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]

June 24, 2024

https://rafaelcalcantara.github.io

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

Outline

Our 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 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 illustrate the new method by analyzing data studied originally by Lindo et al. (2010) to estimate the effect of academic probation on university students' GPA

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:

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

for some cutoff value c.

The distribution of Y, the outcome, conditional on X is assumed to be smooth.

The effect of the treatment is measured by the size of the discontinuity at c.

An illustration

We use BART models to "learn" the two curves

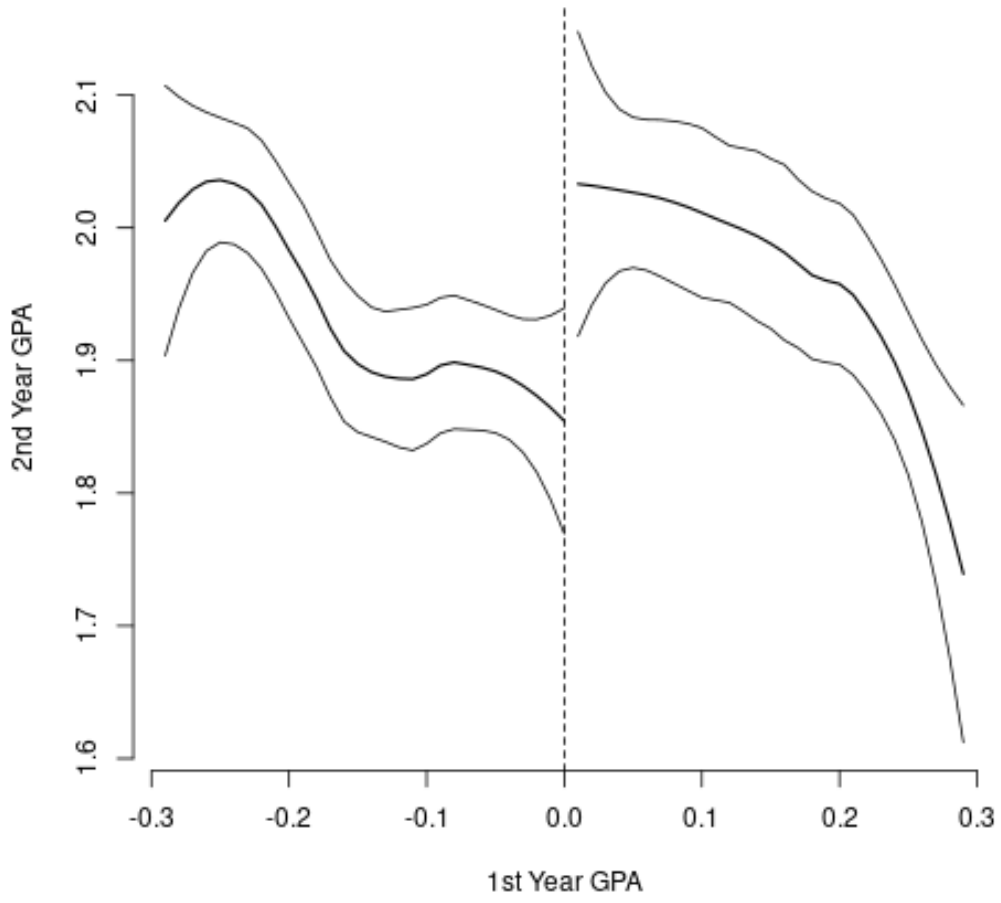


Figure 1: Effect of 1st year GPA cutoff on 2nd year GPA.

RDD - potential outcomes

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

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

The running variable and covariates, (X_i, w_i), are assumed to be unaffected by the treatment.

The motivation behind the RDD is the assumption that individuals who lie just above or just below the cutoff must be very similar except for the treatment assignment.

In this case, the cutoff rule acts as a randomization device for these units. Therefore, interest lies in treatment effects at the cutoff.

RDD - ATE

We consider some comparison between

$$\mathbb{E}[Y_i|Z_i = 0, X_i = c] \quad \text{and} \quad \mathbb{E}[Y_i|Z_i = 1, X_i = c].$$

We focus on the difference in expected potential outcomes:

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

While the second term in expression (2) is never observed, under the assumption that the distribution of Y_i is smooth in X_i, the treatment effect may be estimated as a limit:

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

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 assumption that individuals near the cutoff are similar 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.

Brief BART review

Letting $f(x) = E(Y \mid X = x)$ denote a smooth function of a covariate vector X, the BART model is traditionally written

$$\begin{aligned} Y_i &= f(x_i) + \varepsilon_i \\ &= \sum_{j=1}^{k} g_j(x_i; T_j, m_j) + \varepsilon_i \end{aligned} \tag{3}$$

where $\varepsilon_i \sim \mathrm{N}(0, \sigma^2)$ is a normally distributed additive error term

$g_j(x; T_j, m_j)$: 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, m_j, which records the values $g(\cdot)$ takes on each of those regions

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

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

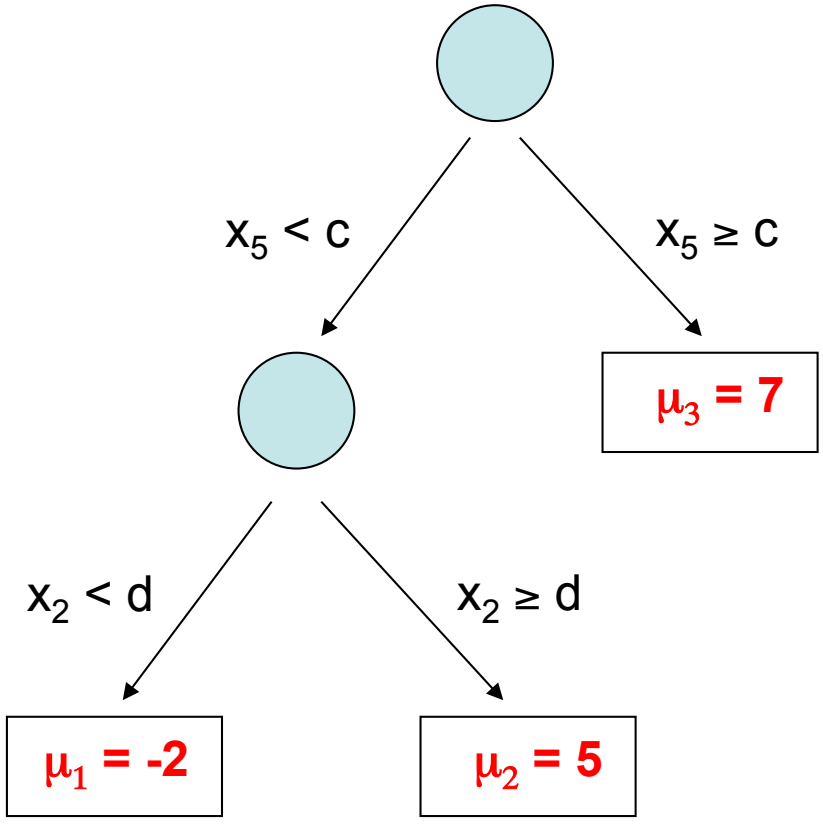
We can then use the draws as our inference for $g(x; \theta)$.

