SEMINAR TERM PAPER

LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN
DEPARTMENT OF STATISTICS

Estimation and Inference of Heterogeneous Treatment Effects using Random Forests

Thomas Jürgensen
Matriculation Nr.: 12369869

Supervised by Dr. Daniel Wilhelm

Munich, August 31st, 2025

Contents

1 Introduction

One of the most important tasks in economics, medicine, statistics, and many other disciplines is estimating the causal effect of a treatment or intervention. Assuming that a treatment's impact is largely consistent across people and observations, a large portion of the literature has historically concentrated on estimating average treatment effects (ATEs). However, on many different scenarios, the effects of some sort of treatment are intrinsically heterogeneous, which means that they differ greatly among different individuals, observations or sub-groups. For instance, a newly developed medical intervention might be effective for younger patients but ineffective for elderly patients, or its effect might be different for men than it is for women. Similarly, an educational program may have little effect on high achievers but improve results for students with inadequate prior preparation.

The development of adaptable techniques that can reveal individual differences in treatment response has been spurred by the growing interest in these so-called **heterogeneous treatment effects**. However, there are a number of difficulties in precisely estimating them. In high-dimensional settings or when the form of heterogeneity is complex, traditional statistical techniques like regression with interaction terms or subgroup analysis can lose their reliability. Moreover, a lot of machine learning techniques are very good at predicting results, but they are not made for drawing conclusions about causality, and they usually don't have reliable instruments for measuring uncertainty.

In response to this methodological gap, Wager and Athey (2018) propose the **causal forest**, an adaptation of Breiman's random forest algorithm (Breiman 2001) tailored for the estimation of **conditional average treatment effects** (or CATEs). Their approach not only offers the flexibility of nonparametric machine learning but also provides a theoretical framework for statistical inference, including pointwise confidence intervals for individual treatment effects. This is made possible through key innovations such as the use of **honest** trees, which separate the data used for tree construction from the data used for estimation, and a consistent variance estimation technique based on the infinitesimal jackknife for random forests developed by Wager, Hastie, and Efron (2013). The causal forest method thus represents a significant advancement in the field of causal inference. It bridges the gap between machine learning's ability to capture complex relationships and the statistical rigor needed for credible statistical inference.

The remainder of this term paper will provide a detailed overview of Wager and Athey's methodology, the theoretical guarantees supporting causal forests, empirical performance as demonstrated in an empirical example, and a discussion of the method's limitations and potential extensions.

1.1 Literature Review

Following the development of causal forests, a number of studies have explored their application, refinement, and theoretical foundations. Athey and Wager (2019) applied the method to education data, highlighting practical considerations such as accounting for clustered errors and discussing how causal forests use propensity scores to be more robust to confounding. Davis and Heller (2017) demonstrated its usefulness in randomized trials for youth employment programs, using estimated CATEs to identify subgroups with the largest responses to the intervention, and Lechner (2018) proposed modifications to extend the causal forest framework to models involving multiple treatments, enhancing flexibility in policy evaluation and stratified causal effect estimation.

Further theoretical and applied developments have broadened the scope of the method. Gavrilova, Langørgen, and Zoutman (2025) developed a difference-in-differences causal forest, which provide consistent estimates with a parallel trend assumption. The methodology has also begun to see broader uptake in clinical and epidemiological research: Susukida et al. (2024) used causal forests to explore heterogeneous treatment effects in psychosocial interventions for substance use disorder, revealing nuanced findings like limited overall heterogeneity but potential subgroup-specific effects.

A particularly important recent contribution was made by Cattaneo, Klusowski, and Tian (2022), who offer a rigorous theoretical perspective on recursive partitioning methods for pointwise inference. They demonstrate that adaptive (non-honest) trees may fail to achieve even polynomial convergence rates and can be unreliable for inference, whereas random forests transform weak base learners into estimators with near optimal convergence properties through subsampling and random feature selection. Their results provide strong theoretical justification that supports the theoretical necessity of honesty, subsampling, and randomness for pointwise valid causal inference via forest methods.

2 Methodology Overview

Following the motivation outlined in Section 1, this second chapter introduces the methodology developed by Wager and Athey (2018): the **causal forest**. This method builds upon the traditional machine learning technique, the random forest, in order to estimate heterogeneous treatment effects in a flexible and statistically principled way.

