ERASMUS UNIVERSITEIT ROTTERDAM

Erasmus School of Economics

Bachelor Thesis Financial Econometrics

Heterogeneous Treatment Effects:

Comparing causal forests to other tree-based methods

Author:

Qamar Suleri

Student ID:

493256

Supervisor:

Jens Klooster

Second Assessor:

Eoghan O'Neill

July 4, 2021

Abstract

Adequate estimation of heterogeneous treatment effects is crucial in many different settings. Examples include personalised medicine, consumer-specific ad campaigns, and government/corporate policy evaluation. For this purpose, Wager and Athey (2018) have proposed causal forests, which have quickly gained popularity due to their shown theoretical properties. However, other tree-based methods have shown the potential to outperform causal forests. Using various simulated experiments, we find that local linear forests achieve performance better than or equal to causal forests, and X-learners based on random forests are superior when the treatment group is much smaller than the control group.

The views stated in this thesis are those of the author and not necessarily those of the supervisor, second assessor, Erasmus School of Economics or Erasmus University Rotterdam.

Contents

1 Introduction **2**

2 Methodology **3**

2.1 Potential outcomes framework 4

2.2 k-nearest neighbors (k-NN) 4

2.3 Causal forests (CF) 5

2.4 Local linear forests (LLF) 7

2.5 X-learners based on random forests (XRF) 10

2.6 X-learners based on Bayesian additive regression trees (XBART) 11

3 Simulation study **12**

4 Results **14**

5 Conclusion **19**

References **21**

Appendix **24**

1 Introduction

Computational capacity, algorithmic innovation and the amount of available data have all been growing exponentially (Kurzweil, 2010), resulting in a large machine learning repertoire. However, when regarding (causal) inference, this repertoire is far more limited. Nevertheless, using data to infer the causal effect of a certain treatment is useful in numerous different settings. Examples include medical studies developing personalized medicine (Neto et al., 2017; Yazdani and Boerwinkle, 2015), marketing-based studies regarding the impact of an ad campaign on consumer behavior (Norman et al., 2016; Tye et al., 1987; Varian, 2016) and economic studies evaluating corporation and government policies (Kreif and DiazOrdaz, 2019; Panizza and Presbitero, 2014; Yee, 1996).

When estimating heterogeneous treatment effects, the main challenge is that for each observation, only the outcome given the treatment status is observed, whereas this is not the case for the potential counterfactual. Accordingly, different nonparametric machine learning methods have been proposed, which include kernel smoothing methods, series methods, and nearest neighbor matching (Crump et al., 2008; Lee, 2009; Willke et al., 2012).

In this paper we focus on tree-based methods, as especially this literature has been expanding. For example, various papers regarding (modified) random forests (Breiman et al., 1984) have recently been published (Friedberg et al., 2020; Mentch and Hooker, 2016; Scornet et al., 2015; Wager and Athey, 2018). Causal forests have particularly become popular, as Wager and Athey (2018) show that predictions made by these adapted random forests are both asymptotically unbiased and normally distributed. They find causal forests to significantly outperform the aforementioned classical methods, particularly when the used data contains noisy covariates. However, using local linear forests, which aim to better capture local trends in the data generating process (DGP) by combining local linear regressions with random forests, Friedberg et al. (2020) show that causal forests can potentially be outperformed.

Moreover, Künzel et al. (2017) have introduced X-learners which are algorithms that can build on tree-based methods in such a manner, that they yield more robust treatment effect estimates when the treatment and control group differ in size. Using random forests and Bayesian additive regression trees (BART) (Chipman et al., 2010) as base models, Künzel et al. (2017) show how X-learners can be adaptive to many different DGPs in terms of local trends and sparsity. Similar to random forests, BART is a popular approach for the estimation of heterogeneous treatment effects, as it can handle datasets comprising many covariates and produce consistent confidence intervals (Dorie et al., 2019; Green and Kern, 2012; Hill, 2011; Hill and Su, 2013). Furthermore, Hill (2011) has shown BART's success in identifying heterogeneous treatment effects by winning the 2016 Atlantic Causal Inference Conference Data Challenge.

In consequence, the question arises of how the performances of causal forests, local linear forests, and X-learners based on random forests and BART, match up to each other in different settings. Hence, in this paper, we aim to further compare these methods. In addition, we use k-nearest neighbors as a baseline, since trees and forests essentially are adaptive nearest neighbor methods (Wager and Athey, 2018). Accordingly, we present the following research question:

RQ: *What is the difference in performance between causal forests and other tree-based methods when estimating heterogeneous treatment effects?*

We use simulated experiments to compare the performance of the methods in terms of mean squared error (MSE) and coverage. We extend the results of Friedberg et al. (2020), by showing that local linear forests achieve performance better than or equal to causal forests in various settings. However, for small datasets, we find that causal forests obtain better coverage than local linear forests. Moreover, we find that X-learners based on random forests can perform as well as local linear forests for data comprising many covariates. Additionally, we find that X-learners based on random forests obtain superior performance when the treatment group is much smaller than the control group, in line with the results of Künzel et al. (2017). Moreover, we find that causal forests consistently achieve good coverage for datasets consisting of a small number of covariates but decrease more quickly in coverage rate than the other methods when the underlying dataset gets larger. Lastly, despite the results of Hill (2011), we find X-learners based on BART to be outperformed by the other methods for all our experiments.

This paper is arranged in the following manner. In Section 2, we describe all methods. Subsequently, we show the simulation set-up used to evaluate the performance of the methods in Section 3. Thereafter, we present and discuss the obtained results in Section 4. Lastly, in Section 5, we yield our conclusion, accompanied by some suggestions for future research.

2 Methodology

In this section, we first describe the framework used to estimate heterogeneous treatment effects. Thereafter, the models: k-nearest neighbors, causal forests, local linear forests, X-learners based on random forests, and X-learners based on BART are presented, respectfully.

2.1 Potential outcomes framework

As specified by Neyman (1923) and Rubin (1974), we follow the potential outcomes framework. We use a training dataset containing n i.i.d. samples, where sample $i \leq n$ consists of a structure (Y_i, X_i, W_i). Hence, each sample contains a response $Y_i \in \mathbb{R}$, a covariate vector $X_i \in [0,1]^d$ comprising d covariates, and a treatment indicator $W_i \in \{0,1\}$ which equals 1 if the sample is part of the treatment group and 0 otherwise. Next, we postulate that $Y_i^{(1)}$ and $Y_i^{(0)}$ are the potential responses of sample i, with and without treatment respectively. Then, as specified by Wager and Athey (2018), we aim to approximate the treatment effect at point x for sample i, which is defined as

$$\tau(x) = \mathbb{E}[Y_i^{(1)} - Y_i^{(0)} | X_i = x]. \tag{1}$$

Nonetheless, we only observe one of the two hypothetical outcomes, as $Y_i = W_i Y_i^{(1)} + (1 - W_i) Y_i^{(0)}$. Hence, to estimate the treatment effect, we need to imply additional restrictions on the DGP. Consequently, we assume unconfoundedness as specified by Rosenbaum and Rubin (1983):

$$\{Y_i^{(1)}, Y_i^{(0)}\} \perp\!\!\!\perp W_i | X_i, \tag{2}$$

which means that the potential response of sample i is independent of the treatment assignment, conditional on the covariates. As a result, assuming unconfoundedness insinuates that closeby observations in the covariate-space can be treated as being drawn from a stochastic DGP. Hence, methods based on nearest neighbor matching will generally give consistent estimates of the treatment effect (Wager and Athey, 2018).