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.

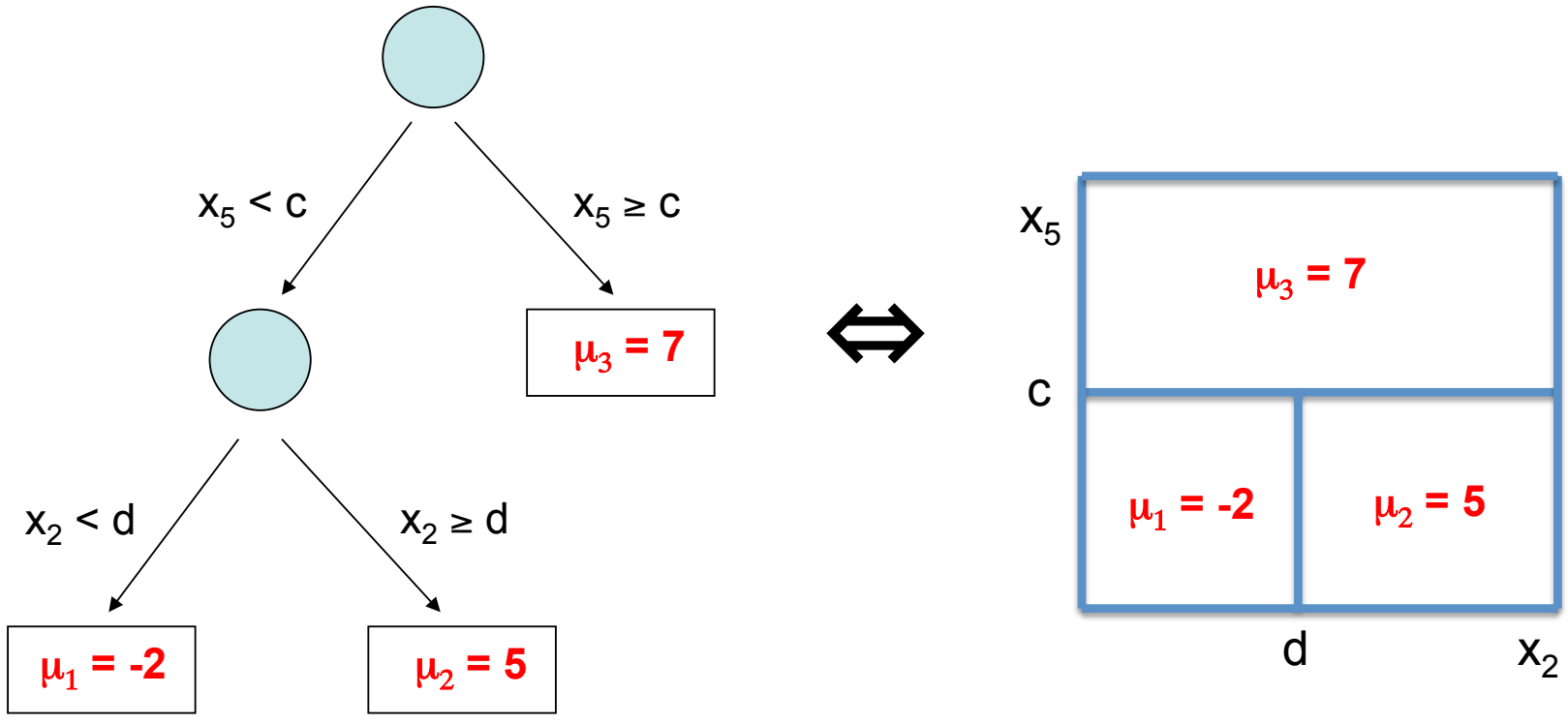
Let $g(x; \theta)$, $\theta = (T, M)$ be a regression tree function that assigns a μ value to x.



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.

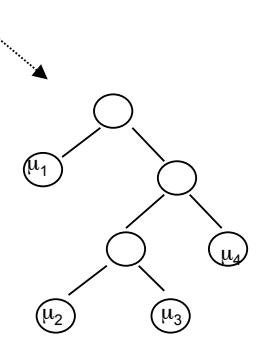
Bayesian model averaging (BMA) applied to the posterior arising from a Bayesian single-tree model.

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

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



$m = 200, 1000, \ldots,$ 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:

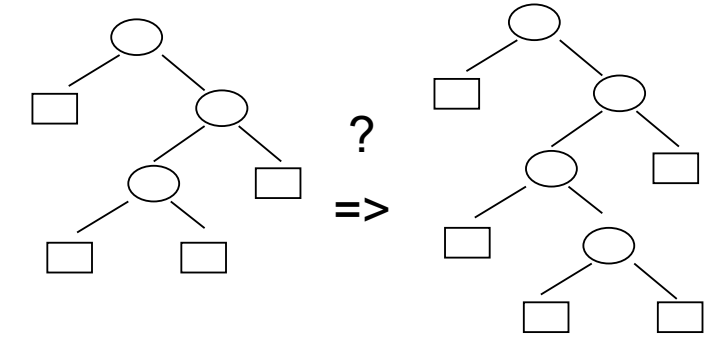
$$\begin{aligned} (T_i, M_i) \quad &| \quad (T_1, M_1, \ldots, T_{i-1}, M_{i-1}, T_{i+1}, M_{i+1}, \ldots, T_m, M_m, \sigma) \\ \sigma \quad &| \quad (T_1, M_1, \ldots, \ldots, T_m, M_m) \end{aligned}$$

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.

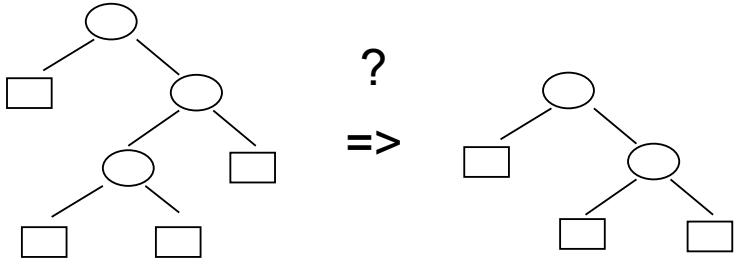
We integrate out M to draw T and then draw $M|T$.

