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

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

April 29, 2025

https://rafaelcalcantara.github.io

[1]The University of Texas at Austin
[2]SoMSS, Arizona State University
[3]Insper Institute of Education and Research

Outline

Our contribution

- We propose a modification of the BART model of Chipman et al. (2010) in which the constant leaf node predictions are replaced by RDD regressions (*i.e.* regressions on the running variable and treatment dummy)
- 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

We conceptualize the treatment effect estimation problem via a quartet of random variables (Y, X, Z, U)

- Y: outcome variable;
- X: running variable;
- Z: treatment assignment indicator variable;
- U: additional, possibly unobserved, causal factors.

What specifically makes this correspond to an RDD is that we stipulate that $Z = \mathbb{I}(X > c)$, for cutoff c. We assume $c = 0$ without loss of generality.

Causal DAG

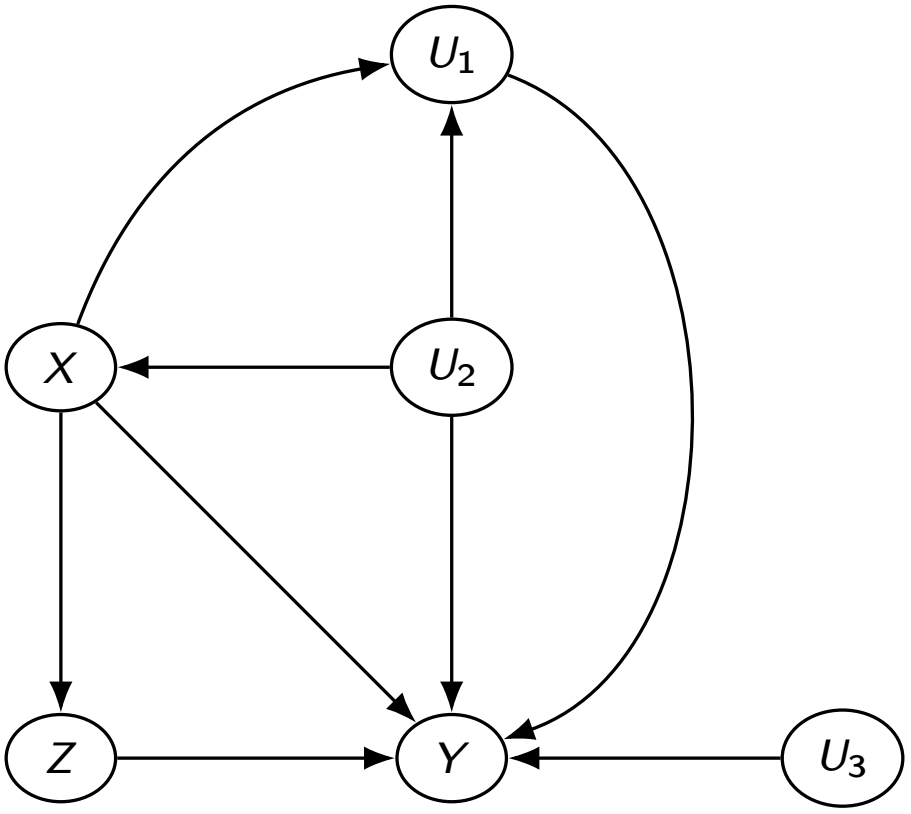


Figure 1: A causal directed acyclic graph representing the general structure of a regression discontinuity design problem, where $Z = \mathbb{I}(X > 0)$. Here $U = (U_1, U_2, U_3)$ is depicted in terms of three distinct components, exhaustively illustrating the various relationships that could obtain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

Causal DAG

Two key features of this diagram:

1. X blocks the impact of U on Z: in other words, X satisfies the back-door criterion for learning causal effects of Z on Y;

2. X and U are not descendants of Z.

Functional Causal Model

We may express Y as some function of its graph parents, the random variables (X, Z, U):

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

This formulation relates to the potential outcomes framework straightforwardly:

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

Note that this construction implies the *consistency* condition: $Y = Y^1 Z + Y^0(1 - Z)$. Likewise, this construction implies the *no interference* condition because each Y_i is considered to be produced with arguments (X_i, Z_i, U_i) and not those from other units j; in particular, in constructing Y_i, F does not take Z_j for $j \neq i$ as an argument.

Identification

Next, we define the following conditional expectations

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

with which we can define the treatment effect function

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

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

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

the right-hand-sides of which can be estimated from sample data, which we supposed to be i.i.d realizations of (Y_i, X_i, Z_i)

Identification

Because $Z = \mathbb{I}(X > 0)$ we can in fact only learn $\mu_1(x)$ for $X > 0$ and $\mu_0(x)$ for $X < 0$. In potential outcomes terminology, conditioning on X satisfies ignorability,

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

but not *strong ignorability*, because overlap is violated.

For continuous X, it is possible to estimate $\tau(0)$ as the difference between $\mu_1(0) - \mu_0(0)$, so long as one is willing to assume that $\mu_1(x)$ and $\mu_0(x)$ are both suitably smooth functions of x: any inferred discontinuity at $x = 0$ must therefore be attributable to treatment effect (Hahn et al., 2001)

