A Constrained BART Model for Identifying Heterogeneous Treatment Effects in RDD

Rafael Alcantara[1] Meijia Wang[2]
P. Richard Hahn[2] Hedibert Lopes[1]-[2]

November 29, 2023

https://rafaelcalcantara.github.io

[1]Insper Institute of Education and Research
[2]SoMSS, Arizona State University

Outline

Contribution

- We propose a modification of the Bayesian Causal Forest model (Hahn et al., 2020) — itself an extension of the BART model of Chipman et al. (2010) — which uses a novel regression tree prior that incorporates the unique structure of regression discontinuity designs
- We show that unmodified BART and BCF models estimate RDD treatment effects poorly, while our modified model accurately recovers treatment effects at the cutoff
- At the same time, the model retains the inherent flexibility of all BART-based models, allowing it to effectively explore heterogeneous treatment effects
- We also show that heterogeneity poses a threat to the performance of the local polynomial estimator

Regression Discontinuity Designs - Motivation

Thistlethwaite and Campbell (1960): motivational effects of public recognition in a national scholarship competition in academic outcomes

Treatment: the Certificate of Merit, an award which is widely publicized among colleges, universities and other agencies

Assignment: score in a national exam

Confounding: latent student characteristics could make it more likely for the student to score higher and hence, more likely to receive the award, but also would lead to a higher likelihood of a student observing more positive academic gains

Regression Discontinuity Designs - Motivation

Solution: Because of the deterministic assignment rule, scores completely deconfound the data

Problem: Complete lack of overlap; impossible to construct causal contrasts without further assumptions

Fundamental RDD assumption: Introducing smoothness assumptions about the potential outcomes distribution near the cutoff

Regression Discontinuity Designs

Let Z be a binary treatment variable and X be a variable defining the treatment assignment, *i.e.* X is the the running variable:

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

for some cutoff value c.

RDD - potential outcomes

Let $Y_i(z_i)$ denote the potential outcome when $Z_i = z_i$. We observe only

$$Y_i = Y_i(1)Z_i + Y_i(0)(1 - Z_i). \tag{1}$$

We focus on the difference in expected potential outcomes:

$$\tau_S := \mathbb{E}[Y_i(Z_i = 1) \mid X_i = c, \mathrm{w}_i] - \mathbb{E}[Y_i(Z_i = 0) \mid X_i = c, \mathrm{w}_i] \tag{2}$$

Under the assumption that the distribution of Y_i is smooth in X_i, at least at $X = c$, the treatment effect may be estimated as a limit:

$$\tau_S = \lim_{x \downarrow c} \mathbb{E}[Y_i | X_i = x, \mathrm{w}_i] - \lim_{x \uparrow c} \mathbb{E}[Y_i | X_i = x, \mathrm{w}_i].$$

An illustration

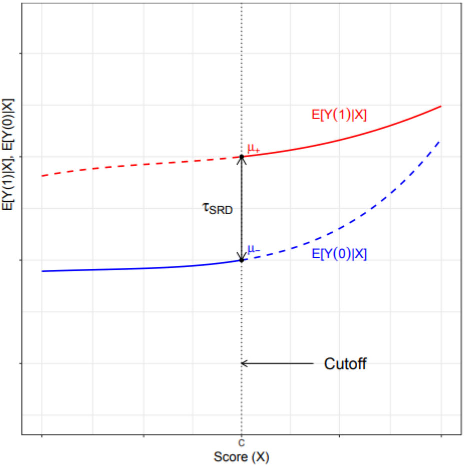


Figure 1: RDD Example

RDD

The treatment effect can be estimated by estimating conditional expectation *functions* $\mathbb{E}[Y_i|X_i, \mathrm{w}_i]$, both above and below the cutoff and taking a difference at the point $X = c$.

The most common estimation strategy is to perform a local polynomial regression of Y on X with a bandwidth choice that asymptotically minimizes the mean-squared error (MSE) of the predictions (Hahn et al., 2001; Imbens and Kalyanaraman, 2012).

Controlling for covariates can increase precision in the estimation and make the continuity assumption more credible (Calonico et al., 2019).

CATE

Estimation of conditional average treatment effects (CATE) from RDD data is a bit more subtle, as interacting many covariates with the running variable quickly leads to high-variance estimators.

Our contribution: In this respect, Bayesian regression trees, which incorporate interactions in a data-driven but regularized way, are a natural framework to pursue.

Basic BART

- Bayesian "sum-of-trees" model where each tree is constrained by a regularization prior to be a weak learner, and fitting and inference are accomplished via an iterative Bayesian backfitting MCMC algorithm that generates samples from a posterior.
- BART is a nonparametric Bayesian regression approach which uses dimensionally adaptive random basis elements.
- Motivated by ensemble methods in general, and boosting algorithms in particular, BART is defined by a statistical model: a prior and a likelihood.
- By keeping track of predictor inclusion frequencies, BART can also be used for model-free variable selection.

Nonlinear regression

We want to “fit” the fundamental model:

$$y_i = g(x_i; \theta) + \epsilon_i$$

BART is a Markov Monte Carlo Method that draws from

$$g(x; \theta) | (x, y)$$

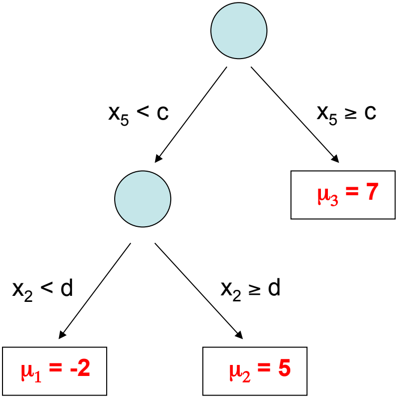
We can then use the draws as our inference for $g(x; \theta)$.

A regression tree model

Let T denote the tree structure including the decision rules.

Let $M = \{\mu_1, \mu_2, \ldots, \mu_b\}$ denote the set of bottom node μ's.

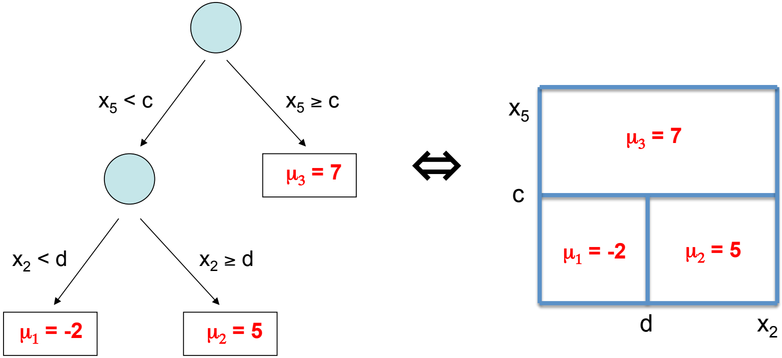
Let $g(x; \theta)$, $\theta = (T, M)$ be a regression tree function that assigns a μ value to x.



