Learning Conditional Average Treatment Effects in Regression Discontinuity Designs using Bayesian Additive Regression Trees

Rafael Alcantara[1], P. Richard Hahn[2], Carlos Carvalho[1], and Hedibert Lopes[3]

[1]The University of Texas at Austin
[2]Arizona State University
[3]Insper

Abstract

BART (Bayesian additive regression trees) has been established as a leading supervised learning method, particularly in the field of causal inference. This paper explores the use of BART models for learning conditional average treatment effects (CATE) from regression discontinuity designs, where treatment assignment is based on whether an observed covariate (called the running variable) exceeds a pre-specified threshold. A purpose-built version of BART that uses linear regression leaf models (of the running variable and treatment assignment dummy) is shown to outperform off-the-shelf BART implementations as well as a local polynomial regression approach and a CART-based approach. The new method is evaluated in thorough simulation studies as well as an empirical application looking at the effect of academic probation on student performance.

Keywords: Bayesian additive regression trees, BART, CATE, conditional average treatment effects, RDD, regression discontinuity design

Rafael Alcantara gratefully acknowledges financial support from FAPESP grant 2021/13137-0 during his PhD in Business Economics at Insper under the supervision of Hedibert Lopes. Hedibert Lopes also acknowledges partial financial support from FAPESP grants 2023/02538-0 and 2024/01027-4.

1 Introduction

Regression discontinuity designs (RDD), originally proposed by Thistlethwaite and Campbell (1960), are widely used in economics and other social sciences to estimate treatment effects from observational data. Such designs arise when treatment assignment is based on whether a particular covariate — referred to as the running variable — lies above or below a known value, referred to as the cutoff value. Because treatment is deterministically assigned as a known function of the running variable, RDDs are trivially deconfounded: treatment assignment is independent of the outcome variable, given the running variable (because treatment is conditionally constant).

However, estimation of treatment effects in RDDs is more complicated than simply controlling for the running variable, because doing so introduces a complete lack of overlap, which is the other key condition needed to justify regression adjustment for causal inference. Nonetheless, treatment effects *at the cutoff* may still be identified. Specifically, it is well-known that treatment effects at the cutoff can be estimated from RDDs as the magnitude of a discontinuity in the conditional mean response function at that point (Hahn et al., 2001).

This paper investigates the use of Bayesian additive regression tree models (Chipman et al., 2010; Hahn et al., 2020) for the purpose of estimating conditional average treatments effects (CATE) at the cutoff, conditional on observed covariates other than the running variable. To the best of our knowledge, such data-driven CATE estimation has not been a focus of the existing RDD literature and we are the first to propose BART for this purpose.

1.1 Previous work

The inclusion of covariates in RDD models has been studied by a great number of authors, but largely from the perspective of obtaining precision gains for average treatment effect (ATE) estimation (at the cutoff), mostly in the context of linear models, and mostly from a frequentist perspective. To our knowledge, only two previous works look at nonlinear data-driven CATE estimation. Becker et al. (2013) extend the traditional local regression to include interaction terms between the treatment dummy and smooth basis functions of additional covariates — a direct extension of traditional RDD methods for the ATE. Reguly (2021) propose a modified CART (classification and regression tree) algorithm in which the tree is split using all features available *except* for the running variable; then, within each leaf the algorithm performs a separate regression for treated and untreated units, and the leaf-specific ATE parameter is obtained as the difference between the intercepts of the two regressions. The many ways that the approach developed in this paper differs from Reguly (2021) will be revisited after the new approach has been spelled out in detail.

Prominent examples of Bayesian estimators for RDDs include Chib et al. (2023), who estimate the response curves with global splines where observations are weighted by their distance to the cutoff; Karabatsos and Walker (2015), who propose approximating the conditional expectations by an infinite mixture of normals; and Branson et al. (2019), who propose a Gaussian process prior for the expectations, in which observations are also weighted by their distance to the cutoff. All of these methods consist of global approximations of the outcome curves, while in some cases emphasizing units near the cutoff to obtain better predictions in that region. Although in principle these methods

could perform CATE estimation, none of these papers evaluate performance that way. Moreover, the steep computational demands of these approaches make thorough comparisons infeasible.

Our work is also a contribution to the burgeoning field of extensions to BART. BART models have been greatly extended in the years since 2010, to include heteroskedastic variants (Pratola et al., 2020; Murray, 2021), classification (Murray, 2021), conditional density estimation (Orlandi et al., 2021), causal inference (Hill, 2011; Hahn et al., 2020), variable selection (Linero, 2018), monotonicity constraints (Chipman et al., 2022; Papakostas et al., 2023), survival analysis (Sparapani et al., 2016), partial identification (Hahn et al., 2016) among others.

2 Background

To keep the paper relatively self-contained, we briefly review the basics of regression discontinuity designs and BART, and cast the RDD problem from a functional causal model perspective that is convenient for BART modeling.

2.1 Regression Discontinuity Designs

We conceptualize the treatment effect estimation problem via a quartet of random variables (Y, X, Z, U). The variable Y is the outcome variable; the variable X is the running variable; the variable Z is the treatment assignment indicator variable; and the variable U represents additional, possibly unobserved, causal factors. What specifically makes this correspond to an RDD is that we stipulate that $Z = \mathbb{I}(X > c)$, for cutoff c. For the remainder of this paper we assume $c = 0$ without loss of generality.

Figure 1 depicts a causal diagram representing the assumed causal relationships between these variables. Two key features of this diagram are one, that X blocks the impact of U on Z: in other words, X satisfies the back-door criterion for learning causal effects of Z on Y (more about this shortly; for details about the back-door criterion in causal graphs, see section 3.3 in Pearl (2009)). And two, X and U are not descendants of Z.

Using this causal diagram, we may express Y as some function of its graph parents, the random variables (X, Z, U):

$$Y = F(X, Z, U).$$

In principle, we may obtain draws of Y by first drawing (X, Z, U) according to their joint distribution and then applying the function F. Similarly, we may relate this formulation to the potential outcomes framework straightforwardly:

$$\begin{aligned} Y^1 &= F(X, 1, U), \\ Y^0 &= F(X, 0, U). \end{aligned} \tag{1}$$

Here, draws of (Y^1, Y^0) may be obtained (in principle) by drawing (X, Z, U) from their joint distribution and using only the (X, U) elements as arguments in the above two equations, "discarding" the drawn value of Z. Note that this construction implies the *consistency* condition: $Y = Y^1 Z + Y^0(1-Z)$. Likewise, this construction implies the *no interference* condition because each Y_i is considered to be

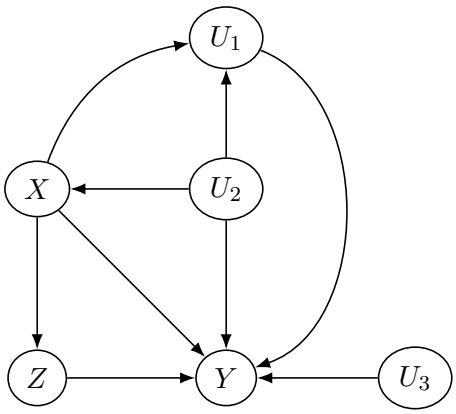

Figure 1: A causal directed acyclic graph representing the general structure of a regression discontinuity design problem, where $Z = \mathbb{I}(X > 0)$. Here $U = (U_1, U_2, U_3)$ is depicted in terms of three distinct components, exhaustively illustrating the various relationships that could obtain between X, Y, Z, and U while still preserving acyclicity and the necessary conditional independence relationships for causal identification.

produced with arguments (X_i, Z_i, U_i) and not those from other units j; in particular, in constructing Y_i, F does not take Z_j for $j \neq i$ as an argument.

Next, we define the following conditional expectations

$$\begin{aligned} \mu_1(x) &= \mathbb{E}[F(x, 1, U) \mid X = x], \\ \mu_0(x) &= \mathbb{E}[F(x, 0, U) \mid X = x], \end{aligned} \tag{2}$$

with which we can define the treatment effect function

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

Because X satisfies the back-door criterion, μ_1 and μ_0 are estimable from the data, meaning that

$$\begin{aligned} \mu_1(x) &= \mathbb{E}[F(x, 1, U) \mid X = x] = \mathbb{E}[Y \mid X = x, Z = 1], \\ \mu_0(x) &= \mathbb{E}[F(x, 0, U) \mid X = x] = \mathbb{E}[Y \mid X = x, Z = 0], \end{aligned} \tag{3}$$

the right-hand-sides of which can be estimated from sample data, which we supposed to be independent and identically distributed realizations of (Y_i, X_i, Z_i) for $i = 1, \ldots, n$. However, because $Z = \mathbb{I}(X > 0)$ we can in fact only learn $\mu_1(x)$ for $X > 0$ and $\mu_0(x)$ for $X < 0$. In potential outcomes terminology, conditioning on X satisfies ignorability,

$$(Y^1, Y^0) \perp\!\!\!\perp Z \mid X,$$

but not *strong ignorability*, because overlap is violated. Overlap would require that

$$0 < \mathbb{P}(Z = 1 \mid X = x) < 1 \quad \forall x,$$

which is clearly violated by the RDD assumption that $Z = \mathbb{I}(X > 0)$. Consequently, the overall ATE, $\bar{\tau} = \mathbb{E}(\tau(X))$, is unidentified, and we must content ourselves with estimating $\tau(0)$, the conditional average effect at the point $x = 0$, which we estimate as the difference between $\mu_1(0) - \mu_0(0)$. This is possible for continuous X so long as one is willing to assume that $\mu_1(x)$ and $\mu_0(x)$ are both suitably smooth functions of x: any inferred discontinuity at $x = 0$ must therefore be attributable to treatment effect. See Hahn et al. (2001) for the seminal exposition of continuity-based identification in RDD from the potential outcomes perspective.