An illustration

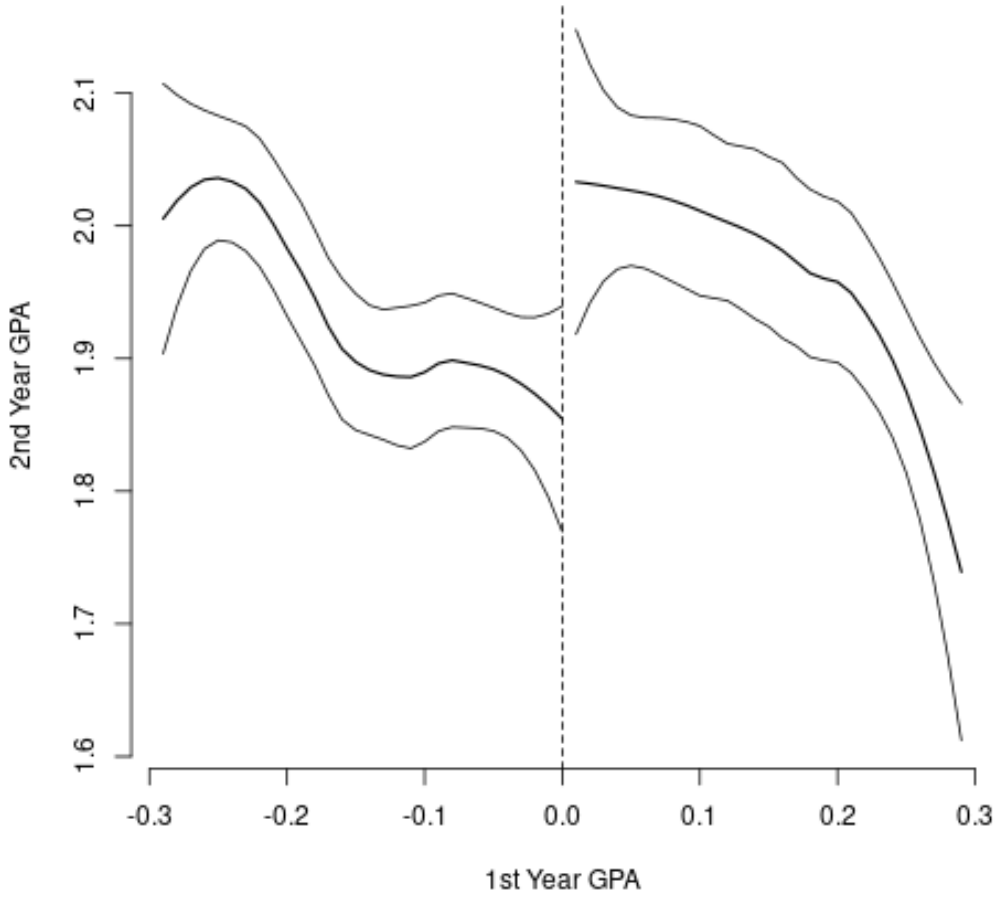


Figure 2: Effect of 1st year GPA cutoff on 2nd year GPA.

CATE in RDD

Our focus is learning $\tau(0, \mathrm{w})$ for some covariate vector w. Let the observed covariate vector $W = \varphi(U)$ be an observable function of the (possibly unobservable) causal factors U. We may then define our potential outcome means as

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

and our treatment effect function as

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

CATE in RDD

We consider our data to be independent and identically distributed realizations (Y_i, X_i, Z_i, W_i) for $i = 1, \ldots, n$

We must assume that $\mu_1(x, \mathrm{w})$ and $\mu_0(x, \mathrm{w})$ are suitably smooth functions of x, *for every* w; in other words, for each value of w the usual continuity-based identification assumptions must hold.

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

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

$$\begin{aligned} Y_i &= f(x_i) + \varepsilon_i \\ &= \sum_{j=1}^{k} g_j(x_i; T_j, m_j) + \varepsilon_i \end{aligned} \tag{5}$$

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

BART for causal inference

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 (Hahn et al., 2020):

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.

BARDDT

We propose a BART model where the trees are allowed to split on (x, w) but where each leaf node parameter is a vector of regression coefficients tailored to the RDD context

Let ψ denote the following basis vector:

$$\psi(x, z) = \begin{bmatrix} 1 & zx & (1-z)x & z \end{bmatrix}. \tag{6}$$

Let $b_j(x, \mathrm{w})$ denote the node in the jth tree which contains the point (x, w); then the prediction function for tree j is defined to be:

$$g_j(x, \mathrm{w}, z) = \psi(x, z)\Gamma_{b_j(x,\mathrm{w})} \tag{7}$$

for a leaf-specific regression vector $\Gamma_{b_j} = (\eta_{b_j}, \lambda_{b_j}, \theta_{b_j}, \Delta_{b_j})^t$.

BARDDT

Let n_{b_j} denote the number of data points allocated to node b in the jth tree and Ψ_{b_j} denote the $n_{b_j} \times 4$ matrix, with rows equal to $\psi(x, z)$ for all $(x_i, z_i) \in b_j$. The model for observations assigned to leaf b_j can be expressed in matrix notation as:

$$\begin{aligned} {}_{b_j} \mid \Gamma_{b_j}, \sigma^2 &\sim \mathrm{N}(\Psi_{b_j}\Gamma_{b_j}, \sigma^2) \\ \Gamma_{b_j} &\sim \mathrm{N}(0, \Sigma_0), \end{aligned} \tag{8}$$

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

BARDDT

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

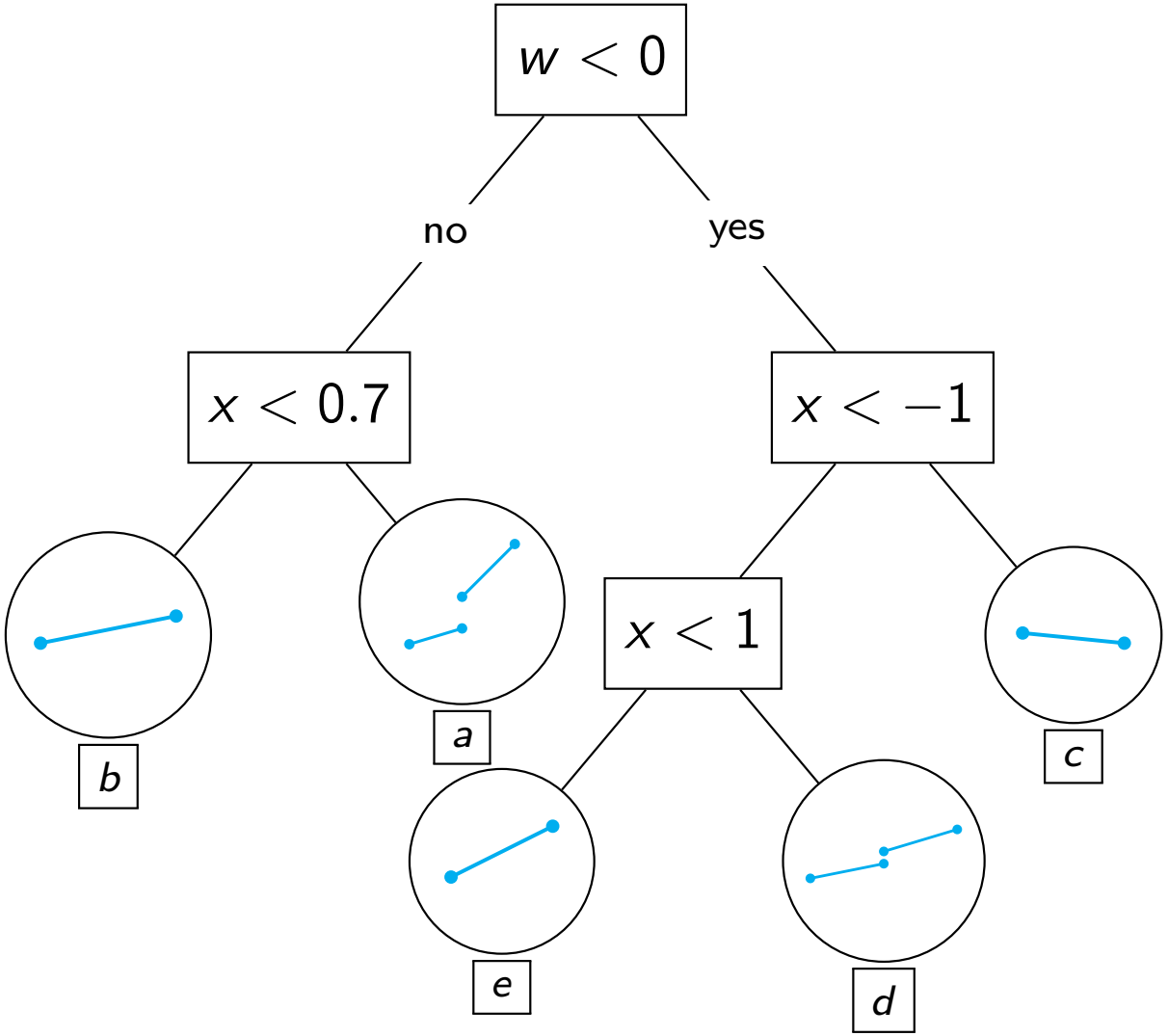
$$
\begin{aligned}
\tau(0, \mathrm{w}) &= \mathbb{E}[Y^1 \mid X = 0, W = \mathrm{w}] - \mathbb{E}[Y^0 \mid X = 0, W = \mathrm{w}] \\
&= \mathbb{E}[Y \mid X = 0, W = \mathrm{w}, Z = 1] - \mathbb{E}[Y \mid X = 0, W = \mathrm{w}, Z = 0] \\
&= \sum_{j=1}^{J} g_j(0, \mathrm{w}, 1) - \sum_{j=1}^{J} g_j(0,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