2.2 *k*-nearest neighbors (*k*-NN)

As the other methods used in this paper consist of adaptive nearest neighbor estimators, we set the nonadaptive k-NN to be our baseline model. k-NN is built on the assumption that similar observations exist in a close proximity of each other. In order to estimate the relationship between the covariates and the dependent variable, the k nearest points in the data are averaged. In this paper, we determine the proximity between data points based on the Euclidean distance, which is the most widely applied metric in this setting. Consequently, the treatment effect is estimated as

$$\widehat{\tau}(x) = \frac{1}{k} \sum_{i \in S_1(x)} Y_i - \frac{1}{k} \sum_{i \in S_0(x)} Y_i, \tag{3}$$

where S_1 (S_0) contains the k nearest neighbors to data point x in the treatment (control) group. Subsequently, following Wager and Athey (2018), we produce confidence intervals by constructing the estimated treatment effect $\widehat{\tau}(x)$ as normally distributed with mean $\tau(x)$ and variance $\dfrac{\widehat{V}(S_0) + \widehat{V}(S_1)}{k(k-1)}$, where $\widehat{V}(S_{0/1})$ is the sample variance for $S_{0/1}$.

2.3 Causal forests (CF)

A causal forest is a specific type of random forest, which is based on regression trees. After the fashion of k-NN, regression trees essentially look for observations that behave similarly and partition them together. That is, regression trees use the covariates to explain the variation of the dependent variable by persistently partitioning the data into more comparable categories (De'ath and Fabricius, 2000). Starting with the full data set, a regression tree picks a covariate to be used for partitioning the data. Thereafter, different covariates are chosen step-by-step to sequentially "slice up" the covariate-space. This partitioning is continued until a prespecified threshold is reached. Finally, the average value of the dependent variable within each terminal partition, called a leaf, is used to approximate the underlying function of the dependent variable.

At every new partition, the covariate which can most properly discriminate among the potential outcomes is chosen to partition the data. This ensures that the data is optimally partitioned into similar categories. Moreover, this can be achieved by using the classification and regression tree (CART) algorithm of Breiman et al. (1984), which minimizes the impurity at the beginning of each partition. The impurity is equal to the prediction loss, which is a measure that takes the value zero for entirely equivalent data points, and increases as data points become less similar (De'ath and Fabricius, 2000).

As an example, Figure 1 displays a tree which uses the covariates *size* and *value* to partition the observations. As seen in the left diagram, the tree first uses *size* to split the data, and then uses *value* for further partitioning. The according thresholds ($size = 0.5$ and $value = 0.3$) are chosen such that the impurity is minimized. As seen in the right diagram, this process coincides with rectangularly "slicing up" the covariate-space. Lastly, the predictions for each leaf are made based on the observations it contains.

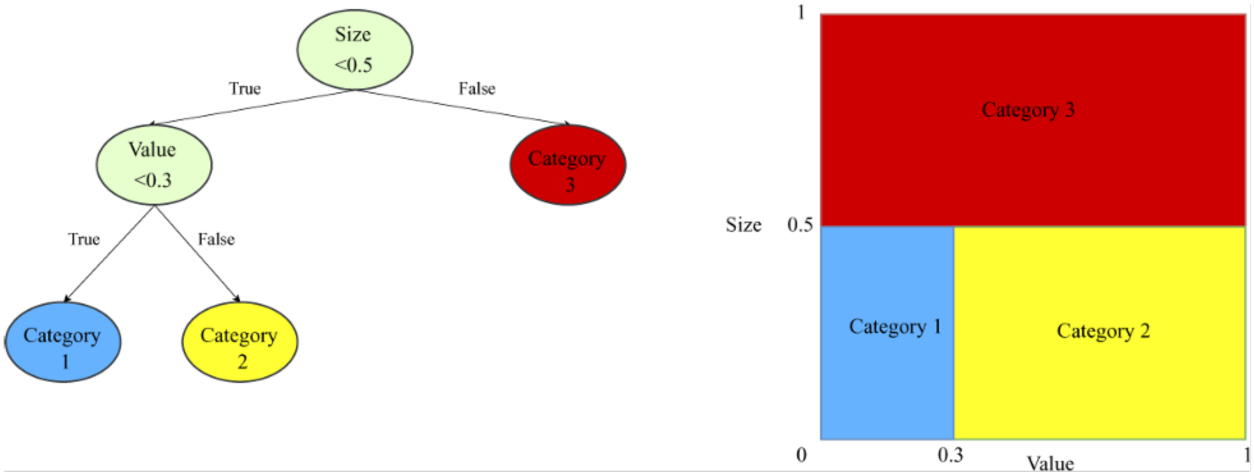


Figure 1: A regression tree (Gu et al., 2020)

This figure displays two representations of the same regression tree. Two covariates (*size* and *value*) are used to partition the observations into three different categories.

When using causal trees, we aim for the leaves to be little enough that its pairs (Y_i, W_i) for sample i in leaf $L(x)$ can be viewed as being drawn from a stochastic DGP. Then, the treatment effect can be estimated as

$$\widehat{\tau}(x) = \frac{1}{|\{i : W_i = 1, X_i \in L\}|} \sum_{\{i:W_i=1,X_i \in L\}} Y_i - \frac{1}{|\{i : W_i = 0, X_i \in L\}|} \sum_{\{i:W_i=0,X_i \in L\}} Y_i, \quad \forall \quad x \in L. \quad (4)$$

Since trees are prone to overfitting, they must be heavily regularized. Therefore, we use a causal forest, which is an ensemble method that combines forecasts $\widehat{\tau}_b(x)$ from B different trees into a single forecast $\widehat{\tau}(x) = \frac{1}{B}\sum_{b=1}^{B} \widehat{\tau}_b(x)$ and hence reduces the variance of the estimated prediction function. Here, the causal forest combines the trees in such a way that each tree depends on the values of a random vector which is assumed to be sampled i.i.d. for all trees in the causal forest (Breiman, 2001). The random vector contains s samples where $\frac{s}{n} \ll 1$.

In addition, causal trees are subject to a powerful restriction called honesty. Wager and Athey (2018) define a tree grown on a training sample $(Z_1 = (X_1, Y_1), \dots, Z_s = (X_s, Y_s))$ to be honest if (**1**) the splitting rule of the tree disregards responses $Y_1, \dots, Y_s$ (propensity tree); or (**2**) the tree divides the training sample into two halves of which only one is used when placing the splits (double sample tree). In other words, a tree satisfies the honesty criterion when it contains no data point that is used both to determine the splitting rule and to estimate the treatment effect. Propensity trees and double sample trees are further described in Appendices A and B, respectively.

