Some methods for heterogeneous treatment effect estimation in high-dimensions

Scott Powers, Junyang Qian, Kenneth Jung, Alejandro Schuler,
Nigam Shah, Trevor Hastie and Robert Tibshirani
Stanford University

November 10, 2017

Abstract

When devising a course of treatment for a patient, doctors often have little quantitative evidence on which to base their decisions, beyond their medical education and published clinical trials. Stanford Health Care alone has millions of electronic medical records (EMRs) that are only just recently being leveraged to inform better treatment recommendations. These data present a unique challenge because they are high-dimensional and observational. Our goal is to make personalized treatment recommendations based on the outcomes for past patients similar to a new patient. We propose and analyze three methods for estimating heterogeneous treatment effects using observational data. We compare the performance of these methods in simulation with the gradient forest of Athey et al. (2017).

1 Introduction

In February 2017, at the Grand Rounds of Stanford Medicine, one of us (NS) unveiled a new initiative — the Informatics Consult. Through this service, clinicians can submit a consultation request online and receive a report based on insights drawn from hundreds of millions of electronic medical records (EMRs) from Stanford Health Care. While the system is in its early stages, a future version will include treatment recommendations: helping a doctor to choose between treatment options for a patient, in cases where there is no randomized controlled trial (RCT) which compares the options. This announcement was met with excitement from the doctors in attendance, considering that they generally need to make decisions without any support from quantitative evidence (about 95% of the time) (Shah, 2016). Building such a system is a priority in many medical centers in the U.S. and around the world.

The problem setting on which this paper focuses is when a doctor is presented with a patient who has some medical ailment, and the doctor is considering one or more treatment options. A relevant question from the patient's perspective is,

what is the effect of these treatments on patients like me? Devising a meaningful definition for "patients like me" is especially difficult given the high-dimensional nature of the problem: We observe thousands of features describing each patients, any of which could be used to describe patient similarity. The other significant complication is that our goal is to infer causal effects from observational data. The task of mining EMRs to support physician decision-making is what motivates this paper. We propose and study methods for estimation and inference of heterogeneous treatment effects, for both randomized experiments and observational studies. We focus on the case of a choice between two treatments, which for the purposes of this manuscript we label as "treatment" and "control".

In detail, we have an $n \times p$ matrix of features $\mathbf{X}$, a treatment indicator vector $\mathbf{T} \in \{0,1\}^n$, and a vector of quantitative responses $\mathbf{Y} \in \mathbb{R}^n$. Let X_i denote the ith row of $\mathbf{X}$, likewise T_i and Y_i. We assume the n observations (X_i, T_i, Y_i) are sampled i.i.d. from some unknown distribution. The number of treated patients is $N_1 = |\{i : T_i = 1\}|$, and the number of control patients is $N_0 = |\{i : T_i = 0\}|$. We adopt the Neyman–Rubin potential outcomes model (Splawa-Neyman et al., 1990; Rubin, 1974): each patient i has potential outcomes $Y_i^{(1)}$ and $Y_i^{(0)}$, only one of which is observed. $Y_i^{(1)}$ is the response that the patient would have under treatment, and $Y_i^{(0)}$ is the response the patient would have under control. Hence the outcome that we actually observe is $Y_i = Y_i^{(T_i)}$. We consider both randomized controlled trials, where T_i is independent of all pre-treatment characteristics,

$$\left(X_i,\, Y_i^{(0)},\, Y_i^{(1)}\right) \perp\!\!\!\perp T_i, \tag{1}$$

and observational studies, where the distribution of T_i is dependent on the covariates. This scenario is discussed in further detail in Section 2.1.

We describe four important functions for modelling data of this type. The first is the propensity function, which gives the probability of treatment assignment, conditional on covariates:

$$\pi(x) \equiv \mathbb{P}(T = 1 | X = x). \tag{2}$$

The next two functions are the conditional mean functions: the expected response given treatment and the expected response given control.

$$\mu_1(x) \equiv \mathbb{E}[Y | X = x, T = 1] \quad \text{and} \quad \mu_2(x) \equiv \mathbb{E}[Y | X = x, T = 0].$$

The fourth function, and the one of greatest interest, is the treatment effect function, which is the difference between the two conditional means:

$$\tau(x) \equiv \mu_1(x) - \mu_0(x).$$

We seek regions in predictor space where the treatment effect is relatively large or relatively small. This is particularly important for the area of personalized medicine, where a treatment might have a negligible effect when averaged over all patients but could be beneficial for certain patient subgroups.

An outline of this paper is as follows. Section 2 reviews related work. In Section 3 we describe the two main high-level approaches to the estimation of heterogeneous treatment effects: transformed outcome regression and conditional mean regression. In Section 4 we introduce *pollinated transformed outcome* (PTO) forests, while *causal boosting* is proposed in Section 5. *Causal MARS* is the focus of Section 6. In Section 7 we report the results of a simulation study comparing all of these methods, and a real data application is illustrated in Section 8. We end with a discussion.

2 Related work

Early work on heterogeneous treatment effect estimation (Gail and Simon, 1985) was based on comparing pre-defined subpopulations of patients in randomized experiments. To characterize interactions between a treatment and continuous covariates, Bonetti and Gelber (2004) formalized the subpopulation treatment effect patter plot (STEPP). Sauerbrei et al. (2007) proposed an efficient algorithm for flexible model-building with multivariable fractional polynomial interaction (MFPI) and compared the empirical performance of MFPI with STEPP.

Identifying subgroups within the patient population is becoming especially problematic in high-dimensional data, as in EMRs. In recent years, a great amount of work has been done to apply methods from machine learning to let the data inform what are the important subgroups in terms of treatment effect. Su et al. (2009) proposed interaction trees for adaptively defining subgroups based on treatment effect. Athey and Imbens (2016) proposed causal trees, which are similar, and constructed valid confidence intervals. Wager and Athey (2015) improved on this line of work by growing random forests (Breiman, 2001) from causal trees. These tree-based methods all use shared-basis conditional mean regression in the framework of Section 3. An example of a transformed-outcome estimator is the FindIt method of Imai and Ratkovic (2013) which trains an adapted support vector machine on a transformed binary outcome. Tian et al. (2014) introduced a simple linear model based on transformed covariates and show that it is equivalent to transformed outcome regression in the Gaussian case. In a novel approach, Zhao et al. (2012) used outcome weighted learning to directly determine individualized treatment rules, skipping the step of estimating individualized treatment effects. The problem of estimating heterogeneous treatment effects has also received significant attention in Bayesian literature. Hill (2011) and Green and Kern (2012) approached the problem using Bayesian additive regression trees (Chipman et al., 1998), and Taddy et al. (2016) proposed a method based on Bayesian forests. Chen et al. (2012) developed a Bayesian method for finding qualitative interactions between treatment and covariates, and there are other Bayesian methods for flexible nonlinear modelling of interactive/non-additive relationships between covariates and response (LeBlanc, 1995; Gustafson, 2000).

What all of the above work (except Hill (2011)) have in common is that they assume randomized treatment assignment. Athey and Imbens (2016) discussed

the possibility of adapting their method to observational data but go no further. Wager and Athey (2015) proposed the propensity forest when treatment is not randomized, but this method does not target heterogeneity in the treatment effect. Similarly, Xie et al. (2012) model treatment effect as a function of propensity score, missing out on how it depends on the covariates except through treatment propensity. Crump et al. (2008) devised a nonparametric test for the null hypothesis that the treatment effect is constant across patients, but that is not suited to high-dimensional data. One promising approach which flexibly handles high-dimensional and observational data is the gradient forest of Athey et al. (2017)—we compare the performance of our methods with that of the gradient forest in Section 7.

We are particularly interested in flexible, non-parametric approaches that can handle large numbers of observations and predictors, and model interactions between predictors, which none of these papers deal with (except for Zhao et al. (2012)).

2.1 Propensity score methods

Much of causal inference is based on the propensity score (Rosenbaum and Rubin, 1983), which is the estimated probability that a patient would receive treatment, conditioned on the patient's covariates. If the estimate of the propensity function (2) is $\hat{\pi}(\cdot)$, then the propensity score for a patient with covariate vector x is $\hat{\pi}(x)$. Throughout the present work, we estimate the propensity function using the probability forests of Malley et al. (2012). We are able to do so quickly using the fast implement in the R package `ranger` (Wright and Ziegler, 2015).

For the estimation of a population-average treatment effect (ATE), propensity score methods for reducing bias in observational studies have been established (Austin, 2011). *Propensity score matching* emulates a randomized control trial (RCT) by choosing pairs of patients with similar propensity scores, one each in the treatment and control arms, and discards the unmatched patients. *Stratification on the propensity score* groups patients into bins of similar propensity scores to compute the ATE within each bin. The overall ATE is the average of these treatment effects, weighted by the overall frequency of each bin. *Inverse probability weighting* assigns a weight to each patient equal to the inverse of the propensity score if the patient is treated, or else the inverse of one minus the propensity score if the patient is not treated. Hence patients who tend to be under-represented in their arm are given more weight. Propensity score stratification and inverse probability weighting are discussed in more detail in the appendix, along with an additional method: *transformed outcome averaging*.

The assumption that enables these methods to generate causal conclusions from observational data is known alternatingly across the literature as unconfoundedness, exogeneity or strong ignorability:

$$(Y_i^{(1)}, Y_i^{(0)}) \perp\!\!\!\perp T_i | X_i$$

This is the assumption made in the present work. It means that the relationship between the potential outcomes and treatment must be fully explained by X. There can be no additional unmeasured confounding variable which effects a dependence between potential outcomes and treatment. Note, however, that the outcome itself is not independent of treatment because the treatment determines which potential outcome is observed.

Low et al. (2016) cast doubt on the ability of propensity score methods to adequately account for selection bias in a sophisticated simulation designed to model reality. Nevertheless, we observe in Section 7 that propensity score adjustments improve results in non-randomized simulations, which means that they can be used to help doctors make more informed decisions, so we push forward with the application of propensity scores.

3 Transformed outcome regression and conditional mean regression

Methods for estimating heterogeneous treatment effects generally fall into one of two categories: *transformed outcome regression* or *conditional mean regression*. In this section we describe the two approaches and explain why we prefer conditional mean regression. The propensity transformed outcome method (Section 4) uses a combination of the two approaches, while causal forests (Section 2), causal boosting (Section 5), and causal MARS (Section 6) are all conditional mean regression methods.

