Estimating Treatment Effects with Causal Forests: An Application

Susan Athey
athey@stanford.edu

Stefan Wager
swager@stanford.edu

Stanford University

Abstract

We apply causal forests to a dataset derived from the National Study of Learning Mindsets, and consider resulting practical and conceptual challenges. In particular, we discuss how causal forests use estimated propensity scores to be more robust to confounding, and how they handle data with clustered errors.

1 Methodology and Motivation

There has been considerable recent interest in methods for heterogeneous treatment effect estimation in observational studies (Athey and Imbens, 2016; Athey, Tibshirani, and Wager, 2019; Ding, Feller, and Miratrix, 2016; Dorie, Hill, Shalit, Scott, and Cervone, 2017; Hahn, Murray, and Carvalho, 2017; Hill, 2011; Imai and Ratkovic, 2013; Künzel, Sekhon, Bickel, and Yu, 2017; Luedtke and van der Laan, 2016; Nie and Wager, 2017; Shalit, Johansson, and Sontag, 2017; Su, Tsai, Wang, Nickerson, and Li, 2009; Wager and Athey, 2018; Zhao, Small, and Ertefaie, 2017). In order to help elucidate the drivers of successful approaches to treatment effect estimation, Carlos Carvalho, Jennifer Hill, Avi Feller and Jared Murray organized a workshop at the 2018 Atlantic Causal Inference Conference and asked several authors to analyze a shared dataset derived from the National Study of Learning Mindsets (Yeager et al., 2016).

This note presents an analysis using causal forests (Athey, Tibshirani, and Wager, 2019; Wager and Athey, 2018); other approaches will be discussed in a forthcoming issue of *Observational Studies* with title "Empirical Investigation of Methods for Heterogeneity." All analyses are carried out using the `R` package `grf`, version `0.10.2` (Tibshirani et al., 2018; R Core Team, 2017). Full replication files are available at `github.com/grf-labs/grf`, in the directory `experiments/acic18`.

1.1 The National Study of Learning Mindsets

The National Study of Learning Mindsets is a randomized study conducted in U.S. public high schools, the purpose of which was to evaluate the impact of a nudge-like intervention designed to instill students with a growth mindset[1] on student achievement. To protect student privacy, the present analysis is not based on data from the original study, but rather on data simulated from a model fit to the National Study dataset by the workshop organizers. The present analysis could serve as a pre-analysis plan to be applied to the original National Study dataset (Nosek et al., 2015).

[1] According to the National Study, "A growth mindset is the belief that intelligence can be developed. Students with a growth mindset understand they can get smarter through hard work, the use of effective strategies, and help from others when needed. It is contrasted with a fixed mindset: the belief that intelligence is a fixed trait that is set in stone at birth."

`S3`	Student's self-reported expectations for success in the future, a proxy for prior achievement, measured prior to random assignment
`C1`	Categorical variable for student race/ethnicity
`C2`	Categorical variable for student identified gender
`C3`	Categorical variable for student first-generation status, i.e. first in family to go to college
`XC`	School-level categorical variable for urbanicity of the school, i.e. rural, suburban, etc.
`X1`	School-level mean of students' fixed mindsets, reported prior to random assignment
`X2`	School achievement level, as measured by test scores and college preparation for the previous 4 cohorts of students
`X3`	School racial/ethnic minority composition, i.e., percentage of student body that is Black, Latino, or Native American
`X4`	School poverty concentration, i.e., percentage of students who are from families whose incomes fall below the federal poverty line
`X5`	School size, i.e., total number of students in all four grade levels in the school
Y	Post-treatment outcome, a continuous measure of achievement
W	Treatment, i.e., receipt of the intervention

Table 1: Definition of variables measured in the National Study of Learning Mindsets

Our analysis is based on data from $n = 10,391$ children from a probability sample of $J = 76$ schools.[2] For each child $i = 1, ..., n$, we observe a binary treatment indicator W_i, a real-valued outcome Y_i, as well as 10 categorical or real-valued covariates described in Table 1. We expanded out categorical random variables via one-hot encoding, thus resulting in covariates $X_i \in \mathbb{R}^p$ with $p = 28$. Given this data, the workshop organizers expressed particular interest in the three following questions:

1. Was the mindset intervention effective in improving student achievement?

2. Was the effect of the intervention moderated by school level achievement (`X2`) or pre-existing mindset norms (`X1`)? In particular there are two competing hypotheses about how `X2` moderates the effect of the intervention: Either it is largest in middle-achieving schools (a "Goldilocks effect") or is decreasing in school-level achievement.

3. Do other covariates moderate treatment effects?

We define causal effects via the potential outcomes model (Imbens and Rubin, 2015): For each sample i, we posit potential outcomes $Y_i(0)$ and $Y_i(1)$ corresponding to the outcome we would have observed had we assigned control or treatment to the i-th sample, and assume that we observe $Y_i = Y_i(W_i)$. The average treatment effect is then defined as $\tau = \mathbb{E}\left[Y_i(1) - Y_i(0)\right]$, and the conditional average treatment effect function is $\tau(x) = \mathbb{E}\left[Y_i(1) - Y_i(0) \,\middle|\, X_i = x\right]$.

This dataset exhibits two methodological challenges. First, although the National Study itself was a randomized study, there seems to be some selection effects in the synthetic data used here. As seen in Figure 1, students with a higher expectation of success appear to be more likely to receive treatment. For this reason, we analyze the study as an observational rather than randomized study. In order to identify causal effects, we assume unconfoundedness, i.e., that treatment assignment is as

[2]Initially, 139 schools were recruited into the study using a stratified probability sampling method (Gopalan and Tipton, 2018). Of these 139 recruited schools, 76 agreed to participate in the study; then, students were individually randomized within the participating schools. In this note, we do not discuss potential bias from the non-randomized selection of 76 schools among the 139 recruited ones.

