Research Article

Steven Siwei Ye*, Yanzhen Chen, and Oscar Hernan Madrid Padilla

2D score-based estimation of heterogeneous treatment effects

https://doi.org/10.1515/jci-2022-0016
received February 25, 2022; accepted June 22, 2023

Abstract: Statisticians show growing interest in estimating and analyzing heterogeneity in causal effects in observational studies. However, there usually exists a trade-off between accuracy and interpretability for developing a desirable estimator for treatment effects, especially in the case when there are a large number of features in estimation. To make efforts to address the issue, we propose a score-based framework for estimating the conditional average treatment effect (CATE) function in this article. The framework integrates two components: (i) leverage the joint use of propensity and prognostic scores in a matching algorithm to obtain a proxy of the heterogeneous treatment effects for each observation and (ii) utilize nonparametric regression trees to construct an estimator for the CATE function conditioning on the two scores. The method naturally stratifies treatment effects into subgroups over a 2d grid whose axis are the propensity and prognostic scores. We conduct benchmark experiments on multiple simulated data and demonstrate clear advantages of the proposed estimator over state-of-the-art methods. We also evaluate empirical performance in real-life settings, using two observational data from a clinical trial and a complex social survey, and interpret policy implications following the numerical results.

Keywords: observational data, subgroup treatment effects, regression tree, matching

MSC 2020: 62D20, 62G05

1 Introduction

The questions that motivate many scientific studies in disciplines such as economics, epidemiology, medicine, and political science are not associational but causal in nature. In the study of causal inference, many researchers are interested in inferring average treatment effects, which provide a good sense of whether treatment is likely to deliver more benefit than the control among a whole community. However, the same treatment may affect different individuals very differently. Therefore, a substantial amount of works focus on analyzing heterogeneity in treatment effects, of which the term refers to variation in the effects of treatment across individuals. This variation may provide theoretical insights, revealing how the effect of interventions depends on participants' characteristics or how varying features of a treatment alters the effect of an intervention.

In this article, we follow the potential outcome framework for causal inference [1,2], where each unit is assigned into either the treatment or the control group. Each unit has an observed outcome variable with a set of covariates. In randomized experiments and observational studies, it is desirable to replicate a sample as closely as possible by obtaining subjects from the treatment and control groups with similar covariate distributions when estimating causal effects. However, it is almost impossible to match observations exactly the

* **Corresponding author: Steven Siwei Ye**, Department of Statistics, University of California, Los Angeles, USA,
e-mail: stevenysw@g.ucla.edu
Yanzhen Chen: Department of ISOM, Hong Kong University of Science and Technology, Clear Water Bay, Hong Kong,
e-mail: imyanzhen@ust.hk
Oscar Hernan Madrid Padilla: Department of Statistics, University of California, Los Angeles, USA,
e-mail: oscar.madrid@stat.ucla.edu

same in both treatment and control groups in observational studies. To address this problem, it is usually preferred to define prespecified subgroups under certain conditions and estimate the treatment effects varying among subgroups. Accordingly, the conditional average treatment effect (CATE [3]) function is designed to capture heterogeneity of a treatment effect across subpopulations. The function is often conditioned on some component(s) of the covariates or a single statistic, like propensity score [4] and prognostic score [5]. Propensity scores are the probabilities of receiving the treatment of interest; prognostic scores model the potential outcome under a control group assignment. To understand treatment effect heterogeneity in terms of propensity and prognostic scores, we assume that equal or similar treatment effects are observed along some intervals of the two scores.

We target at constructing an estimator for treatment effects that conditions on only two quantities, propensity and prognostic scores, and we assume a piecewise constant structure in treatment effects. We take a step further from score-based matching algorithms and propose a data-driven approach that integrates the joint use of propensity and prognostic scores in a matching algorithm and a partition over the entire population via a nonparametric regression tree. In the first step, we estimate propensity scores and prognostic scores for each observed unit in the data. Secondly, we perform a K-nearest-neighbour matching of units of the treatment and control groups based on the two estimated scores and forth construct a proxy of individual treatment effects for all units. The last step involves growing a binary tree regressed on the two estimated scores.

The complementary nature of propensity and prognostic score methods supports that conditioning on both the propensity and prognostic scores has the potential to reduce bias and improve the precision of treatment effect estimates, and it is affirmed in the simulation studies by Leacy and Stuart [6] and Antonelli et al. [7]. We also demonstrate such advantage for our proposed estimator across almost all scenarios examined in the simulation experiments.

Besides high precision in estimation, our proposed estimator demonstrates its superiority over the state-of-arts methods with a few attractive properties as follows:

- The estimator is computationally efficient. Propensity and prognostic scores can be easily estimated through simple regression techniques. Our matching algorithm based on the two scores largely reduces dimensionality compared to full matching on the complete covariates. Moreover, growing a single regression tree saves much time over other tree-based estimation methods, such as Bayesian additive regression tree models (BART) [8] and random forests [9, 10].
- Many previous works in subgroup analysis, such as studies by Assmann et al. [11] and Abadie et al. [12], set stratification on the sample with a fixed number of subgroups before estimating treatment effects. These approaches require a pre-determination on the number of subgroups contained in the data, and they inevitably introduce arbitrariness into the causal inference. In comparison, our proposed method simultaneously identifies the underlying subgroups in observations through binary split according to propensity and prognostic scores and provides a consequential estimation of treatment effects on each subgroups.
- Although random forests-based methods [9, 10] achieve great performance in minimizing bias in estimating treatment effects, these ensemble methods are often referred to as "black boxes." It is hard to capture the underlying reason why the collective decision with the high number of operations involved is made in their estimation process, especially in the case when there are many features considered in the model. On the contrary, our proposed method provides a straightforward and low-dimensional summary of treatment effects simply based on two scores that are relatively easy to estimate. As a result, given the covariates of an observation, one can easily deduce the positiveness and magnitude of its treatment effect according to its probability of treatment receipt and potential outcome following the structure of the regression tree without carrying out feature selection process.
- Instead of relying on a single feature or a subset of features, identifying subgroups with the joint distribution of all features will be of particular interest to policymakers seeking to target policies on those most likely to benefit. Therefore, many existing researches focus on subgroup stratification based on a single index that combines baseline characteristics. For instance, Abadie et al. [12] utilized potential outcomes without treatment (prognostic scores) for endogenous stratification, and Kent and Hayward [13] stratified experimental subjects based on their predicted probability of certain risks (propensity scores). Our method can be

considered as a natural extension of the two previous works, which allows researchers and policy makers to merge a large number of baseline characteristics into simply two indices and identify the subgroup who are most in need of help in a population.

We review relevant literature on matching algorithms and estimation of heterogeneous treatment effects in Section 2. In Section 3, we provide the theoretical framework and preliminaries for the causal inference model. We propose our method for estimation and prediction in Section 4. Section 5 lists the results of numerical experiments on multiple simulated data sets and two real-world data sets, following with the comparison with state-of-the-art methods in existing literature and the discussion on policy implications under different realistic scenarios.

2 Relevant literature

Statistical analysis of causality can be dated back to Neyman [1]. Causal inference can be viewed as an identification problem [14], for which statisticians are dedicated to learn the true causality behind the data. In reality, however, we do not have enough information to determine the true value due to a limited number of observations for analysis. This problem is also summarized as a “missing data” problem [15], which stems from the *fundamental problem of causal inference* [16], that is, for each unit at most one of the potential outcomes is observed. Importantly, the causal effect identification problem, especially for estimating treatment effects, can only be resolved through assumptions. Several key theoretical frameworks have been proposed over the past decades. The potential outcome framework by Rubin [2], often referred to as the Rubin causal model [16], is a common model of causality in statistics at the moment. Dawid [17] developed a decision theoretic approach to causality that rejects counterfactuals. Pearl [18,19] advocated for a model of causality based on non parametric structural equations and path diagrams.

2.1 Matching

To tackle the “missing data” problem when estimating treatment effects in randomized experiments in practice, matching serves as a very powerful tool. The main goal of matching is to find matched groups with similar or balanced observed covariate distributions [20]. The exact K-nearest-neighbour matching [2] is one of the most common and easiest to implement and understand methods; ratio matching [21–23], which finds multiple good matches for each treated individual, performs well when there is a large number of control individuals. Rosenbaum [24], Gu and Rosenbaum [25], and Zubizarreta [26,27] developed various optimal matching algorithms to minimize the total sum of distances between treated units and matched controls in a global sense. Abadie and Imbens [28] studied the consistency of covariate matching estimators under large sample assumptions. Instead of greedy matching on the entire covariates, propensity score matching (PSM) by Rubin and Thomas [29] is an alternative algorithm that does not guarantee optimal balance among covariates and reduces dimension sufficiently. Imbens [30] improved propensity score matching with regression adjustment. The additional matching on prognostic factors in propensity score matching was first considered by Rubin and Thomas [22]. Later, Leacy and Stuart [6] demonstrated the superiority of the joint use of propensity and prognostic scores in matching over single score-based matching in low-dimensional settings through extensive simulation studies. Antonelli et al. [7] extended the method to fit to high-dimensional settings and derived asymptotic results for the so-called doubly robust matching estimators (DRME). The sequential works by Aikens et al. [31] and Aikens and Baiocchi [32] pioneered in visualizing the relationship between propensity and prognostic scores as auxiliary in the estimation of treatment effects.

2.2 Subclassification

To understand heterogeneity of treatment effects in the data, subclassification, first used by Cochran [33], is another important research problem. The key idea is to form subgroups over the entire population based on characteristics that are either immutable or observed before randomization. Rosenbaum and Rubin [4,34], and Lunceford and Davidian [35] examined how creating a fixed number of subclasses according to propensity scores removes the bias in the estimated treatment effects, and Yang et al. [36] developed a similar methodology in settings with more than two treatment levels. Full matching [37–39] is a more sophisticated form of subclassification that selects the number of subclasses automatically by creating a series of matched sets. Schou and Marschner [40] presented three measures derived using the theory of order statistics to claim heterogeneity of treatment effect across subgroups. Su et al. [41] pioneered in exploiting standard regression tree methods [42] in subgroup treatment effect analysis. Further, Athey and Imbens [43] derived a recursive partition of the population according to treatment effect heterogeneity. Hill [44] was the first work to advocate for the use of BART [45] for estimating heterogeneous treatment effects, followed by a significant number of research papers focusing on the seminal methodology, including Green and Kern [46], Hill and Su [47], and Hahn et al. [8]. Abadie et al. [12] introduced endogenous stratification to estimate subgroup effects for a fixed number of subgroups based on certain quantiles of the prognostic score. More recently, Padilla et al. [48] combined the fused lasso estimator with score matching methods to lead to a data-adaptive subgroup effects estimator.

2.3 Machine learning for causal inference

For the goal of analyzing treatment effect heterogeneity, supervised machine learning methods play an important role. One of the more common ways for accurate estimation with experimental and observational data is to apply regression [49] or tree-based methods [50]. From a Bayesian perspective, Heckman et al. [51] provided a principled way of adding priors to regression models, and Taddy et al. [52] developed Bayesian nonparametric approaches for both linear regression and tree models. The recent breakthrough work by Wager and Athey [9] proposed the causal forest estimator arising from random forests from Breiman [53]. More recently, Athey et al. [10] took a step forward and enhanced the previous estimator based on generalized random forests. Imai and Ratkovic [54] adapted an estimator from the support vector machine classifier with hinge loss [55]. Bloniarz et al. [56] studied treatment effect estimators with lasso regularization [57] when the number of covariates is large, and Koch et al. [58] applied group lasso for simultaneous covariate selection and robust estimation of causal effects. In the meantime, a series of papers including Qian and Murphy [59], Kü nzel et al. [60], and Syrgkanis et al. [61] focused on developing meta-learners for heterogeneous treatment effects that can take advantage of various machine learning algorithms and data structures.

