Modified BART for Learning Heterogeneous Effects in Regression Discontinuity Designs

Rafael Alcantara
The University of Texas at Austin
and
Meijia Wang
Google
and
P. Richard Hahn
Arizona State University
and
Hedibert Lopes
Insper

July 26, 2024

Abstract

This paper introduces BART-RDD, a sum-of-trees regression model built around a novel regression tree prior, which incorporates the special covariate structure of regression discontinuity designs. Specifically, the tree splitting process is constrained to ensure overlap within a narrow band surrounding the running variable cutoff value, where the treatment effect is identified. It is shown that unmodified BART-based models estimate RDD treatment effects poorly, while our modified model accurately recovers treatment effects at the cutoff. Specifically, BART-RDD is perhaps the first RDD method that effectively learns conditional average treatment effects. The new method is investigated in thorough simulation studies as well as an empirical application looking at the effect of academic probation on student performance in subsequent terms (Lindo et al., 2010).

Keywords: Bayesian causal forest, tree ensembles, regression discontinuity design

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. Thus in an RDD, deterministic treatment assignment implies that conditional unconfoundedness is trivially satisfied, given the running variable. However, controlling for the running variable introduces a complete lack of overlap. Identification of treatment effects from RDDs relies on assumptions that permit coping with this lack of overlap.

Previous work has shown that treatment effects can be estimated from RDDs as the magnitude of a discontinuity in the conditional mean response function at the cutoff (Hahn et al., 2001), under the assumption that this conditional mean function is continuous but for the impact of the treatment. This paper investigates the use of Bayesian additive regression tree models (Chipman et al., 2010; Hahn et al., 2020) for the purpose of fitting RDD data with additional covariates for estimating conditional average treatments effects (CATE) at the cutoff. Broadly, our work expands on both frequentist and Bayesian methods for performing RDD analyses that incorporate covariates in addition to the running variable. Relative to earlier works, the method proposed here accommodates a richer set of data generating processes and admits more convenient sensitivity analysis and methods for visualizing heterogeneity. Most importantly, our estimator is one of the few RDD estimators which allow for exploring heterogeneity in a data-driven manner, instead of relying on separate ATE estimation for predetermined subgroups.

1.1 Previous work

The inclusion of covariates in RDD models has been studied by Calonico et al. (2019), who extend the local linear regression to include covariates linearly and discuss the implications of this in terms of bias and variance, and Frölich and Huber (2019) who propose a nonparametric kernel regression method which increases precision and may reduce bias and restore identification in data with discontinuities in the covariate set at the threshold, provided that all relevant discontinuous covariates are included. Additionally, Arai et al. (2024) and Kreiss and Rothe (2023) study RDD regressions with high-dimensional covariate sets. The latter two essentially consist of a pre-selection step where one fits a variable selection model (typically a Lasso) to either the full sample or a subsample closer to the cutoff and then fits the local polynomial estimator of Calonico et al. (2019) to the reduced feature set. These methods inherit the local polynomial's linearity assumption which can lead to high variance estimators in the presence of strong heterogeneity if the running variable and other features interact in more complex ways. The estimator proposed by Frölich and Huber (2019) is more flexible in that regard because it allows for feature-specific kernels, extending the traditional local RDD regression. However, this can render the method computationally infeasible as the dimension of the feature set increases.

The previous works discuss inclusion of covariates from a perspective of obtaining precision gains and barely discuss effect moderation. Regarding effect heterogeneity, Frandsen et al. (2012) and Shen and Zhang (2016) discuss it from the perspective of distributional effects, while Cattaneo et al. (2016) and Bertanha (2020) discuss heterogeneity arising from settings with multiple cutoffs. These works do not focus specifically on heterogeneity in the form of moderation by additional variables. Becker et al. (2013) extend the traditional local regression to include interaction terms between the running variable and additional features. Hsu and Shen (2019) develop hypothesis tests for detecting effect moderation in the local regression setup by means of comparison between the ATE parameter for the whole sample and pre-specified subgroups of interest. These methods still depend on reasonable previous knowledge about potential sources of heterogeneity.

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. As will be discussed later, our method can be seen as an intermediate approach between such global approximations and the local linear regression ubiquitous in the frequentist literature, since we use the entire data to estimate the outcomes but take advantage of the local nature of BART for estimation near the cutoff.

Reguly (2021) proposes what is perhaps the closest in spirit to our method. The author proposes a modification to the basic Classification and Regression Tree (CART) 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 model can be seen as a non-linear polynomial regression where the parameters depend on the covariates via the CART fit. This approach presents two important differences compared to ours. First, it is a single tree method, whereas we propose a forest model. Second, the flexibility of the tree is only used for the additional covariates, but the leaf regressions are still polynomials of the running variable.

2 Background

This paper brings together ideas from many different areas, each with their own terminology and notation. In this section we review the basics of regression discontinuity designs, BART, and Bayesian causal forests.

2.1 Regression Discontinuity Designs

Following Imbens and Lemieux (2008), we frame the RD setting in a potential outcomes model, which can be briefly described as follows. Let Z denote a binary treatment variable and Y_i^z denote the potential outcome of unit i under treatment state $Z_i = z$. The treatment effect for unit i is defined as:

$$\tau_i \coloneqq Y_i^1 - Y_i^0. \tag{1}$$

Let X denote the running variable and W denote a set of additional covariates. Commonly, interest lies on the average and conditional average treatment effect (ATE/CATE):

$$\begin{aligned} \mathbb{E}[\tau_i] &= \mathbb{E}[Y_i^1 - Y_i^0] \\ \mathbb{E}[\tau_i \mid X, W] &= \mathbb{E}[Y_i^1 - Y_i^0 \mid X, W]. \end{aligned} \tag{2}$$

These quantities are of course unobservable since each unit is only observed at a single given treatment state. However, under the following assumptions we can link τ to the observed outcome and covariates (Y, Z, X, W).

Assumption 2.1 (SUTVA) *This assumption has two components: consistency and no-interference, which are represented, respectively, as:*

$$\begin{aligned} &Y = Y^0 + Z(Y^1 - Y^0) \\ &Y_i^1, Y_i^0 \perp\!\!\!\perp Z_j, \end{aligned} \tag{3}$$

for all $i, j \in \{1, \ldots, n\}$, where $i \neq j$.

Assumption 2.2 (Mean conditional unconfoundedness) *Y^1, Y^0 are mean independent of Z conditional on X, W:*

$$\begin{aligned} \mathbb{E}[Y^1 \mid Z, X, W] &= E[Y^1 \mid X, W] \\ \mathbb{E}[Y^0 \mid Z, X, W] &= E[Y^0 \mid X, W]. \end{aligned} \tag{4}$$

Under assumptions (2.1) and (2.2), the CATE is identified as:

$$\mathbb{E}[Y^1 - Y^0 \mid X, W] = \mathbb{E}[Y \mid Z = 1, X, W] - \mathbb{E}[Y \mid Z = 0, X, W]. \tag{5}$$

While the previous assumptions lead to identification of the CATE, one more assumption is necessary for estimation:

Assumption 2.3 (Conditional overlap) *Both treatment states have a positive probability conditional on X, W:*

$$0 < P(Z = z \mid X, W) < 1. \tag{6}$$

In words, assumption (2.3) allows one to compare treated and untreated units in any region of the support of (X, W), leading to the construction of valid causal contrasts.

The distinctive feature of the RDD is that Z is a deterministic function of X:

$$Z_i = \begin{cases} 0, & \text{if} \quad X_i \quad < c \\ 1, & \text{if} \quad X_i \quad \geq c \end{cases}$$

for a known cutoff value c[1].

The deterministic assignment mechanism implies that controlling for X is sufficient to ensure unconfoundedness. However, this control induces a complete lack of overlap. Therefore, treatment effect estimation in the RDD requires additional assumptions to circumvent this issue. In order to discuss these assumptions, we write the expectation of Y given (X, W, Z) as:

$$\begin{aligned} \mathbb{E}[Y \mid X, W, Z] &= \mu(X, W) + \tau(X, W)Z \\ \mu(X, W) &= \mathbb{E}[Y \mid X, W, Z = 0] \\ \tau(X, W) &= \mathbb{E}[Y \mid X, W, Z = 1] - \mathbb{E}[Y \mid X, W, Z = 0]. \end{aligned} \tag{7}$$

Because of the lack of overlap, one can only learn $\mu(X, W)$ in the region $X < c$ and $\mu(X, W) + \tau(X, W)$ in the region $X \geq c$, so that inferences concerning $\tau(X, W)$ at arbitrary x cannot be obtained without further assumptions. We now discuss the kinds of assumptions necessary for estimation in the RDD.

