and $\gamma$ as shown in (7), $\text{cov}(\hat{\beta}, \hat{\gamma}) = \mathbf{0}$. After some algebraic simplification, the Wald test statistic is given by

$$\left( \frac{1}{n_{\tau_L 0}} + \frac{1}{n_{\tau_L 1}} + \frac{1}{n_{\tau_R 0}} + \frac{1}{n_{\tau_R 1}} \right)^{-1} \left[ \left( \log \frac{n_{\tau_L 0} n_{\tau_R 1}}{n_{\tau_L 1} n_{\tau_R 0}} \right)^2 + \frac{\{(\bar{y}_{\tau_L 1} - \bar{y}_{\tau_L 0}) - (\bar{y}_{\tau_R 1} - \bar{y}_{\tau_R 0})\}^2}{\hat{\sigma}^2} \right],$$

where $\hat{\sigma}^2 = \left\{ \sum_{i=1}^n y_i^2 - \left( n_{\tau_L 0} \bar{y}^2_{\tau_L 0} + n_{\tau_L 1} \bar{y}^2_{\tau_L 1} + n_{\tau_R 0} \bar{y}^2_{\tau_R 0} + n_{\tau_R 1} \bar{y}^2_{\tau_R 1} \right) \right\} / n$ is the MLE of $\sigma^2$ in model (11).

### 3.2 Aggregated Grouping

Despite easy interpretability, one single tree model is notoriously unstable in the sense that a minor perturbation of the data could result in substantial changes in the final tree structure. In order to get around this problem, we propose an aggregated grouping method to integrate the stratification results from a number of competitive tree models. The key idea of this method is to obtain an $n \times n$

distance or dissimilarity matrix $\mathbf{D}$ with entries that measure how likely each pair of observations is assigned to different strata. Cluster analysis can then be applied for final grouping.

The procedure is described as follows. Let $\mathcal{L}$ denote the whole data set. At each iteration $b$ for $b = 1, \ldots, B$, generate bootstrap sample $\mathcal{L}^{(b)}$ from $\mathcal{L}$. Divide $\mathcal{L}^{(b)}$ into two parts at random with a ratio of 2:1, the learning sample $\mathcal{L}_1^{(b)}$ and the test sample $\mathcal{L}_2^{(b)}$. Using $\mathcal{L}_1^{(b)}$, a large initial CIT $\mathcal{T}_0^{(b)}$ is grown and pruned. With the aid of the test sample $\mathcal{L}_2^{(b)}$, a best-sized tree $\mathcal{T}_\star^{(b)}$ is selected. Let $K_b = |\widetilde{\mathcal{T}}_\star^{(b)}|$ be the number of terminal nodes in $\mathcal{T}_\star^{(b)}$. Then we apply $\mathcal{T}_\star^{(b)}$ to the whole data $\mathcal{L}$ so that each observation in $\mathcal{L}$ falls into one and only one terminal node of $\mathcal{T}_\star^{(b)}$. Next, we define an $n \times n$ pairwise distance matrix $\mathbf{D}_b = \{d_{ii'}\}$ such that

$$d_{ii'} = \begin{cases} 0 & \text{if observations } \{i, i'\} \text{ fall into the same terminal node of } \mathcal{T}_\star^{(b)}; \\ 1 & \text{otherwise,} \end{cases}$$

for $i, i' = 1, \ldots n$. To compute $\mathbf{D}_b$, first obtain an $n \times K_b$ matrix $\mathbf{Z}_b = (z_{ik})$ such that

$$z_{ik} = \begin{cases} 1 & \text{if observation } i \text{ falls into the } k\text{-th terminal node,} \\ 0 & \text{otherwise,} \end{cases} \tag{12}$$

for $i = 1, \ldots, n$ and $k = 1, \ldots, K_b$. It follows that

$$\mathbf{D}_b = \mathbf{Z}_b \mathbf{Z}_b^t. \tag{13}$$

Next, the distance matrices are integrated by averaging over $B$ iterations: $\mathbf{D} = \sum_{b=1}^{B} \mathbf{D}_b / B$. It can be seen that the entries in $\mathbf{D}$ satisfy the triangle inequality and other properties that are required for being a legitimate distance measures. Finally, we can apply a clustering algorithm on $\mathbf{D}$ in order to obtain the final data stratification. The number of clusters $K$ can be either determined by the clustering algorithm itself or preset as the mode of $K_b$'s. Other techniques for exploring distance or proximity matrices can also be applied, such as multidimensional scaling (MDS; Torgerson 1958). The whole procedure is outlined in Algorithm 1.

**Algorithm 1** Pseudo-Codes for Aggregated Grouping

Set $B \leftarrow$ number of repetitions.
**for** $b = 1$ to $B$ **do**
— Generate bootstrap sample $\mathcal{L}^{(b)}$.
— Randomly divide data $\mathcal{L}^{(b)}$ into $\{\mathcal{L}_1^{(b)}, \mathcal{L}_2^{(b)}\}$ with a ratio of 2:1.
— Grow a large CIT $\mathcal{T}_0^{(b)}$ using $\mathcal{L}_1^{(b)}$ and prune.
— Select the best tree $\mathcal{T}_\star^{(b)}$ using $\mathcal{L}_2^{(b)}$. Let $K_b = |\widetilde{\mathcal{T}}_\star^{(b)}|$.
— Apply $\mathcal{T}_\star^{(b)}$ to data $\mathcal{L}$ and compute distance matrix $\mathbf{D}_b = (d_{ii'})$ such that $d_{ii'} = 1$ if observation pair $\{i, i'\}$ falls into different nodes of $\mathcal{T}_\star^{(b)}$ and 0 otherwise.
**end for**
Obtain $\mathbf{D} \leftarrow 1/B \cdot \sum_{b=1}^{B} \mathbf{D}_b$;
Obtain $K \leftarrow \text{mode}\{K_b : b = 1, \ldots, B\}$.
Apply a clustering algorithm on $\mathbf{D}$ with $K$ clusters.

We also suggest an optional alternative for computing the distance matrix $\mathbf{D}_b$, which is motivated by the amalgamation or node merging idea of Ciampi et al. (1988). It is common that non-neighboring terminal nodes in a final tree structure do not show much differences from each other.

This is because similar patterns in treatment assignment and effect may occur to sub-populations with different characteristics. By taking this issue into consideration of the distance matrix $\mathbf{D}_b$ in Algorithm 1, a more effective way of grouping data may be achieved.

To do so, we first obtain a $K_b \times K_b$ pairwise distance matrix $\mathbf{K}_b = \{\kappa\}$ for all the terminal nodes in $\mathcal{T}_\star^{(b)}$, the best-sized tree obtained in the $b$-th iteration. Here, $\kappa = \kappa(\tau, \tau') \geq 0$ denotes the distance between two terminal nodes $\tau, \tau' \in \widetilde{\mathcal{T}}_\star^{(b)}$, which can be defined as the logworth (i.e., the negative logarithm with base 10) of the p-value obtained from a likelihood ratio test in (9) that compares $\tau$ with $\tau'$. That is,

$$\kappa(\tau, \tau') = -\log_{10}(\text{p-value}).$$

The likelihood ratio test can be conducted using all data in $\mathcal{L}$. The smaller the p-value, the larger the difference between $\tau$ and $\tau'$ is, as reflected by a larger value of $\kappa(\tau, \tau')$. Elements in matrix $\mathbf{D}_b$ are then defined by

$$d_{ii'} = \kappa(\tau(i), \tau(i')),$$

where $\tau(i)$ denotes the terminal node into which the $i$-th observation falls. In matrix form, $\mathbf{D}_b$ can be computed as

$$\mathbf{D}_b = \mathbf{Z}_b \mathbf{K}_b \mathbf{Z}_b^t, \tag{14}$$

where $\mathbf{Z}_b$ is given by (12). The $\mathbf{D}_b$ in (13) can be viewed as a special case of (14) with $\mathbf{K}_b = \mathbf{I}_b$.

With modified $\mathbf{D}_b$ in (14), there are two immediate consequences: first, the distances $d_{ii'}$ in $\mathbf{D}$ may not necessarily satisfy the triangle inequality; secondly, the number of final clusters $K$ can no longer be suggested by the best tree sizes. Instead, it has to be determined by the clustering algorithm itself. Recent work on automatic determination of the optimal number of clusters is exemplified by Tibshirani, Walther, and Hastie (2001) and Wang (2010). Both methods are computationally demanding.

Compared to one single CIT, the aggregated grouping produces a more accurate and stable grouping of data. Its results can help evaluate the instability of CIT. However, one drawback is loss of interpretability of the stratification rules.

### 3.3 Summarizing Strata and ACE Estimation

To summarize the final $K$ strata obtained from either one single CIT or the aggregated grouping method, estimated propensity rate $\hat{e}_k$ and the treatment effect $\widehat{\Delta}_k$ can be obtained for each stratum. Such information helps delineate the heterogeneity structures in both assignment mechanisms and effects of the treatment. Strata with extremely low or high propensities may be excluded from causal inference due to lack of comparison basis. One may take a liberal approach when inspecting differential causal effects across $K$ strata. The use of ACE to summarize treatment effects can be tentatively justified unless strong evidence for qualitative interaction exists. This is similar to the common practice in multi-center trials. While the quantitative treatment-by-center interaction is commonly seen, the overall efficacy of an investigational drug can still be established as long as there is no significant directional change in the comparison. An estimate of ACE, $\widehat{\Delta}$ is given by

$$\widehat{\Delta} = \sum_{k=1}^{K} \frac{n_k}{n} \cdot (\bar{y}_{k1} - \bar{y}_{k0}) \tag{15}$$

with sampling variance approximated by

$$\sum_{k=1}^{K} \frac{n_k^2}{n^2} \cdot \left( \frac{s_{k1}^2}{n_{k1}} + \frac{s_{k0}^2}{n_{k0}} \right), \tag{16}$$

where $(\bar{y}_{k1}, s_{k1}^2)$ are the sample mean and variance of observed $Y$'s in the treated group of the $k$-th stratum and similar definitions apply to other quantities. Additional covariate adjustment within each terminal node can be made and alternative stratification estimates of ACE are available, as summarized and discussed in Lunceford and Davidian (2004).

Propensity score stratification or subclassification seeks subpopulations of form $\Omega_e = \{\omega : e(\mathbf{X}) = e\}$, in which homogeneity of treatment effects, however, can not be guaranteed. Direct comparison of the mean responses could give a distorted estimate of the causal effect in $\Omega_e$. Comparatively, CIT and aggregating grouping offer refined stratification so that the causal effect within each resultant stratum $\Omega_{\mathbf{a}}$ can be correctly captured, which consequently offers improved estimation of ACE. Alternatively, one may try to correct the problem with propensity score stratification by applying additional ANCOVA-typed adjustment within each stratum. It is important to note that ANCOVA does no help with this correction, unless effect modification is incorporated into the model by allowing for treatment-by-covariate interaction terms. This approach would consist of two steps. In the first step, a number of strata are obtained by stratifying propensity scores. In the second step, an extended ANCOVA model that allows for interactions is fit within each stratum. We may adopt an approach explained by Aiken and West (1991) in order to make the overall causal effect in $\Omega_e$ appear as a regression coefficient. This approach fits a linear model of form

$$E(Y_i|T_i, \mathbf{X}_i) = \beta_0 + \delta_e T_i + \beta^t \mathbf{x}_i' + T_i \cdot \gamma^t \mathbf{x}_i'. \tag{17}$$

where $\mathbf{x}_i' = \mathbf{x}_i - E(\mathbf{X}|\Omega_e)$ for $i \in \Omega_e$ denotes the centered covariate vector. Then the parameter $\delta_e$ in (17) coincides with the overall causal effect in $\Omega_e$. Finally, the ACE is estimated by combining $\hat{\delta}_e$'s via (15). The CIT stratification roughly resembles this two-step approach described above, yet with additional advantages. First, the facilitating score offers a unified setting where these two steps are naturally combined. Secondly, how to specify interaction terms in (17) remains a dazzling task, which, however, can be efficiently handled with recursive partitioning in CIT.

### 3.4 Predicting Individual Causal Effects (ICE)

With the advent of research with biobanks, molecular profiling technologies have been greatly advanced to allow for collection of a patient's proteomic, genetic, and metabolic information. Given various information collected on a patient, how to customize treatments to the individual's best needs has posed great challenges to players in the field of personalized medicine, including statisticians. A fine delineation of treatment effects plays a critical role in such endeavors.

For this purpose, we define "*individual causal effect*" (ICE) as a conditional expectation $E(Y_1 - Y_0|\mathbf{x})$, given a subject with $\mathbf{X} = \mathbf{x}$. ICE is conceptually different from the unit level causal effect $Y_1(\omega) - Y_0(w)$. Strictly speaking, ICE makes conditional causal inference at the subpopulation level $\{\omega : \mathbf{X}(\omega) = \mathbf{x}\}$. On the other hand, ICE is the best that one could practically do with available information in order to approximate the unit level causal effect. Especially when $\mathbf{X}$ is high-dimensional and has many continuous components, it is likely that each value $\mathbf{x}$ corresponds uniquely to unit $\omega$ with $\mathbf{X}(\omega) = \mathbf{x}$. In what follows, we devise a powerful method via ensemble CITs to predict ICE by borrowing ideas from random forests (Breiman, 2001).