Subsequently, we can use the infinitesimal jackknife (Wager et al., 2014) to obtain precise estimates of the asymptotic variance. This method is based on repeatedly resampling the sample data and averaging the estimates into one statistic (bootstrapping). It reduces the bootstrap bias by using a leave-one-out approach. That is, it recomputes the estimate by weighing one observation slightly less than the others. After this is done for all observations, all estimates are combined. Furthermore, to obtain pointwise consistent results, we assume that the expected response with and without treatment, specified as

$$\mu_1(x) = \mathbb{E}[Y_i^{(1)}|X_i = x] \quad \text{and} \quad \mu_0(x) = \mathbb{E}[Y_i^{(0)}|X_i = x], \quad \text{respectively,} \tag{5}$$

are Lipschitz continuitious, meaning that they have bounded derivatives. Lipschitz continuity is formally defined in Appendix C. Additionally, we assume there to be overlap, such that

$$\epsilon < \mathbb{P}[W = 1|X = x] < 1 - \epsilon, \quad \text{for some } \epsilon > 0, \tag{6}$$

which ensures that all test samples contain treatment and control indicators in close enough proximity for a sufficiently large number of samples. Then, consistent variance estimates are obtained via

$$\widehat{\text{Var}}(x) = \frac{n(n-1)}{(n-s)^2} \sum_{i=1}^{n} \text{Cov}_*[\widehat{\tau}_b^*(x), I_{ib}^*]^2, \tag{7}$$

where $\widehat{\tau}_b^*$ denotes the approximated treatment effect obtained by tree b and I_{ib}^* denotes an indicator function which is equal to 1 if tree b uses sample i and 0 otherwise. Lastly, the first term is added to remove the bias of estimates based on finite samples.

2.4 Local linear forests (LLF)

As specified by Friedberg et al. (2020), local linear forests generate weights using a random forest so that they can be implemented as a kernel for local linear regression. Local linear forests estimate a model for the response Y by assuming additive errors, generally specified as:

$$Y_i = \mu(X_i) + \varepsilon_i, \tag{8}$$

where $\mu(x_0)$ denotes the conditional mean $\mathbb{E}[Y_i|X_i = x_0]$ at specified point x_0, and ε_i denotes the error term. Using a random forest to achieve the conditional mean, we obtain $\widehat{\mu}(x_0) = \sum_{i=1}^{n} \alpha_i(x_0)Y_i$, where

$$\alpha_i(x_0) = \frac{1}{B}\sum_{b=1}^{B} \frac{I\{X_i \in L_b(x_0)\}}{|L_b(x_0)|}, \tag{9}$$

which denotes the weights obtained by the random forest. Here, L_b is the leaf corresponding to tree b and $I\{X_i \in L_b(x_0)\}$ is an indicator function. Furthermore, $0 \leq \alpha_i(x_0) \leq 1$ and

$$\sum_{i=1}^{n} \alpha_i(x_0) = \begin{cases} 1, & \text{if one or more trees has/have a leaf which contains } x_0; \\ 0, & \text{otherwise.} \end{cases} \tag{10}$$

Using the weights defined in Equation 9, local linear forests apply weighted least squares. In the setting of causal inference, local linear forests first estimate the treatment propensity $e(x_0)$ and main effect $m(x_0)$, which are specified as

$$e(x_0) = \mathbb{P}[W = 1|X = x_0] \quad \text{and} \quad m(x_0)) = \tfrac{1}{2}\mathbb{E}[Y^{(1)} + Y^{(0)}|X = x_0]. \tag{11}$$

Then, we use a ridge regression, which uses L2 penalties to handle highly correlated covariates (Hoerl and Kennard, 1970). It approximates the treatment effect by following

$$\begin{aligned} \Big\{\widehat{\tau}(x_0), \widehat{\theta}_\tau(x_0), \widehat{\alpha}(x_0), \widehat{\theta}_\alpha(x_0)\Big\} = \operatorname{argmin}_{\tau,\theta}\Bigg\{ \sum_{i=1}^{n} \alpha_i(x_0)\Bigg(& Y_i - \widehat{m}^{(-i)}(X_i) - a - (X_i - x_0)\theta_a \\ & - (\tau + \theta_\tau(X_i - x_0))(W_i - \widehat{e}^{(-i)}(X_i))\Bigg)^2 + \lambda_\tau\|\theta_\tau\|_2^2 + \lambda_a\|\theta_a\|_2^2\Bigg\}, \end{aligned} \tag{12}$$

where a denotes the intercept, of which the estimate should be equal to zero if the treatment propensity and the main effect are estimated correctly. Nevertheless, the intercept is still incorporated in an attempt to make the approximation more robust. Moreover, the parameters $\theta_\tau(x_0)$ and $\theta_\alpha(x_0)$ are implemented to cope with the local trend within $X_i - x_0$. To prevent the model from overfitting to this local trend, the ridge parameters λ_τ and λ_a are applied as L2 penalties. That is, they pull all estimates regarding $\theta_\tau(x_0)$ and $\theta_\alpha(x_0)$ closer to zero but withstand from forcing exact zeros anywhere, preventing the coefficients from becoming overly big in magnitude (Gu et al., 2020).

Additionally, the ridge parameters can be tuned using k-fold cross-validation, which is a resampling technique, specified in Algorithm 1. It first splits the data into k samples and picks one. Based on a specific parameter value, it then fits the model on the data contained in the remaining samples. Thereafter, it evaluates the performance of the fitted model on the held out sample. It repeats this process for the other samples and then combines the evaluation scores into one metric. This procedure is then repeated for all parameter values of interest, after which the parameter value is chosen which obtains the best performance.

Algorithm 1: k-fold cross-validation

```
Specify which parameter values to evaluate
Split the randomized dataset into $k$ partitions
Initialize minimum loss $\mathcal{L}_0 = \infty$ and a vector of parameter values $\theta$
for each parameter combination do
    for each partition $p \leq k$ do
        Hold out partition $p$
        Fit the model on the remaining partitions
        Run the model on partition $p$ and calculate the loss $\mathcal{L}_i$[a]
    end
    Calculate the overall loss $\mathcal{L} = \frac{1}{k}\sum_{p=1}^{k} \mathcal{L}_p$
    if $\mathcal{L} \leq \mathcal{L}_0$ then
        $\mathcal{L}_0 \leftarrow \mathcal{L}$
        $\theta \leftarrow$ current parameter combination
    end
end
Result: The optimal set parameter of parameter values $\theta$
```

[a] For the purpose of tuning ridge parameters, we minimize the loss function recommended by Nie and Wager (2020):

$$\widehat{\mathcal{L}}(\widehat{\tau}(\cdot) = \sum_{i=1}^{n} (Y_i - \widehat{m}^{(-i)})(W_i - \widehat{e}^{(-i)}(X_i)). \tag{13}$$

as this eliminates spurious effects in the treatment effect estimates by controlling for correlations between the treatment propensity and main effect.