2.1.1 Conditional average treatment effects in RDD

In this paper, we are concerned with learning not only $\tau(0)$, the "RDD ATE" (e.g. the CATE at $x = 0$), but also RDD CATEs, $\tau(0, \mathrm{w})$ for some covariate vector w. Incorporating additional covariates in the above framework turns out to be straightforward, simply by defining $W = \varphi(U)$ to be an observable function of the (possibly unobservable) causal factors U. We may then define our potential outcome means as

$$\begin{aligned} \mu_1(x, \mathrm{w}) &= \mathbb{E}[F(x, 1, U) \mid X = x, W = \mathrm{w}] = \mathbb{E}[Y \mid X = x, W = \mathrm{w}, Z = 1], \\ \mu_0(x, \mathrm{w}) &= \mathbb{E}[F(x, 0, U) \mid X = x, W = \mathrm{w}] = \mathbb{E}[Y \mid X = x, W = \mathrm{w}, Z = 0], \end{aligned} \tag{4}$$

and our treatment effect function as

$$\tau(x, \mathrm{w}) = \mu_1(x, \mathrm{w}) - \mu_0(x, \mathrm{w}).$$

We consider our data to be independent and identically distributed realizations (Y_i, X_i, Z_i, W_i) for $i = 1, \ldots, n$. Furthermore, we must assume that $\mu_1(x, \mathrm{w})$ and $\mu_0(x, \mathrm{w})$ are suitably smooth functions of x, *for every* w; in other words, for each value of w the usual continuity-based identification assumptions must hold.

With this framework and notation established, CATE estimation in RDDs boils down to estimation of condition expectation functions $\mathbb{E}[Y \mid X = x, W = \mathrm{w}, Z = z]$, for which we turn to BART models.

2.2 Bayesian Additive Regression Trees

The Bayesian Additive Regression Trees model (Chipman et al., 2010), or BART, represents an unknown mean function as a sum of regression trees, where each regression tree is assigned the prior described in Chipman et al. (1998). In this section we describe the model in terms of generic predictor vector X, to match earlier work, but in subsequent sections we will specialize our notation to include (X, W, Z) as in the RDD notation of the previous section.

Letting $f(x) = \mathbb{E}(Y \mid X = x)$ denote the unknown mean function of a covariate vector X, a

BART model with J trees is traditionally written

$$\begin{aligned} Y &= f(x) + \varepsilon, \\ &= \sum_{j=1}^{J} g(x; T_j, \mathbf{m}_j) + \varepsilon, \\ &= \sum_{j=1}^{J} g_j(x) + \varepsilon, \end{aligned} \tag{5}$$

where $\varepsilon \sim \mathrm{N}(0, \sigma^2)$ is a normally distributed additive error term. Here, $g(x; T_j, \mathbf{m}_j)$ denotes a piecewise constant function of x defined by a set of splitting rules T_j that partition the domain $\mathcal{X}$ into B_j disjoint regions, and a vector, $\mathbf{m}_j = (m_{j,1}, \ldots, m_{j,B_j})$, which records the values taken by $g(\cdot)$ on each of those regions. That is, let $b_j(x) : \mathcal{X} \to \{1, \ldots, |\mathbf{m}_j| = B_j\}$ be a function denoting which leaf node of the jth tree contains the point x; then

$$g_j(x) = g(x; T_j, \mathbf{m}_j) = m_{j,b_j(x)}.$$

Therefore, the parameters of a standard BART regression model are $(T_1, \mathbf{m}_1), \ldots, (T_J, \mathbf{m}_J)$ and σ. Chipman et al. (2010) consider priors such that: the tree components $(T_j, \mathbf{m}_j)$ are independent of each other and of σ^2, and the leaf node parameters $m_{j,b}$ are all mutually independent. Furthermore, Chipman et al. (2010) specify the same priors for all trees and leaf node parameters. The model thus consists of the specification of three priors: $p(T)$, $p(\sigma^2)$ and $p(\mathbf{m} \mid T)$.

The tree prior, $p(T)$, is defined by three components. First, the probability that a node with depth d will split is

$$\frac{\alpha}{(1+d)^{\beta}}, \quad \alpha \in (0,1), \beta \in [0, \infty), \tag{6}$$

implying that trees of greater depth have lower prior probability. The prior over cutpoints of the regression trees are uniform on the observed range of each feature and each feature is given equal prior weight. For the prior on the leaf node parameters, $p(\mathbf{m} \mid T)$, Chipman et al. (2010) specify independent Gaussian distributions over the elements of the $\mathbf{m}_j$ vectors: $m_{j,b} \overset{\text{iid}}{\sim} \mathrm{N}(m_0, \sigma_0^2)$. Finally, σ^2 is given an inverse Gamma prior. For further details and justification concerning BART prior specification, see Chipman et al. (2010).

Chipman et al. (2010) construct a Gibbs sampling algorithm to obtain posterior draws of the trees, their leaf parameters, and the residual variance σ^2. Let T_{-j} denote the set of all trees *except* T_j, and similarly for $\mathbf{m}_{-j}$. At each iteration, the algorithm produces J consecutive samples of $(T_j, \mathbf{m}_j, \sigma)$ using the following compositions:

$$T_j \mid T_{-j}, \mathbf{m}_{-j}, \sigma, y, \tag{7}$$

then

$$\mathbf{m}_j \mid T_j, T_{-j}, \mathbf{m}_{-j}, \sigma, y, \tag{8}$$

and finally

$$\sigma \mid T_1, \mathbf{m}_1, \ldots T_J, \mathbf{m}_J, y. \tag{9}$$

Sampling from (7) is simplified by noting that each tree depends on $(T_{-j}, \mathbf{m}_{-j}, y)$ only through a "partial" residual:

$$\mathrm{r}_j = y - \sum_{j' \neq j} g(x; T_{j'}, \mathbf{m}_{j'}). \tag{10}$$

Therefore, the log marginal likelihood of the data in leaf b_j (integrating out the unknown leaf mean parameter m_{j,b_j}) is:

$$l_{b_j} = -\frac{n_{b_j}}{2}\log(2\pi) - n_{b_j}\log(\sigma) + \frac{1}{2}\log\left(\frac{\sigma^2}{n_{b_j}\sigma_\mu^2 + \sigma^2}\right) - \frac{\sum_{i:x_i \in b_j} r_i^2}{2\sigma^2} + \frac{\sigma_\mu^2(\sum_{i:x_i \in b_j} r_i)^2}{2\sigma^2(n_{b_j}\sigma_\mu^2 + \sigma^2)}, \tag{11}$$

which is used to compute a Metropolis-Hastings ratio for accepting or rejecting a proposed tree. Details may be found in Chipman et al. (2010) . Conditional on the tree, sampling the elements of $\mathbf{m}_j$ (step 8) is a standard conjugate update that can be found in any textbook (see, for example, section 2.3 in Gamerman and Lopes (2006))), where the observed "data" is the r_j vector from just above.

3 Bayesian Additive RDD Trees

3.1 BART for Causal Inference

BART may be used to fit the conditional expectations needed for treatment effect estimation in RDDs in a number of different ways. The first way, what is sometimes called an "S-Learner", is simply to include the treatment assignment indicator $Z = \mathbb{I}(X > 0)$ among the feature set along with the other covariates, X and W. This approach was proposed in Hill (2011) in the context of regression adjustment for treatment effect estimation under conditional strong ignorability. The second way, what is sometimes called a "T-Learner" is to fit two individual BART models to the treated $z = 1$ and control $z = 0$ data separately. See Künzel et al. (2019) for a discussion of the S- and T-Learner nomenclature.

Hahn et al. (2020) provide an extensive discussion of potential drawbacks to the S-Learner and T-Learner approaches in the context of BART models, which we now summarize. We will refer to S-BART and T-BART in relation to these strategies. The main problem with S-BART is that there can be many trees, with potentially very different splits, that achieve a similar likelihood evaluation. While this is unobjectionable if the goal is merely to predict $\mathbb{E}[Y \mid X = x, W = \mathrm{w}, Z = z]$ — indeed, the over-parametrization of these models probably accounts for much of their empirical success — different splitting patterns in Z versus X tend to imply different treatment effect estimates. The upshot — borne out by extensive simulation studies — is that S-BART models for causal inference tend to have unpredictable biases as a result of the specific dependency structure among the predictor variables in a given data set. While the T-BART approach successfully addresses this drawback of S-BART, fitting completely separate models to the treated and control data introduces a different problem: regularization of the conditional treatment effect function itself is implicit, and generally too weak when treatment effects are expected to be small relative to variation in the outcome due to other observed features or unobserved factors. In short, T-BART tends to over-fit the data, yielding

CATE estimators with high variance.