To proceed, we first randomly divide data $\mathcal{L}$ into $V$ folds. To ensure similar proportions of individuals in the treatment groups across all folds, stratified sampling with stratification on $T$ can be used. Let $\mathcal{L}_v$ denote the $v$-th fold and $\mathcal{L}_{(v)} = \mathcal{L} - \mathcal{L}_v$ for the remaining data.

**Algorithm 2** Pseudo-Codes for Predicting Personal Causal Effects (ICE)

Set $V$, $B$, and $m$.
Randomly split data $\mathcal{L}$ into $V$ sets $\{\mathcal{L}_1, \ldots, \mathcal{L}_V\}$, with stratification on $T$.
**for** $v = 1$ to $V$ **do**
  Set $\mathcal{L}_{(v)} = \mathcal{L} - \mathcal{L}_v$.
  **for** $b = 1$ to $B$ **do**
    — Generate bootstrap sample $\mathcal{L}_{(v)}^{(b)}$ from $\mathcal{L}_{(v)}$.
    — Grow a CIT $\mathcal{T}_{(v)}^{(b)}$ using $\mathcal{L}_{(v)}^{(b)}$ without pruning. At each split, only $m$ randomly selected variables are used.
    — Estimate causal effects $\hat{\Delta}_\tau$ and propensity $\hat{e}_\tau$ for each $\tau \in \widetilde{\mathcal{T}}_{(v)}^{(b)}$ based on $\mathcal{L}_{(v)}$.
    — Apply $\mathcal{T}_{(v)}^{(b)}$ to data $\mathcal{L}_v$.
    — Compute $\hat{\Delta}_i^{(b)}$ and $\hat{e}_i^{(b)}$ for each $i \in \mathcal{L}_v \cap \tau$, via $\hat{\Delta}_\tau$ and $\hat{e}_\tau$.
  **end for**
  Obtain $\{\hat{\Delta}_i, \hat{e}_i\}$ as averages of $\{(\hat{\Delta}_i^{(b)}, \hat{e}_i^{(b)}) : b = 1, \ldots, B\}$, for each $i \in \mathcal{L}_v$.
**end for**
Merge estimated $\{\hat{\Delta}_i, \hat{e}_i\}$ into data $\mathcal{L}$ using ID key.
**return** $\mathcal{L}$.

We draw $B$ bootstrap samples from $\mathcal{L}_{(v)}$. With each bootstrap sample $\mathcal{L}_{(v)}^{(b)}$, grow a moderately-sized CIT $\mathcal{T}_{(v)}^{(b)}$ without pruning. When constructing $\mathcal{T}_{(v)}^{(b)}$, we adapt the approach in random forests (Breiman, 2001), where only $m$ randomly selected variables and their associated cutoff points are evaluated at each split. This tactic helps improve the predictive performance by de-correlating the tree models in the random forests. For each terminal node $\tau \in \widetilde{\mathcal{T}}_{(v)}^{(b)}$, estimates of the causal effect and propensity,

$$\hat{\Delta}_\tau = \bar{y}_{\tau 1} - \bar{y}_{\tau 0} \quad \text{and} \quad \hat{e}_\tau = n_{\tau 1}/n_\tau,$$

are computed using data in $\mathcal{L}_{(v)}$. Then we apply $\widetilde{\mathcal{T}}_{(v)}^{(b)}$ to $\mathcal{L}_{(v)}$ and predict the ICE $\hat{\Delta}_i^{(b)}$ and propensity score $\hat{e}_i^{(b)}$ for each individual $i \in \mathcal{L}_v$. Specifically,

$$\hat{\Delta}_i^{(b)} = \hat{\Delta}_\tau \quad \text{and} \quad \hat{e}_i^{(b)} = \hat{e}_\tau,$$

if the $i$-th individual falls into the terminal node $\tau$. The final predicted ICE and propensity for the $i$ individual are

$$\hat{\Delta}_i = \frac{1}{B}\sum_{b=1}^{B} \hat{\Delta}_i^{(b)} \quad \text{and} \quad \hat{e}_i = \frac{1}{B}\sum_{b=1}^{B} \hat{e}_i^{(b)}.$$

Their standard errors can also be obtained from the bootstrap repetitions.

The same procedure is repeated for each fold to estimate ICE and propensity scores for all individuals in $\mathcal{L}$. The whole method is described in Algorithm 2. Further exploration can be done with the estimated ICE and propensity scores and some illustrations are given in Section 5. While

we have used a $V$-fold cross-validation approach in Algorithm 2, the method can be directly applied to an independent future sample for predicting ICE. Other features in random forests such as variable importance ranking and partial dependence plots could also be adopted for causal inference.

Some alternative ways of predicting ICE are discussed below. First of all, the standard method for modeling treatment-by-covariates interaction in many application fields is to fit a linear model with first-order cross-product terms, that is,

$$\begin{aligned} y &= \beta_0 + \beta_1 T + \beta_2^t \mathbf{x} + T \cdot \beta_3^t \mathbf{x} + \varepsilon \\ &= \beta_0 + \beta_2^t \mathbf{x} + (\beta_1 + \beta_3^t \mathbf{x}) \cdot T + \varepsilon. \end{aligned} \tag{18}$$

The ICE is formulated as $(\beta_1 + \beta_3^t \mathbf{x})$, which is also linear in $\mathbf{x}$. While this parametric approach is readily available, it relies heavily on linearity and is subject to a greater risk of model misspecification.

Another convenient way for predicting ICE is to use the 'regression estimation' approach, as described in Schafer and Kang (2008). In this approach, we separately fit a predictive model (possibly using machine learning techniques) for $Y_1$ using data in the treated group only and a predictive model for $Y_0$ using data in the untreated group only. Then we apply these models to obtain predicted values $(\hat{y}_{i1}, \hat{y}_{i0})$ for the potential outcomes for every subject in the data. ICE can be estimated as $\hat{\Delta}_i = \hat{y}_{i1} - \hat{y}_{i0}$. Alternatively, the observed response can be used in the calculation so that $\hat{\Delta}_i = y_i - \hat{y}_{i0}$ for the treated group and $\hat{\Delta}_i = \hat{y}_{i1} - y_i$ for the untreated group. Note that this regression estimation method solely involves the outcome models. The underlying rationale is based on the fact that $E(Y_t|\mathbf{X} = \mathbf{x}) = E(Y|\mathbf{X} = \mathbf{x}, T = t)$ for $t = 0, 1$, given strong ignorability and other conditions. However, the prediction across treatment groups heavily involves extrapolation, again due to the imbalance in covariates. When used for ACE estimation, Schafer and Kang (2008) found that it is not among the top performers, but may be possibly improved by incorporating the propensity score into the model.

Estimating ICE also emerges as one intermediate step in some ACE inference procedures including structural nested models introduced by Robins (1989), marginal structural models (see, e.g., Robins 1999), and the targeted maximum likelihood method (see, e.g., van der Laan and Rubin 2006). These procedures are particularly advantageous in dealing with longitudinal observational data where both treatment and covariates are time-varying, but they are also applicable to cross-sectional or 'point treatment' data. Two estimation methods are commonly used in these procedures: the g-computation and the inverse probability of treatment weighting (IPTW). Model (18) is often embedded in either method, for handling effect moderators in IPTW or being used as the $Q$-model in g-computation (see, e.g., Snowden, Rose, and Mortimer 2011) or targeted maximum likelihood (see, e.g., Rosenblum and van der Laan 2011) to model and predict potential outcomes. With g-computation, it is clear that other semiparametric or nonparametric data adaptive methods (as in 'the 'regression estimation' approach) can be flexibly used for predicting potential outcomes for each observation under each possible treatment regimen. See van der Laan, Polley, and Hubbard (2007) and Austin (2012) for examples.

Yet another method for estimate ICE $E(Y_1 - Y_0|\mathbf{X} = \mathbf{x})$ directly is to relax $\mathbf{X} = \mathbf{x}$ to a neighborhood of $\mathbf{x}$, $\mathcal{N}(\mathbf{x})$. Such a neighborhood of $\mathbf{x}$ can be facilitated using either $K$-nearest neighbor (KNN) or, more generally, kernel smoothing. If KNN is used, let $\mathcal{N}_K(\mathbf{x})$ denote the corresponding neighborhood of $\mathbf{x}$. A natural estimate of ICE is given by

$$\frac{\sum_{i:\ \mathbf{x}_i \in \mathcal{N}_K(\mathbf{x})} y_i T_i}{\sum_{i:\ \mathbf{x}_i \in \mathcal{N}_K(\mathbf{x})} T_i} - \frac{\sum_{i:\ \mathbf{x}_i \in \mathcal{N}_K(\mathbf{x})} y_i (1 - T_i)}{\sum_{i:\ \mathbf{x}_i \in \mathcal{N}_K(\mathbf{x})} (1 - T_i)}.$$

This KNN approach assigns weight 1 to $K$ observations within $\mathcal{N}_K(\mathbf{x})$ and weight 0 to others. More generally, we may use kernel smoothing to have weights depending on $\| \mathbf{x}_i - \mathbf{x} \|$ for all data points. To make it more robust to non-random treatment assignment mechanism, it might be possible to incorporate propensity score into the weights as well. While this implementation has not been seen in the literature, it has some promising potentials for its nonparametric nature and deserves further research. On the other hand, a neighborhood defined with high-dimensional data could have poor performance and the computation could be demanding. In addition, interpretation with respect to covariates becomes obscure with nearest neighbor approaches.

Comparatively, the essential ingredient in our ensemble CIT approach is stratified causal estimates within subpopuations $\{\mathbf{x} : \mathbf{a}(\mathbf{x}) = \mathbf{a}\}$, which is intermediary in-between ACE and ICE. We have the convenience to either move forward for ICE with ensemble models or move backward for ACE by integrating stratified results. It is natural to use tree methods for extracting strata. Tree-structured methods are nonparametric in nature and hence more robust to model misspecification. Recursive partitioning excels in efficiently handling interactions and categorical variables and provides meaningful interpretations. Besides, ensemble models usually performs better in predictive modeling. With that being said, a comprehensive comparison study of these alternative approaches in predicting ICE would be interesting for future research.

### 3.5 Consistency

In terms of asymptotic properties of recursive partitioning based estimators, Breiman et al. (1984) provided detailed developments of convergence in $r$th mean and uniform convergence on compacts. Gordon and Olshen (1984) established the almost sure convergence under certain constraints. In this section, consistency of the CIT based causal effect estimator is provided in the light of Breiman et al. (1984).

Let the predictor space $\mathbb{X} \in \mathbb{R}^p$ be Euclidean. A tree structure $\mathcal{T}$ partitions $\mathbb{X}$ into a number of disjoint sets or terminal nodes $\{\tau : \tau \in \widetilde{\mathcal{T}}\}$. Again, $\tau(\mathbf{x})$ denotes the terminal node where $\mathbf{x}$ falls into. Let $d(\cdot)$ be the diameter of a set, that is, $d_n(\tau) = \sup_{\mathbf{x},\mathbf{x}' \in \tau} \| \mathbf{x} - \mathbf{x}' \|$, where $\| \cdot \|$ is the Euclidean norm. With the observed data of size $n$, let $k_n$ be nonnegative constants such that, with probability one,

$$n_\tau \geq k_n \log n \quad \text{for any } \tau \in \widetilde{\mathcal{T}},$$

where, same as before, $n_{\tau 1}$ is used to denote the number of subjects in node $\tau$ that are assigned to the treated group, that is, $n_{\tau 1} = \sum_{i \in \tau} T_i$, and $n_{\tau 0}$ for the control group. Suppose that $\mathbf{a}(\cdot)$ is a weak facilitating score and $\bar{\mathbf{a}}_\tau = \sum_{i \in \tau} \mathbf{a}(\mathbf{x}_i)$ denotes its mean vector in node $\tau$. Let $(Y_1, Y_0, Y, T, \mathbf{x}) \in \tau$ represent a new observation that is independent of current data $\{(y_i, T_i, \mathbf{x}_i) : i = 1, \ldots, n\}$. The following theorem establishes the mean square risk consistency for $(\bar{y}_{\tau 1} - \bar{y}_{\tau 0})$, the causal effect estimate based on direct comparison of sample means in the terminal node $\tau = \tau(\mathbf{x})$.