For some $\epsilon > 0$, let $\mathcal{X}_\varepsilon^- = \{x \mid c - \epsilon < x < c\}$, $\mathcal{X}_\varepsilon^+ = \{x \mid c \leq x < c + \epsilon\}$, and $\mathcal{X}_\epsilon = \mathcal{X}_\epsilon^- \cup \mathcal{X}_\epsilon^+$. Suppose that, for $X \in \mathcal{X}_\epsilon$, the treatment effect function is independent from the treatment variable conditional on X. Then:

$$\begin{aligned} &\mathbb{E}[\mathbb{E}[Y \mid X, W, Z = 1] \mid X \in \mathcal{X}_\epsilon] - \mathbb{E}[\mathbb{E}[Y \mid X, W, Z = 0] \mid X \in \mathcal{X}_\epsilon] \\ =&\mathbb{E}[\tau(X, W) \mid X \in \mathcal{X}_\epsilon^+] + (\mathbb{E}[\mu(X, W) \mid X \in \mathcal{X}_\epsilon^+] - \mathbb{E}[\mu(X, W) \mid X \in \mathcal{X}_\epsilon^-]). \end{aligned} \tag{8}$$

Suppose that:

$$\begin{aligned} \mathbb{E}[\mu(X, W) \mid X \in \mathcal{X}_\epsilon^+] = \mathbb{E}[\mu(X, W) \mid X \in \mathcal{X}_\epsilon^-] = \mathbb{E}[\mu(X, W) \mid X \in \mathcal{X}_\epsilon] \\ \mathbb{E}[\tau(X, W) \mid X \in \mathcal{X}_\epsilon^+] = \mathbb{E}[\tau(X, W) \mid X \in \mathcal{X}_\epsilon^-] = \mathbb{E}[\tau(X, W) \mid X \in \mathcal{X}_\epsilon]. \end{aligned} \tag{9}$$

[1] Our focus lies on the so-called "sharp" RDD, in which case there is perfect compliance — as opposed to the "fuzzy" RDD, in which case compliance is imperfect — so the perfect compliance assumption is implicit throughout the text.

Then, the ATE[2] is identified inside this region:

$$\begin{aligned}&\mathbb{E}[\mathbb{E}[Y \mid X, W, Z = 1] \mid X \in \mathcal{X}_\epsilon] - \mathbb{E}[\mathbb{E}[Y \mid X, W, Z = 0] \mid X \in \mathcal{X}_\epsilon] \\ =&\mathbb{E}[\tau(X, W) \mid X \in \mathcal{X}_\epsilon].\end{aligned} \tag{10}$$

This is the basis of the continuity-based identification approach introduced by Hahn et al. (2001). Under that setting, if these conditions can be assumed to hold at least as $\epsilon \to 0$ — *i.e.* if the expectation of the μ and τ functions are continuous at $X = c$ — the ATE at this point is identified as $\mathbb{E}[\tau(X = c, W) \mid X = c]$.

If interest lies in identification of the CATE in some region of the feature set $W = w$, we need similar assumptions about the expectations conditional on W. Suppose that, for all $x_- \in \mathcal{X}_\epsilon^-$ and $x_+ \in \mathcal{X}_\epsilon^+$:

$$\begin{aligned}\mu(X = x_-, W = w) = \mu(X = x_+, W = w) \\ \tau(X = x_-, W = w) = \tau(X = x_+, W = w).\end{aligned} \tag{11}$$

Then, the CATE at $W = w$ is identified in the region $\mathcal{X}_\epsilon$ by $\tau(X, W = w)$. As before, if these equalities hold as $\epsilon \to 0$, *i.e.* if μ and τ are continuous at $X = c$, the CATE for $W = w$ is identified at that point. Because we propose an estimator for the CATE, (11) is assumed to hold for the remainder of the text. However, it is worth emphasizing that only (9) is required for identification of the ATE, so that, even if $\tau(X, W)$ does not identify any CATE, estimates of this function can still be used to produce ATE estimates if the relevant assumptions hold.

To introduce some of the challenges faced by tree models in the RDD context, consider the treatment effect estimate in a single tree model for a partition in the tree fit that contains $X = c$, denoted by $\mathcal{B}$. For points inside that partition, define $\mathcal{X}_+^{\mathcal{B}} = \{x \mid c \leq x \leq \overline{x}\}$,

[2] As is commonly done in the RDD literature, we refer to the CATE conditional only on X as the ATE and use CATE only when conditioning on W as well

$\mathcal{X}_-^{\mathcal{B}} = \{x \mid \underline{x} \leq x < c\}$, where $\underline{x}$ and $\overline{x}$ are the smallest and largest values of X inside the partition, respectively, and $\mathcal{X}^{\mathcal{B}} = \mathcal{X}_+^{\mathcal{B}} \cup \mathcal{X}_-^{\mathcal{B}}$. Then:

$$\begin{aligned} &E[Y \mid X \in \mathcal{X}^{\mathcal{B}}, W, Z = 1] - E[Y \mid X \in \mathcal{X}^{\mathcal{B}}, W, Z = 0] \\ =&\mu(X \in \mathcal{X}_+^{\mathcal{B}}, W) - \mu(X \in \mathcal{X}_-^{\mathcal{B}}, W) + \tau(X \in \mathcal{X}_+^{\mathcal{B}}, W). \end{aligned} \tag{12}$$

This means that the bias for the cutoff treatment effect estimate inside this partition is given by:

$$\tau_{\text{bias}} = \tau(X = c, W) - \tau(X \in \mathcal{X}_+^{\mathcal{B}}, W) + \mu(X \in \mathcal{X}_+^{\mathcal{B}}, W) - \mu(X \in \mathcal{X}_-^{\mathcal{B}}, W). \tag{13}$$

Equation (13) shows how the bias in a tree model is determined by the composition of the leaf nodes. In particular, it implies that the bias goes to zero as $\underline{x} \to c$ and $\overline{x} \to c$. In words, although nodes that are too tight around $X = c$ could lead to an increase in variance due to the decreasing number of available points in the leaves, nodes that contain too wide regions around the cutoff could lead to extremely biased estimates if μ and τ feature a wide range of values in that partition. Therefore, when considering a split in a tree, minimal variation of the prognostic and treatment effect functions around the cutoff inside the generated leaves should be a key component of the tree growth process. This is the essence of the BART-RDD model, which will be discussed in more detail in section 3.

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 assumed to be drawn from the tree prior described in Chipman et al. (1998). Letting $f(x) = \mathbb{E}(Y \mid X = x)$ denote a smooth function of a covariate vector X, the BART model is traditionally written

$$\begin{aligned} Y_i &= f(x_i) + \varepsilon_i, \\ &= \sum_{j=1}^{k} g_j(x_i; T_j, \mathbf{m}_j) + \varepsilon_i, \end{aligned} \tag{14}$$

where $\varepsilon_i \sim \mathrm{N}(0, \sigma^2)$ is a normally distributed additive error term. Here, each $g_j(x; T_j, \mathbf{m}_j)$ denotes a piecewise function of x defined by a set of splitting rules T_j that partitions the domain of x into disjoint regions, and a vector, $\mathbf{m}_j$, which records the values $g(\cdot)$ takes on each of those regions. Therefore, the parameters of a standard BART regression model are $(T_1, \mathbf{m}_1), \ldots, (T_k, \mathbf{m}_k)$ and σ.