Transformed outcome regression is based on the same idea as transformed outcome averaging, which is laid out in detail in the appendix. Given the data described in Section 1, we define the *transformed outcome* as

$$Z \equiv T\frac{Y}{\pi(X)} + (1-T)\frac{-Y}{1-\pi(X)}.$$

This quantity is interesting because, as shown in the appendix, for any covariate vector x, $\mathbb{E}[Z|X = x] = \tau(x)$. So the transformed outcome gives us for each patient an unbiased estimate of the personalized treatment effect for that patient. Using this, we can simply use the tools of supervised learning to estimate a regression function for the mean of Z given X. The weakness of this approach is that while Z is unbiased for the treatment effect, its variance can be large due the presence of the propensity score, which can be close to zero or one, in the denominator.

An alternative approach, conditional mean regression is based on the idea that because $\tau(x)$ is defined as the difference between $\mu_1(x)$ and $\mu_0(x)$, if we can get good estimates of these conditional mean functions, then we have a good estimate of the treatment effect function. Estimating the functions $\mu_1(x)$ and $\mu_0(x)$ are supervised learning problems. If they are both estimated perfectly, then there is no need to bother with propensity scores. The problem is that in practice we never estimate either function perfectly, and differences between the

covariate distributions in the two treatment groups can lead to bias in treatment effect estimation if propensity scores are ignored.

We compare these two approaches with a simple example: Consider the task of estimating an ATE using data from a randomized trial. This may seem far removed from heterogeneous treatment effect estimation, but we will describe how two of our methods are based on estimating local ATEs for subpopulations in our data. In this case, the transformed outcome is

$$Z = T\frac{Y}{1/2} + (1-T)\frac{-Y}{1/2} = 2TY - 2(1-T)Y,$$

and the corresponding estimate of the ATE is

$$\hat{\tau}_{\text{TO}} = \frac{1}{n}\sum_{i=1}^{n} Z_i = \frac{2N_1\bar{Y}_1 - 2N_0\bar{Y}_0}{N_1+N_0} = \frac{N_1}{n/2}\bar{Y}_1 - \frac{N_0}{n/2}\bar{Y}_0,$$

where $\bar{Y}_1$ is the average response of patients who received treatment and $\bar{Y}_0$ is the average response of control patients. Meanwhile the conditional mean estimator of the ATE would be

$$\hat{\tau}_{\text{CM}} = \bar{Y}_1 - \bar{Y}_0.$$

Here we are implicitly assuming that neither N_1 nor N_0 is zero. It is worth noting that

$$\hat{\tau}_{\text{TO}} = \hat{\tau}_{\text{CM}} + \frac{N_1 - N_0}{n}(\bar{Y}_1 + \bar{Y}_0),$$

so if $N_1 = N_0$ or $\bar{Y}_1 + \bar{Y}_0 = 0$, then $\hat{\tau}_{\text{TO}} = \hat{\tau}_{\text{CM}}$. However N_1, N_0, $\bar{Y}_1$ and $\bar{Y}_0$ are all random. Given a fixed sample size n, N_1 follows a Binomial$(n, 1/2)$ distribution (truncated to exclude 0 and n), and N_0 is the difference between n and N_1. Suppose $\bar{Y}_1$ and $\bar{Y}_0$ have normal distributions with variances inversely proportional to sample size:

$$\bar{Y}_1 \sim \text{Normal}(\mu_1, \sigma^2/N_1) \qquad \text{and} \qquad \bar{Y}_0 \sim \text{Normal}(\mu_0, \sigma^2/N_0).$$

Note that both $\hat{\tau}_{\text{CM}}$ and $\hat{\tau}_{\text{TO}}$ are unbiased for $\tau \equiv \mu_1 - \mu_0$, but the two estimators have different variances. Conditioning on N_1, the variance of $\hat{\tau}_{\text{CM}}$ is

$$\mathbb{E}[(\hat{\tau}_{\text{CM}} - \tau)^2 | N_1] = \mathbb{V}(\bar{Y}_1 - \bar{Y}_0 | N_1) = \sigma^2/N_1 + \sigma^2/N_0$$

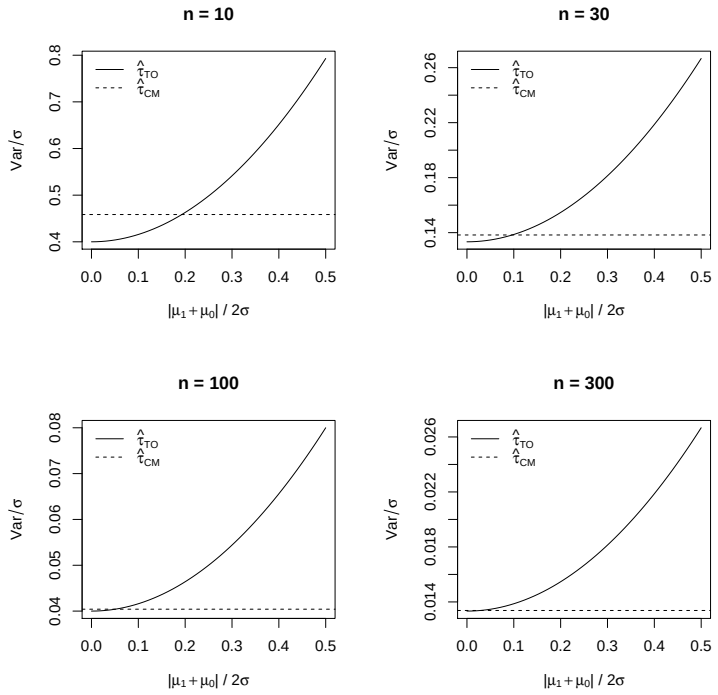
while the variance of $\hat{\tau}_{\text{TO}}$ given N_1 is

$$\mathbb{E}[(\hat{\tau}_{\text{TO}} - \tau)^2 | N_1] = \mathbb{V}(\hat{\tau}_{\text{TO}} | N_1) + (\mathbb{E}[\hat{\tau}_{\text{TO}} - \tau | N_1])^2 = \frac{4}{n}\sigma^2 + \left(\frac{N_1 - N_0}{n}\right)^2 (\mu_1 + \mu_0)^2.$$

So the key is the ratio of the main effect $(\mu_1 + \mu_0)/2$ to the noise level σ. If

$$\left|\frac{\mu_1 + \mu_0}{2\sigma}\right| < \sqrt{\frac{N_1^{-1} + N_0^{-1} - 4n^{-1}}{(N_1 - N_0)^2}},$$

Figure 1: *The variance of two ATE estimators for $n = 10$, 30, 100 and 300, as the ratio of the absolute main effect $|\mu_1 + \mu_0|/2$ to the noise level σ increases from 0 to 0.5.*



then $\hat{\tau}_{\mathrm{TO}}$ has less variance. If the inequality is reversed, then $\hat{\tau}_{\mathrm{CM}}$ has less variance. Marginalizing over the truncated binomial distribution of N_1 is difficult to do analytically, but we can numerically estimate the marginal variance of each estimator for any $n > 1$. Figure 1 illustrates the results for a few different choices of n.

We observe that for small n, $\hat{\tau}_{\mathrm{TO}}$ can have slightly smaller variance than $\hat{\tau}_{\mathrm{CM}}$ if the absolute value of the main effect is close to zero. But this advantage tends to zero as n increases, and $\hat{\tau}_{\mathrm{TO}}$ has much greater variance if the main effect is large. In conclusion, we prefer the conditional mean estimator because of the potentially high variance of the transformed outcome estimator. This is reflected in the following sections as all of our methods use some version of conditional mean regression.

3.1 Shared-basis conditional mean regression

In high-dimensional data it is often necessary to choose a subset of variables to include in a model. Beyond that, nonparametric methods adaptively choose

transformations of variables. Collectively, we refer to the variables and transformations selected as the basis of the regression. In conditional mean regression it is to be expected that the selected basis be different between the two regression functions. This can cause differences between the conditional means attributable not to evidence of a heterogeneous treatment effect but rather due to chance in basis selection.

To address this all of our methods jointly choose the same basis for both conditional mean regressions. In detail, this shared basis is chosen adaptively to best explain heterogeneity in the treatment effect, rather than explaining the variance in either treatment group. How exactly this shared basis is determined is different for each method.

4 Pollinated transformed outcome (PTO) forests

We first present the idea of a pollinated transformed outcome (PTO) forest in detail and then explain the various components.

In step 1 we compute an unbiased point estimate of the treatment effect for each individual; then in step 2, we fit a random forest using this effect as the outcome. In principal, this should estimate our personalized treatment effect. However, we don't trust these estimates too much, because the outcome can be highly variable. But we will put faith in the trees they produced.

Thus in step 3, we "pollinate" the trees separately with the treated and untreated populations. That is, we send data down each tree and compute new predictions for each terminal node. In step 4, the difference $z_i = G_1(x_i) - G_0(x_i)$ gives us an estimate of the treatment effect. Finally in step 5, we then post-process these predictions by fitting one more forest, primarily for interpretation.

Figure 2 illustrates the benefits of cross-pollination. In this example $n = 100, p = 50$ and the response is simulated in each arm according to $Y_i \sim \mathcal{N}(1 - X_{i1} + X_{i2}, 1)$ for treated patients and $Y_i \sim \mathcal{N}(X_{i1} + X_{i2}, 1)$ for untreated patients. Hence the true personalized treatment effect for patient i is $1 + 2X_{i1}$. In the top row the treatment is randomly assigned, while in the bottom row, the probability of treatment assignment is $(1 + e^{X_{i1}+X_{i2}})^{-1}$. The raw estimates correspond to a random forest (as in step 2) grown to predict the transformed outcome. The pollinated estimates correspond to re-estimating (as in step 3) the means of the leaves within each arm. We observe that in each case, the pollination improves the estimates.

5 Causal boosting

An alternative to a random forest for least squares regression is boosted trees. Boosting builds up a function approximation by successively fitting weak learners to the residuals of the model at each step. In this section we generalize least squares boosting for regression (Friedman, 2001) to the problem of heterogeneous treatment effect estimation.