BARDDT

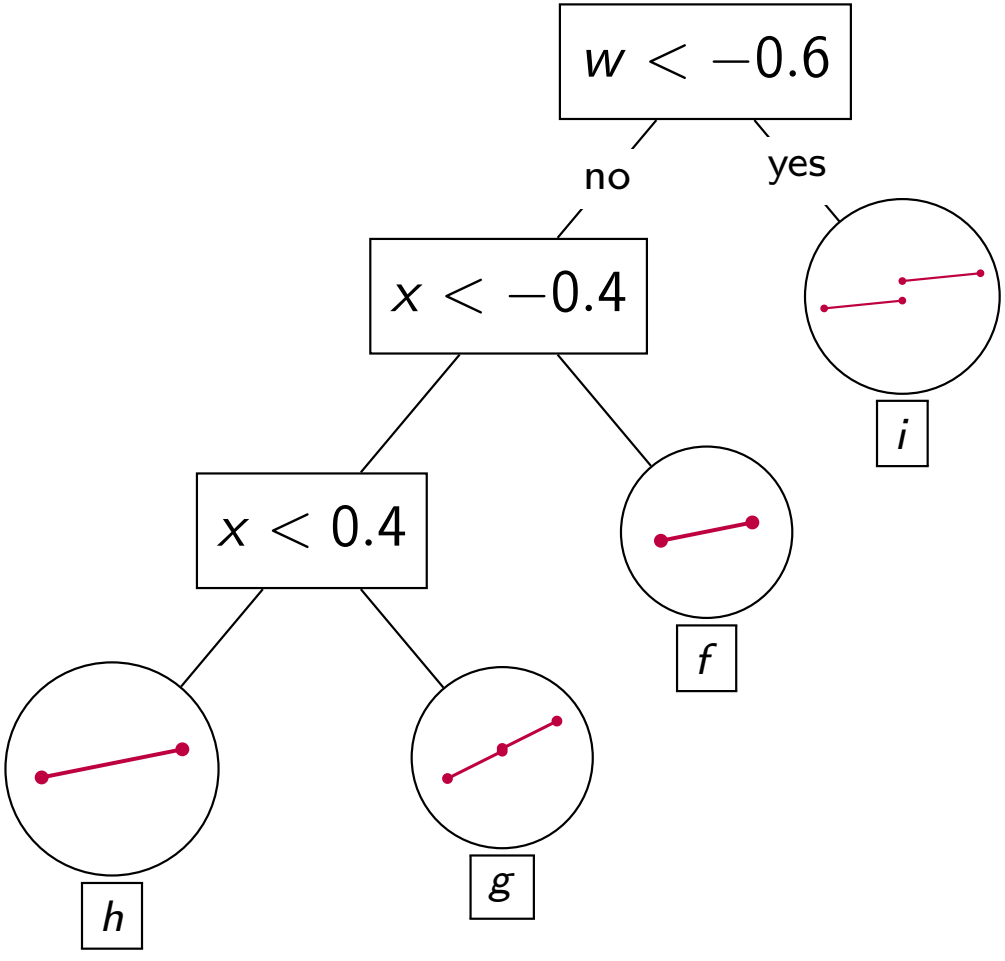
Posterior sampling from this model proceeds nearly identically to the traditional BART Gibbs sampler, but with a modified log marginal likelihood, which generalizes the one from constant-leaf BART to one from a node-level regression with the basis defined above.

