Adaptive Estimation of Partially Identified Treatment Effects

Maria Nareklishvili *

Abstract

This paper proposes a non-parametric algorithm for estimating and inferring partially identified treatment effects. In particular, we introduce multivariate random forests that can be used to fit the bounds of treatment effects identified as the solution to a set of local moment equations. To detect heterogeneous subgroups, multivariate random forests adaptively search for subsets of data that exhibit the highest variation in the treatment effect bounds. We provide consistency guarantees for the estimators of the treatment effect bounds and derive their asymptotic normality under certain regularity conditions and sample splitting assumptions. Simulation experiments and applications to the National Longitudinal Survey of Youth reveal significant heterogeneity in the effect of the Head Start program on years of schooling.

Keywords: Partial identification, Random forests, Treatment effects.

JEL classification: C14, C31

*Department of Economics, University of Oslo, Oslo, Norway (Correspondence to: maria.nareklishvili@frisch.uio.no). This paper has benefitted from the constructive feedback of Knut Røed, Edwin Leuven, Alexander Torgovitsky, Thibaut Lamadon, Thomas Wiemann, Phillip Heiler, Nicholas Polson, California Econometrics Conference participants, Seminars at the University of Oslo, University of Aarhus, Ragnar Frisch Center for Economic Research, as well as the Computational Workshop participants at the University of Chicago. All the remaining errors are my own. I acknowledge the funding by the Research Council of Norway, under project number 280350/GE.

1. Introduction

A feature common to a wide range of social science phenomena is the heterogeneity of treatment effects across units of analysis. Individuals differ not only in their background characteristics but also in how they respond to a particular treatment, intervention, or stimulation (Chernozhukov et al., 2017; Athey and Imbens, 2016; Athey and Wager, 2019). Randomized control trials, commonly used to infer the policy effect on an outcome, can be costly, unethical, or impossible to implement. In that case, subjects can self-select into the treatment (Heckman, 1979). For example, years of education typically correlate with the error of earnings, and conventional estimation methods, such as ordinary least squares, are biased.

The aim of this paper is to study how the effects of a policy or treatment can vary among different subgroups of the population in observational data, when a randomized control trial is not available. Specifically, we look at whether the treatment effect bounds can be made more precise and informative by conditioning on observable characteristics; whether the conditional bounds reflect subgroups of the population that gain or lose the most from the policy; and lastly, we assess the statistical and economic significance of any heterogeneity or nonlinearity observed in the treatment effect bounds.

We use bounded support of the outcome and the monotonic treatment selection assumption (as described by Manski, 1990; Manski and Pepper, 2000, 2009) to identify and estimate the range of treatment effects for different subgroups of the population. The monotonic treatment selection assumption means that the selection into a binary treatment is either positive or negative in a given subset of the population. As an example, imagine that the treatment is participation in a preschool program, and the outcome is years of schooling. In this case, the monotonic treatment selection implies that, for any potential treatment level, those who participated in the program are expected to have either higher or lower mean years of schooling compared to those who did not.

In this paper, we propose multivariate random forests to estimate and infer the bounds of treatment effects. This method extends the causal forest approach of Athey and Wager (2019) to a multivariate setting, and is particularly well-suited for partially identified parameters. Specifically, using local moment equations, the random trees in the multivariate forest search for subsets of the data that show the greatest variation in the treatment effect bounds. This can provide valuable insights into how the effects of a treatment or a policy may vary

across different subgroups of the population. Our analysis of the large sample theory shows that, under standard assumptions about the distribution of the data and the use of sample splitting, the multivariate forest estimates are asymptotically normally distributed.

The first contribution of this paper is the introduction of a flexible direction for the monotonic treatment selection assumption. The existing work in the literature often assumes a fixed direction for this assumption, regardless of the observed values of the covariates (Jiang et al., 2014; De Haan and Leuven, 2020). The predefined sign of the unconditional treatment selection can be economically justified in many settings (Beresteanu and Manski, 2000; De Haan, 2011, 2017), but it may be difficult to do so for specific subgroups of the population. In this article, the selection into the treatment has the same sign as the observed difference in mean outcomes between the treated and control groups. We conjecture and show that this added flexibility can increase the coverage and informativeness of the treatment effect bounds.

The second contribution of this paper lies at the intersection of two cultures. It uses machine learning tools to deduce statistical inference on multiple correlated coefficients (Athey and Imbens, 2016; Athey et al., 2019). The novelty of the random forests in this paper is the ability to detect and measure heterogeneity in partially identified treatment effects. Specifically, we show that multivariate random forests have desirable large sample properties and construct the loss function tailored to the bounds of treatment effects. The trees in multivariate random forests are grown based on a combination of two loss functions: the inverse covariance weighted local loss function (Segal and Xiao, 2011) and the asymmetric least absolute deviation loss (Koenker and Hallock, 2001; Hao et al., 2007; Chernozhukov and Hansen, 2006). By minimizing these loss functions simultaneously, we can increase the coverage and informativeness of the treatment effect bounds (Nekipelov et al., 2018). The moment function in this article is similar to the ones introduced by Athey and Imbens (2016) and Chernozhukov et al. (2018). However, it differs from them by its multivariate nature and allows us to model multiple correlated means and quantiles of the outcomes, re-scaled to their corresponding covariance.

In this article, we use the National Longitudinal Survey of Youth (De Haan and Leuven, 2020) to conduct simulation experiments and empirical analyses of early childhood education programs such as Head Start. The treatment is binary and indicates the Head Start (or any other program) participation, and the outcome is years of schooling. Data contain background characteristics, such as age, gender, race, and parental education. According to

simulation experiments, multivariate random forests in this study are twice as informative and precise as their univariate counterpart proposed by Athey and Imbens (2016). The results align with those of De Haan and Leuven (2020), indicating that treatment effect bounds are significantly heterogeneous. In particular, we find that the average treatment effect is negative for participants with highly educated parents and no siblings, but we cannot rule out the positive effect of the program for individuals with disadvantaged family backgrounds. Overall, our analysis suggests that participants who are most in need of the program are more likely to receive the highest benefits. These findings provide important insights into the effectiveness of early childhood education programs, particularly for disadvantaged subgroups of the population. By examining the heterogeneity of treatment effect bounds, we can gain a better understanding of the factors that influence educational attainment and inform future policy decisions.

2. Related Literature Review

2.1. Random Forests

Despite the extensive practical use of random forests, most early work on the theoretical side of random forests concentrates on stylized or simplified versions of the original method. Breiman (2001) offers an upper bound on the generalization error of forests in terms of correlation and strength of the individual trees. This is followed by Breiman (2004) that focuses on a stylized version of the original algorithm. Lin and Jeon (2006) highlight an interesting connection between random forests and a particular class of nearest neighbor predictors and establish the lower bound of the expected mean squared error for nonadaptive forests (i.e., independent of the training set). Meinshausen and Ridgeway (2006) show the consistency of random forests in the context of conditional quantile prediction. A paper by Biau (2012) proves the consistency of random forests which requires independence of the candidate splits and the predicted leaf outcome. Denil et al. (2014), Wager (2014) and Scornet et al. (2015) propose consistency of random forests that are the closest to the original algorithm in a sparse feature space.

Athey and Imbens (2016) and Wager and Athey (2018) take a step forward in forest exploration by introducing causal forests for heterogeneous treatment effect analysis. They consider a univariate treatment effect estimator and prove the asymptotic normality of the estimated treatment effects based on the Hajek projection of a U-statistic (Korolyuk

and Borovskich, 2013). Athey et al. (2019) introduce moment conditions of the outcome variable and generalize the method to a broader class of the parameters. Another interesting work by Nekipelov et al. (2018) shows the uniform consistency of random forests with multiple correlated parameters for classification problems. A paper by Li (2020) generalizes asymptotic theory for random forests to network data. The author investigates correlated coefficients across multiple subsets (i.e., leaves) of a given tree. Additionally, Wang et al. (2021) introduce random forests for the instrumental variable approach.

Chernozhukov and Hansen (2006) propose quantile instrumental variable regression for heterogeneous treatment effect analysis. A closely related article by Chernozhukov et al. (2018) introduces a generic machine learning approach to estimate and infer key features of heterogeneous treatment effects in randomized experiments. They proxy conditional average treatment effects by a given machine learning approach and post-process them for inferring treatment effects. Their approach is also valid for high-dimensional data. Additionally, Belloni et al. (2017) discuss inference on heterogeneous treatment effects based on high-performing machine learning methods.

Multivariate random forests complement the existing theoretical work by extending the large sample theory to a multivariate setup. Multivariate random forests have the ability to jointly predict the means and quantiles of the outcome variables for estimation and inference on partially identified treatment effects. In this paper, the multivariate random forest relies on local moment equations that are generalizations of Chernozhukov et al. (2018); Chernozhukov and Hansen (2006) and Athey and Imbens (2016) to partially identified parameters. Specifically, we assume, each personalized treatment effect bound is the sum of the group-level treatment effect bound and the error. The goal is to minimize the deviation between the unobserved personalized bounds, and the observed group-level means of these bounds. Furthermore, since the treatment effect bounds consist of outcome bounds (or quantiles), we simultaneously minimize the asymmetric least absolute deviation loss. The moment function in this study is also similar to generative adversarial networks, where we minimize the deviation of the covariates with respect to their expected value (Creswell et al., 2018; Liu et al., 2021).

2.2. Partial Identification

Set identification of treatment effects has a long history (Molinari, 2020; Heckman, 2000). Manski (1990) identifies bounds of the average treatment effect under bounded support of

the outcome. Balke and Pearl (1997) propose bounds of the average treatment effects with a random selection and non-compliance. Horowitz and Manski (2000) develop nonparametric bounds of the treatment effects under a non-random treatment selection. Manski and Pepper (2000) evoke weak monotonicity assumptions of the mean potential outcomes to set identify treatment effects. A closely related paper by Beresteanu and Manski (2000) offers conditional bounds based on the monotonicity assumptions proposed in a series of papers by Charles Manski. Their method requires a low-dimensional covariate space, and the monotonicity assumption has a predetermined direction for each observation.

Alternatively, Lee (2009) proposes a trimming procedure for identifying the bounds of treatment effects. The author assumes that the treatment effect on selection has the same sign for all subjects. Semenova (2020) generalizes treatment effect bounds by Lee (2009) and allows pretreatment covariates to determine the sign of this effect. The weak conditional monotonicity assumptions by Semenova (2020) induce higher flexibility to incorporate a large set of covariates and tighten the bounds. Heiler (2022) builds on works of Lee (2009) and Semenova (2020) and shows that conditional Lee bounds exploit higher variation across subsets of the population. The author estimates conditional Lee bounds based on the local moment functions and provides theoretical guarantees for the estimators of these bounds.

Other related methods are offered by Heckman and Vytlacil (1999), Shaikh and Vytlacil (2011). Heckman and Vytlacil (1999) quantify conditional bounds of different treatment effect parameters in a latent variable framework. Shaikh and Vytlacil (2011) build on the work of Heckman and Vytlacil (2001) and Heckman and Vytlacil (1999), and offer bounds of the conditional average treatment effect under a binary treatment and a binary outcome. Extensive theoretical work for identification and inference on partially identified parameters is proposed by Chernozhukov et al. (2007); Fan and Park (2010); Romano and Shaikh (2010); Beresteanu et al. (2011); Huber and Mellace (2015); Mogstad and Torgovitsky (2018); Torgovitsky (2019).

The bounds in this work are based on ideas combined from Manski (1990); Manski and Pepper (2000); Lee (2009); Chernozhukov et al. (2017). The generalized monotonic treatment selection assumption in this article is weaker than the one of one-directional treatment selection. Additionally, the method allows for the incorporation of a large number of covariates as long as the dimension of these covariates grows at a slower rate relative to the number of observations.

3. Illustrative Example

In many cases, the selection into a treatment is fixed in the population. For example, De Haan and Leuven (2020) assume a negative selection into the Head Start program in order to identify the bounds of the program's effect on various labor market outcomes. The negative selection into the treatment implies that, on average, individuals who participate in the program have a weakly lower outcome compared to those who do not participate in the program, regardless of their potential treatment status. De Haan and Leuven (2020) assume that the negative selection into the treatment holds conditional on an individual's observable characteristics. De Haan and Leuven (2020) demonstrate that, when conditioned on background information such as parental education, race, and gender, the bounds of the Head Start effect on earnings are highly informative. In this paper, we allow for the direction of treatment selection to depend on the covariates. Specifically, when the difference in the observed mean outcomes of the treated and control groups is positive within a given subset of the population, the selection into the treatment is also positive. Conversely, when this quantity is negative, the selection into the treatment is also negative.

Figure 1 illustrates the heterogeneity of treatment effect bounds under two different assumptions about the direction of treatment selection. On the left, the selection into the treatment is fixed to negative across all dimensions of the covariate space, while on the right, a flexible monotonic treatment selection assumption is used. We employ causal forests (Wager and Athey, 2018) to estimate the bounds under the fixed monotonicity assumption (left), and multivariate random forests to estimate them under the flexible monotonicity assumption (right). The analysis is based on data from the National Longitudinal Survey of Youth (NLSY, De Haan and Leuven, 2020).

According to Figure 1, the coverage under a flexible monotonicity assumption increases from 0.513 to 1.000, while the mean squared error decreases from 1.140 to 0.577. This indicates that the bounds are twice as likely to be informative under the flexible assumption compared to the fixed assumption. However, the improvement in the bounds comes at the cost of decreased heterogeneity, as the bounds obtained under a flexible monotonicity assumption are smoother than those obtained under a fixed assumption. An additional example is provided in Appendix 12.1 to further illustrate the desirable finite sample properties of random forests.

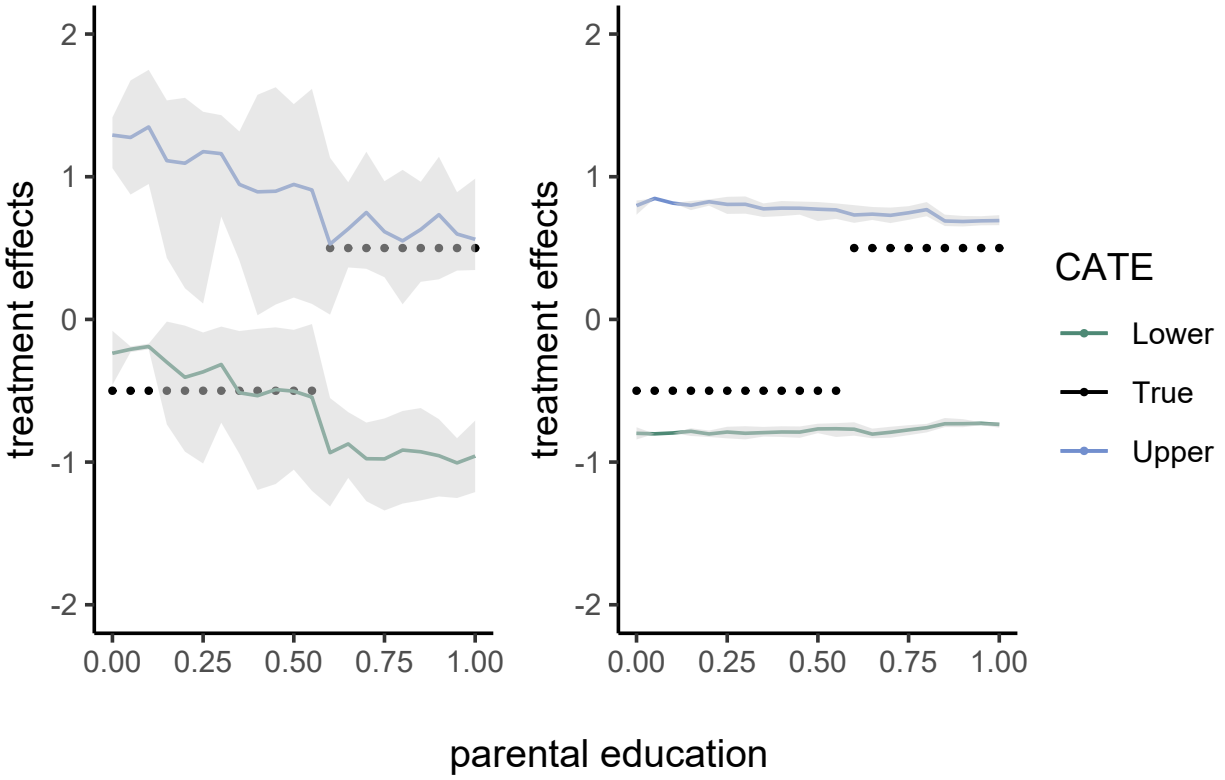


Figure 1. The bounds of the conditional average treatment effects when the selection into the treatment is fixed to negative (left), and when the selection into the treatment changes across different subsets of the covariate space (right). The green and blue lines represent the mean of the bounds across parental education, while the dotted lines are the simulated values of the Head Start effect on the outcome. The dashed gray area covers the 0.025 and 0.975 quantiles of the treatment effects. The left plot is obtained by estimating the bounds separately using causal forests, and the right plot is obtained by estimating the bounds jointly based on multivariate random forests, as detailed in Section 8.

4. Problem Formulation and the Setup

This section introduces the framework for partial identification of treatment effects with an endogenous treatment.

4.1. Treatment Effect Bounds

Consider a binary treatment $D_i \in \{0, 1\}$, and assume it correlates with the error of outcome ε_i for each individual $i = 1, \ldots, N$. We adopt the Neyman-Rubin potential outcomes framework (Rubin, 1974, 2006). Let the potential treatment and outcome be denoted by d and $Y_i(d) \in \mathbb{R}$ for $d \in \{0, 1\}$, respectively. The outcome of each subject i is given as $Y_i \equiv Y_i(1)D_i + Y_i(0)(1 - D_i)$. Assume $X_i \in \mathcal{X}$ is a $p-$dimensional vector of covariates.

Manski and Pepper (2009) focus on the bounds of the average treatment effect when the selection into the treatment D_i is not random. In this paper, the quantity of interest is the conditional average treatment effect

$$\theta(x) = \mathbb{E}(Y_i(1) - Y_i(0)|X_i = x).$$

The individual-level coefficients $\theta_i = Y_i(1) - Y_i(0)$ cannot be estimated without additional assumptions for two reasons. First, each individual is only observed in one treatment condition, not both. Second, the treatment assignment D_i is not random, so standard methods, such as ordinary least squares, cannot be used to identify θ_i. Formally, we can decompose $\theta(x)$ as:

$$\begin{aligned}\mathbb{E}\big(Y_i(1)|X_i = x\big) = \mathbb{E}\big(Y_i(1)|D_i = 1, X_i = x\big) \cdot P(D_i = 1|X_i = x) + \\ \mathbb{E}\big(Y_i(1)|D_i = 0, X_i = x\big) \cdot P(D_i = 0|X_i = x),\end{aligned} \tag{4.1}$$

$$\begin{aligned}\mathbb{E}\big(Y_i(0)|X_i = x\big) = \mathbb{E}\big(Y_i(0)|D_i = 0, X_i = x\big) \cdot P(D_i = 0|X_i = x) + \\ \mathbb{E}\big(Y_i(0)|D_i = 1, X_i = x\big) \cdot P(D_i = 1|X_i = x).\end{aligned} \tag{4.2}$$

Estimating the potential outcomes is challenging because each subject i is only observed in one treatment state. We can observe the average outcome of treated individuals, $\mathbb{E}(Y_i(1)|D_i = 1, X_i = x)\big)$, and of untreated individuals, $\mathbb{E}(Y_i(0)|D_i = 0, X_i = x)\big)$. We also observe the proportion of treated observations, $P(D_i = 1 \mid X_i = x)$, and of untreated observations, $P(D_i = 0 \mid X_i = x)$. However, $Y_i(1)$ and $Y_i(0)$ are unobserved for an untreated and treated subject, respectively. Therefore, the mean potential outcomes for the treated and untreated groups, $\mathbb{E}(Y_i(0)|D_i = 1, X_i = x)$ and $\mathbb{E}(Y_i(1)|D_i = 0, X_i = x)$, respectively, are unknown.

We employ assumptions introduced by Manski (1990), Manski and Pepper (2000) and Manski and Pepper (2009) that allow partial identification of the treatment effects.

Assumption 4.1 (**Bounded Support of the Outcome**). *Conditional on X_i, outcome $Y_i \in \mathbb{R}$ has bounded support, $Y_i|X_i \in [Y^L(X_i), Y^U(X_i)]$.*

Assumption 4.2 (**Monotonic Treatment Selection**). *Under any potential treatment status, treated units have the highest or the lowest conditional mean outcome:*

$$\begin{aligned}&\mathbb{E}\big(Y_i(d)|D_i = 1, X_i = x\big) \geq \mathbb{E}\big(Y_i(d)|D_i = 0, X_i = x\big) \text{ or} \\ &\mathbb{E}\big(Y_i(d)|D_i = 1, X_i = x\big) \leq \mathbb{E}\big(Y_i(d)|D_i = 0, X_i = x\big).\end{aligned}$$

Assumption 4.1 implies that the lower and upper bounds of the outcome, $Y^L(X_i)$ and $Y^U(X_i)$, depend on the covariates. Assumption 4.1 is automatically satisfied when the outcome is binary or ordinal. After a careful choice of the number of subgroups, Assumption 4.1 can be credible when the outcome is continuous. Assumption 4.2 is intuitive in many settings, as it allows for the possibility that certain individuals may be more likely to receive the treatment due to their characteristics. For example, highly motivated students may be more likely to graduate from a higher educational institution, implying that treated individuals are positively selected. Alternatively, non-treated individuals who inherited the family business may, on average, have higher earnings compared to treated individuals, implying that subjects are negatively selected into the treatment.

In this article, the sign of the observed mean difference in the outcomes of the treated and untreated groups determines the direction of the selection. Specifically, if $\mathbb{E}\big(Y_i|D_i = 1, X_i = x) \geq \mathbb{E}\big(Y_i|D_i = 0, X_i = x)$, the selection is positive, and the opposite if $\mathbb{E}\big(Y_i|D_i = 1, X_i = x) < \mathbb{E}\big(Y_i|D_i = 0, X_i = x)$.

Remark 1. *When $\mathbb{E}(Y_i|D_i = 1, X_i = x) \approx \mathbb{E}(Y_i|D_i = 0, X_i = x)$, the sign of the treatment selection can be misclassified. The underlying assumption of the proofs and the results in this article is that the misclassification error is on the order $o(1)$. An alternative is to provide the confidence interval robust to misclassification (see Heiler, 2022), which is beyond the scope of this paper.*

Assumptions 4.1-4.2 yield partial identification of the unknown expected treatment effects.

Proposition 4.1. *Let Assumptions 4.1-4.2 hold. Then $\theta^L(x) \leq \theta(x) \leq \theta^U(x)$, where $\theta^L(x)$ and $\theta^U(x)$ are the sharp lower and upper bounds of the conditional average treatment effects (CATE) at a point $x \in \mathcal{X}$. If Assumption 4.2 is negative, then :*

$$
\begin{aligned}
\theta^L(x) &= \mathbb{E}\big(Y_i(1)|D_i = 1, X_i = x\big) - \mathbb{E}\big(Y_i(0)|D_i = 0, X_i = x\big) \\
\theta^U(x) &= \mathbb{E}\big(Y_i(1)|D_i = 1, X_i = x\big) \times P(D_i = 1|X_i = x) + \\
&\quad + Y^U(x) \times P(D_i = 0|X_i = x) - \mathbb{E}\big(Y_i(0)|D_i = 0, X_i = x\big) \times \\
&\quad \times P(D_i = 0|X_i = x) - Y^L(x) \times P(D_i = 1|X_i = x).
\end{aligned}
$$

On the other hand, if Assumption 4.2 is weakly positive:

$$\begin{aligned}\theta^L(x) = \mathbb{E}\big(Y_i(1)|D_i = 1, X_i = x\big) \times P(D_i = 1|X_i = x) + \\ Y^L(x) \times P(D_i = 0|X_i = x) - \mathbb{E}\big(Y_i(0)|D_i = 0, X_i = x\big) \times \\ \times P(D_i = 0|X_i = x) - Y^U(x) \times P(D_i = 1|X_i = x).\end{aligned}$$

$$\theta^U(x) = \mathbb{E}\big(Y_i(1)|D_i = 1, X_i = x\big) - \mathbb{E}\big(Y_i(0)|D_i = 0, X_i = x\big).$$

Proof. The proof is in Appendix 12.2. ■

Appendix 12.3 introduces an additional monotonic instrumental variable (Manski and Pepper, 2009) to tighten the suggested bounds. [1]

Remark 2. *The bounds in Proposition 4.1 are sharp conditional on x. However, their expected values across the covariate space are not. Therefore, we sacrifice sharpness by estimating these bounds based on methods, such as random forests. For additional details, see the paper by Semenova (2020) who generalizes and introduces sharp lee bounds (Lee, 2009) aggregated across the subsets of covariates.*

4.2. Quantile Treatment Effect Bounds

Non-parametric bounds in Proposition 4.1 are often wide and uninformative. The edges of the outcome represent the most extreme values of the outcome distribution. We consider different quantiles of the outcome to tighten the bounds.

If we order the outcome variable from the smallest to largest values, the outcome bounds in Assumption 4.1 are equivalent to the smallest and largest quantiles. Define the $\alpha-$quantile $(0 \leq \alpha \leq 1)$ of the outcome as

$$Q_\alpha(x) = \inf\{y : \mathcal{F}(y|X_i = x) \geq \alpha\},$$

where $\mathcal{F}(y|X_i = x) = P(Y_i \leq y|X_i = x)$ is the probability that, for $X_i = x$, Y_i is smaller than $y \in \mathbb{R}$. Then for a given x, Assumption 4.1 implies that

$$Y^L(x) = Q_0(x),$$
$$Y^U(x) = Q_1(x),$$

[1] To tighten the bounds and relax Assumption 4.1, we can impose the monotonic treatment response assumption described by Manski and Pepper (2000).

The mean potential outcome bounds that depend on the minimum and maximum values of the outcome are also called "no assumption" or "worst-case" bounds because no assumptions are made about the effect of treatment on the outcome (Lee, 2005; De Haan, 2011). If data consist of outliers, these bounds are typically wide and uninformative (Lechner, 1999). We conjecture and show that the deviation of α from 0 and 1 can substantially tighten the bounds.