2.4 Applied work

On the application side, the estimation of heterogeneous treatment effects is particularly an intriguing topic in causal inference with broad applications in scientific research. Gaines and Kuklinski [62] estimated heterogeneous treatment effects in randomized experiments in the context of political science. Dehejia and Wahba [63] explored the use of propensity score matching for nonexperimental causal studies with application in economics. Dahabreh et al. [64] investigated heterogeneous treatment effects to provide the evidence base for precision medicine and patient-centred care. Zhang et al. [65] proposed the survival causal tree method to discover patient subgroups with heterogeneous treatment effects from censored observational data. Rekkas et al. [66] examined three classes of approaches to identify heterogeneity of treatment effect within a randomized clinical trial, and Tanniou et al. [67] rendered a subgroup treatment estimate for drug trials.

3 Preliminaries

Before we introduce our method, we need to provide some mathematical background for treatment effect estimation. We follow Rubin's framework on causal inference [2] and assume a superpopulation or distribution $\mathcal{P}$ from which a realization of n independent random variables is given as the training data. That is, we are given $\{(Y_i(0), Y_i(1), X_i, Z_i)\}_{i=1}^{n}$ independent copies of $(Y(1), Y(0), X, Z)$, where $X_i \in \mathbb{R}^d$ is a d-dimensional covariate or feature vector, $Z_i \in \{0, 1\}$ is the treatment-assignment indicator, $Y_i(0) \in \mathbb{R}$ is the potential outcome of unit i when i is assigned to the control group, and $Y_i(1)$ is the potential outcome when i is assigned to the treatment group.

One important and commonly used measure of causality in a binary treatment model is the average treatment effect (ATE; [30]), that is, the mean outcome difference between the treatment and control groups. Formally, we write the ATE as

$$\text{ATE} \coloneqq \mathbb{E}[Y(1) - Y(0)].$$

With the n units in the study, we further define the individual treatment effect (ITE) of unit i denoted by D_i as follows:

$$D_i \coloneqq Y_i(1) - Y_i(0).$$

Then, an unbiased estimate of the ATE is the sample average treatment effect

$$\bar{Y}(1) - \bar{Y}(0) = \frac{1}{n}\sum_{i=1}^{n} D_i.$$

However, we cannot observe D_i for any unit because a unit is either in the treatment group or in the control group, but not in both.

To analyze heterogeneous treatment effects, it is natural to divide the data into subgroups (e.g., by gender, or by race) and investigate if the average treatment effects are different across subgroups. Therefore, instead of estimating the ATE or the ITE directly, statisticians seek to estimate the conditional average treatment effect (CATE), defined by

$$\tau(x) \coloneqq \mathbb{E}[Y(1) - Y(0) \mid X = x]. \tag{1}$$

The CATE can be viewed as an ATE in a subpopulation defined by $\{X = x\}$, i.e. the ATE conditioned on membership in the subgroup.

We also recall the propensity score [4], denoted by $e(X)$, and defined as follows:

$$e(X) = \mathbb{P}(Z = 1 \mid X).$$

Thus, $e(X)$ is the probability of receiving treatment for a unit with covariate X. In addition, we consider prognostic scores, denoted by $p(X)$, for potential outcomes. The prognostic score is first defined by Hansen [5] as any quantity $p(X)$ that satisfies

$$Y(0) \perp\!\!\!\perp X \mid p(X),$$

and in this article, we use the conventional definition as the predicted outcome under the control condition [31]:

$$p(X) = E[Y(0) \mid X].$$

We are interested in constructing a 2d summary of treatment effects based on propensity and prognostic scores. Instead of conditioning on the entire covariates or a subset of it in the CATE function, we express our estimand, named as scored-based subgroup CATE, by conditioning on the two scores:

$$\tau(x) \coloneqq \mathbb{E}[Y(1) - Y(0) \mid e = e(x), p = p(x)]. \tag{2}$$

For interpretability, we assume that treatment effects are piecewise constant over a 2d grid of propensity and prognostic scores. Specifically, there exists a partition of intervals $\{I_1^e, \ldots, I_s^e\}$ of $[0, 1]$ and another partition of intervals $\{I_1^p, \ldots, I_t^p\}$ of $\mathbb{R}$ such that for any $i \in \{1, \ldots, s\}$ and $j \in \{1, \ldots, t\}$, we have

$$\tau(x) \equiv C_{i,j} \quad \text{for } x \quad \text{s.t. } e(x) \in I_i^e, p(x) \in I_j^p,$$

where $C_{i,j} \in \mathbb{R}$ is a constant.

Moreover, our estimation of treatment effects relies on the following assumptions:

Assumption 1. Throughout the article, we maintain the stable unit treatment value assumption (SUTVA [49]), which consists of two components: no interference and no hidden variations of treatment. Mathematically, for unit $i = 1, \ldots, n$ with outcome Y_i and treatment indicator Z_i, it holds that

$$Y_i(Z_1, Z_2, \ldots, Z_n) = Y_i(Z_i).$$

Thus, the SUTVA requires that the potential outcomes of one unit should be unaffected by the particular assignment of treatments to the other units. Furthermore, for each unit, there are no different forms or versions of each treatment level, which lead to different potential outcomes.

Assumption 2. The assumption of probabilistic assignment holds. This requires the assignment mechanism to imply a nonzero probability for each treatment value, for every unit. For the given covariates X and treatment-assignment indicator Z, we must have

$$0 < \mathbb{P}(Z = 1 \mid X) < 1,$$

almost surely.

This condition, regarding the joint distribution of treatments and covariates, is also known as overlap in some literature (see Assumption 2.2 in the studies by Imbens [30] and D'Amour et al. [68]), and it is necessary for estimating treatment effects everywhere in the defined covariate space. Note that $\mathbb{P}(Z_i = 1 \mid X_i)$ is the propensity score. In other words, Assumption 2 requires that the propensity score, for all values of the treatment and all combinations of values of the confounders, be strictly between 0 and 1.

Assumption 3. We make the assumption that

$$(Y(0), Y(1)) \perp\!\!\!\perp Z \mid e(X), p(X)$$

holds.

This assumption is an implication of the usual unconfoundedness assumption:

$$(Y(0), Y(1)) \perp\!\!\!\perp Z \mid X. \tag{3}$$

Combining Assumption 2 and that in Equation (3), the conditions are typically referred as *strong ignorability* defined in Rosenbaum and Rubin [4]. Strong ignorability states which outcomes are observed or missing is independent of the missing data conditional on the observed data. It allows statisticians to address the challenge that the "ground truth" for the causal effect is not observed for any individual unit. We rewrite the conventional assumption by replacing the vector of covariates x with the joint of propensity score $e(x)$ and $p(x)$ to accord with our estimation target.

Provided that Assumptions 1–3 hold, it follows that

$$\mathbb{E}[Y(z) \mid e = e(x), p = p(x)] = \mathbb{E}[Y \mid e = e(x), p = p(x), Z = z],$$

and thus our estimand (2) is equivalent to

$$\tau(x) = \mathbb{E}[Y \mid e = e(x), p = p(x), Z = 1] - \mathbb{E}[Y \mid e = e(x), p = p(x), Z = 0]. \tag{4}$$

Thus, in this article, we focus on estimating (4), which is equivalent to (2) if the aforementioned assumptions hold, but might be different if Assumption 3 is violated.

4 Methodology

We now formally introduce our proposal of a three-step method for estimating heterogeneous treatment effects and the estimation rule for a given new observation. We assume a sample of size n with covariate X, treatment indicator Z, and outcome variable Y, where the notations are inherited from the previous section. Generally, we consider a low-dimensional setup, where the sample size n is larger than the covariate dimension d. An extension of our proposed method to the high-dimensional case is discussed in this section as well.

4.1 Step 1

We first estimate propensity and prognostic scores for all observations in the sample. For propensity score, we apply a logistic regression on the entire covariate X and the treatment indicator Z by solving the optimization problem

$$\hat{\alpha} = \operatorname*{arg\,min}_{\alpha \in \mathbb{R}^d} - \sum_{i=1}^{n} [Z_i \log \sigma(X_i^\top \alpha) + (1 - Z_i) \log(1 - \sigma(X_i^\top \alpha))], \tag{5}$$

where $\sigma(x) = \frac{1}{1 + \exp(-x)}$ is the logistic function. With the coefficient vector $\hat{\alpha}$, we compute the estimated propensity scores $\hat{e}$ by

$$\hat{e}_i = \sigma(X_i^\top \hat{\alpha}).$$

For prognostic score, we restrict to the controlled group and regress the outcome variable Y on the covariate X through ordinary least squares: we solve

$$\hat{\theta} = \operatorname*{arg\,min}_{\theta \in \mathbb{R}^d} \sum_{i:Z_i=0} (Y_i - X_i^\top \theta)^2, \tag{6}$$

and we estimate prognostic scores as

$$\hat{p}_i = X_i^\top \hat{\theta}.$$

4.2 Step 2

Next, we perform a nearest-neighbour matching based on the two estimated scores from the previous step. We adapt the notation from Abadie and Imbens [28], and use the Mahalanobis distance norm as the distance metric in the matching algorithm. Mahalanobis distance matching, first proposed by Hansen [69], has been widely used in matching algorithms for causal inference [6,31]. It has been found to perform well when the dimension of covariates are relatively low [20], and thus, it is more suitable than using standard Euclidean distance in our case when we have only propensity scores and prognostic scores as covariates.

Formally, for the units i and j with estimated propensity scores $\hat{e}_i$, $\hat{e}_j$ and propensity scores $\hat{p}_i$, $\hat{p}_j$, we define the score-based Mahalanobis distance between i and j by

$$d(i,j) = \left[\begin{pmatrix} \hat{e}_i \\ \hat{p}_i \end{pmatrix} - \begin{pmatrix} \hat{e}_j \\ \hat{p}_j \end{pmatrix}\right]^\top \Sigma^{-1} \left[\begin{pmatrix} \hat{e}_i \\ \hat{p}_i \end{pmatrix} - \begin{pmatrix} \hat{e}_j \\ \hat{p}_j \end{pmatrix}\right],$$

where Σ denotes the variance-covariance matrix of $(\hat{e}, \hat{p})^\top$.

Let $j_k(i)$ be the index $j \in \{1, 2, \ldots, n\}$ that solves $Z_j = 1 - Z_i$ and

$$\sum_{l:Z_l=1-Z_i} \mathbf{1}\{d(l, i) \leq d(j, i)\} = k,$$

where $\mathbf{1}\{\cdot\}$ is the indicator function. This is the index of the unit that is the kth closest to unit i in terms of the distance between two scores, among the units with the treatment opposite to that of unit i. We can now construct the K-nearest-neighbour set for unit i by the set of indices for the first K matches for unit i,

$$\mathcal{J}_K(i) = \{j_1(i), \ldots, j_K(i)\}.$$

We then compute

$$\tilde{Y}_i = (2Z_i - 1)\left(Y_i - \frac{1}{K}\sum_{j\in\mathcal{J}_K(i)} Y_j\right). \tag{7}$$

Intuitively, the construction of $\tilde{Y}$ gives a proxy of the ITE on each unit. We find K matches for each unit in the opposite treatment group based on the similarity of their propensity and prognostics scores, and the mean of the K matches is used to estimate the unobserved potential outcome for each unit.