In the next section, we introduce a modified BART model — different from both S-BART and T-BART — that it is markedly better at CATE estimation than either one, and also better than local polynomial regression and the CART approach of Reguly (2021) as well.

3.2 The BARDDT Model

For RDD, we propose that a linear model in the leaf is a viable strategy for overcoming the problems with T-BART and S-BART described above. We build on the work of Chipman et al. (2002), Gramacy and Lee (2008), and Starling et al. (2020), by proposing a BART model where the trees are allowed to split on (x, w) but where each leaf node parameter is a vector of regression coefficients tailored to the RDD context (rather than a scalar constant as in default BART). In one sense, such a model can be seen as implying distinct RDD ATE regressions for each subgroup determined by a given tree; however, this intuition is only heuristic, as the entire model is fit jointly as an ensemble of such trees. Instead, we motivate this model as a way to estimate the necessary conditional expectations via a parametrization where the conditional treatment effect function can be explicitly regularized, as follows.

Let ψ denote the following basis vector:

$$\psi(x, z) = \begin{bmatrix} 1 & zx & (1-z)x & z \end{bmatrix}. \tag{12}$$

To generalize the original BART model, we define $g_j(x, \mathrm{w}, z)$ as a piecewise linear function as follows. Let $b_j(x, \mathrm{w})$ denote the node in the jth tree which contains the point (x, w); then the prediction function for tree j is defined to be:

$$g_j(x, \mathrm{w}, z) = \psi(x, z)\Gamma_{b_j(x,\mathrm{w})} \tag{13}$$

for a leaf-specific regression vector $\Gamma_{b_j} = (\eta_{b_j}, \lambda_{b_j}, \theta_{b_j}, \Delta_{b_j})^t$. Therefore, letting n_{b_j} denote the number of data points allocated to node b in the jth tree and Ψ_{b_j} denote the $n_{b_j} \times 4$ matrix, with rows equal to $\psi(x, z)$ for all $(x_i, z_i) \in b_j$, the model for observations assigned to leaf b_j, can be expressed in matrix notation as:

$$\begin{aligned} \mathbf{Y}_{b_j} \mid \Gamma_{b_j}, \sigma^2 &\sim \mathrm{N}(\Psi_{b_j}\Gamma_{b_j}, \sigma^2) \\ \Gamma_{b_j} &\sim \mathrm{N}(0, \Sigma_0), \end{aligned} \tag{14}$$

where we set $\Sigma_0 = \frac{0.033}{J}\mathrm{I}$ as a default (for x vectors standardized to have unit variance in-sample).

This choice of basis entails that the RDD CATE at w, $\tau(0, \mathrm{w})$, is a sum of the $\Delta_{b_j(0,\mathrm{w})}$ elements across all trees $j = 1, \ldots, J$:

$$
\begin{aligned}
\tau(0,\mathrm{w}) &= \mathbb{E}[Y^1 \mid X=0, W=\mathrm{w}] - \mathbb{E}[Y^0 \mid X=0, W=\mathrm{w}] \\
&= \mathbb{E}[Y \mid X=0, W=\mathrm{w}, Z=1] - \mathbb{E}[Y \mid X=0, W=\mathrm{w}, Z=0] \\
&= \sum_{j=1}^{J} g_j(0,\mathrm{w},1) - \sum_{j=1}^{J} g_j(0,\mathrm{w},0) \\
&= \sum_{j=1}^{J} \psi(0,1)\Gamma_{b_j(0,\mathrm{w})} - \sum_{j=1}^{J} \psi(0,0)\Gamma_{b_j(0,\mathrm{w})} \\
&= \sum_{j=1}^{J} \Big(\psi(0,1) - \psi(0,0)\Big)\Gamma_{b_j(0,\mathrm{w})} \\
&= \sum_{j=1}^{J} \Big((1,0,0,1) - (1,0,0,0)\Big)\Gamma_{b_j(0,\mathrm{w})} \\
&= \sum_{j=1}^{J} \Delta_{b_j(0,\mathrm{w})}.
\end{aligned}
\tag{15}
$$

As a result, the priors on the Δ coefficients directly regularize the treatment effect. We set the tree and error variance priors as in the original BART model.

Posterior sampling from this model proceeds nearly identically to the traditional BART Gibbs sampler, but with a modified log marginal likelihood, which for a node b_j is:

$$
\begin{aligned}
l_{b_j} = &-\frac{n_{b_j}}{2}\log(2\pi) - n_{b_j}\log(\sigma) - \frac{1}{2}\log\left(\det\left(\mathrm{I} + \frac{\Sigma_0 \Psi_{b_j}^t \Psi_{b_j}}{\sigma^2}\right)\right) \\
&- \frac{r_{b_j}^t r_{b_j}}{2\sigma^2} + \frac{1}{2}\frac{r_{b_j}^t \Psi_{b_j}}{\sigma^2}\left(\Sigma_0^{-1} + \frac{\Psi_{b_j}^t \Psi_{b_j}}{\sigma^2}\right)^{-1}\frac{\Psi_{b_j}^t r_{b_j}}{\sigma^2},
\end{aligned}
\tag{16}
$$

where r_{b_j} is a n_{b_j} vector containing the partial residuals, as defined in (10), for the points in b_j. Note that this expression generalizes (11); the two expressions become equivalent if the basis vector Ψ_{b_j} in the above expression is a single column of ones.

Likewise, the parameter sampling follows a standard conditionally (on σ^2) conjugate linear regression update, independently for each leaf of the current tree which we omit here as it can be found in standard references (for example, section 2.3.3 in Gamerman and Lopes (2006)).

Figures 2 through 4 provide a graphical depiction of how the BARDDT model fits a response surface and thereby estimates CATEs for distinct values of w. For simplicity only two trees are used in the illustration, while in practice dozens or hundreds of trees may be used (in our simulations and empirical example, we use 50 trees).

An interesting property of BARDDT can be seen in this small illustration — by letting the regression trees split on the running variable, there is no need to separately define a "bandwidth" as is used in the polynomial approach to RDD. Instead, the regression trees automatically determine (in the course of posterior sampling) when to "prune" away regions away from the cutoff value. There are two notable features of this approach. One, different trees in the ensemble are effectively using

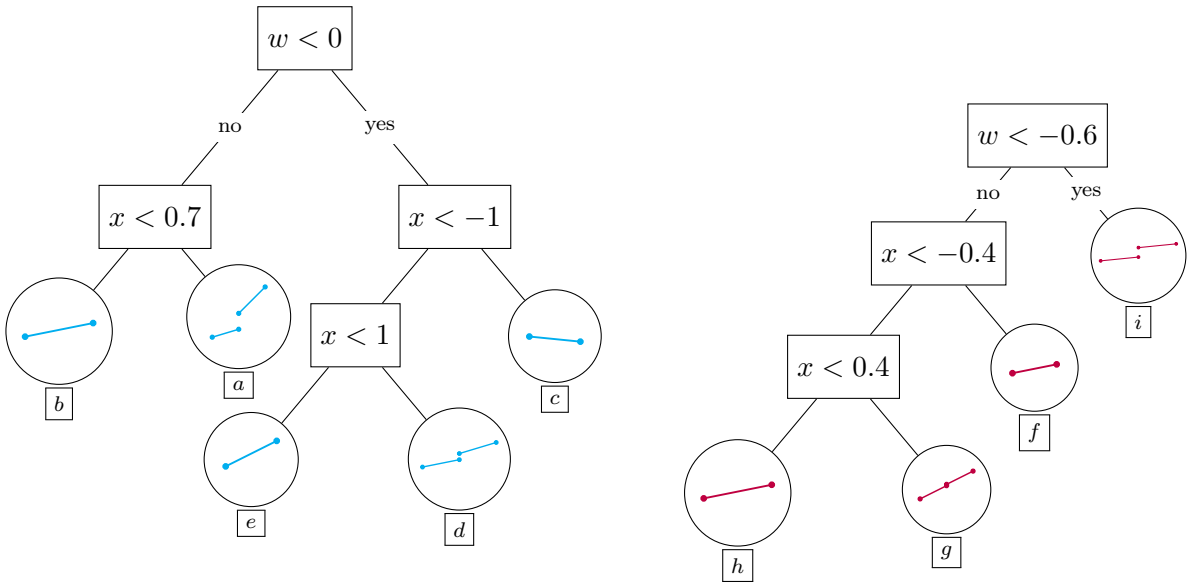

Figure 2: Two regression trees with splits in x and a single scalar w. Node images depict the $g(x, \mathrm{w}, z)$ function (in x) defined by that node's Γ coefficients. The vertical gap between the two line segments in a node that contain $x = 0$ is that node's contribution to the CATE at $X = 0$. Note that only such nodes contribute for CATE prediction at $x = 0$

different local bandwidths and these fits are then blended together. For example, in the bottom panel of figure 3, we obtain one bandwidth for the region $d+i$, and a different one for regions $a+g$ and $d+g$. Two, for cells in the tree partition that do not span the cutoff, the regression within that partition contains no causal contrasts — all observations either have $Z = 1$ or $Z = 0$. For those cells, the treatment effect coefficient is ill-posed and in those cases the posterior sampling is effectively a draw from the prior; however, such draws correspond to points where the treatment effect is unidentified and none of these draws contribute to the estimation of $\tau(0, \mathrm{w})$ — for example, only nodes $a + g$, $d + g$, and $d + i$ in figure 3 provide any contribution. This implies that draws of Δ corresponding to nodes not predicting at $X = 0$ will always be draws from the prior, which has some intuitive appeal.