In the context of estimating heterogeneous treatment effects, the BART model can be used in one of two ways: by fitting a single model for the whole sample, with the treatment variable included as just another covariate (see Hill (2011)); or by fitting a separate model for treatment and control units. Following the recent literature on machine learning estimators for causal inference, we name these approaches, respectively, "S-BART" and "T-BART"[3]. Hahn et al. (2020) note that both the S-BART and T-BART estimators are not ideal for causal inference because the models lead to an implicit prior over treatment effects which is hard to control. The authors propose the Bayesian Causal Forest (BCF) model, which reparametrizes the response as a sum of a prognostic and a treatment effect component, each of which is assigned a separate BART prior:

$$\begin{aligned} & Y_i = \mu(X_i, \mathrm{w}_i) + \tau(X_i, \mathrm{w}_i) b_{z_i} + \varepsilon_i, \quad \varepsilon_i \sim N(0, \sigma^2), \\ & b_0 \sim \mathrm{N}(0, 1/2), \quad b_1 \sim \mathrm{N}(0, 1/2). \end{aligned} \tag{15}$$

where $\mu(\cdot)$ is the prognostic function and $\tau(\cdot)$ is the treatment effect function[4].

[3] The "S" stands for single, and the "T" for two; see Künzel et al. (2019)

[4] This terminology is motivated by the case where $b_0 = 0$ and $b_1 = 1$, in which case $\mu(x) = \mathbb{E}(Y^0 \mid X = x)$ and $\tau(x) = \mathbb{E}(Y^1 \mid X = x) - \mathbb{E}(Y^0 \mid X = x)$.

The original BART algorithm searches the tree space by making small changes at a given tree at each step, which can make it very slow, especially for high-dimensional data, a property evidently inherited by the BCF model. He and Hahn (2023) propose the XBART algorithm, which uses the same tree-splitting criteria that define the BART prior, but samples a new tree at each step, exploring the tree space more effectively. Krantsevich et al. (2023) propose the XBCF algorithm, which adapts the XBART algorithm to the BCF model.

Our proposal consists of an adaptation of the XBCF algorithm to the RDD context. Because our strategy consists of adding constraints at the individual tree level, we will not focus on specific details about the construction of the BART prior, the BCF model or the XBART and XBCF algorithms. We refer the interested reader to the appendix for a much more thorough discussion on these topics.

3 Bayesian Regression Trees for RDD

Unlike local polynomial regression methods, a BART-based approach to RDD does not have to pre-specify a set of global basis functions nor must it entirely discard data outside of a neighborhood of the cutoff. These features are particularly useful when incorporating additional covariates W for the purpose of CATE estimation. However, this flexibility comes at a cost and estimation can go wrong in one of two ways. First, a BART-based T-Learner may give poor estimates of $\mathbb{E}(Y \mid X = c, W)$ because tree ensembles with constant leaf models are known to extrapolate poorly (but see Starling et al. (2021) and Wang et al. (2024) for alternatives, which we do not pursue here.) Second, a BART-based S-Learner may estimate the *response surface* at $X = c$ reasonably well, but still provide biased estimates of the *treatment effect* because some of its individual trees end up using data very far from the cutoff. These flaws will be depicted in numerical examples below.

To overcome these problems, we introduce novel splitting constraints, which ensure that the data used to make predictions at $X = c$ warrant a causal interpretation. Specifically, we impose the constraint that our ensembles must be composed of trees where any partition containing the point $(x = c, w)$ is estimated from data "close enough" to the cutoff from both sides.

3.1 Splitting Constraints for RDD with Regression Trees

The proposed constraints have two distinct, though related, goals. First, we need the treatment-control contrast – upon which $\tau(x = c, w)$ will be estimated – to be well-defined: for this we require observations from both treatment arms (e.g. overlap). Second, because we cannot rely on observations far from the cutoff to estimate the treatment effect, we insist that a partition that includes $x = c$ have a strong majority of its observations within a narrow, user-defined band about the cutoff. This constraint defines a set of suitably *localized* basis functions from which to perform causal inference at the cutoff.

More formally, these constraints can be expressed as follows. For a user-defined bandwidth parameter $h > 0$, we assume that the potential outcome mean function does not vary abruptly inside the interval $[c - h, c + h]$, which we refer to as the "identification strip". Let $B \subset \mathcal{X}$ be a hypercube corresponding to a node in a regression tree and let N_b denote the number of observations falling within B. Further, let n_l denote the number of observations in $B \cap [c - h, c)$ and n_r denote the number of observations in $B \cap [c, c + h]$. For user-specified variables $N_{Omin} \in \mathbb{N}^+$ and $\alpha \in (0, 1)$, the leaf node region B is valid if it satisfies the following condition:

$$\begin{aligned}&\big(\forall w \mid (x=c, w) \notin B\big)\cup \\ &\bigg((\exists w \mid (x=c, w) \in B) \cap (\min{(n_l, n_r)} \geq N_{Omin}) \cap ((n_l + n_r)/N_b \geq \alpha) \bigg).\end{aligned} \tag{16}$$

The initial clause says that any node which does not make predictions at the cutoff remains entirely unrestricted; the second clause says that any node that *does* make predictions at $x = c$ has to have both i) a minimum number of observations within the cutoff region on either side of the cutoff, as well as ii) not too many observations, proportionally, outside of the identification strip.

3.2 Stochastic search for valid partitions

For nodes predicting at the cutoff, the two conditions (i and ii) above have qualitatively different ramifications for the stochastic search for valid partitions. In particular, the first condition, if unsatisfied, can never become satisfied by further branching, while the second condition, if unsatisfied, *can* be satisfied by further branching, by trimming away observations outside of the identification strip. This observation motivates us to use XBART/XBCF rather than standard BART MCMC for our model fitting. An unmodified local random walk would violate recurrence because certain valid states can only be reached by passing through invalid states; as a practical matter, reaching valid partitions by a random walk would be highly inefficient. By utilizing the Grow-From-Root algorithm from XBART, described in the appendix, passing through invalid states to reach favorable valid states is a simple matter of not terminating the growth process at an invalid state. Specifically, never accept a partition that violates condition i, and never stop at a partition that violates condition ii.

In practice, this new stochastic search procedure is implemented by modifying the likelihood calculation in steps 8 and 10 of the GFR algorithm for XBCF as follows. Consider a candidate split with cutpoint c_{jk} which splits the current node into left and right nodes. Let $B_x^{(l)}$ denote the range of x which the left node covers and similarly define $B_x^{(r)}$. Let n_{ll} and n_{lr} denote the number of observations such that $x \in [c-h, c)$ and $x \in [c, c+h]$ in the left node, and n_{rl} and n_{rr} denote the same quantities in the right node, respectively. If,

$$c \in B_x^{(l)} \quad \text{and} \quad \max(n_{ll}, n_{lr}) < N_{Omin} \tag{17}$$

or if

$$c \in B_x^{(r)} \quad \text{and} \quad \max(n_{rl}, n_{rr}) < N_{Omin}, \tag{18}$$

this split violates condition (i). Therefore, we consider this an invalid partition and set $L(c_{jk}) = 0$. If the split does not violate condition (i), we calculate its likelihood as in the GFR algorithm. For condition (ii), we check whether:

$$c \in B_x \quad \text{and} \quad \frac{n_l + n_r}{N_b} < \alpha. \tag{19}$$

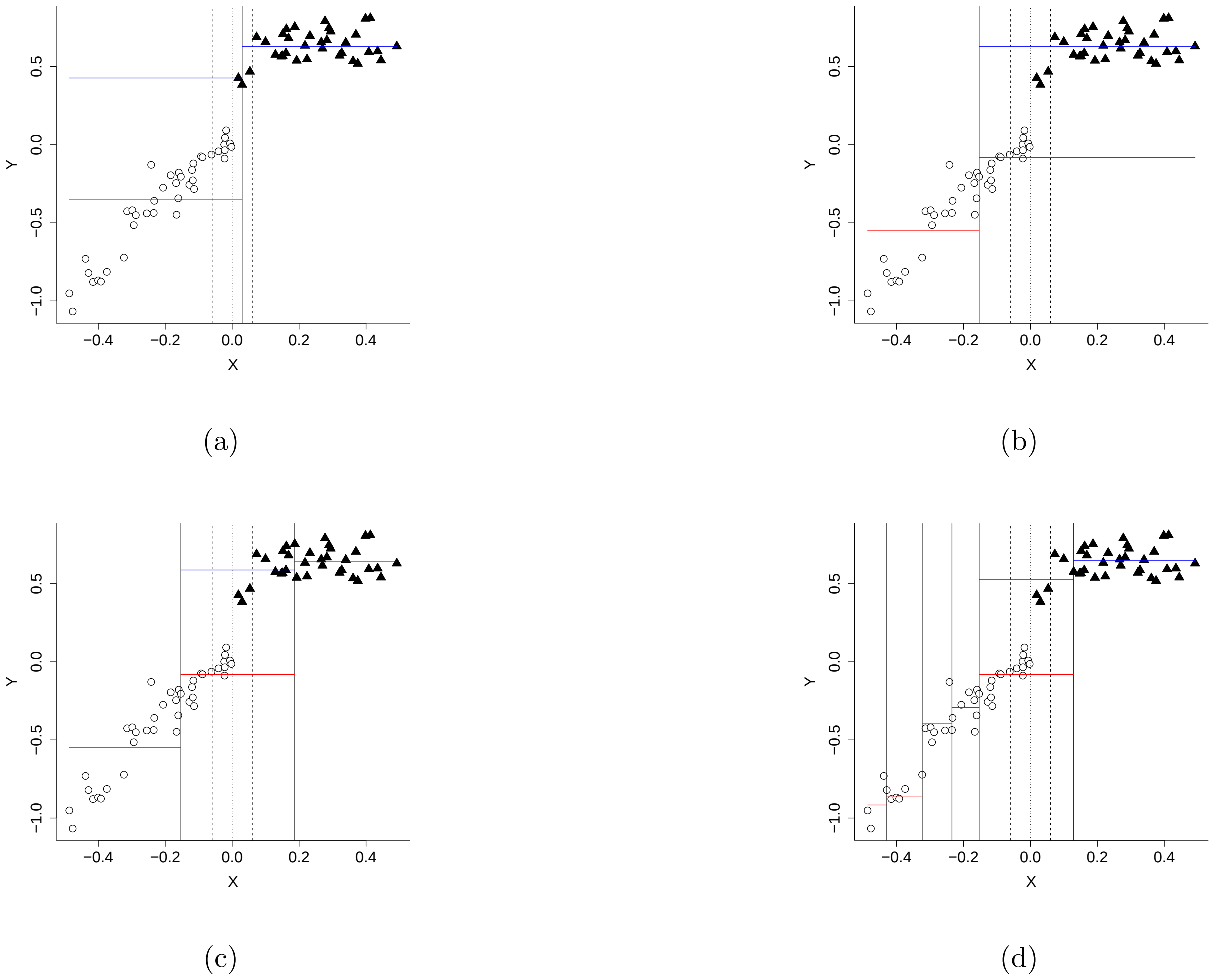
If so, we set the likelihood of the no-split option $L(\emptyset) = 0$ unless there are no other valid splits, in which case we set $L(\emptyset) = 1$. In the latter case, we end up with a tree that is still invalid, as it violates condition ii; our implementation monitors for this eventuality and find that it rarely if ever occurs in most data sets.

3.3 Illustration of the constraints and search

The impact of expression (16) on the fitted trees may be visualized by considering a concrete example. Consider a tree fit with only the running variable (X). Figure 1 plots X against some outcome Y for a dataset with 75 observations, presenting different partitions in X. For this example, the cutoff is $c = 0$ – denoted by the dotted line – and the ATE at that point is equal to 0.5. We consider a window of $h = 0.06$, denoted by the dashed lines in the plots. The treated units ($x \geq c$) are denoted by black triangle dots and the control units are denoted by white round dots. Splits in X are denoted by solid lines. For each partition, we represent the inferred potential outcome as a red line for untreated and blue line for treated outcomes. For the partitions that include both types of observations — *i.e.* points from both sides of the cutoff — we represent both potential outcomes.

Panel 1a shows a split which is invalid since it cuts through the identification strip, leading to a node that contains only one point to the right of the cutoff in that region. The ATE at the cutoff for that tree is predicted to be 0.78. Panel 1b presents a split which only violates condition (ii), since it does not cut through the identification strip, but features a node with too many points outside the strip. The ATE at the cutoff for that tree is predicted to be 0.7. This tree can be made valid by 'trimming out' points too far from the cutoff in the right node. Panel 1c presents an additional split that does exactly that. The ATE at the cutoff for that tree is predicted to be 0.67. Finally, panel 1d presents another tree, with a couple of additional splits to the left of the identification strip, and a split to the right that's closer to the strip. Since the new nodes generated do not include the identification strip, they are all potentially valid. The ATE at the cutoff for that tree is predicted to be 0.6.

This example highlights problems that unmodified BART models face in the context of the RDD. Note that all trees above are valid under the standard BART prior. If the nodes that contain $X = c$ include points close to the cutoff from both sides, but many points far from it, these trees will only lead to reasonable causal contrasts if Y is relatively constant with respect to X. Otherwise, such trees should exhibit strong bias if the prognostic or treatment effect functions vary substantially, as is the case in the previous example, which illustrates the bias described in equation (13). In fact, as we move closer towards the kinds



(a)

(b)

(c)

(d)

Figure 1: Tree examples: panel (a) shows a tree with a split that violates condition i; panel (b) presents a tree with a split that violates condition ii; the tree in panel (c) is an example of the kind of tree that would be accepted by the algorithm; the tree in panel (d) is the same as the one in panel (c) with some additional splits that do not contain the identification strip

of trees that would be accepted by BART-RDD — *i.e.* moving from the first panel to the last — we decrease the bias in the predicted ATE at the cutoff. The important distinction here is that, while all BART-based models could reach reasonable trees, only BART-RDD is guaranteed to do so by rejecting trees that do not behave appropriately.

3.4 Prior elicitation

The question remains as to how one should set the relevant parameters in order to obtain good predictions. A completely general rule cannot be expected, as the impact each parameter has in the estimation is highly data-dependent. Nonetheless, it is instructive to consider what kinds of restrictions our parameters imply to the tree search process and how they impact prior bias.

First, consider the bandwidth parameter h. On the one hand, h should be set as low as possible for two reasons. First, setting h too high makes it more likely that a given tree would cut through the identification strip. In particular if h is such that the strip covers all the support of X, the algorithm would only accept trees that do not use X for partitions. Given the essential role that X plays in the outcome distribution in an RDD setting, not using it for the tree splits would lead to severely biased trees. Additionally, since the goal is to make inference enforcing smoothness over X at a specific point ($X = c$), one should only use points as close to c as possible to obtain better approximations of the true function around that point.

On the other hand, there is also a limit to how low one can set h for any particular dataset. In particular, if h is so small that there are no points inside the identification strip, any tree will produce nodes with an empty overlap region and, thus, be invalid. This means that h also interacts with N_{Omin} in that extreme: even if there are points in the identification strip, if there are less than N_{Omin} points, the same phenomenon happens, making all trees invalid.

Next, we turn to N_{Omin}: if it is set to 0, the trees could produce nodes which contain the overlap region but have no points inside it. Thus, predictions near the cutoff could be based only on observations too far from the cutoff, which would undermine the constraint associated with this parameter. Setting N_{Omin} too high could be too restrictive since there would be fewer valid nodes containing the overlap region, which could bias the individual posterior distributions or at least make it harder to detect heterogeneity in the data, since we could have very short trees.

Finally, we note that if α is set too low, we allow for many points outside the strip to influence the results from nodes that do include the strip. However, setting it high is not necessarily a problem since, as discussed in the previous section, nodes that do not satisfy this criterion can be made to satisfy it by simply 'trimming' the outer region of nodes containing the identification strip. It is important to note, however, that setting α too high could lead to many forced splits in the boundaries of the identification strip, which can lead to an increase in variance if these additional splits are not particularly relevant. Therefore, this parameter should also be chosen carefully.