4.3 Step 3

The last step involves denoising of the point estimates of the individual treatment effects $\tilde{Y}$ obtained from Step 2. The goal is to partition all units into subgroups such that the estimated treatment effects would be constant over some 2d intervals of propensity and prognostic scores (see the left of Figure 1).

To perform such stratification, we grow a regression tree on $\tilde{Y}$, denoted by T, and the regressors are the estimated propensity scores $\hat{e}$ and the estimated prognostic scores $\hat{p}$ from Step 1. We follow the very general rule of binary recursive partitioning to build the tree T: allocate the data into the first two branches, using every possible binary split on every covariate; select the split that minimizes Gini impurity, and continue the optimal splits over each branch along the covariate's values until the minimum node size is reached. To avoid overfitting, we set the minimum node size as 20 in our model. Choosing other criteria such as information gain instead of Gini impurity is another option for splitting criteria. A 10-fold cross validation is also performed at meantime to prune the large tree T for deciding the value of cost complexity. Cost complexity is the minimum improvement in the model needed at each node. The pruning rule follows that if one split does not improve the overall error of the model by the chosen cost complexity, then that split is decreed to be not worth pursuing (see more details in Section 9.2 in the study by Hastie et al. [70]).

The final tree T (see the right plot of Figure 1) contains a few terminal nodes, and these are the predicted treatment effects for all units in the data. The values exactly represent a piecewise constant stratification over the 2d space of propensity and prognostic scores.

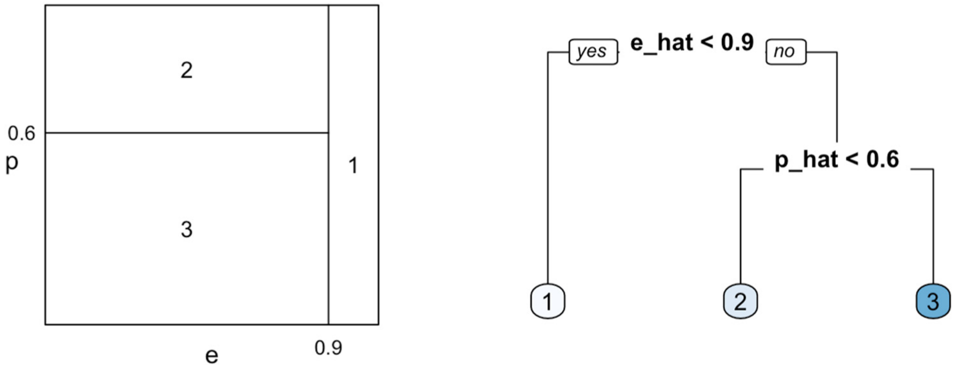

Figure 1: *Left:* a hypothetical partition over the 2d space of propensity and prognostic scores with the true values of piecewise constant treatment effects; *Right:* a sample regression tree T constructed in Step 3.

4.4 Estimation on a new unit

After we obtain the regression tree model T in Step 3, we can now estimate the value of the individual treatment effect corresponding to a new unit with covariate x_{new}.

We first compute the estimated propensity and prognostic scores for the new observation by

$$\hat{e}_{\text{new}} = \sigma(x_{\text{new}}^{\top}\hat{\alpha}), \quad \hat{p}_{\text{new}} = x_{\text{new}}^{\top}\hat{\theta},$$

where $\hat{\alpha}$ and $\hat{\theta}$ are the solutions to equations (5) and (6), respectively. Then with the estimated propensity score $\hat{e}_{\text{new}}$ and prognostic score $\hat{p}_{\text{new}}$, we can obtain an estimate of the treatment effect for this unit following the binary predictive rules in the tree T.

4.5 High-dimensional estimator

In a high-dimensional setting where the covariate dimension d is much larger than the sample size n, we can estimate the propensity and prognostic scores by adding a lasso (l1-based) penalty [57] instead. This strategy was first proposed and named as DRME by Antonelli et al. [7]. The corresponding optimization problems for the two scores can be written as follows:

$$\hat{\alpha} = \underset{\alpha \in \mathbb{R}^d}{\arg\min} - \sum_{i=1}^{n} [Y_i \log \sigma(X_i^{\top}\alpha) + (1 - Y_i) \log(1 - \sigma(X_i^{\top}\alpha))] + \lambda_1 \sum_{j=1}^{d} |\alpha_j|,$$

$$\hat{\theta} = \underset{\theta \in \mathbb{R}^d}{\arg\min} \sum_{i:Z_i=0} (Y_i - X_i^{\top}\theta)^2 + \lambda_2 \sum_{j=1}^{d} |\theta_j|.$$

The selection of the tuning parameters λ_1 and λ_2 can be determined by any information criteria (AIC, BIC, etc.). In practice, we use 10-fold cross-validation (CV) to select the value of λ. Then we perform the K-nearest-neighbour matching based on propensity and prognostic scores to obtain the estimates of individual treatment effects using equation (7).

We extend the aforementioned estimator with our proposal of applying a regression tree on the estimated propensity and prognostic scores. The procedure for estimation of subgroup heterogeneous treatment effects and estimation on a new unit remains the same as Step 3 for low-dimensional setups.

Remark. The choice of the number of nearest neighbours is a challenging problem. Updating distance metrics for every observation is computationally expensive, and choosing a value that is too small leads to a higher influence of noise on estimation. With regards to the application of nearest neighbour matching in causal inference, Abadie and Imbens [28] derived large sample properties of matching estimators of average treatment effects with a fixed number of nearest neighbours, but the authors did not provide any details on how to select the exact number of neighbours. Conventional settings on the number of nearest neighbours in current literature is to set $K = 1$ (one-to-one matching [20, 71]). However, Ming and Rosenbaum [72] suggested that in observational studies, substantially greater bias reduction is possible through matching with a variable number of controls rather than exact pair matching.

In Appendix A, we conduct a simulation study following one of the generative models from Section 5 to show how sensitive estimation accuracy is to the number of nearest neighbours selected and setting K to a large number other than 1 is more “sensible” to reduce estimation bias. Although it is usually difficult to select a perfect value of K in practice, simply setting $K \approx \log(n)$ as suggested by Brito et al. [73] leads to reasonable results for a data sample of size n. Throughout all our experiments in the next section, setting K to the integer closest to the value $\log(n)$ provides estimates with high accuracy and does not require too much computational cost.

4.6 Computational complexity

Our method is composed of three steps as introduced before. We first need to implement a logistic regression for estimating propensity score for a sample with size n and ambient dimension d. To solve a logistic regression optimization problem involves using the proximal Newton method, and its computational complexity is of # of iterations $\cdot\, O(nd)$ [74]. In general, the algorithm converges after a small number of iterations, and we can safely assume it is of a constant order. As a result, the overall time complexity of solving a logistic regression is $O(nd)$. The estimation of prognostic scores also requires a computational cost of $O(nd)$, and for high-dimensional settings, the cost of solving the solution via coordinate descent remains at $O(nd)$ [75]. The complexity of a K-nearest-neighbour matching based on the two estimated scores in the second step is of $O(Kn)$ [76], and the selection of $K \approx \log(n)$ leads to a complexity of $O(n \log n)$. In the third step, we grow a regression tree based on two estimated scores, and it requires a computational complexity of $O(n \log n)$.

The overall computational complexity of our method depends on the comparison between the order of d and $\log(n)$. Our method attains a computational complexity of $O(nd)$ if the order of d is greater than that of $\log(n)$. Otherwise, the complexity becomes $O(n \log n)$.

5 Experiments

In this section, we will examine the performance of our proposed estimator (PP) in a variety of simulated and real data sets. The baseline estimators we compete against are leave-one-out endogenous stratification (ST; [12]), causal forest (CF [9]), single-score matching including PSM, and prognostic-score matching. Note that in the original research by Abadie et al. [12], the authors restricted their attention to randomized experiments, because this is the setting where endogenous stratification is typically used. However, they mentioned the possibility of applying the method on observational studies. We take this into consideration and make their method as one of our competitors.

We implement our methods in R, using the packages "MatchIt" for K-nearest-neighbour Mahalanobis distance matching and "caret" for growing a nonparametric regression tree with cross validation over complexity parameter. Throughout, we set the number of nearest neighbours, K, to be the closest integer to $\log(n)$, where n is the sample size. For causal forest, we directly use the R package "grf" developed by Athey et al. [10], following with an automatic cross-validation to select values of hyperparameters [77] and a K-fold cross-fitting under a conventional setting of $K = 10$ recommended by Nie and Wager [78]. Software that replicate all the simulations is available on the authors' Github page.

We evaluate the performance of each method according to two aspects, accuracy and uncertainty quantification. The results for single-score matching algorithms are not reported in this article because of very poor performance throughout all scenarios.

5.1 Simulated data

We first examine on the following simulated data sets under six different data generation mechanisms. We obtain insights from the simulation study by Leacy and Stuart [6] for the models considered in Scenarios 1–3. The propensity score and outcome (prognosis) models in Scenarios 1 and 3 are characterized by additivity and linearity (main effects only), but with different piecewise constant structures in the true treatment effects over a 2d grid of the two scores. We add nonadditivity and nonlinear terms to both propensity and prognosis models in Scenarios 2. In other words, both propensity and prognostic scores are expected to be misspecified in these two models if we apply generalized linear models directly in estimation. Scenario 4 comes from Abadie et al. [12], with a constant treatment effect over all observations. Scenario 5 is modified from the study by Wager and Athey [9] (see Equation 27 there), in which the propensity model follows a continuous distribution instead of a linear structure. In addition, the true treatment effect is not a function of propensity and prognostic scores. In Scenario 6, we break the mold assumption of nicely delineated subgroups in treatment effects over 2d space of the two scores, and use a smooth function instead. A high-dimensional setting ($d \gg n$) is examined in Scenario 7, where the generative model inherits from Scenario 1. We also include a table summarizing all simulation scenarios in Appendix B to help the audience interpret the result of the simulation study more readily.

We first introduce some notations used in the experiments: the sample size n, the ambient dimension d, as well as the following functions:

$$\begin{aligned}&\text{true treatment effect: } \tau^*(X) = \mathbb{E}[Y(1) - Y(0)|X],\\&\text{treatment propensity: } e(x) = \mathbb{P}(Z = 1|X = x),\\&\text{treatment logit: } \operatorname{logit}(e(x)) = \log\left(\frac{e(x)}{1 - e(x)}\right).\end{aligned}$$

Throughout all the models we consider, we maintain the unconfoundedness assumption discussed in Section 3, generate the covariate X following a certain distribution, and entail homoscedastic Gaussian noise ε.