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.

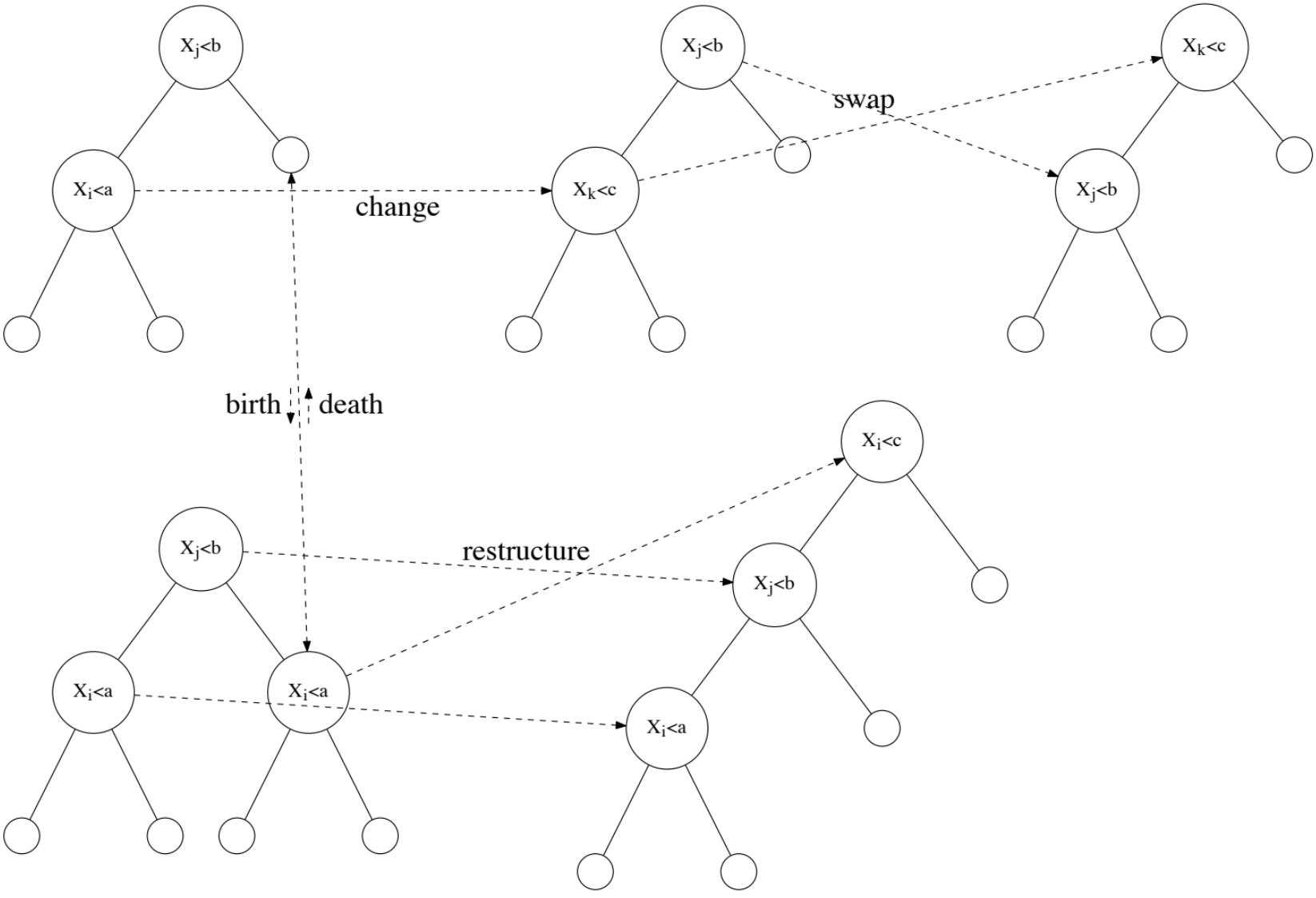


propose a more complex tree

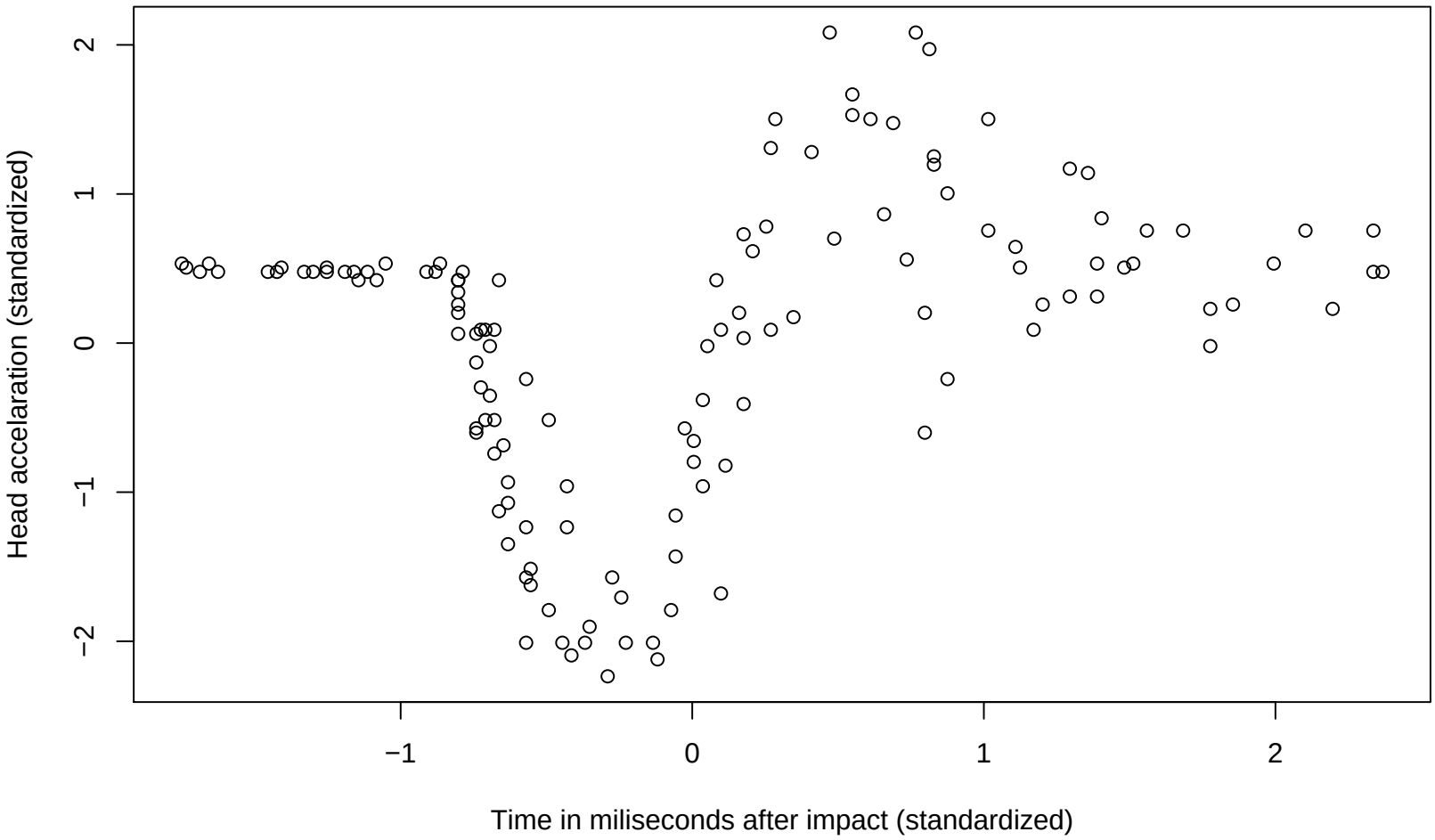


propose a simpler tree

Tree moves


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

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

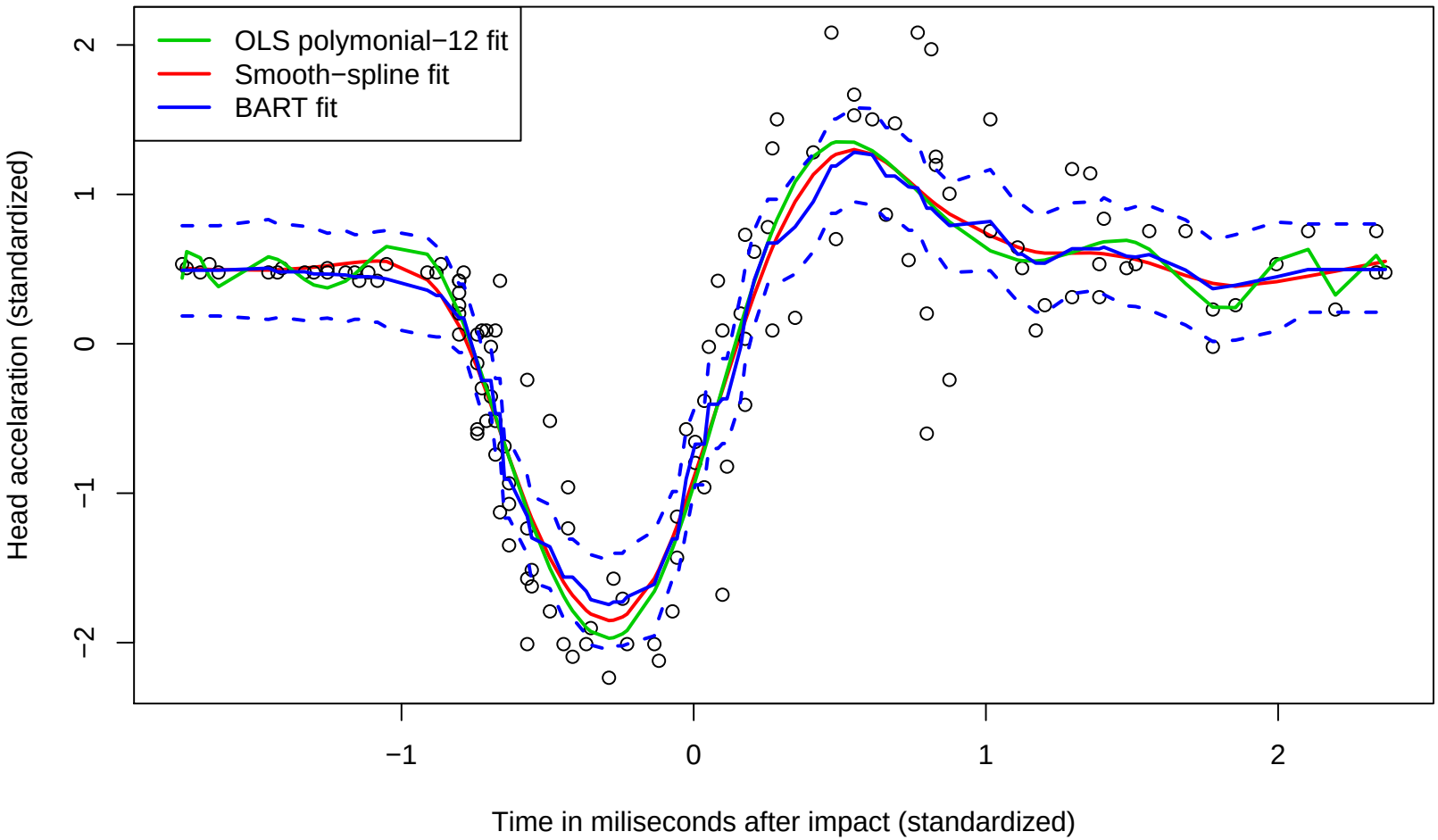
```
install.packages("BART")
library(MASS)
library(BART)
xt = mcycle$times[1:132]
yt = mcycle$accel[1:132]
xt = (xt-mean(xt))/sqrt(var(xt))
yt = (yt-mean(yt))/sqrt(var(yt))

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)

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

T-learners: Two BART models (Künzel et al., 2019).

These approaches are not ideal in common causal inference settings:

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

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

BCF

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

$$Y_i = \mu(X_i, \mathrm{w}_i) + \tau(X_i, \mathrm{w}_i) b_{z_i} + \varepsilon_i, \quad \varepsilon_i \sim N(0, \sigma^2), \qquad (4)$$

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.

When $b_0 = 0$ and $b_1 = 1$,

$\mu(x) = \mathbb{E}(Y^0 \mid X = x)$ and $\tau(x) = \mathbb{E}(Y^1 \mid X = x) - \mathbb{E}(Y^0 \mid X = x)$.

The ATE can be expressed as

$$\mathbb{E}(Y^1 \mid X = x) - \mathbb{E}(Y^0 \mid X = x) = (b_1 - b_0)\tau(x). \qquad (5)$$

XBART and XBCF

He and Hahn (2021) propose the accelerated Bayesian additive regression trees (XBART) algorithm for BART-like models. XBART grows new trees recursively, but stochastically, at each step while using a similar set of cutpoints and splitting criteria as BART, which allows for much faster exploration of the posterior space.