Algorithm 1: PTO forest

1. Build a depth-controlled propensity random (regression) forest $\hat{\pi}$ using the treatment indicator as the response. Use regression trees, so that we estimate the probability of the terminal-node means. If the data are known to have come from a randomized trial, do not build a random forest and instead define $\hat{\pi}$ to be identically equal to the probability of treatment assignment.

2. Define the transformed outcome by

$$Z_i = T_1 \frac{Y_i}{\hat{\pi}(X_i)} + (1 - T_i) \frac{-Y_i}{\hat{\pi}(X_i)}.$$

3. For the randomized treatment setting, define the transformed outcome by

$$\delta_i = (2T_i - 1)Y_i.$$

   Note that if $\hat{\pi}(X_i)$ is the true probability of receiving treatment given covariates X_i, then $E[Z_i|T_i, X_i] = \tau(X_i)$, the true conditional treatment effect (see appendix for details).

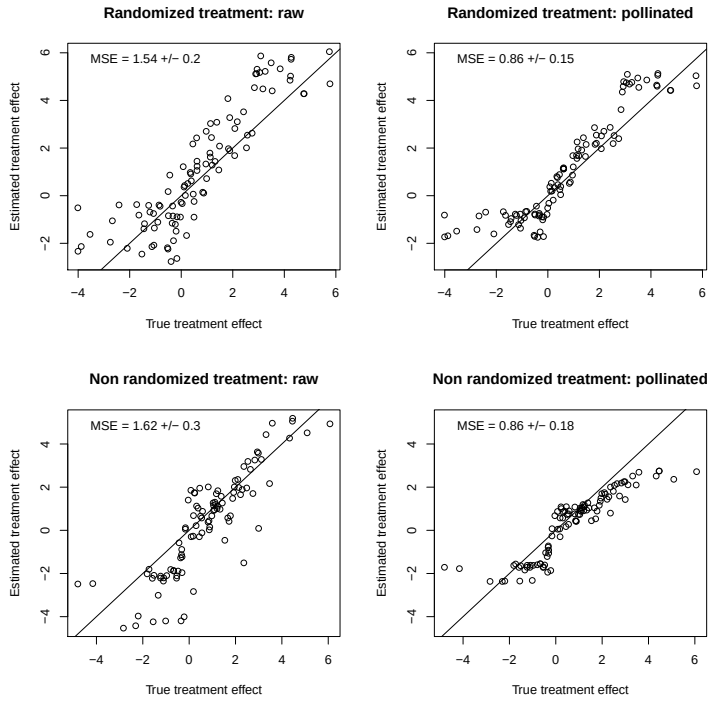
4. Grow a depth-controlled random forest G_{TOF} to δ_i.

5. Pollinate G_{TOF} separately with the data in the treated group and the control group to produce two regression forests G_1 and G_0, respectively. This entails sending each observation in the treatment group down each tree in the forest to determine its terminal node and re-estimating the response in that node to be the average of its observations. The same is done for the control group.

6. Compute $\delta_i = G_1(X_i) - G_0(X_i)$.

7. Optionally, fit a random forest S to δ_i and return S, which predicts the treatment effect $\hat{\tau}(x) = S(x)$. This optional layer of regression also helps with the interpretability of the results, yielding importance scores for variables as they relate directly to the estimated treatment effect.

Figure 2: *A comparison of raw and pollinated transformed outcome forests. Each method is applied to a randomized simulation and a non-randomized simulation, and we visually compare the estimated treatment effect with the true treatment effect We see that in each case, the pollination improves the estimates.*



Given data of the form $(X_i, Y_i), i = 1, ..., n$, least squares boosting starts with a regression function $\hat{F}(x) = 0$ and residuals $R_i = Y_i - \hat{F}(x_i)$. We fit a regression tree to R_i, yielding predictions $\hat{f}_1(x)$. Then we update $\hat{F}(x) \leftarrow \hat{F}(x) + \epsilon \cdot \hat{f}_1(x)$, and $R_i \leftarrow R_i - \epsilon \cdot \hat{f}_1(x_i)$ and repeat this (say) a few hundred times. The final prediction is simply $\hat{F}(x)$, a sum of trees shrunk by ϵ.

For our current problem, our data has the form $(X_i, T_i, Y_i), i = 1, ..., n$ with $T_i \in \{0, 1\}$. For now assume randomized treatment assignment. In the next subsection we show to handle the non-randomized case. Here is how we propose to generalize least squares boosting. As with causal forests Wager and Athey (2015), our building block is a causal tree, which returns a function $\hat{g}(x, t)$. The estimated causal effect for an observation $X = x$ is $\hat{\tau}(x) = \hat{g}(x, 1) - \hat{g}(x, 0)$. This is a standard causal tree, except that for each terminal node, we return the pair of treatment-specific means rather than the treatment effect. In other words, if observation $X_i = x$ gets you into terminal node k, where the pair of estimated means are $\hat{\mu}_{1k}$ (treated) and $\hat{\mu}_{0k}$ (untreated), then these are the values returned, respectively, for $\hat{g}(x, 1)$ and $\hat{g}(x, 0)$. The algorithm is summarized in Algorithm

2 below.

Algorithm 2: Causal Boosting

1. Set the outcome $R_i = Y_i$, and define $\hat{G}_0(x,t) = 0$.
2. Do $k = 1, ..., K$
   (a) Fit a causal tree $\hat{g}_k$ to data (X_i, R_i, T_i).
   (b) Set
$$\begin{aligned} R_i &\leftarrow R_i - \epsilon \cdot \hat{g}_k(X_i, T_i) \\ \hat{G}_k &\leftarrow \hat{G}_{k-1} + \epsilon \cdot \hat{g}_k. \end{aligned}$$
3. Return $\hat{G}_K(x, T)$.

The estimated treatment effect for any observation x is $\hat{G}_K(x,1) - \hat{G}_K(x,0)$.

Note that this generalizes to loss functions other than squared error. For example, if the causal tree was trained for a binary outcome, then each terminal node would return a pair of logits $\hat{\eta}_{1k} = \text{logit}[\Pr(Y = 1|X = x, T = 1)]$ and $\hat{\eta}_{0k} = \text{logit}[\Pr(Y = 1|X = x, T = 0)]$. Thus $\hat{G}_K(x,T)$ would be a function that returned a pair of logits at x, and hence treatment success probabilities. The treatment effect would be the appropriate function of these (difference, log-odds ratio). Other enhancements to boosting, such as stochastic boosting, are also applicable in the setting.

Note that causal boosting is not strictly a gradient boosting algorithm, because there is no loss function for which we are evaluating the gradient at each step, in order to minimize this loss. Rather, causal boosting is an adaptation of gradient boosting on the observed response, with a different function in each arm of the data. The adaptation is that we use causal trees as our weak learners instead of a standard regression technique. This tweak encourages the learned function to find treatment effect heterogeneities.

5.1 Cross-validation for causal boosting

Unlike random forests, gradient boosting algorithms can over-fit the training data as the number of trees increases (Hastie et al., 2009). This is because each successive tree is not built independently of the previous ones but rather with the goal of fitting to the residuals of the previous trees. Whereas a random forest will only benefit from using more trees, the number of trees in gradient boosting is itself an important parameter which needs to be tuned.

Complicating matters, the usual cross-validation framework does not apply to the setting of estimating a heterogeneous treatment effect because in this setting each observation does not come with a response corresponding directly to the function we are interested in estimating. We don't observe a response

τ_i for the i^{th} patient. What we observe is either $Y_i^{(0)}$ or $Y_i^{(1)}$, depending on whether or not the patient received the treatment.

We describe our approach in the context of a held-out validation set, but this fully specifies our cross-validation procedure. Cross-validation is simply validation done by partitioning the training set into several folds and averaging the results obtained by holding out each fold as a validation set and training on all other folds. The data in this context are a training set $(\mathbf{X}^{tr}, \mathbf{T}^{tr}, \mathbf{Y}^{tr})$ and a validation set $(\mathbf{X}^{v}, \mathbf{T}^{v}, \mathbf{Y}^{v})$. After training causal boosting on $(\mathbf{X}^{tr}, \mathbf{T}^{tr}, \mathbf{Y}^{tr})$, we are left with a sequence of models $G_1(x, T), ..., G_K(x, T)$, and we would like to evaluate the performance of each of these.

To validate the performance of each of these models, we use a pollination of the causal boosting model much like step 3 of the PTO forest. We run through the causal boosting algorithm again, making all the same splits as in the original training. The difference is in how we estimate the value returned in each node of each shallow causal tree. As in causal forests and in step 3 of the transformed outcome forest, we use $(\mathbf{X}^{(tr)}, \mathbf{T}^{(tr)}, \mathbf{Y}^{(tr)})$ to populate the nodes of the constituent causal trees and estimate the ATE within each node. The residuals r_i from the causal boosting algorithm are initialized to be the y_i from the validation set and are updated according to these re-fitted trees. The result is a new "honest" sequence of models $H_1(x, T), ..., H_K(x, T)$.

We are ready to define our validation error for each of the original models $G_1(x, T), ..., G_K(x, T)$. The validation error for a causal boosting model with k trees is given by

$$\sum_{x \in v} \left(\{G_k(x, 1) - G_k(x, 0)\} - \{H_K(x, 1) - H_K(x, 0)\} \right)^2 .$$

We have several remarks to make about this form. $G_k(x, 1) - G_k(x, 0)$ is the estimated treatment effect at x, for causal boosting with k trees. $H_K(x, 1) - H_K(x, 0)$ is the estimated treatment effect correspond to the maximum number of trees, *using the responses from the validation set*. For a large number of trees, we can be sure that this is over-fitting to the response, and this is the analog of traditional cross-validation, which compares predictions on the validation set with observed response in the validation set. This observed response, corresponding to the saturated model, is as over-fitted as possible. Intuitively, we are comparing our estimated treatment effect for each validation point against another estimate, which uses the same structure as the model fit to find similar patients and estimate the treatment effect based on those similar patients, some of whom will have received treatment, some of who will have received control. The better the structure is that causal boosting has learned for the heterogeneous treatment effect, the more the local ATE in the training set will mirror the local ATE in the validation set. For the results in Section 7, we use this procedure to do cross-validation for causal boosting.