We evaluate the accuracy of an estimator $\tau(X)$ by the mean-squared error for estimating $\tau^*(X)$ at a random example X, with $m = n/10$, defined by

$$\operatorname{MSE}(\hat{\tau}(X)) \coloneqq \frac{1}{m}\sum_{i=1}^{m}[\hat{\tau}_i(X) - \tau_i^*(X)]^2.$$

We record the averaged mean squared error (MSE) over 1,000 Monte Carlo trials for each scenario. For each Monte Carlo trial, we use train-test split to evaluate model performance. In detail, we first obtain n training sample and train models on this set. We then generate a separate testing sample under the same data generation mechanism with sample size $n/10$. We make predictions on the test data using the trained models and compute the MSEs from different models accordingly. In terms of uncertainty quantification, we measure the coverage probability of $\tau(X)$ with a target coverage rate of 0.95. For endogenous stratification and our proposed method, we use nonparametric bootstrap with replacement to construct the empirical quantiles for each unit. The details on the implementation of nonparametric bootstrap methods are presented in Appendix C. For causal forest, we construct 95% confidence intervals by estimating the standard errors of estimation using the “grf” package. We do not include the results for Scenario 7, a high-dimensional setting, because the asymptotic normality guarantees of nonparametric bootstrap and random forests are not valid in high dimensions [9,79].

Scenario 1. With $d \in \{2, 10, 50\}$, $n \in \{1{,}000, 5{,}000\}$, for $i = 1, \ldots, n$, we generate the data as follows:

$$\begin{aligned}&Y_i = p(X_i) + Z_i \cdot \tau_i^* + \varepsilon_i,\\&\tau_i^* = \mathbf{1}_{\{e(X_i)<0.6, p(X_i)<0\}},\\&\operatorname{logit}(e(X_i)) = X_i^\top\beta^e,\\&p(X_i) = X_i^\top\beta^p,\\&X_i \overset{\text{i.i.d.}}{\sim} \mathcal{U}[0, 1]^d,\\&\varepsilon_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, 1),\end{aligned}$$

where β^e and β^p are randomly generated from $\mathcal{U}[-1, 1]^d$ in each iteration.

Scenario 2. We now add some interaction and nonlinear terms to the propensity and prognostic models in Scenario 1, while keeping the setups of the covariate X, the response Y, the true treatment effect τ^*, and the error term ε unchanged. We set $d = 10$ and $n = 5{,}000$ in this case.

$$\begin{aligned}\operatorname{logit}(e(X_i)) = {} & X_i^\top\beta^e + 0.5X_{i1}X_{i3} + 0.7X_{i2}X_{i4} + 0.5X_{i3}X_{i5}\\ & + 0.7X_{i4}X_{i6} + 0.5X_{i5}X_{i7} + 0.5X_{i1}X_{i6}\\ & + 0.7X_{i2}X_{i3} + 0.5X_{i3}X_{i4} + 0.5X_{i4}X_{i5}\\ & + 0.5X_{i5}X_{i6} + X_{i2}^2 + X_{i4}^2 - X_{i7}^2\\ p(X_i) = {} & X_i^\top\beta^p + 0.5X_{i1}X_{i3} + 0.7X_{i2}X_{i4} + 0.5X_{i3}X_{i8},\\ & +0.7X_{i4}X_{i9} + 0.5X_{i8}X_{i10} + 0.5X_{i1}X_{i9}\\ & + 0.7X_{i2}X_{i3} + 0.5X_{i3}X_{i4} + 0.5X_{i4}X_{i8}\\ & + 0.5X_{i8}X_{i9} + X_{i3}^2 + X_{i4}^2 - X_{i10}^2.\end{aligned}$$

Again for each simulation, β^e and β^p are randomly generated from $\mathcal{U}[-1,1]^d$.

Scenario 3. In this case, we define the true treatment effect with a more complicated piecewise constant structure over the 2d grid, under the same model used in Scenario 1, with $d = 10$ and $n = 5{,}000$:

$$\tau_i^* = \begin{cases} 0 & \text{if } e(X_i) \le 0.6, p(X_i) \le 0,\\ 1 & \text{if } e(X_i) \le 0.6, p(X_i) > 0 \quad \text{or} \quad e(X_i) > 0.6, p(X_i) \le 0,\\ 2 & \text{if } e(X_i) > 0.6, p(X_i) > 0.\end{cases}$$

Scenario 4. Setting $d = 10$ and $n = 5{,}000$, the data are generated as follows:

$$\begin{aligned} & Y_i = 1 + \beta^\top X_i + \varepsilon_i,\\ & X_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \mathbf{I}_{d\times d}),\\ & \varepsilon_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, 100 - d),\end{aligned}$$

where $\beta = (1, \ldots, 1)^\top \in \mathbb{R}^d$. Moreover, the treatment indicators for the simulations are such that $\sum_i Z_i = \lceil n/2 \rceil$. By construction, the vector of treatment effects satisfies $\tau^* = 0$.

Scenario 5. The data satisfy

$$\begin{aligned} \tau_i^* &= (X_{i1} + X_{i2} + \ldots + X_{i10})/10,\\ Y_i &= 2X_i^\top\mathbf{e}_1 - 1 + Z_i \cdot \tau_i^* + \varepsilon_i,\\ Z_i &\sim \operatorname{Binom}(1, e(X_i)),\\ X_i &\overset{\text{i.i.d.}}{\sim} \mathcal{U}[0,1]^d,\\ e(X_i) &= \frac{1}{4}[1 + \beta_{2,4}(X_i^\top\mathbf{e}_1)],\\ \varepsilon_i &\overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1),\end{aligned}$$

where $\mathbf{e}_1 = (1, 0, \ldots, 0)^\top$. We compare the results of different methods under the setting: $d = 10, n = 5{,}000$.

Scenario 6. For this scenario, we examine the performance when the treatment effect takes on a smooth interval of propensity and prognostic scores. We use the same model in Scenario 1 for generating the two scores with $d = 10$ and $n = 5{,}000$, and we inherit and modify the generative function (28) from the study by Wager and Athey [9] for treatment effects as follows:

$$\tau_i^* = \zeta(e(X_i)) \cdot \zeta(p(X_i)), \quad \zeta(x) = 1 + \frac{1}{1 + e^{-20(x-1/3)}}.$$

Scenario 7. In the last case, we study the performance of different estimators on a high-dimensional data. The data model follows

$$\begin{aligned} & X_i \overset{\text{i.i.d.}}{\sim} \mathcal{U}[0,1]^d,\\ & Y_i = p(X_i) + Z_i \cdot \tau_i^* + \varepsilon_i,\\ & \tau_i^* = \mathbf{1}_{\{e(X_i)<0.6, p(X_i)<0\}},\\ & \operatorname{logit}(e(X_i)) = 0.4X_{i1} + 0.9X_{i2} - 0.4X_{i3} - 0.7X_{i4} - 0.3X_{i5} + 0.6X_{i6},\\ & p(X_i) = 0.9X_{i1} - 0.9X_{i2} + 0.2X_{i3} - 0.2X_{i4} + 0.9X_{i5} - 0.9X_{i6},\\ & \varepsilon_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1).\end{aligned}$$

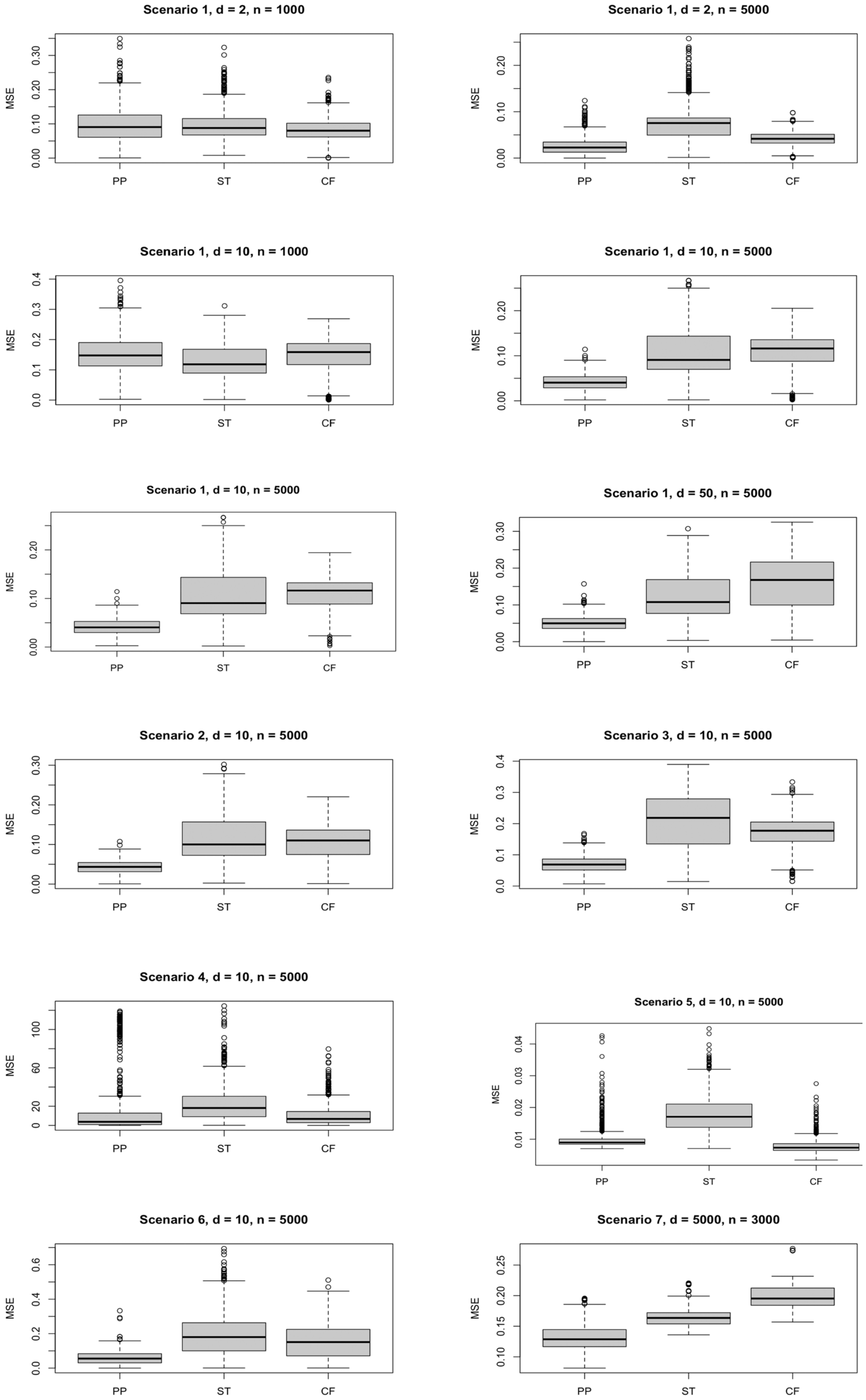

Figure 2: Comparison of mean squared errors over 1,000 Monte Carlo simulations under different generative models for our proposed method (PP), endogenous stratification (ST), and causal forest (CF).