Let $0 \leq \alpha_1 \leq 1$ and $0 \leq \alpha_0 \leq 1$. Under a positive treatment selection, the quantiles of the outcome in this paper are given as

$$Q_{\alpha_1}(x) = \inf\{y : \mathcal{F}(y|X_i = x) \geq \alpha_1 \ \& \ D_i = 1\},$$
$$Q_{\alpha_0}(x) = \inf\{y : \mathcal{F}(y|X_i = x) \geq \alpha_0 \ \& \ D_i = 0\},$$

whereas, under a negative treatment selection:

$$Q_{\alpha_1}(x) = \inf\{y : \mathcal{F}(y|X_i = x) \geq \alpha_1 \ \& \ D_i = 0\},$$
$$Q_{\alpha_0}(x) = \inf\{y : \mathcal{F}(y|X_i = x) \geq \alpha_0 \ \& \ D_i = 1\}.$$

For example, assume the selection into the treatment is positive. In that case, if $\alpha_1 = 0.9$, then the upper bound of the mean potential outcome for the control group, $\mathbb{E}(Y_i(1)|D_i = 0, X_i = x)$, is the 90th percentile of the treatment group outcome. The intuition behind the idea is to trim the outcome variable in the treatment and control groups and estimate the bounds of the outcome (Lee, 2009). When the selection into the treatment is negative, the treatment effect bounds become:

$$\theta^L(x) = \mathbb{E}\big(Y_i(1)|D_i = 1, X_i = x\big) - \mathbb{E}\big(Y_i(0)|D_i = 0, X_i = x\big) \tag{4.3}$$
$$\theta^U(x) = \mathbb{E}\big(Y_i(1)|D_i = 1, X_i = x\big) \times P(D_i = 1|X_i = x) +$$
$$Q_{\alpha_1}(x) \times P(D_i = 0|X_i = x) - \mathbb{E}\big(Y_i(0)|D_i = 0, X_i = x\big) \times$$
$$\times P(D_i = 0|X_i = x) - Q_{\alpha_0}(x) \times P(D_i = 1|X_i = x).$$

Bounds are the same under a positive selection into the treatment as in Proposition 4.1 where we replace $Y^U(x)$ and $Y^L(x)$ with $Q_{\alpha_1}(x)$ and $Q_{\alpha_0}(x)$, respectively.

5. Methodology

In this section, we adopt the approach of Wager and Athey (2018) and introduce the notation and definitions for a generic $m-$dimensional outcome variable. Next, we generalize the theoretical properties to partially identified parameters.

Assume the available training data $A_i = (Y_i, X_i)_{i=1}^N$ consist of the $m-$dimensional outcome $Y_i \in \mathbb{R}^m$ and a p-dimensional vector $X_i \in \mathcal{X}$. A tree recursively partitions the feature (covariate) space $\mathcal{X}$ and makes axis-aligned splits to estimate the conditional mean vector of the outcome $\mu(x) = \mathbb{E}(Y_i|X_i = x)$ at a test point x. An axis-aligned split is a pair $s = (j, c)$, where $j = 1, \ldots, p$ is a specific covariate (splitting coordinate), and $c \in \mathbb{R}$ is the value of this variable (splitting index). The partitioning is performed recursively such that the algorithm begins with considering the set $\mathcal{P}^{(0)} = \mathcal{X} \in R^p$ (*parent node* of the tree). For this set, we select the splitting coordinate $j : 1 \leq j \leq p$ and the splitting index c such that $\mathcal{P}^{(0)}$ is split into two children (non-overlapping rectangles denoted as nodes):

$$\mathcal{P}^{(1,1)} = \mathcal{P}^{(0)} \cap \{\widetilde{x} \in \mathcal{P}^{(0)} : \widetilde{x}_j \leq c\} \text{ and } \mathcal{P}^{(1,2)} = \mathcal{P}^{(0)} \cap \{\widetilde{x} \in \mathcal{P}^{(0)} : \widetilde{x}_j > c\}, \tag{5.1}$$

where $\widetilde{x}_j$ is the j-th coordinate of the vector $\widetilde{x}$ from the training data.

After the first split, the process is repeated for $\mathcal{P}^{(1,1)}$ and $\mathcal{P}^{(1,2)}$ separately until a given stopping criterion is met. The sequential splitting process is based on data that belong to the corresponding partition.

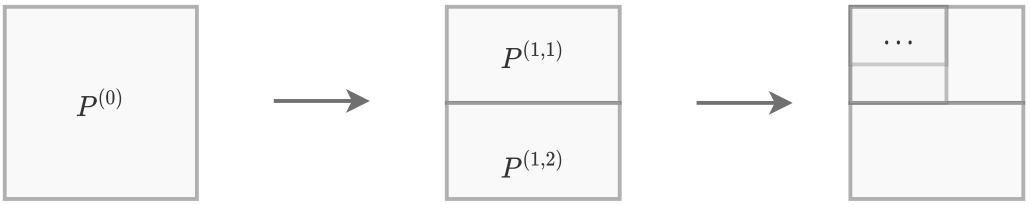


Figure 2. The splitting process of a tree.

The sequence of k splits induces the partition of $\mathcal{X}$ which we denote by Π. This partition consists of non-overlapping rectangular regions ℓ which we call leaves (or *terminal nodes*) of the tree. Figure 2 illustrates the process graphically. The union of the partitions is a covariate space:

$$\Pi = \{\ell_1, \ell_2, \ldots, \ell_{|\Pi|}\} \quad \text{and} \quad \cup_{n=1}^{|\Pi|} \ell_n = \mathcal{X}.$$

A common approach to finding the optimal splitting parameters $s = (j, c)$ in a regression setting is to minimize the sum of mean squared errors of L (left) and R (right) partitions:

$$(j, c) = \arg\min_{\tilde{j}, \tilde{c}} \sum_{i:X_i \in L} (Y_i - \mu_L)^T (Y_i - \mu_L) + \sum_{i:X_i \in R} (Y_i - \mu_R)^T (Y_i - \mu_R), \tag{5.2}$$

where μ_L and μ_R denote vectors of means of the outcome variable in partitions L and R, respectively.

Given the collection of $\{\ell_n\}_{n=1}^{|\Pi|}$ terminal nodes that form a partition of a feature space $\mathcal{X}$, a prediction of a tree at a generic test point x is defined as:

$$\widetilde{T} = [\widetilde{T}_1(x, \xi, A_i, \ldots, A_N), \widetilde{T}_2(x, \xi, A_i, \ldots, A_N), \ldots, \widetilde{T}_M(x, \xi, A_i, \ldots, A_N)], \tag{5.3}$$

with

$$\begin{aligned}
\widetilde{T}_1(x, \xi, A_i, \ldots, A_N) &= \sum_{n=1}^{|\Pi|} 1(x \in \ell_n) \frac{1}{N_{\ell_n}} \sum_{i: X_i \in \ell_n} Y_{i1}, \\
\widetilde{T}_2(x, \xi, A_i, \ldots, A_N) &= \sum_{n=1}^{|\Pi|} 1(x \in \ell_n) \frac{1}{N_{\ell_n}} \sum_{i: X_i \in \ell_n} Y_{i2}, \\
&\vdots \\
\widetilde{T}_M(x, \xi, A_i, \ldots, A_N) &= \sum_{n=1}^{|\Pi|} 1(x \in \ell_n) \frac{1}{N_{\ell_n}} \sum_{i: X_i \in \ell_n} Y_{iM}.
\end{aligned}$$

ξ is an external source of randomization, to allow for the randomized split selection procedures. $\mathbf{1}(x \in \ell_n)$ is an indicator function and equals one if a point $x \in \ell_n$, and zero otherwise. N_{ℓ_n} denotes the number of observations in a terminal node ℓ_n. A tree $\mathcal{T}(x, \xi, A_1, \ldots, A_N)$ is therefore a prediction at a point x based on data $\{A_i\}_{i=1}^N$ and a randomization parameter ξ. Lewis (2000) and Kingsford and Salzberg (2008) give a more detailed intuitive overview of classification and regression trees. Trees are easy to interpret and implement. They are not sensitive to outliers and missing data. Trees, however, have a high variance, are unstable, and are prone to overfitting. For these reasons, it is difficult to decide on the optimal tree structure. To overcome these issues, Breiman (2001) introduces the random forest algorithm.

Define $s < N$ to be a subset of size s from a population $i = \{1, \ldots, N\}$. $s = N^\beta$, where β is sufficiently close to 1 (Wager and Athey, 2018). Following Breiman (2001) and Wager and Athey (2018), we define the random forest estimator as the average of the tree estimators aggregated over all the possible size-s subsamples of the training data (marginalized over the auxiliary noise ξ). Specifically, the prediction of a multivariate random forest estimator at a particular test data point x is defined as

$$\mathcal{MF}(x, A_1, \ldots, A_N) = \frac{1}{\binom{N}{s}} \sum_{1 \le i_1 \le \cdots \le i_s \le N} \mathbb{E}_\xi \mathcal{T}(x, \xi, A_{i_1}, \ldots, A_{i_s}), \tag{5.4}$$

where $i_1, \ldots, i_s$ are the size-s subsamples of the population $\{i = 1, \ldots, N\}$. In practice, we estimate such a random forest by Monte Carlo averaging:

$$\mathcal{MF}(x, A_1, \ldots, A_N) \approx \frac{1}{B} \sum_{b=1}^{B} \mathcal{T}(x, \xi^*, A_1^*, \ldots A_N^*) \tag{5.5}$$

where $\{A_1^*, \dots A_N^*\}$ is drawn without replacement from $\{A_1, \dots A_N\}$. ξ^* is an auxiliary noise in a given sample and B is the number of sub-samples. $\mathcal{MF}(x, A_1, \dots, A_N)$ is a $1 \times M$ vector. Therefore, most of the arithmetic operations in this article are defined coordinate-wise in $\mathbb{R}^M$.

6. Large Sample Properties

In this section, we introduce underlying assumptions, large sample theory, and the method of moments estimator. Large sample theory of multivariate random forests strongly follows Wager and Athey (2018) and Athey et al. (2019). In particular, we discuss the assumptions and asymptotic properties of conventional random forest estimators in a multivariate setup. Then we show that the results generalize well to the bounds proposed in Subsection 4.1. We also extend the proofs to various quantiles and products of the outcome variables.

6.1. Assumptions

Random forests for partially identified parameters rest on similar assumptions introduced by Wager and Athey (2018). The first assumption we impose is the "honesty" of a tree.

Assumption 6.1 (**Honesty**). *Conditional on the observed covariates X_i, the outcome Y_{im} and the splitting parameters (the splitting coordinates and the splitting indices, $s = (j, c)$) are independent of each other. In particular, for each subject i where Y_{im} participates in the final prediction:*

$$\mathcal{F}(Y_{im}|X_i, s) = \mathcal{F}(Y_{im}|X_i).$$

$\mathcal{F}$ denotes the density of the corresponding m-th outcome variable for $m = 1, \dots, M$.

This assumption can be satisfied in multiple ways. As described by Athey and Imbens (2016), the first approach is to split the sample into two different partitions, train (S^{tr}) and estimation (S^{est}) data. We can use observations in S^{tr} and the features $X_i \in S^{est}$ to determine the splitting coordinates and indices (s) of the trees. However, the predicted outcome stems from the estimation sample S^{est}. The second approach is to find the optimal splitting parameters based on the features only:

$$(j, c) = \arg\min_{\tilde{j}, \tilde{c}} \sum_{i: X_i \in L} ||X_i - \mu_L(X_i)||^2 + \sum_{i: X_i \in R} ||X_i - \mu_R(X_i)||^2. \tag{6.1}$$

$\mu_L(X_i)$ and $\mu_R(X_i)$ denote the centers of the mass of X_i in each left and right partition, respectively. This approach is similar to a clustering algorithm that divides data into two distinct classes. The method in this article is based on the former.

Assumption 6.2 (**Random Split Trees**). *At each recursive step, the probability that the next split occurs at j-th covariate is bounded below by π/d for $\pi \in (0, 1]$, for all $j = 1, \ldots, p$.*

To guarantee consistency, the leaves of the trees have to become small in all dimensions of the feature space as N gets large. Based on Meinshausen and Ridgeway (2006) and Wager and Athey (2018), Assumption 6.2 guarantees that for all splitting steps, each variable is selected with probability at least π/d for some $0 < \pi \leq 1$.

Assumption 6.3 (**The Splitting Algorithm is (α, k)-regular**). *There exists $\alpha > 0$, where each split leaves at least a fraction α of the available training examples on each side of the split, and moreover, the splitting ceases at a node when it contains less than k observations for some $k \in \mathbb{N}$.*

Assumption 6.3 ensures that each half-space contains a sufficient number of observations. Wager and Walther (2015) show that Assumption 6.3 implies that half-spaces are also large in Euclidean volume. Assumption 6.3 places an upper bound on the number of observations in terminal nodes. In particular, a tree is fully grown to depth k, so that there are between $[k, 2k-1]$ observations in each terminal node of the tree. An immediate consequence of this assumption is that the variance of the tree estimator (at any test point x) is bounded above.

Assumption 6.4 (**Distributional Assumptions on the Data Generating Process**). *The features X_i are supported on the unit cube $X_i \in [0, 1]^p$, with a density that is bounded away from 0 and ∞. First and second moments, $\mathbb{E}(Y_{im}|X_i = x)$, $\mathbb{E}\big(\big(Y_{im}\big)^2|X_i = x\big)$, are Lipschitz-continuous for each m-th outcome variable, respectively ($m = 1, \ldots, M$). Furthermore, $Var(Y_{im}|X_i = x)$ is bounded away from 0 (i.e., $inf_{x \in \mathcal{X}} Var(Y_{im}|X_i = x) > 0$).*

Lipschitz-continuity and the bounded variances are commonly employed assumptions in the literature (Wager and Athey, 2018; Biau, 2012). The results of the paper do not explicitly depend on the distributional assumptions of X_i, however, they affect the constants that we carry throughout this paper (Lemma 2 and Theorem 3 in Section 3.2 in Wager and Athey, 2018).

Assumption 6.5 (**Overlap**). *We assume that for some $0 < \epsilon < 1$ and all $x \in [0, 1]^p$:*

$$\epsilon < \mathbb{P}(D_i = 1|X_i = x) < 1 - \epsilon.$$

Assumption 6.5 guarantees that for large enough N, there will be enough treated and control observations at any test point x.

6.2. Consistency and Asymptotic Normality

A random forest estimator is a U-statistic (Hoeffding, 1961; Korolyuk and Borovskich, 2013). A standard approach to investigate the large sample theory of random forests is to obtain the lower bound of its' Höeffding decomposition. Van der Vaart (1998); Hájek (1968) describe Höeffding decomposition (also known as the Hajek projection) in a univariate case.

Consider a vector-valued function $T \in \mathbb{R}^M$ which is measurable and permutation symmetric, that is, $T(\pi x) = T(x)$ for all $\pi \in \Pi$ (a tree in this setting). Then the Hajek projection of this function is defined as:

$$\dot{T} = \mathbb{E}(T) + \sum_{i=1}^{N} \big[\mathbb{E}(T|X_i) - \mathbb{E}(T)\big] = \sum_{i=1}^{N} \mathbb{E}(T|X_i) - (N-1)\mathbb{E}(T). \tag{6.2}$$

Intuitively, (6.2) is a projection of T onto the linear subspace of all random variables of the form $\sum_{i=1}^{N} g_i(X_i)$ with arbitrary measurable functions $g_i : \mathbb{R}^d \mapsto \mathbb{R}$ (such that $\mathbb{E}(g_i^2(X_i) < \infty$ for $i = 1, \ldots, N$).

It is clear that the conditional expectation of $\dot{T}$ in (6.2) is equal to the conditional expectation of T:

$$\begin{aligned} \mathbb{E}(\dot{T}|X_i) &= \mathbb{E}(T|X_i), \text{ and} \\ \mathbb{E}(\dot{T}) &= \mathbb{E}(T). \end{aligned} \tag{6.3}$$

Now consider the multivariate random forest estimator, $\mathcal{MF}(x, A_1, \ldots, A_N) \in \mathbb{R}^M$, and let the corresponding vector of means be μ. Moreover, let $\dot{\mathcal{MF}}(x, A_1, \ldots, A_N)$ and Σ denote the Hajek projection of the multivariate random forest estimator, and the covariance matrix of the Hajek projection, respectively. Assume also that the trees in $\dot{\mathcal{MF}}(x, A_1, \ldots, A_N)$ are symmetric and the observations are $i.i.d$ as before. Then Lemma 6.1 holds:

Lemma 6.1. *The Hajek projection of a multivariate random forest estimator and the covariance of this projection are given as:*

$$\begin{aligned} \dot{\mathcal{MF}}(x, A_1, \ldots, A_N) - \mu &= \frac{s}{N} \sum_{i=1}^{N} \big(\widetilde{T}_1(A_i) - \mu\big), \\ \Sigma &= \frac{s}{N} \mathbb{V}(\dot{\widetilde{T}}) \in \mathbb{R}^{M \times M}, \end{aligned}$$

where $\dot{\widetilde{T}} = \sum_{i=1}^{s} \widetilde{T}_1(A_i)$ *with* $\widetilde{T}_1(a) = \mathbb{E}_{\xi,A_2,\ldots,A_N}\widetilde{T}(x,\xi,a,A_2,\ldots,A_N)$ *is the Hajek projection of a tree* $\widetilde{T}(x,A_1,\ldots,A_N) = \mathbb{E}_{\xi}\widetilde{T}(x,\xi,A_1,\ldots,A_N) \in \mathbb{R}^M$. $s = N^{\beta}$ *as before and* M *is the dimension of the outcome variables in each terminal node.* $\mathbb{V}$ *denotes the covariance matrix of the projected elements of the tree.*

Proof. See Appendix 12.4. ■

Figure 8 in Appendix 12.4 visualizes the projection of random forests onto the subspace formed by all variables of the form $\sum_{i=1}^{N} g_i(X_i)$. Required conditions for the Lindeberg central limit theorem to hold are met (Billingsley, 2008; DiCiccio and Romano, 2020), therefore, the Hajek projection of the multivariate random forest estimator is asymptotically normally distributed:

$$\Sigma^{-1/2}\big(\dot{\mathcal{MF}}(x,A_1,\ldots,A_N) - \mu\big) \xrightarrow{d} \mathcal{N}(0,I),$$

where 0 is a $\mathbb{R}^M$ vector of zeros and $I_{M\times M}$ is an identity matrix. Our objective is to prove that the multivariate random forest estimator is asymptotically normal. By adding and subtracting $\Sigma^{-1/2}\dot{\mathcal{MF}}(x,A_1,\ldots,A_N)$ to $\Sigma^{-1/2}\big(\mathcal{MF}(x,A_1,\ldots,A_N) - \mu\big)$, we can easily verify that the multivariate random forest estimator is related to its projection the following way:

$$\Sigma^{-1/2}\big(\mathcal{MF}(x,A_1,\ldots,A_N) - \mu\big) = \Sigma^{-1/2}\big(\mathcal{MF}(x,A_1,\ldots,A_N) - \dot{\mathcal{MF}}(x,A_1,\ldots,A_N)\big) + \Sigma^{-1/2}\big(\dot{\mathcal{MF}}(x,A_1,\ldots,A_N) - \mu\big).$$

The objective of this article is to show that:

$$\Sigma^{-1/2}\big(\mathcal{MF}(x,A_1,\ldots,A_N) - \dot{\mathcal{MF}}(x,A_1,\ldots,A_N)\big) \xrightarrow{p} 0.$$

Then by Slutsky's theorem, the multivariate random forest estimator is asymptotically normally distributed.

Wager and Athey (2018) derive the lower bound of the variance of $\dot{\mathcal{MF}}(x,A_1,\ldots,A_N)$ and show that it converges to zero. We adopt the same approach in this study. The objective is to show that $\Sigma^{-1/2}\big(\mathcal{MF}(x,A_1,\ldots,A_N) - \dot{\mathcal{MF}}(x,A_1,\ldots,A_N)\big)$ converges in squared mean. For notational convenience, we denote $\mathcal{MF}(x,A_1,\ldots,A_N)$ and $\dot{\mathcal{MF}}(x,A_1,\ldots,A_N)$ as $\mathcal{MF}$ and $\dot{\mathcal{MF}}$, respectively.

Lemma 6.2. *The mean squared difference of* $\mathcal{MF}$ *and* $\dot{\mathcal{MF}}$ *has the upper bound:*

$$\mathbb{E}\big(\mathcal{MF} - \dot{\mathcal{MF}}\big)^T\Sigma^{-1}\big(\mathcal{MF} - \dot{\mathcal{MF}}\big) \leq \frac{s}{N} tr\left(\big(\mathbb{V}(\dot{\widetilde{T}})\big)^{-1}\mathbb{V}(\widetilde{T})\right),$$

where tr is a trace operator, and $\mathbb{V}(\widetilde{T})$ *and* $\mathbb{V}(\dot{\widetilde{T}})$ *denote the variance of a multivariate tree and its' Hajek projection, respectively.*

Proof. See Appendix 12.5. ■

Under Assumptions 6.1-6.5, Theorem 6.3 shows that $\frac{s}{N} tr\left(\left(\mathbb{V}(\dot{\widetilde{T}})\right)^{-1}\mathbb{V}(\widetilde{T})\right)$ approaches zero in the limit.

Theorem 6.3. *The entries of* $\mathbb{V}(\widetilde{T})$ *are bounded and its diagonal elements are bounded away from zero. Moreover, the lower bound of the off-diagonal terms of* $\mathbb{V}(\dot{\widetilde{T}})$ *are on the order of* $o\big(\frac{1}{\log^p(s)}\big)$*. The upper bound in Lemma 6.2 converges to zero in the limit:*

$$\frac{s}{N} tr\left(\left(\mathbb{V}(\dot{\widetilde{T}})\right)^{-1}\mathbb{V}(\widetilde{T})\right) \to 0.$$

Proof. See Appendix 12.6. ■

By Slutsky's theorem, Theorem 6.3 implies that the multivariate random forest estimator is asymptotically normally distributed. The proof is equivalent when the outcomes are treatment effects (see Appendix 12.6).

It was shown that the random forest approximates the means of different outcomes $\mathbb{E}(Y_{im}|X_i = x)$ for $m = 1, \ldots, M$. Meinshausen and Ridgeway (2006) show that the random forests can approximate the full conditional density of the outcome variable, and provide consistency of the estimator. As defined by Meinshausen and Ridgeway (2006), consider the density of the $m-$th outcome variable:

$$\mathbb{E}(y|X_i = x) = \mathbb{P}(Y_{im} \leq y|X_i = x) = \mathbb{E}(1_{\{Y_{im} \leq y\}}|X_i = x).$$

Just as $\mathbb{E}(Y_{im}|X_i = x)$ is approximated by the weighted mean over the observations of Y_i, define the approximation to $\mathbb{E}(1_{\{Y_{im} \leq y\}}|X_i = x)$ by a tree as

$$\hat{F}(y|X_i = x) = \widetilde{T}(x, \xi, A_i, \ldots, A_N) = \sum_{n=1}^{|\Pi|} 1(x \in \ell_n) 1_{\{Y_{im} \leq y\}}.$$

The multivariate random forest estimator $\mathcal{MF}(x, A_1, \ldots, A_N)$ is the average of multiple trees that estimate quantiles of the outcomes instead of the means. Under the proposed assumptions, Proposition guarantees asymptotic normality of the estimator.

Proposition 6.1. *The quantile regression forest with multiple outcome variables is consistent and asymptotically normally distributed:*

$$\Sigma^{-1}(\mathcal{MF}(x, A_1, \ldots, A_N) - \mu) \xrightarrow{d} \mathcal{N}(0, I).$$

Proof. See Appendix 12.8. ■

6.3. Implementation: Method of Moments

Define the parameter model for the bounds of treatment effects as

$$\theta^B(X_i) = \theta^B(X_i, \Pi) + v_i^B, \text{ where } \theta^B(X_i, \Pi) = \mathbb{E}\big[\theta^B(X_i)|X_i \in \ell(x, \Pi)\big].$$

$\theta^B(X_i, \Pi)$ is the conditional expectation of the treatment effect bounds in a given subset $\ell(x, \Pi)$. Assume, the latent error v_i^B is uncorrelated with the expected bound, $\theta^B(X_i, \Pi)$ for $B \in \{L, U\}$. Proposition 4.1 provides the estimands of $\theta^B(X_i, \Pi)$. In practice, to estimate the bounds, we replace the estimands with their unbiased sample analogs. Consider the tree predictions under a negative treatment selection. Then the predicted lower bound at a point x is given by:

$$\tilde{\theta}^L(x) = \frac{1}{|i : D_i = 1, X_i \in \ell_n|} \sum_{\{i:D_i=1, X_i \in \ell_n\}} Y_i - \frac{1}{|i : D_i = 0, X_i \in \ell_n|} \sum_{\{i:D_i=0, X_i \in \ell_n\}} Y_i.$$

For simplicity of notation, assume $P(D_i|X_i = x) = p_D$ for $D \in \{0, 1\}$. Further, assume, $\frac{1}{|i:D_i=1, X_i \in \ell_n|} \sum_{\{i:D_i=1, X_i \in \ell_n\}} Y_i = \bar{Y}^{(1)}$ and $\frac{1}{|i:D_i=0, X_i \in \ell_n|} \sum_{\{i:D_i=0, X_i \in \ell_n\}} Y_i = \bar{Y}^{(0)}$. Then the predicted upper bound at a point x is given as

$$\tilde{\theta}^U(x) = \bar{Y}^{(1)} \cdot p_1 + Q_{\alpha_1}(x) \cdot p_0 - \bar{Y}^{(0)} \cdot p_0 - Q_{\alpha_0}(x) \cdot p_1.$$

$Q_{\alpha_1}(x)$ and $Q_{\alpha_0}(x)$ denote the upper and lower quantiles of the outcome as before, respectively. Predictions are determined similarly under a positive treatment selection assumption. The proof in Subsection 12.8 in the Appendix generalizes the asymptotic theory for the proposed bounds.

The estimation and inference on partially identified parameters rely on the existence of conditional moment functions in each subset of the covariate space. Define the population conditional moment functions as

$$m(x, \theta) = \mathbb{E}\big[\big(\rho(A_i, \theta^\ell)|X_i = x\big] = 0. \tag{6.4}$$

θ^ℓ is a vector-valued parameter in a ℓ-th partition. The first assumption we impose is the existence of a solution in each ℓ-th subset [2].