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

BARDDT - Illustration I



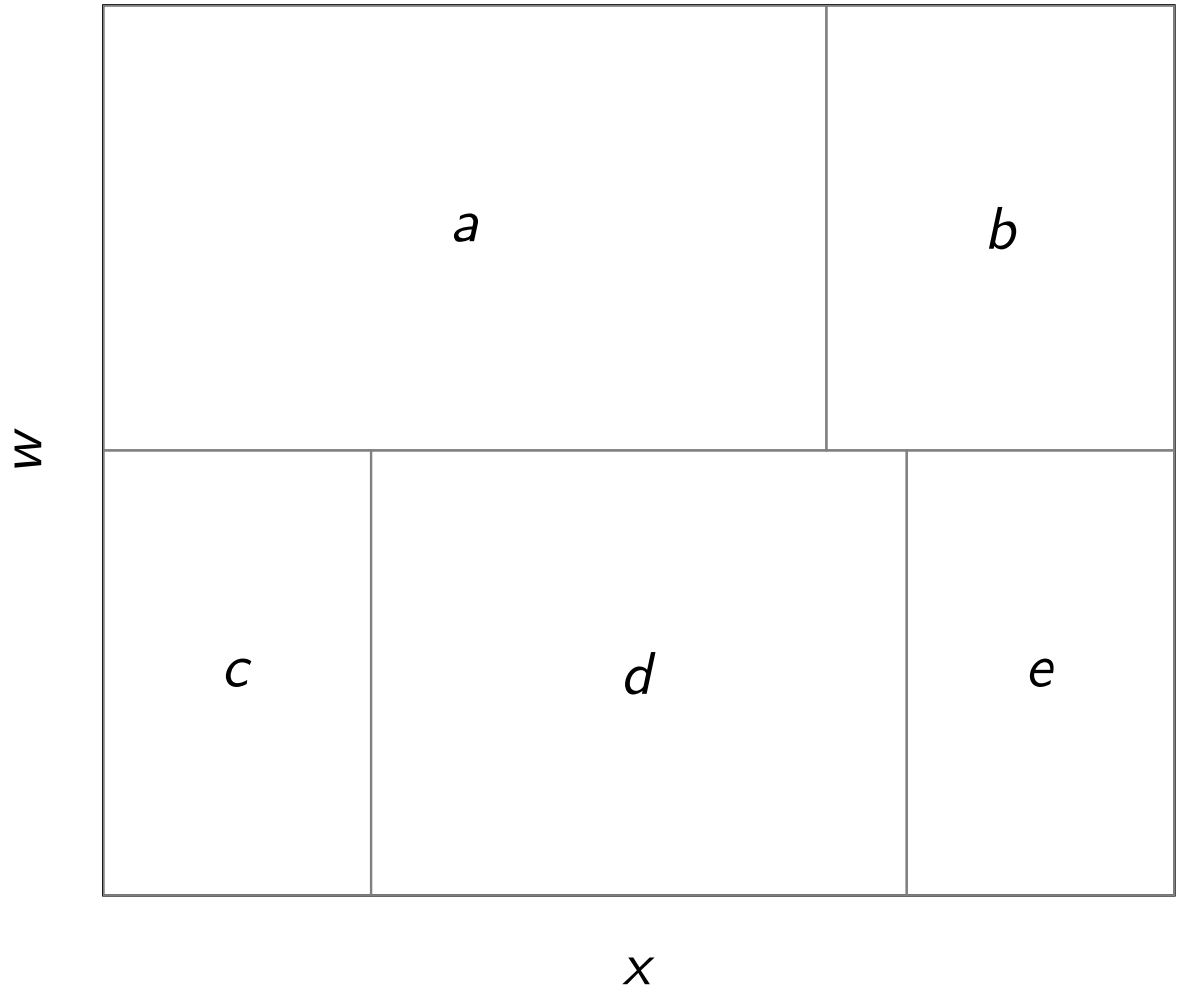
BARDDT - Illustration II



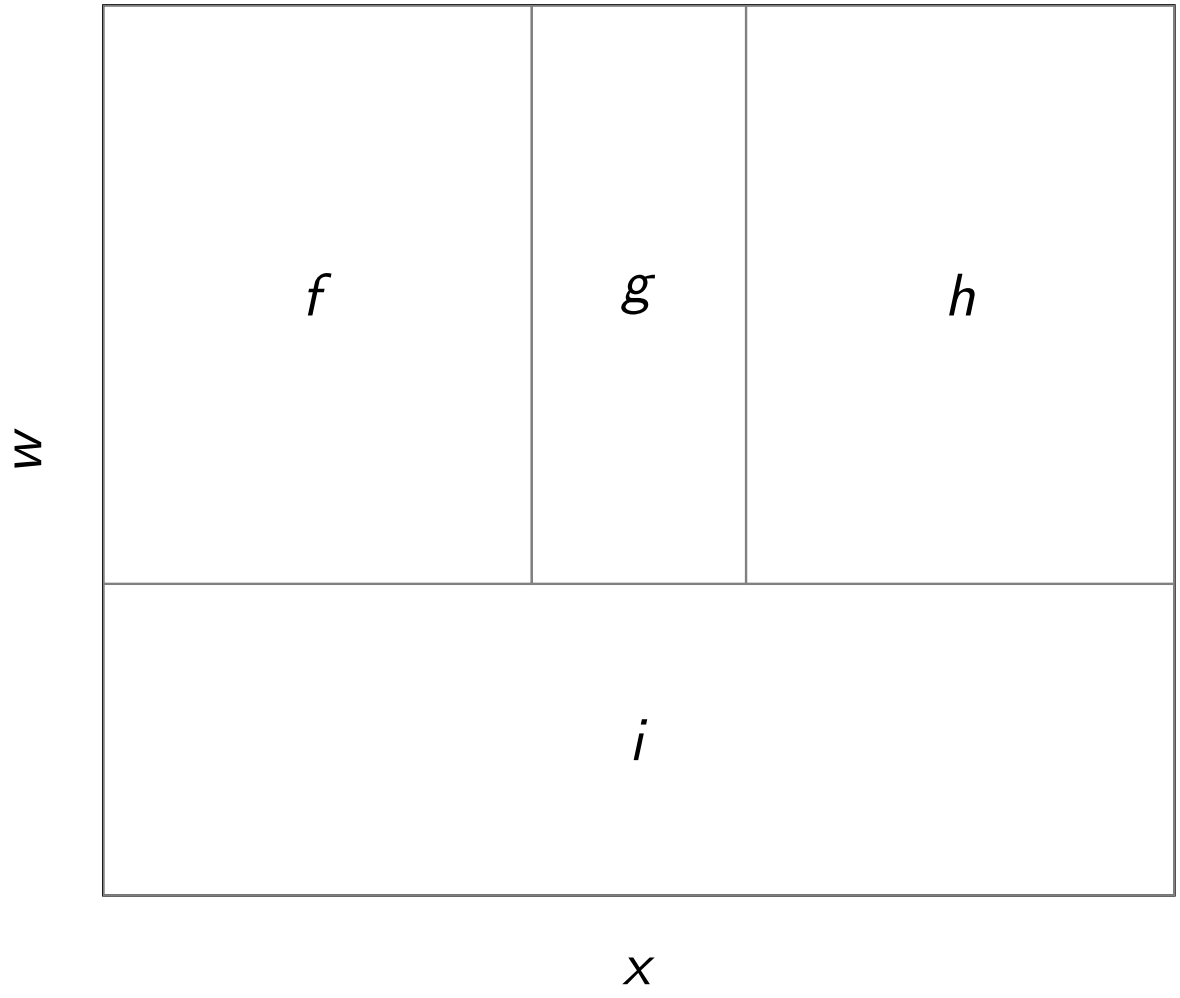
BARDDT - Illustration III

Figure 3: Two regression trees with splits in x and a single scalar w. Node images depict the $g(x, \mathrm{w}, z)$ function (in x) defined by that node's Γ coefficients. The vertical gap between the two line segments in a node that contain $x = 0$ is that node's contribution to the CATE at $X = 0$. Note that only such nodes contribute for CATE prediction at $x = 0$