A single tree model:

$$y_i = g(x_i; \theta) + \epsilon_i.$$

A coordinate view of $g(x;\theta)$



Easy to see that $g(x;\theta)$ is just a step function.

Turning the Bayesian crank

To get the draws, we will have to:

- Put a prior on $g(x;\theta)$.
- Specify a Markov chain whose stationary distribution is

$$p\left(g(x;\theta)|(x,y)\right).$$

Ensemble methods

Various methods which combine a set of tree models, so called ensemble methods, have attracted much attention, each of which use different techniques to fit a linear combination of trees.

- Bagging (Breiman, 1996)
- Random forests (Breiman, 2001)
- Boosting (Friedman, 2001)
- Bayesian model averaging (Chipman, George and McCulloch, 1998)

Bagging and random forests use randomization to create a large number of independent trees, and then reduce prediction variance by averaging predictions across the trees. Boosting fits a sequence of single trees, using each tree to fit data variation not explained by earlier trees in the sequence.

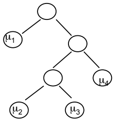
Bayesian model averaging (BMA) applied to the posterior arising from a Bayesian single-tree model.

Key references

1. Breiman (1996) Bagging predictors
   Machine Learning, 26, 123-140.
2. Hastie and Tibshirani (2000) Bayesian Backfitting
   Statistical Science, 15(3), 196-223.
3. Friedman (2001) Greedy function approximation: A gradient boosting machine
   Annals of Statistics, 29, 1189-1232.
4. Breiman (2001) Random forests
   Machine Learning, 45, 5-32.
5. Chipman, George and McCulloch (1998) Bayesian CART model search
   Journal of the American Statistical Association, 93, 935-960.
6. Efron, Hastie, Johnstone and Tibshirani (2004) Least angle regression
   Annals of Statistics, 32, 407-499.

The BART model

$$Y = g(x;T_1,M_1) + g(x;T_2,M_2) + \ldots + g(x;T_m,M_m) + \sigma z, \quad z \sim N(0,1)$$



$m = 200, 1000, \ldots, \text{big}, \ldots$

$f(x|\cdot)$ is the sum of all the corresponding μ's at each bottom node.

Such a model combines additive and interaction effects.

Complete the model with a regularization prior

The prior of the BART model can be written as

$$\pi(\theta) = \pi((T_1, M_1), (T_2, M_2), \ldots, (T_m, M_m), \sigma).$$

π wants:

- Each T small.
- Each μ small.
- "nice" σ (smaller than least squares estimate).

We refer to π as a regularization prior because it keeps the overall fit small.

In addition, it keeps the contribution of each $g(x; T_i, M_i)$ model component small.

BART MCMC

The model/prior is described by

$$Y = g(x;T_1,M_1) + \ldots + g(x;T_m,M_m) + \sigma\, z$$
plus
$$\pi((T_1,M_1),\ldots.(T_m,M_m),\sigma)$$

First, it is a "simple" Gibbs sampler:

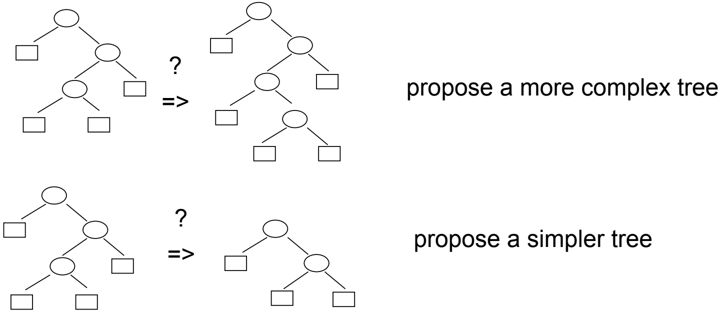
$$\begin{array}{rcl} (T_i, M_i) & | & (T_1, M_1, \ldots, T_{i-1}, M_{i-1}, T_{i+1}, M_{i+1}, \ldots, T_m, M_m, \sigma) \\ \sigma & | & (T_1, M_1, \ldots, \ldots, T_m, M_m) \end{array}$$

To draw $(T_i, M_i)|\cdot$ we subract the contributions of the other trees from both sides to get a simple one-tree model.

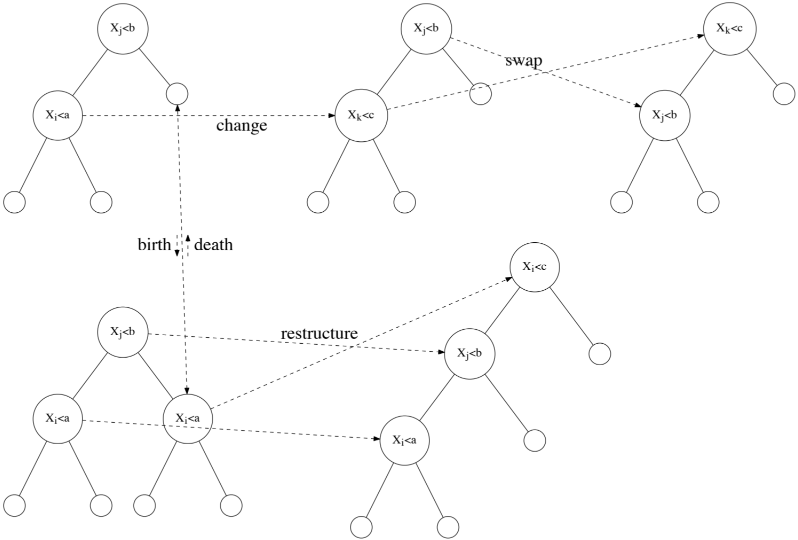
We integrate out M to draw T and then draw $M|T$.

Birth-death moves

To draw T we use a Metropolis-Hastings with Gibbs step.
We use various moves, but the key is a "birth-death" step.