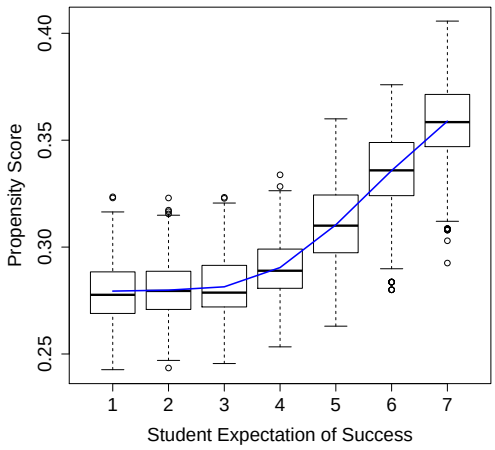


Figure 1: Visualizing estimated treatment propensities against student expectation of success.

good as random conditionally on covariates (Rosenbaum and Rubin, 1983)

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

To relax this assumption, one could try to find an instrument for treatment assignment (Angrist and Pischke, 2008), or conduct a sensitivity analysis for hidden confounding (Rosenbaum, 2002).

Second, the students in this study are not independently sampled; rather, they are all drawn from 76 randomly selected schools, and there appears to be considerable heterogeneity across schools. Such a situation could arise if there are unobserved school-level features that are important treatment effect modifiers; for example, some schools may have leadership teams who implemented the intervention better than others, or may have a student culture that is more receptive to the treatment. If we want our conclusions to generalize outside of the 76 schools we ran the experiment in, we must run an analysis that robustly accounts for the sampling variability of potentially unexplained school-level effects. Here, we take a conservative approach, and assume that the outcomes Y_i of students within a same school may be arbitrarily correlated within a school (or "cluster"), and then apply cluster-robust analysis tools (Abadie, Athey, Imbens, and Wooldridge, 2017).

The rest of this section presents a brief overview of causal forests, with an emphasis of how they address issues related to clustered observations and selection bias. Causal forests are an adaptation of the random forest algorithm of Breiman (2001) to the problem of heterogeneous treatment effect estimation. For simplicity, we start below by discussing how to make random forests cluster-robust in the classical case of non-parametric regression, where we observe pairs (X_i, Y_i) and want to estimate $\mu(x) = \mathbb{E}\left[Y_i \,\middle|\, X_i = x\right]$. Then, in the next section, we review how forests can be used for treatment effect estimation in observational studies.

1.2 Cluster-Robust Random Forests

When observations are grouped in unevenly sized clusters, it is important to carefully define the underlying target of inference. For example, in our setting, do we want to fit a model that accurately reflects heterogeneity in our available sample of $J = 76$ schools, or one that we hope will generalize to students from other schools also? Should we give more weight in our analysis to schools from which we observe more students?

Here, we assume that we want results that generalize beyond our J schools, and that we give each school equal weight; quantitatively, we want models that are accurate for predicting effects on

a new student from a new school. Thus, if we only observed outcomes Y_i for students with school membership $A_i \in \{1, ..., J\}$ we would estimate the global mean as $\hat{\mu}$ with standard error $\hat{\sigma}$, with

$$\hat{\mu}_j = \frac{1}{n_j} \sum_{\{i:A_i=j\}} Y_i, \quad \hat{\mu} = \frac{1}{J} \sum_{j=1}^{J} \hat{\mu}_j, \quad \hat{\sigma}^2 = \frac{1}{J(J-1)} \sum_{j=1}^{J} (\hat{\mu}_j - \hat{\mu})^2, \tag{2}$$

where n_j denotes the number of students in school j. Our challenge is then to use random forests to bring covariates into an analysis of type (2). Formally, we seek to carry out a type of non-parametric random effects modeling, where each school is assumed to have some effect on the student's outcome, but we do not make assumptions about its distribution (in particular, we do not assume that school effects are Gaussian or additive).

At a high level, random forests make predictions as an average of b trees, as follows: (1) For each $b = 1, ..., B$, draw a subsample $\mathcal{S}_b \subseteq \{1, ..., n\}$; (2) Grow a tree via recursive partitioning on each such subsample of the data; and (3) Make predictions

$$\hat{\mu}(x) = \frac{1}{B} \sum_{b=1}^{B} \sum_{i=1}^{n} \frac{Y_i \, \mathbf{1}\left(\{X_i \in L_b(x),\, i \in \mathcal{S}_b\}\right)}{|\{i : X_i \in L_b(x),\, i \in \mathcal{S}_b\}|}, \tag{3}$$

where $L_b(x)$ denotes the leaf of the b-th tree containing the training sample x. In the case of out-of-bag prediction, we estimate $\hat{\mu}^{(-i)}(X_i)$ by only considering those trees b for which $i \notin \mathcal{S}_b$. This short description of forests of course leaves many details implicit. We refer to Biau and Scornet (2016) for a recent overview of random forests and note that, throughout, all our forests are "honest" in the sense of Wager and Athey (2018).

When working with clustered data, we adapt the random forest algorithm as follows. In step (1), rather than directly drawing a subsample of observations, we draw a subsample of clusters $\mathcal{J}_b \subseteq \{1, ..., J\}$; then, we generate the set $\mathcal{S}_b$ by drawing k samples at random from each cluster $j \in \mathcal{J}_b$.[3] The other point where clustering matters is when we want to make out-of-bag predictions in step (3). Here, to account for potential correlations within each cluster, we only consider an observation i to be out-of-bag if its cluster was not drawn in step (1), i.e., if $A_i \notin \mathcal{J}_b$.

1.3 Causal Forests for Observational Studies

One promising avenue to heterogeneous treatment effect estimation starts from an early result of Robinson (1988) on inference in the partially linear model (Nie and Wager, 2017; Zhao, Small, and Ertefaie, 2017). Write $e(x) = \mathbb{P}\left[W_i \,\middle|\, X_i = x\right]$ for the propensity score and $m(x) = \mathbb{E}\left[Y_i \,\middle|\, X_i = x\right]$ for the expected outcome marginalizing over treatment. If the conditional average treatment effect function is constant, i.e., $\tau(x) = \tau$ for all $x \in \mathcal{X}$, then the following estimator is semiparametrically efficient for τ under unconfoundedness (1) (Chernozhukov et al., 2018a; Robinson, 1988):

$$\hat{\tau} = \frac{\frac{1}{n} \sum_{i=1}^{n} \left(Y_i - \hat{m}^{(-i)}(X_i)\right) \left(W_i - \hat{e}^{(-i)}(X_i)\right)}{\frac{1}{n} \sum_{i=1}^{n} \left(W_i - \hat{e}^{(-i)}(X_i)\right)^2}, \tag{4}$$

assuming that $\hat{m}$ and $\hat{e}$ are $o(n^{-1/4})$-consistent for m and e respectively in root-mean-squared error, that the data is independent and identically distributed, and that we have overlap, i.e., that propensities $e(x)$ are uniformly bounded away from 0 and 1. The $^{(-i)}$-superscripts denote "out-of-bag" or "out-of-fold" predictions meaning that, e.g., Y_i was not used to compute $\hat{m}^{(-i)}(X_i)$.