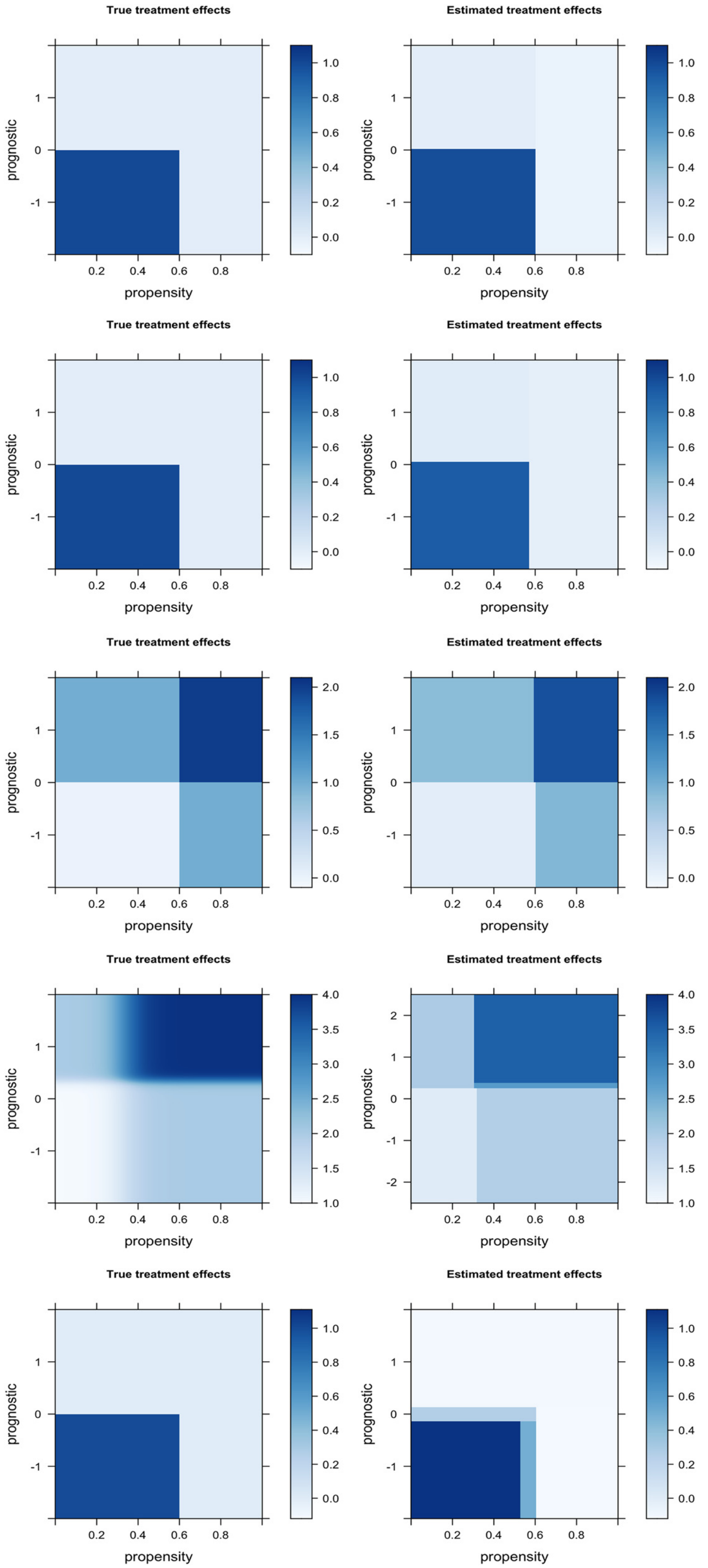

Figure 3: One instance comparison between the true treatment effects and the estimates from the score-based method for Scenarios 1, 2, 3, 6, and 7 (from the top to the bottom). Plots on the left column depict the true treatment effect on the 2d grid of propensity and prognostic scores, and the right plots demonstrate the estimated treatment effects from our proposed method over the same grid.

We select $n = 3{,}000$ and $d = 5{,}000$ for analysis.

The boxplots that depict the distribution of MSEs obtained under all scenarios are presented in Figure 2. We can see that for Scenario 1, our proposed estimator achieves better accuracy when the sample size n is large, and it is the best among the three estimators in these cases. The good performance of our method is consistent when we assume a more complex partition on the defined 2d grid as in Scenario 3. In addition, variation in accuracy, measured by the difference between the upper quartiles and the lower quartiles (also referred as the interquartile ranges) in each boxplot becomes smaller, accompanied with the increase in n. The reason for the enhanced performance when applying our method on a large-size data is the nature of nearest neighbour algorithm, which benefits more in learning the boundaries of clusters accurately with a larger sample size. Moreover, causal random forest suffers more in accuracy especially as d gets larger [9], while our method does not impact significantly because it focuses on a low-dimensional space of the two scores instead of the overall covariate space. In Scenario 2, we introduce nonadditivity and nonlinear terms into the data model. Although linear assumptions are violated for both propensity and prognostic models, our method performs better compared to the other two methods regarding accuracy and variability. For a potential outcome model with randomized assignment of treatment and constant treatment effects, as in Scenario 4, our method has the best accuracy among all candidates, even though large noise is added to the true signal. In Scenario 5, our estimator performs slightly worse than causal forests in accuracy since the true treatment function is not directly dependent on propensity and prognostic scores, but it still outperforms endogenous stratification. However, for the previous two cases, our method shows larger variation in accuracy than the others. One of the reasons for this phenomenon is due to the misspecification in estimating the true scores and the large variance in constructing tree models. When the assumption of a sharp stratification of treatment effects in the 2d space of the two scores is invalid, as shown in Scenario 6, our method maintains its superiority over the benchmarks. With subgroups identified over the 2d interval of the scores, the mean squared errors are reduced through our estimator comparably better than the other methods. In a high-dimensional setting such as Scenario 7, we consider modified methods with lasso-regularized regressions for both our methodology and endogenous stratification, and our method remains its high performance as in the low-dimensional setups since the piecewise constant structure in treatment effects in the setting does not change.

We now take a careful look at the visualization comparison between the true treatment effect and the predictions obtained from our method for Scenarios 1, 2, 3, 6, and 7. We confine both the true signal and the predictive model in a 2d grid scaled on the true propensity and prognostic scores, as shown in Figure 3. It is not surprising that our proposed estimators provide a descent recovery of the piecewise constant partition in the true treatment effects over the 2d grid as in Scenarios 1, 2, 3, and 7, with only a small difference in the magnitude of treatment effects. In the setting when treatment effects are smoothly distributed across the grid of the two scores as in Scenario 6, our estimation does not exactly capture the soft boundary of the shape, but still provides us with a simplified delineation that help us understand general heterogeneity in treatment effects over the 2d space.

Table 1: Reported coverage rates with a target confidence level of 0.95 and within iteration standard error estimates (in parentheses)

Scenario	n	d	PP	ST	CF
1	1,000	2	0.996 (0.177)	0.748 (0.158)	0.969 (0.128)
	5,000	10	0.987 (0.116)	0.731 (0.062)	0.736 (0.061)
2	1,000	10	0.994 (0.230)	0.753 (0.152)	0.797 (0.157)
	5,000	10	0.956 (0.141)	0.619 (0.064)	0.829 (0.224)
3	1,000	10	0.989 (0.382)	0.420 (0.163)	0.769 (0.297)
	5,000	10	0.949 (0.164)	0.307 (0.075)	0.750 (0.252)
4	1,000	10	1.000 (28.621)	0.933 (16.578)	0.993 (18.067)
	5,000	10	1.000 (6.693)	0.998(6.049)	0.973 (5.957)
5	1,000	10	1.000 (0.524)	0.950 (0.353)	0.978 (0.429)
	5,000	10	0.845 (0.336)	0.622 (0.287)	0.716 (0.321)
6	1,000	10	0.995 (0.472)	0.495 (0.161)	0.865 (0.474)
	5,000	10	0.999 (0.179)	0.578 (0.089)	0.821 (0.259)

With regard to uncertainty quantification, we examine coverage rates with a target confidence level of 0.95 for each method under different scenarios, and the corresponding results, along with the within iteration standard error estimates, are recorded in Table 1. It is quite clear that our proposed method achieves nominal coverage over the other two methods in almost all scenarios. It is worth to notice that although our estimator achieves a better coverage rate than the benchmarks, it suffers from large variation due to the nature of nearest-neighbour matching, especially in the case with small sample size. In general, our method that incorporates both propensity and prognostic scores in identifying subgroups and estimating treatment effects show its advantage in reducing bias, but the resulting large variance is worrisome and it remains space for future study to solve this issue.

We also compare the computational cost of our proposed scored-based method and causal forest, with simulated data from Scenario 1. Endogenous stratification is not considered in this comparison because the method exerts an arbitrary subclassification into a predetermined number of subgroups, and it does not involve a cross-validation process. We record the averaged time consumed to obtain the estimate over 1,000 simulations for both methods. Figure 4 demonstrates that our proposed method (PP) is comparably time-efficient over its competitor, and CF can be very expensive in operational time for large-size problems.

These results highlight the promise of our method for accurate estimation of subgroups in heterogeneous treatment effects, all while emphasizing avenues for further work. An immediate challenge is to control the bias in estimation of propensity and prognostic scores. Using more powerful models instead of simple linear regression in estimation is a good way to reduce bias by enabling the trees to focus more closely on the coordinates with the greatest signal. The study of splitting rules for trees designed to estimate causal effects is still in its infancy and improvements may be possible.

5.2 Real data analysis

To illustrate the behaviour of our estimator, we apply our method on the two real-world data sets, one from a clinical study and the other from a complex social survey. Propensity score-based methods are frequently used for confounding adjustment in observational studies, where baseline characteristics can affect the outcome of policy interventions. Therefore, the results from our method are expected to provide meaningful implications for these real data sets. However, one potential limitation of our proposed tree-based approach is that it imposes

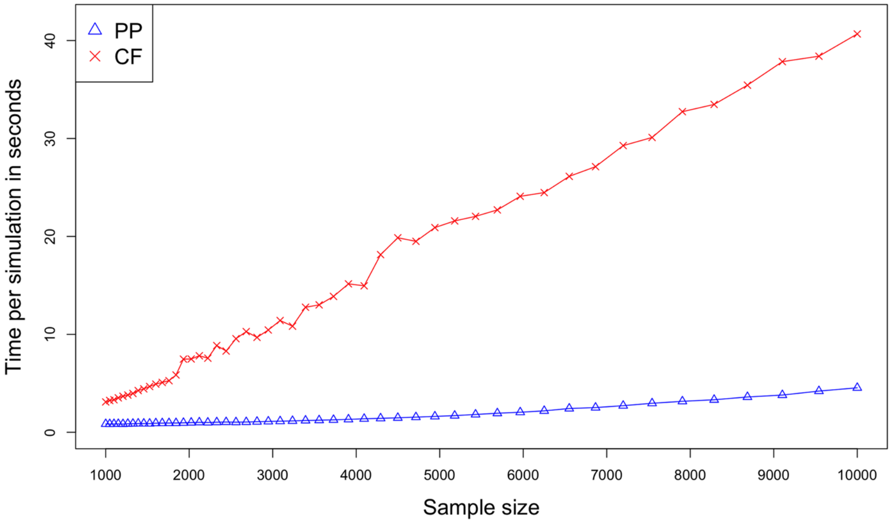

Figure 4: A plot of time per simulation in seconds for our proposed method (PP) and CF against problem size n (50 values from 1,000 up to 10,000). For each method, the time to compute the estimate for one simulated data is averaged over 1,000 Monte Carlo simulations.

arbitrary dichotomization with sharp thresholds and thus may obscure potential continuous gradient over the space. For real data analysis in general, our method provides a reasonable simplification that is easy to interpret for the audience and gives some framework for thinking about when such a simplification would be appropriate.