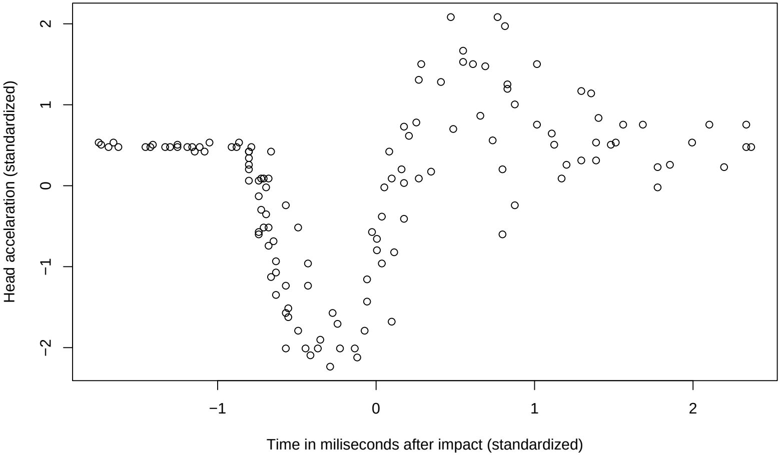


Tree moves



X_j<b
X_i<a
change
X_j<b
X_k<c
swap
X_k<c
X_j<b
birth
death
X_j<b
X_i<a
X_i<a
restructure
X_i<c
X_j<b
X_i<a

motorcycle dataset



Smooth spline

The goal is to find $g(\cdot)$ that minimizes

$$\sum_{i=1}^{n}(y_i - g(x_i))^2 + \lambda \int g''(t)^2 dt$$

for tuning parameter $\lambda > 0$.

The basis functions for a global cubic polynomial are $B_i(x) = x^{i-1}$ for $i = 1, 2, 3, 4$, so

$$g(x) = \sum_{j=1}^{4} \beta_j B_j(x)$$

Splines are piecewise cubic polynomials: $B_1(x) = 1$, $B_2(x) = x$ and

$$B_{2+i}(x) = \frac{(x - x_i)_+^3 - (x - x_n)_+^3}{x_n - x_i} - \frac{(x - x_{n-1})_+^3 - (x - x_n)_+^3}{x_n - x_{n-1}}$$

R code

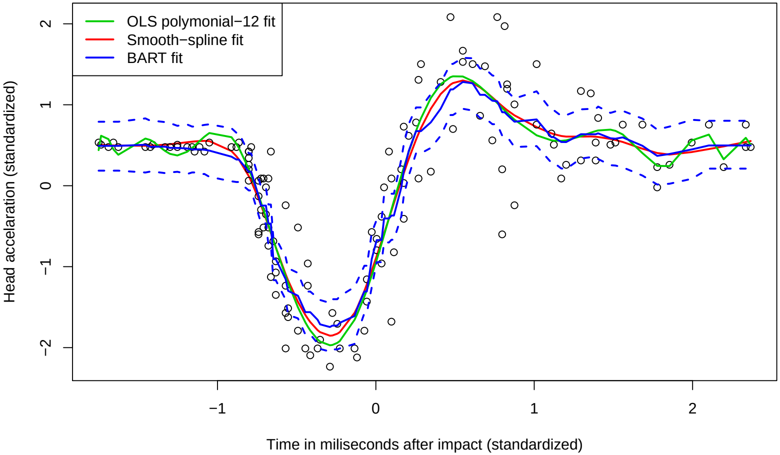
```
install.packages("BART")
library(MASS)
library(BART)
xt = mcycle$times[1:132]
yt = mcycle$accel[1:132]
xt = (xt-mean(xt))/sqrt(var(xt))
yt = (yt-mean(yt))/sqrt(var(yt))

d=12
xx = NULL
for (i in 1:d)
  xx = as.matrix(cbind(xx,xt^i))
xx = (xx - matrix(apply(xx,2,mean),n,d,byrow=TRUE))%*%diag(sqrt(1/apply(xx,2,va

# OLS, smooth spline and BART fits
linear.fit = lm(yt~xx-1)
fit = smooth.spline(xt,yt)
bart.fit   = wbart(xt,yt)
bart.q     = t(apply(bart.fit$yhat.train,2,quantile,c(0.05,0.5,0.95)))

plot(fit,xlab="Time in miliseconds after impact (standardized)",
     ylab="Head accelaration (standardized)",type="l",lwd=2,col=2,
     xlim=range(xt),ylim=range(yt))
points(xt,yt)
lines(xt,linear.fit$fit,col=3,lwd=2)
```

lm, smooth.spline and wbart in action



References

1. Chipman, George and McCulloch (2010) BART: Bayesian Additive Regression Trees. *The Annals of Applied Statistics*, 4(1), 266-298.

2. Taddy, Gramacy and Polson (2011) Dynamic Trees for Learning and Design. *Journal of the American Statistical Association*, 106(493), 109-123.

3. Pratola, Chipman, Higdon, McCulloch and Rust (2014) Parallel BART. *Journal of Computational and Graphical Statistics*, 23, 830-852.

4. Lakshminarayanan, Roy and Teh (2015) Particle Gibbs for BART. *Proceedings of the 18th Conference on Artificial Intelligence and Statistics*.

5. Kapelner and Bleich (2016) `bartMachine`: machine learning with BART. *Journal of Statistical Software*, 70(4).

6. Pratola (2016) Efficient Metropolis-Hastings Proposal Mechanisms for BART models. *Bayesian Analysis*, 11(3), 885-911.

7. Hernández, Raftery, Pennington and Parnell (2017) BART using BMA. *Statistics and Computing*.

8. Linero (2017) Bayesian Regression Trees for High Dimensional Prediction and Variable Selection. *Journal of the American Statistical Association*.

9. Pratola, Chipman, George and McCulloch (2017) Heteroscedastic BART Using Multiplicative Regression Trees.

Bayesian Causal Forest (BCF)

BART for causal inference:

S-learners: BART with treatment as covariate (Hill, 2011).

T-learners: Separate BART models for treated and untreated units

Problems:

S-learner: degree of regularization depends on the joint distribution of the control variables and the treatment variable.

T-learner: regularization of the treatment effect is necessarily weaker than regularization of each individual model.

BCF

Bayesian Causal Forest (BCF) model (Hahn et al., 2020): fits two BART models simultaneously to a reparametrized response function:

$$Y_i = \mu(X_i, \mathrm{w}_i) + \tau(X_i, \mathrm{w}_i) b_{z_i} + \varepsilon_i, \quad \varepsilon_i \sim N(0, \sigma^2), \tag{3}$$

where $b_0 \sim \mathrm{N}(0, 1/2)$ and $b_1 \sim \mathrm{N}(0, 1/2)$.

$\mu(\cdot)$ is a prognostic function and $\tau(\cdot)$ a treatment effect function.

The ATE can be expressed as

$$\mathbb{E}(Y^1 \mid X = x) - \mathbb{E}(Y^0 \mid X = x) = (b_1 - b_0)\tau(x). \tag{4}$$

XBART and XBCF

BART MCMC algorithm is very inefficient

Accelerated Bayesian additive regression trees (XBART) algorithm (He and Hahn, 2021): grows new trees recursively, but stochastically, at each step

Accelerated Bayesian causal forest (XBCF) algorithm (Krantsevich et al., 2023): adaptation of XBART to the reparametrized model of BCF

Our method consists of an adaptation of the XBCF algorithm to the RDD setting.