[3] If $k \leq n_j$ for all $j = 1, ..., J$, then each cluster contributes the same number of observations to the forest as in (2). In `grf`, however, we also allow users to specify a value of k larger than the smaller n_j; and, in this case, for clusters with $n_j \leq k$, we simply use the whole cluster (without duplicates) every time $j \in \mathcal{J}_b$. This latter option may be helpful in cases where there are some clusters with a very small number of observations, yet we want $\mathcal{S}_b$ to be reasonably large so that the tree-growing algorithm is stable.

Although the original estimator (4) was designed for constant treatment effect estimation, Nie and Wager (2017) showed that we can use it to motivate an "R-learner" objective function for heterogeneous treatment effect estimation,

$$\hat{\tau}(\cdot) = \text{argmin}_\tau \left\{ \sum_{i=1}^{n} \left(\left(Y_i - \hat{m}^{(-i)}(X_i) \right) - \tau(X_i) \left(W_i - \hat{e}^{(-i)}(X_i) \right) \right)^2 + \Lambda_n \left(\tau(\cdot) \right) \right\}, \tag{5}$$

where $\Lambda_n \left(\tau(\cdot) \right)$ is a regularizer that controls the complexity of the learned $\hat{\tau}(\cdot)$ function. A desirable property of this approach is that, if the true conditional average treatment effect function $\tau(\cdot)$ is simpler than the main effect function $m(\cdot)$ or the propensity function $e(\cdot)$, e.g., qualitatively, if $\tau(\cdot)$ allows for a sparser representation than $m(\cdot)$ or $e(\cdot)$, then the function $\hat{\tau}(\cdot)$ learned by optimizing (5) may converge faster than the estimates for $\hat{m}(\cdot)$ or $\hat{e}(\cdot)$ used to form the objective function.

Causal forests as implemented in `grf` can be seen as a forest-based method motivated by the R-learner (5). Typically, random forests (Breiman, 2001) are understood as an ensemble method: A random forest prediction is an average of predictions made by individual trees. However, as discussed in Athey, Tibshirani, and Wager (2019), we can equivalently think of random forests as an adaptive kernel method; for example, we can re-write the regression forest from (3) as

$$\hat{\mu}(x) = \sum_{i=1}^{n} \alpha_i(x) Y_i, \quad \alpha_i(x) = \frac{1}{B} \sum_{b=1}^{B} \frac{\mathbf{1}\left(\{X_i \in L_b(x),\, i \in \mathcal{S}_b\} \right)}{\left| \{i : X_i \in L_b(x),\, i \in \mathcal{S}_b\} \right|}, \tag{6}$$

where, qualitatively, $\alpha_i(x)$ is a data-adaptive kernel that measures how often the i-th training example falls in the same leaf as the test point x. This kernel-based perspective on forests suggests a natural way to use them for treatment effect estimation based on (4) and (5): First, we grow a forest to get weights $\alpha_i(x)$, and then set

$$\hat{\tau} = \frac{\sum_{i=1}^{n} \alpha_i(x) \left(Y_i - \hat{m}^{(-i)}(X_i) \right) \left(W_i - \hat{e}^{(-i)}(X_i) \right)}{\sum_{i=1}^{n} \alpha_i(x) \left(W_i - \hat{e}^{(-i)}(X_i) \right)^2}. \tag{7}$$

Athey, Tibshirani, and Wager (2019) discuss this approach in more detail, including how to design a splitting rule for a forest that will be used to estimate predictions via (7). Finally, we address clustered observations by modifying the random forest sampling procedure in an analogous way to the one used in Section 1.2.

Concretely, the `grf` implementation of causal forests starts by fitting two separate regression forests to estimate $\hat{m}(\cdot)$ and $\hat{e}(\cdot)$. It then makes out-of-bag predictions using these two first-stage forests, and uses them to grow a causal forest via (7). Causal forests have several tuning parameters (e.g., minimum node size for individual trees), and we choose those tuning parameters by cross-validation on the R-objective (5), i.e., we train causal forests with different values of the tuning parameters, and choose the ones that make out-of-bag estimates of the objective minimized in (5) as small as possible.

We provide an exact implementation of our treatment effect estimation strategy with causal forests in Algorithm 1. We train the `Y.forest` and `W.forest` using default settings, as their predictions are simply used as inputs to the causal forest and default parameter choices often perform reasonably well with random forests.[4] For our final causal forest, however, we deploy some tweaks for improved precision. Motivated by Basu, Kumbier, Brown, and Yu (2018), we start by training a pilot random forest on all the features, and then train a second forest on only those features that saw a reasonable number of splits in the first step.[5] This enables the forest to make more splits on the most important

[4] The nuisance components `Y.hat` or `W.hat` need not be estimated by a regression forest. We could also use other predictive methods (e.g., boosting with cross-fitting) or use oracle values (e.g., the true randomization probabilities for `W.hat` in a randomized trial). If we simply run the command `causal_forest(X, Y, W)` without specifying `Y.hat` or `W.hat`, then the software silently estimates `Y.hat` or `W.hat` via regression forests.

[5] Given good estimates of `Y.hat` and `W.hat`, the construction (7) eliminates confounding effects. Thus, we do not need to give the causal forest all features X that may be confounders. Rather, we can focus on features that we believe may be treatment modifiers; see Zhao, Small, and Ertefaie (2017) for a further discussion.