BARDDT differs from Reguly (2021) — which is, to the best of our knowledge, the only other tree-based CATE estimator for RDD — in three important ways:

- BARDDT is a sum of many regression trees, rather than a single tree,
- the BARDDT estimator is based on Bayesian posterior mean[1] rather than a single optimization-based model fit,
- BARDDT trees are permitted to split in the running variable (as mentioned in the previous paragraph).

[1]As a partition model, BART-based estimates of conditional expectations have points of discontinuity. Although RDD demands that μ_1 and μ_0 are smooth functions, this identification condition is on the DGP, not on the estimator. BART is a consistent estimator of that underlying smooth function (He and Hahn, 2023; Saha, 2023) even if its estimates are not.

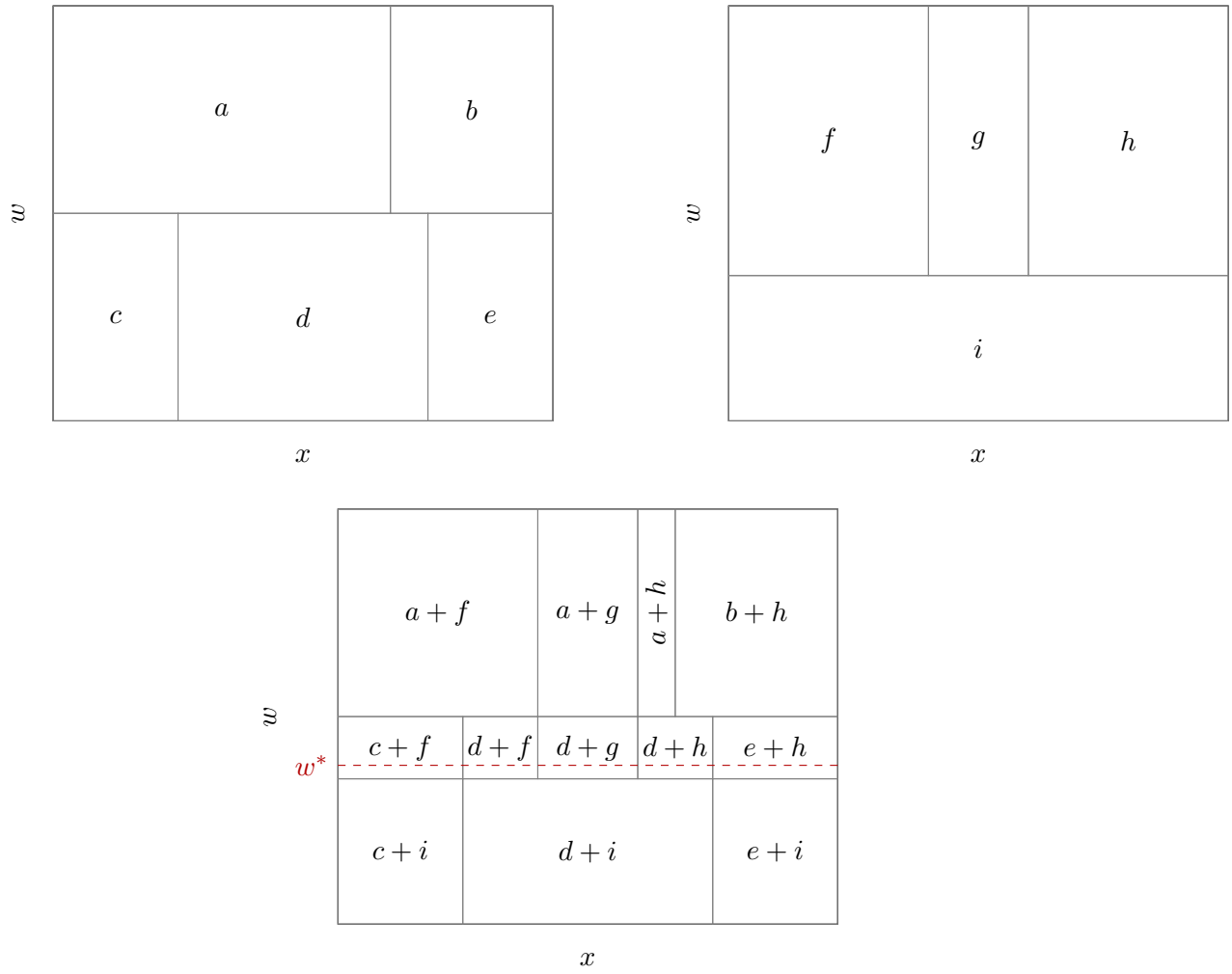

Figure 3: The two top figures show the same two regression trees as in the preceding figure, now represented as a partition of the x-w plane. Labels in each partition correspond to the leaf nodes depicted in the previous picture. The bottom figure shows the partition of the x-w plane implied by the sum of the two trees; the red dashed line marks point $W = w^*$ and the combination of nodes that include this point

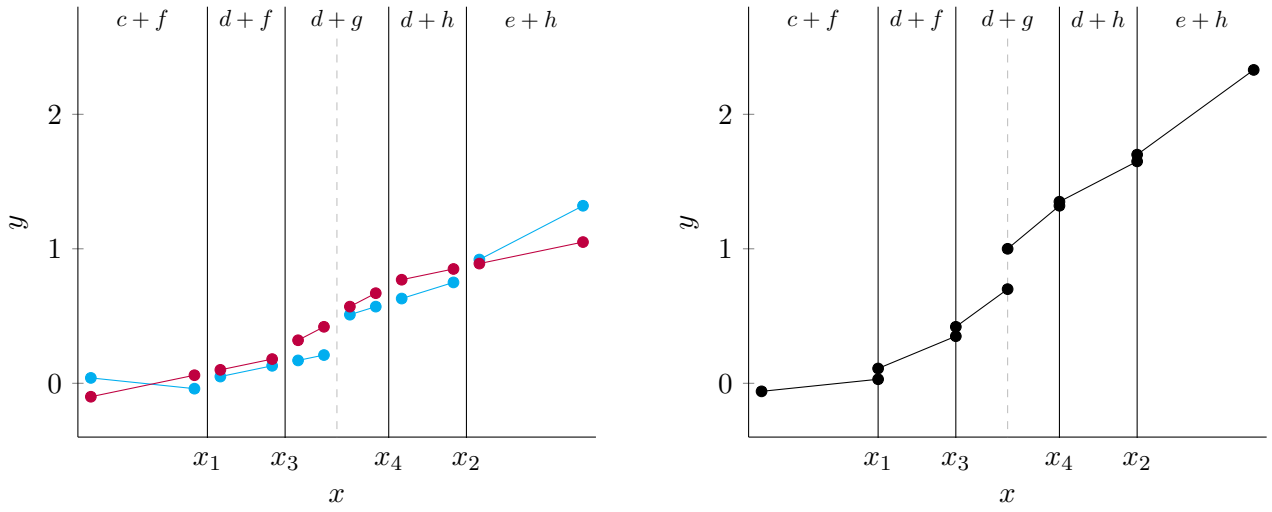

Figure 4: Left: The function fit at $W = w^*$ for the two trees shown in the previous two figures, shown superimposed. Right: The aggregated fit achieved by summing the contributes of two regression tree fits shown at left. The magnitude of the discontinuity at $x = 0$ (located at the dashed gray vertical line) represents the treatment effect at that point. Different values of w will produce distinct fits; for the two trees shown, there can be three distinct fits based on the value of w.

4 Simulation studies for CATE estimation in RDDs

This section describes a parametrized protocol for simulating data for evaluating CATE estimation methods in RDDs. Modifiable code implementing this approach is available at the Github repository associated with the paper.

There are three ingredients to any simulation-based statistical method evaluation procedure: the estimand, the evaluation criteria, and the data generating process.