Subsequently, we use the random forest delta method as specified by Athey et al. (2019) to obtain confidence intervals for the estimates. This method approximates the asymptotic probability distribution for a function of an asymptotically normally distributed random variable based on its finite sample variance. First, we obtain a local solution:

$$\sum_{i=1}^{n} \alpha_i(x_0)\phi\Big(X_i, Y_i, \widehat{\mu}(x_0), \widehat{\theta}(x_0)\Big) = 0, \tag{14}$$

where the gradient of the log-likelihood (score) function

$$\phi(X_i, Y_i, \mu, \theta) = \nabla_{\mu,\theta}\frac{1}{2}\left(\left(Y_i - \Delta_i \begin{pmatrix}\mu\\ \theta\end{pmatrix}\right)^2 + \lambda\|\theta\|_2^2\right) \tag{15}$$

is used. Here, the centered matrix Δ_i is structured such that $\Delta_{i,1} = 1$ and $\Delta_{i,j+1} = x_{i,j} - x_{0,j}$. Then, the variance can be approximated as

$$\widehat{\text{Var}}\Big[\Big(\widehat{\mu}(x_0), \widehat{\theta}(x_0)\Big)\Big] = \widehat{V}(x_0)^{-1}\widehat{H}_n(x_0)\Big(\widehat{V}(x_0)^{-1}\Big)', \tag{16}$$

where

$$V(x_0) = \nabla_{\mu,\theta}\mathbb{E}[\phi(X_i, Y_i, \mu, \theta)|x_0 = x_0] \tag{17}$$

and

$$H_n(x_0) = \text{Var}\left[\sum_{i=1}^{n} \alpha_i(x_0)\phi\Big(X_i, Y_i, \mu^*(x_0)\theta^*(x_0)\Big)\right], \tag{18}$$

denote the gradient and the variance of the expected value of the score function at its minimum, respectively.

2.5 X-learners based on random forests (XRF)

As specified by Künzel et al. (2017), X-learners based on random forests use a three-stage algorithm. In stage **1**, honest random forests, as described in Section 2.3, are used to approximate the conditional expected responses μ_0 and μ_1.

Next, in stage **2**, the individual treatment effect $D_i = Y_i^{(1)} - Y_i^{(0)}$ is estimated for each sample $i \leq n$. The approximated responses in stage **1** are used to estimate the individual treatment effects as

$$\widetilde{D}_i^{(0)} = \widehat{\mu}_0(X_i^{(0)}) - Y_i^{(1)} \quad \text{and} \quad \widetilde{D}_i^{(1)} = Y_i^{(0)} - \widehat{\mu}_0(X_i^{(1)}), \tag{19}$$

for samples with and without treatment, respectively. If the estimated responses are accurate, such that $\widehat{\mu}_1 = \mu_1$ and $\widehat{\mu}_0 = \mu_0$, we obtain that $\tau(x) = \mathbb{E}[\widetilde{D}^{(1)}|X_i = x] = \mathbb{E}[\widetilde{D}^{(0)}|X_i = x]$. However, as this rarely is achievable in practice, both the estimated individual effects $\widetilde{D}^{(1)}$ and $\widetilde{D}^{(0)}$ are used to estimate the treatment effect.

Consequently, in stage **3**, $\widetilde{D}^{(1)}$ and $\widetilde{D}^{(0)}$ are used to obtain the approximated treatment effect for treated samples ($\widehat{\tau}_1$) and untreated samples ($\widehat{\tau}_1$). Finally, the treatment effect is computed as a weighted average of $\widehat{\tau}_1$ and $\widehat{\tau}_0$:

$$\widehat{\tau}(x) = g(x)\widehat{\tau}_0(x) + (1 - g(x))\widehat{\tau}_1(x), \tag{20}$$

where the function $g \in [0, 1]$ represents a random forest that is used to obtain the weights. In short, the advantage of X-learners is that they can use information from treated (untreated) samples to obtain better estimators for the untreated (treated) samples.

Additionally, Künzel et al. (2017) show that the bootstrapping method in Algorithm 2 can be used to obtain confidence intervals for X-learners. First, it resamples the sample data and fits the model a large number of times. Then, it combines the estimates of the treatment effects and calculates the mean and the standard deviation. Lastly, it determines the confidence interval based on the chosen quantiles.

Algorithm 2: Bootstrapped confidence interval

```
for $b \in \{1, \dots, B\}$ do
    $s$ = sample(1 : $n$, replace = $T$, size = $\frac{n}{2}$)
    $x_b^* = x_s$
    $w_b^* = w_s$
    $y_b^* = y_s$
    $\widehat{\tau}_b^*(p)$ = random.forest($x_b^*, w_b^*, y_b^*$)[$p$]
end
$\tau(p)$ = random.forest($x_b, w_b, y_b$)[$p$]
$\sigma$ = st.dev.($\{\widehat{\tau}_b^*(p)\}_{b=1}^{B}$)
Result: $[\widehat{\tau}(p) - q_{\alpha/2}\sigma, \widehat{\tau}(p) + q_{1-\alpha/2}\sigma]$
```

2.6 X-learners based on Bayesian additive regression trees (XBART)

X-learners based on Bayesian additive regression trees (BART) use the same three-stage approach as shown in Section 2.5, but simply replace the usage of random forests with BART. This method is based on regression trees as well but uses different regularization than forests. Similar to local linear forests, BART assumes additive errors. As specified by Hill (2011), the response Y_i is estimated as

$$Y_i = \sum_{k=1}^{m} g(W_i, X_i; T_k, M_k) + \varphi_k, \tag{21}$$

where $g(\cdot)$ denotes a regression tree, m is the number of trees and the errors terms φ_i are i.i.d. $\mathcal{N}(0, \sigma^2)$. Moreover, every leaf j of a tree T is accompanied by a parameter μ_j which denotes the mean of the responses of all samples within that leaf. In addition, the set M contains the mean of all leaves $\{\mu_1, \mu_2, \ldots, \mu_K\}$, where K is the number of leaves. Then, for a tree model (T, M) and a pair (W_i, X_i), we define $g(W_i, X_i; T, M)$ as the mean which is acquired after partitioning the covariate-space based on (W_i, X_i).

The intuition behind this sum-of-trees model is similar to that of gradient boosting. Namely, it starts by fitting a single tree $g(W_i, X_i; T_1, M_1)$ and then deducts it from the observed response $Y_i = y$ to form residuals. Then, a new tree is fitted to these residuals as an attempt to explain more of the response. Subsequently, this process is repeated for a total of m times.

As this procedure of repeatedly fitting a tree on the remaining residuals could lead to overfitting, each new fit is multiplied by a certain shrinkage rate before being deducted, which ensures more robustness. Additionally, a regularization prior is used to oppose overfitting. By specifying parameters such as the total number of trees, the tree size and the shrinkage rate in advance, a regularization prior forces each tree to be responsible for only a modest part of the final forecast (Carvalho et al., 2019).

3 Simulation study

We follow a similar experimental setup as Wager and Athey (2018), such that we describe our experiments with regard to the sample size n, the number of covariates d, the main effect $m(x)$ and treatment propensity $e(x)$ which are specified in Equation 10, and the treatment effect $\tau(x)$ which is specified in Equation 1. Regarding all our experiments, we assume unconfoundedness as stated in Equation 2, use uniformly distributed covariates $X \sim \mathcal{U}([0, 1]^d)$, and obtain normally distributed homoscedastic responses $Y_{(0/1)} \sim \mathcal{N}(\mathbb{E}[Y_{(0/1)}|X], 1)$. Moreover, we evaluate the performance of the methods based on their expected MSE for estimating the treatment effect, and their corresponding expected coverage, which is set to have a target of 95%.