5.2.1 Right heart catheterization (RHC) analysis

While randomized control trials are widely encouraged as the ideal methodology for causal inference in clinical and medical research, the lack of randomized data due to high costs and potential high risks leads to the studies based on observational data. In this section, we are interested in examining the association between the use of RHC during the first 24 h of care in the intensive care unit (ICU) and the short-term survival conditions of the patients. RHC is a procedure for directly measuring how well the heart is pumping blood to the lungs. RHC is often applied to critically ill patients for directing immediate and subsequent treatment. However, RHC imposes a small risk of causing serious complications when administering the procedure. Therefore, the use of RHC is controversial among practitioners, and scientists want to statistically validate the causal effects of RHC treatments. The causal study using observational data can be dated back to Connors et al. [80], where the authors implemented propensity score matching and concluded that RHC treatment lead to lower survival than not performing the treatment. Later, Hirano and Imbens [81] proposed a more efficient propensity-score-based method, and the recent study by Loh and Vansteelandt [82] using a modified propensity score model suggested that RHC significantly affected mortality rate in a short-term period.

A dataset for analysis was first used in the study by Connors et al. [80], and it is suitable for the purpose of applying our method because of its extremely well-balanced distribution of confounders across levels of the treatment [83]. The treatment variable Z in the data indicates whether a patient received a RHC within 24 hours of admission. The binary outcome Y is defined based on whether a patient died at any time up to 180 days since admission. The original data consisted of 5,735 participants with 73 covariates. We preprocess the full data in the way suggested in the Hirano and Imbens [81] and Loh and Vansteelandt [82], by removing all observations that contain null values in covariates, dropping the singular covariate in the reduced data, and encoding categorical variables into dummy variables. The resulted data contains 2,707 observations and 72 covariates, with 1,103 in the treated group ($Z = 1$) and 1,604 in the control group ($Z = 0$). Among the 72 observed covariates, there are 21 continuous, 25 binary, and 26 dummy variables transformed from the original six categorical variables.

The result of the prediction model from our proposed method is reported in Figure 5. We observe that the sign of estimated treatment effects varies depending on the value of the propensity score and prognostic score. This particular pattern implies that RHC procedures indeed offer both benefits and risks in affecting patients'

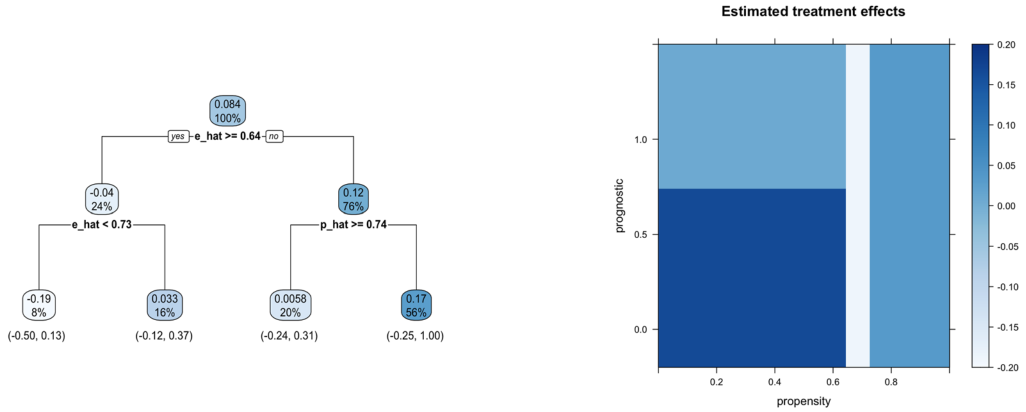

Figure 5: Prediction model for RHC data. *Left:* a tree diagram of predictions on each split based on the two scores, with 95% confidence intervals provided on the bottom. *Right:* a 2d plot of prediction stratification on the intervals of propensity and prognostic scores, with a colour scale on magnitude of treatment effects.

short-term survival conditions. Specifically, we are interested in the occurrence of large positive treatment effects (increase in chance of death) from the estimation. An estimated treatment effect of 0.17 (with a 95% confidence interval between −0.25 and 1.00) is observed on the group of patients with propensity scores less than 0.64 and prognostic scores less than 0.74, and this group accounts for 56% of the entire sample. For the rest of the sample, we observe relatively small treatment effects in magnitude. Under the scenario of RHC data, a smaller propensity score means that the patient is less likely to receive RHC procedures after admitting to the ICU, and it is related to the availability of RHC procedures at the hospital to which the patient is admitted. A smaller prognostic score tells that the patient has lower underlying chance of death. One possible explanation for this significant positive treatment on this certain group is that drastic change in treatment procedures that were applied to patients who do not actually need the aggressive style of care largely undermine patients' health conditions after admission and increase the mortality rate. In summary, our findings generally agree with the results and explanations presented in the study by Connors et al. [80], and they offer some insights for practitioners to decide whether they should apply RHC procedures to patients.

5.2.2 National medical expenditure survey (NMES)

For the next experiment, we analyze a complex social survey data. In many complex surveys, data are not usually well balanced due to potential biased sampling procedure. To incorporate score-based methods with complex survey data requires an appropriate estimation on propensity and prognostic scores. DuGoff et al. [84] suggested that combining a propensity score method and survey weighting is necessary to achieve unbiased treatment effect estimates that are generalizable to the original survey target population. Austin et al. [85] conducted numerical experiments and showed that greater balance in measured baseline covariates and decreased bias is observed when natural retained weights are used in propensity score matching. Therefore, we include sampling weight as an baseline covariate when estimating propensity and prognostic scores in our analysis.

In this study, we aim to answer the research question: how smoking habit affects medical expenditure over lifetime, and we use the same data set as given in the study by Johnson et al. [86], which is originally extracted from the 1987 NMES. The NMES included detailed information about frequency and duration of smoking with a large nationally representative data base of nearly 30,000 adults, and that 1987 medical costs are verified by multiple interviews and additional data from clinicians and hospitals. A large amount of literature focus on applying various statistical methods to analyze the causal effects of smoking on medical expenditures using the NMES data. In the original study by Johnson et al. [86], the authors first estimated the effects of smoking on certain diseases and then examined how much those diseases increased medical costs. In contrast, Rubin [87], Imai and van Dyk [88], and Zhao et al. [89] proposed to directly estimate the effects of smoking on medical expenditures using propensity-score-based matching and subclassification. Hahn et al. [8] applied Bayesian regression tree models to assess heterogeneous treatment effects.

For our analysis, we explore the effects of extensive exposure to cigarettes on medical expenditures, and we use pack-year as a measurement of cigarette measurement, the same as given in the studies by Imai and van Dyk [88] and Hahn et al. [8]. Pack-year is a clinical quantification of cigarette smoking used to measure a person's exposure to tobacco, defined by

$$\text{Pack-year} = \frac{\text{Number of cigarettes per day}}{20} \times \text{Number of years smoked}.$$

Following that, we determine the treatment indicator Z by the question whether the observation has a heavy lifetime smoking habit, which we define to be greater than 17 pack-years, the equivalent of 17 years of pack-a-day smoking.

The subject-level covariates X in our analysis include age at the times of the survey (between 19 and 94), age when the individual started smoking, gender (male, female), race (white, black, other), marriage status (married, widowed, divorced, separated, never married), education level (college graduate, some college, high school graduate, other), census region (Northeast, Midwest, South, West), poverty status (poor, near poor, low income, middle income, high income), seat belt usage (rarely, sometimes, always/almost always), and sample weight. We select the natural logarithm of annual medical expenditures as the outcome variable Y to maintain the assumption of heteroscedasticity in random errors. We preprocess the raw data set by omitting any observations with missing values in the covariates and excluding those who had zero medical expenditure. The resulting restricted data set contains 7,903 individuals, with 4,014 in the treated group (Z = 1) and 3,889 in the controlled group (Z = 0).

The prediction model obtained from our method, as shown in Figure 6, is simple and easy to interpret. We derive a positive treatment effect across the entire sample, and the effect becomes significant when the predicted potential outcome is relatively low (less than 5.7). These results indicate that more reliance on smoking will deteriorate one's health condition, especially for those who currently do not have a large amount of medical expenditure. Moreover, we observe a significant positive treatment effect of 1.4 (with a 95% confidence interval between 0.56 and 2.79), and in other words, a certain and substantial increase in medical expenditure for the subgroup with propensity score less than 0.17. It is intuitive to assume that a smaller possibility of engaging in excessive tobacco exposure is associated with healthier living styles. This phenomenon is another evidence that individuals who are more likely to stay healthy may suffer more from excessive exposure to tobacco products. In all, these results support policymakers and social activists who advocate for nationwide smoking ban.

6 Discussion

Our method is different from existing methods on estimating heterogeneous treatment effects in a way that we incorporate both matching algorithms and nonparametric regression trees in estimation, and the final estimate can be regarded as a 2d summary on treatment effects. Moreover, our method exercises a simultaneous stratification across the entire population into subgroups with the same treatment effects. Subgroup treatment effect analysis is an important but challenging research topic in observational studies, and our method can be served as an efficient tool to reach a reasonable partition.

Our numerical experiments on various simulated and real-life data lay out empirical evidence of the superiority of our estimator over state-of-the-art methods in accuracy and our proposal provides insightful interpretation on heterogeneity in treatment effects across different subgroups. We also discovered that our method is powerful in investigating subpopulations with significant treatment effects. Identifying representative subpopulations that receive extreme results after treatment is a paramount task in many practical

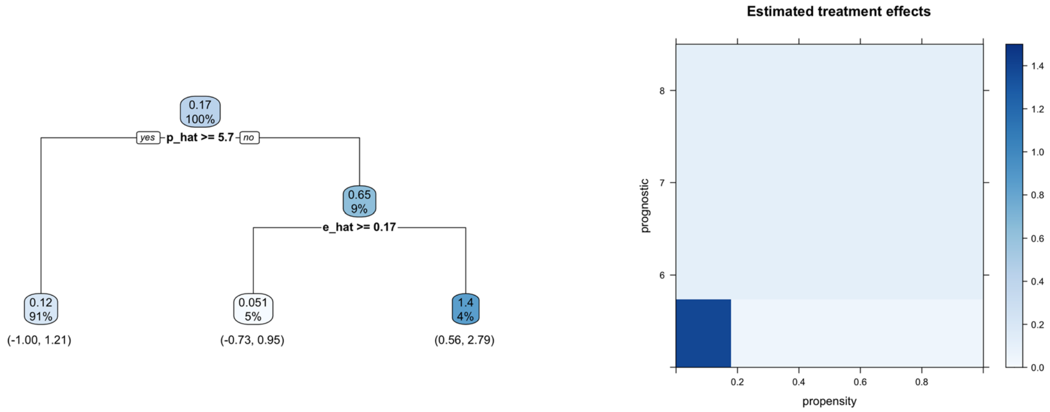

Figure 6: Prediction model for NMES data. *Left:* a tree diagram of predictions on each split based on the two scores, with 95% confidence intervals provided on the bottom. *Right:* a 2d plot of prediction stratification on the intervals of propensity and prognostic scores, with a colour scale on magnitude of treatment effects.

contexts. Through empirical experiments on two real-world datasets from observational studies, our method demonstrates its ability in identifying these significant effects.