Assumption 6.6. $\sup_{x \in \mathcal{X}} \big|\big|\mathbb{E}\big[\big(\rho(A_i, \theta^\ell)|X_i = x\big]\big|\big| = o(1)$.

Assumption 6.6 guarantees that the error in each partition does not diverge to infinity. Next, we consider Assumption 6.7 that guarantees that the inverse variance of the parameters is positive definite.

[2]Parameters can be heterogeneous across a subset of the covariate space $\mathcal{Z} \in \mathbb{R}^{N \times b}$, where $b \leq p$. In this study, $\mathcal{Z}$ coincides with $\mathcal{X}$.

Assumption 6.7. *For each subset ℓ there exists a matrix $\Omega(X_i) \in \mathbb{R}^{p \times p}$ such that eigenvalues of $\Omega(X_i)$ are uniformly bounded by a constant λ and*

$$\mathbb{E}\left[\Omega(X_i)\frac{\partial \rho(A_i, \theta^{\ell})}{\partial \theta^{\ell}}\right] \tag{6.5}$$

is strictly positive definite.

Under Assumptions 6.6-6.7, the unknown moment of interest is:

$$M_{\ell}(X_i, \theta) = \mathbb{E}\Big(\Omega(X_i)m(X_i, \theta)1(X_i \in \ell(x, \Pi)\Big). \tag{6.6}$$

We then estimate the conditional expectation, that yields $M(x, \theta)$. The optimal parameters minimize the corresponding sample analogue:

$$\widehat{\theta}_{\ell} = \arg\inf_{\theta} \widehat{M}_{\ell}(x, \theta) = \frac{1}{N_{\ell}} \sum_{i: X_i \in \ell(x, \Pi)} \kappa(x) m(x, \theta), \tag{6.7}$$

where N_{ℓ} is the number of observations in a leaf ℓ and $\kappa(x)$ is the sample analog of the weight. Segal and Xiao (2011) identify the optimal partitions of the data space based on the covariance weighted node impurity measure. Multivariate random forests inherit a similar loss function. The idea is to minimize the squared deviation between the personalized bounds $\theta(X_i)$ and their conditional group-level means $\theta(X_i, \Pi)$. Additionally, as the treatment effect bounds depend on the upper and lower quantiles of the outcome, $Q_{\alpha_1}(X_i)$ and $Q_{\alpha_0}(X_i)$, respectively, we simultaneously minimize the asymmetric mean absolute deviation:

$$Q_{\alpha_b}(x) = \arg\min_{q} \mathbb{E}(L_{\alpha_b}(Y_i, q)|X_i = x),$$

where $b \in \{0, 1\}$, q is the value of the outcome for α_b-th quantile, and

$$L_{\alpha_b}(y, q) = \begin{cases} \alpha_b \times |y - q| \text{ if } y > q \\ (1 - \alpha_b) \times |y - q| \text{ if } y \leq q \end{cases}.$$

Proposition 6.2 summarizes the method of moments estimator.

Proposition 6.2. *The optimal parameters of interest minimize the squared error, weighted by the inverse covariance matrix:*

$$\theta^*(x, S^{est}, \Pi) = \arg\inf_{\widetilde{\theta}} \mathbb{E}_{S^{te}, S^{est}, S^{tr}}\Bigg[\Big[\big(\theta(X_i) - \widetilde{\theta}(X_i, S^{est}, \Pi)\big)^T \Sigma^{-1} \big(\theta(X_i) - \widetilde{\theta}(X_i, S^{est}, \Pi)\big) - \theta(X_i)^T \Sigma^{-1} \theta(X_i)\Big] | X_i = x\Bigg] + \sum_{b} \mathbb{E}(L_{\alpha_b}(Y_i, q)|X_i = x).$$

The corresponding sample analog is given as

$$\hat{\theta}(x, S^{est}, \Pi) = \arg\max_{\widetilde{\theta}} \frac{1}{N^{tr}} \sum_{\ell} N_{\ell}^{tr} \big(\widetilde{\theta}(X_i, \Pi)^T \widehat{\Sigma}^{-1} \widetilde{\theta}(X_i, \Pi) | X_i = x \big) - \frac{1}{N^{tr}} \sum_{\ell} \sum_{b} N_{\ell}^{tr} L_{\alpha_b}(Y_i, q) | X_i = x \big). \tag{6.8}$$

Proof. See Appendix 12.10. ■

$\theta^*(x, S^{est}, \Pi)$ is a $M \times 1$ vector of population bounds predicted at an arbitrary test point x and a terminal node ℓ_n. $\widehat{\theta}(x, S^{est}, \Pi)$ is a $M \times 1$ vector of the corresponding sample analog. $\widetilde{\theta}(X_i, S^{tr}, \Pi)$ denotes the unbiased estimator of the treatment effect bounds, measured in the training sample. $\widehat{\Sigma}$ is the estimator of the covariance matrix of $\widetilde{\theta}(X_i, S^{tr}, \Pi)$, conditional on the number of observations in the estimation sample, N^{est} (equal to N^{tr} in this paper). Algorithm 1 summarizes the method.

Algorithm 1: Multivariate Random Forests for Partially Identified Treatment Effects

Require: number of trees (M_m), tree depth, number of leaves $|\Pi|$, the number of observations for each bootstrapped data set (s), number of observations in each leaf, data ($\mathcal{X}^{(0)}$, $\mathcal{Y}$, D).

Ensure: Predicted bounds.

1. Divide data into train (S^{tr}), estimation (S^{est}) and test samples (S^{te}).

if $\mathbb{E}(Y_{im}|D_i = 1, X_i = x) \geq \mathbb{E}(Y_{im}|D_i = 0, X_i = x)$ **then**

a) Assume a positive treatment selection.

else

b) Assume a negative treatment selection.

end if

for a) or b):

2. Identify the optimal partitions based on S^{tr} and the loss function in Proposition 6.2.

3. Estimate unbiased sample analogues of the bounds (Proposition 4.1) by using S^{est}.

5. Predict the bounds for an unknown observation x.

6.4. Inference

We evaluate the finite-sample performance of multivariate random forests based on two different measures:

$$\text{Coverage} = \frac{\sum_{i=1}^{N^{te}} \mathbb{1}\big[\widehat{\theta}^L(X_i, \Pi) \leq \theta_i \leq \widehat{\theta}^U(X_i, \Pi)\big]}{N^{te}},$$

$$\text{EMSE} = \frac{1}{2 \times N^{te}} \sum_{i=1}^{N^{te}} \big[\big(\widehat{\theta}^L(X_i, \Pi) - \theta_i\big)^2 + \big(\widehat{\theta}^U(X_i, \Pi) - \theta_i\big)^2\big],$$

where N^{te} is the number of observations in the test data. $\mathbb{1}\big(\widehat{\theta}^L(X_i, \Pi) \leq \theta_i \leq \widehat{\theta}^U(X_i, \Pi)\big)$ equals 1 if the simulated θ_i is within the estimated bounds, and zero otherwise. Coverage measures the frequency with which the estimated bounds recover known parameters. The expected mean squared error (EMSE) is a measure of the distance between the estimated bounds and the known parameter values θ_i. A high coverage indicates a high degree of credibility for the estimator in a given setting. On the other hand, a high value of EMSE indicates that the bounds are located far from the known parameter values, and are therefore uninformative.

Inference of the bounds is based on the infinitesimal jackknife. Athey and Wager (2019); Wager et al. (2014) show that the infinitesimal jackknife leads to a consistent estimator of the variance of the parameters. Moreover, Jennrich (2008) discusses that the infinitesimal jackknife is consistent even if the covariance structure is incorrectly specified.

Let $g = 1, \ldots, G$ be the g-th bootstrapped sample. We use a tree Π_g and the corresponding estimation sample S_g^{est} to obtain the bound $\hat{\theta}_g^B(x, S_g^{est}, \Pi_g)$ at a generic test point x. Next, the average of the individual tree estimates is $\hat{\theta}^B\big(x, \{S_g^{est}\}_{g=1}^G\}, \{\Pi_g\}_{g=1}^G\big) = \frac{1}{G}\sum_{g=1}^G \hat{\theta}_g^B(x, S_g^{est}, \Pi_g)$. Define N_{ig} as the number of times an observation i appears in the g-th bootstrapped sample (either the train or the estimation sample, S^{tr} or S^{est}). Then the following variance can lead to valid confidence intervals:

$$Var\big[\hat{\theta}^B\big(x, \{S_g^{est}\}_{g=1}^G\}, \{\Pi_w\}_{w=1}^W\big)\big] = \sum_{i=1}^N \Delta_g - \frac{N}{G^2}\sum_{g=1}^G \big[\theta_g^B(x, S_g^{est}, \Pi_g) - \hat{\theta}^B\big(x, \{S_g^{est}\}_{g=1}^G\}\big)\big]^2, \tag{6.9}$$

where $\Delta_g = \left[\dfrac{\sum_{g=1}^G (N_{ig} - 1)\big[\hat{\theta}_g^B(x, S_g^{est}, \Pi_g) - \hat{\theta}^B\big(x, \{S_g^{est}\}_{g=1}^G\}, \{\Pi_g\}_{g=1}^G\big)\big]}{G}\right]^2$.

The confidence interval based on the jackknife variance is valid for the bounds of the treatment effect. Imbens and Manski (2004) propose the confidence interval of the treatment effects directly under the "super-efficiency" condition. Specifically, they assume that

$\theta^U(x) > \theta^L(x)$. Stoye (2009) takes a step further and allows the weak inequality $\theta^U(x) \geq \theta^L(x)$. The generalization of the confidence interval given by Heiler (2022) allows for misspecification, such as $\theta^L(x) < \theta^U(x)$. In this work, we estimate confidence intervals of the treatment effect bounds, but do not report them as the standard deviations are small. Trees in the multivariate forest are built under the restriction that $\theta^U(x) > \theta^L(x)$.

Appendix 12.11 proposes a procedure for measuring the statistical significance of the heterogeneity of the bounds. Appendix 12.13 additionally illustrates details regarding the covariance matrix implemented in the algorithm.

7. Data

The analysis in this article is based on data from the National Longitudinal Survey of Youth 1979 (NLSY79; see Rothstein et al., 2019; De Haan and Leuven, 2020). This survey includes data on a sample of individuals who were between the ages of 14 and 22 in 1979 and were born between 1960 and 1964. These individuals were living in the United States at the time of the survey and were interviewed annually until 1994, and every other year thereafter.

The outcome variable is the years of schooling, reported in 1994 when the individuals were in their early 30s. The treatment is a Head Start participation indicator. The respondents were asked whether they attended Head Start or any other preschool program as a child.

Basic background information, such as age (birth year), gender, and race, is available in the data. Data also include parental education. Since education is more often missing for the father than for the mother, the main analysis uses the highest reported completed grade of either the mother or the father as a measure of parental education. The variable consists of the following categories: less than high school, some high school, high school, 1–3 years of college, and 4 years or more of college.

Overall, the sample contains 4434 individuals. For the analysis, we divide data into train and test partitions. Train data consist of 79% of the randomly chosen subjects, stratified across the Head Start participation. Moreover, for stable predictive performance, continuous variables are standardized. De Haan and Leuven (2020) provide a detailed summary of the variables.

8. Simulated Outcome

In this section, we design Monte Carlo simulations to generate the outcome, years of schooling, and investigate the properties of multivariate random forests. In each simulation setup, the Head Start effect on years of schooling is positive for individuals with highly educated parents, and negative for the participants with low-educated parents. The outcome is a function of the treatment and its interaction with parental education.

$$\begin{aligned} \mathrm{Y}_i &= 13.240 - 0.5 \cdot D_i + \mathbb{1}\{\text{Parental Education}_i \geq 0.6\} \cdot D_i + X_i^R \cdot 0 + \varepsilon_i, \\ \varepsilon_i &\sim \mathcal{N}(0, 1 + D_i). \end{aligned}$$

Y_i denotes years of schooling for a participant i. D_i is a binary variable and equals one for the Head Start participants, and zero otherwise. $\mathbb{1}\{\text{Parental Education}_i \geq 0.6\}$ denotes a binary variable which equals one when the standardized parental education is more than 0.6, and zero otherwise. X_i^R are the characteristics (black, female, Hispanic, number of siblings, age) that do not influence the outcome. In each out of 100 experiments, the number of trees equals 200, the maximum depth of the tree is 10, and the minimum number of observations within each leaf is 5. Trees are grown on a subset of size 3472, and test data consist of 932 individuals. The characteristics of individuals correlate with each other. For example, parental education is highly correlated with both age and the number of siblings, with correlation coefficients of -0.39 and -0.43, respectively. The task of the multivariate random forests is, therefore, to recover the most important characteristics, parental education in this case, with the highest probability, followed by the other correlated characteristics such as age and number of siblings.

8.1. Splitting Accuracy

Figure 3 illustrates the variable importance based on multivariate random forests. Specifically, the plot shows the density of the effect of each independent variable on the upper and lower bounds of the years of schooling. As shown in the figure, the effects of parental education, number of siblings, and age on the outcome bounds are significantly different from zero, indicating that these variables are important in explaining the heterogeneity in the treatment effect bounds. In contrast, the other variables do not explain a significant amount of variation in the outcome bounds. This is likely due to the fact that the variables, such as black, Hispanic, and female, have a low correlation with parental education, with

correlation coefficients of -0.05, 0.03, and 0.02, respectively. Thus, Figure 3 verifies that the algorithm is able to effectively differentiate between variables that are redundant and those that are relevant for explaining the heterogeneity in the treatment effect bounds.

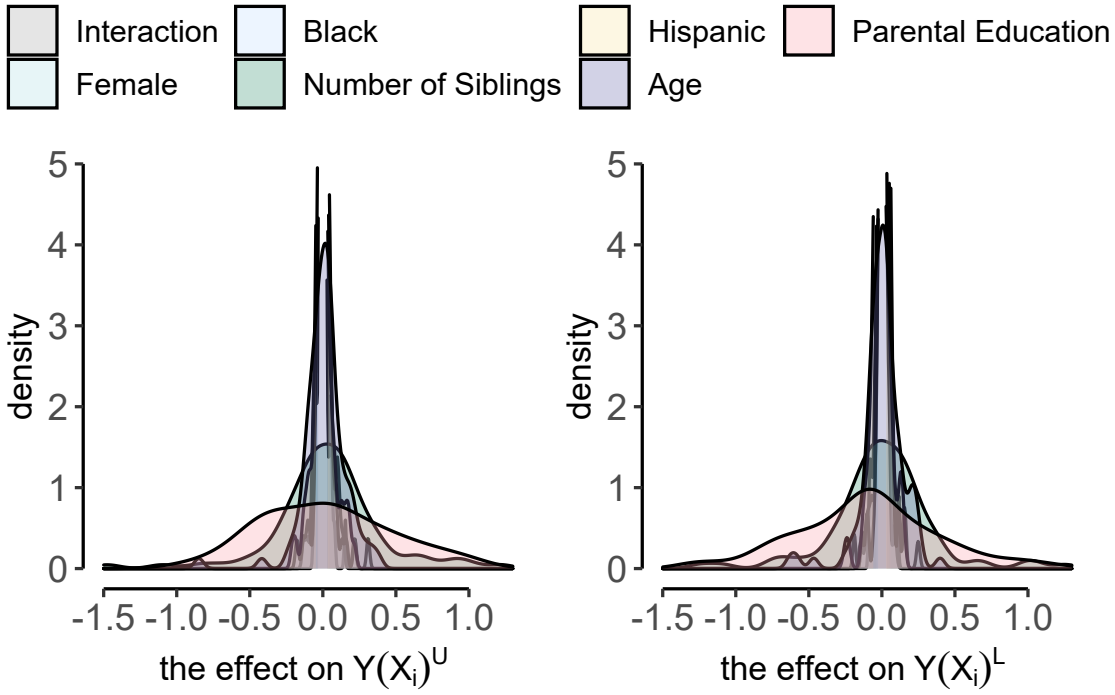


Figure 3. The density of the effect of each independent variable on the upper and lower bound of the outcome, years of schooling. The number of experiments equals 100.

Additionally, Figure 9 (Appendix 12.14) illustrates a single randomly chosen tree in a forest. The bounds of the treatment effects are uninformative. However, the tree successfully recovers parental education and age. Tables IV and V in Appendix 12.14 show that the proposed loss function for multivariate random forests is associated with the highest coverage, informativeness, and heterogeneity of the bounds. Lastly, Table III in Appendix 12.14 shows that the estimated treatment effect bounds capture statistically significant heterogeneity and nonlinearity inherent to the covariate space.

8.2. Additional Simulation Designs

Additional simulation designs closely follow the ones proposed by Farbmacher et al. (2020), Wang et al. (2021) and Athey and Imbens (2016). We test the algorithm for seven different simulated experiments. In each setup, treatment effects are heterogeneous across two variables, and the other three variables are redundant. Each design is based on the following

structural equation setup:

$$
\begin{aligned}
D_i &= \mathbf{1}(0.5 + X_i \cdot \beta + Z_i \cdot 0.2 + \varepsilon_{i1} > 0), \\
Y_i &= 0.3 + X_i \cdot \beta + D_i \cdot \theta_i + \varepsilon_{i2}, \\
\varepsilon_{i1} &\sim \mathcal{N}(0, 1), \\
\varepsilon_{i2} &\sim 0.5 \cdot \varepsilon_{i1} + 0.5 \cdot \mathcal{N}(0, 1), \\
X_i &\sim N(0, I_5), \\
\beta &= [0.01, 0.1, 0.1, 0.05, 0.01],
\end{aligned}
$$

where I_5 is a 5×5 identity matrix. Z_i is a random variable and X_i is a 5×1 vector of covariates.

Table II. The coverage and mean squared error of multivariate random forests for seven different simulated experiments. N is the number of observations and θ_i denotes the personalized treatment effect.

	N	3000	5000	3000	5000
Design	θ_i	Coverage		MSE	
1	0.3	1.000	1.000	0.796	0.811
2	$0.3 \cdot (X_1 + X_2)$	0.863	0.862	0.665	0.626
3	$0.3 \cdot (X_1 X_2)$	0.883	0.916	0.328	0.342
4	$0.3 \cdot \big(X_1 \cdot \mathbb{1}(X_1 > 0) +$ $+ X_2 \cdot \mathbb{1}(X_2 > 0)\big)$	0.972	0.969	0.659	0.680
5	$0.3 \cdot \big(\mathbb{1}(X_1 > 0) + \mathbb{1}(X_2 > 0)\big)$	1.000	1.000	0.516	0.516
6	Same as 2, but no intercept	0.947	0.950	0.808	0.779
7	$0.3 \cdot (X_1^2 X_2^2)$	0.880	0.891	1.076	1.161

Table II summarizes the results for different values of the treatment effects, θ_i. In each case, the multivariate random forest recovers known parameters with a coverage strictly greater than 85%.

9. Head Start Effect on Years of Schooling

In this section, we quantify the bounds of the effect of Head Start participation on years of schooling, and investigate plausible heterogeneity captured within these bounds. Figure 4 shows that the density of the upper bound spans from negative to positive values. The

dashed vertical lines in the figure represent the unconditional, non-parametric bounds of the treatment effects. These unconditional bounds provide a baseline for comparison to the conditional bounds, which take into account the values of the independent variables. As shown in the figure, the conditional bounds are often more informative than the unconditional bounds for certain subsets of the population [3].

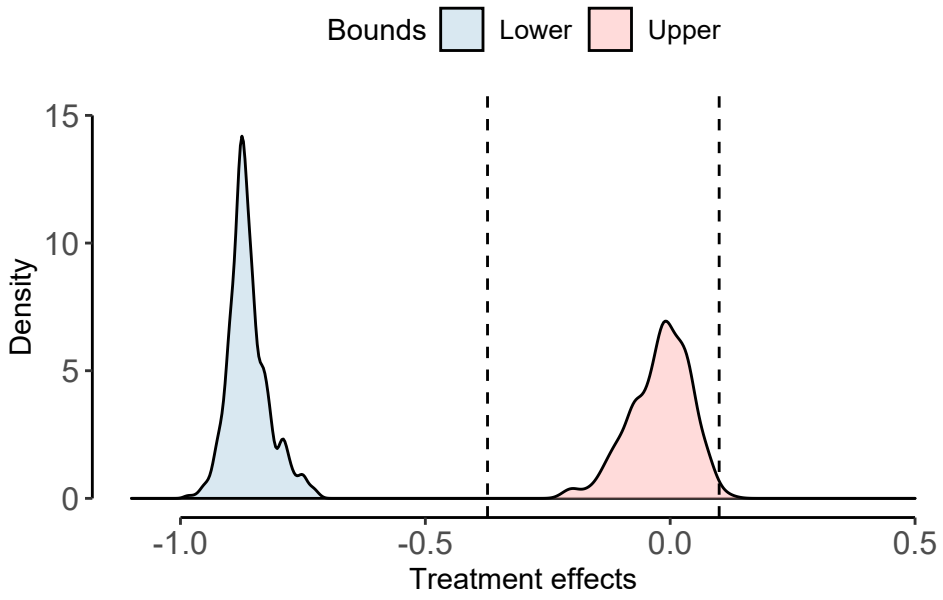


Figure 4. Test data distributions of the bounds of the Head Start effect on years of schooling. The top and bottom quantiles of the outcome variable equal $\alpha_1 = 0.6$ and $\alpha_0 = 0.3$, respectively. Dashed lines represent unconditional treatment effect bounds based on the same values of the outcome quantiles.

We explore the relationship between the bounds of the treatment effects and different family background characteristics of the program participants. Figure 5 displays the bounds of the treatment effect, averaged across different standardized values of parental education, the number of siblings, and age. The upper left panel of the figure shows that certain participants with low-educated parents can, on average, gain more from the Head Start program than those with highly educated parents. Furthermore, the Head Start effect on years of schooling does not exhibit significant heterogeneity across age levels, and the positive effect of the program is not ruled out for participants with a high number of siblings.

[3]Tightness of the bounds depends on the quantiles of the outcome variable. Suggested quantiles are chosen to maximize the variance of the bounds and minimize the loss function simultaneously. The provided grid of the upper quantiles ranges between [0.5 and 1), and the grid of the lower quantiles is between (0, 0.4].

Figure 6 illustrates the relationship between the bounds of the Head Start effect on years of schooling and parental education and the number of siblings for female and male participants. As shown in the figure, the effect of the program is negative for both female and male participants with highly educated parents (top left panel). However, the positive effect of the program is not ruled out for participants with low-educated parents (Hispanics and blacks exhibit the same pattern).

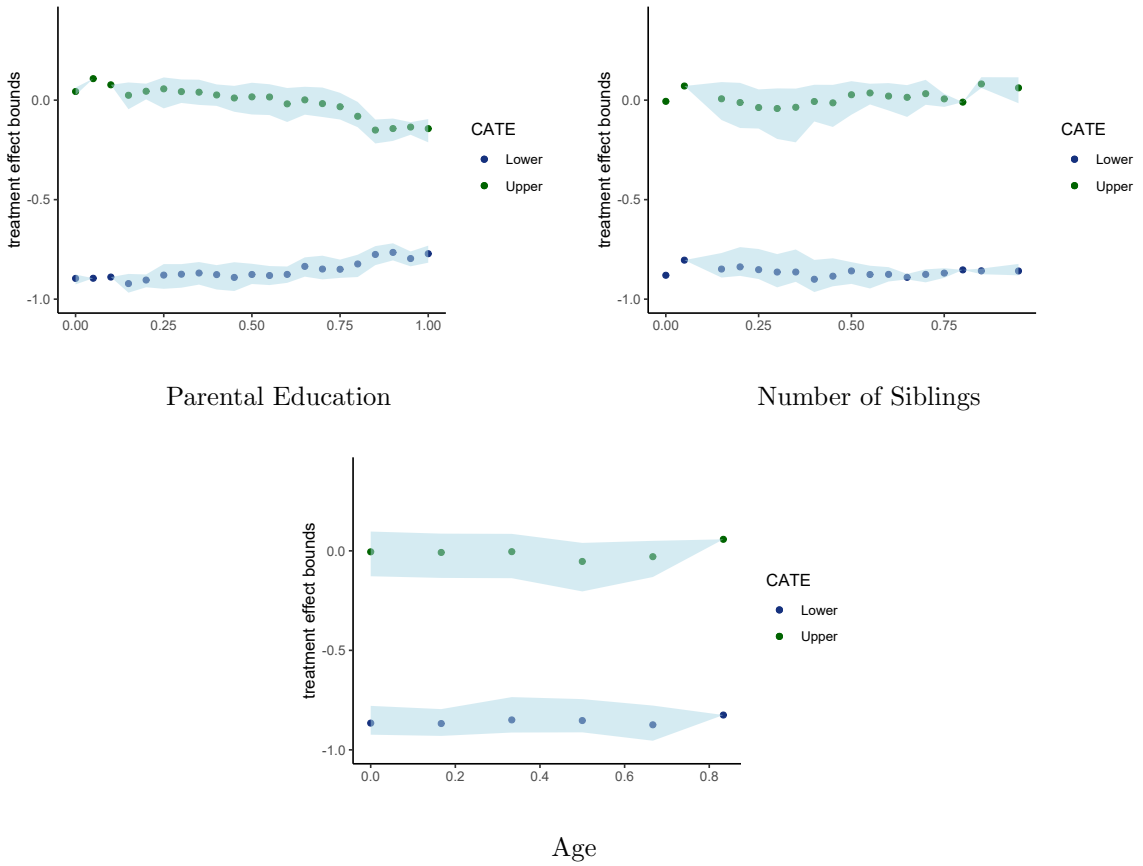


Figure 5. The average bounds of the Head Start participation effect on earnings for unique values parental education, number of siblings, and age. These results are based on predictions in the test data with 932 participants. The blue shaded area covers 0.025 and 0.975 quantiles of the treatment effect bounds.

The relationship between the bounds of the Head Start effect on years of schooling and the number of siblings is reversed for female and male participants. Specifically, female participants with many siblings can potentially gain more from the program than those with a single or no siblings, while male participants do not show significantly different patterns in terms of the effect of the program on years of schooling (bottom right panel of Figure 6).

In addition, the results in Table VI in Appendix 12.15 indicate that, on average, Head

Start participation has a significantly negative effect on the upper and lower bounds of the outcome. This suggests that, on average, individuals who participate in Head Start may have lower potential outcomes compared to those who do not participate. This could be due to negative selection into the treatment group.