5.2 Within-leaf propensity adjustment

When the goal is to estimate not an ATE but rather an individualized treatment effect, the propensity score methods described in Section 2.1 and in the appendix do not immediately extend. Consider for example propensity score stratification. Because each patient belongs to only one stratum of propensity score, we can not average treatment effect estimates for a patient across strata. Technically, if we were to fit a causal boosting model within each stratum, each of these models would be able to make a prediction for the query patient. But then all but one of these models would be unwisely extrapolating outside of its training set to make this prediction. An alternative to propensity score stratification, inverse probability weighting is still viable, but the volatility of this method is exacerbated by the attempt to estimate a varying treatment effect, rather than a constant one.

Within each leaf of a causal tree, however, we estimate an ATE. This is where causal boosting adjusts for non-random treatment assignment, using propensity score stratification to reduce the bias in the estimate of the within-leaf ATE. Before initiating the causal boosting algorithm, we begin by evaluating the propensity score for each patient, which is an estimate of probability of being assigned the treatment, conditioned on the observed covariates. Any binomial regression technique could be used here. We fit a probability forest (Malley et al., 2012), which is similar to a random forest for classification (Breiman, 2001) except that each tree returns a probability estimate rather than a classification. The trees are combined by averaging the probability estimates and not by majority vote. We denote the treatment assignment probability as a function of the covariates by $\pi(x) \equiv \mathbb{P}(T = 1|X = x)$ and the corresponding propensity scores by $\hat{\pi}_i \equiv \hat{\pi}(x_i)$.

We group the patients into S strata of similar propensity scores denoted $1, ..., S$. For example, there could be $S = 10$ strata, with the first comprising $\hat{\pi} \in [0, 0.1)$ and the last comprising $\hat{\pi} \in [0.9, 1]$, with equal-length intervals in between. We use $s_i \in \{1, ..., S\}$ to denote the stratum to which patient i belongs. Hence the data that we observe within each leaf of a causal tree are of the form $(X_i, s_i, T_i, Y_i) \in \mathbb{R}^p \times \{1, ..., S\} \times \{0, 1\} \times \mathbb{R}$. We use n_ℓ to denote the number of patients in leaf ℓ and index these patients by $i = 1, ..., n_\ell$. The propensity-adjusted ATE estimate in leaf ℓ is given by

$$\hat{\tau}_\ell = \frac{\sum_{s=1}^{S} n_{s\ell}(\bar{Y}_{1s\ell} - \bar{Y}_{0s\ell})}{\sum_{s=1}^{S} n_{s\ell}}, \text{ where } \bar{Y}_{ts\ell} = \frac{\sum_{i=1}^{n_\ell} \mathbb{I}_{\{T_i = t \wedge s_i = s\}} Y_i}{n_{ts\ell}} \tag{3}$$

is the mean response among the treatment ($t = 1$) or control ($t = 0$) group in stratum s, and $n_{ts\ell} = \sum_{i=1}^{n_\ell} \mathbb{I}_{\{s_i = s\}}$ is the corresponding number of patients in leaf ℓ for $t \in \{0, 1\}, s \in \{1, ..., S\}$. Finally, $n_{s\ell} = n_{1s\ell} + n_{0s\ell}$.

The estimated variance of $\hat{\tau}_\ell$ is

$$\widehat{\text{Var}}(\hat{\tau}_\ell) = \frac{\sum_{s=1}^{S} n_{s\ell}^2 \hat{\sigma}_{s\ell}^2}{(\sum_{s=1}^{S} n_{s\ell})^2}, \text{ where } \hat{\sigma}_{s\ell}^2 = \frac{s_{1s\ell}^2}{n_{1s\ell}} + \frac{s_{0s\ell}^2}{n_{0s\ell}},$$

and $s^2_{ts\ell}$ is the sample variance of the response for arm t of stratum s in leaf ℓ.

Hence, for two candidate daughter leaves ℓ and r of the same parent, The natural extension of the squared T-statistic splitting criterion from Athey and Imbens (2016) is

$$\frac{|\hat{\tau}_\ell - \hat{\tau}_r|}{\sqrt{\widehat{\mathrm{Var}}(\hat{\tau}_\ell) + \widehat{\mathrm{Var}}(\hat{\tau}_r)}}.$$

This is the propensity-stratified splitting criterion used by causal boosting. This criterion could also be used by a causal forest as it applies directly to its constituent causal trees.

We use this propensity adjustment not only for determining the split in a causal tree but also for estimating the treatment effect in the node. Specifically, the causal tree returns two values in each leaf: the propensity-adjusted mean response in the treatment and control groups.

$$\frac{\sum_{s=1}^{S} n_{s\ell}\bar{Y}_{1sl}}{\sum_{s=1}^{S} n_{s\ell}} \quad \text{and} \quad \frac{\sum_{s=1}^{S} n_{s\ell}\bar{Y}_{0sl}}{\sum_{s=1}^{S} n_{s\ell}}.$$

6 Causal MARS

One drawback to tree-based methods is that because they use the average treatment effect within each leaf as the prediction for that leaf, there could be high bias in this estimate. This is especially problematic when it comes to confidence interval construction for personalized treatment effects. The variance of the estimated treatment effect is relatively straightforward to estimate, but the bias presents more of a challenge.

Multivariate adaptive regression splines (MARS, Friedman (1991)) can be thought of as a modification to CART which alleviates this bias problem. MARS starts with the constant function $f(x) = \beta_0$ and considers adding pairs of functions of the form $\{(x_j - c)_+, (c - x_j)_+\}$ and also the products of variables in the model with these pairs, choosing the pair which lead to the greatest drop in training error when they are added to their model, with regression coefficients estimated via OLS. The difference between this and CART is that in CART the pairs of functions considered are of the form $\{\mathbb{I}_{\{x_j - c \geq 0\}}, \mathbb{I}_{\{c - x_j > 0\}}\}$, and when a product with one of the included terms in chosen, it replaces the included term in the model (Hastie et al., 2009).

We propose causal MARS as the adaptation of MARS to the task of treatment effect estimation. We fit two MARS models in parallel in the two arms (treatment and control) of the data, at each step choosing the same basis functions to add to each model. The criterion that we use identifies the best basis in terms of explaining treatment effect: we compare the drop in training error from including the basis in both models with different coefficients to the drop in training error from including the basis in both models with the *same* coefficient in each model. The steps of causal MARS are as follows. The parameter D

controls the maximum dimension of the regression basis, and in practice we use 11 in our examples. Algorithm 3 has the details.

Algorithm 3: Causal MARS

1. Define $\mathcal{F} = \{\{(x_j - c)_+, (c - x_j)_+\} : c \in \{\mathbf{X}_{ij}\}, j \in \{1, ..., p\}\}$.
2. Initialize $\mathcal{B} = \{1\}$.
3. For d in $1, ..., D$: (growing the model)
   (a) For each pair of functions $\{f, g\} \in \{\{b(x)f^*(x), b(x)g^*(x)\} : b \in \mathcal{B}, \{f^*, g^*\} \in \mathcal{F}\}$:
      i.
$$RSS_\mu = \min_{\beta^1, \beta^0} \sum_{i=1}^{n} \left(y_i - \sum_{b \in \mathcal{B}} (\beta_b^1 b(x_i) \mathbb{I}_{\{t_i=1\}} + \beta_b^0 b(x_i) \mathbb{I}_{\{t_i=0\}}) - \sum_{h \in \{f,g\}} \beta_h h(x_i) \right)^2$$
      ii.
$$RSS_\tau = \min_{\beta^1, \beta^0} \sum_{i=1}^{n} \left(y_i - \sum_{b \in \mathcal{B}} (\beta_b^1 b(x_i) \mathbb{I}_{\{t_i=1\}} + \beta_b^0 b(x_i) \mathbb{I}_{\{t_i=0\}}) - \sum_{h \in \{f,g\}} (\beta_h^1 h(x_i) \mathbb{I}_{\{t_i=1\}} + \beta_h^0 h(x_i) \mathbb{I}_{\{t_i=0\}}) \right)^2$$
      iii.
$$dRSS = RSS_\tau - RSS_\mu$$
   (b) Choose $\{f, g\}$ which maximize $dRSS$ and add them to $\mathcal{B}$.
4. Backward deletion: delete terms one at a time, using the same criterion as in the forward stepwise 3(a). Use the out-of-bag error to estimate the optimal model size.

To reduce the variance of causal MARS, we perform bagging by taking B bootstrap samples of the original dataset and fitting the causal MARS model to each one. The estimated treatment effect for an individual is the average of the estimates for this individual by the B models.

Note that the algorithm described above applies to the randomized case, not observational data. Given S propensity strata and membership $s \in 1, ..., S$, for each patient, we use the same basis functions within each stratum but different

regression coefficients. Within each stratum, the coefficients are estimated separately from the coefficients in other strata. Given the entry criterion $dRSS_s$ and number of patients n_s in each stratum, we combine these into a single criterion $\sum_s n_s dRSS_s$. This is the *propensity-adjusted* causal MARS.

6.1 Confidence intervals

One advantage of the bagging-based methods—causal forest and causal MARS—is that in the process of computing the treatment-effect estimates, one gets at no extra cost the computations necessary to estimate the variance of the estimators. Each of the bagged models is based on its own bootstrap re-sampling of the data, so for each patient we have B re-sampled treatment effect estimates, where B is the number of bags. We propose using the quantiles of these estimates as the confidence interval for each patient. To construct a $(1-\alpha)$ confidence interval, we use the $\alpha/2$ and $1-\alpha/2$ quantiles of the bootstrapped estimates as lower and upper bounds, respectively.