Although our method shows its outstanding performance in estimating treatments effects under the piecewise constant structure assumption, it remains meaningful and requires further study to develop more accurate recovery of such structure. For example, a potential shortcoming of using conventional regression trees for subclassification is that the binary partition over the true signals is not necessarily unique. Using some variants of CART, like optimal trees [90] and dyadic regression trees [91], would be more appropriate for estimation under additional assumptions. Another particular concern with tree-based methods is the curse of dimensionality. Applying other nonparametric regression techniques, such as K-nearest-neighbour fused lasso [92] and locally estimated scatterplot smoothing, may offer a compromised solution to deal with this issue and learn a more smoothed structure in treatment effects other than a rectangular partition on 2d data. It is also worth improving the estimation of propensity and prognostic scores using similar nonparametric based methods if a piecewise constant assumption hold for the two scores as well.

Moreover, we acknowledge that fitting propensity score and prognostic scores normally would demand more than main terms linear and logistic regressions as used in the current method to have hope of eliminating bias in estimating the two scores. The key takeaway of our proposal is not to estimate the two scores more accurately, but to provide a lower-dimensional summary on heterogeneous treatment effects. If more foreknowledge on the structure of the underlying functions of the two scores can be incorporated, we can certainly select a more ideal algorithm or machine learning model for estimation. To compromise the lack of such information under a usual scenario, a SuperLearner ensemble estimator, proposed by van der Laan et al. [93], may offer a more promising solution for estimating the two scores by taking a weighted average of multiple prediction models. Yet, an important challenge for future work is to design rules that can automatically choose the best model to fit the data.

Another question to consider is the source of heterogeneity in causal effects. Our proposal that utilizes two scores can be seen as a reasonable starting point to analyze heterogeneity from a low-dimensional perspective, but it cannot comprehensively capture the sources of such effects. For instance, it is seemingly possible to exist a covariate that leads to substantial heterogeneous treatment effects on the response but does not itself have a large effect on propensity or prognostic scores. There is no particular guiding theories about this issue in current literature, and thus, it remains huge space for further investigation.

Acknowledgments: The authors would like to thank three anonymous referees and the Editor for their constructive comments that improved the quality of this article.

Funding information: Steven Siwei Ye was partially funded with the UCLA 2022–23 Faculty Career Development Award awarded to Oscar Madrid Padilla.

Author contributions: Steven Siwei Ye worked on all the different parts of the project, Oscar Madrid Padilla, and Yanzhen Chen worked in the methodology and applications.

Conflict of interest: The authors state no conflict of interest.

Data availability statement: The data and R codes for implementing the method introduced in the article are publicly available on one of the author's Github page (https://github.com/stevenysw/causal_pp).

References

[1] Neyman J. On the application of probability theory to agricultural experiments. Ann Agricult Sci. 1923;10:1–51.

[2] Rubin DB. Estimating causal effects of treatments in randomized and nonrandomized studies. J Educ Psychol. 1974;66(5):688–701.

[3] Hahn J. On the role of the propensity score in efficient semiparametric estimation of average treatment effects. Econometrica. 1998;66(2):315–31.

[4] Rosenbaum PR, Rubin DB. The central role of the propensity score in observational studies for causal effects. Biometrika. 1983;70(1):41–55.

[5] Hansen BB. The prognostic analogue of the propensity score. Biometrika. 2008;95(2):481–8.

[6] Leacy FP, Stuart EA. On the joint use of propensity and prognostic scores in estimation of the average treatment effect on the treated: a simulation study. Stat Med. 2014;33(20):3488–508.

[7] Antonelli J, Cefalu M, Palmer N, Agniel D. Doubly robust matching estimators for high dimensional confounding adjustment. Biometrics. 2018;74(4):1171–9.

[8] Hahn PR, Murray JS, Carvalho CM. Bayesian regression tree models for causal inference: regularization, confounding, and heterogeneous effects. Bayesian Anal. 2020;15(3):965–1056.

[9] Wager S, Athey S. Estimation and inference of heterogeneous treatment effects using random forests. J Amer Stat Assoc. 2018;113(523):1228–42.

[10] Athey S, Tibshirani J, Wager S. Generalized random forests. Ann Stat. 2019;47(2):1148–78.

[11] Assmann SF, Pocock SJ, Enos LE, Kasten LE. Subgroup analysis and other (mis)uses of baseline data in clinical trials. Lancet. 2000;355(9209):1064–9.

[12] Abadie A, Chingos MM, West MR. Endogenous stratification in randomized experiments. Rev Econ Stat. 2018;C(4):567–80.

[13] Kent DM, Hayward RA. Limitations of applying summary results of clinical trials to individual patients: the need for risk stratification. J Amer Med Assoc. 2007;298(10):1209–12.

[14] Keele L. The statistics of causal inference: a view from political methodology. Political Anal. 2015;23(3):313–35.

[15] Ding P, Li F. Causal inference: a missing data perspective. Stat Sci 2018;33(2):214–37.

[16] Holland PW. Statistics and causal Inference. J Amer Stat Assoc. 1986;81(396):945–60.

[17] Dawid AP. Causal inference without counterfactuals. J Amer Stat Assoc. 2000;95(450):407–24.

[18] Pearl J. Causal diagrams for empirical research. Biometrika 1995;82(4):669–710.

[19] Pearl J. Causality: models, reasoning, and inference. Cambridge, UK: Cambridge University Press; 2009.

[20] Stuart EA. Matching methods for causal inference: a review and a look forward. Stat Sci Rev J Inst Math Stat. 2010;25(1):1–21.

[21] Smith H. Matching with multiple controls to estimate treatment effects in observational studies. Sociol Methodol. 1997;27(1):325–53.

[22] Rubin DB, Thomas N. Combining propensity score matching with additional adjustments for prognostic covariates. J Amer Stat Assoc. 2000;95(450):573–85.

[23] Ming K, Rosenbaum PR. A note on optimal matching with variable controls using the assignment algorithm. J Comput Graph Stat. 2001;10(3):455–63.

[24] Rosenbaum PR. Optimal matching for observational studies. J Amer Stat Assoc.1989;84(408):1024–32.

[25] Gu XS, Rosenbaum PR. Comparison of multivariate matching methods: structures, distances, and algorithms. J Comput Graph Stat. 1993;2(4):405–20.

[26] Zubizarreta JR. Using mixed integer programming for matching in an observational study of kidney failure after surgery. J Amer Stat Assoc. 2012107(500):1360–71.

[27] Zubizarreta JR, Keele L. Optimal multilevel matching in clustered observational studies: a case study of the effectiveness of private schools under a large-scale voucher system. J Amer Stat Assoc. 2017;112(518):547–60.

[28] Abadie A, Imbens GW. Large sample properties of matching estimators for average treatment effects. Econometrica. 2006;74(1):235–67.

[29] Rubin DB, Thomas N. Matching using estimated propensity scores: relating theory to practice. Biometrics. 1996;52(1):249–64.

[30] Imbens GW. Nonparametric estimation of average treatment effects under exogeneity: a review. Rev Econ Stat. 2004;86:4–29.

[31] Aikens RC, Greaves D, Baiocchi M. A pilot design for observational studies: using abundant data thoughtfully. Stat Med. 2020;39(30):4821–40.

[32] Aikens RC, Baiocchi M. Assignment-control plots: a visual companion for causal inference study design. Amer Stat. 2023;77(1):72–84.

[33] Cochran WG. The effectiveness of adjustment by subclassification in removing bias in observational studies. Biometrics. 1968;24(2):295–313.

[34] Rosenbaum PR, Rubin DB. Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. Amer Stat. 1985;39(1):33–38.

[35] Lunceford JK, Davidian M. Stratification and weighting via the propensity score in estimation of causal treatment effects: a comparative study. Stat Med. 2004;23(19):2937–60.

[36] Yang S, Imbens GW, Cui Z, Faries DE, Kadziola Z. Propensity score matching and subclassification in observational studies with multi-level treatments. Biometrics. 2016;72(4):1055–65.

[37] Rosenbaum PR. A characterization of optimal designs for observational studies. J R Stat Soc Ser B (Stat Methodol). 1991;53(3):597–610.

[38] Hansen BB. Full matching in an observational study of coaching for the SAT. J Amer Stat Assoc. 2004;99(467):609–18.

[39] Stuart EA, Green KM. Using full matching to estimate causal effects in non-experimental studies: examining the relationship between adolescent marijuana use and adult outcomes. Development Psychol. 2008;44(2):395–406.

[40] Schou IM, Marschner IC. Methods for exploring treatment effect heterogeneity in subgroup analysis: an application to global clinical trials. Pharmaceut Stat. 2015;14(1):44–55.

[41] Su X, Tsai C, Wang H, Nickerson DM, Li B. Subgroup analysis via recursive partitioning. J Machine Learn Res. 2009;10:141–58.

[42] Breiman L, Friedman J, Stone CJ, Olshen RA. Classification and regression trees. New York: Taylor & Francis; 1984.

[43] Athey S, Imbens GW. Recursive partitioning for heterogeneous causal effects. Proc Nat Acad Sci. 2016;113(27):7353–60.

[44] Hill JL. Bayesian nonparametric modeling for causal inference. J Comput Graph Stat.. 2011;20(1):217–40.

[45] Chipman HA, George EI, McCulloch RE. BART: Bayesian additive regression trees. Ann Appl Stat. 2010;4(1):266–98.

[46] Green DP, Kern HL. Modeling heterogeneous treatment effects in survey experiments with Bayesian additive regression trees. Public Opinion Quarter. 2012;76(3):491–511.

[47] Hill JL, Su Y. Assessing lack of common support in causal inference using Bayesian nonparametrics: implications for evaluating the effect of breastfeeding on childrenas cognitive outcomes. Ann Appl Stat. 2013;7(3):1386–420.

[48] Padilla OHM, Ding P, Chen Y, Ruiz G. A causal fused lasso for interpretable heterogeneous treatment effects estimation. 2021. arXiv: 2110.00901.

[49] Imbens GW, Rubin DB. Causal inference for statistics, social, and biomedical sciences: an introduction. Cambridge, UK: Cambridge University Press; 2015.

[50] Imai K, Strauss A. Estimation of heterogeneous treatment effects from randomized experiments, with application to the optimal planning of the get-out-the-vote campaign. Politic Anal. 2011;19:1–19.

[51] Heckman JJ, Lopes HF, Piatek R. Treatment effects: a Bayesian perspective. Econom Rev. 2014;33(1–4):36–67.

[52] Taddy M, Gardner M, Chen L, Draper D. A nonparametric Bayesian analysis of heterogenous treatment effects in digital experimentation. J Business Econom Stat. 2016;34(4):661–72.

[53] Breiman L. Random forests. Machine Learning. 2001;45:5–32.

[54] Imai K, Ratkovic M. Estimating treatment effect heterogeneity in randomized program evaluation. Ann Appl Stat. 2013;7(1):443–70.

[55] Wahba G. Soft and hard classification by reproducing kernel Hilbert space methods. Proc Nat Acad Sci. 2002;99(26):16524–30.

[56] Bloniarz A, Liu H, Zhang C, Sekhon JS, Yu B. Lasso adjustments of treatment effect estimates in randomized experiments. Proc Nat Acad Sci. 2016;113(27):7383–90.

[57] Tibshirani R. Regression shrinkage and selection via the lasso. J R Stat Soc Ser B (Stat Meth) 1996;58(1):267–88.

[58] Koch B, Vock DM, Wolfson J. Covariate selection with group lasso and doubly robust estimation of causal effects. Biometrics. 2018;74(1):8–17.