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

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

XBART and XBCF

The new model is almost the same as (4) except that XBCF allows the error variance to change for each treatment status:

$$\begin{aligned} Y_i &= a\mu(x_i) + b_{z_i}\tilde{\tau}(x_i) + \epsilon_i, \quad \epsilon_i \sim N(0, \sigma^2_{z_i}) \\ a &\sim \mathrm{N}(0,1), \quad b_0, b_1 \sim \mathrm{N}(0, 1/2), \end{aligned} \tag{6}$$

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 is particularly well-suited to the RDD context.

BART-RDD

Just as BCF was developed to address shortcomings of off-the-shelf BART implementations for treatment effect estimation assuming conditional unconfoundedness, here we propose to modify the BCF model to cope with challenges that are unique to regression discontinuity designs.

Our strategy is to ensure that the data used to make predictions at $X = c$ warrant a causal interpretation, i.e. $\mu(x = c, w)$ 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

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

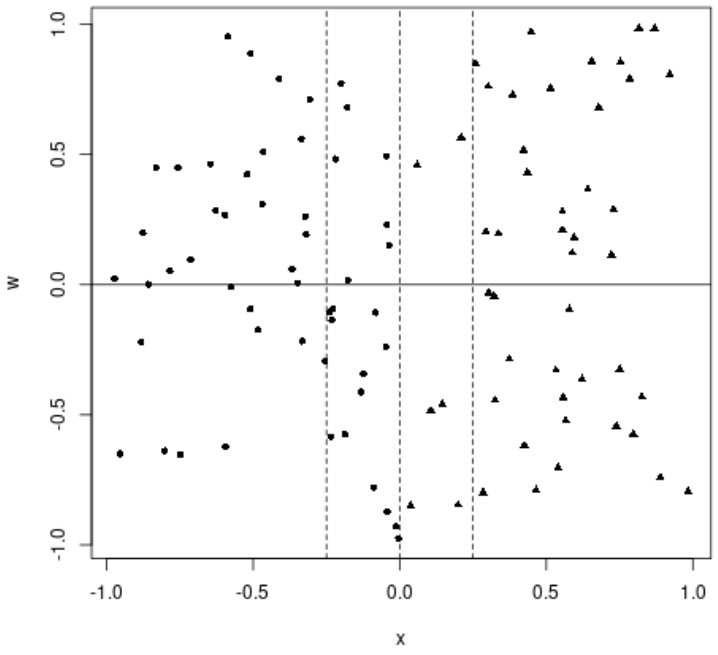
The initial clause says that any node which does not make predictions at the cutoff remains entirely unrestricted;

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.

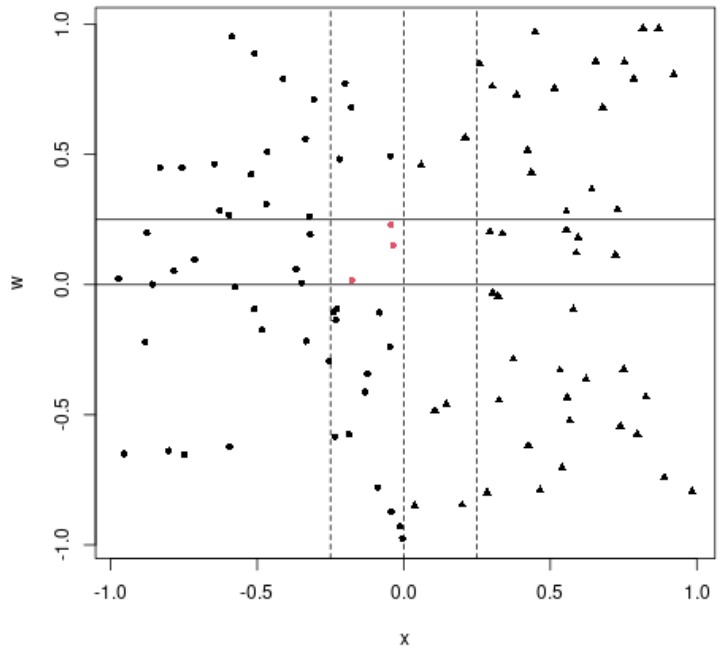
Illustration

- 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



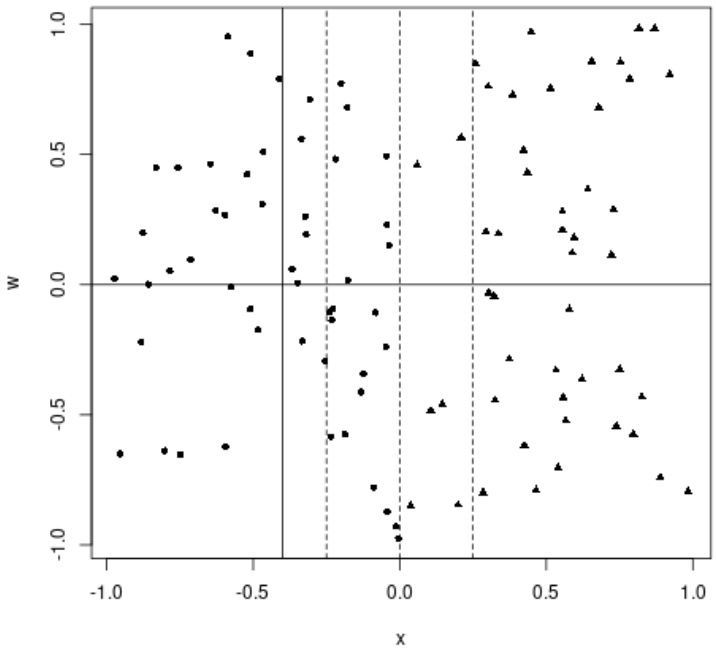
(a)



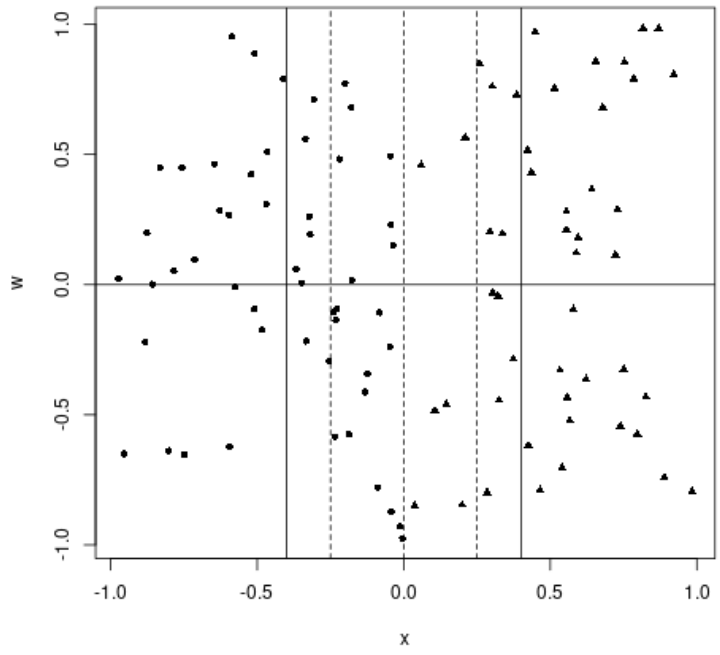
(b)

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 the condition (ii) is violated: both nodes contain the identification strip, but are highly populated by points outside of it