**Theorem 5** *Suppose that*

$$\max \left\{ E|Y_1|^{2+\varepsilon},\ E|Y_0|^{2+\varepsilon} \right\} \leq M < \infty \text{ for some } \varepsilon > 0 \text{ and } M > 0, \tag{19}$$

$0 < e(\mathbf{x}) < 1$, *and treatment assignment is strongly ignorable. Assume that* $E(Y_1|\mathbf{a})$ *and* $E(Y_0|\mathbf{a})$ *are continuous in* $\mathbf{a}$ *and* $\mathbf{a}(\mathbf{x})$ *is continuous in* $\mathbf{x}$. *Further assume that*

$$\lim_{n \to \infty} k_n = \infty. \tag{20}$$

*and*

$$\lim_{n\to\infty} d_n(\tau) = 0 \tag{21}$$

*in probability. Then*

$$\lim_{n\to\infty} E\,|(\bar{y}_{\tau 1} - \bar{y}_{\tau 0}) - E\{Y_1 - Y_0 \,|\, \mathbf{a}(\mathbf{x}) = \bar{\mathbf{a}}_\tau\}|^2 = 0. \tag{22}$$

The results in Theorem 5 can be improved to $L_r$ convergence for any $r \geq 1$ if we change the assumption (19) to

$$E|Y_1|^{r+\varepsilon} \leq M < \infty \text{ and } E|Y_0|^{r+\varepsilon} \leq M < \infty.$$

This can be immediately seen from the proof provided in Appendix C, where all the arguments we have used hold in $L_r$ spaces. Toth and Eltinge (2010) has recently proved asymptotic design $L_2$ consistency of tree-based estimator when applied to complex survey data, following similar arguments in Gordon and Olshen (1978, 1980). It is worth noting that the Horvitz-Thompson ( 1952) typed estimator via inverse probability weighting has fundamental use in both causal inference with observational data and in estimation the superpopulation mean with stratified survey data.

These convergence results for recursive partitioning are obtained without dependence on the specifics of the algorithm. Unfortunately, no theoretical justifications have been obtained so far for the splitting rules and pruning algorithms (p. 327; Breiman et al. 1984). Moreover, one of key assumptions for consistency requires that the mesh size of $\tau$ goes to 0 when the sample size gets large, as implied by assumption (21). This is an unappealing constrain to practical applications.

## 4. Simulated Experiments

In this section, simulation experiments are performed to first understand and assess CIT and make comparisons with other methods and then investigate how CIT performs under misspecification.

### 4.1 Performance of CIT

In terms of applications of tree methods relevant to treatment effect assessment, there have been two major developments serving different purposes: 1) propensity trees (PT) that estimate the propensity score $e(\mathbf{X})$, as studied by McCaffrey, Ridgeway, and Morral (2004) and Lee, Lessler, and Stuart (2010); and 2) interaction trees (IT) for subgroup analysis (Su et al., 2009). An interaction tree explicitly models the treatment-by-covariates interactions for detecting differential treatment effects. However, this method was developed for experimental data and does not take the non-randomized treatment assignment into account. As we shall demonstrate, failure or inadequacy to account for propensity information may lead to misleading interaction results, in that the superficial difference in treatment effects might have been caused merely by heterogenous treatment selection mechanisms.

We generate data with the following steps.

1. Generate $X_1, \ldots X_5$ independently from Unif(0,1) and create threshold variables $Z_j = 1_{\{X_j \leq 0.5\}}$ for $j = 1, \ldots, 5$.

2. Set $\text{logit}(\pi) = a_0 + a_1 Z_1 + a_2 Z_2$ with $\text{logit}(\pi) = \log\{\pi/(1-\pi)\}$. Generate $T \sim \text{Bernoulli}(\pi)$.

3. Set $\mu = b_0 + b_1 T + b_2 Z_2 + b_3 Z_3 + b_4 Z_4 + b_5 T \cdot Z_4$ and generate $Y \sim \mathcal{N}(\mu, \sigma^2)$ with $\sigma = 1$.

In addition to the response variable $Y$ and treatment indicator $T$, each data set involves five covariates. In the above simulation strategy, covariate $X_1$ is an exposure or treatment predictor involved in the propensity model only, $X_2$ is a confounder that relates to both $T$ and $Y$, $X_3$ is a response predictor or prognostic factor, $X_4$ is an effect-modifier, and $X_5$ is a totally irrelevant covariate. All the covariate values are rounded at the second decimal place.

By applying different values for the coefficients $a_i,\ i = 0,1,2$, and $b_j,\ j = 0,\ldots 5$, we can obtain different model configurations, for example, containing either interaction or confounding terms, both, or neither. We can also investigate how these tree methods handle covariates that play different types of roles in the causal pathway between $T$ and $Y$. Specifically we consider the following five model configurations:

$$\begin{array}{lll}
\text{Model A.} & \mathbf{a} = \{a_j\} = (2,0,0)', & \mathbf{b} = \{b_j\} = (2,2,0,0,0,0)'; \\
\text{Model B.} & \mathbf{a} = \{a_j\} = (2,2,-4)', & \mathbf{b} = \{b_j\} = (2,2,2,2,2,2)'; \\
\text{Model C.} & \mathbf{a} = \{a_j\} = (2,0,-4)', & \mathbf{b} = \{b_j\} = (2,2,2,0,2,2)'; \\
\text{Model D.} & \mathbf{a} = \{a_j\} = (2,2,-4)', & \mathbf{b} = \{b_j\} = (2,2,2,0,0,0)'; \\
\text{Model E.} & \mathbf{a} = \{a_j\} = (2,2,-4)', & \mathbf{b} = \{b_j\} = (2,2,0,0,2,2)'.
\end{array}$$

Model A is a null model, where the covariates have no influence on the treatment effect. This model helps investigate the size issue or the type I error rate. Model B is equipped with all structures. Nevertheless, a massive tree with 16 terminal nodes is needed in order to fully represent the model structure. Model C also contains both confounding effect of $X_2$ and interacting effect of $X_4$, while neither $X_1$ nor $X_3$ is involved. In this case, a tree with four terminal nodes is expected. Model D mainly involves the confounder $X_2$, plus the exposure predictor $X_1$. Lastly, the active components in Model E are the effect modifier $X_4$ and the prognostic factor $X_3$.

For each simulated data set, all three tree methods, CIT, IT, and PT, are applied. Only one sample size is reported and the test sample method is used to select the optimal tree size, with 600 observations for the training sample and 400 observations for the test sample. Both AIC and BIC are used for the tree model selection. For each final tree selected, we record the optimal tree size and the splitting variables involved in the final tree structure. Table 1 presents the summarized results over 200 simulation runs.

We first examine the results from the null Model A. When BIC is used, all three tree methods seem rather conservative in committing Type I errors, implying that unsolicited signals are unlikely to be identified. With AIC, the empirical size, that is, the rate of giving false tree signals, is $(100 - 90.5)\% = 9.5\%$ for CIT, $(100 - 88.5)\% = 11.4\%$ for IT, and $(100 - 98.5)\% = 1.5\%$ for PT.

Next, Model B contains all the components that are related to the treatment and the response. Experimenting with this model provides an overall picture of what patterns each tree method tends to recognize. It can be seen that CIT yields the largest tree models by mostly catching the effects of $X_2$, $X_3$, and $X_4$. The treatment predictor $X_1$ is completely missed out by BIC and occasionally (32% of the time) selected by AIC. Note that $X_1$ is neither a confounder nor a modifier to the treatment effect. Due to the smaller penalty for mode complexity, AIC tends to select larger trees than BIC. As expected, the final propensity trees are split by both $X_1$ and $X_2$. The average final tree size of IT is 2.92, compared to its expected value 2. It is interesting to note that IT frequently gets confused by the confounding effect of $X_2$.

Model C contains only the components that actively influence the causal effects, namely, the confounder $X_2$ and the effect-modifier $X_4$. Both are perfectly identified by CIT. PT performs well

| | | Selection | Final Tree Size | | | | | | | Splitting Variables | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Model | Method | Criterion | 1 | 2 | 3 | 4 | 5 | 6 | $\geq 7$ | $X_1$ | $X_2$ | $X_3$ | $X_4$ | $X_5$ |
| **A** | CIT | AIC | 90.5 | 5.5 | 1.5 | 1.5 | 1.0 | 0.0 | 0.0 | 4.5 | 2.0 | 2.5 | 2.0 | 2.0 |
| | | BIC | 100.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| | IT | AIC | 88.5 | 6.0 | 4.5 | 1.0 | 0.0 | 0.0 | 0.0 | 2.5 | 4.0 | 3.5 | 3.0 | 3.0 |
| | | BIC | 100.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| | PT | AIC | 98.5 | 0.5 | 1.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.5 | 0.5 | 0.5 | 1.0 | 0.0 |
| | | BIC | 100.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| **B** | CIT | AIC | 0.0 | 4.0 | 0.0 | 19.5 | 5.5 | 39.5 | 31.5 | 32.0 | 96.0 | 96.0 | 100.0 | 1.5 |
| | | BIC | 0.0 | 4.0 | 0.0 | 26.0 | 0.0 | 61.5 | 8.5 | 0.0 | 96.0 | 96.0 | 100.0 | 0.0 |
| | IT | AIC | 0.0 | 6.5 | 25.0 | 50.5 | 13.0 | 3.0 | 2.0 | 7.0 | 93.5 | 6.0 | 100.0 | 8.0 |
| | | BIC | 0.0 | 40.5 | 28.5 | 27.0 | 3.5 | 0.5 | 0.0 | 1.0 | 59.5 | 0.5 | 100.0 | 1.0 |
| | PT | AIC | 0.0 | 0.5 | 33.5 | 61.5 | 4.0 | 0.5 | 0.0 | 99.5 | 100.0 | 0.5 | 2.0 | 1.5 |
| | | BIC | 0.0 | 1.0 | 51.5 | 46.5 | 1.0 | 0.0 | 0.0 | 99.0 | 100.0 | 0.0 | 0.5 | 0.0 |
| **C** | CIT | AIC | 0.0 | 0.0 | 4.5 | 90.5 | 5.0 | 0.0 | 0.0 | 1.0 | 100.0 | 2.0 | 100.0 | 0.5 |
| | | BIC | 0.0 | 0.0 | 4.5 | 95.5 | 0.0 | 0.0 | 0.0 | 0.0 | 100.0 | 0.0 | 100.0 | 0.0 |
| | IT | AIC | 0.0 | 47.0 | 35.5 | 14.0 | 1.5 | 1.5 | 0.5 | 0.5 | 52.5 | 0.5 | 100.0 | 2.5 |
| | | BIC | 0.0 | 55.5 | 33.0 | 100.0 | 1.0 | 0.5 | 0.0 | 0.0 | 44.5 | 0.0 | 100.0 | 0.0 |
| | PT | AIC | 0.0 | 97.5 | 1.0 | 1.5 | 0.0 | 0.0 | 0.0 | 1.0 | 100.0 | 1.0 | 0.5 | 0.5 |
| | | BIC | 0.0 | 100.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 100.0 | 0.0 | 0.0 | 0.0 |
| **D** | CIT | AIC | 0.0 | 1.0 | 83.5 | 11.0 | 2.0 | 1.5 | 1.0 | 99.0 | 100.0 | 2.0 | 2.5 | 4.5 |
| | | BIC | 0.0 | 10.5 | 89.5 | 0.0 | 0.0 | 0.0 | 0.0 | 89.5 | 100.0 | 0.0 | 0.0 | 0.0 |
| | IT | AIC | 1.5 | 43.0 | 31.5 | 18.0 | 4.5 | 1.0 | 0.5 | 5.5 | 98.5 | 3.0 | 4.5 | 2.0 |
| | | BIC | 0.2 | 54.5 | 28.0 | 14.0 | 1.5 | 0.0 | 0.0 | 1.0 | 98.0 | 0.5 | 1.5 | 0.0 |
| | PT | AIC | 0.0 | 0.0 | 33.5 | 62.0 | 2.5 | 1.5 | 0.5 | 100.0 | 100.0 | 2.0 | 2.5 | 1.0 |
| | | BIC | 0.0 | 1.0 | 49.5 | 49.0 | 0.5 | 0.0 | 0.0 | 99.0 | 100.0 | 0.0 | 0.0 | 0.0 |
| **E** | CIT | AIC | 0.0 | 0.0 | 2.5 | 89.0 | 8.5 | 0.0 | 0.0 | 1.5 | 3.0 | 100.0 | 100.0 | 2.0 |
| | | BIC | 0.0 | 0.0 | 2.5 | 97.5 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 100.0 | 100.0 | 0.0 |
| | IT | AIC | 0.0 | 90.0 | 9.0 | 1.0 | 0.0 | 0.0 | 0.0 | 2.5 | 1.5 | 2.5 | 100.0 | 3.5 |
| | | BIC | 1.5 | 97.5 | 1.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.5 | 0.0 | 0.0 | 98.5 | 0.5 |
| | PT | AIC | 96.5 | 3.5 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 1.5 | 0.5 | 1.0 | 0.5 | 0.0 |
| | | BIC | 98.5 | 1.5 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 1.0 | 0.5 | 0.0 | 0.0 | 0.0 |

Table 1: Simulation Results Based on the Test Sample Method: Relative frequencies (in percentages) of the final tree sizes in 200 runs identified by the causal inference tree (CIT), interaction tree (IT), and propensity tree (PT). Only one set of sample sizes is reported, with 600 observations forming the learning sample and 400 observations for the test sample.

in identifying the confounder $X_2$ while IT succeeds in recognizing the effect-modifier $X_4$. The same interesting phenomenon as with Model B occurs again: IT wrongly selects $X_2$ quite often. This will further be elaborated in Model D.