Algorithm 1 Estimating treatment effects with causal forests.

```
Y.forest = regression_forest(X, Y, clusters = school.id)
Y.hat = predict(Y.forest)$predictions
W.forest = regression_forest(X, W, clusters = school.id)
W.hat = predict(W.forest)$predictions

cf.raw = causal_forest(X, Y, W,
                       Y.hat = Y.hat, W.hat = W.hat,
                       clusters = school.id)
varimp = variable_importance(cf.raw)
selected.idx = which(varimp > mean(varimp))

cf = causal_forest(X[,selected.idx], Y, W,
                   Y.hat = Y.hat, W.hat = W.hat,
                   clusters = school.id,
                   samples_per_cluster = 50,
                   tune.parameters = TRUE)
tau.hat = predict(cf)$predictions
```

features in low-signal situations. Second, we increase the `samples_per_cluster` parameter (called k in Section 1.2) to increase the number of samples used to grow each tree. Finally, the option `tune.parameters = TRUE` has the forest cross-validate tuning parameters using the R-objective rather than just setting defaults.

2 Workshop Results

We now use our causal forest as trained in Algorithm 1 to explore the questions from Section 1.1.

2.1 The average treatment effect

The first question asks about the overall effectiveness of the intervention. The package `grf` has a built-in function for average treatment effect estimation, based on a variant of augmented inverse-propensity weighting (Robins, Rotnitzky, and Zhao, 1994). With clusters, we compute an average treatment effect estimate $\hat{\tau}$ and a standard error estimate $\hat{\sigma}^2$ as follows:

$$\begin{aligned} \hat{\tau}_j &= \frac{1}{n_j} \sum_{\{i:A_i=j\}} \widehat{\Gamma}_i, \quad \hat{\tau} = \frac{1}{J}\sum_{j=1}^{J} \hat{\tau}_j, \quad \hat{\sigma}^2 = \frac{1}{J(J-1)} \sum_{j=1}^{J} (\hat{\tau}_j - \hat{\tau})^2, \\ \widehat{\Gamma}_i &= \hat{\tau}^{(-i)}(X_i) + \frac{W_i - \hat{e}^{(-i)}(X_i)}{\hat{e}^{(-i)}(X_i)\left(1-\hat{e}^{(-i)}(X_i)\right)} \left(Y_i - \hat{m}^{(-i)}(X_i) - \left(W_i - \hat{e}^{(-i)}(X_i)\right)\hat{\tau}^{(-i)}(X_i)\right). \end{aligned} \tag{8}$$

See Section 2.1 of Farrell (2015) for a discussion of estimators with this functional form, and Section 2.4 of Athey, Imbens, and Wager (2018) for a recent literature review. The value of cross-fitting is stressed in Chernozhukov et al. (2018a). An application of this method suggests that the treatment had a large positive on average.

```
ATE = average_treatment_effect(cf)
paste("95% CI for the ATE:", round(ATE[1], 3),
      "+/-", round(qnorm(0.975) * ATE[2], 3))
> "95% CI for the ATE: 0.247 +/- 0.04"
```

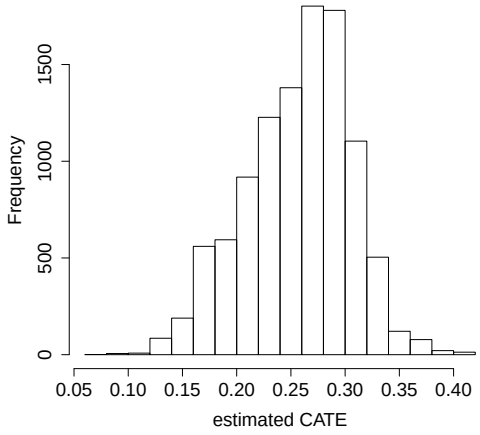


Figure 2: Histogram of out-of-bag CATE estimates from a causal forest trained as in Algorithm 1.

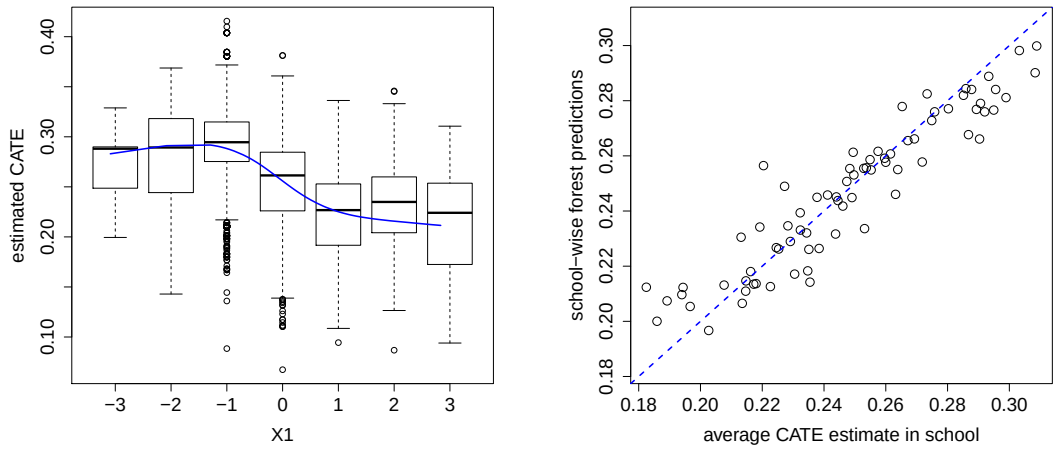
2.2 Assessing treatment heterogeneity

The next two questions pertain to treatment heterogeneity. Before addressing questions, however, it is useful to ask whether the causal forest has succeeded in accurately estimating treatment heterogeneity. As seen in Figure 2, the causal forest CATE estimates obviously exhibit variation; but this does not automatically imply that $\hat{\tau}^{(-i)}(X_i)$ is a better estimate of $\tau(X_i)$ than the overall average treatment effect estimate $\hat{\tau}$ from (8). Below, we seek an overall hypothesis test for whether heterogeneity in $\hat{\tau}^{(-i)}(X_i)$ is associated with heterogeneity in $\tau(X_i)$.

A first, simple approach to testing for heterogeneity involves grouping observations according to whether their out-of-bag CATE estimates are above or below the median CATE estimate, and then estimating average treatment effects in these two subgroups separately using the doubly robust approach (8). This procedure is somewhat heuristic, as the "high" and "low" subgroups are not independent of the scores $\widehat{\Gamma}_i$ used to estimate the within-group effects; however, the subgroup definition does not directly depend on the outcomes or treatments (Y_i, W_i) themselves, and it appears that this approach can provide at least qualitative insights about the strength of heterogeneity.