The results in Figures 6 and 5 are consistent with the findings of De Haan and Leuven (2020). The authors investigate the effect of Head Start on wage income and education by parental education, gender, and race. De Haan and Leuven (2020) find that the lower bound of the effect of Head Start on years of schooling is smaller for women than for men, and that individuals with disadvantaged parental backgrounds tend to benefit the most from the program. These results support the conclusion that Head Start is effective at improving outcomes for those who are most in need of the program.

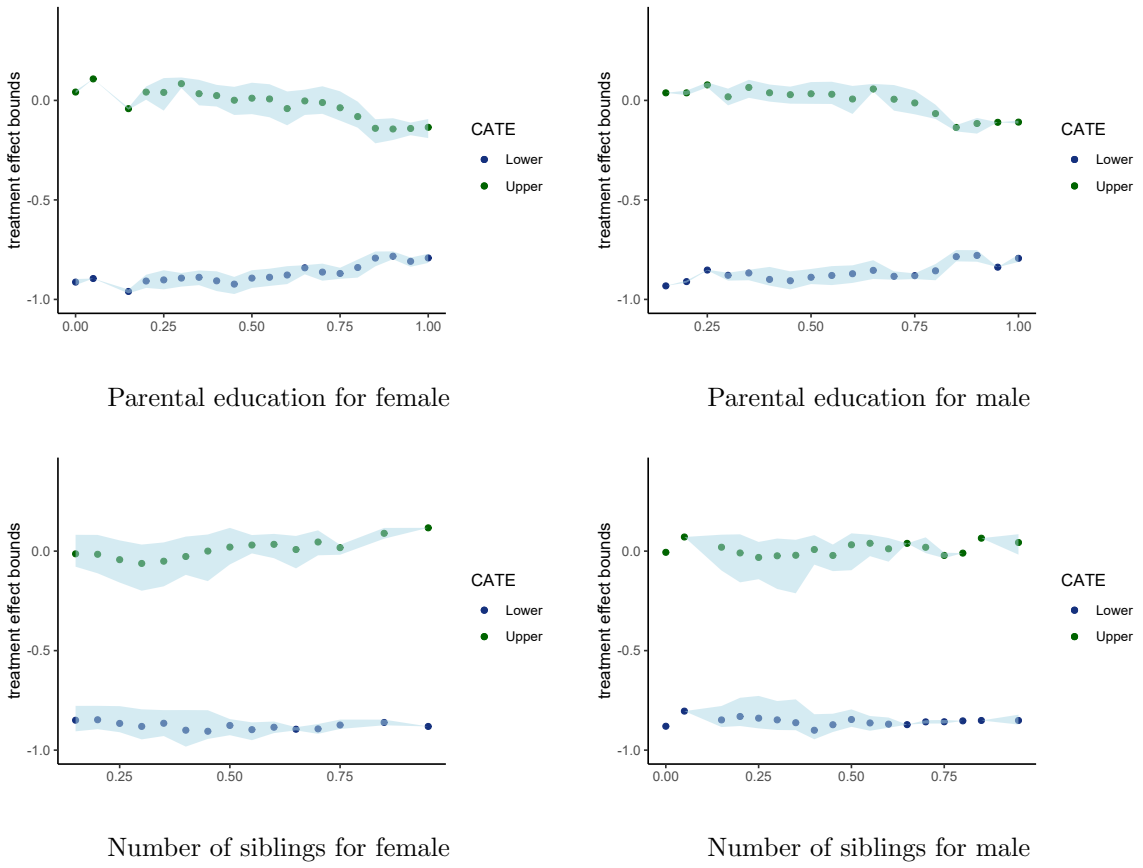


Figure 6. The bounds of the Head Start effect on years of schooling, averaged across different (standardized) levels of parental education for female (upper left), for male (upper right), and across the standardized number of siblings for female (lower left), and male (lower right). The dashed area represents the area between the 0.025 and 0.975 quantiles.

10. Conclusion and Future Directions

This paper introduces multivariate random forests for estimating and inferring heterogeneity in the bounds of treatment effects. The method explores and detects subsets of a population with the highest variation in the treatment effect bounds. Under several regularity conditions and assumptions about the data generating process, the method is consistent and the estimated parameters are asymptotically normally distributed. These results also extend to products of different parameters and quantiles of the outcome. Our findings demonstrate that the nonlinearities and heterogeneities captured by the treatment effect bounds can inform personalized policy analysis and decision-making.

One contribution of this paper is to introduce flexibility in the direction of the treatment selection. This allows a practitioner to avoid predetermined unidirectional sign of the treatment selection, which can often be difficult to justify for chosen strata. Another contribution of this article is the investigation of the large sample theory of multivariate random forests. Asymptotic guarantees allow us to infer multiple parameters of interest, including the treatment effect bounds. Supplementary material covers additional details, extensions, and generalizations of the method to settings beyond conventional nonparametric bounds.

To illustrate the properties of our method, we apply it to the effect of the Head Start preschool program on years of schooling (De Haan and Leuven, 2020). Our results are consistent with the findings of De Haan and Leuven (2020). We show that, on average, the program has a negative effect on schooling for participants with highly educated parents and a low number of siblings. In contrast, we do not reject the hypothesis of a positive effect for participants with a disadvantaged family background. These findings suggest that individuals who are most in need of the program are more likely to benefit from it. We also conduct Monte Carlo simulations to demonstrate the desirable finite sample performance of our proposed method.

This article focuses on partially identified treatment effects. A useful extension of the method is to design multivariate forests for conditional marginal treatment effects (Athey et al., 2019). Multivariate random forests can be generalized to network structures in data (Li, 2020) where the coefficients correlate across leaves. Moreover, in the future, we plan to empirically investigate partial identification with anomalous data. Xiong et al. (2021) show consistent identification of the treatment effects from multiple sources of data, weighted by Hessian, sample size, and inverse variance. Their work has implications for anomaly detec-

tion, as we can imagine anomalous and representative instances as two different sources of data. By comparison, Yadlowsky et al. (2021) introduce a general score-weighted treatment effect, where the score depends on the covariates. To separate anomalous covariates from the representative ones, a useful extension of this paper is to introduce the anomaly detection score (Liu et al., 2008). Specifically, in fully grown trees, Liu et al. (2008) measure the length of the branches in each tree. According to the authors, the shortest path length has a high probability of being an anomaly. This separation will enable us to investigate differences between the bounds of treatment effects that reflect anomalous covariates and the ones based on the non-anomalous covariate values. Lastly, the performance of the method can be increased by a deep structure of random forests. Nareklishvili et al. (2022) offer a deep partial least squares method to extract relevant features among many independent regressors. This method can be generalized to multivariate random forests when the covariates are normally distributed.

References

Athey, S. and Imbens, G. (2016). Recursive partitioning for heterogeneous causal effects. *Proceedings of the National Academy of Sciences*, 113(27):7353–7360.

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

Athey, S. and Wager, S. (2019). Estimating treatment effects with causal forests: An application. *Observational Studies*, 5(2):37–51.

Balke, A. and Pearl, J. (1997). Bounds on treatment effects from studies with imperfect compliance. *Journal of the American Statistical Association*, 92(439):1171–1176.

Belloni, A., Chernozhukov, V., Fernández-Val, I., and Hansen, C. (2017). Program evaluation and causal inference with high-dimensional data. *Econometrica*, 85(1):233–298.

Beresteanu, A. and Manski, C. F. (2000). Bounds for stata: draft version 1.0. *manuscript, Northwestern University*.

Beresteanu, A., Molchanov, I., and Molinari, F. (2011). Sharp identification regions in models with convex moment predictions. *Econometrica*, 79(6):1785–1821.

Biau, G. (2012). Analysis of a random forests model. *The Journal of Machine Learning Research*, 13(1):1063–1095.

Billingsley, P. (2008). *Probability and measure*. John Wiley & Sons.

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

Breiman, L. (2004). Consistency for a simple model of random forests.

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., and Newey, W. (2017). Double/debiased/neyman machine learning of treatment effects. *American Economic Review*, 107(5):261–65.

Chernozhukov, V., Demirer, M., Duflo, E., and Fernandez-Val, I. (2018). Generic machine learning inference on heterogeneous treatment effects in randomized experiments, with an application to immunization in india. Technical report, National Bureau of Economic Research.

Chernozhukov, V. and Hansen, C. (2006). Instrumental quantile regression inference for structural and treatment effect models. *Journal of Econometrics*, 132(2):491–525.

Chernozhukov, V., Hong, H., and Tamer, E. (2007). Estimation and confidence regions for parameter sets in econometric models 1. *Econometrica*, 75(5):1243–1284.

Creswell, A., White, T., Dumoulin, V., Arulkumaran, K., Sengupta, B., and Bharath, A. A. (2018). Generative adversarial networks: An overview. *IEEE signal processing magazine*, 35(1):53–65.

De Haan, M. (2011). The effect of parents' schooling on child's schooling: a nonparametric bounds analysis. *Journal of Labor Economics*, 29(4):859–892.

De Haan, M. (2017). The effect of additional funds for low-ability pupils: A non-parametric bounds analysis. *The Economic Journal*, 127(599):177–198.

De Haan, M. and Leuven, E. (2020). Head start and the distribution of long-term education and labor market outcomes. *Journal of Labor Economics*, 38(3):000–000.

Denil, M., Matheson, D., and De Freitas, N. (2014). Narrowing the gap: Random forests in theory and in practice. In *International conference on machine learning*, pages 665–673. PMLR.

DiCiccio, C. and Romano, J. P. (2020). Clt for u-statistics with growing dimension. Technical report, Technical report, Standford University.

Fan, Y. and Park, S. S. (2010). Sharp bounds on the distribution of treatment effects and their statistical inference. *Econometric Theory*, 26(3):931–951.

Farbmacher, H., Guber, R., and Klaassen, S. (2020). Instrument validity tests with causal forests. *Journal of Business & Economic Statistics*, pages 1–10.

Goel, A., Khanna, S., Raghvendra, S., and Zhang, H. (2014). Connectivity in random forests and credit networks. In *Proceedings of the twenty-sixth annual ACM-SIAM symposium on Discrete algorithms*, pages 2037–2048. SIAM.

Hájek, J. (1968). Asymptotic normality of simple linear rank statistics under alternatives. *The Annals of Mathematical Statistics*, pages 325–346.

Hao, L., Naiman, D. Q., and Naiman, D. Q. (2007). *Quantile regression*. Number 149. Sage.

Heckman, J. J. (1979). Sample selection bias as a specification error. *Econometrica: Journal of the econometric society*, pages 153–161.

Heckman, J. J. (2000). Causal parameters and policy analysis in economics: A twentieth century retrospective. *The Quarterly Journal of Economics*, 115(1):45–97.

Heckman, J. J. and Vytlacil, E. (2001). Policy-relevant treatment effects. *American Economic Review*, 91(2):107–111.

Heckman, J. J. and Vytlacil, E. J. (1999). Local instrumental variables and latent variable models for identifying and bounding treatment effects. *Proceedings of the national Academy of Sciences*, 96(8):4730–4734.

Heiler, P. (2022). Estimating heterogeneous bounds for treatment effects under sample selection and non-response. *arXiv preprint arXiv:2209.04329*.

Hoeffding, W. (1961). The strong law of large numbers for u-statistics. Technical report, North Carolina State University. Dept. of Statistics.

Horowitz, J. L. and Manski, C. F. (2000). Nonparametric analysis of randomized experiments with missing covariate and outcome data. *Journal of the American Statistical Association*, 95(449):77–84.

Horvitz, D. G. and Thompson, D. J. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American statistical Association*, 47(260):663–685.

Huber, M. and Mellace, G. (2015). Sharp bounds on causal effects under sample selection. *Oxford bulletin of economics and statistics*, 77(1):129–151.

Imbens, G. W. and Manski, C. F. (2004). Confidence intervals for partially identified parameters. *Econometrica*, 72(6):1845–1857.

Jennrich, R. I. (2008). Nonparametric estimation of standard errors in covariance analysis using the infinitesimal jackknife. *Psychometrika*, 73(4):579–594.

Jiang, Z., Chiba, Y., and VanderWeele, T. J. (2014). Monotone confounding, monotone treatment selection and monotone treatment response. *Journal of causal inference*, 2(1):1–12.

Kingsford, C. and Salzberg, S. L. (2008). What are decision trees? *Nature biotechnology*, 26(9):1011–1013.

Koenker, R. and Hallock, K. F. (2001). Quantile regression. *Journal of economic perspectives*, 15(4):143–156.

Kong, Y. and Yu, T. (2018). A deep neural network model using random forest to extract feature representation for gene expression data classification. *Scientific reports*, 8(1):1–9.

Korolyuk, V. S. and Borovskich, Y. V. (2013). *Theory of U-statistics*, volume 273. Springer Science & Business Media.

Lechner, M. (1999). Nonparametric bounds on employment and income effects of continuous vocational training in east germany. *The Econometrics Journal*, 2(1):1–28.

Lee, D. S. (2009). Training, wages, and sample selection: Estimating sharp bounds on treatment effects. *The Review of Economic Studies*, 76(3):1071–1102.

Lee, M.-J. (2005). *Micro-econometrics for policy, program and treatment effects*. OUP Oxford.

Lewis, R. J. (2000). An introduction to classification and regression tree (cart) analysis. In *Annual meeting of the society for academic emergency medicine in San Francisco, California*, volume 14. Citeseer.

Li, K. (2020). Asymptotic normality for multivariate random forest estimators. *arXiv preprint arXiv:2012.03486*.

Lin, Y. and Jeon, Y. (2006). Random forests and adaptive nearest neighbors. *Journal of the American Statistical Association*, 101(474):578–590.

Liu, F. T., Ting, K. M., and Zhou, Z.-H. (2008). Isolation forest. In *2008 eighth ieee international conference on data mining*, pages 413–422. IEEE.

Liu, Q., Xu, J., Jiang, R., and Wong, W. H. (2021). Density estimation using deep generative neural networks. *Proceedings of the National Academy of Sciences*, 118(15):e2101344118.

Manski, C. and Pepper, J. (2000). Monotone treatment response, with an application to the returns to schooling. *Econometrica*, 68:997–1012.

Manski, C. F. (1990). Nonparametric bounds on treatment effects. *The American Economic Review*, 80(2):319–323.

Manski, C. F. and Pepper, J. V. (2009). More on monotone instrumental variables. *The Econometrics Journal*, 12:S200–S216.

Meinshausen, N. and Ridgeway, G. (2006). Quantile regression forests. *Journal of Machine Learning Research*, 7(6).

Mogstad, M. and Torgovitsky, A. (2018). Identification and extrapolation of causal effects with instrumental variables. *Annual Review of Economics*, 10:577–613.

Molinari, F. (2020). Microeconometrics with partial identification. *Handbook of econometrics*, 7:355–486.

Nareklishvili, M., Polson, N., and Sokolov, V. (2022). Deep partial least squares for iv regression. *arXiv preprint arXiv:2207.02612*.

Nekipelov, D., Novosad, P., and Ryan, S. P. (2018). Moment forests.

Peccati, G. (2004). Hoeffding-anova decompositions for symmetric statistics of exchangeable observations. *The Annals of Probability*, 32(3):1796–1829.

Romano, J. P. and Shaikh, A. M. (2010). Inference for the identified set in partially identified econometric models. *Econometrica*, 78(1):169–211.

Rothstein, D. S., Carr, D., and Cooksey, E. (2019). Cohort profile: the national longitudinal survey of youth 1979 (nlsy79). *International Journal of Epidemiology*, 48(1):22–22e.

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

Rubin, D. B. (2006). *Matched sampling for causal effects*. Cambridge University Press.

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

Segal, M. and Xiao, Y. (2011). Multivariate random forests. *Wiley interdisciplinary reviews: Data mining and knowledge discovery*, 1(1):80–87.

Semenova, V. (2020). Better lee bounds. *arXiv preprint arXiv:2008.12720*.

Shaikh, A. M. and Vytlacil, E. J. (2011). Partial identification in triangular systems of equations with binary dependent variables. *Econometrica*, 79(3):949–955.

Stoye, J. (2009). More on confidence intervals for partially identified parameters. *Econometrica*, 77(4):1299–1315.

Torgovitsky, A. (2019). Partial identification by extending subdistributions. *Quantitative Economics*, 10(1):105–144.

Van der Maaten, L. and Hinton, G. (2008). Visualizing data using t-sne. *Journal of machine learning research*, 9(11).

Van der Vaart, A. (1998). Cambridge series in statistical and probabilistic mathematics. *Asymptotics Statistics*.

Wager, S. (2014). Asymptotic theory for random forests. *arXiv preprint arXiv:1405.0352*.

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.

Wager, S. and Walther, G. (2015). Adaptive concentration of regression trees, with application to random forests. *arXiv preprint arXiv:1503.06388*.

Wang, G., Li, J., and Hopp, W. J. (2021). An instrumental variable forest approach for detecting heterogeneous treatment effects in observational studies. *Management Science*.

Xiong, R., Koenecke, A., Powell, M., Shen, Z., Vogelstein, J. T., and Athey, S. (2021). Federated causal inference in heterogeneous observational data. *arXiv preprint arXiv:2107.11732*.

Yadlowsky, S., Fleming, S., Shah, N., Brunskill, E., and Wager, S. (2021). Evaluating treatment prioritization rules via rank-weighted average treatment effects. *arXiv preprint arXiv:2111.07966*.

11. Notation

11.1. Partial Identification of Treatment Effects

1 $\{Y_i, D_i, X_i\}_{i=1}^N$ - Data: outcome, treatment, covariates for each individual i.

2 d, $Y_i(d) \in \mathbb{R}$ - Potential treatment and outcome for $d \in \{0, 1\}$.

3 ℓ, ℓ_n - Any partition (node), and a terminal node, correspondingly.

4 $\mathcal{X}$ - Feature space.

5 $n = 1 \dots |\Pi|$ - Number of terminal nodes (leafs) in a partition.

6 $\Pi = \{\ell_1, \ell_2, \dots, \ell_{|\Pi|}\}$ - Partition of the feature space where $\cup_{n=1}^{|\Pi|} \ell_n = \mathcal{X}$.

7 $\ell(x, \Pi)$ - Subgroup (leaf) such that $\ell \in \Pi$ and $x \in \ell$.

8 $\theta_0(X_i)$, $\theta(X_i)$, ε_i - Individual-level intercept, ATE, error, respectively, when Z_i is not available.

9 $\theta_0(X_i, \Pi)$, $\theta(X_i, \Pi)$ - Subgroup-level intercept, treatment effects, the means of $\theta_0(X_i)$ and $\theta(X_i)$ in a given subset, respectively. E.g., $\theta(X_i, \Pi) = \mathbb{E}(\theta(X_i)|X_i \in \ell(x, \Pi))$.

10 $[Y^L(X_i), Y^U(X_i)]$ - Conditional lower and upper bounds of Y_i.

11 $\theta^B(X_i)$, $\theta^B(X_i, \Pi)$, v_i^B - Individual-level bound of the treatment effect, the subgroup-level bound of the treatment effect, error for $B \in \{L, U\}$.

12 $|i : X_i \in \ell_n, D_i = a|$ - The number of observations of the treated (a = 1) and untreated (a = 0) groups in a leaf ℓ_n.

11.2. Trees and Multivariate Random Forests

1 S^{tr}, S^{est}, S^{te} - Train, estimation and test data samples, respectively.

2 N^{tr}, N^{est}, N^{te} - Number of observations in train, estimation and test data samples, respectively.

3 $\mu(x)$ - Mean of the outcome Y_i at a test point x.

4 $\mu(X_i, \Pi)$, $\widetilde{\mu}(X_i, S^{est}, \Pi)$ - Subsample-level mean of the outcome, and the corresponding estimator from the estimation sample.

5 $\mathbb{V}\big(\widetilde{\mu}(X_i, S^{est}, \Pi)\big)$ - Variance of $\widetilde{\mu}(X_i, S^{est}, \Pi)$.

6 $S^2_{S^{tr}}(X_i, \Pi)$ - Estimator of the variance of $\widetilde{\mu}(X_i, S^{est}, \Pi)$ based on train data.

7 $s = (j, c)$ - A pair of splitting covariate, value.

8 $\mathcal{P}^{(0)}$, $\mathcal{P}^{(1,1)}$, $\mathcal{P}^{(1,2)}$ - Parent node of the tree, two subsequent nodes (children) in a tree, respectively.

9 $\widetilde{x}_j$ - $j-$th corrdinate of the vector $\widetilde{x}$ from the train data.

10 μ_L, μ_R - Means of the outcome in the left and right nodes, respectively.

11 $\mathcal{T}(x, \xi, A_1, \ldots, A_N)$ - Prediction of a tree at a test point x, for a randomization parameter ξ, and data $A_i = \{Y_i, X_i\}_{i=1}^N$.

12 $s = N^\beta$ $(\beta > 0)$ - Subset size from a population of size N.

13 $\mathcal{F}(x, \xi, A_1, \ldots, A_N)$ - prediction of a forest.

14 $\widetilde{\theta}(X_i, S^{tr}, \Pi)$, $\widetilde{\theta}(X_i, S^{est}, \Pi)$ - Estimator of the treatment effect based on train and estimation samples, respectively.

15 $S^2_{s^{tr}_{treat}}$, $S^2_{s^{tr}_{control}}$ - Estimators of the variances of the mean outcomes in the treated and control groups.

16 p - The share of treated observations in the train data.

17 $\mathcal{CT}(x, \xi, A_1, \ldots, A_N)$ - Prediction of a causal tree.

18 $Y_i^{(1)}$, $Y_i^{(0)}$ - Observed outcomes for the treated and untreated groups.

19 $\mathcal{F}(Y_{im}|X_i, s) = \mathcal{F}(Y_{im}|X_i)$ - Density of the $m-$th outcome variable, where $m = 1 \ldots, M$.

20 $\mu_L(X_i)$, $\mu_R(X_i)$ - Centers of the mass of X_i in the left and right nodes.

21 $0 < \pi \leq 1$ - Probability that the $j-$th coordinate is selected for splitting.

22 α, k - fraction of the available training observations, and minimum number of observations in each leaf, respectively.

23 $\widetilde{T} = [\widetilde{T}_1, \widetilde{T}_2, \ldots, \widetilde{T}_M]$ - Prediction of a multivariate tree at a test point x.

24 $i_1, \ldots, i_s$ - $\frac{n!}{s!(n-s)!}$ subsets of size s from $1, \ldots, N$.

25 $\mathcal{MF} = \mathcal{MF}(x, A_1, \ldots, A_N)$ - $1 \times M$ vector of predicted outcomes of a multivariate random forest.

26 $T(\pi x) = T(x)$ - A permutation symmetric function.

27 $g_i : \mathbb{R}^d \mapsto \mathbb{R}$ - An arbitrary measurable function.

28 $\dot{T}$ - A Hajek projection of T.

29 μ - $1 \times M$ vector of means of $\mathcal{MF}$.

30 $\dot{\mathcal{MF}} = \dot{\mathcal{MF}}(x, A_1, \ldots, A_N)$ - A Hajek projection of multivariate random forests.

31 Σ, $\widehat{\Sigma}$ - Covariance matrix of $\dot{\mathcal{MF}}$ and the corresponding sample analogue.

32 $\mathbb{V}(\dot{\widetilde{T}})$, $\mathbb{V}(\widetilde{T})$ - Covariance of the Hajek projection of a multivariate tree, and of a multivariate tree, respectively.

33 θ^ℓ - Parameter in an $\ell-$th partition.

34 $\rho(A_i, \theta^\ell)$ - Unconditional moment function.

35 $m(x, \theta) = \mathbb{E}\big(\rho(A_i, \theta^\ell)|X_i = x\big)$ - Conditional moment function.

36 $M_\ell(x,\theta)$, $\widehat{M}_\ell(x,\theta)$ - Population conditional moment function, and the sample analogue, respectively.

37 $\Omega(X_i)$, $\kappa(X_i)$ - Unknown positive definite matrix and the corresponding sample analogue.

38 $\theta^\star(X_i, S^{est}, \Pi)$ - Optimal population parameter in the estimation sample.

11.3. A Test for Heterogeneity

1 $Y_0^B(X_i)$, $b_0^B(X_i)$, $s_0^B(X_i)$ - Unknown bounds of the outcome, of intercept, and of the treatment effect.

2 $Y^B(X_i)$, $b(X_i)$, $s(X_i)$ - Proxies of $Y_0^B(X_i)$, $b_0^B(X_i)$, $s_0^B(X_i)$, respectively.

3 U_i^B - Errors of the outcome bounds for $B \in \{L, U\}$.

4 $p = \mathbb{P}$ - Probability.

5 $p(X_i)$ - Propensity score.

6 $D_i - p(X_i)$ - Demeaned treatment.

7 $(D_i - p(X_i))(s^B - \mathbb{E}s^B)$ - Interaction^B.

8 β_1^B, β_2^B - The effects of the demeaned treatment and an interaction on outcome.

9 α^B - Nuisance parameter in the equation of outcome bounds.

10 $\widehat{\beta}_1^B$, $\widehat{\beta}_2^B$ $\widehat{\alpha}^B$ - Estimates of β_1^B, β_2^B, α^B, respectively.

11 $w(X_i) = [p(X_i)(w - p(X_i)]^{-1}$ - Weight in the best linear predictor of $s_0^B(X_i)$.

12 $H_i = H_i(D_i, X_i)$ - Horvitz-Thompson Transformation.

13 β_0^B, β^B - The intercept and the effect of D_i on $Y^B(X_i)$, respectively.

14 $g = 1, \ldots, G$ - $g-$th bootstrapped sample.

15 $\widehat{\theta}_g^B(x, S_g^{est}, \Pi_g)$ - Estimate of $\theta^B(x, S^{est}, \Pi)$ for the $g-$th bootstrapped sample.

16 N_{ig} - The number of times an observation i appears in the $g-$th sample (train and estimation).

17 $\widehat{\theta}^B\big(x, \{S_g^{est}\}_{g=1}^G, \{\Pi_g\}_{g=1}^G\big)$ - Mean of $\widehat{\theta}_g^B(x, S_g^{est}, \Pi_g)$ across all bootstrapped samples G.