Model D is basically a propensity model, involving both the exposure predictor $X_1$ and the confounder $X_2$ only. In this case, CIT and PT provide equivalent results. Aiming at differential treatment effects, IT is supposed to have a null tree structure. However, we can see that most of time IT ends up with one or more splits on $X_2$. To gain insight, Figure 1 plots the splitting statistic used in both IT and CIT versus each cutoff point for $X_2$ in a single split of the data. The splitting statistic used in IT is a squared $t$ test statistic for interaction; thus the best cutoff point corresponds

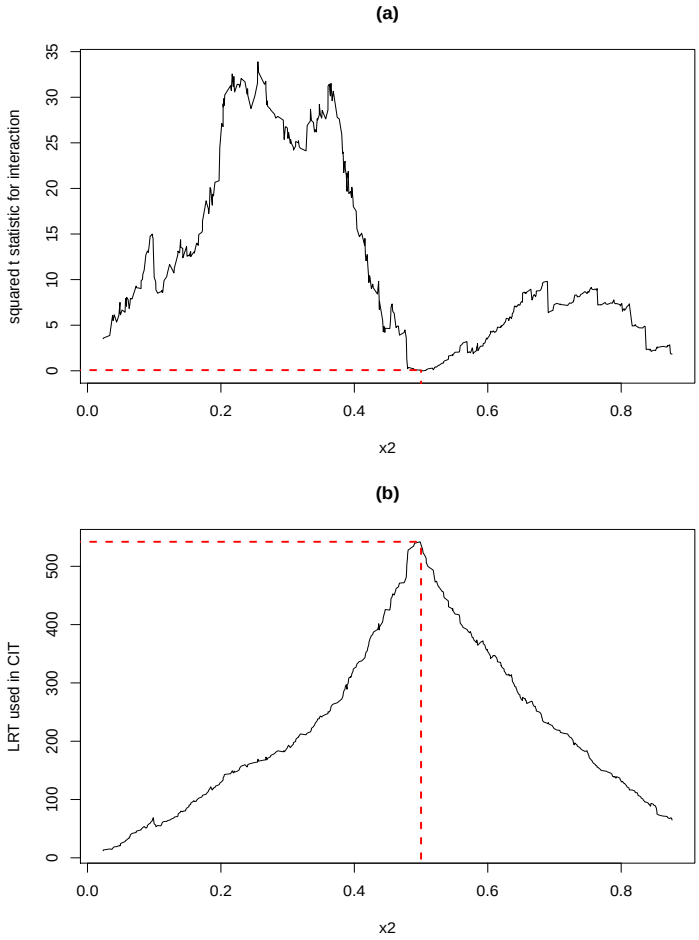



Figure 1: Plot of splitting statistic versus cutoff point on the confounder $X_2$: (a) $t$ test statistic (squared) for interaction used in IT; (b) likelihood ratio test statistic (up to some constant) used in CIT. Data were generated from Model D in Section 4.

to the maximum of splitting statistics. It is interesting to note in Figure 1(a) that the splitting statistic actually reaches its minimum at $X_2 = 0.5$, the only place where the treatment comparison is unbiased. At other cutoff points, the splitting statistic as a measure of interaction misleadingly inflates due to lack of adjustment for propensity. On the contrary, this does not cause a problem for CIT, which correctly selects the right cutoff point 0.5 as shown in Figure 1(b). Therefore, in order to identify differential causal effects correctly, it is critical to take confounders into consideration; otherwise, the estimation bias owing to imbalance of confounders between treatment groups may become overwhelming and eventually lead to misleading conclusions about the differential causal effects.

Finally, Model E is essentially an outcome regression model, in which both the prognostic factor $X_3$ and the effect-modifier $X_4$ are involved. It can be seen that CIT functions similarly to IT in

| | Effect | Propensity | Case I | | Case II | | Case III | | Case IV | |
|---|---|---|---|---|---|---|---|---|---|---|
| Group | $\Delta_k$ | $e_k$ | $\widehat{\Delta}_k$ | $\hat{e}_k$ | $\widehat{\Delta}_k$ | $\hat{e}_k$ | $\widehat{\Delta}_k$ | $\hat{e}_k$ | $\widehat{\Delta}_k$ | $\hat{e}_k$ |
| 1 | $-1.940$ | 0.156 | $-2.077$ | 0.189 | $-3.246$ | 0.502 | $-2.005$ | 0.171 | $-2.106$ | 0.201 |
| 2 | 2.067 | 0.866 | 1.924 | 0.861 | 1.925 | 0.861 | $-0.098$ | 0.857 | 1.916 | 0.860 |
| 3 | $-1.938$ | 0.843 | $-1.987$ | 0.829 | $-2.629$ | 0.676 | $-1.016$ | 0.840 | $-2.006$ | 0.824 |

Table 2: Simulation Results for Assessing Sensitivity of CIT to Misspecification. Four scenarios are considered. In Case I, variables $\{X_1, X_2, X_3, X_4\}$ are used; In Case II, the confounder $X_2$ is omitted; In Case III, the effect-modifier $X_3$ is omitted; In Case IV, the collider $X_5$ is included. The estimated treatment effect and propensity for each group were averaged over 100 runs.

detecting treatment-by-covariate interactions. CIT also identifies splits on the prognostic factor $X_3$. It comes as no surprise that PT, concerning propensity only, gives a null tree for most of the time.

### 4.2 Sensitivity under Misspecification

We next investigate how CIT performs under misspecified scenarios where an important confounder or effect-modifier is omitted or when a collider is included. We design an experiment with the following data generation scheme:

1. Generate $X_1, \ldots, X_4$ independently from Unif(0,1) and create threshold variables $Z_j = 1_{\{X_j \le 0.5\}}$ for $j = 1, \ldots, 4$.

2. Generate $W_1$ and $W_2$ independently from Bernoulli(0.5) and hence simulate $X_5 \sim \mathcal{N}(2W_1 + 2W_2, 1)$.

3. Set $\text{logit}(\pi) = 0.5 - Z_1 Z_2 + W_1$. Generate $T \sim \text{Bernoulli}(\pi)$.

4. Set $\mu = 2 + 2Z_1 Z_2 - 2T + 4Z_1 Z_3 T + W_2$ and generate $y \sim \mathcal{N}(\mu, 1)$.

The observed data consist of repetitions of $\{Y, X_1, \ldots, X_4\}$. With the above configuration, $X_1$ is both a confounder and an effect-moderator; $X_2$ is a confounder; $X_3$ is a moderator; $X_4$ is irrelevant; and $X_5$ is a collider with the $M$ diagram model (see, e.g., Figure 2(a) in Greenland 2003). The data essentially involve three groups with either different propensities or treatment effects. Observations in Group 1 satisfies $Z_1 Z_2 = 1$; Group 2 is characterized by $(1 - Z_1)Z_3 = 1$; and Group 3 contains the others.

In order to assess sensitivity, an independent validation set with $5,000$ observations is first generated. Based on true grouping, the causal effect and propensity for each group are computed and presented in Table 2. Next, a total of 100 simulation runs are considered. For each simulation run, a training set with 600 observations and a test set with 400 observations are generated, on which basis CITs are constructed using different sets of variables. In Case I, variables $\{X_1, X_2, X_3, X_4\}$ are used; Case II uses $\{X_1, X_3, X_4\}$ with confounder $X_2$ omitted; In Case III, $\{X_1, X_2, X_4\}$ are used by omitting the moderator $X_3$; In Case IV, $\{X_1, X_2, X_3, X_4, X_5\}$ are used by including the collider $X_5$. Each final CIT (based on BIC) is applied to the validation set to compute the individual causal effect $\widehat{\Delta}_i$ and propensity $\hat{e}_i$ for each observation in the validation set. The predicted ICEs and propensities are

aggregated for each group, based on the true grouping. The grouped causal effect and propensity estimates are then averaged over 100 simulation runs. The results are also presented in Table 2. It can be seen that, in Case I, CIT does very well in estimating treatment effects and propensities. In both Case II and Case III, substantial bias is present in estimating the treatment effects. The results for Case IV suggest that the collider $X_5$ also introduces bias. However, compared to the bias from omitting a confounder or moderator, the bias from including a collider is much smaller. This is consistent with the conclusions in Greenland (2003).

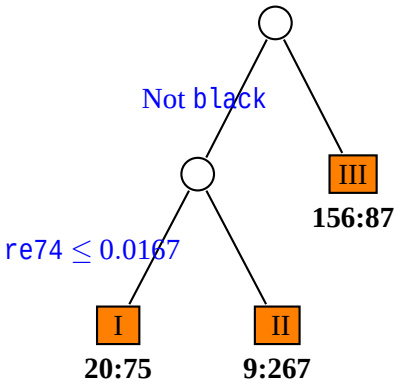
## 5. Analysis of NSW Data

As an illustration, we revisit the NSW data set extensively analyzed by LaLonde (1986) and Dehejia and Wahba (1999), where the objective is to assess the impact of the National Supported Work (NSW) Demonstration on post-intervention income levels. The NSW demonstration was a labor training program implemented in the mid-1970s to provide work experiences for a period of 6-18 months to individuals facing economic and social difficulties. NSW itself was designed as a randomized controlled study where subjects were randomly assigned to two treatment groups: the NSW-exposed group and the unexposed group.

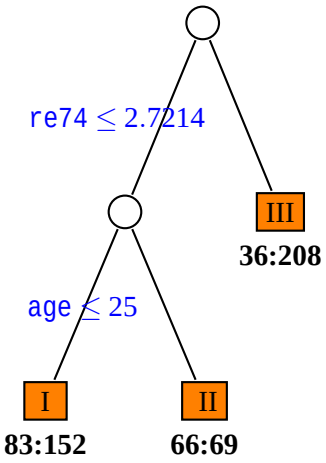
With a rather innovative approach that later on became influential, LaLonde (1986) compiled a composite data set by taking subjects in the NSW-exposed group only and then obtaining the nonexperimental control group from other sources, including the Panel Study of Income Dynamics (PSID) and the Current Population Survey (CPS) databases. His aim was to examine the extent to which nonexperimental estimators can replicate the unbiased experimental estimate of the treatment impact. He concluded that nonexperimental estimators are either inaccurate relative to the experimental benchmark or sensitive to model specification. Since then, the mixed NSW data have been analyzed by various authors with alternative approaches. Among others, Dehejia and Wahba (1999) obtained estimates of the treatment effect that are close to the experimental benchmark estimate or the 'gold' standard using propensity score matching and stratification.

Most of these previous works are focused on estimating the ACE of NSW. Here we shall apply the CIT methods to explore the variabilities of its effects, in addition to dealing with the nonrandom treatment assignments. There are several versions of the data with varying sources for obtaining the control or unexposed group, available from `http://www.nber.org/~rdehejia/nswdata.html`. The data set we use is the one available in the R package `MatchIt` contributed by Ho et al. (2007, 2011). This is a subset restricted to males who had 1974 earnings available, for the reasons explained in Dehejia and Wahba (1999). There are 614 observations (185 treated and 429 control) and 10 variables in the data, which include the treatment assignment indicator. A brief description and some summary statistics of these variables are provided in Table 3. The outcome variable is `re78`, the 1978 earnings. All covariates but `educ` are severely unbalanced between the participants actively exposed to NSW and those in the unexposed group selected from other survey databases.

(a) Propensity Tree




(b) Interaction Tree




(c) Causal Inference Tree

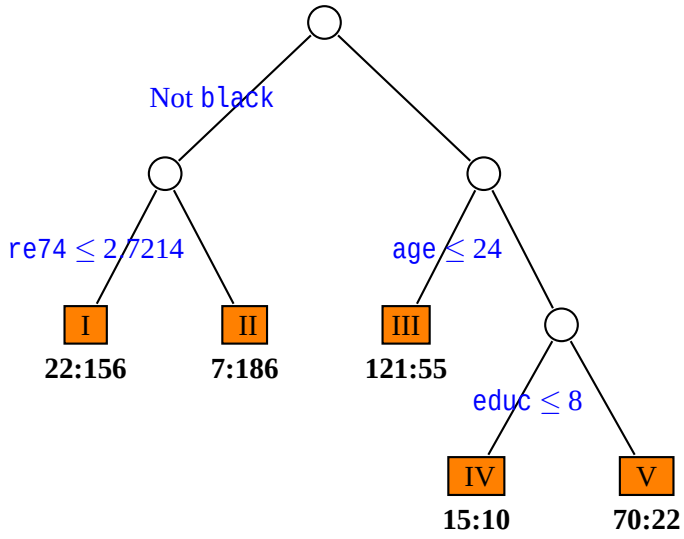



Figure 2: Final Tree Models for the NSW Data: (a) Propensity Tree; (b) Interaction Tree; (c) Causal Inference Tree.