2.1 Defining Causal Forests

2.1.1 Estimating Conditional Average Treatment Effects

Estimating the Conditional Average Treatment Effect (CATE), which is defined as

$$\tau(x) = \mathbb{E}[Y(1) - Y(0)|X = x], \tag{1}$$

is a common goal in causal inference. Here, $Y(1)$ and $Y(0)$ represent the potential outcomes under treatment and control, respectively, and X is a vector of observed covariates. The objective is to estimate the expected causal effect of receiving the treatment for any individuals with a given set of characteristics $X = x$. However, estimating $\tau(x)$ is inherently challenging because only one of the two potential outcomes is observed for each individual, since each person can only be either treated or not treated. To make the estimation feasible, the standard assumptions of **unconfoundedness** and **overlap** are imposed:

- **Unconfoundedness**: According to this assumption, the treatment assignment is essentially random, conditional on the covariates X, and is independent of the potential outcomes, i.e., $Y(1), Y(0) \perp W|X$, meaning that treatment assignment W, conditional on covariates X, is essentially random.
- **Overlap**: For every covariate x, both treated and control units are observed because the probability of receiving treatment is positive for all values of X but not equal to 1, i.e., $0 < \mathbb{P}(W = 1|X = x) < 1$ for all x, ensuring that both treated and control units are observed for every covariate x.

The CATE $\tau(x)$ can now be estimated from the observed data because of these assumptions, which enable us to treat observational data as though it had originated from a randomized exper-

iment, conditional on covariates.

2.1.2 Causal Trees

The decision tree, a popular and extensively used machine learning algorithm that recursively splits the covariate space into disjoint regions (also called "leaves"), within which a basic model is applied, is a logical place to start. Conventional regression trees are designed to forecast outcomes rather than estimate causal effects. Causal trees, on the other hand, are intended to directly estimate treatment effects. The difference in average outcomes between treated and control units is used to estimate the treatment effect in each leaf:

$$\hat{\tau}(x) = \left(\frac{1}{n_1} \sum_{i:W_i=1, X_i \in L} Y_i \right) - \left(\frac{1}{n_0} \sum_{i:W_i=0, X_i \in L} Y_i \right), \tag{2}$$

where L is the leaf that contains the covariate x, n_1 is the number of treated units in leaf L, and n_0 is the number of control units in leaf L. It is crucial to note that, because of the unconfoundedness assumption, we can simply calculate the difference in average outcomes between treated and control units, and this difference would recover a causal effect. Without this assumption, the difference in observed outcomes would not necessarily reflect the true causal effect, as there could be confounding factors that influence both treatment assignment and outcomes.

Although causal trees offer flexible estimates of the localized effects of treatments, a single tree is highly sensitive to data perturbations and prone to high variance, as is well known from traditional machine learning. The causal forest algorithm is based on an ensemble of trees that are used to assess this problem.

2.1.3 Causal Forests

A causal forest is an ensemble of many causal trees, each built on a random subsample of the data. By aggregating the treatment effect estimates from many such trees, causal forests reduce variance and produce more stable treatment effect estimates. Formally, the causal forest estimator is given by:

$$\hat{\tau}(x) = \frac{1}{B} \sum_{b=1}^{B} \hat{\tau}_b(x), \tag{3}$$

where $\hat{\tau}_b(x)$ is the treatment effect estimate from the b-th causal tree, and B is the total number of trees grown in the forest.

However, an important issue arises in the construction of causal trees and forests. If the same data is used both to determine the tree structure (i.e., the locations of splits) and to estimate treatment effects within the resulting leaf nodes, the resulting estimates may suffer from overfitting. This is especially problematic for inference, since overfitting would lead to biased treatment effect estimates. More critically, this kind of bias would undermine the validity of any subsequent statistical inference, resulting in invalid and unreliable confidence intervals.

2.1.4 Honest Trees

To address this problem, Wager and Athey (2018) introduce the concept of **honest** trees, a design choice that separates the tasks of model selection and estimation.

Definition 1 (Honest Trees)[1]: *A tree is called **honest** if it uses separate subsamples for two distinct purposes:*

- *One subsample is used to determine the tree structure, that is, where to place the splits,*
- *The other is used to estimate the treatment effects within each leaf.*

This separation prevents the model from overfitting during tree construction, ensuring that the treatment effect estimates remain unbiased.