12. Appendix

12.1. Additional Example

When the dimension of the covariate space is small, a parametric regression model may be used for heterogeneous treatment effect analysis. To illustrate this, consider the treatment D_i is higher education and Y_i is earnings. Assume, $X_i = \{X_{i1}, X_{i2}, X_{i3}\} \in \{0, 1\}$. Since we do not know which characteristics and their interactions induce heterogeneity in the

treatment effects, we include each of them in the regression equation:

$$\begin{aligned} Y_i = {} & \alpha_0 + \alpha_1 X_{i1} + \alpha_2 X_{i2} + \alpha_3 X_{i3} + \alpha_4 X_{i1} X_{i2} + \alpha_5 X_{i1} X_{i3} \\ & + \alpha_6 X_{i2} X_{i3} + \alpha_7 X_{i1} X_{i2} X_{i3} + \\ & \beta_0 D_i + \beta_1 X_{i1} D_i + \beta_2 X_{i2} D_i + \beta_3 X_{i3} D_i + \beta_4 X_{i1} X_{i2} D_i + \\ & \beta_5 X_{i1} X_{i3} D_i + \beta_6 X_{i2} X_{i3} D_i + \beta_7 X_{i1} X_{i2} X_{i3} D_i. \end{aligned}$$

Treatment effects may differ across the subgroups of the population. For example, consider two different subsets of the covariate space:

$$\begin{aligned} &\text{Subgroup 1} : \{X_{i1} = 1, X_{i2} = 1, X_{i3} = 1\}; \\ &\text{Subgroup 2} : \{X_{i1} = 1, X_{i2} = 1, X_{i3} = 0\}. \end{aligned}$$

We can test whether the coefficients stemming from Subgroup 1 differ from the ones in Subgroup 2:

$$\beta_0 + \beta_1 + \beta_2 + \beta_3 + \beta_4 + \beta_5 + \beta_6 + \beta_7 \neq \beta_0 + \beta_1 + \beta_2 + \beta_3 + \beta_4.$$

This approach inherits several major challenges. First, the number of terms in the regression equation increases exponentially with the dimension of the feature vector. With P binary regressors, we end up with 2^P unique combinations that could influence ATE. Hence, the estimation of such a model is often infeasible. Second, the regression approach does not specifically identify subgroups that induce heterogeneous treatment effects; it is unclear which subsets induce heterogeneous treatment effects and which observations belong to the specific subgroups. Lastly, the effects of different regressors are linearly additive in the setup. In practice, they may depend on the outcome in a more complex nonlinear way.

To shed the light on the problem, we simulate a high-dimensional covariate space with a network structure. The goal of the simulation is to identify whether the random forest method (Breiman, 2001) detects the relevant dimensions of the feature space that inherit heterogeneous treatment effects. Each node in the network represents one feature. Nodes are connected to each other, forming a network. The network follows a scale-free power-law degree distribution. That means, only a few features in the network have a high probability of having a large number of "neighbours". The distance between two features is defined by the shortest path in the network. For more details, see Kong and Yu (2018).

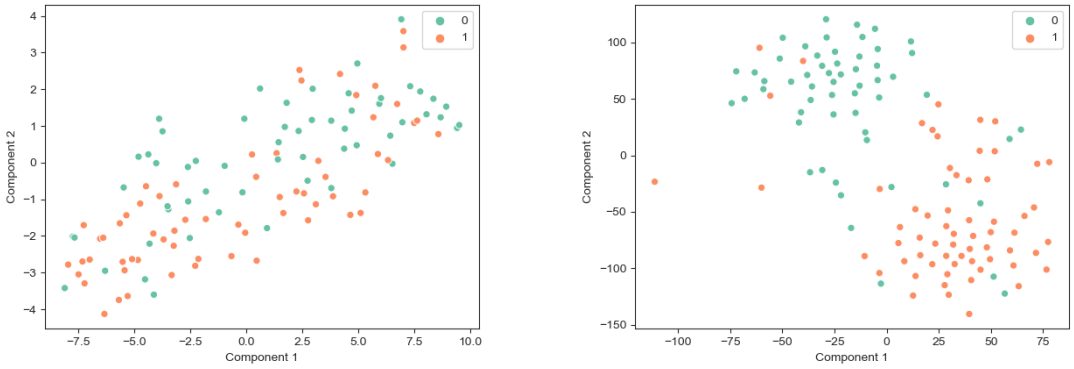
First, we consider a 600×600 distance matrix D consisting of pair-wise distances among predictors. The distance between a predictor and itself is zero, hence diagonal elements of D are equal to zero. Then the following rule transforms the distance matrix into a covariance matrix:

$$\Sigma_{i,j} = 0.7^{D_{i,j}}, \ \ i, j = 1, ..., 600.$$

Next, we generate a 600×600 feature matrix from a normal distribution $\mathbf{X} \sim \mathcal{N}(0, \Sigma)$ and the treatment effects from a simple logistic function:

$$\theta(X_i) = \mathbb{1}\left(\frac{\exp(\mathbf{X_s}\boldsymbol{\beta})}{1 + \exp(\mathbf{X_s}\boldsymbol{\beta})} > 0.5\right),$$

where $\mathbf{X}_s$ denotes a matrix with 600 observations and 6 randomly selected predictors from a set $\mathbf{X}$. $\beta \sim N(0, \Sigma_\beta)$ is a 6×1 vector of the effect of $\mathbf{X}_s$ on $\theta(X_i)$ (Σ_β and Σ denote covariance matrices of the parameters β and features $\mathbf{X}$, respectively).



Original Random Forests

Figure 7. Feature representation of random forests based on the T-SNE algorithm. The left plot represents the separability between the two classes of treatment effects based on the originally simulated covariates. The right plot illustrates the separability of two classes of treatment effects based on individual decision tree predictions. Representation of treatment effects is based on test data.

Trained random forests consist of 200 trees. The T-SNE algorithm (Van der Maaten and Hinton, 2008) in Figure 7 represents the two classes of treatment effects as a function of the original features (left), and decision tree predictions (right). The random forest method improves the representation of the features as the classes of treatment effects become more separable and evident.

12.2. Proposition 4.1

Proof. The proof is based on Manski and Pepper (2000) and Manski (1990) under the positive treatment selection. The proof is analogous for a negative treatment selection. Based on the law of iterated expectations, the potential outcomes can be decomposed as

$$\begin{aligned}\mathbb{E}\big(Y_i(1)|X_i=x\big) =& \mathbb{E}\big(Y_i(1)|D_i=1,X_i=x\big)\times P(D_i=1|X_i=x)+ \\ & \mathbb{E}\big(Y_i(1)|D_i=0,X_i=x\big)\times P(D_i=0|X_i=x), \\ \mathbb{E}\big(Y_i(0)|X_i=x\big) =& \mathbb{E}\big(Y_i(0)|D_i=0,X_i=x\big)\times P(D_i=0|X_i=x)+ \\ & \mathbb{E}\big(Y_i(0)|D_i=1,X_i=x\big)\times P(D_i=1|X_i=x).\end{aligned} \tag{12.1}$$

Estimating these potential outcomes is complicated as the individual is observed only in one particular state. We observe the average outcome of observations who received the treatment, $\mathbb{E}(Y_i(1)|D_i=1,X_i=x)$, and of those who did not, $\mathbb{E}(Y_i(0)|D_i=0,X_i=x)$. We also observe the proportion of treated observations, $P(D_i=1|X_i=x)$, and of untreated ones, $P(D_i=0|X_i=x)$. However, for an untreated individual, $Y_i(1)$ (e.g., earnings if she was instead treated) is not observed. For a treated individual, $Y_i(0)$ (e.g., earnings if she was instead untreated) is unknown. Hence, the mean potential outcomes, $\mathbb{E}(Y_i(0)|D_i=1,X_i=x)$ and $\mathbb{E}(Y_i(1)|D_i=0,X_i=x)$, are unknown for treated or untreated groups, respectively.

Assumption 4.1 implies that the unknown quantities lie between $Y^L(X_i)$ and $Y^U(X_i)$. Hence the bounds of the mean potential outcomes become

$$\begin{aligned}&\mathbb{E}\big(Y_i(1)|D_i=1,X_i=x\big)\times P(D_i=1|X_i=x)+Y^L(X_i)\times P(D_i=0|X_i=x) \\ &\qquad\qquad \leq \mathbb{E}\big(Y_i(1)|X_i=x\big) \leq \\ &\mathbb{E}\big(Y_i(1)|D_i=1,X_i=x\big)\times P(D_i=1|X_i=x)+Y^U(X_i)\times P(D_i=0|X_i=x), \\ &\mathbb{E}\big(Y_i(0)|D_i=0,X_i=x\big)\times P(D_i=0|X_i=x)+Y^L(X_i)\times P(D_i=1|X_i=x) \\ &\qquad\qquad \leq \mathbb{E}\big(Y_i(0)|X_i=x\big) \leq \\ &\mathbb{E}\big(Y_i(0)|D_i=0,X_i=x\big)\times P(D_i=0|X_i=x)+Y^U(X_i)\times P(D_i=1|X_i=x).\end{aligned}$$

Additionally, Assumption 4.2 implies that $\mathbb{E}\big(Y_i(1)|D_i=1,X_i=x\big) \geq \mathbb{E}\big(Y_i(1)|D_i=0,X_i=x\big)$ and $\mathbb{E}\big(Y_i(0)|D_i=1,X_i=x\big) \geq \mathbb{E}\big(Y_i(0)|D_i=0,X_i=x\big)$. Therefore, we obtain that the observed outcome for the treated observations, $\mathbb{E}\big(Y_i(1)|D_i=1,X_i=x\big)$, is the upper bound of $\mathbb{E}\big(Y_i(1)|X_i=x\big)$, whereas the observed outcome for the non-treated individuals, $\mathbb{E}\big(Y_i(0)|D_i=0,X_i=x\big)$ is the lower bound of $\mathbb{E}\big(Y_i(0)|X_i=x\big)$. By subtracting

the upper (lower) bound of $\mathbb{E}(Y_i|X_i = x)$ from the lower (upper) bound of $\mathbb{E}(Y_i|X_i = x)$ the proof of Proposition 4.1 is complete.

■

12.3. Self-Selection with a Monotonic Instrumental Variable

Often, a researcher has access to an additional variable (instrument) that can strongly influence the selection into the treatment. Consider a binary monotonic instrumental variable $Z_i \in \{0, 1\}$ which satisfies the following assumption:

Assumption 12.1 (**Monotonic Instrumental Variable**).

$$\mathbb{E}(Y_i(d)|Z_i = 1, X_i = x) \geq \mathbb{E}(Y_i(d)|Z_i = 1, X_i = x), \text{ or}$$
$$\mathbb{E}(Y_i(d)|Z_i = 1, X_i = x) \leq \mathbb{E}(Y_i(d)|Z_i = 1, X_i = x).$$

Additionally, assume, the outcome linearly depends on the treatment D_i, and the instrument has a strong effect on the outcome.

Assumption 12.2 (**Linear Response Model**). $Y_i = \theta_0(X_i) + D_i\theta(X_i) + \varepsilon_i$.

Assumption 12.1 implies that, under any potential treatment state, those who are assigned to the instrument level one have weakly higher or lower mean potential outcomes. De Haan and Leuven (2020) show that this assumption, jointly with the monotonic treatment selection, can tighten the nonparametric bounds. Partial identification of treatment effects does not require Assumption 12.2, however, it can tighten the bounds (Manski and Pepper, 2009).

As shown by Manski and Pepper (2009), under a positive monotonicity of the instrument and Assumption 12.2,

$$\mathbb{E}(Y_i - D_i\theta(x)|Z_i = 1, X_i = x) \geq \mathbb{E}(Y_i - D_i\theta(x)|Z_i = 0, X_i = x).$$

This yields partial identification of treatment effects for each data point x:

$$\begin{aligned}\theta^L(x) = \mathbb{E}\big(Y_i(1)|D_i = 1, X_i = x\big) \times P(D_i = 1|X_i = x) + \\ Y^L(x) \times P(D_i = 0|X_i = x) - \mathbb{E}\big(Y_i(0)|D_i = 0, X_i = x\big)\times \\ \times P(D_i = 0|X_i = x) - Y^U(x) \times P(D_i = 1|X_i = x).\end{aligned} \tag{12.2}$$

$$\theta^U(x) = \frac{\mathbb{E}(Y_i|Z_i = 1, X_i = x) - \mathbb{E}(Y_i|Z_i = 0, X_i = x)}{\mathbb{E}(D_i|Z_i = 1, X_i = x) - \mathbb{E}(D_i|Z_i = 0, X_i = x)}. \tag{12.3}$$

To the contrary, when the selection into the treatment is negative, the bounds are given as

$$\theta^L(x) = \frac{\mathbb{E}(Y_i|Z_i = 1, X_i = x) - \mathbb{E}(Y_i|Z_i = 0, X_i = x)}{\mathbb{E}(D_i|Z_i = 1, X_i = x) - \mathbb{E}(D_i|Z_i = 0, X_i = x)}.$$
$$\theta^U(x) = \mathbb{E}\big(Y_i(1)|D_i = 1, X_i = x\big) \times P(D_i = 1|X_i = x) +$$
$$Y^U(x) \times P(D_i = 0, X_i = x) - \mathbb{E}\big(Y_i(0)|D_i = 0, X_i = x\big) \times$$
$$\cdot P(D_i = 0|X_i = x) - Y^L(x) \times P(D_i = 1|X_i = x).$$

12.4. Lemma 6.1

Proof. Define the Hajek projection of the multivariate random forest estimator:

$$\dot{\mathcal{MF}}(x, A_1, \ldots, A_N) - \mu = \sum_{i=1}^{N} \mathbb{E}\big(\mathcal{MF}(x, A_1, \ldots, A_N) - \mu|A_i\big) = \tag{12.4}$$
$$\frac{1}{\binom{N}{s}} \sum_{i=1}^{N} \mathbb{E}\Bigg(\sum_{1 \le i_1 \le \cdots \le i_s \le N} \mathbb{E}_\xi \widetilde{T}(x, \xi, A_{i_1}, \ldots, A_{i_s}) - \mu|A_i \Bigg),$$

where $\binom{N}{s}$ is the number of $i_1 \le \cdots \le i_s$ size-s subsets from $1, \ldots, N$ observations.

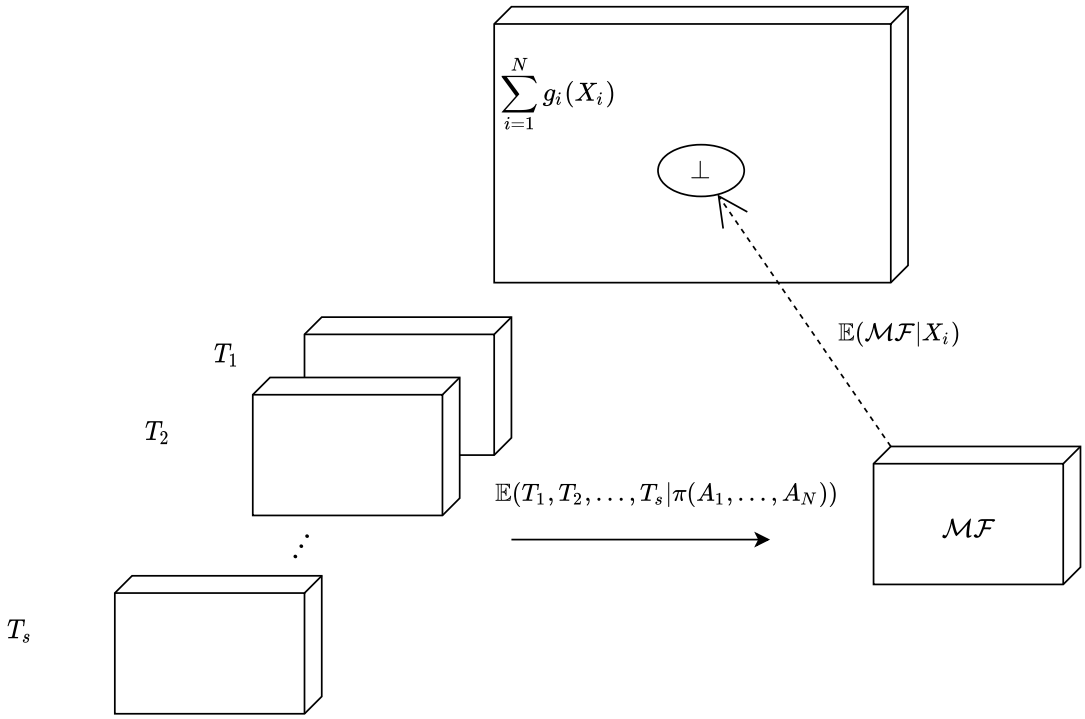


Figure 8. Hajek Projection of a (vector-valued) U-statistic, formed by the expectation of symmetric functions (i.e., trees), and aggregated over subsamples $i_1, \ldots, i_s$. $\pi(A_1, \ldots, A_N)$ denotes permutations of the data and $\perp$ is the orthogonality. Projection is the expectation of $\mathcal{MF}$ conditional on covariates X_i. It can be shown that $\mathbb{E}\big[(T - \dot{T}) \sum_{i=1}^{N} g_i(X_i)\big] = 0$.

When the observation i is not in samples $1 \leq i_1 \leq \cdots \leq i_s$, then the conditional expectation of the tree (aggregated over the randomization) is the same as the unconditional one. Therefore:

$$\mathbb{E}\big(\mathbb{E}_\xi \widetilde{T}(\xi, A_{i_1}, \ldots, A_{i_s}) | A_i\big) = \mathbb{E}_{\xi, A_{i_1}, \ldots, A_{i_s}} \widetilde{T}(x, \xi, A_{i_1}, \ldots A_{i_s}) = \mu.$$

Overall, there are $\binom{N-1}{s-1}$ samples that contain observation i. Moreover, the sequence of observations is *i.i.d.* and the trees are permutation symmetric. Therefore, for each sample,

$$\mathbb{E}\big(\mathbb{E}_\xi \widetilde{T}(x, \xi, A_{i_1}, \ldots, A_{i_s}) - \mu | A_i\big) = \widetilde{T}_1(A_i) - \mu, \tag{12.5}$$

where $\widetilde{T}_1(a) = \mathbb{E}_{\xi, A_2, \ldots, A_N} \widetilde{T}(x, \xi, a, A_2, \ldots, A_N)$.

Incorporating (12.5) in (12.4) yields:

$$\mathcal{M}\dot{\mathcal{F}}(x, A_1, \ldots, A_N) - \mu = \frac{\binom{N-1}{s-1}}{\binom{N}{s}} \sum_{i=1}^{N} \big(\widetilde{T}_1(A_i) - \mu\big) = \frac{s}{N} \sum_{i=1}^{N} \big(\widetilde{T}_1(A_i) - \mu\big). \tag{12.6}$$

Since the observations $A_1, \ldots, A_N$ are i.i.d, the same property holds for $\widetilde{T}_1(A_i)$. By taking the expectation of both sides in (12.6), we can easily verify that $\mathbb{E}\big(\mathcal{M}\dot{\mathcal{F}}(x)\big) = \mu$ where $\mathcal{M}\dot{\mathcal{F}}(x) = \mathcal{M}\dot{\mathcal{F}}(x, A_1, \ldots, A_N)$. Define Σ to be the covariance matrix of $\mathcal{M}\dot{\mathcal{F}}(x, A_1, \ldots, A_N)$. Then:

$$\Sigma = \mathbb{V}\bigg[\frac{s}{N} \sum_{i=1}^{N} \big(\widetilde{T}_1(A_i) - \mu\big)\bigg] = \frac{s^2}{N} \mathbb{V}\big(\widetilde{T}_1(A_i)\big) = \frac{s}{N} \mathbb{V}\big(\sum_{i=1}^{s} \widetilde{T}_1(A_i)\big) = \frac{s}{N} \mathbb{V}(\dot{\widetilde{T}}) \in \mathbb{R}^{M \times M}, \tag{12.7}$$

where $\dot{\widetilde{T}} = \sum_{i=1}^{s} \widetilde{T}_1(A_i)$ is the Hajek projection of a tree $\widetilde{T}(x, A_1, \ldots, A_N) = \mathbb{E}_\xi \widetilde{T}(x, \xi, A_1, \ldots, A_N) \in \mathbb{R}^M$. Note that, a tree $\widetilde{T}$ is symmetric in its arguments and observations $i = 1, \ldots, N$ are *i.i.d.* Therefore, the Hajek projection of a tree estimator reduces to $\sum_{i=1}^{s} \widetilde{T}_1(A_i)$ (as in (12.6)). We disregard the second (constant) term, as it does not enter in the variance $\mathbb{V}$. Note that since the statistic $\widetilde{T}_1(A_i)$ is a vector, the operation $\mathbb{V}$ applies coordinate-wise.

■

12.5. Lemma 6.2

Proof. Define the mean squared deviation of the multivariate forest estimator and its projection:

$$\mathbb{E}(\mathcal{M}\mathcal{F} - \mathcal{M}\dot{\mathcal{F}})^T \Sigma^{-1} (\mathcal{M}\mathcal{F} - \mathcal{M}\dot{\mathcal{F}}) = \mathbb{E}\big[\mathrm{tr}\Sigma^{-1} (\mathcal{M}\mathcal{F} - \mathcal{M}\dot{\mathcal{F}})(\mathcal{M}\mathcal{F} - \mathcal{M}\dot{\mathcal{F}})^T\big] = \tag{12.8}$$
$$\mathrm{tr}\Sigma^{-1} \mathbb{E}(\mathcal{M}\mathcal{F} - \mathcal{M}\dot{\mathcal{F}})(\mathcal{M}\mathcal{F} - \mathcal{M}\dot{\mathcal{F}})^T = \mathrm{tr}\Sigma^{-1/2} \mathbb{V}\big(\mathcal{M}\mathcal{F} - \mathcal{M}\dot{\mathcal{F}}\big)\Sigma^{-1/2}.$$

Assume there exist functions $\widetilde{T}_i$, such that the following equality holds:

$$\mathbb{E}\big(\widetilde{T}_i(X_i \in B)|X_i \notin B)\big) = 0. \tag{12.9}$$

Equation (12.9) is the necessary condition for the weak independence of the exchangeable sequences of X_i. Assume, $\widetilde{T}_i(X_i \in B)$ are symmetric, square-integrable, vector-valued functions. Then each $\widetilde{T}_i$ and $\widetilde{T}_{i'}$ are pairwise independent. Since $i = 1, \ldots, N$ is an exchangeable $(i.i.d)$ sequence, Theorem 6 of Peccati (2004) applies. In addition, Proposition 1 of Li (2020) applies to our case as well. We define Höeffding decomposition of a multivariate U-statistic:

$$\mathcal{MF} - \dot{\mathcal{MF}} = \frac{1}{\binom{N}{s}}\Bigg[\sum_{i<j}\binom{N-2}{s-2}\big(\widetilde{T}_2(A_i, A_j) - \mu\big) + \sum_{i<j<m}\binom{N-3}{s-3}\big(\widetilde{T}_3(A_i, A_j, A_m) - \mu\big) + \ldots \tag{12.10}$$

where $\widetilde{T}_1, \widetilde{T}_2 \ldots$ are second, third and higher order projections of a tree $\widetilde{T}$ that meet the following conditions:

$$\mathbb{E}\big(\widetilde{T}_i - \mu\big)^T\Sigma^{-1}\big(\widetilde{T}_{i'} - \mu\big) = 0 \text{ for each } i \neq i', \text{ and} \tag{12.11}$$

$$\mathbb{E}\big(\widetilde{T}_i - \mu\big)^T\Sigma^{-1}\big(\widetilde{T}_i - \mu\big) \leq \mathbb{E}\big(\widetilde{T} - \mu\big)^T\Sigma^{-1}\big(\widetilde{T} - \mu\big), \tag{12.12}$$

where $\widetilde{T}_i$ and $\widetilde{T_{i'}}$ are the i- th and i'-th projections of the tree, with $i = 1, \ldots, N$.

We fix the variance (Σ) of the multivariate random forest estimator. Moreover, we notice that $\binom{N}{s} \geq \binom{N-1}{s-1} \geq \binom{N-2}{s-2} \geq \binom{N-3}{s-3} \geq \ldots$. Therefore:

$$\mathcal{MF} - \dot{\mathcal{MF}} \leq \frac{s}{N}\Bigg[\sum_{i<j}\big(\widetilde{T}_2(A_i, A_j) - \mu\big) + \sum_{i<j<m}\big(\widetilde{T}_3(A_i, A_j, A_m) - \mu\big) + \ldots\Bigg], \tag{12.13}$$

where $\frac{s}{N} = \frac{\binom{N-1}{s-1}}{\binom{N}{s}}$. Based on Equation (12.12), the variance of $\mathcal{MF} - \dot{\mathcal{MF}}$ has an upper bound:

$$\mathbb{V}\big(\mathcal{MF} - \dot{\mathcal{MF}}\big) \leq \left(\frac{s}{N}\right)^2\mathbb{V}(\widetilde{T}). \tag{12.14}$$

In (12.7) we derived $\Sigma = \frac{s}{N}\mathbb{V}(\dot{\widetilde{T}})$. Plugging the value of Σ and (12.14) in (12.8) leads to the upper bound of the squared deviation:

$$\mathbb{E}(\mathcal{MF} - \dot{\mathcal{MF}})^T\Sigma^{-1}(\mathcal{MF} - \dot{\mathcal{MF}}) \leq tr\Bigg(\big(\frac{s}{N}\mathbb{V}(\dot{\widetilde{T}})\big)^{-1/2}\left(\frac{s}{N}\right)^2\mathbb{V}(\widetilde{T})\big(\frac{s}{N}\mathbb{V}(\dot{\widetilde{T}})\big)^{-1/2}\Bigg) = \tag{12.15}$$

$$\frac{s}{N}tr\Bigg(\big(\mathbb{V}(\dot{\widetilde{T}})\big)^{-1}\mathbb{V}(\widetilde{T})\Bigg)$$

In the final equality, we use the cyclic property of the trace operator: $tr(XYZ) = tr(YZX) = tr(ZXY)$.

■

12.6. Theorem 6.3

Proof. Bounded elements of $\mathbb{V}(\widetilde{T})$ directly follow from the proposed assumptions. According to Assumption 6.3, the number of observations in each terminal node is bounded above. This implies that the variance of the tree is bounded above by constant times $\mathbb{V}(Y_i|X_i = x)$. Moreover, Assumption 6.4 guarantees that $\mathbb{V}(Y_i|X_i = x)$ is bounded away from zero.