(a) Continuous Variables

| Variable | | All | | NSW Exposed | | Unexposed | | P-value | |
|---|---|---|---|---|---|---|---|---|---|
| Name | Description | mean | sd | mean | sd | mean | sd | two-sample *t* | Wilcoxon |
| age | Age in years | 27.36 | 9.88 | 25.82 | 7.16 | 28.03 | 10.79 | 0.0107 | 0.5195 |
| educ | Schooling years | 10.27 | 2.63 | 10.35 | 2.01 | 10.24 | 2.86 | 0.6330 | 0.7920 |
| re74 | 1974 earnings | 4,557.55 | 6,477.96 | 2,095.57 | 4,886.62 | 5,619.24 | 6,788.75 | 0.0000 | 0.0000 |
| re75 | 1975 earnings | 2,184.94 | 3,295.68 | 1,532.06 | 3,219.25 | 2,466.48 | 3,292.00 | 0.0012 | 0.0000 |
| re78 | 1978 earnings | 6,792.83 | 7,470.73 | 6,349.14 | 7,867.40 | 6,984.17 | 7,294.16 | 0.3342 | 0.2818 |

(b) Discrete Variables

| Variable | | Frequency | | P-value | |
|---|---|---|---|---|---|
| Name | Description | NSW Exposed | Unexposed | $\chi^2$ | Fisher's Exact |
| black | 0 - No | 29 | 342 | 0.0000 | 0.0000 |
| | 1 - African-American | 156 | 87 | | |
| hispan | 0 - No | 174 | 368 | 0.0053 | 0.0026 |
| | 1 - of Hispanic origin | 11 | 61 | | |
| married | 0 - No | 150 | 209 | 0.0000 | 0.0000 |
| | 1 - Yes | 35 | 220 | | |
| nodegree | 0 - No | 54 | 173 | 0.0113 | 0.0106 |
| | 1 - Has a high school degree. | 131 | 256 | | |

Table 3: Variable description and summary statistics for the NSW data set. All earnings are expressed in U.S. dollars.

(a) Propensity Tree

| | NSW Group | | | Unexposed Group | | | Estimated |
|---|---|---|---|---|---|---|---|
| Node | size | mean | sd | size | mean | sd | Propensity |
| I | 20 | 8.1423 | 6.6646 | 75 | 5.2302 | 6.3981 | 21.05% |
| II | 9 | 6.0534 | 4.9218 | 267 | 8.1712 | 7.6170 | 3.26% |
| III | 156 | 6.1363 | 8.1435 | 87 | 4.8534 | 6.2017 | 64.20% |

(b) Interaction Tree

| | NSW Group | | | Unexposed Group | | | Treatment Effect | |
|---|---|---|---|---|---|---|---|---|
| Node | size | mean | sd | size | mean | sd | estimate | s.e. |
| I | 83 | 5.0392 | 5.1160 | 152 | 5.2804 | 5.5401 | −0.2412 | 0.7192 |
| II | 66 | 8.4894 | 10.3819 | 69 | 3.4528 | 5.8233 | 5.0366 | 1.4576 |
| III | 36 | 5.4455 | 7.0965 | 208 | 9.4007 | 8.0201 | −3.9552 | 1.3070 |

(c) Causal Inference Tree

| | NSW Group | | | Unexposed Group | | | Estimated | Treatment Effect | |
|---|---|---|---|---|---|---|---|---|---|
| Node | size | mean | sd | size | mean | sd | Propensity | estimate | s.e. |
| I | 22 | 8.1431 | 6.3676 | 156 | 4.8438 | 5.6728 | 12.35% | 3.2993 | 1.3118 |
| II | 7 | 5.4539 | 5.3997 | 186 | 9.7759 | 8.0259 | 3.62% | −4.3221 | 3.0634 |
| III | 71 | 4.6987 | 4.8043 | 55 | 4.8545 | 5.9303 | 56.35% | −0.1558 | 0.9564 |
| IV | 15 | 3.8662 | 3.9130 | 10 | 1.0999 | 2.8541 | 60.00% | 2.7663 | 1.4438 |
| V | 70 | 8.0809 | 10.7408 | 22 | 6.5570 | 7.3371 | 76.09% | 1.5239 | 2.4565 |

Table 4: Summary statistics for the terminal nodes: (a) the final propensity tree (PT); (b) the final interaction tree (IT); and (c) the final causal inference tree (CIT). The means and standard deviations are given in thousand dollars.

We applied three tree procedures to the data: PT, IT, and CIT. The final tree structures, all selected by BIC, are plotted in Figure 2. Considering the moderate sample size, a bootstrap method was used for final tree selection. In Figure 2, the internal nodes are denoted by circles. The splitting rule is given under each internal node. Observations satisfying the rule go to the left child node and observations not satisfying go to the right child node. The terminal nodes are denoted by rectangles and renamed by Roman numerals inside. Underneath each terminal node is the number of exposed subjects versus the number of unexposed subjects within the terminal node. Some summary statistics for the terminal nodes in each final tree are provided in Table 4.

Figure 2(a) gives the final PT structure, which delineates a meaningful heterogeneity in propensity. It is clear that African Americans were more likely to participate in this labor program. PT also identifies a group, terminal node II, with extremely low propensity (3.26%). This group is characterized by people who were not African Americans and had some income in 1974. However, this PT model tells nothing about differential treatment effects.

Figure 2(b) displays the final IT structure. Variables `re74` and `age` stand out as determinants of differential causal effects. Apparently remarkable differential treatment effects seem to exist across the three terminal nodes based on Table 4(b). However, since the method does not adjust

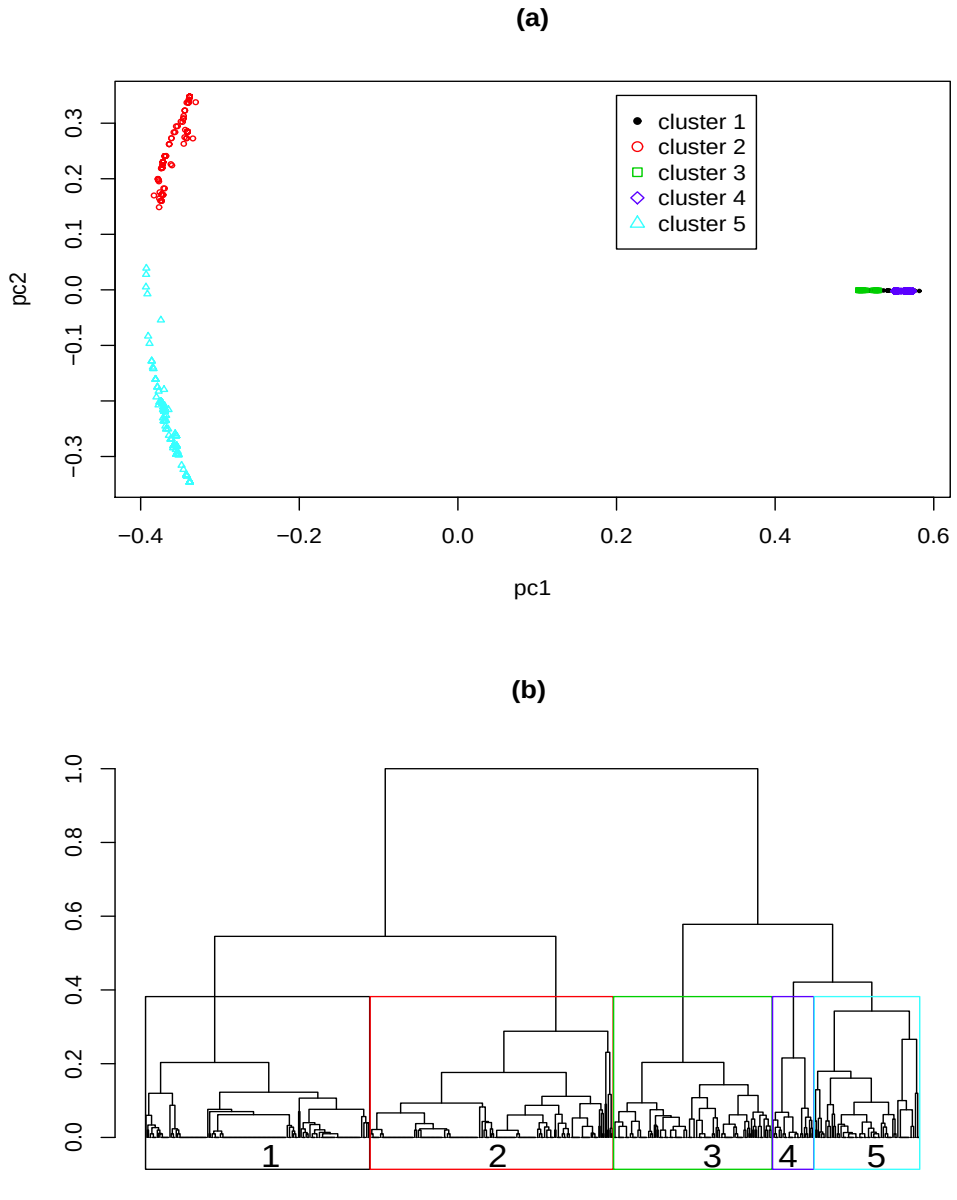



Figure 3: Aggregated Grouping for the NSW Data: (a) Multidimensional scaling (MDS) plot; (b) Dendrogram for hierarchical clustering with average linkage. The distance matrix was computed by aggregating 100 bootstrap samples.

for heterogeneous propensity, the results are not reliable. Hence we make no further attempt in interpreting.

Figure 2(c) presents the final CIT model, which has a more comprehensive structure. It is interesting to see that the left-half of the tree resembles the PT tree in Figure 2(a). In particular, the CIT comes up with a similar terminal node II, which contains non-African Americans with income higher than \$2,721 in 1974. Since CIT accounts for both propensity and differential causal effects, it is valid to estimate the NSW effect via direct comparison of sample means within each terminal node. Table 4(c) provides the relevant quantities. CIT also identifies some interesting patterns of differential treatment effects. The surprising comparison occurs to terminal node II, where the NSW-exposed group had a lower average income than the unexposed group with a mean difference of \$4,322. However, this should not be a point of great concern due to its very low propensity 3.62%.

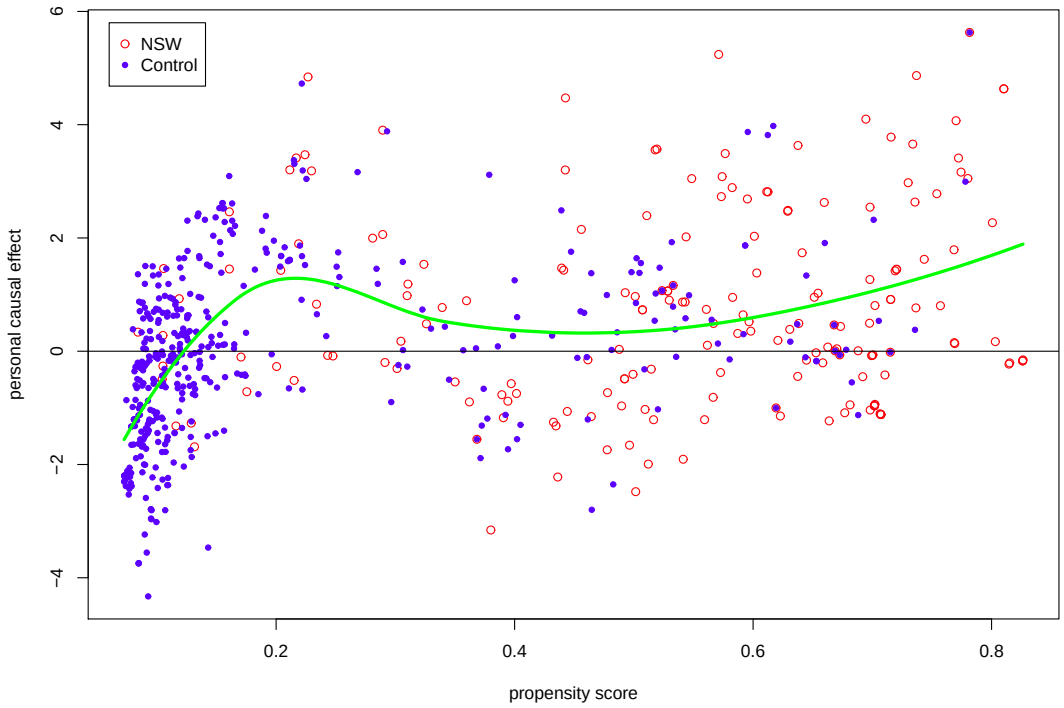



Figure 4: Plot of the Estimated Personal Causal Effects vs. Propensity Scores for the NSW Data. Referring to Algorithm 2, $B = 1000$ bootstrap samples were used in a three-fold cross validation procedure and the parameter $m$ was set as 3.