4.1 The estimand

Generically, our estimand is the CATE function at $x = 0$; i.e. $\tau(0, \mathrm{w})$. The key practical question is which values of w to consider. Some values of w will not be well-represented near $x = 0$ and so no estimation technique will be able to estimate those points effectively. As such, to focus on feasible points — which will lead to interesting comparisons between methods — we recommend restricting the evaluation points to the observed w_i such that $|x_i| \leq \delta$, for some $\delta > 0$. In our example, we use $\delta = 0.1$ for a standardized x variable. Therefore, our estimand of interest is a vector of treatment effects:

$$\tau(0, \mathrm{w}_i) \quad \forall i \;\text{ such that }\; |x_i| \leq \delta. \tag{17}$$

4.2 Estimation loss function

For our evaluation criteria we will consider average root-mean-squared estimation error, expressed as a fraction of a default ATE estimator:

$$\text{CATE RMSE} = \frac{\sqrt{\sum_{i:|x_i|\leq\delta} \left(\hat{\tau}(0, \mathrm{w}_i) - \tau(0, \mathrm{w}_i)\right)^2}}{\sqrt{\sum_{i:|x_i|\leq\delta} \left(\hat{\tau}(0) - \tau(0, \mathrm{w}_i)\right)^2}}. \tag{18}$$

This performance metric judges the ability of $\hat{\tau}(0, \mathrm{w})$ to estimate CATEs relative to a baseline ATE estimator (at $x = 0$), thereby allowing us to tell if methods are doing better than would be possible just by assuming homogeneous effects. It also permits a unitless performance measure, so that relative accuracy across methods can be compared in a standardized way across data generating processes of varying outcome scales, which can affect the implicit difficulty of the estimation problem.

4.3 Data generating process

The goal of our simulation study is to understand how various methods perform at estimating CATEs across a variety of DGPs. More particularly, we would like to be able to characterize what aspects of a DGP make a causal inference problem hard or easy so that we may identify methods which adapt to variation in the "intrinsic" problem difficulty. To approach this problem we will take an analysis of variance (ANOVA) perspective (Hahn et al., 2018, 2019), tailored to the RDD context.

Because an RDD only identifies the treatment effect at $x = 0$, the relevant signal to noise ratios vis-a-vis treatment effect estimation are conditional on $x = 0$; accordingly, we will design our DGP so that it is explicitly parametrized in terms of conditional variances at $x = 0$. Data will be simulated consistent with the causal diagram in Figure 1: W and X will be generated , followed by Y given W and X.

4.3.1 Generating (W, X)

Our simulation studies will consider W to be fixed in advance and we will consider replications over (X, Y). The covariates W can be empirical data from a real-world application or can be simulated. Here, for illustration purposes, we generate W according to a mean-zero multivariate Gaussian distribution with a Toeplitz covariance matrix, with entries ranging from 0 to 2. For example, for $p = 5$ the covariance would be:

$$\mathbb{C}\text{ov}(W) = \begin{pmatrix} 2 & \frac{3}{2} & 1 & \frac{1}{2} & 0 \\ \frac{3}{2} & 2 & \frac{3}{2} & 1 & \frac{1}{2} \\ 1 & \frac{3}{2} & 2 & \frac{3}{2} & 1 \\ \frac{1}{2} & 1 & \frac{3}{2} & 2 & \frac{3}{2} \\ 0 & \frac{1}{2} & 1 & \frac{3}{2} & 2 \end{pmatrix}.$$

We then draw X according to a Gaussian distribution centered at a linear combination of the $W = \mathrm{w}$ values:

$$X \mid W = \mathrm{w} \sim \mathrm{N}(\gamma_0 + \mathrm{w}^t\gamma, \nu)$$

where γ_0 is the marginal mean and γ is a p-dimensional vector of regression coefficients. For our demonstration here we use $\gamma_0 = 1$; this choice was made so that X is not centered at the cutoff, which we thought would be unrealistic. For γ we use an evenly weighted coefficient vector such that $\mathbb{C}\text{ov}(X) = 1$ and $\mathbb{C}\text{or}(X, W^t\gamma) = \rho$, for some pre-specified value of ρ; these constraints also determine the value of ν. Full details can be found in the Github repository associated with this paper.

Setting γ to the zero vector implies that X and W are independent, which is an interesting special case. But being able to test the performance of CATE estimators under varying degrees of association between the running variable X and moderators W is important and this linear model is a simple test case for that. To summarize, γ_0 and ρ are important parameters in our DGP, governing how concentrated around the cutoff the data are and the strength of the association between the running variable and the moderators.

4.3.2 Generating Y, given W and X

To begin, we use the following "treatment effect parametrization" when specifying our DGP.

$$\mathbb{E}(Y \mid X = x, W = \mathrm{w}, Z = z) = \mu(x, \mathrm{w}) + \tau(x, \mathrm{w})z, \tag{19}$$

which relates to the notation in Section 2.1 by taking $\mu(x, \mathrm{w}) \equiv \mu_0(x, \mathrm{w})$ and $\tau(x, \mathrm{w}) \equiv \mu_1(x, \mathrm{w}) - \mu_0(x, \mathrm{w})$. This parametrization allows us to generate our data directly in terms of the treatment effect function; we may specify the average magnitude and complexity of $\tau(x, \mathrm{w})$ explicitly. In this paper we will consider only homoskedastic errors in our DGP:

$$Y_i = \mu(x_i, \mathrm{w}_i) + \tau(x_i, \mathrm{w}_i)z_i + \sigma\epsilon_i, \tag{20}$$

for a mean-zero Gaussian error term; extensions to heteroskedastic and/or non-normal errors are straightforward.

Before describing our specific choices for $\mu(x, \mathrm{w})$ and $\tau(x, \mathrm{w})$, we will discuss a strategy for fixing some properties of these functions averaged over W, given $X = 0$. Specifically, we consider the following properties/quantities: $\min_{\mathrm{w}} \tau(0, \mathrm{w})$, $\mathbb{V}(\tau(0, W) \mid X = 0)$, and $\mathbb{V}(\mu(0, W) \mid X = 0)$. We use Monte Carlo simulation to compute these quantities for "template" functions $\mu^\star$ and $\tau^\star$ and then devise linear transformations of those template functions to achieve desired relationships between them. That is, we take a large sample from $W \mid X = 0$ and compute the above quantities based on that simulated data. For (W, X) draw as described above, $W \mid X = 0$ is a multivariate Gaussian with

$$\begin{aligned} \mathbb{E}(W \mid X = 0) &= -\gamma_0\gamma, \\ \mathbb{V}(W \mid X = 0) &= \Sigma_W - \Sigma_W\beta\beta^t\Sigma_W^t. \end{aligned} \tag{21}$$

Using this strategy, we fix $\mathbb{V}(\mu(0, W) \mid X = 0) = 1$ and specify our DGP in terms of the following

parameters

$$\begin{aligned}\sqrt{\mathbb{V}(\tau(0,W) \mid X=0)} &= k_2,\\ \sqrt{\mathbb{V}(Y \mid X=0,W)} = \sigma &= k_4\\ \min_{\mathrm{w}} \tau(0,\mathrm{w}) = \underline{\tau}_0 &= k_5.\end{aligned} \tag{22}$$

These values, along with γ_0 and ρ mentioned in the previous section, are the key parameters in our DGP.

Last, but not least, we must specify $\mu^\star$ and $\tau^\star$. Letting $\mathrm{w} = (w_1 \dots w_p)$ be realizations of a length p random vector W, define $w^\star = \frac{\sum_{j=1}^p w_j}{\sqrt{p}}$. Our template functions are:

$$\begin{aligned}\mu^\star(x,\mathrm{w}) &= k_1(x+1)^3 + (w^\star+2)^2\left(\mathrm{sign}(x+1)\sqrt{|(x+1)|}\right)^{k_3},\\ \tau^\star(\mathrm{w}) &= \Phi(2w_1+3)/2 + \phi(w_1),\end{aligned} \tag{23}$$

where $\Phi(\cdot)$ and $\phi(\cdot)$ are the cumulative distribution and probability density functions, respectively, of a standard normal random variable. The variables k_1 and k_3 are further parameters for variation; the parameter k_1 controls how much variability of μ will be due to X versus to W and k_3 determines whether or not μ is additive in X and W or if there is an interaction ($k_3 = 1$) or not ($k_3 = 0$).

These choices of $\mu^\star$ and $\tau^\star$ provide nontrivial nonlinearities while being relatively easy to grok. Furthermore, they are designed such that the dimension of W can be modified without changing the function definition, but still using all available dimensions in the definition of μ. The treatment effect function is restricted to depend on a single element of W to facilitate plotting.

Many variations of these functions were explored in the preparation of this paper; these specific choices nicely illustrate the simulation procedure. Considering a variety of template functions is of course recommended and should be chosen depending on domain knowledge to investigate their impact on the performance of candidate CATE estimation procedures in a use-case-relevant way.

4.4 Estimation methods

To demonstrate our simulation protocol we will compare the following methods:

- BARDDT
- S-BART
- T-BART
- a local polynomial estimator
- RD-Tree (Reguly, 2021).

All three BART variants were fit with 50 trees each (two forests of 50 trees for T-BART), with tree depth parameters set as in Chipman et al. (2010): $\alpha = 0.95, \beta = 2$ and fit using the `stochtree` package. Further, the CATE estimator in all cases was the vector of posterior means of $\tau(0, \mathrm{w}_i)$ for i such that $|x_i| \leq 0.1$.