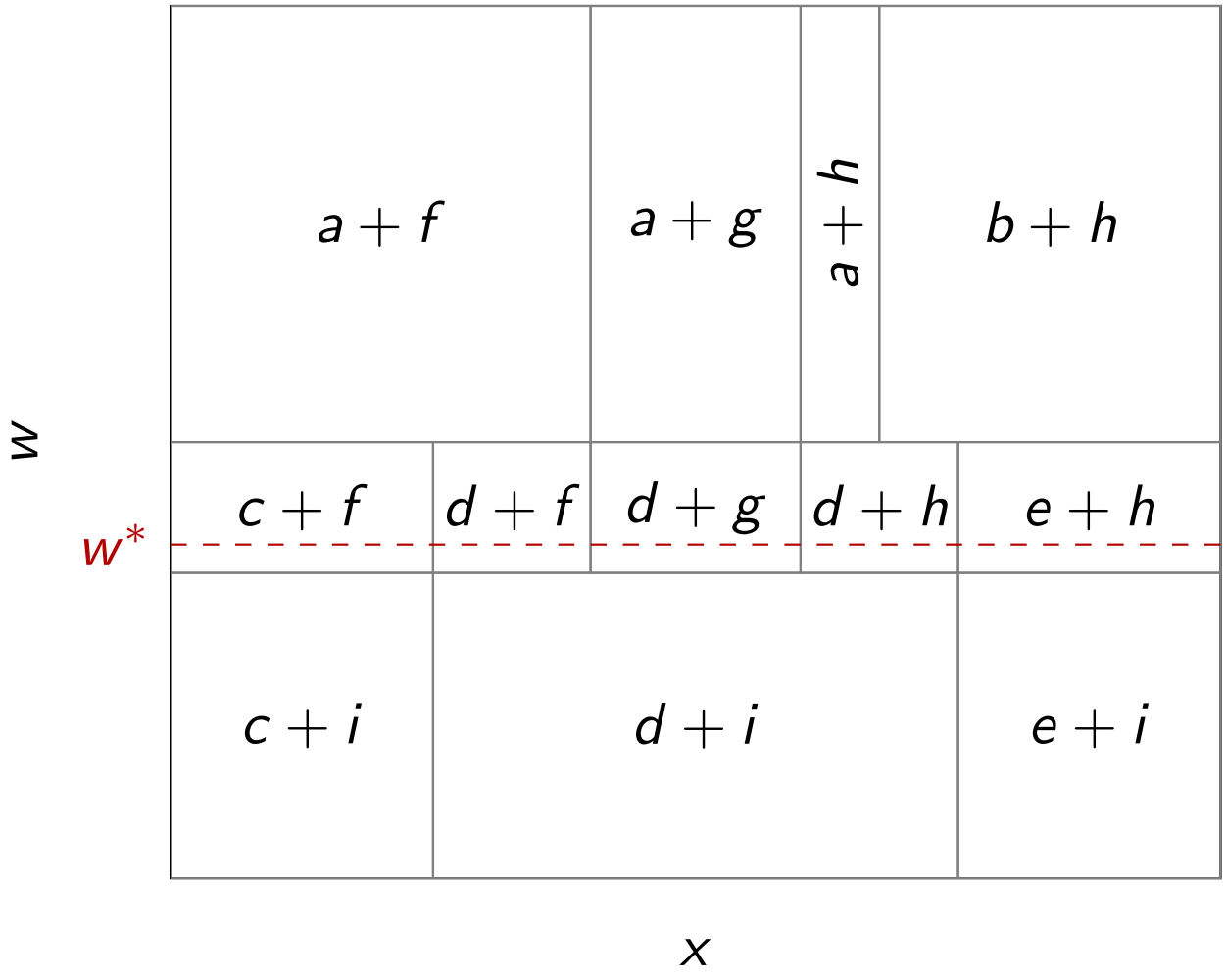
BARDDT - Illustration IV



BARDDT - Illustration V



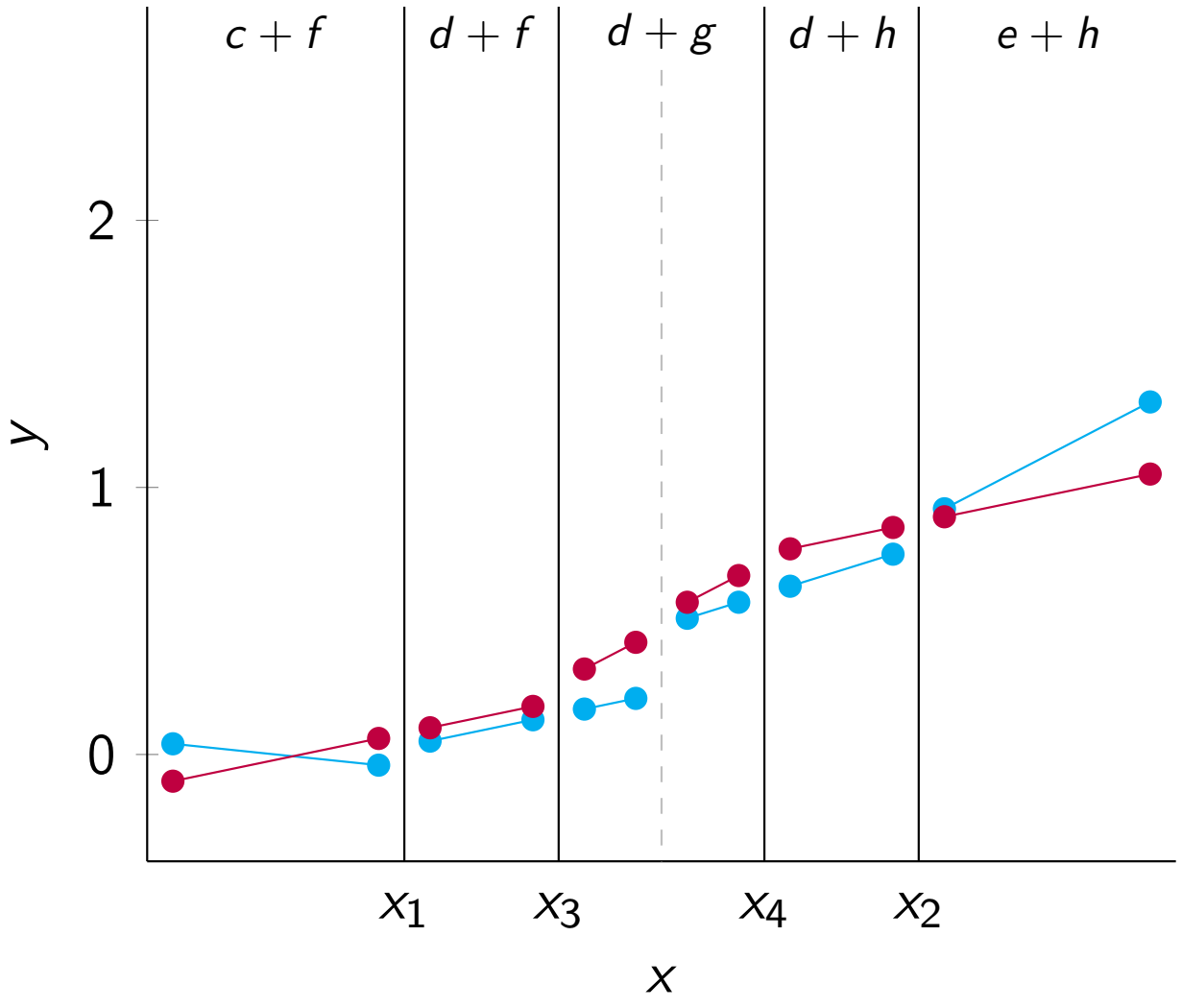
BARDDT - Illustration VI



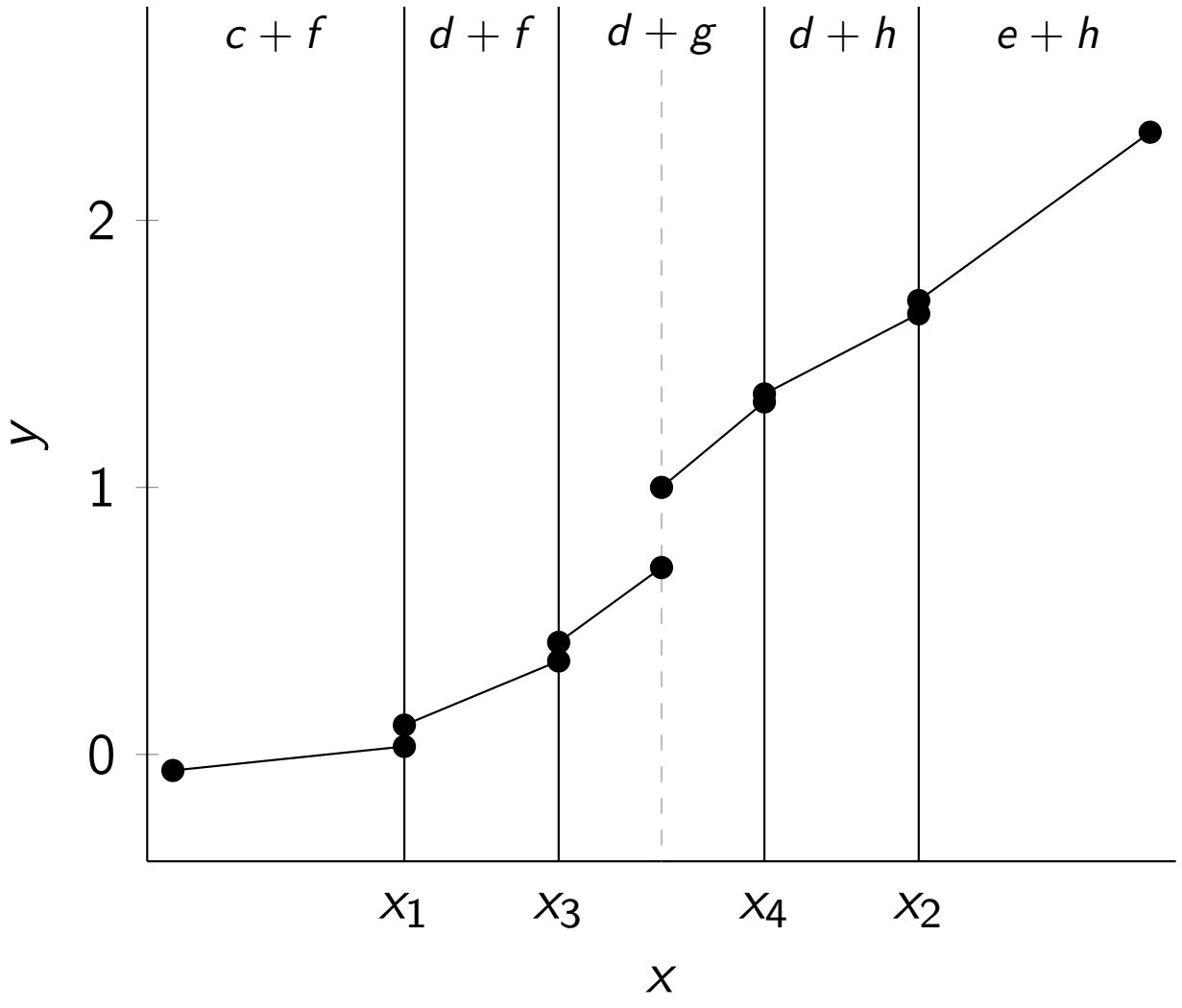
BARDDT - Illustration VII

Figure 4: The two top figures show the same two regression trees as in the preceding figure, now represented as a partition of the x-w plane. Labels in each partition correspond to the leaf nodes depicted in the previous picture. The bottom figure shows the partition of the x-w plane implied by the sum of the two trees; the red dashed line marks point $W = w^*$ and the combination of nodes that include this point

BARDDT - Illustration VIII



BARDDT - Illustration IX



BARDDT - Illustration X

Figure 5: Left: The function fit at $W = w^*$ for the two trees shown in the previous two figures, shown superimposed. Right: The aggregated fit achieved by summing the contributes of two regression tree fits shown at left. The magnitude of the discontinuity at $x = 0$ (located at the dashed gray vertical line) represents the treatment effect at that point. Different values of w will produce distinct fits; for the two trees shown, there can be three distinct fits based on the value of w.

Simulations - Estimand

Generically, our estimand is $\tau(0, \mathrm{w})$. To focus on feasible points, we recommend restricting the evaluation points to the observed w_i such that $|x_i| \leq \delta$, for some $\delta > 0$. Therefore, our estimand of interest is a vector of treatment effects:

$$\tau(0, \mathrm{w}_i) \quad \forall i \text{ such that } |x_i| \leq \delta. \tag{10}$$

In our example, we use $\delta = 0.1$ for a standardized x variable.

Simulations - Estimation loss function

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

$$\text{CATE RMSE} = \frac{\sqrt{\sum_{i:|x_i|\leq\delta} \left(\hat{\tau}(0, \mathrm{w}_i) - \tau(0, \mathrm{w}_i)\right)^2}}{\sqrt{\sum_{i:|x_i|\leq\delta} \left(\hat{\tau}(0) - \tau(0, \mathrm{w}_i)\right)^2}}. \qquad (11)$$

This performance metric judges the ability of $\hat{\tau}(0, \mathrm{w})$ to estimate CATEs relative to a baseline ATE estimator (at $x = 0$)

- Check if methods are doing better than would be possible just by assuming homogeneous effects
- Relative accuracy can be compared in a standardized way across data generating processes of varying outcome scales

DGP

Because an RDD only identifies the treatment effect at $x = 0$, the relevant signal to noise ratios vis-a-vis treatment effect estimation are conditional on $x = 0$

Accordingly, we will design our DGP so that it is explicitly parametrized in terms of conditional variances at $x = 0$

Data will be simulated consistent with the causal diagram in Figure 1: W and X will be generated , followed by Y given W and X.