We also try a second test for heterogeneity, motivated by the "best linear predictor" method of Chernozhukov, Demirer, Duflo, and Fernandez-Val (2018b), that seeks to fit the CATE as a linear function of the the out-of-bag causal forest estimates $\hat{\tau}^{(-i)}(X_i)$. Concretely, following (4), we create two synthetic predictors, $C_i = \bar{\tau}(W_i - \hat{e}^{(-i)}(X_i))$ and $D_i = (\hat{\tau}^{(-i)}(X_i) - \bar{\tau})(W_i - \hat{e}^{(-i)}(X_i))$ where $\bar{\tau}$ is the average of the out-of-bag treatment effect estimates, and regress $Y_i - \hat{m}^{(-i)}(X_i)$ against C_i and D_i. Then, we can interpret the coefficient of D_i as a measure of the quality of the estimates of treatment heterogeneity, while C_i absorbs the average treatment effect. If the coefficient on D_i is 1, then the treatment heterogeneity estimates are well calibrated, while if the coefficient is D_i significant and positive, then at least we have evidence of a useful association between $\hat{\tau}^{(-i)}(X_i)$ and $\tau(X_i)$. More formally, one could use the p-value for the coefficient of D_i to test the hypothesis that the causal forest succeeded in finding heterogeneity; however, we caution that asymptotic results justifying such inference are not presently available.

Below, we show output from running both analyses (note that all results are cluster-robust, where each cluster gets the same weight). The overall picture appears somewhat mixed: Although point estimates are consistent with the presence of heterogeneity, neither detection is significant. Thus, at least if we insist on cluster-robust inference, any treatment heterogeneity that may be present appears to be relatively weak, and causal forests do not identify subgroups with effects that obviously stand



(a) variation with school-level mindset **(b)** evaluating forest trained on $\hat{\tau}_j$ from (8)

Figure 3: Panel **(a)** plots students' CATE estimates against school-level mindset `X1`. Panel **(b)** compares estimates from a regression forest trained to predict the per-school doubly robust treatment effect estimates $\hat{\tau}_j$ from (8) using school-level covariates, to school-wise averages of the causal forest estimates $\hat{\tau}^{(-i)}(X_i)$ trained as in Algorithm 1.

out. We discuss the role of cluster-robustness further in Section 3.1.

```
# Compare regions with high and low estimated CATEs
high_effect = tau.hat > median(tau.hat)
ate.high = average_treatment_effect(cf, subset = high_effect)
ate.low = average_treatment_effect(cf, subset = !high_effect)
paste("95% CI for difference in ATE:",
      round(ate.high[1] - ate.low[1], 3), "+/-",
      round(qnorm(0.975) * sqrt(ate.high[2]^2 + ate.low[2]^2), 3))
> "95% CI for difference in ATE: 0.053 +/- 0.071"

# Run best linear predictor analysis
test_calibration(cf)
>                         Estimate  Std. Error  t value  Pr(>|t|)
> mean.prediction         1.007477   0.083463 12.0710    <2e-16 ***
> differential.prediction 0.321932   0.306738  1.0495     0.294
```

2.3 The effect of `X1` and `X2`

Although our omnibus tests did not find strong evidence of treatment heterogeneity, this does not mean there is no heterogeneity present. Researchers had pre-specified interest in heterogeneity along two specific variables, namely pre-existing mindset (`X1`) and school-level achievement (`X2`), and it is plausible that a test for heterogeneity that focuses on these two variables may have more power than the agnostic tests explored above.

Both `X1` and `X2` are school-level variables, so we here design tests based on the per-school doubly robust treatment effect estimates $\hat{\tau}_j$ computed in (8). As seen below, this more targeted analysis

uncovers notable heterogeneity along `X1`, i.e., schools with larger values of `X1` appear to experience larger effects than schools with smaller values of `X1`. Conversely, we do not see much heterogeneity along `X2`, whether we divide schools into 2 subgroups (to test the monotone hypothesis) or into 3 subgroups (to test the goldilocks hypothesis).

Although the p-value for heterogeneity along `X1` is not small enough to withstand a Bonferroni test, it seems reasonable to take the detection along `X1` seriously because heterogeneity along `X1` was one of two pre-specified hypotheses. Interestingly, we also note that `X1` was the most important variable in the causal forest: The final causal forest was trained on 9 "selected" variables, and spent 24% of its splits on `X1` with splits weighted by depth (as in the function `variable_importance`). The left panel of Figure 3 plots the relationship between `X1` and $\hat{\tau}^{(-i)}(X_i)$.

```
dr.score = tau.hat + W / cf$W.hat *
  (Y - cf$Y.hat - (1 - cf$W.hat) * tau.hat) -
  (1 - W) / (1 - cf$W.hat) * (Y - cf$Y.hat + cf$W.hat * tau.hat)
school.score = t(school.mat) %*% dr.score / school.size

school.X1 = t(school.mat) %*% X$X1 / school.size
high.X1 = school.X1 > median(school.X1)
t.test(school.score[high.X1], school.score[!high.X1])
> t = -3.0205, df = 72.087, p-value = 0.00349
> 95 percent confidence interval: -0.1937 -0.0397

school.X2 = (t(school.mat) %*% X$X2) / school.size
high.X2 = school.X2 > median(school.X2)
t.test(school.score[high.X2], school.score[!high.X2])
> t = 1.043, df = 72.431, p-value = 0.3004
> 95 percent confidence interval: -0.0386 0.1234

school.X2.levels = cut(school.X2,
  breaks = c(-Inf, quantile(school.X2, c(1/3, 2/3)), Inf))
summary(aov(school.score ~ school.X2.levels))
>                   Df Sum Sq Mean Sq F value Pr(>F)
> school.X2.levels  2  0.085 0.04249   1.365  0.262
> Residuals        73  2.272 0.03112
```

2.4 Looking for school-level heterogeneity

Our omnibus test for heterogeneity from Section 2.2 produced mixed results; however, when we zoomed in on the pre-specified covariates `X1` and `X2` in Section 2.3, we uncovered interesting results. Noticing that both `X1` and `X2` are school-level (as opposed to student-level) covariates, it is natural to ask whether an analysis that only focuses only on school-level effects may have had more power than our original analysis following Algorithm 1.