The local polynomial estimator is trained on data points within the bandwidth obtained with the `rdrobust` package (Calonico et al., 2015). Ordinary-least-squares is used to fit a fourth degree

polynomial in each feature of W, interacted with X and Z, and an additive third degree polynomial in X, or, in Wilkinson notation (Wilkinson and Rogers, 1973):

$$Y \sim \left(\sum_{j=1}^{p} \text{poly}(W_j, 4) \right) \cdot X \cdot Z + \text{poly}(X, 3), \tag{24}$$

where $\text{poly}(A, p) = \sum_{j=1}^{p} A^j$. The choice of this particular polynomial is of course open to discussion, but any specific choice will likewise be subject to critique; a main benefit of tree-based regressions is that they side-step this decision.

Model parameters for RD-Tree were set as suggested in Reguly (2021). For further details on the method and its parameters, please refer to the original text and the example script at https://github.com/regulyagoston/RD_tree/.

4.5 Comparisons

The results in this section are based on configurations of the DGP described in the previous section which can be roughly separated into two groups: "easy" and "hard". For the easy setting, $(k_1 = 1, k_2 = 1, k_3 = 0, k_4 = 0.1)$: prognostic and treatment variation are comparable magnitudes, μ is separable in x and w, and low noise. For the "hard" setting $(k_1 = 5, k_2 = 0.25, k_3 = 1, k_4 = 0.5)$: prognostic variation is twenty times larger than treatment variation, μ is non-separable in x and w, and noise is high. Results are based on 100 replications of size $n = 4000$ for each DGP configuration.

4.5.1 Overall results

Figure 5 shows boxplots of the CATE RMSE per replication for three configurations each of the "easy" and "hard" setups. S-BART results are excluded from these plots because it had significantly worse performance in all scenarios, distorting the plotting scale and making comparing the other methods more difficult visually. Recall that a value greater than or equal to 1 means the estimator is doing no better at CATE prediction than estimating the ATE and assuming treatment effect homogeneity.

Table 1 presents the average RMSE obtained by all estimators. BARDDT is always better than the other methods. T-BART performs well in the easier settings, but suffers a near tenfold increase in its average RMSE in the settings with high noise and non-separable μ. The local polynomial estimator usually performs worse than BARDDT, even in the settings in which it is competitive on average, and it frequently obtains much larger errors in individual replications. RD-Tree is competitive with the polynomial for second-best in the harder DGPs, but is the second worst performer in the easier cases.

4.5.2 Individual fits

In addition to the aggregate results, such as those reported in Table 1 and depicted in Figure 5, it is often instructive to consider individual fits compared to the ground-truth, which is available to us in simulation studies. The qualitative behavior — the specific ways that the models misfit certain types of data — are often persistent across replications, but can be seen in individual fits. Figure 6

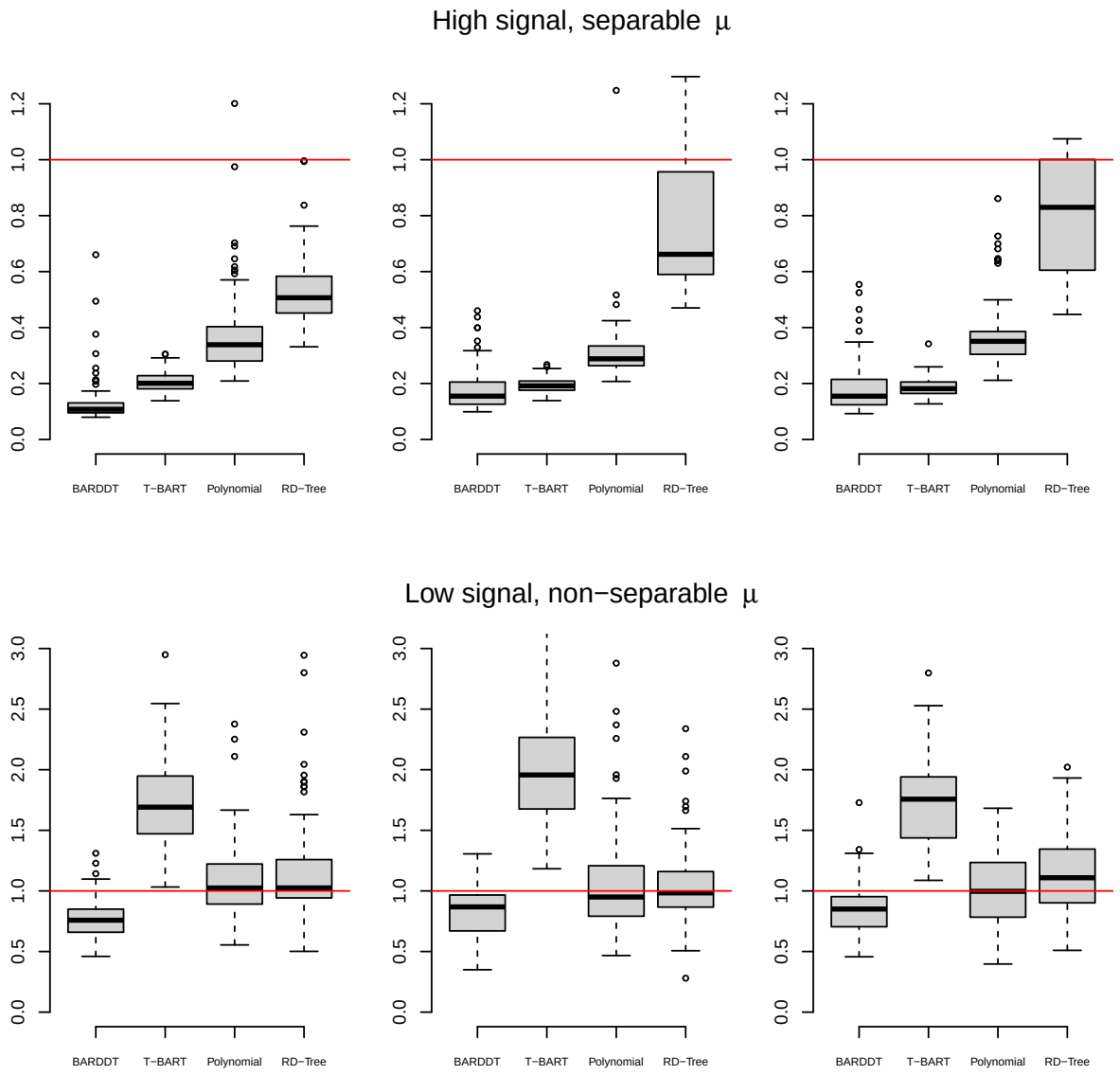

Figure 5: RMSE results per estimator. The upper row corresponds to the 'easy' setup described in the text; the lower row corresponds to the 'hard' setup. For the easy setting, the values for the additional parameters are $(k_5, p, \rho) = \{(0, 2, 0.5), (0, 4, 0), (1, 2, 0)\}$; for the hard setting, the values are $(k_5, p, \rho) = \{(0, 4, 0.5), (1, 2, 0.5), (1, 4, 0)\}$. The CATE RMSE values are divided by the RMSE obtained by predicting the CATE with an estimate ATE. Thus, the red line at 1 indicates the point above which the methods are worse than this naive estimator. BARDDT produces lower RMSE than the other estimators on average and is most often better than the homogenous estimator.

k_1	k_2	k_3	k_4	k_5	p	ρ	BARDDT	T-BART	S-BART	Polynomial	RD-Tree
1.00	1.00	0.00	0.10	0.00	2.00	0.50	0.13	0.21	1.32	0.37	0.53
1.00	1.00	0.00	0.10	0.00	4.00	0.00	0.18	0.19	1.47	0.31	0.74
1.00	1.00	0.00	0.10	1.00	2.00	0.00	0.18	0.19	1.67	0.37	0.80
5.00	0.25	1.00	0.50	0.00	4.00	0.50	0.77	1.70	1.49	1.08	1.16
5.00	0.25	1.00	0.50	1.00	2.00	0.50	0.84	2.00	2.83	1.06	1.03
5.00	0.25	1.00	0.50	1.00	4.00	0.00	0.84	1.75	2.47	1.02	1.14

Table 1: Average RMSE per DGP, also divided here by the RMSE of the naive ATE estimator

presents the CATE fits for on an individual data set, for one easy and one hard DGP. The qualitative fits of BARDDT are in line with expectations, while the other methods exhibit undesirable behavior in certain cases.

- Although T-BART performs well under the easier regime, even there it still exhibits high variance CATE estimates. T-BART's high variance becomes more pronounced under the harder DGP, resulting in substantially higher RMSE relative to both BARDDT and the polynomial model.

- The extreme bias shift exhibited by S-BART in the low noise setting is reminiscent of the regularization-induced confounding (RIC) problem, described by Hahn et al. (2020). Broadly, the lesson here is that S-BART has unpredictable biases in causal inference problems. It does comparatively well in the high noise case, but only because it rarely splits in that case, collapsing to a homogenous treatment model, which outperforms the overfitting T-BART and polynomial models in this regime.

- The fits for the easier setup show that, even with high signal, the polynomial model struggles with extrapolation at the boundaries of the support of w_1. At the same time, the polynomial model also presents a sizable increase in variance under high noise, as seen on the fits for the harder regime.

- RD-Tree appears to "under-split" on W, leading to a too-coarse fit of the CATE function, especially in the low-noise regime. This behavior is to be expected with a single CART fit, a problem that additive tree models, like BART, were explicitly designed to address.