If it is agreed that terminal node II be excluded from consideration due to lack of comparison basis and the minor negative effect of NSW in terminal node III be ignored, then one may tentatively conclude the absence of qualitative interactions. Using Equations (15)-(16) and information in Table 4(c), the ACE is estimated as \$1,845 $\pm$ \$809, which is very close to the benchmark randomized

experiment estimate of \$1,794 $\pm$ \$633. As a comparison, the unadjusted estimate is \$635 $\pm$ \$677 and the ANCOVA estimate is \$1,548 $\pm$ \$781 after adjusting for all covariates. It is worth mentioning that the ANCOVA estimate varies dramatically when different sets of variables are included in the model. Using nonparametric matching method (Ho et al., 2007) implemented in the `MatchIt` package, the subclassification estimate (with 5 subclasses) is \$1,237 $\pm$ \$1,196 and the optimal matching (Rosenbaum, 1989) based estimate is \$1,366 $\pm$ \$720.

Next, we applied the aggregated grouping method described in Algorithm 1. We obtained an averaged distance matrix from 100 bootstrap samples. The modal number of optimal tree sizes is 5. The classical MDS (Gower, 1966) was used to explore the the distance matrix. Figure 3(a) provides the resultant plot when the data are represented in a two-dimensional space. Agglomerative hierarchical clustering with average linkage was then used to determine the final clusters. See Figure 3(b) for the dendrogram. The cluster membership specification was also added to the MDS plot in Figure 3(a). It can be seen that Cluster 2 and Cluster 5 are distant from other three clusters. Table 5(a) shows the correspondence between the five clusters and the five CIT terminal nodes. It can be seen that overall they match well, except for minor inconsistency between clusters 4 & 5 and terminal nodes IV & V. This indicates that the CIT structure is relatively stable. The summary statistics for the five clusters are outlined in Table 5(b), showing a pattern similar to Table 4(c). After removing Cluster 2, the estimate of ACE is \$1,897 $\pm$ \$807. We would like to emphasize that the excluded Group II in CIT can be explained by the fact that people who were not black and had some income in 1974 seemed unlikely (with estimated propensity 3.62%) to participate the NSW intervention program. This easy interpretation is no longer available with Cluster 2 obtained from the aggregated grouping procedure.

Finally, ensemble CITs were used to estimate the ICE and propensity score for each individual. Referring to Algorithm 2, three-fold ($V = 3$) cross-validation with $B = 1,000$ bootstrap samples (with stratification on treatment) was used in the analysis; and at each split, $m = 3$ variables were randomly selected as candidate splitting variables. Figure 4 plots the estimated ICE vs. propensity scores. The interpretation for ICE is the difference between what an individual would have earned in 1978 if he had attended NSW, compared to the 1978 earnings if he had not attended. It can be seen that the area with low propensity (below .10) is dominated by subjects in the control and their associated personal effects of NSW are quite mixed. Other than that, the intervention program seem to have an overall positive effect. Figure 5 summarizes the results for each treatment-by-stratum combination, in which the five strata obtained from aggregated grouping are used. It can be seen that both propensity and individual causal effects are reasonably homogeneous within each stratum, even though the individuals were from different treatment groups.

## 6. Extension to Ordinal/Continuous Treatments

The concept and properties of the facilitating score can be extended to scenarios where the treatment variable is nominal (Lechner, 1999) or ordinal (Imbens, 2000). Suppose that the treatment variable $T$ is allowed to range within $\Im$, where $\Im$ is a discrete set with ordered or unordered values. Let $Y_t = Y_t(\omega)$ denote the potential outcome if unit $\omega$ was assigned to the treatment level $t$. Let $e_t(\mathbf{X}) = \Pr\{T = t|\mathbf{X}\}$ be the generalized propensity score (GPS). A generalized weak facilitating score can be defined as below.

**Definition 6** *A generalized weak facilitating score* $\mathbf{a}(\mathbf{X})$ *is a q-dimensional (*$0 < q \leq p$*) function of* $\mathbf{X}$ *such that (i)* $\mathbf{X} \perp\!\!\!\perp T \,|\, \mathbf{a}(\mathbf{X})$ *and (ii)* $E(Y_t - Y_{t'}|\mathbf{a}(\mathbf{X})) = E(Y_t - Y_{t'}|\mathbf{X})$ *for any* $t, t' \in \Im$.

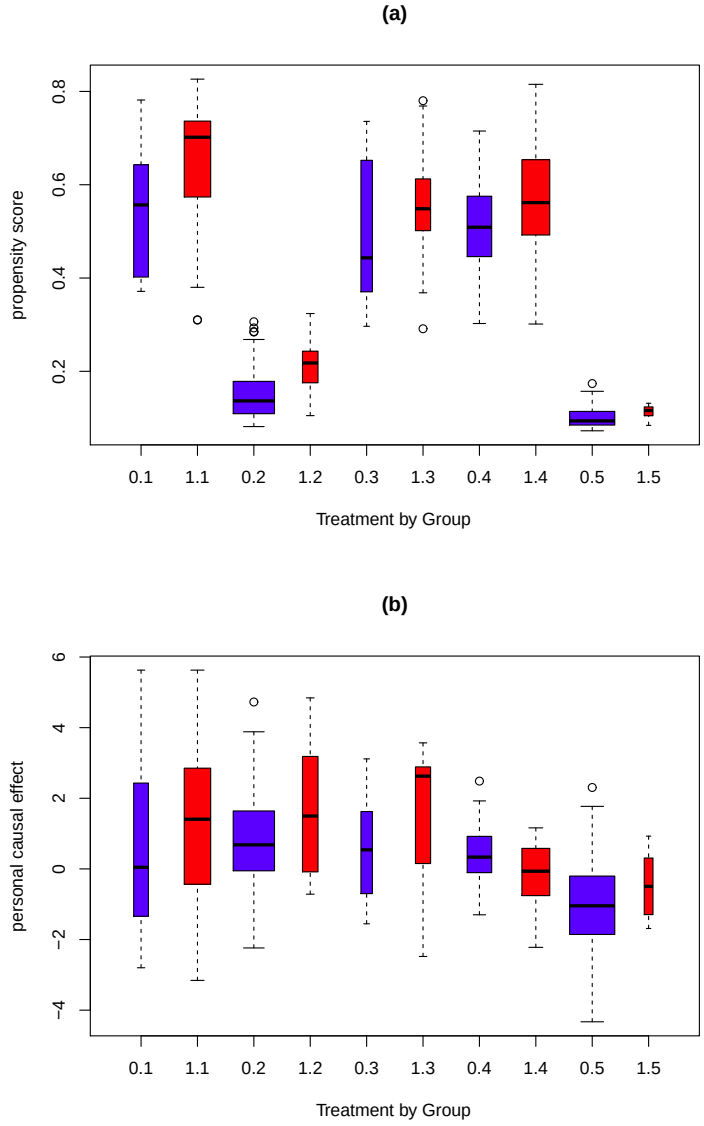



Figure 5: Parallel Box-Plots of (a) the Propensity Scores and (b) the Estimated ICE for Each of the Treatment$\times$Stratum Combinations. The '0.k' combination corresponds to individuals in Stratum $k$ who did not attend the NSW program while '1.k' corresponds to those in Stratum $k$ who did, for $k = 1, \ldots, 5$. The width of each box has been made proportional to the sample size in each combination.

Condition (ii) is equivalent to saying that $E(Y_t - Y_{t'}|\mathbf{a}(\mathbf{X}) = \mathbf{a})$ is independent of $\mathbf{X}$. The following theorem provides a basis for its usage. It shows that, if the joint distribution of $(Y, T)$ can be modeled through a vector-valued function $\mathbf{h}(\mathbf{X})$, then $\mathbf{h}(\mathbf{X})$ is a generalized weak facilitating score and direct estimates of causal effects can be obtained by conditioning on $\mathbf{h}(\mathbf{X}) = \mathbf{h}$.

(a) Correspondence

| | Cluster | | | | |
|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 |
| I | 178 | 0 | 0 | 0 | 0 |
| II | 0 | 193 | 0 | 0 | 0 |
| III | 0 | 0 | 126 | 0 | 0 |
| IV | 0 | 0 | 0 | 22 | 3 |
| V | 0 | 0 | 0 | 11 | 81 |

(b) Summary of Five Groups

| | NSW Group | | | Unexposed Group | | | Estimated | Treatment Effect | |
|---|---|---|---|---|---|---|---|---|---|
| Node | size | mean | sd | size | mean | sd | Propensity | estimate | s.e. |
| 1 | 22 | 8.1431 | 6.3676 | 156 | 4.8438 | 5.6728 | 12.36% | 3.2993 | 1.3118 |
| 2 | 7 | 5.4539 | 5.3997 | 186 | 9.7759 | 8.0259 | 3.63% | −4.3221 | 3.0634 |
| 3 | 71 | 4.6987 | 4.8043 | 55 | 4.8545 | 5.9303 | 56.35% | −0.1558 | 0.9564 |
| 4 | 21 | 4.1514 | 4.3182 | 12 | 2.4027 | 5.5116 | 63.64% | 1.7487 | 1.7283 |
| 5 | 64 | 8.3825 | 11.0826 | 20 | 6.3210 | 7.1054 | 76.19% | 2.0614 | 2.6383 |

Table 5: Results for the five groups obtained from aggregated grouping: (a) correspondence between the obtained groups and the five CIT terminal nodes; (b) summary statistics. The distance matrix was computed from 100 bootstrap samples and hierarchical clustering with average linkage was used for determining the final groups.

**Theorem 7** *Assume that the conditional joint density of* $(Y,T)$ *given* $\mathbf{X}$, $f_{Y,T|\mathbf{X}}(Y,T|\mathbf{X})$, *can be written as* $f_{Y,T|\mathbf{X}}(Y,T|\mathbf{X}) = g(Y,T,\mathbf{h}(\mathbf{X}))$ *for some function* $g(\cdot)$. *In other words,* $(Y,T) \perp\!\!\!\perp \mathbf{X} \,|\, \mathbf{h}(\mathbf{X})$. *Further assume that treatment assignment is strongly ignorable so that* $Y_t \perp\!\!\!\perp 1_{\{T=t\}} \,|\, \mathbf{X}$ *for any* $t \in \Im$. *When* $0 < e_t(\mathbf{X}) < 1$, *we have*

*(1).* $\mathbf{h}(\mathbf{X})$ *is a generalized weak facilitating score.*

*(2). Concerning the causal effect in subpopulation* $\Omega_{\mathbf{h}} = \{\omega : \mathbf{h}(\mathbf{X}(\omega)) = \mathbf{h}\}$,

$$\begin{aligned} E(Y_t - Y_{t'}|\mathbf{h}(\mathbf{X}) = \mathbf{h}) &= E\{Y_t|T = t, \mathbf{h}(\mathbf{X}) = \mathbf{h}\} - E\{Y_{t'}|T = t', \mathbf{h}(\mathbf{X}) = \mathbf{h}\} \\ &= E\{Y|T = t, \mathbf{h}(\mathbf{X}) = \mathbf{h}\} - E\{Y|T = t', \mathbf{h}(\mathbf{X}) = \mathbf{h}\} \end{aligned}$$

*is independent of* $\mathbf{X}$.

The proof of Theorem 7 is deferred to Appendix A. As stressed by Lechner (1999) and Imbens (2000), GPS $e_t(\mathbf{X})$ does not have a causal interpretation. However, the reinforced assumption $f_{Y,T|\mathbf{X}}(Y,T|\mathbf{X}) = g(Y,T,\mathbf{h}(\mathbf{X}))$ implies that $e_t(\mathbf{X})$ can be fully characterized by $\mathbf{h}(\mathbf{X})$ or its components. This is analogous to the assumption of *uniquely parameterized propensity function* in Imai and van Dyk (2004), where a parametric form is prescribed for $e_t(\mathbf{X})$. To estimate $\mathbf{h}(\mathbf{X})$, a multinomial or cumulative logit model can be used for propensity and the outcome can be modeled with

multiple linear regression. The above results also can be extended to continuous treatment variables with arguments similar to Hirano and Imbens (2005).

## 7. Discussion

Embedded in Rubin's causal model, we have introduced a new concept, the facilitating score, to help tackle the heterogeneity in both propensity and causal effects. The facilitating score is a finer balancing score of Rosenbaum and Rubin (1983), plus additional conditions for dealing with differential causal effects. It supplies a framework that promotes joint modeling of $(Y,T)$ for a better understanding of causal effects. Accordingly we have devised recursive partitioning methods to aid in causal inference at different levels.