Here, we examine this question by fitting models to the school-level estimates $\hat{\tau}_j$ from (8) using only school level covariates. We considered both an analysis using a regression forest, as well as classical linear regression modeling. Both methods, however, result in conclusions that are in line with the ones obtained above. The strength of the heterogeneity found by the regression forest trained on the $\hat{\tau}_j$ as measured by the "calibration test" is comparable to the strength of the heterogeneity found by our original causal forest; moreover, as seen in the right panel of Figure 3, the predictions made by this regression forest are closely aligned with school-wise averaged predictions from the original causal forest. Meanwhile, a basic linear regression analysis uncovers a borderline amount of effect modification along `X1` and nothing else stands out.

The overall picture is that, by looking at the predictor `X1` alone, we can find credible effect modification that is correlated negatively with `X1`. However, there does not appear to be strong enough heterogeneity for us to be able to accurately fit a more complex model for $\tau(\cdot)$: Even a linear model for effect modification starts to suffer from low signal, and it is not quite clear whether `X1` is an effect modifier after we control for the other school-level covariates.

```
# Regression forest analysis
school.forest = regression_forest(school.X, school.score)
school.pred = predict(school.forest)$predictions
test_calibration(school.forest)
>                           Estimate Std. Error t value Pr(>|t|)
> mean.prediction            0.998765   0.083454 11.9679   <2e-16 ***
> differential.prediction    0.619299   0.706514  0.8766   0.3836

# Ordinary least-squares analysis
coeftest(lm(school.score ~ school.X), vcov = vcovHC)
>                Estimate Std. Error t value Pr(>|t|)
> (Intercept)  0.2434703  0.0770302  3.1607 0.002377 **
> X1          -0.0493032  0.0291403 -1.6919 0.095377 .
> X2           0.0143625  0.0340139  0.4223 0.674211
> X3           0.0092693  0.0264267  0.3508 0.726888
> X4           0.0248985  0.0258527  0.9631 0.339019
> X5          -0.0336325  0.0265401 -1.2672 0.209525
> XC.1        -0.0024447  0.0928801 -0.0263 0.979081
> XC.2         0.0826898  0.1052411  0.7857 0.434845
> XC.3        -0.1376920  0.0876108 -1.5716 0.120818
> XC.4         0.0408624  0.0820938  0.4978 0.620313
```

3 Post-workshop analysis

Two notable differences between the causal forest analysis used here and a more direct machine-learning-based analysis were our use of cluster-robust methods, and of orthogonalization for robustness to confounding as in (7). To understand the value of these features, we revisit some analyses from Section 2 without them.

3.1 The value of clustering

If we train a causal forest on students without clustering by school, we obtain markedly different results from before: The confidence interval for the average treatment effect is now roughly half as long as before, and there appears to be unambiguously detectable heterogeneity according to the `test_calibration` function. Moreover, as seen in the left panel of Figure 4, the CATE estimates $\hat{\tau}^{(-i)}(X_i)$ obtained without clustering are much more dispersed than those obtained with clustering (see Figure 2): The sample variance of the $\hat{\tau}^{(-i)}(X_i)$ increases by a factor 5.82 without clustering.

It appears that these strong detections without clustering are explained by excess optimism from ignoring variation due to idiosyncratic school-specific effects, rather than from a true gain in power from using a version of causal forests without clustering. The right panel of Figure 4 shows per-school estimates of $\hat{\tau}^{(-i)}(X_i)$ from the non-cluster-robust causal forest, and compares them to predictions for the mean CATE in the school obtained in a way that is cluster-robust. The differences are striking: For example, the left-most school in the right panel of Figure 4 has non-cluster-robust $\hat{\tau}^{(-i)}(X_i)$ estimates that vary from 0.26 to 0.36, whereas the cluster-robust estimate of its mean CATE was roughly 0.2. A simple explanation for how this could come about is that students in the school happened to have

unusually high treatment effects, and that the non-cluster-robust forest was able to overfit to this school-level effect because it does not account for potential correlations between different students in the same school.

To gain deeper insights into the behavior of non-cluster robust forests, we tried a 5-fold version of this algorithm where the forests themselves are not cluster-robust, but the estimation folds are cluster aligned. Specifically, we split the schools into 5 folds; then, for each fold, we fit a causal forest without clustering on observations belonging to schools in the 4/5 other folds, and made CATE estimates on the held out fold. Finally, re-running a best linear prediction test on out-of-fold predictions as in the `test_calibration` function, we found at best tenuous evidence for the presence of heterogeneity (in fact, the resulting t-statistic for heterogeneity, 0.058, was weaker than the one in Section 2.2). In other words, if we use evaluation methods that are robust to clustering, then the apparent gains from non-cluster-robust forests wash away.

Thus, it appears that different schools have very different values of $\hat{\tau}_j$; however, most of the school-wise effects appear to be idiosyncratic, and cannot be explained using covariates. In order to gain insights that generalize to new schools we need to cluster by school; and, once we do so, much of the apparent heterogeneity between schools ends up looking like noise.