Clearly, it is nontrivial how these considerations might interact in a given sample. This reflects the immense delicacy of the regression discontinuity design itself, rather than a limitation intrinsic to the BART-RDD proposed; all RDD methods require grappling with how to set tuning parameters. Our proposed approach is via a *prior predictive elicitation* procedure. Specifically, we recommend, for a given sample (x, w, y), the following:

1. Generate s samples of a synthetic data from a known DGP using (x, w)

2. Fit BART-RDD to each sample for different values of the prior parameters

3. Choose the parameter values which lead to lower RMSE values for the ATE in those s synthetic samples.

We find that generating the synthetic data from a simple model with no treatment effect heterogeneity and relatively small ATE leads to finding good values for the prior parameters even when the true data exhibits strong heterogeneity or large effects. This is also a reasonable prior for the treatment effects, unless one has strong reason to believe in a more complex scenario. In the simulation studies to follow, we use this procedure to choose h, N_{Omin} and α. A detailed analysis of this procedure may be found in the supplemental appendix, where we discuss its application in the context of the simulations.

4 Simulation studies

4.1 Setup

In order to investigate the properties of the BART-RDD algorithm, we perform a simulation study comparing its performance to an S-learner BART fit (S-BART), a T-learner BART fit (T-BART), the robust bias-corrected local linear regression (LLR), as implemented by Calonico et al. (2015), and the cubic spline estimator (CGS) of Chib et al. (2023). The goal of this exercise is twofold. First, we want to investigate how BART-RDD compares with off-the-shelf implementations of BART applied to the RDD context. We are able to show that our modification does in fact make the BART prior more suited to this context. Second, we want to compare BART-RDD to estimators that were designed specifically to the RDD context, in particular the local polynomial estimator, by far the most commonly used in the literature, and the cubic splines estimator which is possibly the closest in spirit to that in the Bayesian literature. Besides showing BART-RDD is more suited to the RDD setup than unmodified BART models, we also show that BART-RDD generally performs better and never far worse than the estimators designed specifically for the RDD.

Let X denote the running variable, W an additional set of features, Z the treatment indicator and Y a continuous outcome. We investigate 500 samples of size 1000 of variations of the following DGP:

$$
\begin{array}{lcr}
X \sim 2\mathcal{B}(2,4) - 0.75 & Z = \mathbf{1}(X \geq c) & c = 0 \\
W_1 \sim U(-0.1, 0.1) & \mu(X,W) = \dfrac{\mu_0(X,W)}{\sigma(\mu_0(X,W))}\delta_\mu & \bar{\tau} = \{0.2, 0.5\} \\
W_2 \sim \mathcal{N}(0, 0.2) & \tau(X,W) = \bar{\tau} + \dfrac{\tau_0(X,W)}{\sigma(\tau_0(X,W))}\delta_\tau & \delta_\mu = \{0.5, 1.25\} \\
W_3 \sim \text{Binomial}(1, 0.4) & Y = \mu(X,W) + \tau(X,W)Z + \varepsilon & \delta_\tau = \{0.1, 0.3\} \\
W_4 \sim \text{Binomial}(1, p(x)) & & \varepsilon \sim \mathcal{N}(0,1),
\end{array}
$$

where $\mathcal{B}(2,4)$ denotes a Beta distribution with parameters 2 and 4, $p(x)$ denotes the Gaussian probability density of x with mean c and standard deviation 0.5, and:

$$
\begin{aligned}
\mu_0(X,W) &= 3x^5 - 2.5x^4 - 1.5x^3 + 2x^2 + 3x + 2 + \frac{1}{2}\sum_{p=1}^{4}(w_p - E[w_p]) \\
\tau_0(X,W) &= -0.1x + \frac{1}{4}\sum_{p=1}^{4}(w_p - E[w_p])
\end{aligned}
\tag{20}
$$

For the BART-based models, we fit the model for the whole sample and consider only points inside the BART-RDD window for prediction. Let $S = \{x \mid c - h < x < c + h\}$. We then obtain CATE estimates at $X = c$ for the N_S points in S. For the ATE, we focus on the sample ATE in S, $\mathbb{E}[\tau(X = c, W) \mid X \in S]$, which we estimate by averaging over the CATE estimates in our window:

$$
\bar{\tau} = \frac{1}{N_s}\sum_{i: x_i \in S} \tau(x_i = c, w_i). \tag{21}
$$

Some remarks on the design of this DGP: Although there are other parameters that affect the performance of any estimator, the spread in μ and τ were the only factors that we found to have distinct impacts on different estimators. In other words, the effect of other DGP characteristics in the results were common across estimators in the expected ways. We control these features in the data through the parameters $(\delta_\mu, \delta_\tau)$. The particular choices for these parameters were made in an attempt to replicate realistic behavior in μ and τ. Particularly, we made sure that there are generally no sign changes in the individual treatment effects and that the spread in the prognostic component is larger than the spread in the treatment effects.

4.2 Results

4.2.1 Comparison of ATE Estimates

We begin by examining its performance on the ATE for comparison with other methods and because a good CATE learner should be able to provide good ATE estimates as well. Table 1 and figure 2 present the RMSE for the ATE point estimate produced by each estimator[5].

Table 1: RMSE - ATE

$\bar{\tau}$	δ_μ	δ_τ	BART-RDD	S-BART	T-BART	CGS	LLR
0.2	0.5	0.1	0.114	0.214	0.253	0.370	0.233
0.2	0.5	0.3	0.114	0.228	0.264	0.388	0.243
0.2	1.25	0.1	0.226	0.298	0.424	0.411	0.234
0.2	1.25	0.3	0.250	0.321	0.440	0.445	0.255
0.5	0.5	0.1	0.158	0.257	0.249	0.387	0.247
0.5	0.5	0.3	0.147	0.250	0.258	0.372	0.239
0.5	1.25	0.1	0.251	0.397	0.432	0.437	0.251
0.5	1.25	0.3	0.247	0.402	0.429	0.443	0.245

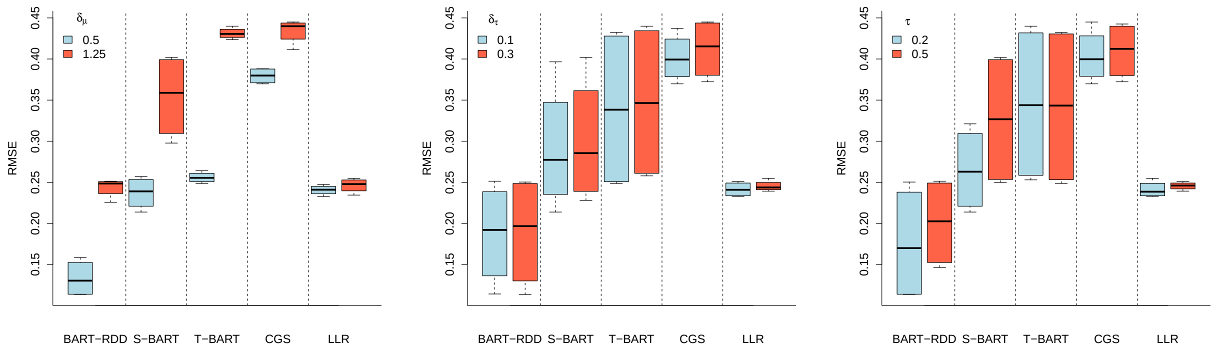
The results indicate that high variation in μ makes estimation harder for all estimators, although the difference is not so sizeable for LLR. In that setting, BART-RDD and LLR perform similarly. However, when δ_μ is low, BART-RDD clearly outperforms all estimators. Regarding the other BART-based estimators, S-BART and T-BART perform similarly, but

[5] In the case of the Bayesian estimators, we consider the posterior mean as the point estimate.

the former is less sensitive to high variability in μ. Overall, CGS is the worst performer in terms of the RMSE.