Generating (W, X)

Our simulation studies will consider W to be fixed in advance and we will consider replications over (X, Y)

We generate W according to a mean-zero multivariate Gaussian distribution with a Toeplitz covariance matrix, with entries ranging from 0 to 2

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

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

where γ_0 is the marginal mean and γ is a p-dimensional vector of regression coefficients

Generating Y given (W, X)

We generate Y according to:

$$Y_i = \mu(x_i, \mathrm{w}_i) + \tau(x_i, \mathrm{w}_i) z_i + \sigma \epsilon_i, \tag{12}$$

for a mean-zero Gaussian error term

We consider the following quantities: $\min_{\mathrm{w}} \tau(0, \mathrm{w})$, $\mathbb{V}(\tau(0, W) \mid X = 0)$, and $\mathbb{V}(\mu(0, W) \mid X = 0)$

We take a large sample from $W \mid X = 0$ and compute the above quantities based on that simulated data. For (W, X) draw as described above, $W \mid X = 0$ is a multivariate Gaussian with

$$\begin{aligned} \mathbb{E}(W \mid X = 0) &= -\gamma_0 \gamma, \\ \mathbb{V}(W \mid X = 0) &= \Sigma_W - \Sigma_W \beta \beta^t \Sigma_W^t. \end{aligned} \tag{13}$$

Generating Y given (W, X)

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

$$\begin{aligned} \sqrt{\mathbb{V}(\tau(0, W) \mid X = 0)} &= k_2, \\ \sqrt{\mathbb{V}(Y \mid X = 0, W)} = \sigma &= k_4 \\ \min_{\mathsf{w}} \tau(0, \mathsf{w}) &= \tau_0 = k_5. \end{aligned} \tag{14}$$

Finally, letting $\mathsf{w} = (w_1 \dots w_p)$ be realizations of a length p random vector W, define $w^\star = \frac{\sum_{j=1}^{p} w_j}{\sqrt{p}}$. Our template functions are:

$$\begin{aligned} \mu^\star(x, \mathsf{w}) &= k_1(x+1)^3 + (w^\star + 2)^2 \left(\text{sign}(x+1)\sqrt{|(x+1)|}\right)^{k_3}, \\ \tau^\star(\mathsf{w}) &= \Phi(2w_1 + 3)/2 + \phi(w_1), \end{aligned} \tag{15}$$

where $\Phi(\cdot)$ and $\phi(\cdot)$ are the cumulative distribution and probability density functions, respectively, of a standard normal random variable

Generating Y given (W, X)

Parameter k_1 controls how much variability of μ will be due to X versus to W and k_3 determines whether or not μ is additive in X and W or if there is an interaction ($k_3 = 1$) or not ($k_3 = 0$)

Methods

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

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

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

Methods

The local polynomial estimator is trained on data points within the bandwidth obtained with the `rdrobust` package (Calonico et al., 2015). Ordinary-least-squares is used to fit a fourth degree polynomial in each feature of W, interacted with X and Z, and an additive third degree polynomial in X:

$$Y \sim \left(\sum_{j=1}^{p} \text{poly}(W_j, 4) \right) \cdot X \cdot Z + \text{poly}(X, 3), \tag{16}$$

where $\text{poly}(A, p) = \sum_{j=1}^{p} A^j$.

Model parameters for RD-Tree were set as suggested in Reguly (2021)

Results

The results are based on configurations of the DGP which can be roughly separated into two groups: "easy" and "hard"

For the easy setting, $(k_1 = 1, k_2 = 1, k_3 = 0, k_4 = 0.1)$: prognostic and treatment variation are comparable magnitudes, μ is separable in x and w, and low noise

For the "hard" setting $(k_1 = 5, k_2 = 0.25, k_3 = 1, k_4 = 0.5)$: prognostic variation is twenty times larger than treatment variation, μ is non-separable in x and w, and noise is high

Results are based on 100 replications of size $n = 4000$ for each DGP configuration.