All models are executed in the statistical software environment **R**. We use the `FNN` (Beygelzimer et al., 2013) package to perform our k-NN analysis. Moreover, we use the `causalTree` (Athey and Imbens, 2015) and `randomForestCI` (Wager et al., 2014) packages to obtain causal forest estimates and confidence intervals. Additionally, we use the `grf` (Athey et al., 2019) and `glmnet` (Friedman et al., 2010) packages to implement local linear forests. Lastly, we use the `causalToolbox` (Künzel et al., 2017) package to execute the X-learners.

We perform four separate experiments. In Experiment 1, we check whether the models can provide unbiased forecasts even though there is an interaction between the main effect $m(x)$ and the treatment propensity $e(x)$. That is, we set

$$e(X) = \frac{1}{4}(1 + \beta_{2,4}(X_1)), \quad m(X) = 2(X_1) - 1, \tag{22}$$

where $\beta_{a,b}$ denotes the β distribution parameterized by shape parameters a and b. Besides, we set the treatment effect equal to zero. Moreover, we use $n = 500$ samples and range the number of covariates as $d \in \{2, 5, 10, 15, 20, 30\}$. As the aim of this experiment is to accurately embed the propensity into the forecasts, we use propensity trees for the causal forests. In addition, we use 1,000 trees with a sample size of 50.

Next, in Experiment 2, we check to what extent the models succeed in adapting to heterogeneity in the treatment effect while keeping the other functions fixed: $m(x) = 0$ and $e(x) = 0.5$. We specify the treatment effect as a function of the first and second covariate:

$$\tau(x) = \zeta(X_1)\zeta(X_2), \zeta(x) = 1 + \frac{1}{1 + e^{-20(x-\frac{1}{3})}}. \tag{23}$$

We use $n = 5,000$ samples and range the number of covariates as $d \in \{2, 3, 4, 5, 6, 8\}$. Here, we use double-sample trees for the causal forests. In addition, we use 2,000 trees with a sample size of 2,500. Subsequently, for Experiment 3, we repeat the second experiment, but decrease the treatment propensity to $e(x) = 0.05$, as the treatment group can be much smaller than the control group in practice.

Lastly, the set-up for Experiment 4 is in line with the second experiment as well but replaces the treatment effect with a function that has a sharper spike:

$$\tau(x) = \zeta(X_1)\zeta(X_2), \zeta(x) = 1 + \frac{1}{1 + e^{-12(x-\frac{1}{2})}}. \tag{24}$$

This allows us to determine how the models cope with a more volatile treatment effect. Furthermore, we use $n = 10,000$ samples, 10,000 trees and a sample size of 2,000.

4 Results

In this section, we present the results for each method per experiment. We report both the MSE and the coverage per model. Overall, we find that local linear forests (LLF) always attain an MSE which is either equal or lower than that of causal forests (CF). Moreover, in the setting of Experiment 3, where the propensity score is rather low ($e(x) = 0.05$), we find that XRF attains the lowest MSE. Additionally, we find the coverage of causal forests to decrease quicker than the other methods when the number of covariates increases.

Table 1 displays the performance of all models in terms of their MSE (top) and coverage (bottom) in the setting of Experiment 1. These metrics are given for datasets containing different numbers of covariates d. Based on MSE, we observe that causal forests are only slightly outperformed when the data comprises two covariates, as local linear forests attain an MSE of 0.01 and causal forests attain an MSE of 0.02. When the data comprises more than two covariates, local linear forests and causal forests attain the same MSE. For the bigger datasets ($d \geq 15$), XRF attains the lowest MSE as well. In terms of coverage, we observe that causal forests depend relatively more on the size of the dataset, as its coverage decreases from 95% ($d = 2$) to 85% ($d = 30$), whereas the coverage for local linear forests only changes from 92% to 96%. Moreover, both of the X-learners showcase especially high coverage. Furthermore, XBART always attains full coverage, consistently exceeding the target of 95%.

d	**CF**	**LLF**	**XRF**	**XBART**	**10-NN**	**100-NN**
2	0.02	**0.01**	0.05	0.12	0.21	0.09
	0.95	0.92	1.00	1.00	0.93	0.59
5	**0.02**	**0.02**	0.03	0.10	0.24	0.12
	0.95	0.93	1.00	1.00	0.92	0.53
10	**0.02**	**0.02**	0.03	0.34	0.29	0.13
	0.92	0.93	1.00	1.00	0.92	0.53
15	**0.02**	**0.02**	**0.02**	0.19	0.31	0.13
	0.9	0.96	0.95	1.00	0.90	0.47
20	**0.02**	**0.02**	**0.02**	0.22	0.32	0.13
	0.89	0.97	1.00	1.00	0.89	0.48
30	**0.02**	**0.02**	**0.02**	0.19	0.34	0.13
	0.85	0.96	1.00	1.00	0.89	0.48

Table 1: Model performance for Experiment 1

This table reports the MSE (top) and coverage (bottom) for Experiment 1 per the number of covariates d. The results are accumulated over 500 simulations for causal forests (CF), local local forests (LLF), X-learners based on random forests and BART (XRF and XBART, respectively), and k-nearest neighbors based on 10 and 100 neighbors (10-NN and 100-NN, respectively). **Bold** font indicates the lowest attained MSE per covariate level.

Table 2 displays the performance of all models in the setting of Experiment 2, which focuses on capturing the heterogeneity within the treatment effect. Based on MSE, we observe that local linear forests, now more evidently, attain the best performance, always obtaining a value of 0.02. Causal forests only obtain such a low MSE when the data comprises 5 or 6 covariates. XRF does not attain this value but comes close when the dataset gets larger. Nonetheless, in terms of coverage, causal forests do outperform local linear forests, obtaining values between 97% ($d = 2$) and 89% ($d = 8$), whereas local linear forests achieve coverage rates around 86%. Additionally, XRF and 7-NN attain confidence rates around 93% and 92% respectively.

d	**CF**	**LLF**	**XRF**	**XBART**	**7-NN**	**50-NN**
2	0.04	**0.02**	0.05	0.04	0.29	0.04
	0.97	0.85	0.94	1.00	0.93	0.95
3	0.03	**0.02**	0.04	0.06	0.29	0.05
	0.96	0.83	0.95	1.00	0.93	0.91
4	0.03	**0.02**	0.04	0.06	0.3	0.08
	0.95	0.85	0.92	1.00	0.92	0.85
5	**0.02**	**0.02**	0.03	0.07	0.31	0.11
	0.94	0.86	0.93	1.00	0.92	0.76
6	**0.02**	**0.02**	0.03	0.07	0.33	0.15
	0.92	0.88	0.89	1.00	0.91	0.69
8	0.03	**0.02**	0.03	0.07	0.39	0.22
	0.89	0.86	0.92	1.00	0.90	0.56

Table 2: Model performance for Experiment 2

This table reports the MSE (top) and coverage (bottom) for Experiment 2 per the number of covariates d. The results are displayed for each method, accumulated over 40 simulations. **Bold** font indicates the lowest attained MSE per covariate level.