The facilitating score concept and the CIT methods can be useful in personalized medicine and other similar applications. Medical treatment is traditionally centered on standards of care on the basis of large epidemiological cohort studies or randomized trials that are powered for assessing ACE. These studies however do not account for variabilities of individuals in reacting to the treatments and drug-to-drug interactions. The new medical model of personalized medicine or treatments seeks flexible ways that allow for treatment decisions or practices being tailored to individual by integrating post-trial clinical data and new developments in biotechnology to improve healthcare. The collected covariates are often expanded to a more comprehensive consideration of the patient, including medical measurements, family history, social circumstances, environment and behaviors, and biological variables. As a result, the data are often observational and high-dimensional in nature. As demonstrated in the NSW data example, causal inference in observational studies could be very complex, owing to the confounding and interacting effects complicated by covariates. While personalized medicine is the ultimate goal, stratified medicine has been the current approach. Stratified medicine aims to select the best therapy for groups of patients who share common biological characteristics. The proposed CIT method and aggregated grouping can be used seeking strategies for deploying stratified medicines. Insight into a greater degree of personalized treatment can be gained by studying the personal treatment effects with ensemble CITs.

Some limitations of the proposed methods are listed below. First, despite the usefulness of ICE, assessing ICE entails larger data than assessment of ACE in order to have the same level of precision (or variance). There are many trials in research practice that are only powered to detect ACE. For this reason, the proposed methods are best suitable for moderately-sized or large follow-up data collected in post trial periods or extracted from Medicare or Medicaid databases, in which randomization is not available. Secondly, the recursive partitioning methods are highly adaptive or data-driven in nature and often regarded as exploratory or hypothesis-generating. It is important to interpret the results with caution. In addition, the validity of Theorem 7 relies on the assumption of strong ignorability. Like other methods, CIT performance is vulnerable to violated assumptions and model misspecification. Shpitser and Pearl (2008) examines possibly milder conditions to ensure identifiability and facilitate estimation in causal inference. It would be interesting to investigate how to extend the proposed methods under mild conditions.

In terms of future research, Theorem 7 is readily applied to data with binary outcomes. With further research efforts, both the facilitating score and CIT may be extended to other types of outcomes such as censored survival times or longitudinal measurements. It would also be interesting to extend the proposed methods to scenarios when both treatment and confounders are time-varying, as studied in marginal structural models and structural nested models (Robins, 1999), and when some

confounders are unmeasured but there exist some instrumental variables (IV; Angrist, Imbens, and Rubin 1996) that satisfy the strong ignorability and other conditions. In addition, Robins, Rotnitzky, and Zhao (1994) proposed doubly robust (DB) estimation methods to deal with mis-specification in either the response model or the propensity model. Along similar lines, the targeted maximum likelihood (TML; van der Laan and Rubin 2006) is another newly developed causal inference method that enjoys a favorable theoretical property for being doubly robust and locally efficient, meaning that if at least one of the propensity and outcome models is correctly specified, then the TML estimator is consistent and asymptotically normal; if both models are correctly specified it is also efficient. Similar work with facilitating score modeling could be another avenue for future research.

## Acknowledgments

The authors would like to thank the editor and three anonymous referees, whose insightful comments, constructive suggestions, and helpful discussions have greatly improved an earlier version of this manuscript.

## Appendix A. Proof of Theorem 7.

We sketch the proof when $T$ is ordinal or nominal. Theorem 3 follows as a special case when $\Im = \{0,1\}$. Some steps are standard arguments in propensity score theories. We include them for the sake of completeness.

First of all, the conditional probability density function of $T|\mathbf{X}$ is

$$f_{T|\mathbf{X}}(T|\mathbf{X}) = \int_Y f_{Y,T|\mathbf{X}}(Y,T|\mathbf{X})dY = \int_Y g(Y,T,\mathbf{h}(\mathbf{X}))dY.$$

Thus the GPS $e_t(\mathbf{X}) = P(T=t|\mathbf{X}) = g_1(\mathbf{h}(\mathbf{X}))$ for some function $g_1(\cdot)$. Namely, $\mathbf{h}(\mathbf{X})$ is a finer function of $e_t(\mathbf{X})$. For this reason, we denote $e_t(\mathbf{X}) = e_t(\mathbf{h}(\mathbf{X}))$.

Next, since $\mathbf{h}(\mathbf{X})$ is measurable with respect to, $\sigma(\mathbf{X})$, the $\sigma$-algebra generated by $\mathbf{X}$,

$$\Pr\{T=t|\mathbf{X},\mathbf{h}(\mathbf{X})\} = \Pr\{T=t|\mathbf{X},\} = e_t(\mathbf{X}).$$

Let $\delta_t = I\{T=t\}$ be the indicator function of whether $T=t$. By iterated expectation,

$$\begin{aligned}\Pr\{T=t|\mathbf{h}(\mathbf{X})\} &= E(\delta_t|\mathbf{h}(\mathbf{X})) = E\{E(\delta_t|\mathbf{X},\mathbf{h}(\mathbf{X}))|\mathbf{h}(\mathbf{X})\} \\ &= E\{E(\delta_t|\mathbf{X})|\mathbf{h}(\mathbf{X})\} = E\{e_t(\mathbf{X})|\mathbf{h}(\mathbf{X})\} = e_t(\mathbf{X}).\end{aligned}$$

Namely, $\Pr\{T=t|\mathbf{X},\mathbf{h}(\mathbf{X})\} = e_t(\mathbf{X}) = \Pr\{T=t|\mathbf{h}(\mathbf{X})\}$, which implies $T \perp\!\!\!\perp \mathbf{X}|\mathbf{h}(\mathbf{X})$.

Further assuming the treatment assignment is strongly ignorable given $\mathbf{X}$, it follows that the treatment assignment is ignorable given $\mathbf{h}(\mathbf{X})$, that is, $T \perp\!\!\!\perp Y_t|\mathbf{h}(\mathbf{X})$, which can be established by showing

$$\begin{aligned}\Pr\{T=t'|Y_t,\mathbf{h}(\mathbf{X})\} &= E\{\delta_{t'}|Y_t,\mathbf{h}(\mathbf{X})\} = E\{E(\delta_{t'}|\mathbf{X},Y_t,\mathbf{h}(\mathbf{X}))|Y_t,\mathbf{h}(\mathbf{X})\} \\ &= E\{E(\delta_{t'}|\mathbf{X})|Y_t,\mathbf{h}(\mathbf{X})\} \text{ due to strong ignorability} \\ &= E\{e_{t'}(\mathbf{X})|Y_t,\mathbf{h}(\mathbf{X})\} = e_{t'}(\mathbf{X}) = \Pr\{T=t'|\mathbf{h}(\mathbf{X})\}.\end{aligned}$$

To check condition (ii) in Definition 6, consider $E\{Y_t|\mathbf{h}(\mathbf{X})\}$. Since $Y=\sum_t Y_t\delta_t$ and $\delta_t\delta_{t'}=0$ for $t\neq t'$, we have $Y\delta_t=Y_t\delta_t$. Consider

$$\begin{aligned} E\{Y_t|\mathbf{h}(\mathbf{X})\} &= E\{Y_t|\mathbf{h}(\mathbf{X})\}\cdot E\{\delta_t|\mathbf{h}(\mathbf{X})\}/E\{\delta_t|\mathbf{h}(\mathbf{X})\} \\ &= E\{Y_t\delta_t|\mathbf{h}(\mathbf{X})\}/E\{\delta_t|\mathbf{h}(\mathbf{X})\} \text{ by strong ignorability} \\ &= E\{Y\delta_t|\mathbf{h}(\mathbf{X})\}/e_t(\mathbf{X}). \end{aligned}$$

It can be seen that $\mathbf{h}(\mathbf{x})$ is a finer function of both the numerate and denominator in the above expression. Thus $E\{Y_t|\mathbf{h}(\mathbf{X})=\mathbf{h}\}$ is fully determined by $\mathbf{h}$ and no longer relies on the value of $\mathbf{X}$.

Finally, in order to have available causal inference, it is important to note that, for given $t$ and $t'$, $h(\mathbf{x})=\mathbf{h}$ fully determines both $e_t(\mathbf{h})$ and $e_{t'}(\mathbf{h})$. Therefore,

$$\begin{aligned} E\{Y_t-Y_{t'}|\mathbf{h}(\mathbf{X})=\mathbf{h}\} &= E\{Y_t|\mathbf{h}(\mathbf{X})=\mathbf{h},T=t,e_t(\mathbf{h})\}-E\{Y_{t'}|\mathbf{h}(\mathbf{X})=\mathbf{h},T=t',e_{t'}(\mathbf{h})\} \\ &= E\{Y|\mathbf{h}(\mathbf{X})=\mathbf{h},T=t\}-E\{Y|\mathbf{h}(\mathbf{X})=\mathbf{h},T=t'\} \end{aligned}$$

is independent of $\mathbf{X}$. This justifies the direct use of mean response comparison for causal inference in subpopulation $\Omega_{\mathbf{h}}$. ∎

## Appendix B. Proof of Proposition 4.

First of all, condition (i) in Definition 2 holds as $\mathbf{X}\perp\!\!\!\perp T|h_3(\mathbf{X})$. Assuming $(Y_1,Y_0)\perp\!\!\!\perp T\,|\,\mathbf{X}$, it follows that $(Y_1,Y_0)\perp\!\!\!\perp T\,|\,h_3(\mathbf{X})$ under strong ignorability.

Now it suffices to verify condition (ii). Consider

$$\begin{aligned} E\{Y_1\,|\,h_2(\mathbf{X}),h_3(\mathbf{X})\} &= E\{Y_1\,|\,T=1,h_2(\mathbf{X}),h_3(\mathbf{X})\} \\ &= E\{Y\,|\,T=1,h_2(\mathbf{X}),h_3(\mathbf{X})\} \\ &= E\{E(Y|\mathbf{X},T=1)\,|\,h_2(\mathbf{X}),h_3(\mathbf{X})\} \\ &= E\{\gamma_0+\gamma_1+h_1(\mathbf{X})+h_2(\mathbf{X})\,|\,h_2(\mathbf{X}),h_3(\mathbf{X})\} \\ &= \gamma_0+\gamma_1+h_2(\mathbf{X})+E\{h_1(\mathbf{X})\,|\,h_2(\mathbf{X}),h_3(\mathbf{X})\} \end{aligned}$$

Similarly, it can be found that

$$E\{Y_0\,|\,h_2(\mathbf{X}),h_3(\mathbf{X})\}=\gamma_0+E(h_1(\mathbf{X})|h_2(\mathbf{X}),h_3(\mathbf{X})\}.$$

Thus,

$$E\{Y_1-Y_0\,|\,h_2(\mathbf{X})=h_2,\,h_3(\mathbf{X})=h_3\}=\gamma_1+h_2,$$

which is independent of $\mathbf{X}$. ∎

## Appendix C. Proof of Theorem 5.

The following lemma (see, e.g., Chapter 9 of Lin and Bai 2011), derived directly from $C_r$ inequality, will be used in the proof.

**Lemma 8** *Given a sequence $X_1,\ldots,X_n$ of random variables, $\bar{X}_n=\sum_{i=1}^n X_i/n$. Then*

$$E|\bar{X}_n|^r\leq\frac{1}{n}\cdot\sum_{i=1}^n E|X_i|^r \quad \textit{for } r>1.$$

By condition (ii) in Definition 2 of $\mathbf{a}(\mathbf{x})$,

$$\begin{aligned}(\bar{y}_{\tau 1}-\bar{y}_{\tau 0})-E(Y_1-Y_0\,|\,\bar{\mathbf{a}}_\tau) &= (\bar{y}_{\tau 1}-\bar{y}_{\tau 0})-E(Y_1-Y_0\,|\mathbf{x}) \\ &\quad +E(Y_1-Y_0|\mathbf{a}(\mathbf{x}))-E(Y_1-Y_0\,|\,\bar{\mathbf{a}}_\tau) \\ &= \{\bar{y}_{\tau 1}-E(Y_1|\mathbf{x})+E(Y_1\,|\,\mathbf{a}(\mathbf{x}))-E(Y_1|\bar{\mathbf{a}}_\tau)\}- \\ &\quad \{\bar{y}_{\tau 0}-E(Y_0|\mathbf{x})+E(Y_0\,|\,\mathbf{a}(\mathbf{x}))-E(Y_0|\bar{\mathbf{a}}_\tau)\} \\ &= \zeta_1-\zeta_0\end{aligned}$$

For convenience, we have used $\bar{\mathbf{a}}_\tau$ as shorthand for the conditioning event $\{\mathbf{a}(\mathbf{x})=\bar{\mathbf{a}}_\tau\}$. To prove (22), it suffices, by Minkowski's inequality, to verify the mean square or $L_2$ consistency for $\zeta_1$ and $\zeta_0$ separately.