[59] Qian M, Murphy SA. Performance guarantees for individualized treatment rules. Ann Stat 2011;39(2):1180–210.

[60] Kü nzel SR, Sekhon JS, Bickel PJ, Yu B. Metalearners for estimating heterogeneous treatment effects using machine learning. Proc Nat Acad Sci. 2019;116(10):4156–65.

[61] Syrgkanis V, Lei V, Oprescu M, Hei M, Oprescu M, Battocchi K, et al. Machine learning estimation of heterogeneous treatment effects with instruments. Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada. p. 15167–76.

[62] Gaines B, Kuklinski J. Estimation of heterogeneous treatment effects related to self-selection. Amer J Politic Sci. 2011;55(3):724–36.

[63] Dehejia RH, Wahba S. Propensity score matching methods for nonexperimental causal studies. Rev Econ Stat. 2002;84(1):151–61.

[64] Dahabreh IJ, Hayward R, Kent DM. Using group data to treat individuals: understanding heterogeneous treatment effects in the age of precision medicine and patient-centred evidence. Int J Epidemiol. 2016;45(6):2184–93.

[65] Zhang W, Le TD, Liu L, Zhou Z, Li J. Mining heterogeneous causal effects for personalized cancer treatment. Bioinformatics. 2017;33(15):2372–8.

[66] Rekkas A, Paulus JK, Raman G, Wong JB, Steyerberg EW, Rijnbeek PR, et al. Predictive approaches to heterogeneous treatment effects: a scoping review. BMC Med Res Methodol. 2020;20(1):264.

[67] Tanniou J, van der Tweel I, Teerenstra S, Roes KCB. Estimates of subgroup treatment effects in overall nonsignificant trials: to what extent should we believe in them? Pharmaceut Stat. 2017;16(4):280–95.

[68] D’Amour A, Ding P, Feller A, Lei L, Sekhon J. Overlap in observational studies with high-dimensional covariates. J Econ. 2021;221:644–54.

[69] Hansen BB. Bias reduction in observational studies via prognosis scores. Statistics Department, University of Michigan; Ann Arbor, Michigan: Technical Report. 2006. 441.

[70] Hastie T, Tibshirani R, Friedman J. The elements of statistical learning: data mining, inference, and prediction. New York: Springer; 2001.

[71] Austin PC, Schuster T. The performance of different propensity score methods for estimating absolute effects of treatments on survival outcomes: a simulation study. Stat Meth Med Res. 2016;25(5):2214–37.

[72] Ming K, Rosenbaum PR. Substantial gains in bias reduction from matching with a variable number of controls. Biometrics. 2001;56(1):118–24.

[73] Brito MR, Chávez EL, Quiroz AJ, Yukich JE. Connectivity of the mutual k-nearest-neighbour graph in clustering and outlier detection. Stat Probability Lett. 1997;35(1):33–42.

[74] Yuan G, Ho C, Lin C. An improved GLMNET for l1-regularized logistic regression. J Machine Learn Res. 2012;13:1999–2030.

[75] Friedman J, Hastie T, Tibshirani R. Regularization paths for generalized linear models via coordinate descent. J Stat Softw. 2010;33(1):1–22.

[76] von Luxburg U. A tutorial on spectral clustering. Stat Comput. 2007;17(4):395–416.

[77] Athey S, Wager S. Estimating treatment effects with causal forests: an application. Observ Stud. 2019;19(5):37–51.
[78] Nie X, Wager S. Quasi-oracle estimation of heterogeneous treatment effects. Biometrika. 2021;108(2):299–319.
[79] Karoui NE, Purdom E. Can we trust the bootstrap in high-dimensions? The case of linear models. J Machine Learn Res. 2018;19(5):1–66.
[80] Connors AF, Speroff T, Dawson NV, Thomas C, Harrell FE, Wagner D, et al. The effectiveness of right heart catheterization in the initial care of critically ill patients. J Amer Med Assoc. 1996;276(11):889–97.
[81] Hirano K, Imbens GW. Estimation of causal effects using propensity Score weighting: an application to data on Right Heart Catheterization. Health Services Outcomes Res Methodol. 2001;2(3):259–78.
[82] Loh WW, Vansteelandt S. Confounder selection strategies targeting stable treatment effect estimators. Stat Med. 2021;40(3):607–30.
[83] Smith MJ, Mansournia MA, Maringe C, Zivich PN, Cole SR, Leyrat C, et al. Introduction to computational causal inference using reproducible Stata, R, and Python code: A tutorial. Stat Med. 2021;41(2):407–32.
[84] DuGoff EH, Schuler M, Stuart EA. Generalizing observational study results: applying propensity score methods to complex surveys. Health Service Res. 2014;49(1):284–303.
[85] Austin PC, Jembere N, Chiu M. Propensity score matching and complex surveys. Stat Meth Med Res. 2018;27(4):1240–57.
[86] Johnson E, Dominici F, Griswold M, Zeger SL. Disease cases and their medical costs attributable to smoking: an analysis of the national medical expenditure survey. J Econometr. 2003;112(1):135–51.
[87] Rubin DB. Using propensity scores to help design observational studies: application to the tobacco litigation. Health Services Outcomes Res Methodol. 2001;2:169–88.
[88] Imai K, van Dyk DA. Causal inference with general treatment regimes. J Amer Stat Assoc. 2004;99(467):854–66.
[89] Zhao S, van Dyk DA, Imai K. Propensity score-based methods for causal inference in observational studies with non-binary treatments. Stat Meth Med Res 2020;29(3):709–27.
[90] Bertsimas D, Dunn J. Optimal classification trees. Machine Learn. 2017;106(7):1039–82.
[91] Donoho DL. CART and best-ortho-basis: a connection. Ann Stat. 1997;25(5):1870–911.
[92] Padilla OHM, Sharpnack J, Chen Y, Witten D. Adaptive non-parametric regression with the K-NN fused lasso. Biometrika. 2020;107(2):293–310.
[93] van der Laan MJ, Polley EC, Hubbard AE. Super learner. Stat Appl Genetic Mol Biol. 2007;6(1):1–23.
[94] Efron B. Bootstrap methods: another look at the Jackknife. Ann Stat. 1979;7(1):1–26.

Appendix A

Study on the choice of the number of nearest neighbours

In this section, we examine how the number of nearest neighbours in the matching algorithm affects the estimation accuracy. Recall in Step 2 of our proposed method, we implement a K-nearest-neighbour algorithm based on the two estimated scores for a sample of size n. The computational complexity of this K-NN algorithm is of $O(Kn)$. Although a larger K produce a promisingly higher estimation accuracy, more computational costs become the corresponding side effect. Moreover, the issue of "underfitting" – the increase of bias occurs when we make such selection. It is because a selection of too large K defeats the underlying principle behind nearest neighbour algorithms – that neighbours that are nearer share similar densities. Therefore, a smart choice of K is essential to balance the trade-off among accuracy, computational expense, and generability.

We follow the same generative model in Scenario 1 from Section 5 and compute the averaged mean squared error over 1,000 Monte Carlo simulations for $K = 1, \ldots, 50$ with a fixed sample size $n = 5{,}000$. The results in Figure 6 show that the averaged MSE continuously decreases as the number of nearest neighbours K selected in the matching algorithm grows. However, the speed of improvement in accuracy gradually slows down when K exceeds 10, which is close to log(5,000). This suggests that an empirical choice of $K \approx \log(n)$ is sufficient to produce a reasonable estimate on the target parameter, and this choice is more "sensible" than the conventional setting of $K = 1$ (A1).

Appendix B

A summary table of simulation studies

In this appendix, we include a table to provide a text summary of all simulation models considered in Section 5.1. We add some notes to help the audience understand similarities, differences, and sources of these data generative models.

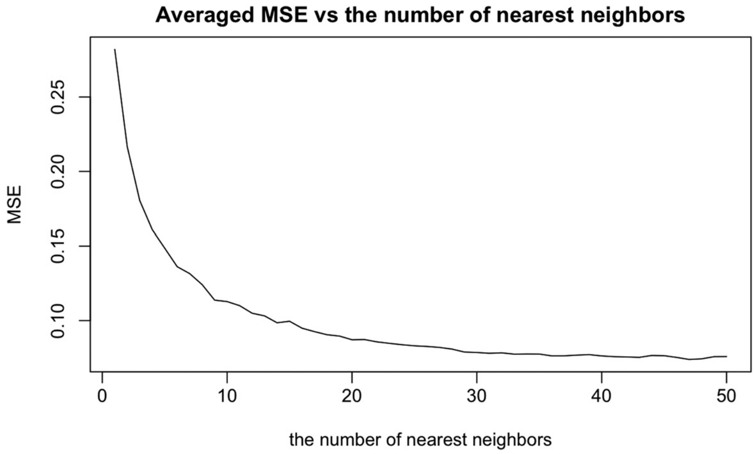

Figure A1: The plot of averaged MSE against the number of nearest neighbours.

Scenario	Propensity model	Prognostic model	Treatment effect model	Note
1	Linearity in covariates	Linearity in covariates	A sharp delineation over the 2d grid of propensity and prognostic scores with two subgroups	
2	Linear model with additional interactions and nonlinear terms	Linear model with additional interactions and nonlinear terms	A sharp delineation over the 2d grid of propensity and prognostic scoreswith two subgroups	The same treatment effect model as in Scenario 1; both propensity and prognostic scores are misspecified with linear model
3	Linearity in covariates	Linearity in covariates	A sharp delineation over the 2d grid of propensity and prognostic scores with four subgroups	The same propensity and prognostic models as in Scenario 1; more subgroups in treatment effects
4	Random with a fixed number of treated units	Linearity in covariates with an intercept	Constant zero	The same setting from Abadie et al. [12]
5	A smooth function of single covariate	Proportional to single covariate	A function of the covariates	Modified from Equation 27 of Wager and Athey [9]
6	Linearity in covariates	Linearity in covariates	A smooth function of the two scores	Modified from Equation 28 of Wager and Athey [9]
7	Linearity in a few covariates with sparsity in others	Linearity in a few covariates with sparsity in others	A sharp delineation over the 2d grid of propensity and prognostic scores with two subgroups	A high-dimensional setting corresponding to Scenario 1

Appendix C

Non-parametric bootstrap in simulation studies

In Section 5, we use nonparametric bootstrap to construct confidence intervals for endogenous stratification and our proposed method. We use these bootstrap samples to compute coverage rates with respect to a target level of 95% as a measurement of uncertainty. The bootstrap method, introduced by [94], is a simple but powerful tool to obtain a robust nonparametric estimate of the confidence intervals through sampling from the empirical distribution function of the observed data. It is worth noting that causal forests or generalized random forests can compute confidence intervals quickly without requiring a bootstrap approach on which our method relies.

In this appendix, we introduce the details on how we implement nonparametric bootstrap for the purpose of computing coverage rates in the simulation experiments. For each scenario in Section 5, we start with generating a sample following the defined data generation model with a sample size n. Next, we create 1,000 random samples with replacement from this single set of data, also with the sample size n. We then apply both methods on these simulation repetitions and obtain a series of estimations on each unit in the original set. Following these estimations, we calculate the corresponding 2.5 and 97.5% quantiles for all units in the original sample. Coverage rates of a 95% prediction level are thus the frequencies of the original units falling inside the intervals between the two quantiles computed in the previous step.