XBCF

The new model is almost the same as (3) except that XBCF allows the error variance to change for each treatment status:

$$\begin{aligned} Y_i &= a\mu(x_i) + b_{z_i}\tilde{\tau}(x_i) + \epsilon_i, \quad \epsilon_i \sim N(0, \sigma^2_{z_i}) \\ a &\sim \mathrm{N}(0,1), \quad b_0, b_1 \sim \mathrm{N}(0, 1/2), \end{aligned} \tag{5}$$

where $\mu(x)$ and $\tilde{\tau}(x)$ are two XBART forests and $\tau = (b_1 - b_0)\tilde{\tau}$.

The key innovation from He and Hahn (2021) is the so-called "Grow-From-Root" stochastic tree-fitting algorithm, which we adapt to the RDD context.

BART-RDD

Ensure that the data used to make predictions at $X = c$ warrant a causal interpretation:

- $\mu(x = c, w)$ and $\tau(x = c, w)$ must be composed of trees where any partition containing the point $(x = c, w)$ has a corresponding function evaluation that has been estimated from causally valid contrasts

Assuming continuous conditional expectations, this is possible if the estimation is based on data close enough to the cutoff.

The BART-RDD model developed here satisfies this criterion by explicitly imposing it during the tree growing process.

BART-RDD: Splitting Constraints

We define an 'identification strip' around the cutoff, ($[c - h, c + h]$), such that:

- Any node which does not contain that region remains entirely unrestricted
- Any node that *does* contain it has to have both:
  1. A minimum number of observations within the region on either side of the cutoff; and
  2. Not too many observations, proportionally, outside of the identification strip

Splitting Constraints

More formally, these constraints can be expressed as follows:

- Define a bandwidth parameter $h > 0$
- Assume that the potential outcome mean function does not vary abruptly inside the interval $[c - h, c + h]$
- 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
- Let n_l denote the number of observations in $B \cap [c - h, c)$ and n_r denote the number of observations in $B \cap [c, c + h]$

Splitting Constraints

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:

$$A \cup (C \cap D \cap E)$$

where

$$\begin{aligned}
A &= \big(\forall w \mid (x = c, w) \notin B\big) \\
C &= (\exists w \mid (x = c, w) \in B) \\
D &= (\min(n_l, n_r) \geq N_{Omin}) \\
E &= ((n_l + n_r)/N_b \geq \alpha)
\end{aligned}$$

Splitting constraints

A split that violates condition E can be satisfied by further branching, ‘trimming’ observations from outside the strip

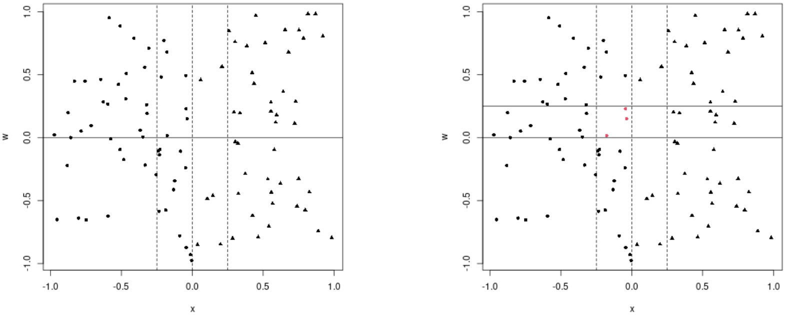
A split that violates condition D can never be satisfied by further branching

We set the likelihood of nodes that violate condition D to zero and force partitions that violate condition E to split until condition E is not violated anymore

Illustration

- Suppose there is only one additional covariate W besides the running variable X, and $X, W \overset{\text{iid}}{\sim} U(-1, 1)$
- Figure 2 presents different possible partitions of a dataset with 100 observations under this DGP
- For this example, we considered $h = 0.25$ – denoted by the dashed lines in the plots – and set $c = 0$ – denoted by the dotted line
- The treated units ($x \geq c$) are denoted by triangle dots and the control units are denoted in round dots

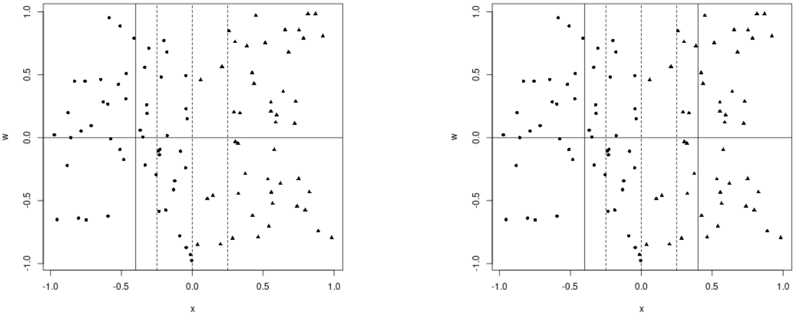
Illustration



(a) (b)

Figure 2: Tree examples

Illustration



(a)

(b)

Figure 3: Tree examples

Illustration

- Panel 2a presents an initial split at $w = 0$
- This partition is not valid because condition E is violated: both nodes contain the identification strip, but are highly populated by points outside of it
- However, condition E is not violated because both nodes feature at least one point inside the identification strip from both sides of the cutoff
- Therefore, our algorithm forces the tree to keep splitting instead of outright rejecting the split

Illustration

- Panel 2b presents a second split in W
- This split leads to a partition where one of the nodes features data inside the left side of the identification strip region but not from the right side (such points are highlighted), violating condition D for any N_{Omin}
- In this instance, the algorithm rejects that split by attributing a likelihood of 0 to it

Illustration

- Panel 3a starts with the same split at $W = 0$ as before and then considers an additional split at $X = -0.4$ for both regions $W < 0$ and $W \geq 0$, leading to a tree with four nodes
- First, note that the nodes to the left of $X = -0.4$ are unrestricted since they do not include the identification strip
- For the other two nodes, condition D is not violated, but condition E is
- In this instance, the algorithm would accept the splits and force the tree to continue splitting until condition E is also met

Illustration

- Finally, panel 3b presents the same partition as 3a with an additional split at $X = 0.4$ for both $W < 0$ and $W \geq 0$
- This partition does not violate any of the conditions, meaning these splits would not be rejected and the tree would not be forced to split (although it could keep splitting if the no-split condition is not chosen and there are still valid splits).