5 The Effect of Academic Probation on Educational Outcomes

We turn now to an empirical illustration based on Lindo et al. (2010), who analyze data on college students enrolled in a large Canadian university in order to evaluate the effectiveness of an academic probation policy. Students who present a grade point average (GPA) lower than a certain threshold at the end of each term are placed on academic probation and must improve their GPA in the subsequent term or else face suspension. We are interested in how being put on probation or not, Z,

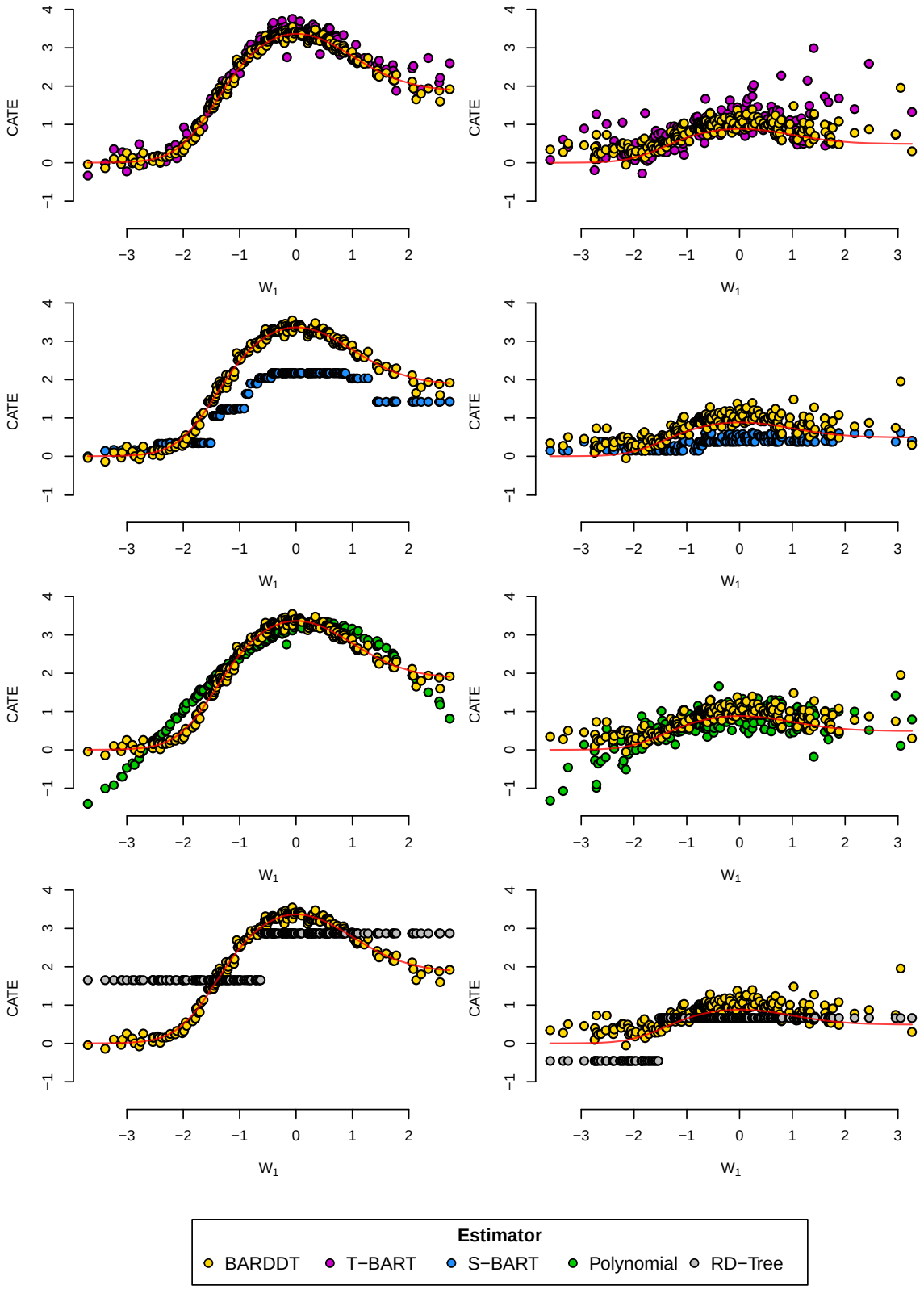

Figure 6: Each panel presents the CATE fits for one illustrative sample for one easy and one hard DGP setting, plotted against the true CATE, shown as a red curve. Left panel corresponds to the parameter configuration in the first row of table 1, right panel corresponds to the fourth row. The BARDDT fit is shown in gold in all of the plots for ease of comparison. From top to bottom we have T-BART (purple), S-BART (blue), polynomial (green), and RD-Tree (gray).

affects students' GPA, Y, at the end of the current term. The running variable, X, is the negative distance between a student's previous-term GPA and the probation threshold, so that students placed on probation ($Z = 1$) have a positive score and the cutoff is 0. Potential moderators, W, are:

- gender ('male'),
- age upon entering university ('age_at_entry')
- a dummy for being born in North America ('bpl_north_america'),
- the number of credits taken in the first year ('totcredits_year1')
- an indicator designating each of three campuses ('loc_campus' 1, 2 and 3), and
- high school GPA as a quantile w.r.t the university's incoming class ('hsgrade_pct').

Figure 7 presents a summary of the CATE posterior produced by BARDDT for this application. This picture is produced fitting a regression, using W as the predictors, to the individual posterior mean CATEs:

$$\bar{\tau}_i = \frac{1}{M} \sum_{h=1}^{M} \tau^{(h)}(0, \mathrm{w}_i), \tag{25}$$

where h indexes each of M total posterior samples. As in our simulation studies, we restrict our posterior analysis to use w_i values of observations with $|x_i| \leq \delta = 0.1$ (after normalizing X to have standard deviation 1 in-sample). For the Lindo et al. (2010) data, this means that BARDDT was trained on $n = 40,582$ observations, of which 1,602 satisfy $x_i \leq 0.1$, which were used to generate the effect moderation tree from Figure 7.

The resulting effect moderation tree indicates that course load (credits attempted) in the academic term leading to their probation is a strong moderator. Contextually, this result is plausible, both because course load could relate to latent character attributes that influence a student's responsiveness to sanctions and also because it could predict course load in the current term, which would in turn have implications for the GPA (i.e. it is harder to get a high GPA while taking more credit hours). The tree also suggests that effects differ by campus, and age and gender of the student. These findings are all prima facie plausible as well.

To gauge how strong these findings are statistically, we can zoom in on isolated subgroups and compare the posteriors of their subgroup average treatment effects. This approach is valid because in fitting the effect moderation tree to the posterior mean CATEs we in no way altered the posterior itself; the effect moderation tree is a posterior summary tool and not any additional inferential approach; the posterior is obtained once and can be explored freely using a variety of techniques without vitiating its statistical validity. Investigating the most extreme differences is a good place to start: consider the two groups of students at opposite ends of the treatment effect range discovered by the effect moderation tree:

Group A a male student that entered college older than 19 and attempted at least 5 credits in the first year (leftmost leaf node, colored red, comprising 128 individuals)

Group B a student of any gender who entered college younger than 19 and attempted more than 4, but less than 5 credits in the first year (rightmost leaf node, colored gold, comprising 108

individuals).

Subgroup CATEs are obtained by aggregating CATEs across the observed w_i values for individuals in each group; this can be done for individual posterior samples, yielding a posterior distribution over the subgroup CATE:

$$\bar{\tau}_A^{(h)} = \frac{1}{n_A} \sum_{i:\mathrm{w}_i} \tau^{(h)}(0, \mathrm{w}_i), \tag{26}$$

where h indexes a posterior draw and n_A denotes the number of individuals in the group A. Figure 8 presents a contour plot for a bivariate kernel density estimate of the joint CATE posterior distribution for subgroups A and B. The contour lines are almost all above the 45° line, indicating that the preponderance of posterior probability falls in the region where the treatment effect for Group B is greater than that of Group A, meaning that the difference in the subgroup treatment effects flagged by the effect moderation tree persist even after accounting for estimation uncertainty in the underlying CATE function.

As always, CATEs that vary with observable factors do not necessarily represent a *causal* moderating relationship. Here, if the treatment effect of academic probation is seen to vary with the number of credits, that does not imply that this association is causal: prescribing students to take a certain number of credits will not necessarily lead to a more effective probation policy, it may simply be that the type of student to naturally enroll for fewer credit hours is more likely to be responsive to academic probation. An entirely distinct set of causal assumptions are required to interpret the CATE variations themselves as causal. All the same, uncovering these patterns of treatment effect variability are crucial to suggesting causal mechanism to be investigated in future studies.

6 Summary

Reliable CATE estimation is important for making the most of our observational data sets. As RDD continues to gain popularity in industry — for example, as a byproduct of business decisions being made based on an observed index — being able to use these data to explore subgroup treatment effects is a big advantage. In this paper, we have demonstrated that a BART ensemble of treed linear regressions — which we call BARDDT — estimates RDD CATEs successfully and markedly better than available alternatives and have demonstrated how to interpret the resulting estimates on a reanalysis of a policy evaluation question from education (Lindo et al., 2010). Software for fitting BARDDT is freely available in the `stochtree` package, available in both `R` and Python.