Table 3 shows the performance of all models with regards to Experiment 3, which focuses on capturing the heterogeneity within the treatment effect when the treatment group consists of only 5% of the data. We observe that XRF always obtains the lowest MSE. Causal forests only match XRF when the data contains 2 covariates. Furthermore, causal forests display similar performance as 50-NN by yielding drastically lower MSEs as the dataset becomes larger. Local linear forests however yield similar MSEs as XRF, obtaining the same value for datasets containing 2 or 4 covariates. In terms of coverage, causal forests produce drastically decreasing rates as well, changing from 92% ($d = 2$) to 57% ($d = 8$). XRF on the other hand shows increasing coverage rates as the dataset becomes larger, changing from 89% ($d = 2$) to 94% ($d = 8$).

Table 4 shows the performance of all models with regards to Experiment 4, which focuses on capturing the heterogeneity of a more volatile treatment effect. We observe a similar outcome as in the first experiment. Namely, we find that causal forests and local linear forests attain equal MSEs, but when the data comprises 2 covariates, local linear forests (MSE = 0.01) slightly outperform

d	CF	LLF	XRF	XBART	7-NN	50-NN
2	**0.06**	**0.06**	**0.06**	0.08	0.29	0.10
	0.92	0.80	0.89	1.00	0.93	0.81
3	0.10	0.07	**0.06**	0.08	0.32	0.08
	0.82	0.75	0.93	1.00	0.92	0.84
4	0.13	**0.07**	**0.07**	0.09	0.34	0.13
	0.73	0.79	0.93	1.00	0.91	0.75
5	0.13	0.08	**0.06**	0.11	0.34	0.18
	0.71	0.70	0.94	1.00	0.92	0.61
6	0.18	0.08	**0.07**	0.12	0.40	0.23
	0.59	0.74	0.85	1.00	0.88	0.54
8	0.20	0.08	**0.06**	0.12	0.43	0.28
	0.57	0.72	0.94	1.00	0.88	0.45

Table 3: Model performance for Experiment 3

This table reports the MSE (top) and coverage (bottom) for Experiment 3 per the number of covariates d. The results are displayed for each method, accumulated over 40 simulations. **Bold** font indicates the lowest attained MSE per covariate level.

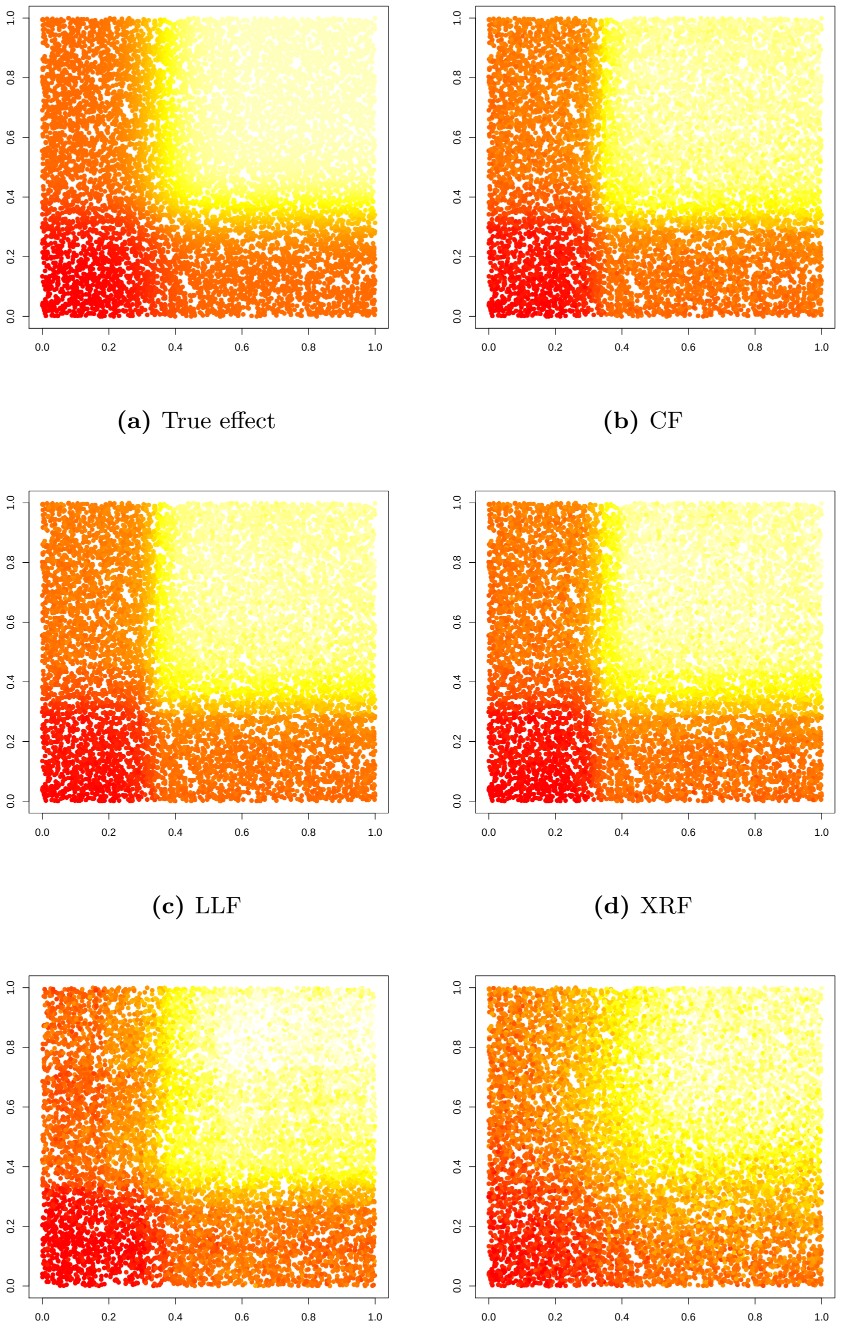
causal forests (MSE $= 0.02$). Furthermore, XRF attains the same MSE for larger datasets ($d \geq 15$). In terms of coverage, causal forests obtain a rate that decreases from 95% ($d = 2$) to 85% ($d = 30$), whereas local linear forests and XRF obtain rates around 87% and 93% respectively.

d	CF	LLF	XRF	XBART	10-NN	100-NN
2	0.02	**0.01**	0.04	0.03	0.2	0.02
	0.93	0.85	0.94	1.00	0.93	0.94
3	**0.02**	**0.02**	0.03	0.04	0.2	0.03
	0.89	0.86	0.95	1.00	0.93	0.9
4	**0.02**	**0.02**	0.03	0.05	0.21	0.05
	0.85	0.88	0.87	1.00	0.93	0.79
5	**0.02**	**0.02**	**0.02**	0.05	0.22	0.09
	0.82	0.89	0.94	1.00	0.93	0.67
6	**0.02**	**0.02**	**0.02**	0.06	0.24	0.15
	0.75	0.86	0.94	1.00	0.92	0.57
8	**0.02**	**0.02**	**0.02**	0.07	0.29	0.27
	0.72	0.88	0.91	1.00	0.9	0.44

Table 4: Model performance for Experiment 4

This table reports the MSE (top) and coverage (bottom) for Experiment 4 per the number of covariates d. The results are displayed for each method, accumulated over 40 simulations. **Bold** font indicates the lowest attained MSE per covariate level.