Illustration - Summary

- We consider only trees that do not cut through the identification strip, are well populated with points in that region from both sides of the cutoff and are tight around that region
- This way, we incorporate the RDD assumption that units sufficiently near the cutoff are similar enough to warrant a causal comparison and use this to create an 'overlap region' around the cutoff
- The shape of the trees is also largely dependent on the data structure. If there are many points with $x \approx c$ we can make the identification strip narrower without being too restrictive on the tree growth especially if the points are well dispersed in regards to the other covariates

Illustration - Summary

- On the contrary, if most points have x far from the cutoff we might need to define a wider identification strip to reasonably explore the tree space
- Finally, it is worth noting that this strategy can be used more generally for any problem where one must fit tree ensembles and enforce smoothness over a specific variable and around a specific point

Parameter settings

We add three new parameters to the BART prior: α, N_{Omin} and h

α shouldn't be set too low (*e.g.* below 0.5), otherwise points far from the cutoff could have a big impact in the estimation at that point

N_{Omin} shouldn't be set too low so that too few points are used to obtain the causal contrasts, and not too high so that nearly any split in W is rejected

Given such considerations, the prior is not very sensitive to these parameters; we recommend a default setting of $\alpha = 0.9$ and $N_{Omin} = 5$, but encourage sensitivity checks in any given sample

Parameter settings

Regarding h, a very tight window could have too few points to obtain good estimates, a very large window could lead to points too far from the cutoff affecting estimation

- **Problem:** the prior is highly sensitive to this parameter and there is no clear guide for what 'too high' or 'too low' means

We develop a prior elicitation heuristic to set h appropriately:

- For a given sample (y, x, w), construct a synthetic RDD model based on (x, w) to generate s samples of y_s
- For each sample, fit the model with a grid of candidate h values
- Calculate the RMSE for each candidate h in the syntehtic samples
- Choose the h value that yields the lowest RMSE

Parameter settings

In other words, we choose h by fine tuning BART to some prior model based on (x, w)

The question, of course, is how to construct this prior model: we suggest a polynomial on X with no heterogeneity, *i.e.* no dependence on W, and small treatment effects, as this is a reasonable and commonly used prior in causal inference settings

In our experiments, this procedure was able to find the 'optimal' region for h even in cases when the true data had strong heterogeneity or large effects

Parameter settings

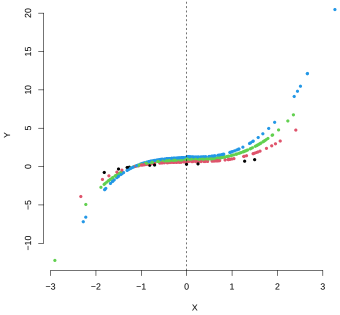
For the h candidates, basing those values on the standard deviation of X (σ_x) has led to the best results in our experiments (for the illustration we will present here, we considered $h \in \{\sigma_x/2, \sigma_x, 2\sigma_x\}$)

Prior exploration - Illustration

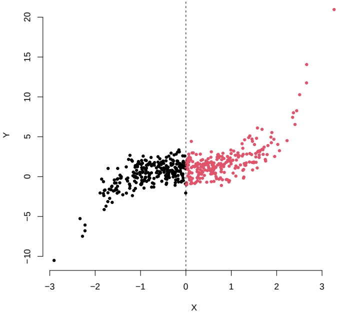
DGP (10 samples)

$$
\begin{aligned}
X &\sim N(0,1)\\
W &\sim B(3,0.7)+1\\
Z &= \mathrm{I}(X \geq 0)\\
\mu(X,W) &= 0.3W + 0.1WX + 0.1W(X+0.05)^2 + 0.2WX^3\\
\tau(X,W) &= 0.03W - 0.2WX + 0.05W(X+0.01)^2 - 0.1WX^3\\
\varepsilon &\sim N(0,1)\\
Y &= \mu(X,W) + \tau(X,W)Z + \varepsilon
\end{aligned}
\tag{6}
$$

Prior exploration - Illustration



(a) $E(Y)$



(b) Y

Prior exploration - Illustration

Prior model (11 samples)

$$
\begin{aligned}
\mu_p(X) &= 0.08 + 0.23X + 0.16X^2 \\
\tau_p(X) &= 0.01 + 0.24X + 0.035X^2 \\
\varepsilon_p &\sim N(0, 0.5^2) \\
Y_s &= \mu_p(X) + \tau_p(X)Z + \varepsilon_p
\end{aligned}
\tag{7}
$$

Prior exploration - Illustration

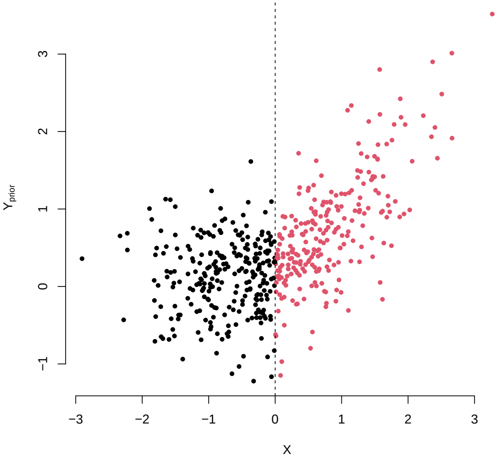


Figure 5: Prior predictive model

Prior exploration - Illustration

- For each of the 10 samples, the procedure selected $h = sd(x)/2$
- Using these values to estimate the "true" model, we obtain an **RMSE** of 0.067, **coverage** of 1 and **interval size** of 0.43

Simulations

Basic setup:

$$
\begin{aligned}
u &\sim U(0,1) \\
W_1 &\sim U(u, u+1) \\
W_2 &\sim U(0, 0.5) \\
W_3 &\sim B(2, u) + 1 \\
W_4 &\sim B(1, 0.6) + 1 \\
X &\sim 2 \times \text{Beta}(2,4) - u - 0.2 \\
Z &\sim \text{I}(X \geq 0) \\
\varepsilon &\sim N(0, \sigma^2) \\
Y_i &= \mu_i(X, W) + \tau_i(X, W)Z + 0.5u + \varepsilon
\end{aligned} \tag{8}
$$

Simulations

Define the following function of W_3 and W_4:

$$f_{34} = \begin{cases} 0.43 & \text{if } W_3 = 1 \cap W_4 = 1 \\ 0.27 & \text{if } W_3 = 2 \cap W_4 = 1 \\ 0.1 & \text{if } W_3 = 3 \cap W_4 = 1 \\ 0.77 & \text{if } W_3 = 1 \cap W_4 = 2 \\ 0.93 & \text{if } W_3 = 2 \cap W_4 = 2 \\ 1.1 & \text{if } W_3 = 3 \cap W_4 = 2 \end{cases} \tag{9}$$