- However, condition (i) 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 (i) 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 (i) is not violated, but condition (ii) is violated for the node $W \geq 0 \cup X \geq -0.4$
- In this instance, the algorithm would accept the splits and force the tree to continue splitting until condition (i) 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 be compared 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

- For setting the prior hyperparameters for a given sample (Y, X, W), we suggest the following prior ellicitation procedure: take (X, W), generate s samples of a known DGP $Y_s(X, W)$, fit the model to each generated sample using different combinations of the parameters and choose the one that leads to the lowest prediction error for this synthetic data

- 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

Simulations

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:

$$\mu(X, W) = \frac{\mu_0(X, W)}{\sigma(\mu_0(X, W))}\delta_\mu$$

$$\tau(X, W) = \bar{\tau} + \frac{\tau_0(X, W)}{\sigma(\tau_0(X, W))}\delta_\tau$$

$$Y = \mu(X, W) + \tau(X, W)Z + \varepsilon$$

$$\bar{\tau} = \{0.2, 0.5\}$$

$$\delta_\mu = \{0.5, 1.25\}$$

$$\delta_\tau = \{0.1, 0.3\}$$

$$\varepsilon \sim \mathcal{N}(0, 1),$$

Simulations

$$\mu_0(X, W) = 3x^5 - 2.5x^4 - 1.5x^3 + 2x^2 + 3x + 2 + \frac{1}{2}\sum_{p=1}^{4}(w_p - E[w_p])$$

$$\tau_0(X, W) = -0.1x + \frac{1}{4}\sum_{p=1}^{4}(w_p - E[w_p]) \tag{7}$$

Estimators

BART-based models:

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

Non-BART models:

- Calonico et al. (2019) (LLR) - local polynomial regression
- Chib et al. (2014) (CGS) - cubic splines on the running variable

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

Summary of Results

ATE estimation:

- Regarding the RMSE, BART-RDD generally outperforms and never lags far behind the other estimators
- Among the non-BART models, LLR stands out as the best, while CGS is much more sensitive to noise
- While LLR, CGS and S-BART usually present coverage above 90%, BART-RDD present coverage that is never below 70% with much tighter intervals

CATE estimation:

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

ATE Results

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

ATE Results

Table 2: Coverage rate for the ATE

$\bar{\tau}$	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	0.940
0.2	0.5	0.3	0.954	0.950	0.724	0.950	0.932
0.2	1.25	0.1	0.782	0.940	0.538	0.970	0.930
0.2	1.25	0.3	0.718	0.950	0.520	0.964	0.938
0.5	0.5	0.1	0.900	0.866	0.828	0.958	0.946
0.5	0.5	0.3	0.920	0.890	0.772	0.962	0.942
0.5	1.25	0.1	0.722	0.870	0.558	0.966	0.918
0.5	1.25	0.3	0.702	0.894	0.572	0.962	0.934

ATE Results

Table 3: Interval sizes for the ATE

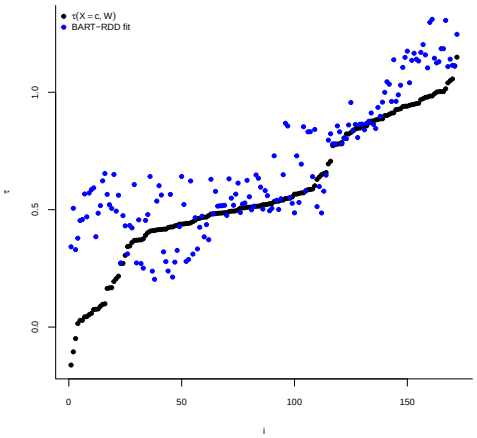
τ	δ_μ	δ_τ	BART-RDD	S-BART	T-BART	CGS	LLR
0.2	0.5	0.1	0.424	0.713	0.719	1.598	0.855
0.2	0.5	0.3	0.442	0.757	0.677	1.604	0.877
0.2	1.25	0.1	0.546	0.970	0.797	1.792	0.863
0.2	1.25	0.3	0.539	1.068	0.794	1.814	0.880
0.5	0.5	0.1	0.536	0.859	0.743	1.604	0.870
0.5	0.5	0.3	0.519	0.913	0.704	1.607	0.870
0.5	1.25	0.1	0.579	1.239	0.810	1.788	0.870
0.5	1.25	0.3	0.567	1.313	0.818	1.798	0.872

CATE Results

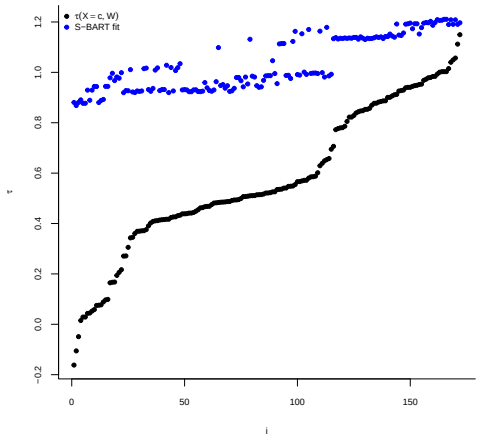
Table 4: RMSE - CATE

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

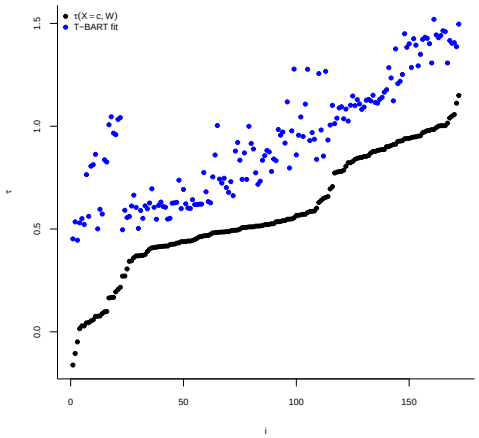
CATE illustration



(a) BART-RDD



(b) S-BART



(c) T-BART

Figure 4: Fit for $\tau(X = c, W)$ for each method when $\delta_\mu = 0.5$, $\delta_\tau = 0.3$ and $\bar{\tau} = 0.5$ versus the true function