Honesty can be implemented in various ways. However, in causal forests, Wager and Athey (2018) suggest the implementation of honest trees through the so-called "double-sample" approach, where each subsample used to grow a causal tree is split into two halves: one half for splitting and tree growing, one the other half for the estimation of treatment effects.

2.2 Asymptotic Theory

As we have seen in Section 2.1, causal forests provide a very flexible and nonparametric method for estimating heterogeneous treatment effects. However, the main theoretical contribution of causal forests is that they permit valid statistical inference. The well-developed asymptotic

[1] This definition is adapted from Wager and Athey (2018).

theory of causal forests ensures both consistency and asymptotic normality of the treatment effect estimates under appropriate conditions and assumptions, in contrast to the majority of machine learning techniques, which are mainly optimized for prediction and do not provide uncertainty quantification. The main theoretical findings in favor of using causal forests for statistical inference are presented in this chapter.

2.2.1 Consistency of the Estimator

The first fundamental result of Wager and Athey (2018) is **pointwise consistency**. This result guarantees that, as the sample size n increases to infinity, the treatment effect estimator $\hat{\tau}(x)$ converges in probability to the true conditional average treatment effect $\tau(x)$ for any fixed covariate $x \in X$, i.e.,

$$\hat{\tau}(x) \xrightarrow{p} \tau(x) \quad \text{as} \quad n \to \infty. \tag{4}$$

This result is very important because it ensures that the causal forest estimator will recover the correct treatment effects at each data point x as the sample size grows, making it a reliable tool for statistical inference. To achieve pointwise consistency, one additional assumption is needed, and that is that both conditional mean functions $\mathbb{E}[Y(0)|X = x]$ and $\mathbb{E}[Y(1)|X = x]$ are Lipschnitz-continuous.

2.2.2 Asymptotic Normality

The second fundamental result of Wager and Athey (2018) is **asymptotic normality**. This result states that, under certain conditions, the treatment effect estimator $\hat{\tau}(x)$ is asymptotically normally distributed around the true treatment effect $\tau(x)$.

Theorem 1 (Asymptotic Normality)[2]:

Let $(\boldsymbol{X}_i, Y_i, W_i)_{i=1}^n$ be n i.i.d. training examples, where:

- $\boldsymbol{X}_i \in [0,1]^d$ *are the covariates,*
- $Y_i \in \mathbb{R}$ *is the observed outcomes,*
- $W_i \in \{0,1\}$ *indicates the binary treatment assignment.*

[2]This theorem is adapted from Wager and Athey (2018).

Suppose that these training examples satisfy the following conditions:

- *The treatment assignment W_i is unconfounded and has overlap,*
- *The conditional means $\mathbb{E}[Y(0)|\boldsymbol{X} = x]$ and $\mathbb{E}[Y(1)|\boldsymbol{X} = x]$ are Lipschnitz-continuous,*
- *The conditional variance is bounded, i.e. $\sup_x \text{Var}(Y|\mathbf{X} = x) < \infty$.*
- *The covariates are **independent** and **uniformly distributed**, i.e. $\boldsymbol{X}_i \sim \mathcal{U}([0,1]^d)$.*

*Given these conditions, let Γ be an **honest** causal forest, where:*

- *Each causal tree is built on a random subsample of size $s_n \propto n^\beta$ for some $\beta_{min} < \beta < 1$, where β_{min} depends on covariate dimension d and regularity parameter α,*
- *each causal tree is **α-regular**, meaning that every split sends at least an α-fraction of the subsample to each child node (in this case, $\alpha \leq 0.2$ is used),*

*Then, for any fixed covariate $x \in [0,1]^d$, the treatment effect estimator $\hat{\tau}(x)$ is **asymptotically normal** and **centered**. That is:*

$$\frac{(\hat{\tau}(x) - \tau(x))}{\sqrt{\text{Var}\left(\hat{\tau}(x)\right)}} \xrightarrow{d} \mathcal{N}(0,1) \quad \text{as} \quad n \to \infty. \tag{5}$$