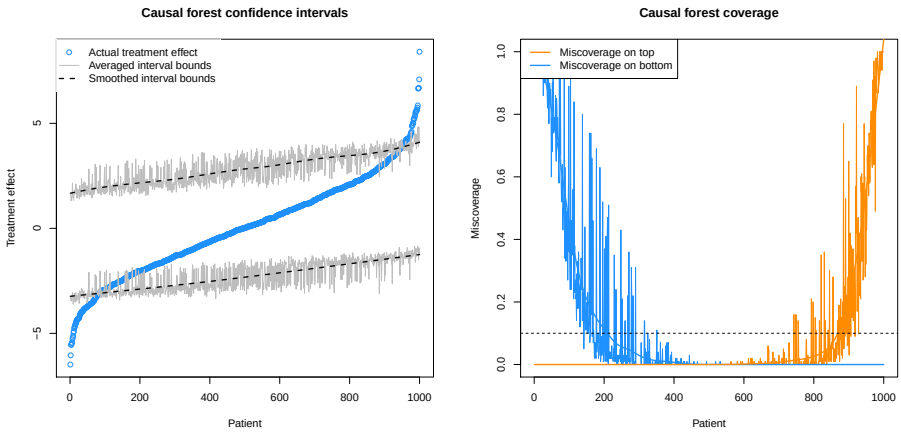
Note that this procedure is targeted at the variability of a single causal tree or a single causal MARS model, but the methods we propose involve averaging these models to reduce their variance. This will make our intervals more conservative because the variance of the bagged models will be lower than the variance of the individual models. However, as the results in this section demonstrate, the conservative nature of these confidence intervals helps with coverage problems due to the inability to fully remove the bias from the treatment effect estimates.

Figure 3 shows confidence interval results for causal forest applied to Simulation 8 in Section 7. That section describes in detail our simulation scheme, but in this section we use it only as an illustration of the confidence interval results. The left figure shows the average upper and lower bounds of the confidence interval for each patient, across 100 simulations. This demonstrates the difficulty with constructing confidence intervals for random forest predictions: Because of the relatively high bias from using the average as the estimate within each leaf, the confidence intervals do not come close to maintaining $(1-\alpha)$ coverage for patients with relatively small or relatively large treatment effects.

This problem for causal forests was the motivation for the development of causal MARS. By using piecewise linear models instead of piecewise constant models, MARS can achieve lower bias than regression trees, which is important for bootstrap confidence-interval construction. Figure 4 shows the results of constructing confidence intervals for the causal MARS estimates in a single simulation. The average confidence intervals are more volatile in Figure 4 than in Figure 3 because causal MARS is a higher-variance method. But we see that the confidence intervals adhere more closely to the true treatment effect for this method than for the causal forest. Examining the coverage, we see that there is still a bias problem for treatment effects near the edges of the range of values, but the miscoverage is closer to 0.5, an improvement of the coverage which approaches 1 for causal forest.

Still, bagged causal MARS has not fully mitigated the bias problem. We see

Figure 3: *Confidence intervals for causal forest in Scenario 8 from Section 7. On the left in blue we plot the true treatment effect for each patient against the index of the patient, sorted by treatment effect. The thin gray lines show the average upper and lower confidence interval bounds for each patient, and the dotted black line smooths over these averages. On the left the thin lines give the miscoverage rate for each patient, and the thick lines smooth over these thin lines. These results reflect 100 simulations using 50 bagged causal trees.*



that the miscoverage on bottom is decreasing with the true treatment effect, and the miscoverage on top is increasing with the true treatment effect. We attempted to address this with a bias correction. We bootstrapped residuals from the fitted model and applied a standard bootstrap bias correction. The results of this correction are shown in Figure 5. Here the confidence intervals adhere even more closely to the true treatment effect, and the coverage is improved. The miscoverage on either side of the confidence interval is capped at 0.2 when smoothed, though the target miscoverage rate is 0.1. We have taken steps toward constructing confidence intervals for personalized treatment effects, but it remains an area for future research.

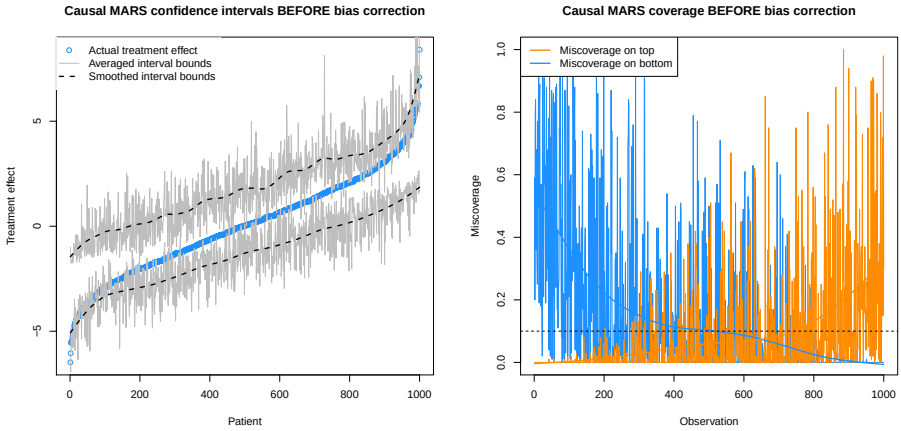
7 Simulation study

In the design of our simulations to evaluate performance of methods for heterogeneous treatment effect estimation, there are four elements to the generation of synthetic data:

1. The number n of patients in the training set, and the number p of features observed for each patient.

2. The distribution $\mathcal{D}_X$ of the feature vectors X_i. Across all scenarios, we draw odd-numbered features independently from a standard Gaussian dis-

Figure 4: *Confidence intervals for causal MARS in Scenario 8 from Section 7. On the left in blue we plot the true treatment effect for each patient against the index of the patient, sorted by treatment effect. The thin gray lines show the average upper and lower confidence interval bounds for each patient across 100 simulations, and the dotted black line smooths over these averages. On the left the thin lines give the miscoverage rate for each patient, and the thick lines smooth over these thin lines. These results reflect 100 simulations using 50 bagged causal MARS models.*



tribution. We draw even-numbered features independently from a standard Bernoulli distribution.

3. The propensity function $\pi(\cdot)$, the mean effect function $\mu(\cdot)$ and the treatment effect function $\tau(\cdot)$. We take the conditional mean effect functions to be $\mu_1(x) = \mu(x) + \tau(x)/2$ and $\mu_0(x) = \mu(x) - \tau(x)/2$.

4. The conditional variance σ_Y^2 of Y_i given X_i and T_i. This corresponds to the noise level, and we choose is to make the percentage of null variance explained of the true model to be roughly 20-25%. This ensures we are comparing the methods on relevant simulations.

Given the elements above, our data generation model is, for $i = 1, ..., n$:

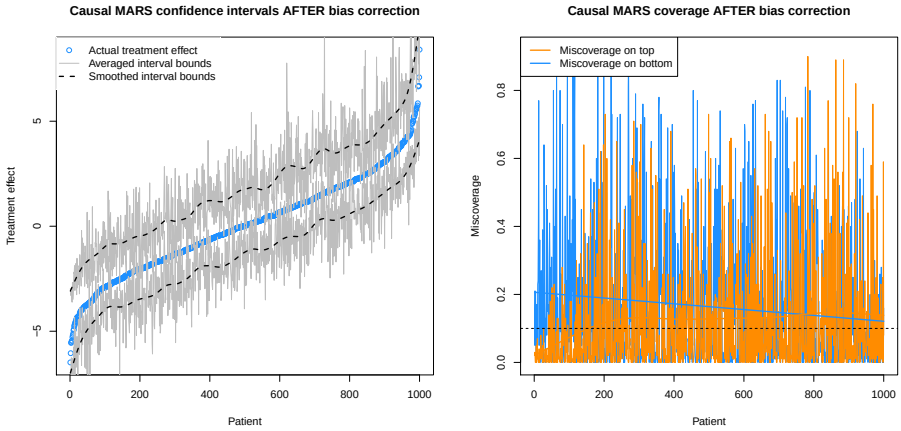
$$X_i \overset{\text{i.i.d.}}{\sim} \mathcal{D}_X$$

$$T_i \overset{\text{ind.}}{\sim} \text{Bernoulli}(\pi(X_i))$$

$$Y_i \overset{\text{ind.}}{\sim} \text{Normal}\left(\mu(X_i) + (T_i - 1/2)\tau(X_i), \sigma_Y^2\right)$$

The third element above, encompassing $\pi(\cdot)$, $\mu(\cdot)$ and $\tau(\cdot)$, is most interesting. Note that $\pi(\cdot)$ and $\mu(\cdot)$ are nuisance functions, and $\tau(\cdot)$ is the function we

Figure 5: *Bias-corrected confidence intervals for causal MARS in Scenario 8 from Section 7. On the left in blue we plot the true treatment effect for each patient against the index of the patient, sorted by treatment effect. The thin gray lines show the average upper and lower confidence interval bounds for each patient across 100 simulations, and the dotted black line smooths over these averages. On the left the thin lines give the miscoverage rate for each patient, and the thick lines smooth over these thin lines. These results reflect 100 simulations using 50 bagged causal MARS models.*



are interested in estimating. In this section, we present two batches of simulations, the first of which represent randomized experiments. The second batch of simulations represent observational studies. Within each set of simulations, we make eight different choices of mean effect function and treatment effect function, meant to represent a wide variety of functional forms: both univariate and multivariate; both additive and interactive; both univariate and multivariate.