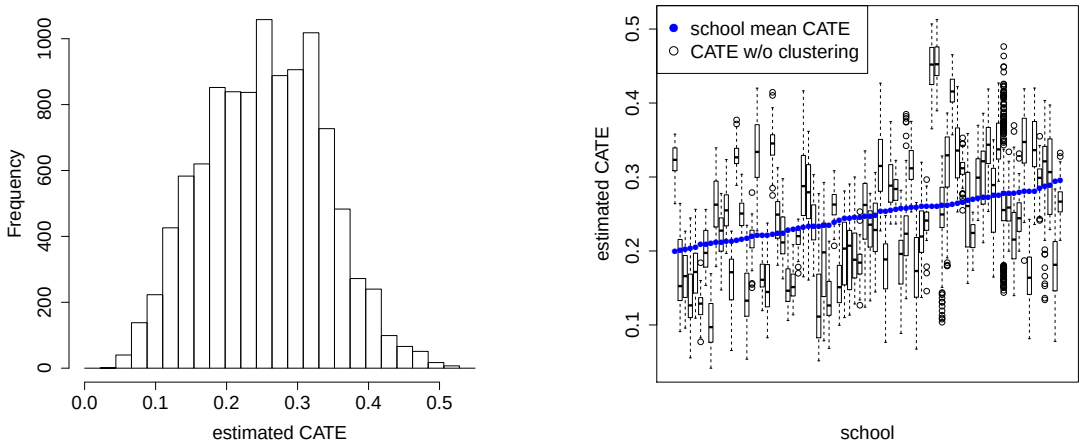
```
cf.noclust = causal_forest(X[,selected.idx], Y, W,
                           Y.hat = Y.hat, W.hat = W.hat,
                           tune.parameters = TRUE)
ATE.noclust = average_treatment_effect(cf.noclust)
paste("95% CI for the ATE:", round(ATE.noprop[1], 3),
      "+/-", round(qnorm(0.975) * ATE.noprop[2], 3))
> "95% CI for the ATE: 0.253 +/- 0.022"

test_calibration(cf.noclust)
>                          Estimate Std. Error t value  Pr(>|t|)
> mean.prediction          1.003796   0.044779 22.4164 < 2.2e-16 ***
> differential.prediction 0.634163   0.132700  4.7789 1.786e-06 ***
```

3.2 The value of orthogonalization

In this dataset, orthogonalization appears to be less important than clustering. If we train a causal forests without estimating the propensity score or, more specifically, using the trivial propensity model $\hat{e}(X_i) = \overline{W} = n^{-1} \sum_{i=1}^{n} W_i$, we uncover essentially the same average treatment effect estimate as with orthogonalization. Moreover, as shown in Figure 5, the causal forests trained with or without orthogonalization yield essentially the same CATE estimates $\hat{\tau}^{(-i)}(X_i)$.

One reason for this phenomenon may be that, here, the most important confounders are also important for predicting Y: In Algorithm 1, the most important predictor for both the W- and Y-forests is `S3`, with 22% of splits and 70% of splits respectively (both weighted by depth as in the `variable_importance` function). Meanwhile, as argued in Belloni, Chernozhukov, and Hansen (2014), orthogonalization is often most important when there are some features that are highly predictive of treatment propensities but not very predictive of Y. Thus, it is possible that the non-orthogonalized forest does well here because we were lucky, and there were no confounders that only had a strong effect the propensity model.

To explore this hypothesis, we present a synthetic example where some variables have stronger effects on W than on Y and see that, as expected, orthogonalization is now important. There is clearly no treatment effect, yet the non-orthogonalized forest appears to find a non-zero effect.



(a) histogram of CATE estimates w/o clustering **(b)** per-school CATE estimates w/o clustering

Figure 4: Panel **(a)** is a histogram of CATE estimates $\hat{\tau}^{(-i)}(X_i)$ trained using a causal forest that does not account for school-level clustering. Panel **(b)** compares per-student predictions $\hat{\tau}^{(-i)}(X_i)$ from a non-cluster-robust causal forest to per-school mean treatment effect predictions from a forest trained on per-school responses as in Section 2.4.

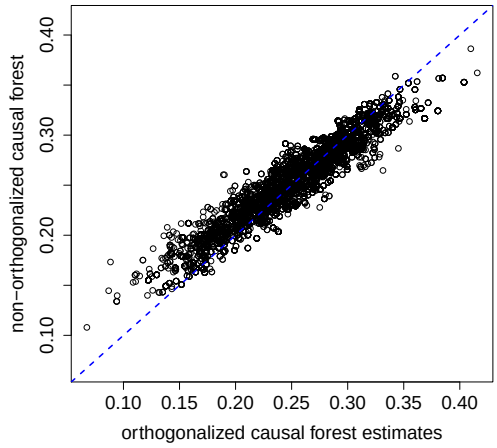


Figure 5: Comparison of estimates from a forest trained with a trivial propensity model $\hat{e}(X_i) = \overline{W} = n^{-1} \sum_{i=1}^{n} W_i$ to predictions from the forest trained as in Algorithm 1.

```
cf.noprop = causal_forest(X[,selected.idx], Y, W,
                          Y.hat = Y.hat, W.hat = mean(W),
                          tune.parameters = TRUE,
                          samples_per_cluster = 50,
                          clusters = school.id)
ATE.noprop = average_treatment_effect(cf.noprop)
paste("95% CI for the ATE:", round(ATE.noprop[1], 3),
      "+/-", round(qnorm(0.975) * ATE.noprop[2], 3))
> "95% CI for the ATE: 0.253 +/- 0.04"

n.synth = 1000; p.synth = 10
X.synth = matrix(rnorm(n.synth * p.synth), n.synth, p.synth)
W.synth = rbinom(n.synth, 1, 1 / (1 + exp(-X.synth[,1])))
Y.synth = 2 * rowMeans(X.synth[,1:6]) + rnorm(n.synth)
Y.forest.synth = regression_forest(X.synth, Y.synth)
Y.hat.synth = predict(Y.forest.synth)$predictions
W.forest.synth = regression_forest(X.synth, W.synth)
W.hat.synth = predict(W.forest.synth)$predictions

cf.synth = causal_forest(X.synth, Y.synth, W.synth,
            Y.hat = Y.hat.synth, W.hat = W.hat.synth)
ATE.synth = average_treatment_effect(cf.synth)
paste("95% CI for the ATE:", round(ATE.synth[1], 3),
      "+/-", round(qnorm(0.975) * ATE.synth[2], 3))
> "95% CI for the ATE: 0.125 +/- 0.151"

cf.synth.noprop = causal_forest(X.synth, Y.synth, W.synth,
            Y.hat = Y.hat.synth, W.hat = mean(W.synth))
ATE.synth.noprop = average_treatment_effect(cf.synth.noprop)
paste("95% CI for the ATE:", round(ATE.synth.noprop[1], 3),
      "+/-", round(qnorm(0.975) * ATE.synth.noprop[2], 3))
> "95% CI for the ATE: 0.220 +/- 0.142"
```

4 Discussion

We applied causal forests to study treatment heterogeneity on a dataset derived from the National Study of Learning Mindsets. Two challenges in this setting involved an observational study design with unknown treatment propensities, and clustering of outcomes at the school level. Causal forests allow for an algorithmic specification that addresses both challenges. Of these two challenges, school-level clustering had a dramatic effect on our analysis. If we properly account for the clustering, we find hints of the presence of treatment heterogeneity (Section 2.3), but accurate non-parametric estimation of $\tau(x)$ is difficult (Section 2.2). In contrast, an analysis that ignores clusters claims to find very strong heterogeneity in $\tau(x)$ that can accurately be estimated (Section 3.1).