Consider

$$\zeta_1=\{\bar{y}_{\tau 1}-E(Y_1|\mathbf{x})\}+\{E(Y_1\,|\,\mathbf{a}(\mathbf{x}))-E(Y_1|\bar{\mathbf{a}}_\tau)\},$$

which has two terms. We examine the second term $\{E(Y_1\,|\,\mathbf{a}(\mathbf{x}))-E(Y_1|\bar{\mathbf{a}}_\tau)\}$ first. If assumptions (19), (20), and (21) hold, then $\bar{\mathbf{a}}_\tau \xrightarrow{p} \mathbf{a}(\mathbf{x})$ by Theorem 12.7 of Breiman et al. (1984, p. 322). Since $E(Y_1|\mathbf{a})$ is assumed continuous in $\mathbf{a}$,

$$E(Y_1|\bar{\mathbf{a}}_\tau) \xrightarrow{p} E(Y_1|\mathbf{a}(\mathbf{x}))$$

by the continuous mapping theorem. Moreover, since $|E(Y_1|\bar{\mathbf{a}}_\tau)| \le E(|Y_1||\bar{\mathbf{a}}_\tau) \le E(|Y_1|) < \infty$, it follows that

$$\lim_n E\,|E(Y_1|\bar{\mathbf{a}}_\tau)-E(Y_1|\mathbf{a}(\mathbf{x}))|^2=0$$

by the dominated (or bounded) convergence theorem.

Next, consider the first term in $\zeta_1$, $\{\bar{y}_{\tau 1}-E(Y_1|\mathbf{x})\}$. Rewrite $\bar{y}_{\tau 1}$ as

$$\bar{y}_{\tau 1}=\frac{\sum_{i\in\tau} Y_i T_i}{\sum_{i\in\tau} T_i}=\frac{\sum_{i\in\tau} Y_i T_i/n_\tau}{\sum_{i\in\tau} T_i/n_\tau}=\frac{\xi_n}{\rho_n}.$$

which is a ratio estimator. The convergence of ratio estimators in the general form was studied by Cramér (1946). Using Theorem 12.7 of Breiman et al. (1984) again, we have $\xi_n \xrightarrow{p} E(YT|\mathbf{x})$ and $\rho_n \xrightarrow{p} e(\mathbf{x})$ in probability. Thus

$$\frac{\xi_n}{\rho_n} \xrightarrow{p} \frac{E(YT|\mathbf{x})}{e(\mathbf{x})}=E(Y_1|\mathbf{x})$$

in probability as well if $e(\mathbf{x})\neq 0$, under the assumption of strong ignorability. To establish its mean square risk consistency, the necessary and sufficient condition is that the random sequence $\{\bar{y}_{\tau 1}^2\}_n$ is uniformly integrable, that is,

$$\lim_{c_0\to\infty}\sup_n E\{\bar{y}_{\tau 1}^2 I(\bar{y}_{\tau 1}>c_0)\}=0.$$

A sufficient condition for uniform integrability (**?**) is that

$$\sup_n E|\bar{y}_{\tau 1}|^{2+\varepsilon}<\infty$$

for some $\varepsilon > 0$. This can be verified because

$$\sup_n E|\bar{y}_{\tau 1}|^{2+\varepsilon} \leq \sup_n \frac{1}{n_{\tau 1}} \sum_{\{i \in \tau:\ T_i=1\}} E|Y|^{2+\varepsilon} \leq M < \infty$$

following from (19) and Lemma 8.

Therefore, $\lim_n E|\zeta_1|^2 = 0$ using Minkowski's inequality again. Similar arguments can be used to show $\lim_n E|\zeta_0|^2 = 0$. This completes the proof of Theorem 5. ∎

## References


L. S. Aiken and S. G. West. *Multiple Regression: Testing and Interpreting Interactions*. Newbury Park, CA: Sage, 1991.

H. Akaike A new look at model identification. *IEEE Transactions an Automatic Control*, 19: 716–723, 1974.

P. C. Austin Using ensemble-based methods for directly estimating causal effects: An investigation of tree-based g-computation. *Multivariate Behavioral Research*, 47:115-135, 2012.

J. D. Angrist, W. Imbens, and D. B. Rubin. Identification of causal effects using instrumental variables. *Journal of the American Statistical Association*, 91:444–455, 1996.

L. Breiman. Random Forests. *Machine Learning*, 45:5–32, 2001.

L. Breiman, J. Friedman, R. Olshen, and C. Stone. *Classification and Regression Trees*. Belmont, CA: Wadsworth International Group, 1984.

A. Ciampi, S. A. Hogg, S. McKinney, and J. Thiffault. RECPAM: a computer program for recursive partition and amalgamation for censored survival data and other situations frequently occurring in biostatistics. I. Methods and program features. *Computer Methods and Programs in Biomedicine*, 26(3):239–256, 1988.

H. Cramér. *Mathematical Methods in Statistics*. Princeton, NJ: Princeton University Press, 1946.

R. H. Dehejia and S. Wahba. Causal effects in nonexperimental studies: Re-evaluating the evaluation of training programs. *Journal of the American Statistical Association*, 94:1053–1062, 1999.

M. Gail and R. Simon. Testing for qualitative interactions between treatment effects and patient subsets. *Biometrics*, 41:361–372, 1985.

L. Gordon and R. Olshen. Asymptotically efficient solutions to the classification problem. *The Annals of Statistics*, 6:515–533, 1978.

L. Gordon and R. Olshen. Consistent nonparametric regression from recursive partitioning schemes. *Journal of Multivariate Analysis*, 10:611–627, 1980.

L. Gordon and R. Olshen. Almost surely consistent nonparametric regression from recursive partitioning schemes. *Journal of Multivariate Analysis*, 15:147–163, 1984.

J. C. Gower. Some distance properties of latent root and vector methods used in multivariate analysis. *Biometrika*, 53:325–328, 1966.

S. Greenland. Quantifying biases in causal models: classical confounding vs collider-stratification bias. *Epidemiology*, 14:300-306, 2003.

K. Imai and D. A. van Dyk. Causal inference with general treatment regimes: generalizing the propensity score. *Journal of the American Statistical Association*, 99:854–866, 2004.

K. Hirano and G. W. Imbens. The propensity score with continuous treatments. In *Applied Bayesian Modeling and Causal Inference from Incomplete-Data Perspectives: An Essential Journey with Donald Rubin's Statistical Family*. New York, NY: Wiley, 2005.

D. Ho, K. Imai, G. King, and E. Stuart. Matching as nonparametric preprocessing for reducing model dependence in parametric causal inference. *Political Analysis*, 15:199–236, 2007.

D. Ho, K. Imai, G. King, and E. Stuart. `Matchit`: Nonparametric preprocessing for parametric causal inference. *Journal of Statistical Software*, 42(8), 2011. `http://gking.harvard.edu/matchit/`.

P. W. Holland. Statistics and causal inference (with discussion). *Journal of the American Statistical Association*, 81:945–970, 1986.

P. W. Holland and D. B. Rubin. Causal inference in retrospective studies. *Evaluation Review*, 12: 203–231, 1988.

D. Horvitz and D. Thompson. A Generalization of sampling without replacement from a finite population. *Journal of the American Statistical Association*, 47:663–685, 1952.

G. Imbens. The role of the propensity score in estimating dose-response functions. *Biometrika*, 87: 706-710, 2000.

K. K. Jain. *Textbook of Personalized Medicine*. New York, NY: Springer, 2009.

S. W. Lagakos. The challenge of subgroup analyses – reporting without distorting. *The New England Journal of Medicine*, 354:1667–1669, 2006.

R. J. LaLonde. Evaluating the econometric evaluations of training programs with experimental data. *American Economic Review*, 76:604–620, 1986.

M. LeBlanc and J. Crowley. Survival trees by goodness of split. *Journal of the American Statistical Association*, 88:457–467, 1993.

M. Lechner. Identification and estimation of causal effects o fmultiple treatments under the conditional independence assumption. In *Econometric Evaluations of Active Labour Market Policies in Europe*, Ed. M. Lechnwe and F. Pfeiffer. Heidelberg: Physica, 1999.

B. K. Lee, J. Lessler, and E. A. Stuart. Improving propensity score weighting using machine learning. *Statistics in Medicine*, 29:337–46, 2010.

K.-C. Li. Sliced inverse regression for dimension reduction (with discussion). *Journal of the American Statistical Association*, 86:316–342, 1991.

Z. Y. Lin and Z. D. Bai. *Probability Inequalities*. Beijing: Sience Press & Berlin Heidelgerg: Springer-Verlag, 2011.

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

D. F. McCaffrey, G. Ridgeway, and A. R. Morral. Propensity score estimation with boosted regression for evaluating causal effects in observational studies. *Psychological Methods*, 9:403–425, 2004.

J. Morgan and J. Sonquist. Problems in the analysis of survey data and a proposal. *Journal of the American Statistical Association*, 58: 415–434, 1963.

J. Neyman. On the application of probability theory to agricultural experiments: Essay on Principles, Section 9. Translated in *Statistical Science*, 5:465–480, 1923.

J. Pearl. *Casuality: Models, Reasoning, and Inference.* Cambridge University Press, 2000.

J. M. Robins. The analysis of randomized and non-randomized AIDS treatment trials using a new approach to causal inference in longitudinal studies. In *Health Service Research Methodology: A Focus on AIDS*. Eds: Sechrest L., Freeman H., Mulley A. Washington, D.C.: U.S. Public Health Service, National Center for Health Services Research, pp. 113–159, 1989.

J. M. Robins. Marginal structural models versus structural nested models as tools for causal inference. In *Statistical Models in Epidemiology: The Environment and Clinical Trials*. Eds: M.E. Halloran and D. Berry, IMA Volume 116, NY: Springer-Verlag, pp. 95-134, 1999.

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

P. R. Rosenbaum. Optimal matching for observational studies. *Journal of the American Statistical Association*, 84:1024–1032, 1989.

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

M. Rosenblum and M. J. van der Laan. Simple examples of estimating causal effects using targeted maximum likelihood estimation. Johns Hopkins University, Dept. of Biostatistics Working Papers. Working Paper 209, 2011. `http://biostats.bepress.com/jhubiostat/paper209`

D. B. Rubin. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66:688–701, 1974.

D. B. Rubin. Inference and missing data. *Biometrika*, 63:581–592, 1976.

D. B. Rubin. Assignment of treatment group on the basis of a covariate. *Journal of Educational Statistics*, 2:1–26, 1977.

D. B. Rubin. Bayesian inference for causal effects: The role of randomization. *The Annals of Statistics*, 7:34–58, 1978.

D. B. Rubin. Causal inference using potential outcomes: Design, modeling, decisions. *Journal of the American Statistical Association*, 100:322–331, 2005.

J. L. Schafer and J. Kang. Average causal effects from nonrandomized studies: a practical guide and simulated example. *Psychological Methods*, 13:279–313, 2008.

G. Schwarz. Estimating the dimension of a model. *Annals of Statistics*, 6: 461–464, 1978.

R. J. Serfling. *Approximation Theorems of Mathematical Statistics*. New York, NY: Wiley-Interscience, 2001.

I. Shpitser and J. Pearl. Complete identification methods for the causal hierarchy. *Journal of Machine Learning Research*, 9:1941–1979, 2008.

J. M. Snowden, S. Rose, and K. M. Mortimer. Implementation of g-computation on a simulated data set: Demonstration of a causal inference technique. *American Journal of Epidemiology*, 173:731–738, 2011.

P. Spirtes, C. N. Glymour, and R. Scheines. *Causation, Prediction, and Search*. 2nd Edition. The MIT Press, 2001.

X. G. Su, C.-L. Tsai, H. Wang, D. M. Nickerson, and B. G. Li. Subgroup analysis via recursive partitioning. *Journal of Machine Learning Research*, 10: 141–158, 2009.

X. G. Su, M. Wang, and J. Fan. Maximum likelihood regression trees. *Journal of Computational and Graphical Statistics*, 13:586–598, 2004.

R. Tibshirani, G. Walther, and T. Hastie. Estimating the number of clusters in a data set via the GAP statistic. *Journal of the Royal Statistical Society, Series B*, 63: 411–423, 2001.

D. Toth and J. L. Eltinge. Building consistent regression trees from complex sample data. Research paper, Office of Survey Methods Research (OSMR), U.S. Bureau of Labor Statistics, 2010. `http://www.bls.gov/osmr/pdf/st100010.pdf`.

W. S. Torgerson. *Theory and Methods of Scaling*. New York: Wiley, 1958.

M. J. van der Laan, E. C. Polley, and A. E. Hubbard. Super learner. *Statistical Applications in Genetics and Molecular Biology*, 6.1, 2007. `http://works.bepress.com/mark_van_der_laan/201`

M. J. van der Laan and D. Rubin. Targeted maximum likelihood learning. U.C. Berkeley Division of Biostatistics Working Paper Series. Working Paper 213, 2006. `http://biostats.bepress.com/ucbbiostat/paper213`.

T. J. VanderWeele and J. M. Robins. Four types of effect modification: A classification based on directed acyclic graphs. *Epidemiology*, 18:561–568, 2007.

T. J. VanderWeele. On the distinction between interaction and effect modification. *Epidemiology*, 20:863–871, 2009.

H. Wang and Y. Xia. Sliced regression for dimension reduction. *Journal of the American Statistical Association*, 103:811–821, 2008.

J. Wang. Consistent selection of the number of clusters via cross validation. *Biometrika*, 97:893–904, 2010.