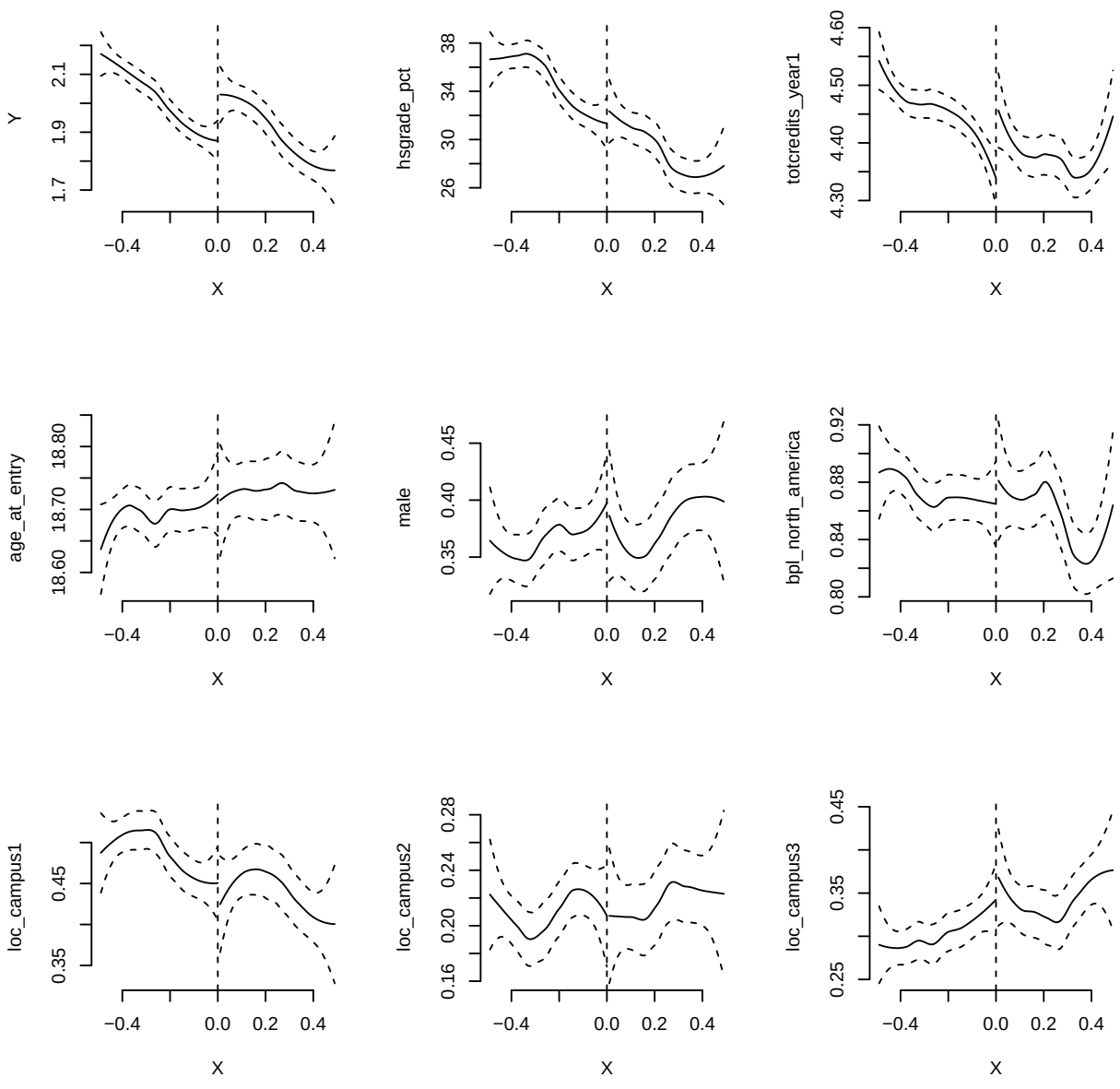
Application: effect of academic probation on education

- 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



Application: BART-RDD vs CKT

Table 5: 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

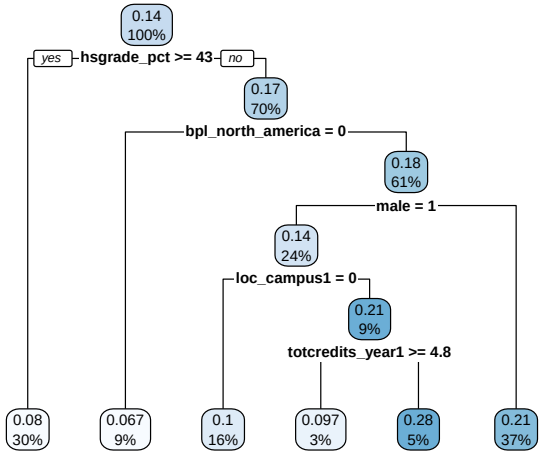
Application: fit-the-fit

- 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')

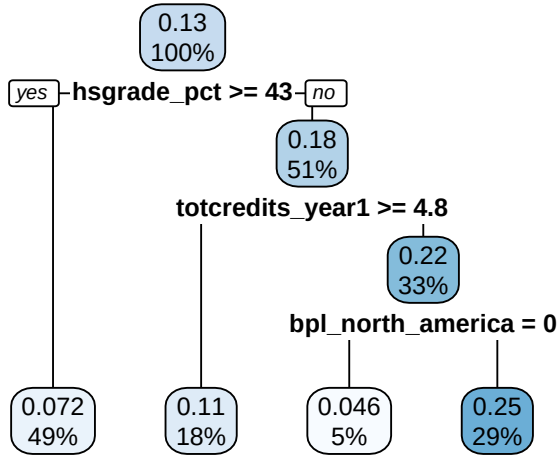
- With this strategy, we allow the data to determine relevant treatment effective modifiers and potential interactions between them

Application: fit-the-fit

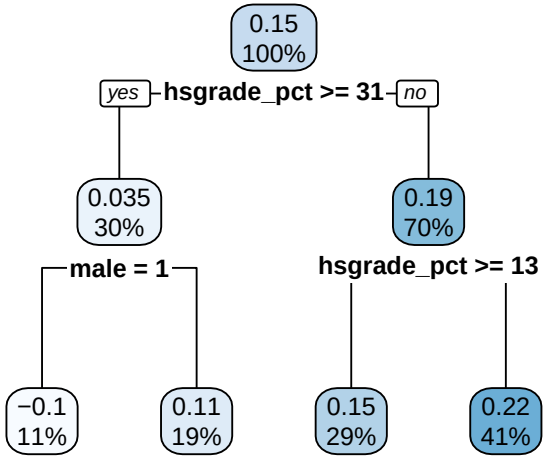
Full sample



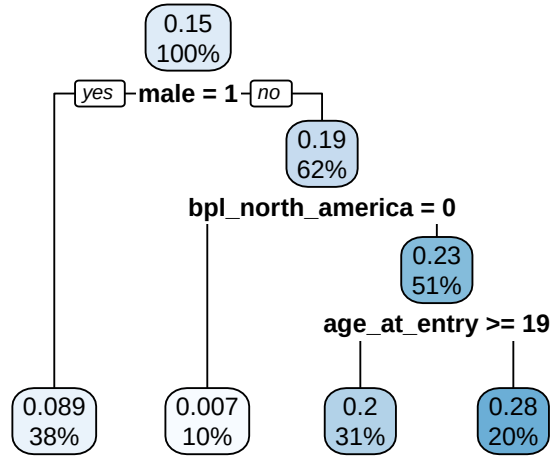
Campus 1



Campus 2



Campus 3



Application: fit-the-fit

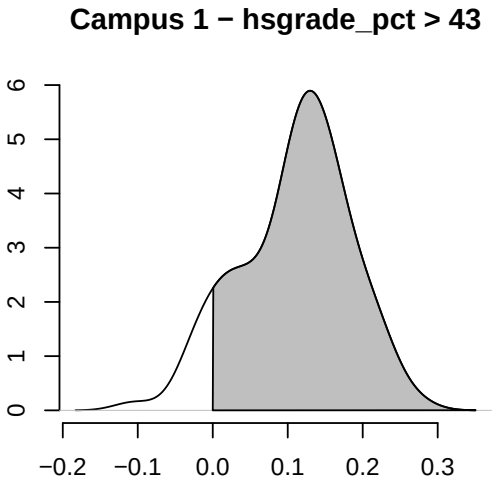
The figure indicates that high school grades are important effect moderators for campus 1 and 2 but not 3, and that the moderators change per campus (for example, credits in year 1 only for campus 1, gender only for campus 2 and age at entry only for campus 3).

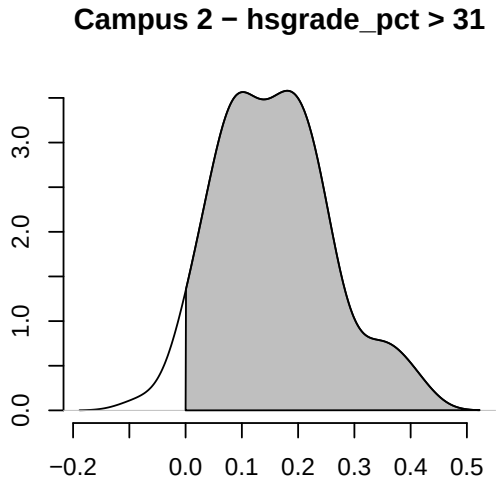
Overall, the effects of the probation policy are decreasing on high school grades, meaning students who performed worst in high school are likely to benefit the most from the policy.

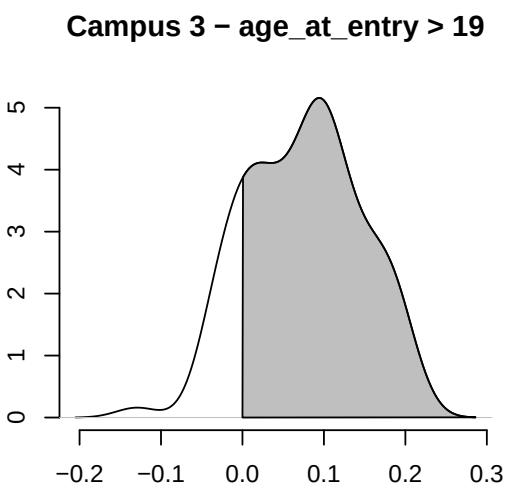
Campus 1 is the central campus and more closely resembles a large university while the other two are composed mainly of part-time and commuter students: it would make sense then that the composition of each campus should affect the effectiveness of the probation policy.

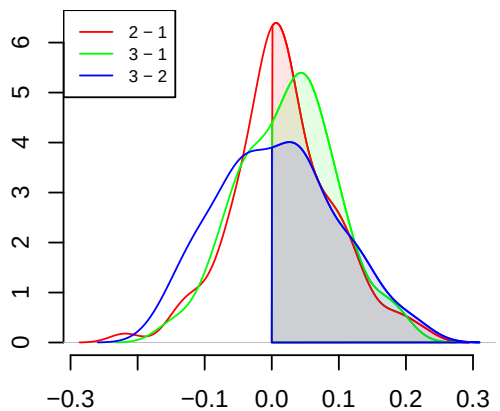
Application - Posterior Comparisons

Details









Conclusion