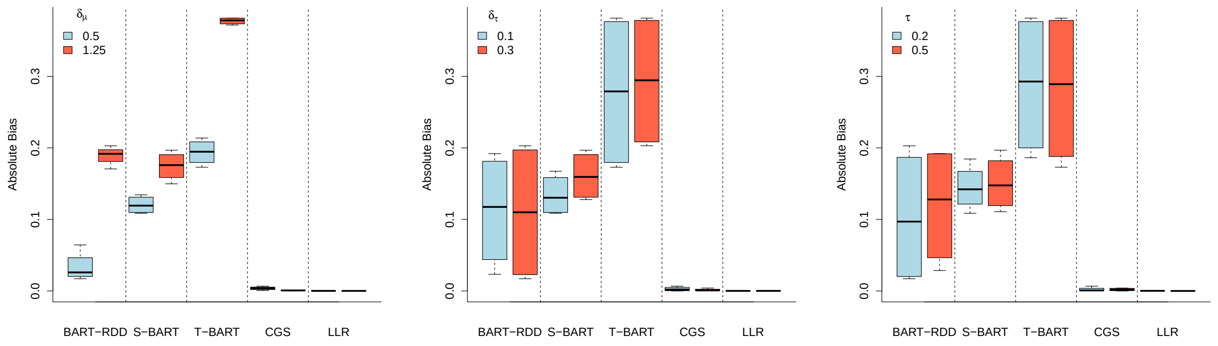
In order to better understand the behavior of the RMSE, figures 3 and 4 present, respectively, the absolute bias and variance for each estimator, separated by the parameter values of the DGPs. This decomposition highlights some important patterns. First, the consistently low bias of the LLR and CGS estimators is remarkable, which means any variation in their RMSE is coming from the estimator variance. For LLR, this should not come as a surprise given this method's focus on reducing bias, but it is worth noting how effective it is in that regard. On the contrary, BART-RDD presents bias comparable to LLR and CGS when heterogeneity in μ is low, but a much greater bias otherwise. This trend is true for all BART-based models, although, for a given value of δ_μ, BART-RDD almost always presents much lower — and never far worse — bias than the others. Finally, δ_μ is the only factor that significantly affects bias for the BART-based models. These results corroborate the bias described in equation (13) for tree-based RDD estimators. In particular, although both variation in μ and τ near the cutoff point can pose problems, the models are potentially much more sensitive to the former, since they require low variation for the prognostic function at both sides of the cutoff. The results also highlight how BART-RDD is particularly effective in decreasing the off-the-shelf BART sensititivity to such issues

Regarding variance, BART-RDD is always the best performer, with a consistently lower variance than the other estimators. LLR presents much larger variance than BART-RDD. T-BART presents a slightly larger variance than BART-RDD, whereas S-BART presents larger variance that is very sensitive to δ_μ. Finally, CGS presents the worst variance in all scenarios, which explains this method's poor RMSE performance.



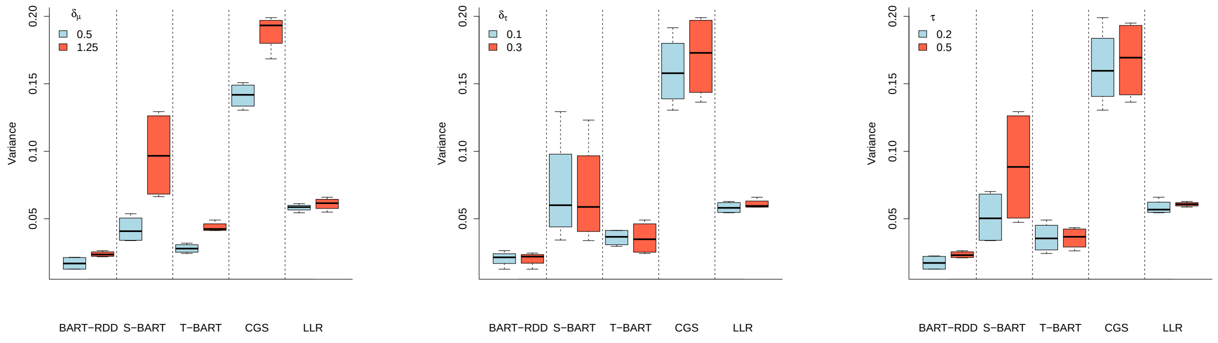
(a) δ_μ (b) δ_τ (c) τ

Figure 2: RMSE for the ATE point estimate produced by each estimator, divided by the different DGP parameters



(a) δ_μ (b) δ_τ (c) τ

Figure 3: Absolute bias for the ATE point estimate produced by each estimator, divided by the different DGP parameters



(a) δ_μ (b) δ_τ (c) τ

Figure 4: Variance for the ATE point estimate produced by each estimator, divided by the different DGP parameters

Although Bayesian intervals are generally not expected to achieve any particular coverage rate, frequentist coverage is a helpful metric to consider. Table 2 presents the coverage rate and interval size (in parenthesis) for each estimator. CGS and LLR present near 95% coverage in all cases, while S-BART presents near 95% coverage in all cases when $\bar{\tau} = 0.2$ and near 90% coverage when $\bar{\tau} = 0.5$. Coverage rates for BART-RDD and T-BART follow a common pattern of decreasing coverage when δ_μ increases. However BART-RDD presents better coverage than T-BART, reaching 95% coverage in one case and never falling below 70% coverage. Meanwhile, T-BART reaches at most 82.8% coverage.

Comparing the interval sizes provides an explanation of the coverage rate behavior. CGS presents, by far, the largest intervals so it is still able to get good coverage despite being the worst in terms of the RMSE. S-BART produces the second largest intervals on average, which also helps compensate the larger bias in some cases, leading to very good coverage generally. LLR produces the third largest intervals on average, which, combined with the relatively good RMSE performance leads to great coverage. T-BART comes next, and the combination of shorter intervals and bad RMSE performance leads to poor coverage.

Finally, BART-RDD produces the shortest intervals. However, because of the really good RMSE performance, it is still able to obtain good coverage in all cases.

Table 2: Coverage rate and interval sizes (in parenthesis) for the ATE

τ	δ_μ	δ_τ	BART-RDD	S-BART	T-BART	CGS	LLR
0.2	0.5	0.1	0.924	0.954	0.798	0.962	0.94
			(0.424)	(0.713)	(0.719)	(1.598)	(0.855)
0.2	0.5	0.3	0.954	0.95	0.724	0.95	0.932
			(0.442)	(0.757)	(0.677)	(1.604)	(0.877)
0.2	1.25	0.1	0.782	0.94	0.538	0.97	0.93
			(0.546)	(0.97)	(0.797)	(1.792)	(0.863)
0.2	1.25	0.3	0.718	0.95	0.52	0.964	0.938
			(0.539)	(1.068)	(0.794)	(1.814)	(0.88)
0.5	0.5	0.1	0.9	0.866	0.828	0.958	0.946
			(0.536)	(0.859)	(0.743)	(1.604)	(0.87)
0.5	0.5	0.3	0.92	0.89	0.772	0.962	0.942
			(0.519)	(0.913)	(0.704)	(1.607)	(0.87)
0.5	1.25	0.1	0.722	0.87	0.558	0.966	0.918
			(0.579)	(1.239)	(0.81)	(1.788)	(0.87)
0.5	1.25	0.3	0.702	0.894	0.572	0.962	0.934
			(0.567)	(1.313)	(0.818)	(1.798)	(0.872)

4.2.2 Comparison of CATE estimates

This section compares the various BART-based models in terms of their CATE estimation (the polynomial estimators do not provide CATE estimates). Table 3 presents the RMSE and coverage for each estimator. The results for the RMSE are qualitatively the same as before for all estimators. Regarding coverage, BART-RDD is the best model, with S-BART and T-BART performing slightly worse. Overall, these results suggest a similar trend as with the ATE: S-BART and T-BART present similar performance, with the latter being more sensitive to variability in μ. BART-RDD comes out as the best estimator among the BART variants in all scenarios but one.