Furthermore, the infinitesimal jackknife[3] (IJ) variance estimator, which is defined as:

$$\hat{V}_{IJ}(x) = \frac{n-1}{n}\left(\frac{n}{n-s}\right)^2 \sum_{i=1}^{n} \left(\text{Cov}_*\left[\hat{\tau}_b^*(x), N_{ib}^*\right]\right)^2, \tag{6}$$

where $\hat{\tau}_b^*(x)$ is the treatment effect estimate given by the b-th tree, and $N_{ib}^* \in \{0,1\}$ indicates whether the training example i was used for the b-th tree, is a consistent estimator of the variance, in the sense that:

$$\frac{\hat{V}_{IJ}(x)}{\text{Var}(\hat{\tau}(x))} \xrightarrow{p} 1 \quad \text{as} \quad n \to \infty. \tag{7}$$

This result establishes causal forests as a method suitable not only for the flexible, nonparametric estimation of heterogeneous treatment effects, but also for conducting **asymptotically valid statistical inference**. The asymptotic normality of the estimator enables the construction

[3]The infinitesimal jackknife is the name of a variance estimation method for random forests developed by Efron (2014) and Wager, Hastie, and Efron (2013).

of confidence intervals around the estimated treatment effect at each point x, thereby allowing researchers to quantify uncertainty in a principled way, i.e.

$$\hat{\tau}(x) \pm z_{1-\frac{\alpha}{2}} \cdot \sqrt{\hat{V}_{IJ}(x)}, \tag{8}$$

where $z_{\alpha/2}$ is the critical value from the standard normal distribution corresponding to the desired confidence level.

However, the conditions and assumptions that underlie the causal forest construction in **Theorem 1** are crucial to the asymptotic normality result. For instance, the estimator may become biased if the honesty condition is broken, which would mean that the same data is used to estimate treatment effects within leaves as well as to choose tree splits. Confidence intervals may become invalid as a result of this bias, which compromises the central limit theorem that underpins sound inference. Consequently, the method's theoretical guarantees, such as consistency and reliable statistical inference, might no longer be valid in the absence of these structural safeguards.

3 Empirical Analysis

4 Discussion

5 Conclusion

A Appendix

References

Athey, Susan, and Stefan Wager. 2019. "Estimating Treatment Effects with Causal Forests: An Application." https://doi.org/10.48550/ARXIV.1902.07409.

Breiman, Leo. 2001. *Machine Learning* 45 (1): 5–32. https://doi.org/10.1023/a:1010933404324.

Cattaneo, Matias D., Jason M. Klusowski, and Peter M. Tian. 2022. "On the Pointwise Behavior

of Recursive Partitioning and Its Implications for Heterogeneous Causal Effect Estimation." https://doi.org/10.48550/ARXIV.2211.10805.

Davis, Jonathan M. V., and Sara B. Heller. 2017. "Using Causal Forests to Predict Treatment Heterogeneity: An Application to Summer Jobs." *American Economic Review* 107 (5): 546–50. https://doi.org/10.1257/aer.p20171000.

Efron, Bradley. 2014. "Estimation and Accuracy After Model Selection." *Journal of the American Statistical Association* 109 (507): 991–1007. https://doi.org/10.1080/01621459.2013.823775.

Gavrilova, Evelina, Audun Langørgen, and Floris T. Zoutman. 2025. "Difference-in-Difference Causal Forests With an Application to Payroll Tax Incidence in Norway." *Journal of Applied Econometrics*, July. https://doi.org/10.1002/jae.70001.

Lechner, Michael. 2018. "Modified Causal Forests for Estimating Heterogeneous Causal Effects." https://doi.org/10.48550/ARXIV.1812.09487.

Susukida, Ryoko, Masoumeh Amin-Esmaeili, Elena Badillo-Goicoechea, Trang Q. Nguyen, Elizabeth A. Stuart, Michael Rosenblum, Kelly E. Dunn, and Ramin Mojtabai. 2024. "Application of Causal Forest Model to Examine Treatment Effect Heterogeneity in Substance Use Disorder Psychosocial Treatments." *International Journal of Methods in Psychiatric Research* 34 (1). https://doi.org/10.1002/mpr.70011.

Wager, Stefan, and Susan Athey. 2018. "Estimation and Inference of Heterogeneous Treatment Effects Using Random Forests." *Journal of the American Statistical Association* 113 (523): 1228–42. https://doi.org/10.1080/01621459.2017.1319839.

Wager, Stefan, Trevor Hastie, and Bradley Efron. 2013. "Confidence Intervals for Random Forests: The Jackknife and the Infinitesimal Jackknife." https://doi.org/10.48550/ARXIV.1311.4555.