Following the treatment effect from Equation 23 which is used in Experiment 2, Figures 2 and 3 visualize to what extent each method captures heterogeneity. We observe that as data comprises more covariates, causal forests less adequately adapt to the high treatment effect displayed in the top right corner. On the other hand, we see that local linear forests and XRF are more consistent in capturing the heterogeneity. Moreover, XBART fails to capture the shape of the treatment effect when the data consists of more covariates and 50-NN totally breaks down, as it no longer represents any shape of the actual treatment effect.



(a) True effect

(b) CF

(c) LLF

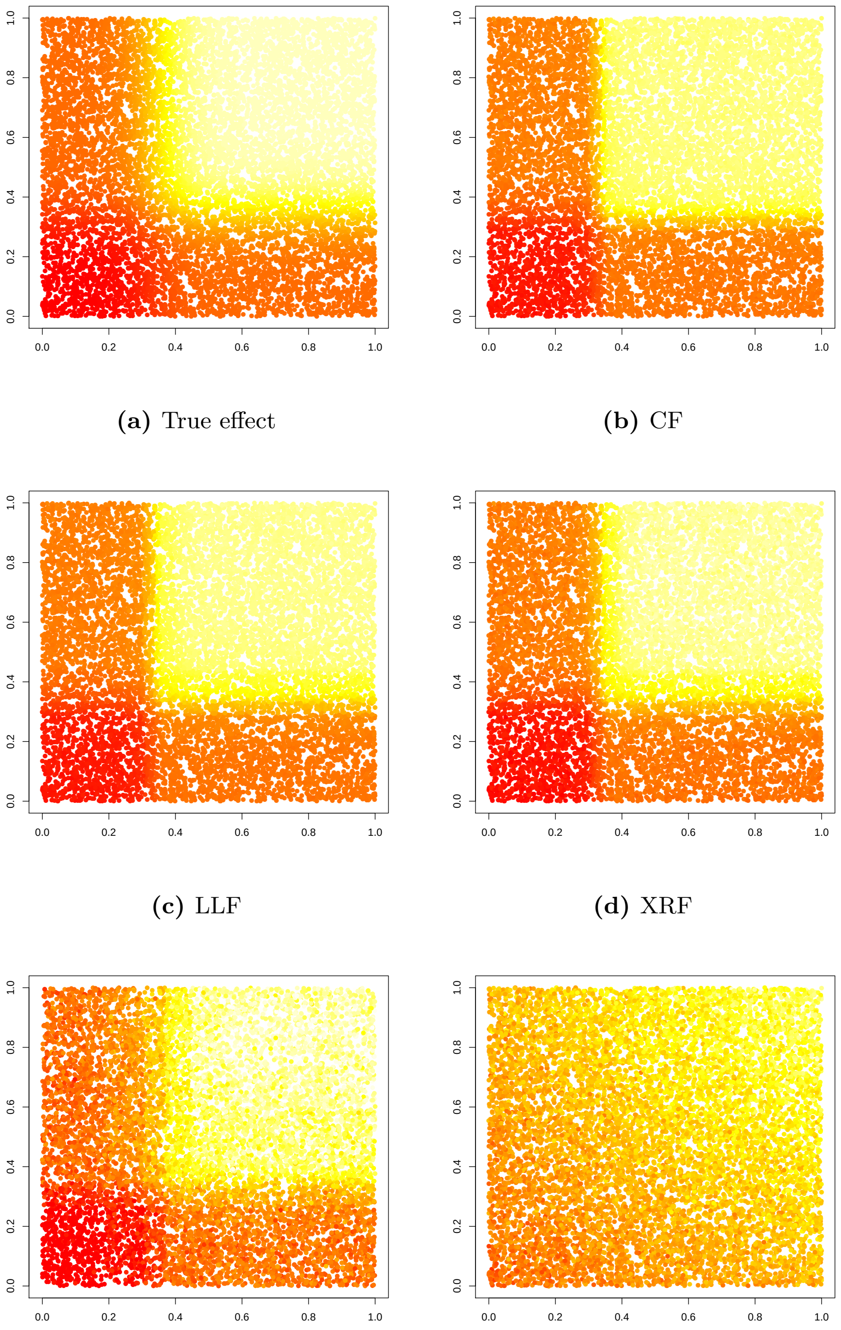
(d) XRF

(e) XBART

(f) 50-NN

Figure 2: Treatment effect comparison (*d*=5)

This figure reports the true treatment effect following Equation 23 and its estimates per method for a dataset comprising 5 covariates. The training and test set both consist of 1,000 samples. The first two coordinates are used to plot the test points, where high and low treatment effects are displayed by light and dark color respectively.



(a) True effect

(b) CF

(c) LLF

(d) XRF

(e) XBART

(f) 50-NN

Figure 3: Treatment effect comparison (*d*=30)

This figure reports the true treatment effect following Equation 23 and its estimates per method for a dataset comprising 30 covariates. The training and test set both consist of 1,000 samples. The first two coordinates are used to plot the test points, where high and low treatment effects are displayed by light and dark color respectively.

5 Conclusion

We have introduced causal forests and compared them to other tree-based methods, being local linear forests, and X-learners based on random forests and BART, in addition to k-nearest neighbors which functioned as a baseline. We focused on four different simulated experiments, to answer the research question: *What is the difference in performance between causal forests and other tree-based methods when estimating heterogeneous treatment effects?*

We have found that local linear forests, in terms of MSE, achieve performance better than or equal to causal forests for all simulated experiments. Moreover, when there is either an interaction between the underlying main effect and treatment propensity, or a sharp spike within the treatment effect, XRF performs as well as local linear forests for data comprising many covariates. When the control and treatment groups are of equal size, XRF does less well in capturing the heterogeneity within the treatment effect than causal forests and local linear forests. Nonetheless, when the treatment group is much smaller than the control group, XRF succeeds the most in capturing the heterogeneity. This performance is closely followed by local linear forests, whereas causal forests are clearly outperformed, especially in the case of larger datasets. In addition, XBART and $k-$NN are outperformed in all scenarios.

In terms of coverage, we have found that causal forests perform well for datasets consisting of a small number of covariates, as they consistently outperform local linear forests. Nevertheless, when datasets get bigger, causal forests yield coverage rates that decrease relatively quicker than in the case of the other methods. On the contrary, XRF consistently yields proper coverage rates, especially when the treatment group is much smaller than the control group. Additionally, when the underlying data comprises more covariates, local linear forests attain better coverage rates than causal forests, as the former are more consistent.

In short, when the treatment and control group are of equal size, we find that local linear forests attain the best performance for large datasets. However, when the dataset is small, we find there to be a trade-off in performances as causal forests attain better coverage and local linear forests attain better MSEs. Lastly, when there is a large difference between the treatment and control group, we find XRF to be superior.

Even though we have found XBART to perform worse than the other tree-based methods, future research could be performed to determine the impact of hyperparameter tuning using k-fold cross-validation on its performance. No additional hyperparameter tuning was performed in this paper, so that the posterior could be interpreted following the standard Bayesian manner (Friedberg et al., 2020). In addition, we focused on a limited range of experiments. Hence, more experiments could be performed to further explore the differences between the models in different settings.