Throughout this proof, we carry the result of Wager and Athey (2018) regarding the order of the variance terms. In particular, Wager and Athey (2018) show that:

$$\mathbb{V}(\dot{\widetilde{T}})_{ii} = \frac{C}{\log^p(s)}, \text{ for some constant } C. \tag{12.16}$$

$\mathbb{V}(\dot{\widetilde{T}})_{ii}$ denotes the diagonal terms of the variance of the projection of a tree estimator. We show that the off-diagonal terms $\mathbb{V}(\dot{\widetilde{T}})_{ij} = o\big(\frac{1}{\log^p(s)}\big)$ for all $i \neq j$.

Start with the definition of a Hajek projection of a tree:

$$\dot{\widetilde{T}} - \mu = \sum_{i=1}^{s} \mathbb{E}(\widetilde{T}|A_i) \tag{12.17}$$

Since the observations are $i.i.d.$, then:

$$\mathbb{V}\big(\dot{\widetilde{T}}\big) = s\mathbb{V}\big(\mathbb{E}(\widetilde{T}|A_1\big). \tag{12.18}$$

Then it is clear to see that:

$$\mathbb{V}\big(\mathbb{E}(\widetilde{T}|A_1)\big) = \mathbb{V}\big(\mathbb{E}(\widetilde{T}|A_1) - \mathbb{E}(\widetilde{T}|X_1)\big) + \mathbb{V}\big(\mathbb{E}(\widetilde{T}|X_1\big). \tag{12.19}$$

Consider m-th outcome variable, where $m = 1, \ldots, M$. Since the tree is honest, the diagonal terms in (12.19) simplify as follows (see the Proof of Theorem 5 in Wager and Athey, 2018):

$$\begin{aligned}\mathbb{V}\big(\mathbb{E}(\widetilde{T}|A_1) - \mathbb{E}(\widetilde{T}|X_1)\big)_{mm} = \mathbb{V}\big(\mathbb{E}(S_{\ell_n}|X_1)(Y_{1m} - \mathbb{E}(Y_{1m}|X_1)\big)_{mm} \approx \\ \mathbb{E}\big[\big(\mathbb{E}(S_{\ell_n}|X_1)\big)^2\big]\mathbb{E}\big[\big(Y_{1m} - \mathbb{E}(Y_{1m}|X_1)\big)^2\big] = \\ \mathbb{E}\big[\big(\mathbb{E}(S_{\ell_n}|X_1)\big)^2\big]Var(Y_m|X_1 = x),\end{aligned} \tag{12.20}$$

and

$$\mathbb{V}\big(\mathbb{E}(\widetilde{T}|X_1\big)_{mm} = \mathbb{E}\big[\big(\mathbb{E}(S_{\ell_n}|X_1)\big)^2\big]Var\big(\mathbb{E}(Y_m|X_1 = x)\big). \tag{12.21}$$

where $\widetilde{T}_m$ is the estimator of a tree at a test point x. S_{ℓ_n} is the indicator function and equals one if $X_1 \in \ell_n(x, \Pi)$, and zero otherwise.

The off-diagonal terms equal to:

$$\mathbb{V}\big(\mathbb{E}(\widetilde{T}|A_1) - \mathbb{E}(\widetilde{T}|X_1)\big)_{mm'} = \mathbb{E}\big[\big(\mathbb{E}(S_{\ell_n}|X_1)\big)^2\big]\mathbb{E}\big[(Y_{1m} - \mathbb{E}(Y_{1m}|X_1))(Y_{1m'} - \mathbb{E}(Y_{1m'}|X_1))\big]. \tag{12.22}$$

According to Assumption 6.4, the variance of each outcome variable is bounded away from zero. Cauchy-Schwarz inequality implies that $|Cov(Y_{im}, Y_{im'}|X_1)|$ is also bounded away from zero [4].

$$|Cov(Y_{1m}, Y_{1m'}|X_1)| \leq \sqrt{Var(Y_{1m}|X_1 = x)Var(Y_{1m'}|X_1 = x)}.$$

Wager and Athey (2018) show that

$$\mathbb{E}\big[\big(\mathbb{E}(S_{\ell_n}|X_1)\big)^2\big] \geq \frac{(p-1)!}{2^{p+1}\log^p(s)} \cdot \frac{1}{ks}, \tag{12.23}$$

where k is the minimum number of observations in a given terminal node. Combining (12.18) and (12.23) yields the order of diagonal and off-diagonal terms:

$$\mathbb{V}\big(\dot{\widetilde{T}}\big)_{mm} = o\bigg(\frac{1}{log^p(s)}\bigg), \text{ and } \mathbb{V}\big(\dot{\widetilde{T}}\big)_{mm'} = o\bigg(\frac{1}{log^p(s)}\bigg). \tag{12.24}$$

Now we prove that $\frac{s}{N}tr\bigg(\big(\mathbb{V}(\dot{\widetilde{T}})\big)^{-1}\mathbb{V}(\widetilde{T})\bigg) \to 0$ in a more general framework. Consider, we have two square matrices C and D with diagonal (c_{ii}, d_{ii}) and non-diagonal terms (c_{ij}, d_{ij}), respectively. Moreover, they exhibit the following properties:

$$1.\ d_{ii} \geq \eta \text{ for some } \eta \in \mathbb{R}^+ \text{ and for all } i = 1, \dots M, \tag{12.25}$$

$$2.\ c_{ii} \geq \frac{d_{ii}}{\log(N)}, \tag{12.26}$$

$$3.\ c_{ij} = o\bigg(\frac{1}{\log(N)}\bigg). \tag{12.27}$$

Then we show that $\frac{s}{N}tr(C^{-1}D) \to 0$. Recall that the Leibniz formula for the determinant is given as follows:

$$det(C) = \sum_{\pi}\bigg(\text{sgn}(\pi)\prod_{i=1}^{M} c_{i,\pi_i}\bigg), \tag{12.28}$$

where π is a permutation function that reorders the set $\{1, \dots, M\}$. Diagonal and off-diagonal terms are on the same order, their product is also on the same order. Therefore,

[4] An alternative argument is to notice that the term in the integrand consists of multiples of the first and second moments of the outcome variables Y_{1m} and $Y_{1m'}$. Since these moments are continuous, they are bounded. Thus, their expectation is also bounded.

$det(C)$ is asymptotically equivalent to either $\prod_{i=1}^{M} c_{ii}$ or $\prod_{i=1}^{M} c_{ij}$ where $i \neq j$. For simplicity, we keep the notation that $det(C) \sim^a \prod_{i=1}^{M} c_{ii}$, where " $\sim^a$ " denotes asymptotic equivalence. Based on Cramer's rule, we can write i-th diagonal term of the inverse of C:

$$(C^{-1})_{ii} = \frac{det(C_{-i})}{det(C)}.$$

C_{-i} is the matrix where we remove the i-th row and the i-th column. By the same argument, $det(C_{-i}) \sim^a \prod_{j=1}^{M-1} c_{jj}$. Then we end up with:

$$(C^{-1})_{ii} \sim^a \frac{\prod_{i=1}^{M-1} c_{jj}}{\prod_{j=1}^{M} c_{ii}} = \frac{1}{c_{ii}}.$$

The i-th diagonal entry of the matrix

$$(C^{-1}D)_{ii} = (c^{-1})_{ii} d_{ii} + \sum_{j \neq i} (c^{-1})_{ij} d_{ji} \sim^a \frac{d_{ii}}{c_{ii}} \leq \log(N).$$

The last equality follows from Property 2 (12.26). Therefore, the trace of $(C^{-1}D)$ is also on the order of $\log(N)$. We take the limit of $\frac{s}{N} tr(C^{-1}D)$, where $s = N^{\beta}$ and $\beta < 1$. L'Hôpital's rule yields:

$$\lim_{N \to \infty} \frac{s}{N} \log(N) = \lim_{N \to \infty} \frac{1}{(1-\beta) N^{1-\beta}} \to 0. \tag{12.29}$$

The proof is complete by letting $C = \left(\mathbb{V}(\dot{\widetilde{T}})\right)^{-1}$ and $D = \mathbb{V}(\widetilde{T})$.

■

12.7. Asymptotic Normality of the Products of Trees

In this section, we examine the large sample theory for trees that can be products of different parameters. We modify the definition of a tree, $\widetilde{T}$. Consider three different mean outcomes in some terminal node ℓ_n defined as before:

$$\widetilde{T}_1(x, \xi, A_i, \ldots, A_N) = \sum_{n=1}^{|\Pi|} 1(x \in \ell_n) \frac{1}{N_{\ell_n}} \sum_{i: X_i \in \ell_n} Y_{i1},$$

$$\widetilde{T}_2(x, \xi, A_i, \ldots, A_N) = \sum_{n=1}^{|\Pi|} 1(x \in \ell_n) \frac{1}{N_{\ell_n}} \sum_{i: X_i \in \ell_n} Y_{i2},$$

$$\widetilde{T}_M(x, \xi, A_i, \ldots, A_N) = \sum_{n=1}^{|\Pi|} 1(x \in \ell_n) \frac{1}{N_{\ell_n}} \sum_{i: X_i \in \ell_n} Y_{i3}.$$

Let the tree estimator be a 2×1 vector:

$$\widetilde{T}(x, \xi, A_i, \ldots, A_N) = \begin{bmatrix} \widetilde{T}_1(x, \xi, A_i, \ldots, A_N) - \widetilde{T}_2(x, \xi, A_i, \ldots, A_N)\widetilde{T}_3(x, \xi, A_i, \ldots, A_N) \\ \widetilde{T}_1(x, \xi, A_i, \ldots, A_N) \end{bmatrix} \tag{12.30}$$

The multivariate random forest estimator is defined as before:

$$\mathcal{MF}(x, A_1, \ldots, A_N) = \frac{1}{\binom{N}{s}} \sum_{1 \leq i_1 \leq \cdots \leq i_s \leq N} \mathbb{E}_\xi \widetilde{T}(x, \xi, A_{i_1}, \ldots, A_{i_s}). \tag{12.31}$$

Proposition 12.1 shows that parameters estimated by the multivariate forest are asymptotically normally distributed.

Proposition 12.1. *The multivariate random forest estimator is asymptotically normally distributed:*

$$\Sigma^{-1}(\mathcal{MF}(x, A_1, \ldots, A_N) - \mu) \xrightarrow{d} \mathcal{N}(0, I),$$

where Σ *is a variance-covariance matrix of* $\dot{\mathcal{MF}}(x, A_1, \ldots, A_N)$*,* I *is a* 2×2 *identity matrix and* 0 *is a* 2×1 *vector of zeroes.*

Proof. Assumptions 6.1 - 6.5 apply in this setting. Moreover, Propositions 6.1 and 6.2 hold for partially identified parameters. To prove that the Proposition 6.3 applies, it is sufficient to show that $\mathbb{V}\big(\dot{\widetilde{T}}\big)_{mm} \geq \frac{\widetilde{T}_{mm}}{\log^p(s)}$ and $\mathbb{V}\big(\dot{\widetilde{T}}\big)_{mm'} = o\left(\frac{1}{\log^p(s)}\right)$.

First, we prove that $\mathbb{V}\big(\dot{\widetilde{T}}\big)_{mm} \geq \frac{\widetilde{T}_{mm}}{\log^p(s)}$. Start with the definition of the Hajek projection of a tree and its variance:

$$\dot{\widetilde{T}}(x, A_1, \ldots, A_N) - \mu = \sum_{i=1}^{s} \mathbb{E}(\widetilde{T}(x, A_1, \ldots, A_N)|A_i), \tag{12.32}$$

$$\mathbb{V}\big(\dot{\widetilde{T}}(x, A_1, \ldots, A_N)\big) = s\mathbb{V}\big(\mathbb{E}\big(\widetilde{T}(x, A_1, \ldots, A_N)|A_1\big)\big), \tag{12.33}$$

Equation (12.33) holds because the observations are *i.i.d.* As before, the variance term can be expanded as:

$$\mathbb{V}\big(\mathbb{E}\big(\widetilde{T}(x, A_1, \ldots, A_N)|A_1\big)\big) = \mathbb{V}\big(\mathbb{E}(\widetilde{T}|A_1) - \mathbb{E}(\widetilde{T}|X_1)\big) + \mathbb{V}\big(\mathbb{E}(\widetilde{T}|X_1)\big). \tag{12.34}$$

We take the expectation of the tree estimator:

$$\mathbb{E}\big(\widetilde{T}(x, \xi, A_i, \ldots, A_N)|A_1\big) = \begin{bmatrix} \mathbb{E}\big(\widetilde{T}_1|A_1\big) - \mathbb{E}\big(\widetilde{T}_2\widetilde{T}_3|A_1\big) \\ \mathbb{E}\big(\widetilde{T}_1|A_1\big) \end{bmatrix} \tag{12.35}$$

For notational convenience, we disregard $(x, \xi, A_i, \ldots, A_N)$ in each tree. We expand the terms in the brackets and then take the variance:

$$\mathbb{E}\big(\widetilde{T}_1|A_1\big) = \mathbb{E}(S_{\ell_n}|X_1)Y_{11} \text{ and} \tag{12.36}$$

$$\mathbb{E}\big(\widetilde{T}_1|X_1\big) = \mathbb{E}(S_{\ell_n}|X_1)\mathbb{E}(Y_1|X_1). \tag{12.37}$$

Note that, because of honesty, the splits are independent of the outcome variable, therefore, $\mathbb{E}(S_{\ell_n}|A_1) = \mathbb{E}(S_{\ell_n}|X_1)$. Moreover, because of honesty, we can take the expectation of the outcome variable and S_{ℓ_n} separately as they do not depend on each other.

$$\mathbb{V}\big(\mathbb{E}(\widetilde{T}_1|A_1) - \mathbb{E}(\widetilde{T}_1|X_1)\big)_{11} = \mathbb{V}\big(\mathbb{E}\big(S_{\ell_n}|X_1\big)\big(Y_{11} - \mathbb{E}(Y_1|X_1)\big)\big)_{11} \approx \mathbb{E}\big[\mathbb{E}\big(S_{\ell_n}|X_1\big)^2\big]Var(Y_1|X_1 = x). \tag{12.38}$$

$$\mathbb{V}\big(\mathbb{E}(\widetilde{T}_1|X_1)\big)_{11} = \mathbb{E}\big[\mathbb{E}\big(S_{\ell_n}|X_1\big)^2\big]Var\big(\mathbb{E}(Y_1|X_1 = x\big). \tag{12.39}$$

The second term in the variance of $\mathbb{E}(\widetilde{T}_1|A_1) - \mathbb{E}(\widetilde{T}_1|X_1)$ is negligibly small, so we ignore it. Now we can expand the product term in a tree $\mathbb{E}\big(\widetilde{T}_2\widetilde{T}_3|A_1\big)$:

$$\begin{aligned}\mathbb{E}\big(\widetilde{T}_2\widetilde{T}_3|A_1\big) &= \mathbb{E}\left[\left(\sum_{n=1}^{|\Pi|} 1(x \in \ell_n) \cdot \frac{1}{N_{\ell_n}^2} \sum_{i:X_i \in N_{\ell_n}} Y_{i2} \sum_{i:X_i \in N_{\ell_n}} Y_{i3}\right)\middle| A_1\right] = \\ &\mathbb{E}(S_{\ell_n}|X_1)\frac{1}{N_{\ell_n}^2}\mathbb{E}\left[\left(\sum_{i:X_i \in N_{\ell_n}} Y_{i2}Y_{i3} + \sum_{i \neq i':X_i \in N_{\ell_n}} Y_{i2}Y_{i'3}\right)\middle| A_1\right] = \\ &\mathbb{E}(S_{\ell_n}|X_1) \cdot \frac{1}{N_{\ell_n}^2} \cdot N_{\ell_n}Y_{21}Y_{31} = \mathbb{E}(S_{\ell_n}|X_1) \cdot \frac{1}{N_{\ell_n}}Y_{21}Y_{31}.\end{aligned} \tag{12.40}$$

In the third equality of (12.40), we use information that the observations are independent, hence the expectation of the product of outcomes for each $i \neq i'$ is zero. Then the variance of $\mathbb{E}\big(\widetilde{T}_2\widetilde{T}_3|A_1\big)$ is given as follows:

$$\mathbb{V}\big(\mathbb{E}\big(\widetilde{T}_2\widetilde{T}_3|A_1\big) - \mathbb{E}\big(\widetilde{T}_2\widetilde{T}_3|X_1\big)\big)_{11} \approx \mathbb{E}\big[\big(\mathbb{E}(S_{\ell_n}|X_1)\big)^2\big]\frac{1}{N_{\ell_n}^2}Var\big(Y_2Y_3|X_1 = x\big), \tag{12.41}$$

$$\mathbb{V}\big(\mathbb{E}\big(\widetilde{T}_2\widetilde{T}_3|X_1\big)\big)_{11} = \mathbb{E}\big[\big(\mathbb{E}(S_{\ell_n}|X_1)\big)^2\big]\frac{1}{N_{\ell_n}^2}Var\big(\mathbb{E}(Y_2Y_3|X_1 = x)\big). \tag{12.42}$$

Overall, the variance of the first parameter in the tree equals:

$$\begin{aligned}&\mathbb{V}\big(\mathbb{E}\big(\widetilde{T}(x, A_1, \ldots, A_N)|A_1\big)\big)_{11} = \\ &\mathbb{E}\big[\mathbb{E}\big(S_{\ell_n}|X_1\big)^2\big]\big[Var(Y_1|X_1 = x) + Var\big(\mathbb{E}(Y_1|X_1 = x\big) + \\ &\frac{1}{N_{\ell_n}^2}Var\big(Y_2Y_3|X_1 = x\big) + \frac{1}{N_{\ell_n}^2}Var\big(\mathbb{E}(Y_2Y_3|X_1 = x)\big)\big].\end{aligned} \tag{12.43}$$

Wager and Athey (2018) prove that

$$\mathbb{E}\big[\mathbb{E}\big(S_{\ell_n}|X_1\big)^2\big] \geq \frac{(p-1)!}{2^{(p+1)\log^p(s)}}\frac{1}{ks}, \tag{12.44}$$

where k is the minimum number of observations in a leaf.

The variance of the first parameter in the tree estimator equals:

$$\mathbb{V}\big(\widetilde{T}_1 - \widetilde{T}_2\widetilde{T}_3|A_1\big)_{11} = \frac{1}{N_{\ell_n}}Var(Y_1|X_1=x) + \frac{1}{N_{\ell_n}^3}Var(Y_2Y_3|X_1=x). \tag{12.45}$$

k is the minimum number of observations in a leaf, therefore, we can replace N_{ℓ_n} with k to obtain the lower bound of the variance terms. Note that

$$\mathbb{V}\big(\dot{\widetilde{T}}\big)_{11} = \frac{(p-1)!}{2^{(p+1)\log^p(s)}}\big(\mathbb{V}\big(\widetilde{T}\big)_{11} + \frac{1}{k}Var\big(\mathbb{E}(Y_1|X_1=x\big) + \frac{1}{k^3}Var\big(\mathbb{E}(Y_2Y_3|X_1=x)\big)\big]. \tag{12.46}$$

Moreover, $(p-1)! \geq 2^{(p+1)}$, thus $\mathbb{V}\big(\dot{\widetilde{T}}\big)_{11} \geq \frac{\mathbb{V}\big(\widetilde{T}\big)_{11}}{\log^p(s)}$. Analogous proof shows that the same holds for the second parameter of the tree.

Cauchy-Schwarz inequality implies the second part of the proof, $\mathbb{V}\big(\dot{\widetilde{T}}\big) = o\bigg(\frac{1}{\log^p(s)}\bigg)$:

$$\big|\mathbb{V}\big(\dot{\widetilde{T}}\big)_{12}\big| \leq \sqrt{\mathbb{V}\big(\dot{\widetilde{T}}\big)_{11}\mathbb{V}\big(\dot{\widetilde{T}}\big)_{22}} = \sqrt{C_1\frac{1}{\log^p(s)}\cdot C_2\frac{1}{\log^p(s)}} = o\bigg(\frac{1}{\log^p(s)}\bigg), \tag{12.47}$$

for some constants C_1 and C_2.

■

12.8. Proposition 6.1

Under the assumption that the Hoeffding decomposition exists for quantile regression forests with multiple outcomes, Lemma 6.1 and Lemma 6.2 and their corresponding proofs in Appendix 12.4 and 12.5 apply directly. The goal is to show that

$$\lim_{N\to\infty}\frac{s}{N}tr\bigg(\big(\mathbb{V}(\dot{\widetilde{T}})\big)^{-1}\mathbb{V}(\widetilde{T})\bigg) = 0. \tag{12.48}$$

In Theorem 6.3, we introduce a Hajek projection of a tree, and show the convergence of the deviation of the multivariate forest and its projection to zero. The proof is analogous, except the variance of the tree is defined as

$$\begin{aligned}\mathbb{V}\big(\mathbb{E}(\widetilde{T}|A_1) - \mathbb{E}(\widetilde{T}|X_1)\big)_{mm} = \mathbb{V}\big(\mathbb{E}(S_{\ell_n}|X_1)(1_{\{Y_{1m}\leq y\}} - \mathbb{E}(1_{\{Y_{1m}\leq y\}}|X_1)\big)_{mm} \approx \\ \mathbb{E}\big[\big(\mathbb{E}(S_{\ell_n}|X_1)\big)^2\big]\mathbb{E}\big[\big(1_{\{Y_{1m}\leq y\}} - \mathbb{E}(1_{\{Y_{1m}\leq y\}}|X_1)\big)^2\big] = \\ \mathbb{E}\big[\big(\mathbb{E}(S_{\ell_n}|X_1)\big)^2\big]Var(1_{\{Y_m\leq y\}}|X_1=x) \leq \mathbb{E}\big[\big(\mathbb{E}(S_{\ell_n}|X_1)\big)^2\big],\end{aligned} \tag{12.49}$$

and

$$\mathbb{V}\big(\mathbb{E}(\widetilde{T}|X_1)\big)_{mm} = \mathbb{E}\big[\big(\mathbb{E}(S_{\ell_n}|X_1)\big)^2\big] Var\big(\mathbb{E}(1_{\{Y_{1m}\leq y\}}|X_1 = x)\big)\mathbb{E}\big[\big(\mathbb{E}(S_{\ell_n}|X_1)\big)^2\big], \tag{12.50}$$

where as defined before, S_{ℓ_n} is an indicator function and equals one if $X_1 \in \ell_n(x, \Pi)$, and zero otherwise.

The off-diagonal terms equal to:

$$\begin{aligned}\mathbb{V}\big(\mathbb{E}(\widetilde{T}|A_1) - \mathbb{E}(\widetilde{T}|X_1)\big)_{mm'} = \\ \mathbb{E}\big[\big(\mathbb{E}(S_{\ell_n}|X_1)\big)^2\big]\mathbb{E}\big[(1_{\{Y_{1m}\leq y\}} - \mathbb{E}(1_{\{Y_{1m}\leq y\}}|X_1))(1_{\{Y_{1m'}\leq y\}} - \mathbb{E}(1_{\{Y_{1m'}\leq y\}}|X_1))\big] \leq \\ \leq \mathbb{E}\big[\big(\mathbb{E}(S_{\ell_n}|X_1)\big)^2\big].\end{aligned} \tag{12.51}$$

The last inequality in each case follows by the fact that the expectation and variance of $1_{\{Y_{1m}\leq y\}}$ is bounded by one from above, as it is a binary variable.

The rest of the proof is analogous to Theorem 6.3.

12.9. Asymptotic Normality for Bounds in Proposition 6.1

The proof is equivalent for forests with the treatment effect as a target outcome. Wager and Athey (2018) show that for the treatment effect as the outcome variable,

$$\mathbb{E}\big[\big(\mathbb{E}(S_{\ell_n}|X_1)\big)^2\big] \geq \frac{(p-1)!}{2^{p+1}\log^p(s)} \cdot \frac{\epsilon}{ks}, \tag{12.52}$$

where, ϵ is a constant from Assumption 6.5. This does not change the results of the proof, as the order of $\mathbb{E}\big[\big(\mathbb{E}(S_{\ell_n}|X_1)\big)^2\big]$ is still $o\left(\frac{1}{\log^p(s)}\right)$.

When the selection into the treatment is negative, and the estimator of the lower bound in Proposition 4.1 is defined as

$$\theta^L(x) = \frac{1}{|i : D_i = 1, X_i \in \ell_n|} \sum_{\{i:D_i=1, X_i\in\ell_n\}} Y_i - \frac{1}{|i : D_i = 0, X_i \in \ell_n|} \sum_{\{i:D_i=0, X_i\in\ell_n\}} Y_i,$$

then (12.52) applies to the lower bound. Moreover, for simplicity, assume $P(D_i = 1|X_i = x) = P(D_i = 0|X_i = x) = p_D$. Define the estimator of the upper bound in Proposition 4.1 as

$$\begin{aligned}\theta^U(x) = p_D \times \left[\frac{1}{|i : D_i = 1, X_i \in \ell_n|} \sum_{\{i:D_i=1, X_i\in\ell_n\}} Y_i - \frac{1}{|i : D_i = 0, X_i \in \ell_n|} \sum_{\{i:D_i=0, X_i\in\ell_n\}} Y_i\right] + \\ + p_D\Big(\left[\frac{1}{|i : D_i = 1, X_i \in \ell_n|} \sum_{\{i:D_i=1, X_i\in\ell_n\}} 1_{\{Y_i\leq y\}} - \frac{1}{|i : D_i = 0, X_i \in \ell_n|} \sum_{\{i:D_i=0, X_i\in\ell_n\}} 1_{\{Y_i\leq y\}}\right] = \\ p_D \times \theta^L(x) + p_D\Big[Q_{\alpha_1}(x) - Q_{\alpha_0}(x)\Big].\end{aligned}$$

Based on our results, $\theta^L(x) \geq \frac{(p-1)!}{2^{p+1}\log^p(s)} \cdot \frac{\epsilon}{ks}$, and $\theta^U(x) \geq \frac{(p-1)!}{2^{p+1}\log^p(s)} \cdot \frac{\epsilon \times p_D}{ks}$. Hence, the result in Theorem 6.3 applies. The proof is analogous when the selection into the treatment is positive.