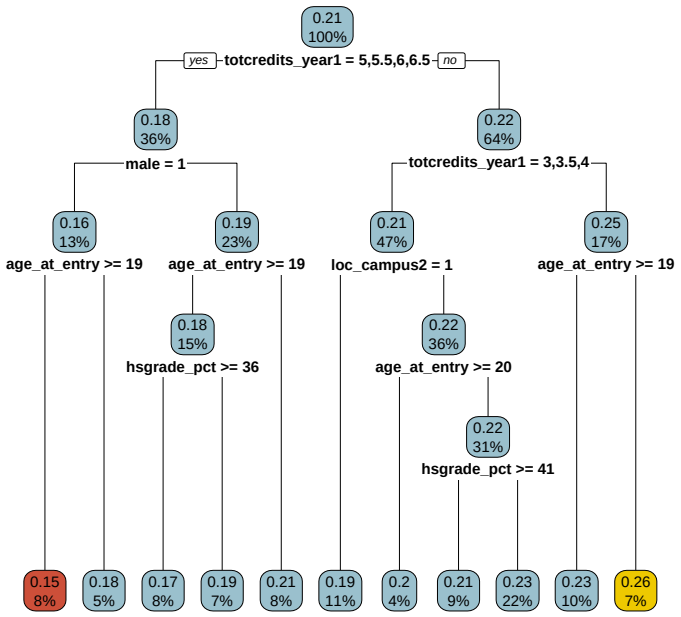

Figure 7: Regression tree fit to posterior point estimates of individual treatment effects: top number in each box is the average subgroup treatment effect, lower number shows the percentage of the total sample in that subgroup; the tree flags credits in first year, gender, and age at entry as important moderators.

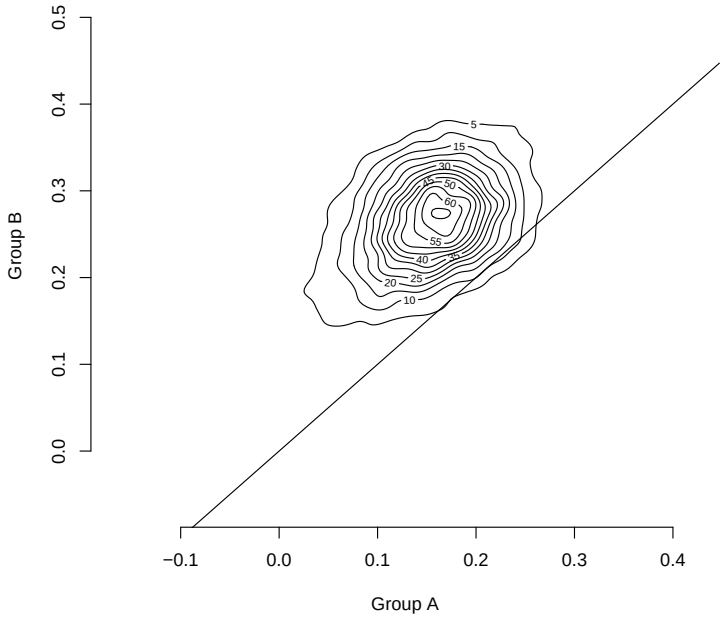

Figure 8: Kernel density estimates for the joint CATE posterior between male students who entered college older than 19 and attempted at least 5 credits in the first year (leftmost leaf node, red) and students who entered college younger than 19 and attempted between 4 and 5 credits in the first year (rightmost leaf node, gold)

References

Becker, S. O., Egger, P. H., and Von Ehrlich, M. (2013). Absorptive capacity and the growth and investment effects of regional transfers: A regression discontinuity design with heterogeneous treatment effects. *American Economic Journal: Economic Policy*, 5(4):29–77.

Branson, Z., Rischard, M., Bornn, L., and Miratrix, L. W. (2019). A nonparametric Bayesian methodology for regression discontinuity designs. *Journal of Statistical Planning and Inference*, 202:14–30.

Calonico, S., Cattaneo, M. D., and Titiunik, R. (2015). rdrobust: An R package for robust nonparametric inference in regression-discontinuity designs. *R J.*, 7(1):38.

Chib, S., Greenberg, E., and Simoni, A. (2023). Nonparametric Bayes analysis of the sharp and fuzzy regression discontinuity designs. *Econometric Theory*, 39(3):481–533.

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

Chipman, H. A., George, E. I., and McCulloch, R. E. (2002). Bayesian treed models. *Machine Learning*, 48:299–320.

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

Chipman, H. A., George, E. I., McCulloch, R. E., and Shively, T. S. (2022). mbart: multidimensional monotone bart. *Bayesian Analysis*, 17(2):515–544.

Gamerman, D. and Lopes, H. F. (2006). *Markov chain Monte Carlo: stochastic simulation for Bayesian inference*. Chapman and Hall/CRC.

Gramacy, R. B. and Lee, H. K. H. (2008). Bayesian treed Gaussian process models with an application to computer modeling. *Journal of the American Statistical Association*, 103(483):1119–1130.

Hahn, J., Todd, P., and Van der Klaauw, W. (2001). Identification and estimation of treatment effects with a regression-discontinuity design. *Econometrica*, 69(1):201–209.

Hahn, P. R., Carvalho, C. M., Puelz, D., He, J., et al. (2018). Regularization and confounding in linear regression for treatment effect estimation. *Bayesian Analysis*, 13(1):163–182.

Hahn, P. R., Dorie, V., and Murray, J. S. (2019). Atlantic causal inference conference (ACIC) data analysis challenge 2017. *arXiv preprint arXiv:1905.09515*.

Hahn, P. R., Murray, J. S., Carvalho, C. M., et al. (2020). Bayesian regression tree models for causal inference: regularization, confounding, and heterogeneous effects. *Bayesian Analysis*.

Hahn, P. R., Murray, J. S., and Manolopoulou, I. (2016). A Bayesian partial identification approach to inferring the prevalence of accounting misconduct. *Journal of the American Statistical Association*, 111(513):14–26.

He, J. and Hahn, P. R. (2023). Stochastic tree ensembles for regularized nonlinear regression. *Journal of the American Statistical Association*, 118(541):551–570.

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

Karabatsos, G. and Walker, S. G. (2015). A Bayesian nonparametric causal model for regression discontinuity designs. In *Nonparametric Bayesian Inference in Biostatistics*, pages 403–421. Springer.

Künzel, S. R., Sekhon, J. S., Bickel, P. J., and Yu, B. (2019). Metalearners for estimating heterogeneous treatment effects using machine learning. *Proceedings of the National Academy of Sciences*, 116(10):4156–4165.

Lindo, J. M., Sanders, N. J., and Oreopoulos, P. (2010). Ability, gender, and performance standards: Evidence from academic probation. *American Economic Journal: Applied Economics*, 2(2):95–117.

Linero, A. R. (2018). Bayesian regression trees for high-dimensional prediction and variable selection. *Journal of the American Statistical Association*, 113(522):626–636.

Murray, J. S. (2021). Log-linear Bayesian additive regression trees for multinomial logistic and count regression models. *Journal of the American Statistical Association*, 116(534):756–769.

Orlandi, V., Murray, J., Linero, A., and Volfovsky, A. (2021). Density regression with Bayesian additive regression trees. *arXiv preprint arXiv:2112.12259*.

Papakostas, D., Hahn, P. R., Murray, J., Zhou, F., and Gerakos, J. (2023). Do forecasts of bankruptcy cause bankruptcy? A machine learning sensitivity analysis. *The Annals of Applied Statistics*, 17(1):711–739.

Pearl, J. (2009). *Causality*. Cambridge university press.

Pratola, M. T., Chipman, H. A., George, E. I., and McCulloch, R. E. (2020). Heteroscedastic BART via multiplicative regression trees. *Journal of Computational and Graphical Statistics*, 29(2):405–417.

Reguly, A. (2021). Heterogeneous treatment effects in regression discontinuity designs. *arXiv preprint arXiv:2106.11640*.

Saha, E. (2023). Theory of posterior concentration for generalized Bayesian additive regression trees. *arXiv preprint arXiv:2304.12505*.

Sparapani, R. A., Logan, B. R., McCulloch, R. E., and Laud, P. W. (2016). Nonparametric survival analysis using Bayesian additive regression trees (BART). *Statistics in Medicine*, 35(16):2741–2753.

Starling, J. E., Murray, J. S., Carvalho, C. M., Bukowski, R. K., Scott, J. G., et al. (2020). BART with targeted smoothing: An analysis of patient-specific stillbirth risk. *Annals of Applied Statistics*, 14(1):28–50.

Thistlethwaite, D. L. and Campbell, D. T. (1960). Regression-discontinuity analysis: An alternative to the ex post facto experiment. *Journal of Educational Psychology*, 51(6):309.

Wilkinson, G. and Rogers, C. (1973). Symbolic description of factorial models for analysis of variance. *Journal of the Royal Statistical Society Series C: Applied Statistics*, 22(3):392–399.