Table 1: *Specifications for the 16 simulation scenarios. The four rows of the table correspond, respectively, to the sample size, dimensionality, mean effect function, treatment effect function and noise level. Simulations 1 through 8 use randomized treatment assignment, meaning $\pi(x) = 1/2$. Simulations 9 through 16 have a bias in treatment assignment, specified by (4).*

Scenarios	1, 9	2, 10	3, 11	4, 12	5, 13	6, 14	7, 15	8, 16
n	200	200	300	300	400	400	1000	1000
p	400	400	300	300	200	200	100	100
$\mu(x)$	$f_8(x)$	$f_5(x)$	$f_4(x)$	$f_7(x)$	$f_3(x)$	$f_1(x)$	$f_2(x)$	$f_6(x)$
$\tau(x)$	$f_1(x)$	$f_2(x)$	$f_3(x)$	$f_4(x)$	$f_5(x)$	$f_6(x)$	$f_7(x)$	$f_8(x)$
σ_Y^2	1	1/4	1	1/4	1	1	4	4

The eight functions that we chose are:

$$f_1(x) = 0 \qquad f_2(x) = 5\mathbb{I}_{\{x_1>1\}} - 5 \qquad f_3(x) = 2x_1 - 4$$

$$\begin{aligned} f_4(x) = {} & x_2x_4x_6 + 2x_2x_4(1-x_6) + 3x_2(1-x_4)x_6 + 4x_2(1-x_4)(1-x_6) + 5(1-x_2)x_4x_6 \\ & +6(1-x_2)x_4(1-x_6) + 7(1-x_2)(1-x_4)x_6 + 8(1-x_2)(1-x_4)(1-x_6) \end{aligned}$$

$$f_5(x) = x_1 + x_3 + x_5 + x_7 + x_8 + x_9 - 2$$

$$f_6(x) = 4\mathbb{I}_{\{x_1>1\}}\mathbb{I}_{\{x_3>0\}} + 4\mathbb{I}_{\{x_5>1\}}\mathbb{I}_{\{x_7>0\}} + 2x_8x_9$$

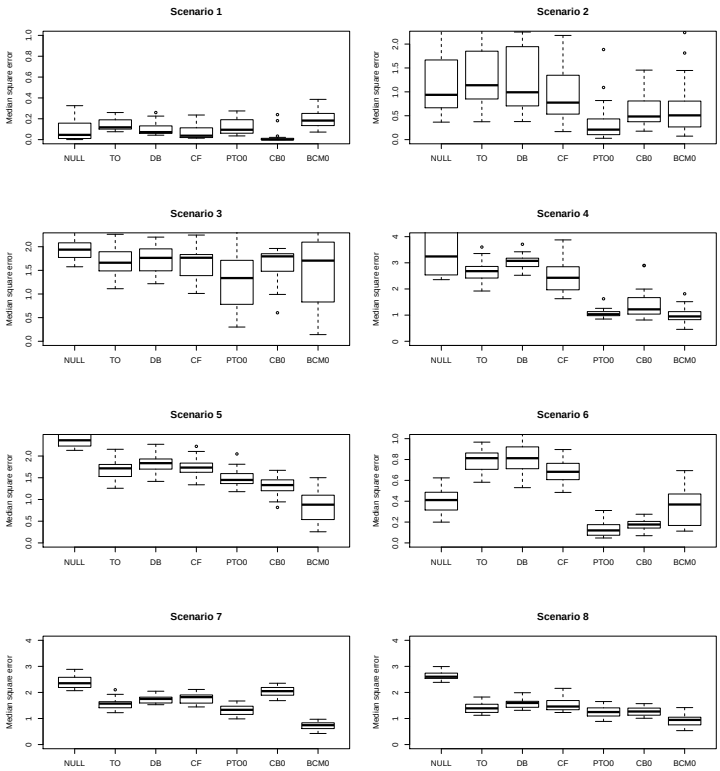
$$f_7(x) = \frac{1}{2}\left(x_1^2 + x_2 + x_3^2 + x_4 + x_5^2 + x_6 + x_7^2 + x_8 + x_9^2 - 11\right)$$

$$f_8(x) = \frac{1}{\sqrt{2}}\left(f_4(x) + f_5(x)\right)$$

Each of the eight functions above is centered and scaled so that with respect to the distribution $\mathcal{D}_X$, each has mean close to zero and all have roughly the same variance. Table 1 gives the mean and treatment effect functions for the eight randomized simulations, in terms of the eight functions above. In these simulations $\pi(x) = 1/2$ for all $x \in \mathbb{R}^p$. In addition to the methods described in Sections 4, 5 and 6, we include results for two additional estimators for comparison. The null estimator is simply the difference $\bar{Y}_1 - \bar{Y}_0$ in mean response between treated and untreated patients. This provides a naive baseline. The other competitor is the gradient forest of Athey et al. (2017), using the `gradient.forest` R package made available online by the authors. The results of the fist batch of simulations are shown in Figure 6.

If we pick "winners" in each of the simulation scenario based on which method has the lowest distribution of errors, causal MARS would win Scenarios

Figure 6: *Results across eight simulated randomized experiments. For details of the generating distributions, see Table 1. The seven estimators being evaluated are: NULL = the null prediction, GF = gradient forest, PTO0 = pollinated transformed outcome forest (using propensity = 1/2), CB0 = causal boosting, CM0 = causal MARS. The vertical blue bar shows the standard deviation of the response, for assessing the practical significance of the difference between the methods' performances.*



5, 7 and 8, tying with the pollinated transformed outcome forest in Scenario 4. The PTO forest would win Scenarios 2 and 3, tying with causal boosting in Scenario 6. In general all of the methods outperform the null estimator except in Scenario 1, when the treatment effect is constant, and in Scenario 6, when the gradient forest perform worst.

The second batch of simulations matches the parameters listed in Table 1: Scenario 9 is like Scenario 1; Scenario 10 is like Scenario 2; and so on. The difference is in the propensity function. For this second batch of simulations, we use

$$\pi(x) = \frac{e^{\mu(x)-\tau(x)/2}}{1+e^{\mu(x)-\tau(x)/2}}. \tag{4}$$

The interpretation of this propensity function is that patients with greater mean effect are more likely to receive the treatment. This resembles a situation in which greater values of the outcome are worse for the patient, and only patients who have need for treatment will receive it. There are many possible forms for the propensity function, but we focus on this one because it is particularly troublesome, and a good estimator of the treatment effect needs to avoid the pitfall of estimating to great an effect because the treated patients have greater mean effect. This is exactly the kind of bias we are most concerned about in observational studies. The results of this second batch of simulations are shown in Figure 7.

In the batch of simulations with biased treatment assignments, propensity-adjusted causal boosting shines. In six of the eight simulations, causal boosting as either the lowest error distribution or is one of the two methods with the lowest error distribution. Curiously, in Scenario 13, unadjusted causal MARS performs very well, but the propensity adjustment ruins this performance. In Scenario 15, PTO forest and gradient forest produce the best results though all of the methods perform well. Overall, across the 16 simulation scenarios, causal boosting and causal MARS stand out as having the best performance.

8 Application

In September 2016, *New England Journal of Medicine* opened The SPRINT Data Analysis Challenge, based on the complete dataset from a randomized trial of a novel intervention for the treatment of high blood pressure (SPRINT Research Group, 2015). The goal was open-ended: to draw novel or clinically useful insights from the SPRINT dataset, possibly in tandem with other publicly available data.

The intervention in the randomized trial (SPRINT Research Group, 2015) was a more intensive control of systolic blood pressure (target 120 mm Hg) than is standard (target 140 mm Hg). The primary outcome of interest was whether the patient experienced any of the following events: myocardial infarction (heart attack), other acute coronary syndrome, stroke, heart failure or death from cardiovascular causes. The trial, which enrolled 9361 patients, ended after a median follow-up period of 3.26 years, when researchers determined at a pre-planned checkpoint that the population-average outcome for the intensive treatment group (1.65% incidence per year) was significantly better than that of the standard treatment group (2.19% incidence per year).

In addition to the primary event, for each patient researchers tracked several other adverse events, as well as 20 baseline covariates recorded at the moment of treatment assignment randomization: 3 demographic variables, 6 medical history variables and 11 lab measurements. The question that we seek to answer in this section is whether we can use these variables to give personalized estimates of treatment effect which are more informative than the population-level average treatment effect. To answer this question, we apply the gradient forest and causal MARS to these data.

Figure 7: *Results across eight simulated observational studies, in which treatment is more likely to be assigned to those with a greater mean effect. The seven estimators being evaluated are: NULL = the null prediction, GF = gradient forest, PTO = pollinated transformed outcome forest, CB1 = causal boosting (propensity adjusted), CB0 = causal boosting, CM1 = causal MARS (propensity adjusted), CM0 = causal MARS. CB0 and CM0 are in gray because they would not be used in this setting. They are provided for reference to assess the effect of the propensity adjustment. The vertical blue bar shows the standard deviation of the response, for assessing the practical significance of the difference between the methods' performances.*

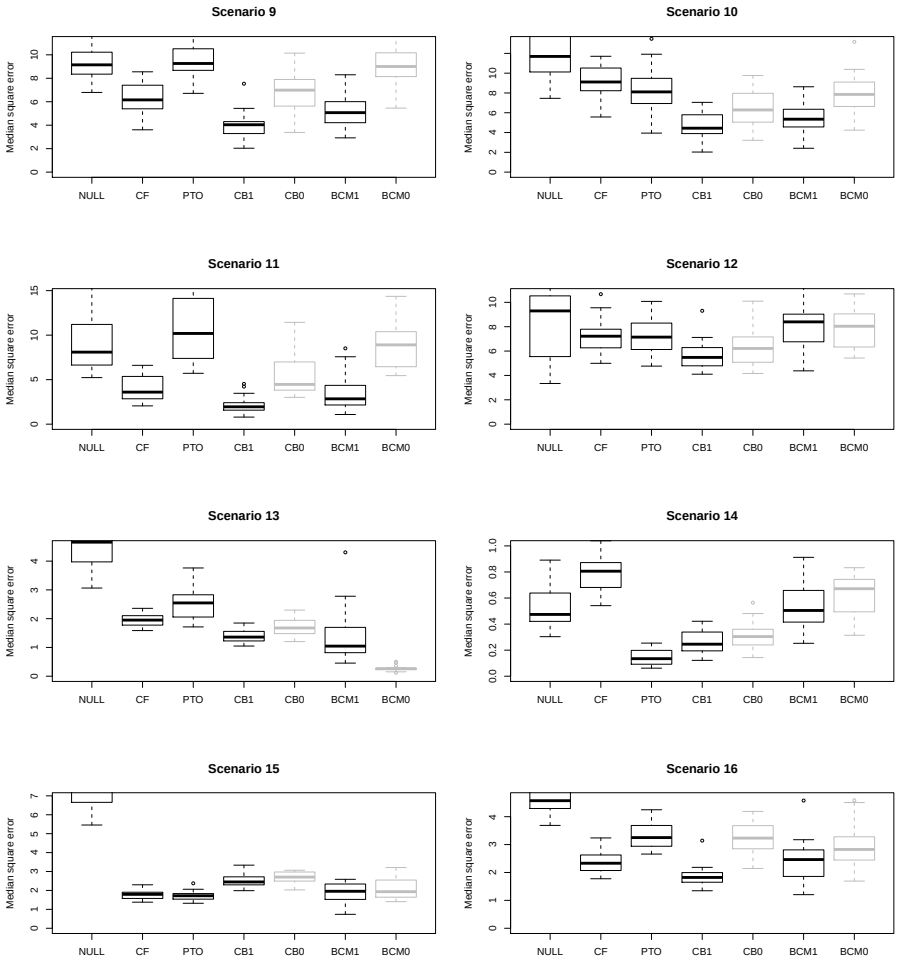
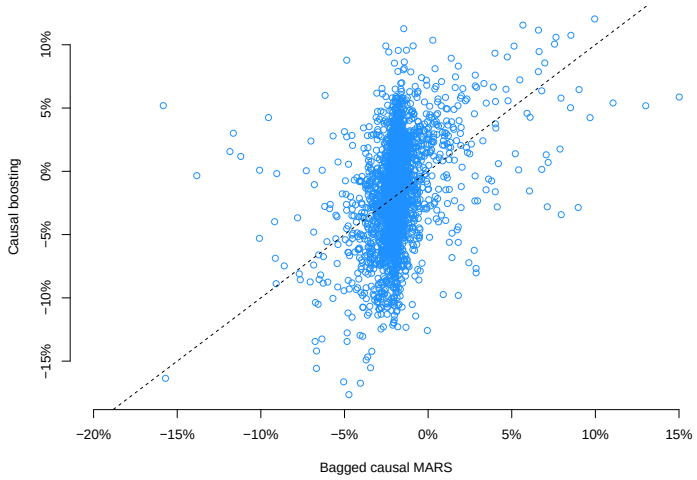