Simulations

Prognostic functions:

$$\begin{aligned}
\mu_1(X,W) &= 0.1875\sin((W_1+W_2)\pi) + 1 + 1.875X \\
&\quad - 1.25X^2 + 1.75X^3 \\
\mu_2(X,W) &= 0.1W_4\sin((W_1+W_2)\pi) + 1 + 0.9W_4 + W_4X \\
&\quad - 0.9W_4X^2 + W_4X^3 \\
\mu_3(X,W) &= 0.2f_{34}\sin((W_1+W_2)\pi) + 1 + 2f_{34} + 2.27f_{34}X \\
&\quad - 1.13f_{34}X^2 + 2f_{34}X^3
\end{aligned} \tag{10}$$

Simulations

Treatment effect functions:

$$\begin{aligned}\tau_1(X, W) &= 0.025\cos((W_1 + W_2)\pi) + 0.05 - 2.8X \\ &\quad + 1.4X^2 - 0.14X^3 \\ \tau_2(X, W) &= 0.0125W_4\cos((W_1 + W_2)\pi) + 0.03 + 0.03W_4 \\ &\quad - 1.8W_4X + 0.9W_4X^2 - 0.09W_4X^3 \\ \tau_3(X, W) &= 0.05f_{34}\cos((W_1 + W_2)\pi) + 0.03 + 0.06f_{34} \\ &\quad - 3.4f_{34}X + 1.7f_{34}X^2 - 0.17f_{34}X^3\end{aligned} \tag{11}$$

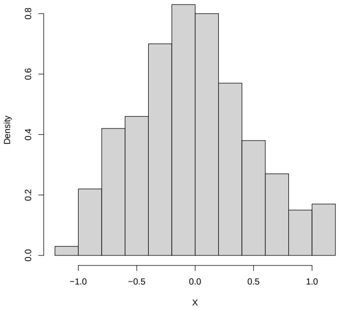
Simulations

Finally, we consider three different noise levels:

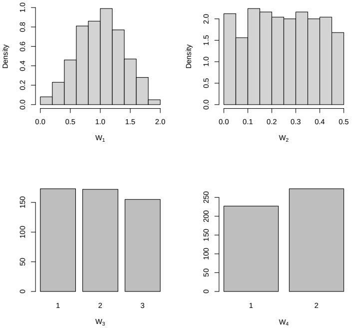
$$\sigma \in \{0.25, 1, 4\} \tag{12}$$

These scenarios showcase small, mild and strong heterogeneity plus low, mild and high signal-to-noise ratio

Simulation Data



(a) X



(b) W

Simulation Data

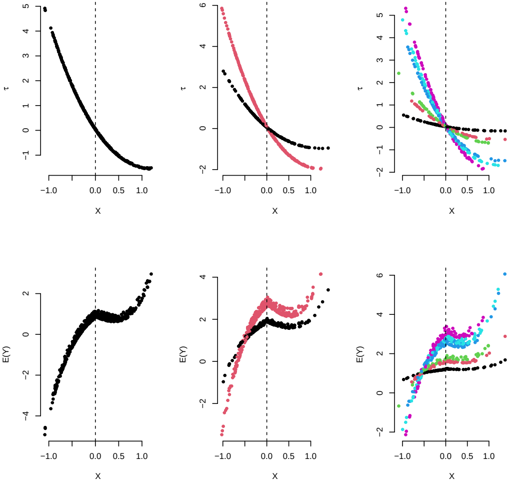


Figure 7: Prognostic and treatment functions

Simulation Data

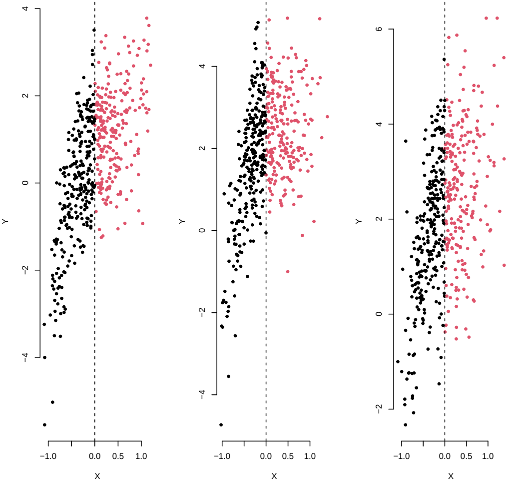


Figure 8: Y

Simulation - Estimators

BART-based models:

- BART-RDD
- S-learner BART (S-BART)
- T-Learner BART (T-BART)
- BCF

Non-BART models:

- Calonico et al. (2019) (CKT) - local polynomial regression

Estimators are compared in terms of RMSE, bias, variance, coverage and interval size

Simulation Results - ATE (Low Noise)

		BART-RDD	BCF	S-BART	T-BART	CKT
	RMSE	0.045	0.071	0.056	0.084	0.102
	Bias	0.026	0.013	-0.011	0.024	0.003
(1)	Variance	0.001	0.005	0.003	0.007	0.010
	Coverage	0.961	0.921	0.937	0.962	0.939
	Size	0.170	0.232	0.256	0.369	0.381
	RMSE	0.041	0.080	0.064	0.092	0.109
	Bias	-0.029	0.011	-0.031	0.013	0.005
(2)	Variance	0.001	0.006	0.003	0.008	0.012
	Coverage	0.930	0.898	0.897	0.941	0.934
	Size	0.150	0.274	0.257	0.367	0.381
	RMSE	0.038	0.076	0.058	0.103	0.137
	Bias	0.011	0.039	-0.017	0.027	0.010
(3)	Variance	0.001	0.004	0.003	0.010	0.019
	Coverage	0.979	0.869	0.945	0.885	0.927
	Size	0.174	0.236	0.266	0.347	0.496

Simulation Results - ATE (Mild Noise)

		BART-RDD	BCF	S-BART	T-BART	CKT
	RMSE	0.073	0.170	0.143	0.210	0.354
	Bias	0.014	0.093	0.030	0.008	0.009
(1)	Variance	0.005	0.020	0.019	0.044	0.125
	Coverage	0.985	0.907	0.996	0.970	0.940
	Size	0.359	0.537	0.775	0.972	1.330
	RMSE	0.091	0.129	0.147	0.221	0.373
	Bias	-0.049	-0.009	-0.012	-0.001	0.009
(2)	Variance	0.006	0.016	0.021	0.049	0.139
	Coverage	0.921	0.962	0.992	0.974	0.940
	Size	0.356	0.525	0.778	1.004	1.330
	RMSE	0.064	0.174	0.149	0.227	0.381
	Bias	0.001	0.101	0.022	0.052	0.019
(3)	Variance	0.004	0.020	0.022	0.049	0.145
	Coverage	0.986	0.884	0.991	0.955	0.934
	Size	0.335	0.533	0.806	0.989	1.371

Simulation Results - ATE (High Noise)