For a more detailed look into the CATE predictions of each model, figures 5 and 6 present the CATE fit for an illustrative sample of the DGP described earlier, with $\delta_\mu = 0.5$ and $\delta_\mu = 1.25$ respectively. We set $\delta_\tau = 0.3$ and $\bar{\tau} = 0.5$ for these examples. The values are presented for units inside the identification strip in ascending order. Two patterns stand

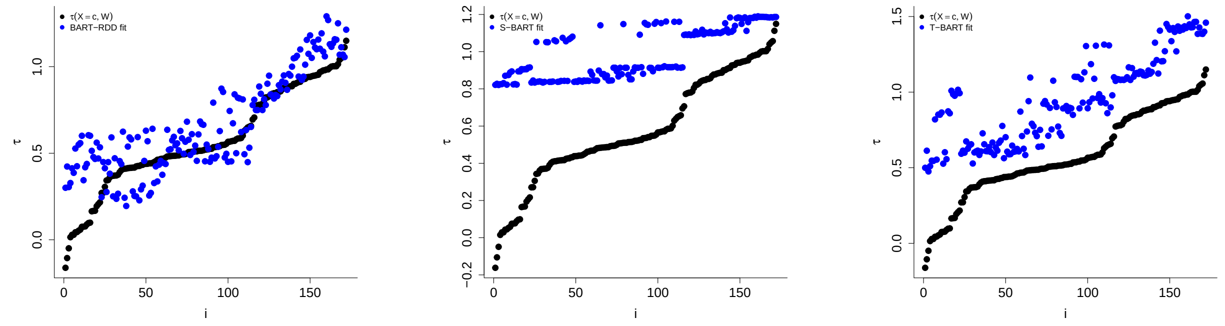
Table 3: RMSE and coverage - CATE

(a) RMSE

τ	δ_μ	δ_τ	BART-RDD	S-BART	T-BART
0.2	0.5	0.1	0.164	0.204	0.280
0.2	0.5	0.3	0.216	0.287	0.298
0.2	1.25	0.1	0.262	0.255	0.445
0.2	1.25	0.3	0.302	0.345	0.463
0.5	0.5	0.1	0.228	0.247	0.281
0.5	0.5	0.3	0.249	0.297	0.295
0.5	1.25	0.1	0.315	0.363	0.451
0.5	1.25	0.3	0.321	0.411	0.452

(b) Coverage

τ	δ_μ	δ_τ	BART-RDD	S-BART	T-BART
0.2	0.5	0.1	0.993	0.969	0.951
0.2	0.5	0.3	0.986	0.904	0.936
0.2	1.25	0.1	0.985	0.949	0.828
0.2	1.25	0.3	0.974	0.897	0.816
0.5	0.5	0.1	0.986	0.919	0.955
0.5	0.5	0.3	0.985	0.933	0.941
0.5	1.25	0.1	0.980	0.909	0.820
0.5	1.25	0.3	0.982	0.922	0.835

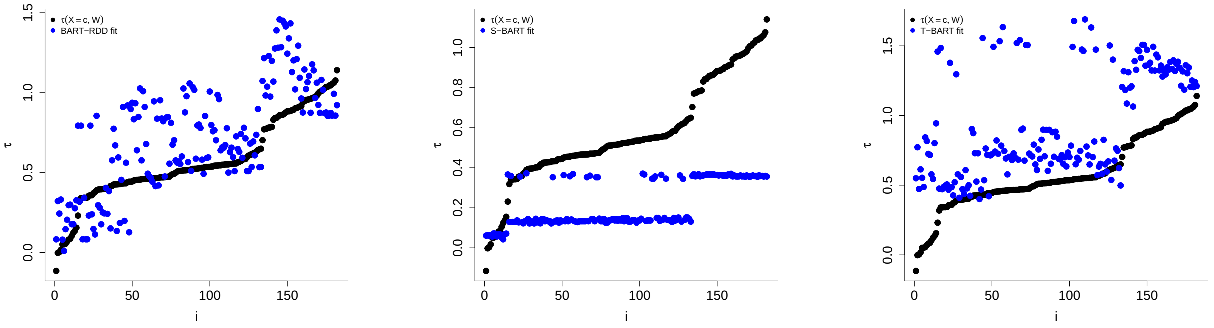


(a) BART-RDD (b) S-BART (c) T-BART

Figure 5: Fit for $\tau(X = c, W)$ for each method when $\delta_\mu = 0.5$ versus the true function

out in these comparisons. First, although increasing δ_μ evidently makes it harder to recover τ in general, the results from BART-RDD are a lot less sensitive to these changes. Second, S-BART seems to have a lot more difficulties in picking up variations in W, producing much more constant CATE estimates than the other methods. Overall, the figures suggest that the BART-RDD CATE predictions are less biased and more able to capture heterogeneity than the unmodified BART models.

The results of this section show how the bias described in (13) for tree-based models shows up empirically. In particular, figures 5 and 6 show how high variation in μ leads to a strong bias in the CATE estimates for unmodified BART models, whereas BART-RDD produces CATE estimates that are much less sensitive to that. Figure 2a then shows that these features carry over to the ATE estimates produced by each model. Overall, the



(a) BART-RDD

(b) S-BART

(c) T-BART

Figure 6: Fit for $\tau(X = c, W)$ for each method when $\delta_\mu = 1$ versus the true function

evidence in this section is clear that BART-RDD is more suited to the RDD setup than its unmodified counterparts.

5 Academic Probation and Student Performance

We now conclude with a detailed empirical application of BART-RDD to illustrate its usage in a real data setting. The data analyzed in this section comes from Lindo et al. (2010) and consists of information on college students enrolled in a large Canadian university. Students who, by the end of each term, present GPA lower than a certain threshold (which differs between the three university campuses) are placed on academic probation and must improve their GPA in the next term and face threat of punishment if they fail to achieve this goal, which can range from 1-year to permanent suspension from the university.

Among the performance outcomes analyzed by Lindo et al. (2010), we focus on GPA in the term after a student is placed on probation (Y). Following the authors, we define the running variable (X) as the negative distance between a student's first-year GPA and the probation threshold, meaning students below the limit have a positive score and the cutoff is 0. Additional student features in the data include gender ('male'), age when student entered the university ('age_at_entry'), a dummy for being born in North America ('bpl_north_america'), attempted credits in the first year ('totcredits_year1'), dummies for which campus each student belongs to ('loc_campus' 1, 2 and 3), and the student's position in the distribution of high school grades of students entering the university in the same year as a measure of high school performance ('hsgrade_pct'). An analysis of some summary statistics of the data is presented in the appendix.

In order to determine the appropriate prior parameters for this sample, we perform the prior elicitation procedure described in section 3.4: fix X, W, generate 10 samples of the DGP described in the appendix, take a grid of candidate values for (N_{Omin}, N_{Opct}, h) and calculate the RMSE over the 10 synthetic samples for each combination in the grid[6]. We set the parameters at the values which yield the lowest RMSE: $(N_{Omin}, N_{Opct}, h) = (5, 0.6, 0.1)$. The full results of the procedure are presented in the appendix.

The CATE and ATE estimates are obtained as described in section 4. We present additional summary statistics for the prediction sample in the appendix as well. We generate 100 draws from the individual-level posterior distribution which, averaging over observations, lead to 100 draws from the ATE posterior distribution. Table 4 presents a summary of the ATE posterior. The distribution is centered at 0.14 with all the posterior mass above zero, indicating strong evidence for positive effects of the probation policy. The 95% credible interval suggests the average effect can be as low as 0.08 and as high as 0.217[7].

We now discuss heterogeneity in the BART-RDD posterior distribution. Figure 7 presents a regression tree fit to posterior point estimates of the individual effects as a

[6] Because of the sample size of over 40,000 points, we are able to explore the prior reasonably with as few as 10 synthetic samples; for data with smaller sample sizes, more synthetic samples might be necessary to clearly distinguish between the candidates.

[7] For comparison, in the supplemental appendix we present the ATE results for the other estimators analyzed in the simulation exercise.

Table 4: BART-RDD posterior summary for the ATE

Mean	SD	2.5%	97.5%	Median	Min	Max
0.140	0.036	0.080	0.217	0.140	0.068	0.253

summarization tool. The summary trees are fit for the full sample and per campus. High-school performance is flagged as an important moderator for the full sample. Looking into each campus separately reveals more heterogeneity. For students who performed poorly in high-school in campus 1, we see additional moderation by birth place and credits attempted in the first year. In campus 2, we can see that the effect for women is larger than for men among those who feature above the 31-st percentile of high-school grades. Finally, for campus 3, the most important moderators are gender, birth place and age at entry.