Figure 8: *Personalized treatment effect estimates from causal boosting and causal MARS. Each circle represents a patient, who gets a personalized estimate from each method. The dashed line represents the diagonal, along which the two estimates are the same.*

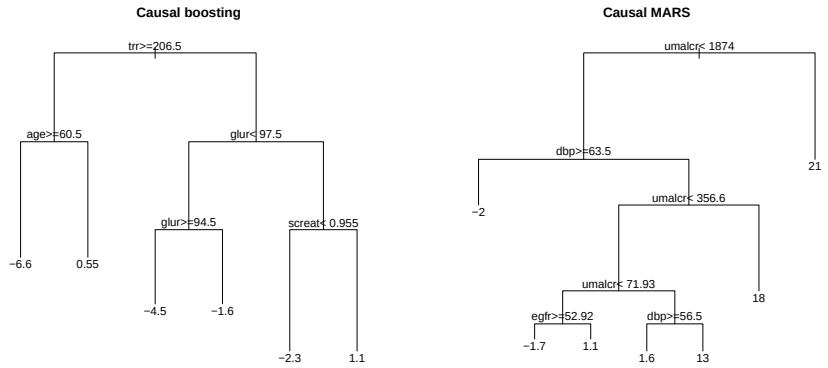


Of the 9361 patients who underwent randomization, 1172 (12.5%) died, discontinued intervention, withdrew consent or were lost to follow-up before the conclusion of the trial. There is little evidence (χ^2 p-value = 31%) that this censorship was more common in either arm of the trial. To extract a binary outcome from these survival data, we use as our response the indicator that a patient experiences the primary outcome within 1000 days of beginning treatment, ignoring patients who were censored before 1000 days. Additionally, we dropped the 1.8% of patients who have at least one lab measure missing. This leaves us with a sample of 7344 patients, which we split into equally sized training and validation sets.

The results of fitting causal boosting and causal MARS on the training sample of 3672 patients are shown in Figure 8. We observe that the two methods yield very different distributions of estimated personalized treatment effects in the aggregate. Causal boosting produces estimates resembling a normal distribution with a standard deviation of about 3.5% risk. In contrast, causal MARS estimates almost all patients to have a treatment effect between −5% risk and +0% risk, but for a small percentage of patients the treatment effect is much greater or much lesser. The tails of this distribution are much heavier than that of a normal distribution. In fact, a very small number of patients (0.4% of the training sample) are not included in this figure because their treatment effect estimate from causal MARS falls outside of the plotted region.

Figure 9 depicts decision trees which summarize the key inferences made by causal boosting and causal MARS. Each leaf gives the average estimated treatment effect for patients who belong to that leaf. Such a decision could be

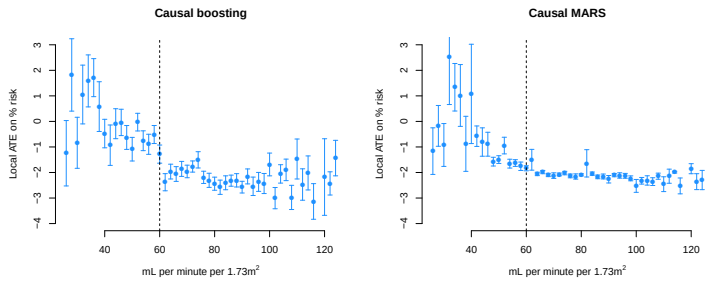
Figure 9: *Decision trees summarizing with broad strokes the inferences of causal boosting and causal MARS. The variables are:* `trr` *triglcerides (mg/dL) from blood draw;* `age` *(years) age at beginning of trial;* `glur` *glucose (mg/dL) from blood draw;* `screat` *creatinine (mg/dL) from blood draw;* `umalcr` *albumin/creatinine ratio from urine sample;* `dbp` *diastolic blood pressure (mm Hg);* `egfr` *estimated glomerular filtration rate (mL/min/1.73m^2). If the inequality at a split is true for a patient, then that patient belongs to the left daughter node.*



reported to a physician to explain the basis for these personalized treatment effect estimates. According to causal boosting, for example, older patients with high triglycerides stand to gain more from the intensive blood pressure treatment than younger patients with high triglycerides. Among patients with low triglycerides and high glucose, those with low creatinine stand to benefit more from the intensive treatment than those with high creatinine. The decision tree for causal MARS makes the extreme claim that for patients with urine albumin/creatinine ratio above 1874, the average treatment effect is a 21% increase in risk. Discussions with practitioners suggest that the distribution of personalized treatment effects estimated by causal boosting is more plausible than that of causal MARS. As such, we focus our interpretation on the results of causal boosting for the reminder of this section.

To simplify the results even more than the decision tree does, we note that for both causal boosting and causal MARS, the two features which correlate most to the personalized treatment effect estimates are estimated glomerular filtration rate (eGFR) and creatinine. These two variables are highly correlated with each other, as creatinine is one of the variables used to estimate GFR. Both are used to assess kidney health, and patients with eGFR below 60 are considered to have chronic kidney disease. Figure 10 shows the relationship between eGFR and the estimated personalized treatment effect from causal boosting. Despite there being no manual notation in the data that there is something special about an eGFR of 60, we have learned from causal boosting that patients below

Figure 10: *Training set personalized treatment effects, estimated via causal boosting, versus estimated glomerular filtration rate (eGFR). Patients are stratified according to eGFR on the x-axis, and each point gives the average personalized treatment effect among patients in that stratum. Error bars correspond to one standard error for the mean personalized treatment effect. The vertical dashed line represents a medical cutoff, below which patients are considered to suffer from chronic kidney disease.*



this cutoff have less to gain from the intensive blood pressure treatment than patients above this cutoff.

Note that we are not only interested in whether a patient's personalized treatment effect is positive or negative. Intensive control of blood pressure comes with side effects and should only be assigned to patients for whom the benefit of reducing the risk of an adverse coronary event is substantial. The results of causal boosting on the training set would suggest that patients with chronic kidney disease have less to gain from this treatment than do other patients.

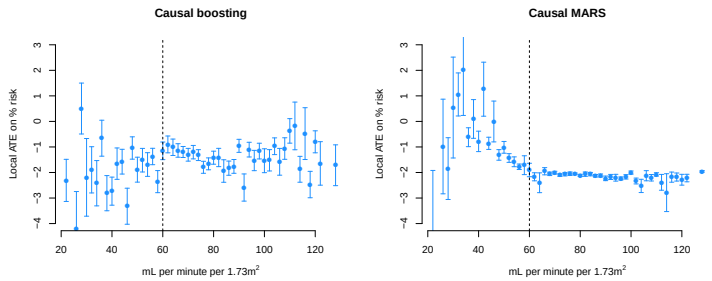
8.1 Validation

The results above tell an interesting story: If you are a patient with chronic kidney disease (eGFR < 60), you are expected to benefit less from intensive blood pressure control. As discussed in Section 5.1, validating treatment effect estimates is challenging because we do not observe the treatment effect for any individual patient. In this section, we make an attempt to validate the more general conclusion from the previous section: that the treatment has less benefit for patients with chronic kidney disease.

Figure 11 shows the results of fitting causal boosting on the held-out validation set of 3672 patients. We see that the relationship between eGFR and estimated treatment effect does not tell the same story as in the training set. In fact, there is no clear relationship between these two variables in the validation set.

It is possible that we have insufficient power in the validation set to identify the relationship between eGFR and treatment effect and that with a larger sample of patients, we would have validated our conclusions from the training set. It is worth noting that the team from Boston University which placed second

Figure 11: *Validation set personalized treatment effects, estimated via causal boosting, versus estimated glomerular filtration rate (eGFR). Patients are stratified according to eGFR on the x-axis, and each point gives the average personalized treatment effect among patients in that stratum. Error bars correspond to one standard error for the mean personalized treatment effect. The vertical dashed line represents a medical cutoff, below which patients are considered to suffer from chronic kidney disease.*



in the SPRINT Data Analysis Challenge made the same finding as shown in the causal boosting results. They found that intensive blood pressure management does not improve primary outcomes for patients with chronic kidney disease. Something that the authors do not address is why they chose to analyze patients with chronic kidney disease. Presumably they used some combination of prior medical knowledge and manual hypothesis selection. In our training set, we came to the same conclusion using causal boosting without the benefit of either of these steps. The dissimilar results on the validation set could be explained by insufficient power.

9 Discussion

We have proposed and compared a number of different methods for estimating heterogeneous treatment effects from high-dimensional covariates. The causal boosting and causal MARS approaches seem particularly promising. More work is needed in refining and testing these methods, and in the construction of reliable confidence intervals for the estimated effects.

10 Acknowledgments

The authors would like to thank Jonathan Taylor and Stefan Wager for helpful discussions, and Susan Athey, Julie Tibshirani and Stefan for sharing their causal forest code.

References

Anderson, K. M., Odell, P. M., Wilson, P. W. and Kannel, W. B. (1991), 'Cardiovascular disease risk profiles', *American Heart Journal* **121**(1), 293–298.