- **Main contributions:** incorporating RDD assumptions into the BART framework and producing reliable CATE estimates
- **Results:** BART-RDD presents lower errors, competitive coverage and smaller intervals than other commonly used estimators based on parametric specifications
- **Limitations:** Sensitivity to prior hyperparameters
- **Extensions:** extrapolating the estimates beyond the cutoff (Wang et al., 2023), modelling non-Gaussian outcomes — *e.g.* discrete or t-distributed — and extending our strategy for settings with multiple cutoffs

Final 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.

Final 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.

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.

Final 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.

Wang, M., He, J., and Hahn, P. R. (2023). Local gaussian process extrapolation for bart models with applications to causal inference. *Journal of Computational and Graphical Statistics*, (just-accepted):1–22.

Application - Posterior Comparisons

Main results 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, which has a 92% posterior mass above 0; the second panel performs the same analysis for the 31-st percentile of high-school grades for students in campus 2, which has a 95% posterior mass above 0; the third panel presents the posterior difference between students that got into college younger versus older than 19 in campus 3, which has a posterior mass of 84% above 0; the last panel presents the posterior differences in the ATE between each campus: there is a 66% posterior probability of a larger effect for campus 3 compared to campus 1, a 59% probability for a larger effect on campus 2 compared to campus 1 and a 54% probability of a larger effect on campus 3 compared to campus 2