		BART-RDD	BCF	S-BART	T-BART	CKT
	RMSE	0.225	0.627	0.351	0.549	1.397
	Bias	-0.019	0.438	0.176	0.114	0.024
(1)	Variance	0.050	0.202	0.092	0.288	1.953
	Coverage	0.989	0.868	0.997	0.980	0.940
	Size	1.244	1.767	1.890	2.665	5.257
	RMSE	0.231	0.486	0.317	0.585	1.460
	Bias	-0.082	0.261	0.077	0.051	0.017
(2)	Variance	0.047	0.168	0.095	0.340	2.135
	Coverage	0.984	0.923	0.998	0.980	0.942
	Size	1.198	1.727	1.950	2.849	5.257
	RMSE	0.198	0.617	0.340	0.600	1.479
	Bias	-0.019	0.454	0.124	0.132	0.056
(3)	Variance	0.039	0.175	0.100	0.343	2.185
	Coverage	0.997	0.871	0.993	0.973	0.932
	Size	1.153	1.732	1.987	2.860	5.271

Simulation exercise

$$
\begin{aligned}
X &\sim 2 \times \text{Beta}(2,4) - 1 \\
W_p &\sim N(0, 0.25^2), \quad p \in \{1,2\} \\
W_p &\sim N\left(\frac{p-1}{p} X, 1\right), \quad p \in \{3,4\} \\
W_5 &\sim \text{Bernoulli}(0.7) \\
Z &= 1(X \geq 0) \\
\varepsilon &\sim N(0,1) \\
\sigma_\mu &= \sqrt{V[\mu_m(0, W)]} \\
\sigma_\tau &= \sqrt{V[\tau_m(0, W)]} \\
\bar{\tau} &= E[\tau_m(0, W)] \\
Y &= \frac{\mu_m(X, W)}{\sigma_\mu} + \left(\xi + \frac{\nu}{\sigma_\tau}(\tau_m(X, W) - \bar{\tau})\right) Z + \kappa\varepsilon.
\end{aligned}
\tag{13}
$$

Simulations

$$
\begin{cases}
\mu_1(X, W) &= 0.1X - 0.2X^2 + 0.5X^3 + \sum_{p=1}^{4} \alpha_p W_p \\
\tau_1(X, W) &= 0.7X + 0.4X^2 - 0.1X^3 + \sum_{p=1}^{4} \beta_p W_p
\end{cases}
$$

$$
\begin{cases}
\mu_2(X, W) &= 0.1X - 0.2X^2 + 0.5X^3 + \sum_{p=1}^{4} \alpha_p W_p + W_5 X \\
\tau_2(X, W) &= 0.7X + 0.4X^2 - 0.1X^3 + \sum_{p=1}^{4} \beta_p W_p + 0.5 W_5 X
\end{cases}
$$

$$
\begin{cases}
\mu_3(X, W) &= \exp X + \sum_{p=1}^{4} \alpha_p \sqrt{|W_p|} \\
\tau_3(X, W) &= \sin X + \sum_{p=1}^{4} \beta_p \sqrt{|W_p|}
\end{cases}
$$

$$
\begin{cases}
\mu_4(X, W) &= \exp X + \sum_{p=1}^{4} \alpha_p \sqrt{|W_p|} + W_5 X \\
\tau_4(X, W) &= \sin X + \sum_{p=1}^{4} \beta_p \sqrt{|W_p|} + 0.5 W_5 X,
\end{cases}
\tag{14}
$$

Simulations

$$
\begin{aligned}
\xi &\in \{0.25, 2\} \\
\kappa &\in \{0.25, 2\} \\
\nu &\in \{0.25, 2\}.
\end{aligned} \tag{15}
$$

$$
\begin{aligned}
\alpha_p &= 2/p \\
\beta_p &= 1/p.
\end{aligned} \tag{16}
$$

Estimators

BART-based models:

- BART-RDD
- S-learner (BART1)
- T-Learner (BART2)
- BCF

Non-BART models:

- Calonico et al. (2019) (CKT) - local polynomial regression
- Chib et al. (2014) (CGS) - cubic splines on the running variable
- Kreiß and Rothe (2021) (KR) - local linear regression to high-dimensional settings

Estimators are compared in terms of RMSE, coverage and interval length

Summary of Results

ATE estimation:

- BART-RDD generally outperforms and never lags far behind the other estimators
- Only the T-learner BART stands out as a reasonable alternative among BART-based models
- Among the non-BART models, CKT stands out as the best, while CGS is competitive but more sensitive to noise
- KR is the worst performer and highly sensitive to noise
- Model complexity plays an important role

CATE estimation:

- BART-RDD clearly outperforms the others in CATE estimation, producing more precise estimates and intervals with comparable size but better coverage

Application: effect of academic probation on education

- We investigate the effect of academic probation in educational outcomes in a large Canadian university (Lindo et al., 2010)
- Students who, by the end of each term, present GPA lower than a certain threshold (which differs between each campus) are placed on academic probation and must improve their GPA in the next term
- Punishment if they fail to achieve this goal can range from 1-year to permanent suspension from the university
- We focus on GPA in the term after a student is placed on probation

Application

- Running variable is the negative distance between a student's GPA and the probation threshold, meaning students below the limit have a positive score and the cutoff is 0
- Additional student features: gender, age, a *dummy* for being born in North America, attempted credits in the first year, *dummies* for which campus each student belongs to, 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.

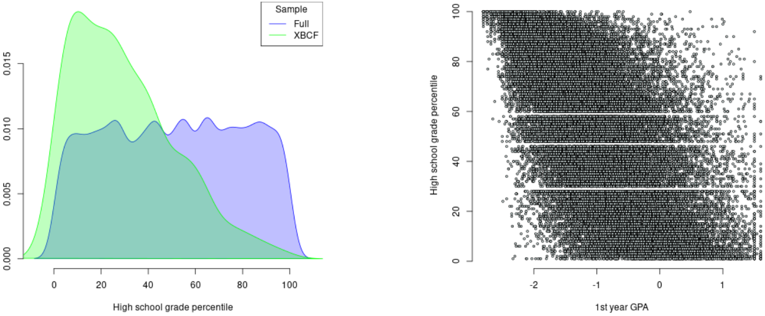
Application

(1) full sample, (2) $h = 0.1$, (3) $h = 0.46$

	(1)		(2)		(3)	
	Mean	Std. Dev	Mean	Std. Dev	Mean	Std. Dev
Next Term GPA	2.57	0.91	1.95	0.81	1.98	0.8
Distance from cutoff	-0.96	0.86	0	0.05	-0.08	0.26
Treatment assignment	0.14	0.35	0.41	0.49	0.36	0.48
High school grade percentile	51	28.71	31.65	22.79	32.76	23.15
Credits attempted in first year	4.58	0.51	4.39	0.54	4.42	0.53
Age at entry	18.66	0.74	18.72	0.75	18.71	0.74
Male	0.38	0.49	0.38	0.48	0.37	0.48
Born in North America	0.87	0.34	0.87	0.34	0.87	0.34
Campus 1	0.59	0.49	0.45	0.5	0.47	0.5
Campus 2	0.17	0.38	0.21	0.41	0.21	0.41
Campus 3	0.24	0.42	0.34	0.47	0.32	0.46