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

Athey, S., Tibshirani, J. and Wager, S. (2017), Solving heterogeneous estimating equations with gradient forests.

Austin, P. C. (2011), 'An introduction to propensity score methods for reducing the effects of confounding in observational studies', *Multivariate Behavioral Research* **46**, 399–424.

Bonetti, M. and Gelber, R. D. (2004), 'Patterns of treatment effects in subsets of patients in clinical trials', *Biostatistics* **5**(3), 465–481.

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

Chen, W., Ghosh, D., Raghunathan, T. E., Norkin, M., Sargent, D. J. and Bepler, G. (2012), 'On Bayesian methods of exploring qualitative interactions for targeted treatment', *Statistics in Medicine* **31**(28), 3693–3707.

Chipman, H. A., George, E. I. and McCulloch, R. E. (1998), 'Bayesian CART model search', *Journal of the American Statistical Assocation* **93**(443), 935–948.

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

Friedman, J. (1991), 'Multivariate adaptive regression splines', *The Annals of Statistics* **19**(1), 1–67.

Friedman, J. (2001), 'Greedy function approximation: A gradient boosting machine', *The Annals of Statistics* **29**(5), 1189–1232.

Gail, M. and Simon, R. (1985), 'Testing for qualitative interactions between treatment effects and patient subsets', *Biometrics* **41**(2), 361–372.

Green, D. P. and Kern, H. L. (2012), 'Modeling heterogeneous treatment effects in survey experiments with Bayesian additive regression trees', *Public Opinion Quarterly* **76**(3), 491–511.

Gustafson, P. (2000), 'Bayesian regression modeling with interactions and smooth effects', *Journal of the American Statistical Association* **95**(451), 795–806.

Hastie, T. J., Tibshirani, R. and Friedman, J. (2009), *The Elements of Statistical Learning: Data mining, inference and prediction*, Springer Series in Statistics, 2nd edn, Springer.

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

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

LeBlanc, M. (1995), 'An adaptive expansion method for regression', *Statistica Sinica* **5**(2), 737–748.

Low, Y. S., Gallego, B. and Shah, N. H. (2016), 'Comparing high-dimensional confounder control methods for rapid cohort studies from electronic health records', *Journal of Comparative Effectiveness Research* **5**(2), 179–192.

Malley, J. D., Kruppa, J., Dasgupta, A., Malley, K. G. and Ziegler, A. (2012), 'Probability machines: consistent probability estimation using nonparametric learning machines', *Methods of Information in Medicine* **51**(1), 74–81.

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

Rubin, D. B. (1974), 'Estimating causal effects of treatments in randomized and nonrandomized studies', *Journal of Educational Psychology* **66**(5), 688–701.

Sauerbrei, W., Royston, P. and Zapien, K. (2007), 'Detecting an interaction between treatment and a continuous covariate: A comparison of two approaches', *Computational Statistics and Data Analysis* **51**(8), 4054–4063.

Shah, N. H. (2016), 'Performing an informatics consult'. Big Data in Biomedicine Conference — Stanford Medicine.
URL: *http://bigdata.stanford.edu/pastevents/2016-presentations.html*

Splawa-Neyman, J., Dabrowska, D. M. and Speed, T. P. (1990), 'On the application of probability theory to agricultural experiments. Essay on principles. Section 9.', *Statistial Science* **5**(4), 465–472.

SPRINT Research Group (2015), 'A randomized trial of intensive versus standard blood-pressure control', *New England Journal of Medicine* **373**(22), 2103–2116.

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

Taddy, M., Gardner, M., Chen, L. and Draper, D. (2016), 'A nonparametric Bayesian analysis of heterogenous treatment effects in digital experimentation', *Journal of Business & Economic Statistics* **34**(4), 661–672.

Tian, L., Alizadeh, A. A., Gentles, A. J. and Tibshirani, R. (2014), 'A simple method for estimating interactions between a treatment and a large number of covariates', *Journal of the American Statistical Association* **109**(508), 1517–1532.

Wager, S. and Athey, S. (2015), Estimation and inference of heterogeneous treatment effects using random forests.

Wright, M. N. and Ziegler, A. (2015), ranger: A fast implementation of random forests for high dimensional data in C++ and R.

Xie, Y., Brand, J. E. and Jann, B. (2012), 'Estimating heterogeneous treatment effects with observational data', *Sociological Methodology* **42**(1), 314–347.

Zhao, Y., Zeng, D., Rush, A. J. and Kosorok, M. R. (2012), 'Estimating individualized treatment rules using outcome weighted learning', *Journal of the American Statistical Association* **107**(499), 1106–1118.

A Appendix

In this appendix we outline the already-established techniques for using propensity score to adjust for bias in treatment assignment for observational studies in which the goal is to estimate a population-average treatment effect (ATE). Define $f(x)$ the marginal feature density, $f_1(x)$ the conditional density of X given $T = 1$ (and likewise $f_0(x)$), where T is binary treatment indicator, and let $\pi_1 = \mathbb{P}(T = 1)$ be the marginal proportion of treated. Let $\mu_1(X) = \mathbb{E}[Y|T = 1, X]$, and likewise $\mu_0(X)$, and $\tau(X) = \mu_1(X) - \mu_0(X)$. Finally, let $\pi(X) = \mathbb{P}(T = 1|X)$ be the treatment propensity.

Transformed outcome averaging

Note that the *transformed outcome*

$$Z \equiv T\frac{Y}{\pi(X)} + (1 - T)\frac{-Y}{1 - \pi(X)}$$

satisfies

$$\begin{aligned}\mathbb{E}[Z|X] &= \mathbb{P}(T = 1|X)\frac{1}{\pi(x)}\mathbb{E}[Y|T = 1, X] - \mathbb{P}(T = 0|X)\frac{1}{1 - \pi(x)}\mathbb{E}[Y|T = 1, X] \\ &= \mathbb{E}[Y|T = 1, X] - \mathbb{E}[Y|T = 0, X] = \mu_1(X) - \mu_0(X) = \tau(X).\end{aligned}$$

Hence if the expectation of Z is evaluated with respect to the distribution of X,

$$\mathbb{E}_X[Z] = \mathbb{E}_X[\mathbb{E}[Z|X]] = \mathbb{E}_X[\tau(X)].$$

In other words, the transformed outcome is unbiased for the ATE. So a natural estimator for the ATE in a sample of patients would be the sample mean of the transformed outcome. This justifies for example using Z as a response to grow a random forest in our pollinated transformed outcome forest.

Propensity score stratification

Note that it is not necessarily the case that $E[Y|T=1] = E[\mu_1(X)|T=1]$ and $E_X[\mu_1(X)]$ are the same; it is possible that conditioning on T changes the distribution of X and consequently the distribution of $\mu_1(X)$. This is the essence of why we cannot ignore non-randomized treatment assignment in observational studies. However, it is the case that

$$\mathbb{E}[Y|T=1,\pi(X)] = \mathbb{E}[\mu_1(X)|\pi(X)].$$

To see this, note that $X \perp\!\!\!\perp T|\pi(X)$ because by assumption $T \sim \text{Binomial}(1,\pi(X))$. Hence the conditional distribution of X given $\pi(X)$ and T is the same as the conditional distribution of X given $\pi(X)$. This implies that

$$\mathbb{E}[Y|T=1,\pi(X)] = \mathbb{E}\left\{\mathbb{E}[Y|T=1,X] \;\middle|\; T=1,\pi(X)\right\} = \mathbb{E}[\mu_1(X)|\pi(X)].$$

What this says is that for fixed $\pi(X)$, the mean response under treatment is unbiased for the conditional expectation of $\mu_1(X)$. This equality holds for any value of X, so the expectations of these two quantities are the same with respect to the distribution of $\pi(X)$:

$$\mathbb{E}_{\pi(X)}\left[\mathbb{E}[Y|T=1,\pi(X)]\right] = \mathbb{E}_{\pi(X)}\left[\mathbb{E}[\mu_1(X)|\pi(X)]\right] = \mathbb{E}_X[\mu_1(X)].$$

This leads to the following estimator for $\mathbb{E}_X[\mu_1(X)]$: Compute the average response for all treated patients for each value of the propensity, and integrate with respect to the distribution of the propensity. In practice, we approximate this by using a rough approximation to the distribution of $\pi(X)$: Define strata (or bins) of the propensity score, for example (0, 0.1], ..., (0.9, 1). Within each stratum, find the average response among treated patients. Then combine these values in a weighted average, weighting according to the frequency of each stratum. This is our estimate of $\mathbb{E}_X[\mu_1(X)]$. We follow the same procedure in the control arm to estimate $\mathbb{E}_X[\mu_0(X)]$, and the difference is our estimate of $\mathbb{E}_X[\tau(X)]$.

Inverse probability weighting

From Bayes' theorem, $f_1(x) = f(x)\pi(x)/\pi_1$. Consider weighting this density with weights proportional to $1/\pi(x)$. The density of this weighted distribution is given by

$$\tilde{f}_1(x) = \frac{\frac{1}{\pi(x)} f(x)\pi(x)/\pi_1}{\int_{\mathbb{R}} \frac{1}{\pi(x)} f(x)\pi(x)/\pi_1} dx = \frac{f(x)/\pi_1}{\int_{\mathbb{R}} f(x)/\pi_1 dx} = \frac{f(x)/\pi_1}{1/\pi_1} = f(x).$$

Hence the weighted conditional distribution of X given $T=1$ is the same as the marginal distribution of X. So the expectation of any function of X with respect to this distribution is the same as with respect to the marginal distribution of X. Specifically, using $\tilde{X}$ to denote the random variable following the weighted density $\tilde{f}_1(x)$,

$$\mathbb{E}_{\tilde{X}}[\mu_1(\tilde{X})] = \mathbb{E}_X[\mu_1(X)].$$

Based on this result, we use the sample mean of the response in the treatment arm, with weights proportional to the inverse of the propensity, as an unbiased estimator for $\mathbb{E}_X[\mu_1(X)]$. Similarly, in the control arm we use weights proportional to $1/(1-\pi(x))$ to get an unbiased estimate for $\mathbb{E}_X[\mu_0(x)]$. The difference between these two is our estimate for $\mathbb{E}_X[\tau(X)]$