This result highlights the need for a deeper discussion of the how to work with clustered observations when modeling treatment heterogeneity. The traditional approach is to capture cluster effects via "fixed effect" or "random effect" models of the form

$$Y_i = m(X_i) + W_i \tau(X_i) + \beta_{A_i} + W_i \gamma_{A_i} + \varepsilon_i, \tag{9}$$

where $A_i \in \{1, ..., J\}$ denotes the cluster membership of the i-th sample whereas β_j and γ_j denote per-cluster offsets on the main effect and treatment effect respectively, and the nomenclature around

fixed or random effects reflects modeling choices for β and γ (Wooldridge, 2010). In a non-parametric setting, however, assuming that clusters have an additive effect on Y_i seems rather restrictive. The approach we took in this note can be interpreted as fitting a functional random effects model

$$Y_i = m_{A_i}(X_i) + W_i \tau_{A_i}(X_i) + \varepsilon_i, \quad \tau(x) = \mathbb{E}\left[\tau_j(x)\right], \tag{10}$$

where each cluster has its own main and treatment effect function, and the expectation above is defined with respect to the distribution of per-cluster treatment effect functions. It would be of considerable interest to develop a better understanding of the pros and cons of different approaches to heterogeneous treatment effect estimation on clustered data.

References

A. Abadie, S. Athey, G. Imbens, and J. Wooldridge. When should you adjust standard errors for clustering? *arXiv preprint arXiv:1710.02926*, 2017.

J. D. Angrist and J.-S. Pischke. *Mostly harmless econometrics: An empiricist's companion*. Princeton university press, 2008.

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

S. Athey, G. W. Imbens, and S. Wager. Approximate residual balancing: debiased inference of average treatment effects in high dimensions. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 80(4):597–623, 2018.

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

S. Basu, K. Kumbier, J. B. Brown, and B. Yu. Iterative random forests to discover predictive and stable high-order interactions. *Proceedings of the National Academy of Sciences*, page 201711236, 2018.

A. Belloni, V. Chernozhukov, and C. Hansen. Inference on treatment effects after selection among high-dimensional controls. *The Review of Economic Studies*, 81(2):608–650, 2014.

G. Biau and E. Scornet. A random forest guided tour. *Test*, 25(2):197–227, 2016.

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

V. Chernozhukov, D. Chetverikov, M. Demirer, E. Duflo, C. Hansen, W. Newey, and J. Robins. Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, 21(1):C1–C68, 2018a.

V. Chernozhukov, M. Demirer, E. Duflo, and I. Fernandez-Val. Generic machine learning inference on heterogenous treatment effects in randomized experiments. Technical report, National Bureau of Economic Research, 2018b.

P. Ding, A. Feller, and L. Miratrix. Randomization inference for treatment effect variation. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 78(3):655–671, 2016.

V. Dorie, J. Hill, U. Shalit, M. Scott, and D. Cervone. Automated versus do-it-yourself methods for causal inference: Lessons learned from a data analysis competition. *arXiv preprint arXiv:1707.02641*, 2017.

M. H. Farrell. Robust inference on average treatment effects with possibly more covariates than observations. *Journal of Econometrics*, 189(1):1–23, 2015.

M. Gopalan and E. Tipton. Is the national study of learning mindsets nationally-representative? *PsyArXiv. November*, 3, 2018.

P. R. Hahn, J. S. Murray, and C. Carvalho. Bayesian regression tree models for causal inference: regularization, confounding, and heterogeneous effects. *arXiv preprint arXiv:1706.09523*, 2017.

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

K. Imai and M. Ratkovic. Estimating treatment effect heterogeneity in randomized program evaluation. *The Annals of Applied Statistics*, 7(1):443–470, 2013.

G. W. Imbens and D. B. Rubin. *Causal Inference in Statistics, Social, and Biomedical Sciences*. Cambridge University Press, 2015.

S. Künzel, J. Sekhon, P. Bickel, and B. Yu. Meta-learners for estimating heterogeneous treatment effects using machine learning. *arXiv preprint arXiv:1706.03461*, 2017.

A. R. Luedtke and M. J. van der Laan. Super-learning of an optimal dynamic treatment rule. *The International Journal of Biostatistics*, 12(1):305–332, 2016.

X. Nie and S. Wager. Quasi-oracle estimation of heterogeneous treatment effects. *arXiv preprint arXiv:1712.04912*, 2017.

B. A. Nosek et al. Promoting an open research culture. *Science*, 348(6242):1422–1425, 2015.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2017. URL https://www.R-project.org/.

J. M. Robins, A. Rotnitzky, and L. P. Zhao. Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association*, 89(427):846–866, 1994.

P. M. Robinson. Root-n-consistent semiparametric regression. *Econometrica*, pages 931–954, 1988.

P. R. Rosenbaum. *Observational Studies*. Springer, 2002.

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

U. Shalit, F. D. Johansson, and D. Sontag. Estimating individual treatment effect: generalization bounds and algorithms. In *ICML*, pages 3076–3085, 2017.

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

J. Tibshirani, S. Athey, R. Friedberg, V. Hadad, L. Miner, S. Wager, and M. Wright. *grf: Generalized Random Forests (Beta)*, 2018. URL https://github.com/grf-labs/grf. R package version 0.10.2.

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

J. M. Wooldridge. *Econometric analysis of cross section and panel data*. MIT press, 2010.

D. S. Yeager et al. Using design thinking to improve psychological interventions: The case of the growth mindset during the transition to high school. *Journal of Educational Psychology*, 108(3): 374, 2016.

Q. Zhao, D. S. Small, and A. Ertefaie. Selective inference for effect modification via the lasso. *arXiv preprint arXiv:1705.08020*, 2017.