12.10. Proposition 6.2

Let Σ be the covariance matrix of $\widetilde{\theta}(X_i, S^{est}, \Pi)$. Following Athey and Imbens (2016), algebraic transformations of the loss function simplify the following way:

$$\mathbb{E}_{S^{tr},S^{est}}\Big[\big(\theta(X_i) - \widetilde{\theta}(X_i, S^{est}, \Pi)\big)^T\Sigma^{-1}\big(\theta(X_i) - \widetilde{\theta}(X_i, S^{est}, \Pi)\big) - \theta(X_i)^T\Sigma^{-1}\theta(X_i)\Big] = \tag{12.53}$$

$$\mathbb{E}_{S^{tr},S^{est}}\Big[\underbrace{\big(\theta(X_i) - \theta(X_i, \Pi)\big)}_{A} + \underbrace{\theta(X_i, \Pi) - \tilde{\theta}(X_i, S^{est}, \Pi)\big)}_{B}{}^T\Sigma^{-1}\big(\underbrace{\theta(X_i) - \theta(X_i, \Pi)}_{A} + \tag{12.54}$$

$$\underbrace{\theta(X_i, \Pi) - \tilde{\theta}(X_i, S^{est}, \Pi)}_{B}\big) - \theta(X_i)^T\Sigma^{-1}\theta(X_i)\Big] =$$

$$\mathbb{E}_{S^{tr}}\Big(\theta(X_i)^T\Sigma^{-1}\theta(X_i) - 2\theta(X_i)^T\Sigma^{-1}\theta(X_i, \Pi) + \tag{12.55}$$

$$\theta(X_i, \Pi)^T\Sigma^{-1}\theta(X_i, \Pi) - \theta(X_i)^T\Sigma^{-1}\theta(X_i)\Big) +$$

$$\mathbb{E}_{X_i,S^{est}}\Big(\big(\theta(X_i, \Pi) - \widetilde{\theta}(X_i, S^{est}, \Pi)\big)^T\Sigma^{-1}\big(\theta(X_i, \Pi) - \widetilde{\theta}(X_i, S^{est}, \Pi)\big)\Big) = \tag{12.56}$$

$$-\mathbb{E}_{X_i}\big(\theta(X_i, \Pi)^T\Sigma^{-1}\theta(X_i, \Pi)\big) + \mathbb{E}(tr(I))_{2\times 2}. \tag{12.57}$$

The second equality follows after taking into account the independence of the train and estimation data, $cov(A, B) = 0$. The final equality is based on the fact that $\theta(X_i, \Pi) = \mathbb{E}(\theta(X_i)|X_i \in \ell(x, \Pi))$, $\mathbb{E}\big(\widetilde{\theta}(X_i, S^{est}, \Pi)\big) = \theta(X_i, \Pi)$, and:

$$\mathbb{E}_{X_i,S^{est}}\Big(\big(\theta(X_i, \Pi) - \widetilde{\theta}(X_i, S^{est}, \Pi)\big)^T\Sigma^{-1}\big(\theta(X_i, \Pi) - \widetilde{\theta}(X_i, S^{est}, \Pi)\big)\Big) =$$

$$tr\Big(\Sigma^{-1}\mathbb{E}\big(\theta(X_i, \Pi) - \widetilde{\theta}(X_i, S^{est}, \Pi)\big)^T\big(\theta(X_i, \Pi) - \widetilde{\theta}(X_i, S^{est}, \Pi)\big)\Big) =$$

$$tr(\Sigma^{-1}\Sigma) = tr(I)_{2\times 2},$$

where $tr(I)_{2\times 2}$ is the trace of a 2×2 identity matrix. Since $\mathbb{E}(tr(I))_{2\times 2}$ does not depend on the parameter of interest, we can disregard it. Hence, the optimal parameter maximizes the

unbiased estimator of the negative mean squared error:

$$\hat{\theta}(x, S^{est}, \Pi) = \arg\max_{\widetilde{\theta}} \frac{1}{N^{tr}} \sum_{\ell} N_{\ell}^{tr} \big(\widetilde{\theta}(X_i, \Pi)^T \widehat{\Sigma}^{-1} \widetilde{\theta}(X_i, \Pi) | X_i = x\big) - \frac{1}{N^{tr}} \sum_{\ell} \sum_{b} L_{\alpha_b}(Y_i, q) | X_i = x\big), \tag{12.58}$$

where the covariance matrix can be estimated as $\hat{\Sigma} = \hat{\Sigma}\big(\tilde{\theta}(X_i, S^{tr}, \Pi) | N^{est}\big)$. In this study, training and estimation samples have an equal number of observations, $N^{tr} = N^{est}$.

12.11. A Test for Heterogeneity

In this section, we adapt the generic machine learning approach by Chernozhukov et al. (2018) and test whether these bounds inherit statistically significant heterogeneity.

Model and Parameters

Consider an alternative specification for the lower and upper bounds of the outcome variable

$$Y^B(X_i) = \theta_0^B(X_i) + D_i \theta^B(X_i) + U_i^B, \tag{12.59}$$

where $B \in \{L, U\}$. $Y^B(X_i)$ is the bound of the outcome variable that can vary across covariates ($B \in \{L, U\}$). Note that $\theta_0^L(x) \leq \mathbb{E}\big(Y_i(0) | X_i = x\big) \leq \theta_0^U(x)$, and Proposition 4.1 (in Subsection 4.1) proves that, under the positive treatment selection,

$$\begin{aligned} \theta_0^L(x) &= \mathbb{E}\big(Y_i(0) | D_i = 0, X_i = x\big), \\ \theta_0^U(x) &= \mathbb{E}\big(Y_i(0) | D_i = 0, X_i = x\big) \times P(D_i = 0 | X_i = x) + Y^U(x) \times P\big(D_i = 1 | X_i = x\big). \end{aligned} \tag{12.60}$$

When the selection into the treatment is negative,

$$\begin{aligned} \theta_0(x)^L &= \mathbb{E}\big(Y_i(0) | D_i = 0, X_i = x\big) \times P(D_i = 0 | X_i = x) + Y^L(x) \times P\big(D_i = 1 | X_i = x\big), \\ \theta_0(x)^U &= \mathbb{E}\big(Y_i(0) | D_i = 0, X_i = x\big). \end{aligned} \tag{12.61}$$

$\theta^L(x) \leq \theta(x) \leq \theta^U(x)$ denotes treatment effect bounds described in Proposition 4.1. Moreover, in each leaf ℓ (with N^{ℓ} number of observations), the outcome bounds are defined as

$$\begin{aligned} Y^L(X_i) &= inf\{Y_1, Y_2, \ldots, Y_{N^{\ell}}\}, \\ Y^U(X_i) &= sup\{Y_1, Y_2, \ldots, Y_{N^{\ell}}\}. \end{aligned} \tag{12.62}$$

Validity of (12.59) relies on the following assumption:

Assumption 12.3 (**Orthogonality of the Errors**). *The error terms* U_i^B *in* (12.59) *are independent of* D_i *and* $\mathbb{E}\big(U_i^B|D_i, X_i\big) = \big(U_i^B|X_i\big) = 0$ *for* $B \in \{L, U\}$.

Assumption 12.3 guarantees that the lower and upper bounds of the outcome variable are valid. Assumption 12.3 does not rule out the correlation of the errors U_i^B with the outcome error through the part unrelated to D_i.

The population random variables $Y^B(X_i)$, $\theta_0^B(X_i)$ and $\theta^B(X_i)$ are unknown. However, instead of focusing on them directly, the aim is to model their *key features*. Define the unbiased estimators (proxies) of the unknown random variables as $\tilde{Y}^B(X_i) = \tilde{Y}^B(X_i, S^{tr}, \Pi)$, $s_0^B(X_i) = \tilde{\theta}_0^B(X_i, S^{est}, \Pi)$ and $s^B(X_i) = \tilde{\theta}^B(X_i, S^{est}, \Pi)$, respectively. We are interested in the conditional treatment effect bounds $\mathbb{E}\big(\theta^B(X_i)|s^B(X_i)\big)$. An objective, therefore, is to find parameters that minimize the best linear predictor of $\theta^B(X_i)$:

$$\text{BLP}\big(\theta^B(X_i)|s^B(X_i)\big) = \arg \min_{f^B(X_i) \in span(1, s^B(X_i))} \mathbb{E}\big(\theta^B(X_i) - f^B(X_i)\big)^2, \tag{12.63}$$

which, if exists, is defined by the projection of $\theta^B(X_i)$ on the linear span of 1 and $s^B(X_i)$.

To estimate the parameters in (12.59), we split data into two partitions, S^{tr} and S^{est}, respectively. As before, to guarantee that Assumption 6.1 (honesty) holds, we grow the trees based on the train partition and estimate the treatment effect bounds and the intercepts in the estimation sample. Then we build the second multivariate forest by using the estimation sample and predict the outcome bounds in the train data. Algorithm 2 summarizes the steps to obtain predicted proxies of the parameters.

The setup in this study is different from the one considered by Chernozhukov et al. (2018). The authors introduce randomized control trials, where for each subject i, Y_i is known, and, conditional on X_i, D_i is unrelated to ε_i. The test in this article applies to partially identified parameters with an endogenous treatment.

Hetergeneity

We consider two strategies for identifying and estimating the best linear predictor of $\theta^B(X_i)$. The first model entails estimating the parameters in a weighted linear regression, while the second strategy relies on Horvitz-Thompson transformation of the outcome (Horvitz and Thompson, 1952).

Algorithm 2: Proxies of the Outcome, Intercept, Treatment Effect Bounds

Require: number of trees (M_m), tree depth, number of leafs $|\Pi|$, number of observations for each bootstrapped data sample (s), number of observations in each leaf, data $\{Y_i, X_i, D_i\}_{i=1}^N$.

Ensure: Predicted proxies of the outcome bounds, $\tilde{Y}^B(x)$, intercept, $s_0^B(x)$, and of treatment effects, $s^B(x)$, at a test point x (for $B \in \{L, U\}$).

1. Divide data into train (S^{tr}), estimation (S^{est}) and test samples (S^{te}).

2. Use S^{tr} to obtain optimal partitions by multivariate random forests defined in Algorithm 1.

3. Use S^{est} to obtain proxies $s_0^B(x, S^{est})$ and $s^B(x, S^{est})$ as in Algorithm 1.

5. Based on S^{est}, build another multivariate random forest and obtain optimal partitions of the covariate space (with the same moment function as in 2. and 3.).

4. Estimate $Y^L(x) = min\{Y_1, Y_2, \ldots, Y_{N_\ell}\}$ and $Y^U(x) = max\{Y_1, Y_2, \ldots, Y_{N_\ell}\}$ in the train sample S^{tr}.

5. Predict $\tilde{Y}^B(x)$, $s_0^B(x)$ and $s^B(x)$ at each test point x.

Weighted Linear Projection

The BLP can be minimized through the weighted linear regression:

$$Y^B(X_i) = X_{1i}\alpha^B + (D_i - p(X_i))\beta_1^B + (D_i - p(X_i))(s^B - \mathbb{E}s^B)\beta_2^B + \varepsilon_i^B,$$
$$\text{with } \mathbb{E}\big[w(X_i)\varepsilon_i^B X_i^B\big] = 0, \text{ and } B \in \{L, U\}. \tag{12.64}$$

The weight is $w(X_i) = \big(p(X_i)(1-p(X_i)\big)^{-1}$ and $s^B = s^B(X_i)$. $p(X_i)$ denotes the propensity score for each individual i. Moreover, $X_i^B = [X_{1i}^B, X_{2i}^B]$ with $X_{1i}^B = [1, s_0^B(X_i)]$ and $X_{2i}^B = \big[(D_i - p(X_i)), (D_i - p(X_i))(s^B - \mathbb{E}s^B)\big]$. α^B is a vector of effects of X_{1i}^B on the outcome bound. Throughout the study, we label $D_i - p(X_i)$ as the "demeaned treatment" and $(D_i - p(X_i))(s^B - \mathbb{E}s^B)$ as the "interaction". Under the weight, $w(X_i)$, the interaction is orthogonal to the demeaned treatment and to other regressors that are functions of X_i. Based on Assumption 12.3, Theorem 12.1 states the main result.

Theorem 12.1. *Consider $x \mapsto s_0^B(x)$ and $x \mapsto s^B(x)$ as fixed maps. Assume, $\mathbb{E}(X_i X_i')$ has a full rank. Moreover, assume, Y_i and X_i have finite second moments, and $Var(s^B(X_i)) > 0$.*

Then, β_1 and β_2 in (12.64) also solve the best linear predictor problem for the target $\theta^B(X_i)$:

$$(\beta_1^B, \beta_2^B)' = \arg\min_{b_1^B, b_2^B} \mathbb{E}\big[\theta^B(X_i) - b_1^B - b_2^B\big(s^B - \mathbb{E}s^B\big)\big]^2, \text{ where}$$

$$\beta_1^B = \mathbb{E}\big[\theta^B(X_i)\big] \text{ and } \beta_2^B = \frac{cov\big(\theta^B(X_i), s^B(X_i)\big)}{Var\big(s^B(X_i)\big)}.$$

Proof. The moment equations, which correspond to $\beta := (\beta_1^B, \beta_2^B)'$, are given by

$$\mathbb{E}\big[w(X_i)(Y^B(X_i) - \alpha' X_{1i}^B - \beta' X_{2i}^B)X_{2i}^B\big] = 0 \tag{12.65}$$

We substitute the (proxy of the) outcome variable by the corresponding model defined in (12.59):

$$\mathbb{E}\big[w(X_i)\big[\big(\theta_0^B(X_i) + D_i\theta^B(X_i) + U_i^B\big) - \alpha' X_{1i}^B - \beta' X_{2i}^B)\big]X_{2i}^B\big] = 0. \tag{12.66}$$

By the definitions $X_{1i}^B = X_{1i}^B(X_i)$ and $X_{2i}^B = X_{2i}^B(X_i, D_i) = \big[D_i - p(X_i), (D_i - p(X_i))(s^B - \mathbb{E}s^B)\big]$ and the law of iterated expectations:

$$\mathbb{E}\big[w(X_i)\theta_0^B(X_i)X_{2i}\big] = \mathbb{E}\big[w(X_i)\theta_0^B(X_i)\underbrace{\mathbb{E}[X_{2i}^B|X_i]}_{=0}\big] = 0,$$

$$\mathbb{E}\big[w(X_i)U_i^B X_{2i}\big] = \mathbb{E}\big[w(X_i)\mathbb{E}\underbrace{\big[U_i^B|D_i, X_i\big]}_{=0} X_{2i}^B(D_i, X_i)\big] = 0,$$

$$\mathbb{E}\big[w(X_i)X_{1i}^B X_{2i}^B\big] = \mathbb{E}\big[w(X_i)X_{1i}^B(X_i)\mathbb{E}\underbrace{\big[X_{2i}^B(D_i, X_i)|X_i\big]}\big] = 0.$$

Note that D_i and $s^B(X_i)$ are independent, as D_i comes from the train data, while $s^B(X_i)$ is the proxy of the treatment effect bounds from the estimation sample. Therefore, $\mathbb{E}\big[X_{2i}^B(D_i, X_i)|X_i\big] = \mathbb{E}\big[D_i - p(X_i)|X_i\big]\mathbb{E}\big[s^B - \mathbb{E}s^B|X_i\big] = (p(X_i) - p(X_i))\big[s^B - \mathbb{E}s^B|X_i\big] = 0$. Moreover, Assumption 12.3 guarantees that $\mathbb{E}\big[U_i^B|D_i, X_i\big] = 0$.

The moment equations simplify to:

$$\mathbb{E}\big[w(X_i)\big(D_i\theta^B(X_i) - \beta' X_{2i}\big)X_{2i}\big] = 0. \tag{12.67}$$

Note that

$$\mathbb{E}\big[(D_i - p(X_i))(D_i - p(X_i))|X_i\big] = p(X_i)(1 - p(X_i)) = w^{-1}(X_i),$$

and $s^B = s^B(X_i)$. The components of X_{2i} are orthogonal by the law of iterated expectations:

$$\mathbb{E}w(X_i)(D_i - p(X_i))(D_i - p(X_i))(s^B - \mathbb{E}s^B) = \mathbb{E}(s^B - \mathbb{E}s^B) = 0.$$

Hence the moment equations become:

$$\mathbb{E}\big(w(X_i)\big[D_i\theta^B(X_i) - \beta_1^B(D_i - p(X_i))\big](D_i - p(X_i))\big) = 0,$$
$$\mathbb{E}\big(w(X_i)\big[D_i\theta^B(X_i) - \beta_2^B(D_i - p(X_i))(s^B - \mathbb{E}s^B)\big](D_i - p(X_i))(s^B - \mathbb{E}s^B)\big) = 0.$$

By solving these equations and the law of iterated expectations, we end up with

$$\beta_1^B = \frac{\mathbb{E}w(X_i)D_i\theta^B(X_i)(D_i - p(X_i))}{\mathbb{E}w(X_i)(D_i - p(X_i))^2} = \frac{\mathbb{E}w(X_i)\theta(X_i)w^{-1}(X_i)}{\mathbb{E}w(X_i)w^{-1}(X_i)} = \mathbb{E}\theta(X_i),$$
$$\beta_2^B = \frac{\mathbb{E}w(X_i)D_i\theta^B(X_i)(D_i - p(X_i))(s^B - \mathbb{E}s^B)}{\mathbb{E}w(X_i)(D_i - p(X_i))^2(s^B - \mathbb{E}s^B)^2} = \frac{\mathbb{E}w(X_i)\theta(X_i)w^{-1}(X_i)(s^B - \mathbb{E}s^B)}{\mathbb{E}w(X_i)w^{-1}(X_i)(s^B - \mathbb{E}s^B)^2} = \frac{Cov(\theta^B(X_i), s^B)}{Var(s^B)}.$$

The conclusion follows by noticing that these coefficients also solve the moment equations

$$\mathbb{E}\big[\theta^B(X_i) - \beta_1^B - \beta_2^B(s^B - \mathbb{E}s^B)\big][1, s^B - \mathbb{E}s^B]' = 0,$$

which is a linear projection of $\theta(X_i)$ on the demeaned s^B and the intercept. Hence the parameters β_1^B and β_2^B solve the best linear prediction problem of $\theta^B(X_i)$ for each $B \in L, U$.

■

Identification of the parameters in (12.64) can be based on the corresponding empirical analog, in this setting, weighted OLS:

$$\hat{Y}^B(X_i) = X_{1i}\widehat{\alpha}^B + (D_i - p(X_i))\widehat{\beta}_1^B + (D_i - p(X_i))(s^B - \mathbb{E}s^B)\widehat{\beta}_2^B + \widehat{\varepsilon}_i^B. \tag{12.68}$$

Remark 3. *When $s^B(X_i)$ is a perfect proxy for $\theta^B(X_i)$, then $\beta_2^B = 1$. Typically, because of the sample-splitting and estimation error, $\beta_2^B \neq 1$. To the contrary, if $s^B(X_i)$ is completely noisy, uncorrelated with $\theta^B(X_i)$, then $\beta_2^B = 0$. Moreover, if there is no heterogeneity, that is, $\theta^B(X_i) = \theta^B$ for some constant θ^B, then $\beta_2^B = 0$. Rejecting the hypothesis that $\beta_2^B = 0$, therefore, indicates that there is heterogeneity in $\theta(X_i)$ and $s^B(X_i)$ is a relevant predictor.*

Horvitz-Thompson Transformation

Another strategy for identifying parameters of BLP is based on the Horvitz-Thompson transformation (Horvitz and Thompson, 1952) of the outcome variable:

$$H_i = H_i(D_i, X_i) = \frac{D_i - p(X_i)}{p(X_i)(1 - p(X_i))},$$

where $p(X_i)$ is the probability of being treated (propensity score). In randomized control trials, the transformed outcome variable $Y(X_i)H_i$ provides an unbiased signal about the

treatment effect: $\mathbb{E}\big(Y_i H_i | X_i\big) = \theta(X_i)$. Note that, when $H_i > 0$ (i.e., $D_i = 1$), then $Y_i^L(X_i)H_i \leq Y_i H_i \leq Y_i^U(X_i)H_i$; when $H_i < 0$ (i.e., $D_i = 0$), $Y_i^L(X_i)H_i \geq Y_i H_i \geq Y_i^U(X_i)H_i$. In practice, we transform the proxy of the outcome bounds $\tilde{Y}_i^B(X_i)$. It follows that

$$\text{BLP}\big[Y_i^B(X_i)H_i | s^B(X_i)\big] = \text{BLP}\big[\theta^B(X_i)|s^B(X_i)\big].$$

Similarly to Chernozhukov et al. (2018), we consider the linear projection of the transformed outcomes:

$$Y^B(X_i)H_i = X_{1i}^B H_i \alpha^B + \beta_1^B + \big(s^B - \mathbb{E}s^B\big)\beta_2^B + \widetilde{\varepsilon}_i^B, \quad \mathbb{E}\big(\widetilde{\varepsilon}_i^B \widetilde{X}_i^B\big) = 0, \tag{12.69}$$

where the terms in $X_{1i}^B = [s_0^B(X_i)]$ or $X_{1i}^B = [s_0^B(X_i), s^B(X_i), p(X_i)]$ are present to reduce the noise. The effect of X_{1i}^B on the unbiased signal of the treatment effect is denoted by β_1^B and $s^B = s^B(X_i)$. The effect of s^B on the outcome is β_2^B.

Theorem 12.2. *Consider $x \mapsto s_0^B(x)$ and $x \mapsto s^B(x)$ as fixed maps. Assume, $\mathbb{E}(X_i X_i')$ has a full rank. Moreover, assume, Y_i and X_i have finite second moments, and $Var(s^B(X_i)) > 0$. Then, β_1 and β_2 in (12.69) also solve the best linear predictor problem for the target $\theta^B(X_i)$:*

$$(\beta_1^B, \beta_2^B)' = \arg\min_{b_1^B, b_2^B} \mathbb{E}\big[\theta^B(X_i) - b_1^B - b_2^B\big(s^B - \mathbb{E}s^B\big)\big]^2, \text{ where}$$

$$\beta_1^B = \mathbb{E}\big[\theta^B(X_i)\big] \text{ and } \beta_2^B = \frac{cov\big(\theta^B(X_i), s^B(X_i)\big)}{Var\big(s^B(X_i)\big)}.$$

Proof. Define $\widetilde{X}_{2i}^B = \widetilde{X}_{2i}^B(X_i) = \big[1, (s^B(X_i) - \mathbb{E}(s^B(X_i))\big]$ and $X_{1i}^B = X_{1i}^B(X_i)$. The moment function with $\beta = (\beta_1^B, \beta_2^B)'$ is given as:

$$\mathbb{E}\big[(Y^B(X_i)H_i - \beta_0' X_{1i}^B H_i - \beta' \widetilde{X}_{2i}^B)\widetilde{X}_{2i}^B\big] = 0. \tag{12.70}$$

We substitute the (proxy of the) outcome variable by the corresponding model defined in (12.59):

$$\mathbb{E}\big[\big(\big(\theta_0^B(X_i) + D_i\theta^B(X_i) + U_i^B\big)H_i - \beta_0' X_{1i}^B H_i - \beta' \widetilde{X}_{2i}^B\big)\widetilde{X}_{2i}^B\big] = 0.$$

By the law of iterated expectations, note that

$$\mathbb{E}\big[\theta_0^B(X_i) H_i \widetilde{X}_2^B(X_i)\big] = \mathbb{E}\big[\theta_0^B(X_i) \underbrace{\mathbb{E}(H_i|X_i)}_{=0} \widetilde{X}_2^B(X_i)\big] = 0,$$

$$\mathbb{E}\big[U_i^B H_i X_{2i}^B\big] = \mathbb{E}\big[\underbrace{\mathbb{E}[U_i^B|X_i, D_i]}_{=0} H_i(D_i, X_i)\widetilde{X}_{2i}^B(X_i)\big] = 0,$$

$$\mathbb{E}\big[X_{1i}^B(X_i) H_i \widetilde{X}_{2i}^B\big] = \mathbb{E}\big[X_{1i}^B(X_i) \underbrace{\mathbb{E}(H_i|X_i)}_{=0} \widetilde{X}_{2i}^B(X_i)\big] = 0.$$

Therefore, the moment equations simplify to:

$$\mathbb{E}\big[(\theta^B(X_i)D_iH_i - \beta'\widetilde{X}^B_{2i})\widetilde{X}^B_{2i}\big] = 0.$$

Since 1 and $s^B(X_i) - \mathbb{E}(s^B(X_i))$ are independent, the normal equations above further simplify to:

$$\mathbb{E}\big[\big(\theta^B(X_i)D_iH_i - \beta_1\big)\big] = 0,$$
$$\mathbb{E}\big[\big(\theta^B(X_i)D_iH_i - \beta_2\big(s^B_i(X_i) - \mathbb{E}(s^B_i(X_i))\big)\big(s^B_i(X_i) - \mathbb{E}(s^B_i(X_i))\big)\big] = 0,$$

By the law of iterated expectations:

$$\mathbb{E}\big[D_iH_i|X_i\big] = \frac{p(X_i)(1-p(X_i))}{p(X_i)(1-p(X_i))} = 1.$$

Therefore, the simplified equations are given as:

$$\mathbb{E}\big[\theta^B(X_i) - \beta^B_1\big] = 0,$$
$$\mathbb{E}\big[\big(\theta^B(X_i) - \beta^B_2(s^B(X_i) - \mathbb{E}(s^B(X_i))\big)\big(s^B(X_i) - \mathbb{E}(s^B(X_i))\big)\big] = 0.$$

Hence:

$$\beta^B_1 = \mathbb{E}\big[\theta^B(X_i)\big],$$
$$\beta^B_2 = \frac{cov\big[\theta^B(X_i), s^B(X_i)\big]}{var(s^B(X_i))}.$$

■

Appendix 12.12 additionally illustrates a test for the aggregate monotonic treatment selection and response assumptions.

12.12. A Test for Monotonic Treatment Selection and Response Assumptions

The difficulty of testing monotonicity assumptions lies in the unobserved mean potential outcomes. To reduce uncertainty regarding the validity of these assumptions, this section proposes a test of the monotonicity assumptions of the mean potential outcomes.

Consider the relation of D_i on $Y^B(X_i)$ through a linear model:

$$Y^B(X_i) = \beta^B_0 + D_i\beta^B + X_i\beta^B_x + \varepsilon^B_i, \tag{12.71}$$

where $Y^B(X_i)$ is a proxy for the bounds of the outcome. β^B_0 is an intercept and β^B is the effect of D_i on the upper or the lower bound of Y_i. Moreover, ε^B_i is uncorrelated with D_i

for each $B \in \{L, U\}$. X_i denotes $N \times p$ control variables and β_x is the corresponding $p \times 1$ vector of their influence on outcome. Note that, the further we partition data, the more $Y^B(X_i)$ resembles the observed outcome Y_i. In that case, D_i is correlated with ε_i^B, and the parameters will be biased. Hence, the number of subgroups should be large, and at the same time, $Y^B(X_i)$ should have enough variation to estimate the model.