Table 1: Descriptive statistics

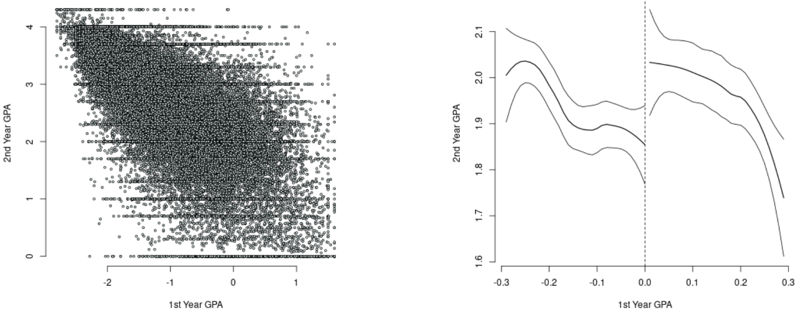
Application



(a) Density (b) Scatter plot

Figure 9: High school grade percentile

Application



(a) Scatter plot

(b) Loess Fit

Figure 10: Second x First Year GPA

Application: BART-RDD vs CKT

	Controls	$\hat{\tau}$	95% CI	h	N
BART-RDD	No	0.11	[0.04,0.17]	0.1	1757
	Yes	0.13	[0.08,0.2]	0.1	1757
CKT	No	0.22	[0.13,0.3]	0.47	8776
	Yes	0.22	[0.12,0.3]	0.46	8776

Table 2: RD Estimates

Application: fit-the-fit

- As in Hahn et al. (2020), we explore the individual effect estimates – the posterior mean of the individual effects – by fitting a CART tree to these estimates based on the covariate set ('fit-the-fit')
- With this strategy, we allow the data to determine relevant treatment effective modifiers and potential interactions between them

Application: fit-the-fit

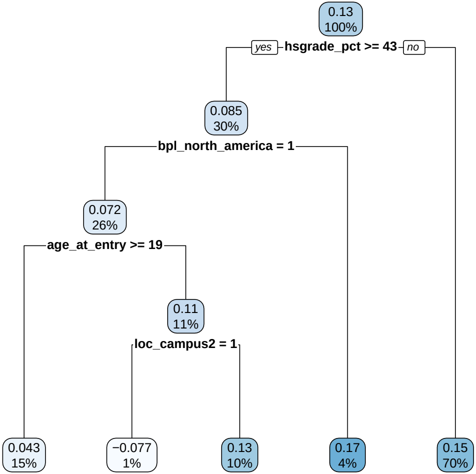


Figure 11: CART trees for individual effect estimates

Application: fit-the-fit

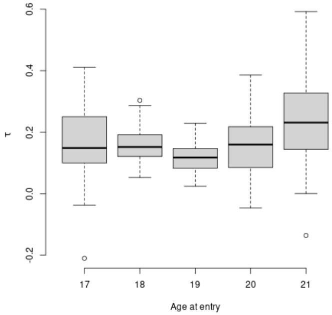
It indicates that high school grades, age and campus location are important effect moderators.

The effects of the probation policy are decreasing on high school grades and age, meaning younger students who performed worst in high school are likely to benefit the most from the policy.

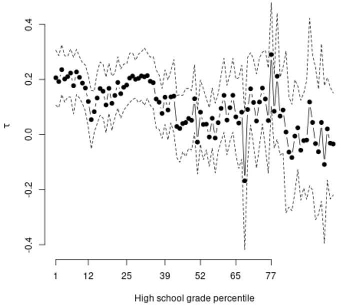
Campus 1 is the central campus and has the lowest acceptance rate (55%) and more closely resembles a large university while the other two have a higher acceptance rate (77%) and are composed mainly of part-time and commuter students.

It would make sense then that the composition of each campus should affect the effectiveness of the probation policy.

Application



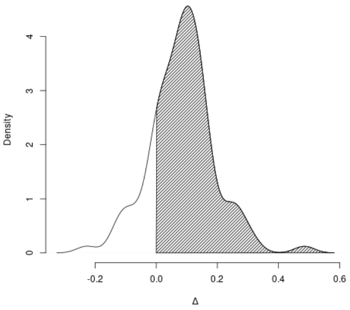
(a) CATE posterior by age



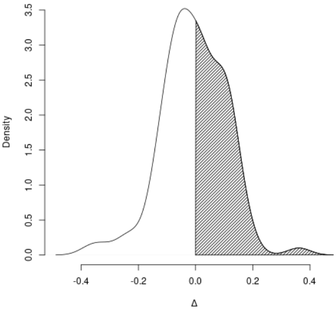
(b) CATE posterior by high school grade percentile

Application

Δ_1: Difference in the posterior distribution for students below 19 in campus 3 versus the other campuses.
Δ_2: Difference in the posterior distribution for students below 19 and below the 34th percentile of high school grades in campus 3 versus the other campuses.



(a) Δ_1



(b) Δ_2

Conclusion

- **Main contributions:** incorporating RDD assumptions into the BART framework and producing reliable ATE and CATE estimates
- **Results:**
  - BART-RDD presents lower errors, competitive coverage and smaller intervals than commonly used polynomial-based estimators
  - ATE variance for BART-RDD is not sensitive to the strength of heterogeneity in the data
  - BCF and S-BART are still good options for CATE estimation; BART-RDD presents better coverage for the CATE at the cost of larger intervals
- **Limitations:** Sensitivity to prior hyperparameters
- **Next steps:** Application to real data Lindo et al. (2010), exploration of CATE results, more formal argument about identification of the BART-RDD tree ensemble

References I

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.

Chib, S., Greenberg, E., and Simoni, A. (2014). Nonparametric bayes analysis of the sharp and fuzzy regression discontinuity designs. *Econometric Theory*, pages 1–53.

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. (2021). Stochastic tree ensembles for regularized nonlinear regression. *Journal of the American Statistical Association*, pages 1–20.

References II

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

Imbens, G. and Kalyanaraman, K. (2012). Optimal bandwidth choice for the regression discontinuity estimator. *The Review of economic studies*, 79(3):933–959.

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. (2021). Inference in regression discontinuity designs with high-dimensional covariates. *arXiv preprint arXiv:2110.13725*.

References III

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.

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.