References

Athey, S. and Imbens, G. (2015). Recursive partitioning for heterogeneous causal effects.

Athey, S., Tibshirani, J., Wager, S., et al. (2019). Generalized random forests. *Annals of Statistics*, 47(2):1148–1178.

Beygelzimer, A., Kakadet, S., Langford, J., Arya, S., Mount, D., and Li, S. (2013). Fnn: fast nearest neighbor search algorithms and applications. *R package version*, 1(1).

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

Breiman, L., Friedman, J., Stone, C. J., and Olshen, R. A. (1984). *Classification and regression trees*. CRC press.

Carvalho, C., Feller, A., Murray, J., Woody, S., and Yeager, D. (2019). Assessing treatment effect variation in observational studies: Results from a data challenge. *arXiv preprint arXiv:1907.07592*.

Chipman, H. A., George, E. I., McCulloch, R. E., et al. (2010). Bart: Bayesian additive regression trees. *The Annals of Applied Statistics*, 4(1):266–298.

Crump, R. K., Hotz, V. J., Imbens, G. W., and Mitnik, O. A. (2008). Nonparametric tests for treatment effect heterogeneity. *The Review of Economics and Statistics*, 90(3):389–405.

De'ath, G. and Fabricius, K. E. (2000). Classification and regression trees: a powerful yet simple technique for ecological data analysis. *Ecology*, 81(11):3178–3192.

Dorie, V., Hill, J., Shalit, U., Scott, M., Cervone, D., et al. (2019). Automated versus do-it-yourself methods for causal inference: Lessons learned from a data analysis competition. *Statistical Science*, 34(1):43–68.

Friedberg, R., Tibshirani, J., Athey, S., and Wager, S. (2020). Local linear forests. *Journal of Computational and Graphical Statistics*, pages 1–15.

Friedman, J., Hastie, T., and Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of statistical software*, 33(1):1.

Green, D. P. and Kern, H. L. (2012). Modeling heterogeneous treatment effects in survey experiments with bayesian additive regression trees. *Public opinion quarterly*, 76(3):491–511.

Gu, S., Kelly, B., and Xiu, D. (2020). Empirical asset pricing via machine learning. *The Review of Financial Studies*, 33(5):2223–2273.

Hill, J. and Su, Y.-S. (2013). Assessing lack of common support in causal inference using bayesian nonparametrics: Implications for evaluating the effect of breastfeeding on children's cognitive outcomes. *The Annals of Applied Statistics*, pages 1386–1420.

Hill, J. L. (2011). Bayesian nonparametric modeling for causal inference. *Journal of Computational and Graphical Statistics*, 20(1):217–240.

Hoerl, A. E. and Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67.

Kreif, N. and DiazOrdaz, K. (2019). Machine learning in policy evaluation: new tools for causal inference. *arXiv preprint arXiv:1903.00402*.

Künzel, S., Sekhon, J., Bickel, P., and Yu, B. (2017). Meta-learners for estimating heterogeneous treatment effects using machine learning.

Kurzweil, R. (2010). Merging with the machines: Information technology, artificial intelligence, and the law of exponential growth. *World Future Review*, 2(1):61–66.

Lee, M.-j. (2009). Non-parametric tests for distributional treatment effect for randomly censored responses. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 71(1):243–264.

Mentch, L. and Hooker, G. (2016). Quantifying uncertainty in random forests via confidence intervals and hypothesis tests. *The Journal of Machine Learning Research*, 17(1):841–881.

Neto, E. C., Prentice, R. L., Bot, B. M., Kellen, M., Friend, S. H., Trister, A. D., Omberg, L., and Mangravite, L. (2017). Towards personalized causal inference of medication response in mobile health: an instrumental variable approach for randomized trials with imperfect compliance.

Neyman, J. (1923). Sur les applications de la théorie des probabilités aux experiences agricoles: Essai des principes. *Roczniki Nauk Rolniczych*, 10:1–51.

Nie, X. and Wager, S. (2020). Quasi-oracle estimation of heterogeneous treatment effects.

Norman, J., Kelly, B., Boyland, E., and McMahon, A.-T. (2016). The impact of marketing and advertising on food behaviours: evaluating the evidence for a causal relationship. *Current Nutrition Reports*, 5(3):139–149.

O'Searcoid, M. (2006). *Metric spaces*. Springer Science & Business Media.

Panizza, U. and Presbitero, A. F. (2014). Public debt and economic growth: is there a causal effect? *Journal of Macroeconomics*, 41:21–41.

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

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of educational Psychology*, 66(5):688.

Scornet, E., Biau, G., Vert, J.-P., et al. (2015). Consistency of random forests. *The Annals of Statistics*, 43(4):1716–1741.

Tye, J. B., Warner, K. E., and Glantz, S. A. (1987). Tobacco advertising and consumption: evidence of a causal relationship. *Journal of public health policy*, 8(4):492–508.

Varian, H. R. (2016). Causal inference in economics and marketing. *Proceedings of the National Academy of Sciences*, 113(27):7310–7315.

Wager, S. and Athey, S. (2018). Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, 113(523):1228–1242.

Wager, S., Hastie, T., and Efron, B. (2014). Confidence intervals for random forests: The jackknife and the infinitesimal jackknife. *The Journal of Machine Learning Research*, 15(1):1625–1651.

Willke, R. J., Zheng, Z., Subedi, P., Althin, R., and Mullins, C. D. (2012). From concepts, theory, and evidence of heterogeneity of treatment effects to methodological approaches: a primer. *BMC medical research methodology*, 12(1):1–12.

Yazdani, A. and Boerwinkle, E. (2015). Causal inference in the age of decision medicine. *Journal of data mining in genomics & proteomics*, 6(1).

Yee, A. S. (1996). The causal effects of ideas on policies. *International organization*, pages 69–108.

Appendix A

Wager and Athey (2018) specify the following procedure for obtaining a propensity tree.

1. Draw a random subsample $\mathcal{I} \in \{1, \ldots, n\}$ of size s without replacement.

2. Use sample $\mathcal{I}$ to train a classification tree where the outcome is the treatment indicator. In other words, use the (X_i, W_i) pairs with $i \in \mathcal{I}$. Given a minimum leaf size c, each leaf of the tree must have c or more observations of each treatment class.

3. Estimate the treatment effect using Equation 4 on the leaf containing x.

Appendix B

Wager and Athey (2018) specify the following procedure for obtaining a double sample tree.

1. Draw a random subsample $\mathcal{K} \in \{1, \ldots, n\}$ of size s without replacement. Then, divide $\mathcal{K}$ into two disjoint sets $\mathcal{I}$ and $\mathcal{J}$ such that they both have size $\frac{s}{2}$.

2. Use recursive partitioning to grow a tree. The splits are chosen using any data from $\mathcal{J}$, but only the features (X) and treatment indicators (W) from $\mathcal{I}$.

3. Estimate leafwise responses using only the observations in $\mathcal{I}$.

Appendix C

A function f from $S \subset \mathbb{R}^n$ into $\mathbb{R}^m$ is Lipschitz continuous at x if there exists a constant K such that

$$||f(y) - f(x)|| \leq K||y - x|| \tag{25}$$

for all $y \in S$ sufficiently near x (O'Searcoid, 2006).