If $\beta^B = 0$ for each $B \in \{L, U\}$, that implies, either the monotonicity assumptions do not hold, or they hold for the mean potential outcomes, but not for the entire distribution of the outcome variable. If $\beta^L > 0$ and $\beta^U > 0$, on average, bounds increase in the levels of D_i. That means treatment shifts the entire distribution of the outcome variable to the right, therefore, on average, the selection into the treatment can be positive. Figure **??** illustrates this case. To the contrary, when $\beta^L < 0$ and $\beta^U < 0$, on average, the selection into the treatment is negative if the mass of the distribution shifts entirely.

It is worth noting that the test is not designed for the unobserved conditional means, $\mathbb{E}(Y_i(d)|D_i, X_i)$, but the observed bounds $\mathbb{E}(Y^B(X_i)|D_i, X_i)$. Therefore, the results of the test hold for the density of the outcome bounds. Nevertheless, if the expected lower and upper bounds of the outcome significantly shift in the levels of D_i, it can be informative about the mean of the outcome.

12.13. Covariance Matrix

The estimator of $\widehat{\Sigma}$ for the bounds in Section 4.1 are straightforward. Assume, $Y^B(X_i)$ is fixed for $B \in \{L, U\}$. Therefore, $\widehat{\Sigma}$ does not depend on them. Then under the positive selection, the variance and the covariance of the lower and upper bounds are:

$$
\begin{aligned}
Var(\widetilde{\theta}^L) = \frac{Var(Y_i^{(D_i=1)})}{N_\ell^{(D_i=1)}} \cdot P(D_i = 1|X_i = x)^2 + \\
\frac{Var(Y_i^{(D_i=0)})}{N_\ell^{(D_i=0)}} \cdot P(D_i = 0|X_i = x)^2, \\
Var(\widetilde{\theta}^U) = \frac{Var(Y_i^{(D_i=1)})}{N_\ell^{(D_i=1)}} + \frac{Var(Y_i^{(D_i=0)})}{N_\ell^{(D_i=0)}}, \\
Cov(\widetilde{\theta}^L, \widetilde{\theta}^U) = \frac{Var(Y_i^{(D_i=0)})}{N_\ell^{(D_i=1)}} \cdot P(D_i = 1|X_i = x) + \\
\frac{Var(Y_i^{(D_i=0)})}{N_\ell^{(D_i=0)}} \cdot P(D_i = 0|X_i = x).
\end{aligned}
$$

The estimators can be derived equivalently when the selection into the treatment is negative.

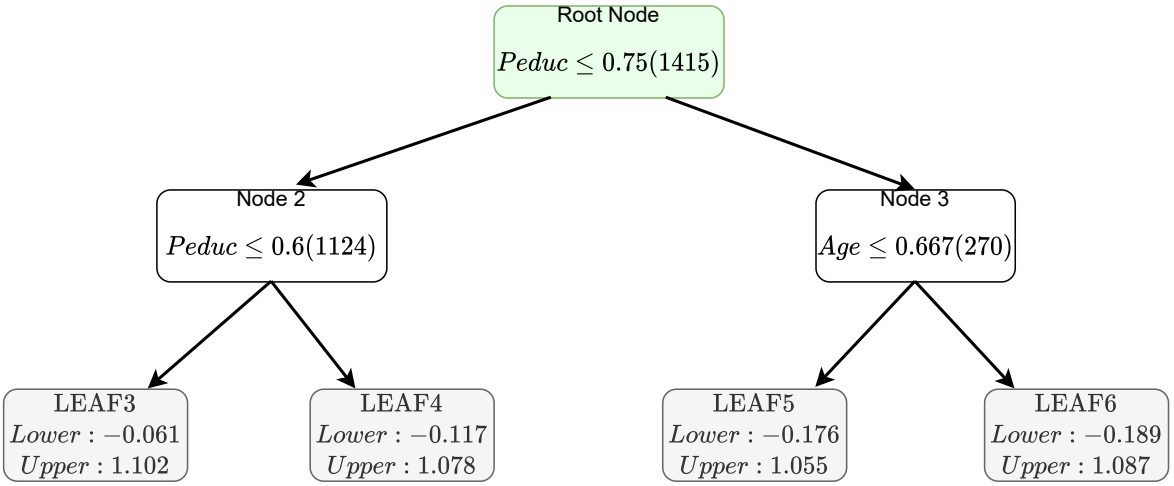


Figure 9. A single (randomly chosen) tree out of 200 trees. The bounds are not informative, however, the tree successfully recovers parental education, the variable that reflects the heterogeneity in the Head Start participation effect on the outcome.

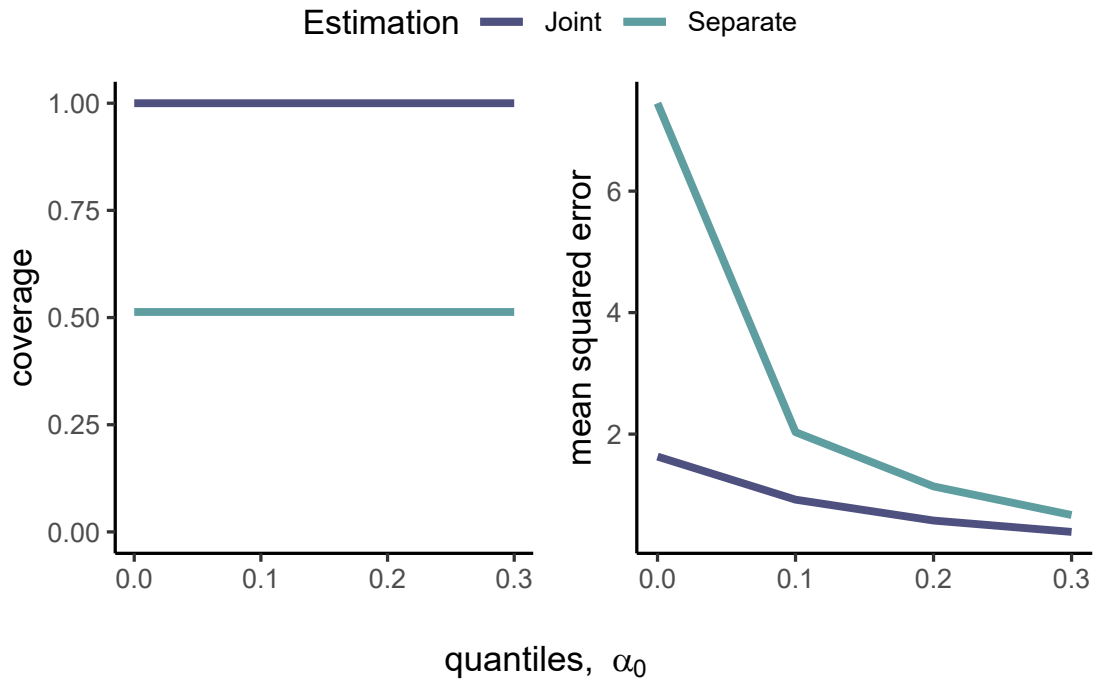


Figure 10. Coverage and the mean squared error for four different quantiles: $\{\alpha_1 = 1 \text{ and } \alpha_0 = 0\}$, $\{\alpha_1 = 0.9 \text{ and } \alpha_0 = 0.1\}$, $\{\alpha_1 = 0.8 \text{ and } \alpha_0 = 0.2\}$, $\{\alpha_1 = 0.7 \text{ and } \alpha_0 = 0.3\}$. Joint estimation of bounds is based on the multivariate random forests where the direction of the treatment selection can vary according to the covariates. Separate estimation of the bounds is based on causal forests (Athey and Wager, 2019) where for each tree the treatment selection is fixed to negative.

We also quantify the statistical significance of the heterogeneity captured in the bounds (based on Algorithm 2). Table III shows the effects of the demeaned Head Start on schooling (β_1^B) and its' interaction with the demeaned bounds of the treatment effects (β_2^B). β_1^B reflects the expected population treatment effect bound, while β_2^B captures the covariance between the estimated and population treatment effect bounds. The effects of the interaction terms on the lower and upper bounds are negative and statistically significant. This implies that the estimated bounds carry important information regarding the subsets of the population.

Table III. A test for heterogeneous treatment effect bounds. The outcome variables are the lower and upper bounds of the outcome (based on 200 trees with a maximum depth of 10). Standard errors are given in parentheses. Figure 11 (Appendix 12.15) shows the density of the outcome bounds.

	Dependent variable:	
	$Y^L(X_i)$	$Y^U(X_i)$
	(1)	(2)
Demeaned Head Start	-0.005^{***}	0.005^{***}
	(0.001)	(0.001)
Interaction^L	-0.208^{***}	
	(0.046)	
Interaction^U		-0.132^{***}
		(0.031)
Constant	7.990^{***}	19.123^{***}
	(0.005)	(0.005)
Observations	932	932
R^2	0.034	0.026
Adjusted R^2	0.032	0.024
Residual Std. Error (df = 929)	0.597	0.645
F Statistic (df = 2; 929)	16.250^{***}	12.383^{***}
Note:	*p<0.1; **p<0.05; ***p<0.01	

Coverage, Informativeness, and Heterogeneity of the Bounds

In this section, we illustrate the coverage, informativeness, and heterogeneity of the treatment effect bounds. The heterogeneity of the bounds is measured in terms of their standard deviation. Table IV illustrates the mean squared error (left) and coverage (right) for different values of the bottom quantiles of the outcome. Specifically, instead of the minimum and maximum values of the outcome, the bounds of treatment effects are based on the bottom and top quantiles of the outcome ($\alpha_0 \in [0, 0.3]$, $\alpha_1 \in [0.7, 1]$, respectively). In each experiment, the values of the error terms change.

Table IV. The mean squared error and coverage of the bounds for 100 values of the upper quantile of the outcome (α_1) uniformly distributed between 1 and 0.7, and the lower quantile of the outcome (α_0) uniformly distributed between 0 and 0.3. The errors are different across experiments. CF corresponds to the causal forest method with the loss function in Athey and Imbens (2016), MCF corresponds to the multivariate random forests with the loss function in Proposition 6.2, MCF diagonal is a multivariate random forest when the variance-covariance is a diagonal matrix (covariance terms are equal to zero).

	MSE			Coverage		
α_0	**MCF**	CF	MCF diagonal	**MCF**	CF	MCF diagonal
0.000	2.966	3.870	3.322	1.000	1.000	1.000
0.031	2.410	2.861	2.509	1.000	0.750	1.000
0.061	1.713	1.964	1.756	0.992	0.628	0.901
0.092	1.241	1.267	1.210	0.976	0.874	0.956
0.122	0.988	0.955	0.954	0.999	0.874	0.983
0.153	0.860	0.839	0.844	0.955	0.743	0.898
0.184	0.750	0.742	0.736	0.828	0.520	0.714
0.214	0.684	0.686	0.669	0.737	0.502	0.652
0.245	0.592	0.601	0.581	0.532	0.426	0.488
0.276	0.525	0.521	0.511	0.372	0.284	0.320
0.300	0.465	0.445	0.452	0.077	0.147	0.075

According to Table IV, multivariate random forests rescaled by the covariance matrix perform the best. One plausible explanation is that the loss function in Athey and Imbens (2016) does not incorporate the asymmetric mean absolute deviation loss. The findings based on Table IV indicate that the combination of the mean squared error (rescaled by the covariance) and the mean absolute deviation loss substantially increases the predictive power.

Table V. Standard deviation (std) of the lower and upper bounds of the treatment effects (y-axis) for various values of α_1 uniformly distributed within 1 and 0.7, and simultaneously α_0 uniformly distributed between 0 and 0.3. The errors are different across experiments. CF corresponds to causal forests with the loss function in Athey and Imbens (2016), MCF corresponds to the multivariate random forests with the loss function proposed in this article (Proposition 6.2), MCF diagonal is multivariate random forests where the variance-covariance (Σ) in the loss is a diagonal matrix (covariance terms are equal to zero).

	Std Upper			Std Lower		
α_0	**MCF**	CF	MCF diagonal	**MCF**	CF	MCF diagonal
0.000	0.091	0.020	0.048	0.101	0.068	0.093
0.031	0.093	0.065	0.090	0.083	0.051	0.074
0.061	0.068	0.058	0.074	0.052	0.041	0.059
0.092	0.057	0.040	0.056	0.051	0.034	0.047
0.122	0.041	0.031	0.041	0.038	0.024	0.039
0.153	0.041	0.030	0.039	0.037	0.024	0.038
0.184	0.041	0.030	0.043	0.032	0.021	0.033
0.214	0.038	0.032	0.038	0.030	0.021	0.031
0.245	0.035	0.030	0.037	0.024	0.019	0.024
0.276	0.033	0.027	0.033	0.020	0.013	0.022
0.300	0.028	0.023	0.029	0.022	0.017	0.022

To measure the heterogeneity of these bounds, Table V illustrates the standard deviation for various quantiles of the simulated outcome variable. Table V shows that the heterogeneity of the bounds is the highest when the outcome takes minimum and maximum values ($\alpha_0 = 0$

and $\alpha_1 = 1$). Moreover, based on Table V, the multivariate random forest captures the highest heterogeneity in the treatment effect bounds. In addition, Figure 10 in Appendix 12.14 shows that the flexible monotonic treatment selection assumption has 50% higher coverage and a lower mean squared error relative to the method with the fixed negative treatment selection across all dimensions of the covariate space.

12.15. Applications to National Longitudinal Study of Youth 1979

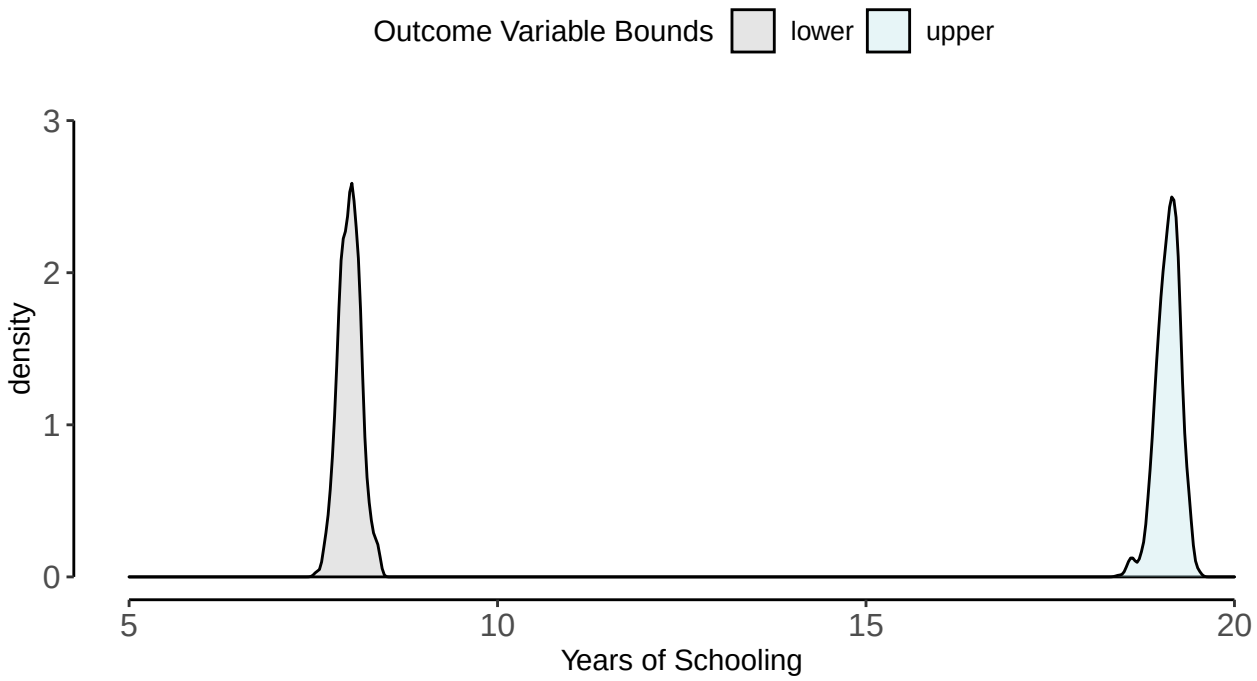


Figure 11. Estimated outcome variable bounds. The random forest consists of 200 trees, with the maximum depth of 10 for each tree.

Table VI. A test of for the monotonicity assumptions for the Head Start participation. Control variables include parental education, age and Hispanic. Y^U and Y^L are the upper and lower bounds of the years of schooling, respectively.

	Dependent variable:	
	Y^U	Y^L
	(1)	(2)
Head Start	-0.025^{**}	-0.022^{*}
	(0.013)	(0.012)
Controls	✓	✓
Constant	18.843^{***}	8.268^{***}
	(0.023)	(0.021)
Observations	932	932
Adjusted R^2	0.180	0.162
Residual Std. Error (df = 927)	0.147	0.136
F Statistic (df = 4; 927)	52.059^{***}	46.143^{***}
Note:	*p<0.1; **p<0.05; ***p<0.01	

13. Extensions

13.1. Network Data

Consider two different nodes as two groups of agents that are allowed to collaborate. Each node is associated with a specific probability of collaborating with each other. Goel et al. (2014) discuss a similar setting with a credit network and the liquidity of the agents. In that case, the parameters of interest are the treatment effect bounds, aggregated across specific leaves:

$$\int_{\ell=1}^{L} w(x)\theta_{\ell}^{B}(x),$$

where $w(x)$ is a given weight at a test point x, and $\theta_{\ell}^{B}(x)$ is a bound for $B \in \{L, U\}$.

Denote two different leaves as ℓ and ℓ'. The aim is to show that

$$\frac{s}{N} tr\big(Var(\dot{\widetilde{T}})^{-1} Var(\widetilde{T})\big) \to 0.$$

Consider the Hajek projection of a tree

$$\dot{\widetilde{T}} - \mu = \sum_{i=1}^{s} \mathbb{E}(\widetilde{T}|A_i), \text{ so that } Var(\dot{\widetilde{T}}) = sVar(\mathbb{E}(\widetilde{T}|A_1).$$

The last equality follows as the data $A_i = (Y_i, X_i)$ are *i.i.d.* Moreover, the variance of the conditional expected tree can be expanded as:

$$Var\mathbb{E}(\widetilde{T}|A_1) = Var\big[\mathbb{E}(\widetilde{T}|A_1) - \mathbb{E}(\widetilde{T}|X_1)\big] + Var\big[\mathbb{E}(\widetilde{T}|X_1)\big].$$

The algorithm is honest, therefore, the difference $\mathbb{E}(\widetilde{T}|A_1) - \mathbb{E}(\widetilde{T}|X_1)$ simplifies to

$$\mathbb{E}(\widetilde{T}_\ell|A_1) - \mathbb{E}(\widetilde{T}_\ell|X_1) = \mathbb{E}(S_\ell|X_1)(Y_{m1} - \mathbb{E}(Y_{m1}|S_\ell = 1, X_1)).$$

$\widetilde{T}_\ell$ is a tree estimate at x_ℓ and S_ℓ is an indicator of whether X_1 and x_ℓ belong to the same terminal node. The outcome Y_{m1} denotes the m-th outcome for the first observation. Note that the covariance matrix in each tree now consists of not only the covariance between the outcomes within a leaf but across terminal nodes as well. Therefore, we focus on the covariance of an outcome from a leaf ℓ with the same outcome from another leaf ℓ' and with another outcome $Y_{m'}$ from another leaf ℓ':

$$\mathbb{E}\big[\mathbb{E}(S_\ell|X_1)\mathbb{E}(S_{\ell'}|X_1)(Y_{m1} - \mathbb{E}(Y_{m1}|S_\ell = 1, X_1))(Y_{m1} - \mathbb{E}(Y_{m1}|S_{\ell'} = 1, X_1))\big],$$
$$\mathbb{E}\big[\mathbb{E}(S_\ell|X_1)\mathbb{E}(S_{\ell'}|X_1)(Y_{m1} - \mathbb{E}(Y_{m1}|S_\ell = 1, X_1))(Y_{m'1} - \mathbb{E}(Y_{m'1}|S_{\ell'} = 1, X_1))\big].$$

The terms $\mathbb{E}(Y_{m1} - \mathbb{E}(Y_{m1}|S_\ell = 1, X_1))(Y_{m1} - \mathbb{E}(Y_{m1}|S_{\ell'} = 1, X_1))$ and $\mathbb{E}(Y_{m1} - \mathbb{E}(Y_{m1}|S_\ell = 1, X_1))(Y_{m'1} - \mathbb{E}(Y_{m'1}|S_{\ell'} = 1, X_1))$ are the polynomials of degree at most two. Since the first and second moments of the outcome are bounded, these terms are also bounded. Next, Cauchy-Schwarz inequality implies

$$\sqrt{\mathbb{E}\big[\mathbb{E}(S_\ell|X_1)\big]^2\mathbb{E}\big[\mathbb{E}(S_{\ell'}|X_1)\big]^2} \leq \sqrt{\mathbb{E}\big[\mathbb{E}(S_\ell|X_1)\mathbb{E}(S_{\ell'}|X_1)}.$$

Therefore, the lower bound of the covariance is on the order $o\left(\frac{1}{log^p(s)}\right)$. Equivalently, we can show that the off-diagonal terms of $Var\big[\mathbb{E}(\widetilde{T}|X_1)\big]$ are on the same order. Then Theorem 6.3 applies, and the asymptotic normality of the parameter of interest holds.

13.2. Anomalous Data

A useful extension of the method is to separate treatment effect bounds that stem from non-anomalous observations x^o from the ones that stem from anomalous observations x^a.

Yadlowsky et al. (2021) discuss the rank-weighted average treatment effect and the targeting operator characteristic (TOC). In an RCT design,

$$TOC(u, S) = \mathbb{E}\big[(Y_i(1) - Y_i(0)|F_S(S(X_i)) \geq 1 - u\big] - \mathbb{E}\big[(Y_i(1) - Y_i(0)\big],$$

where $F_S(S(X_i))$ is a cumulative distribution function of a priority score $S(X_i)$. Intuitively, $TOC(u, S)$ measures the priority of the treatment effects where the priority is reflected in a score that depends on the covariates. The "prioritization rule" $S(X_i)$ is a score that allows a policymaker to rank the treatment effects and draw final implications for policymaking.

In this study, the quantity of interest is the bounds of the treatment effect, conditional on the priority rule

$$TOC^B(u, S) = \mathbb{E}(\theta^B(X_i)|F_S(S(X_i)) \geq 1 - u) - \mathbb{E}(\theta^B(X_i)), \tag{13.1}$$

where for $B \in \{L, U\}$, $\theta^B(X_i)$ is a bound of the treatment effects.

A key idea is to introduce the anomaly score $S(X_i)$. Liu et al. (2008) discuss anomaly detection with random forests. In particular, with a randomly chosen splitting coordinate and value, anomalous covariates have a shorter average path length among the trees compared to non-anomalous observations. The anomaly score is given by

$$S(X_i) = 2^{-\frac{\mathbb{E}(h(x))}{c(n)}}, \tag{13.2}$$

where $\mathbb{E}(h(x))$ is the average path length from a collection of trees. The path length $h(x)$ is measured by the number of nodes x traverses from the root node, until the terminal node. The average path length for a single tree $c(n)$ is defined as an unsuccessful search in a Binary Search Tree:

$$c(n) = 2H(n-1) - (2(n-1)/n),$$

where $H(i)$ is the Harmonic number and can be estimated as $ln(i) + 0.5772156649$ (Euler's constant).

Note that s is monotonic in $\mathbb{E}(h(x))$. In particular

$$\begin{aligned}
&\text{when } \mathbb{E}(h(x)) \mapsto c(n),\ S(x) \mapsto 0.5;\\
&\text{when } \mathbb{E}(h(x)) \mapsto 0,\ S(x) \mapsto 1;\\
&\text{and when } \mathbb{E}(h(x)) \mapsto n-1,\ S(x) \mapsto 0.
\end{aligned}$$

When the score $S(x)$ is close to 1, the instance is an anomaly. If the score of the instance is weakly less than 0.5, it is a representative point. For further details see Liu et al. (2008).

In some cases, a policy maker is only interested in the treatment effect (bounds) that stem only from the representative sample. In that case, we can redefine the tree

$$\mathcal{T}(x, \xi, A_1, \ldots, A_n) = \sum_{n=1}^{|\Pi|} \mathbf{1}(x \in \ell_n)\mathbf{1}(S(x) \leq 0.5)\frac{1}{N_{l_n}} \sum_{i:X_i \in \ell_n} Y_i, \tag{13.3}$$

where $\mathbf{1}(S(x) \leq 0.5)$ is a binary variable and indicates that x is an ordinary, representative covariate. The aim is to show that the asymptotic normality holds for such a tree.

The expected difference between the tree conditional on data A_1 of the first observation, and the covariate X_1 is defined as

$$\mathbb{E}(\widetilde{T}|A_1) - \mathbb{E}(\widetilde{T}|X_1) = \mathbb{E}(I_\ell|X_1)\mathbb{E}(S_\ell|X_1)(Y_{m1} - \mathbb{E}(Y_{m1}|S_\ell = 1, X_1)),$$

where $\mathbb{E}(I_\ell|X_1)$ equals one if the score $S(x)$ is weakly less than 0.5, and zero otherwise. After disregarding the terms that are too small, the variance becomes

$$Var\big[\mathbb{E}(\widetilde{T}|A_1) - \mathbb{E}(\widetilde{T}|X_1)\big] = \mathbb{E}\big[\mathbb{E}(I_\ell|X_1)\big]^2\mathbb{E}\big[\mathbb{E}(S_\ell|X_1)\big]^2 Var(Y_m).$$

As before, $\mathbb{E}\big[\mathbb{E}(S_\ell|X_1)\big]^2$ is $o\left(\frac{1}{\log^p s}\right)$. The upper bound of $\big[\mathbb{E}(I_\ell|X_1)\big]^2$ is one as $S(x)$ is bounded between [0, 1]. A similar result holds for the covariance terms in a multivariate setup. Hence, Theorem 6.3 applies in this setting as well.