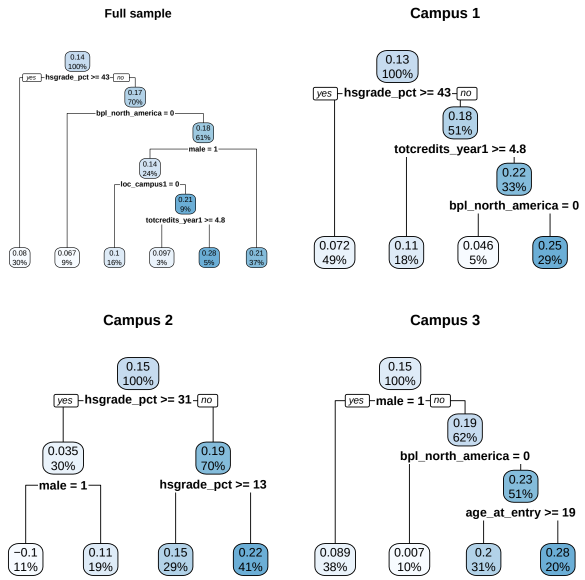


Figure 7: Regression tree fit to posterior point estimates of individual treatment effects: top number in each box is the average subgroup treatment effect, lower number shows the percentage of the total sample in that subgroup

The results described so far are consistent with those presented by Lindo et al. (2010), both in magnitude and, to some degree, in potential sources of treatment effect heterogeneity. In particular, the authors also find a greater effect for students who performed below average in high-school and for women. Our posterior predictions however flag additional features as potential moderators, such as age at entry, birth place and campus, which highlights how depending on pre-specification of relevant subgroups might lead researchers to miss other interesting features of the data. Lindo et al. (2010) note that interpreting these results as true effects requires caution since there is evidence that the probation policy leads to differential dropout rates, which changes the composition of students before and after the evaluation of first-year GPA. However, further discussion on this topic is out of the scope of this project.

We conclude this section with an illustration of how to perform posterior inference about heterogeneity in the effects with the results of our model. Based on the moderators flagged by the summarization trees, we investigate the posterior difference in treatment effects across some subgroups. The first panel in figure 8 presents the posterior difference between students in the bottom 43% versus those in the upper 57% of the high-school grade distribution for campus 1. There is a 92% posterior probability that the treatment effect is larger for the former group. The second panel presents a similar analysis for campus 2, where the threshold was the 31-st percentile of the high-school grade distribution. There is also strong evidence for a larger effect for students lower in that distribution, with a posterior probability of a larger effect of 95%. The third panel presents the posterior difference for students who entered college younger than 19 versus those who entered older than that in campus 3. There is also strong evidence of a larger effect for the former group, with posterior probability of 84%. Finally, the last panel presents the posterior difference in average effects between each campus. The biggest difference is between campus 3 and campus 1, in which case there is a 66% probability of a larger effect for the former. There is a 59% posterior probability that the effect is larger for campus 2 than campus 1 and a 54% posterior probability that the effect is larger for campus 3 than campus 2.

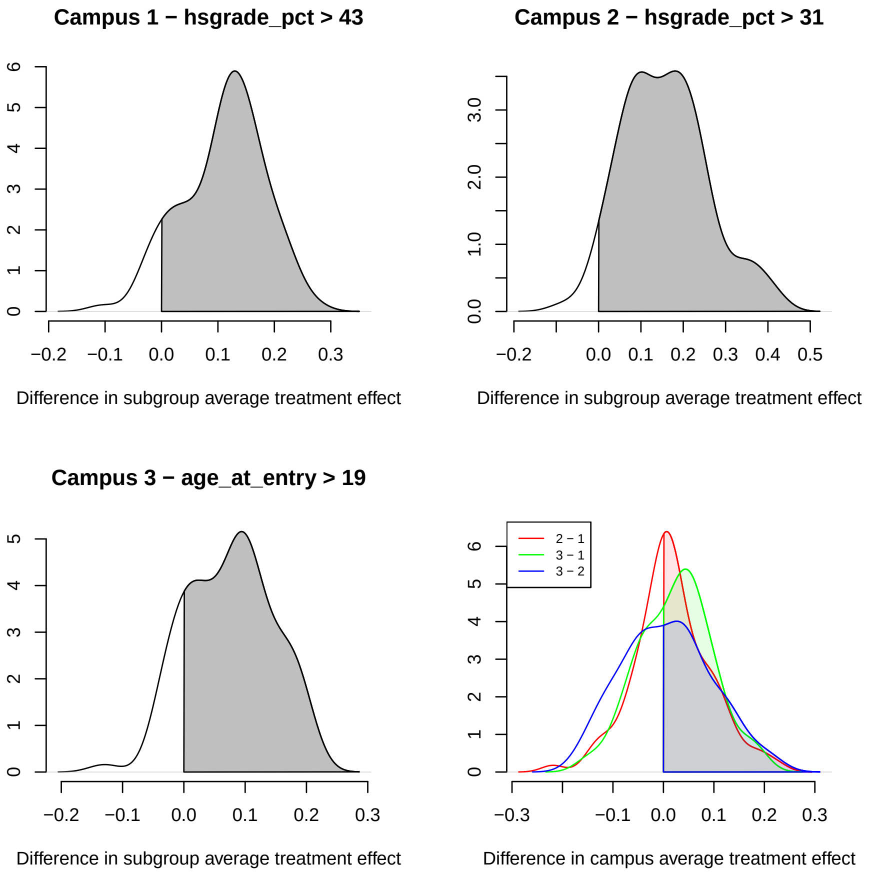


Figure 8: Differences in subgroup treatment effects: the first panel shows the posterior difference between students below and above the 43-rd percentile of high-school grades respectively in campus 1; the second panel performs the same analysis for the 31-st percentile of high-school grades for students in campus 2; the third panel presents the posterior difference between students that got into college younger versus older than 19 in campus 3; the last panel presents the posterior differences in the ATE between each campus

References

Arai, Y., Otsu, T., and Seo, M. H. (2024). Regression discontinuity design with potentially many covariates. *Econometric Theory.*

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.

Bertanha, M. (2020). Regression discontinuity design with many thresholds. *Journal of Econometrics*, 218(1):216–241.

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., Farrell, M. H., and Titiunik, R. (2019). Regression discontinuity designs using covariates. *Review of Economics and Statistics*, 101(3):442–451.

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.

Cattaneo, M. D., Titiunik, R., Vazquez-Bare, G., and Keele, L. (2016). Interpreting regression discontinuity designs with multiple cutoffs. *The Journal of Politics*, 78(4):1229–1248.

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., McCulloch, R. E., et al. (2010). Bart: Bayesian additive regression trees. *The Annals of Applied Statistics*, 4(1):266–298.

Frandsen, B. R., Frölich, M., and Melly, B. (2012). Quantile treatment effects in the regression discontinuity design. *Journal of Econometrics*, 168(2):382–395.

Frölich, M. and Huber, M. (2019). Including covariates in the regression discontinuity design. *Journal of Business & Economic Statistics*, 37(4):736–748.

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

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.

Hsu, Y.-C. and Shen, S. (2019). Testing treatment effect heterogeneity in regression discontinuity designs. *Journal of Econometrics*, 208(2):468–486.

Imbens, G. W. and Lemieux, T. (2008). Regression discontinuity designs: A guide to practice. *Journal of econometrics*, 142(2):615–635.

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.

Krantsevich, N., He, J., and Hahn, P. R. (2023). Stochastic tree ensembles for estimating heterogeneous effects. In *International Conference on Artificial Intelligence and Statistics*, pages 6120–6131. PMLR.

Kreiss, A. and Rothe, C. (2023). Inference in regression discontinuity designs with high-dimensional covariates. *The Econometrics Journal*, 26(2):105–123.

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.

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

Shen, S. and Zhang, X. (2016). Distributional tests for regression discontinuity: Theory and empirical examples. *Review of Economics and Statistics*, 98(4):685–700.

Starling, J. E., Murray, J. S., Lohr, P. A., Aiken, A. R., Carvalho, C. M., and Scott, J. G. (2021). Targeted smooth bayesian causal forests: An analysis of heterogeneous treatment effects for simultaneous vs. interval medical abortion regimens over gestation. *The Annals of Applied Statistics*, 15(3):1194–1219.

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.

Wang, M., He, J., and Hahn, P. R. (2024). Local gaussian process extrapolation for bart models with applications to causal inference. *Journal of Computational and Graphical Statistics*, 33(2):724–735.