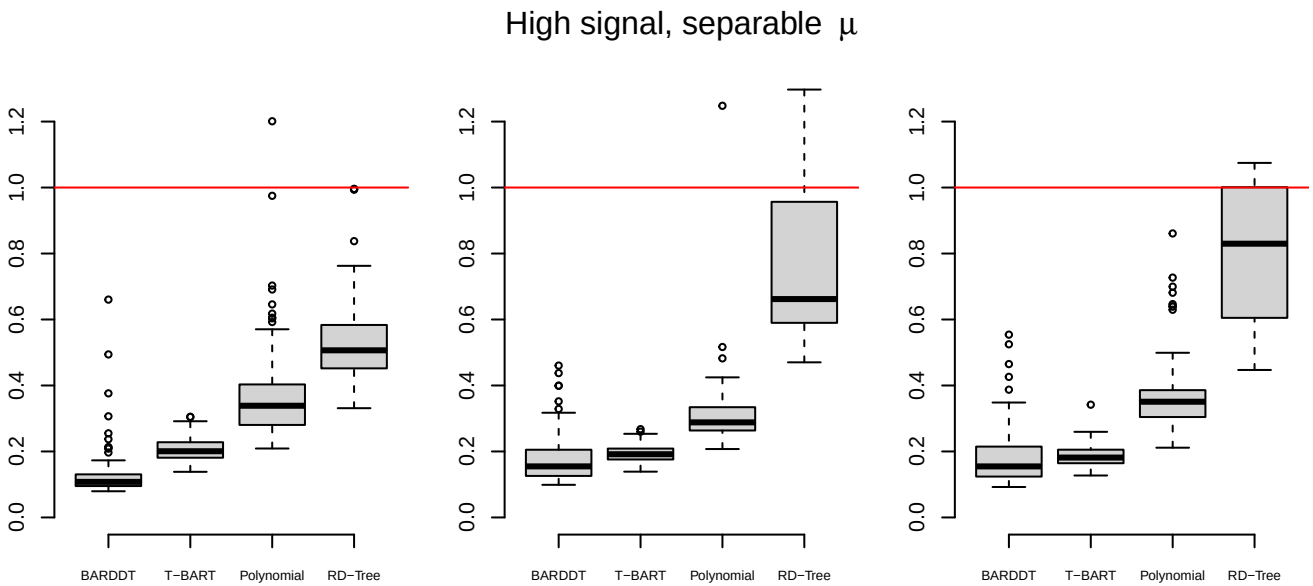
Summary of Results

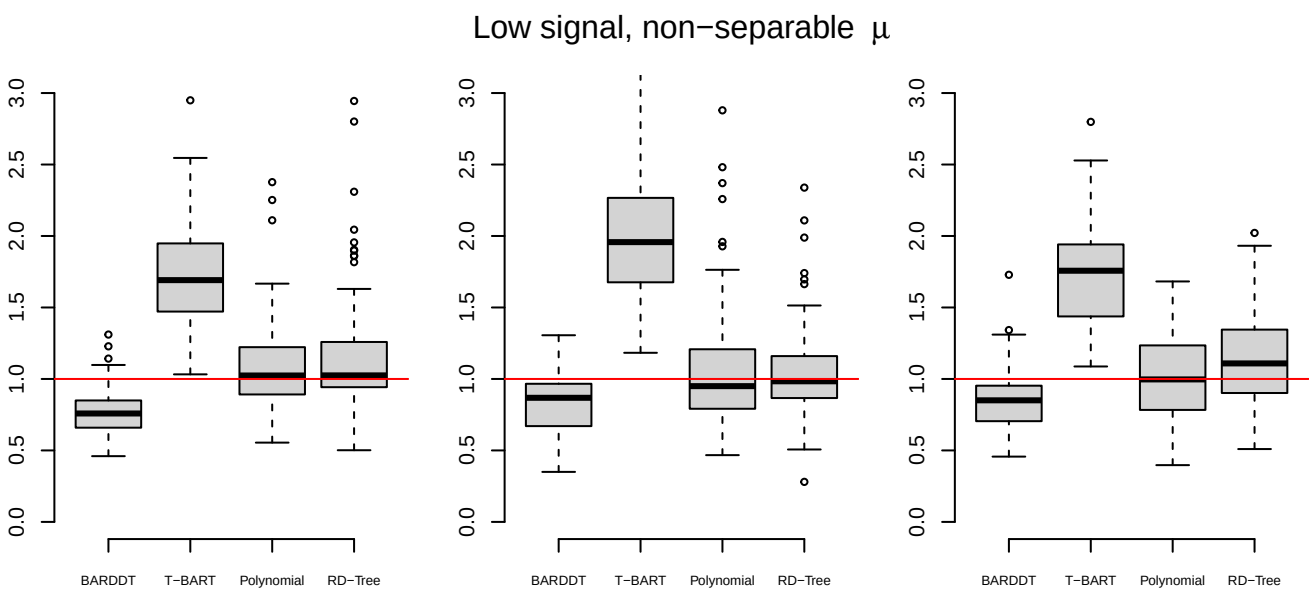
- Although T-BART performs well under the easier regime, even there it still exhibits high variance CATE estimates. T-BART's high variance becomes more pronounced under the harder DGP, resulting in substantially higher RMSE relative to both BARDDT and the polynomial model.
- The extreme bias shift exhibited by S-BART in the low noise setting is reminiscent of the regularization-induced confounding (RIC) problem, described by Hahn et al. (2020). Broadly, the lesson here is that S-BART has unpredictable biases in causal inference problems. It does comparatively well in the high noise case, but only because it rarely splits in that case, collapsing to a homogenous treatment model, which outperforms the overfitting T-BART and polynomial models in this regime.

Summary of Results

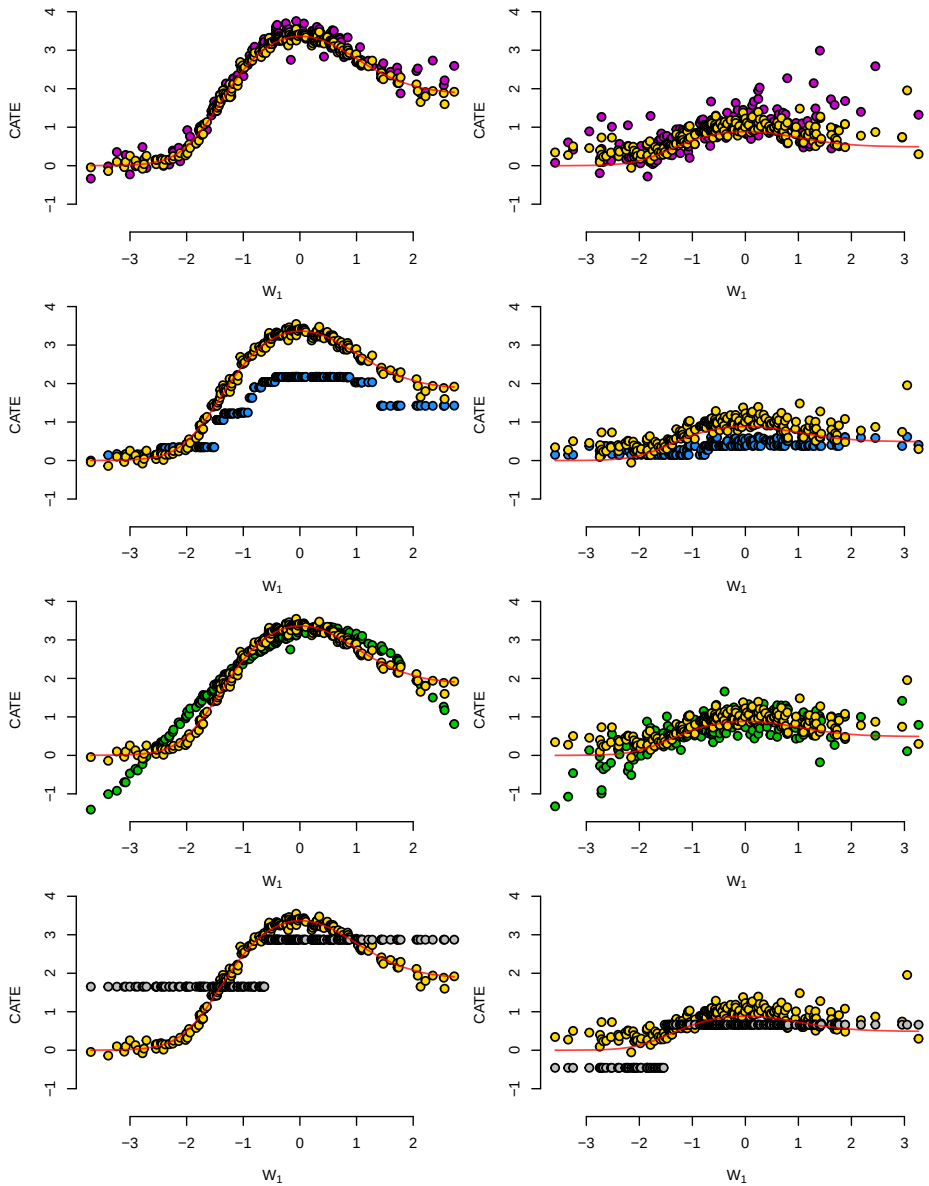
- The fits for the easier setup show that, even with high signal, the polynomial model struggles with extrapolation at the boundaries of the support of w_1. At the same time, the polynomial model also presents a sizable increase in variance under high noise, as seen on the fits for the harder regime.
- RD-Tree appears to "under-split" on W, leading to a too-coarse fit of the CATE function, especially in the low-noise regime. This behavior is to be expected with a single CART fit, a problem that additive tree models, like BART, were explicitly designed to address.

Overall Results





Individual Fits



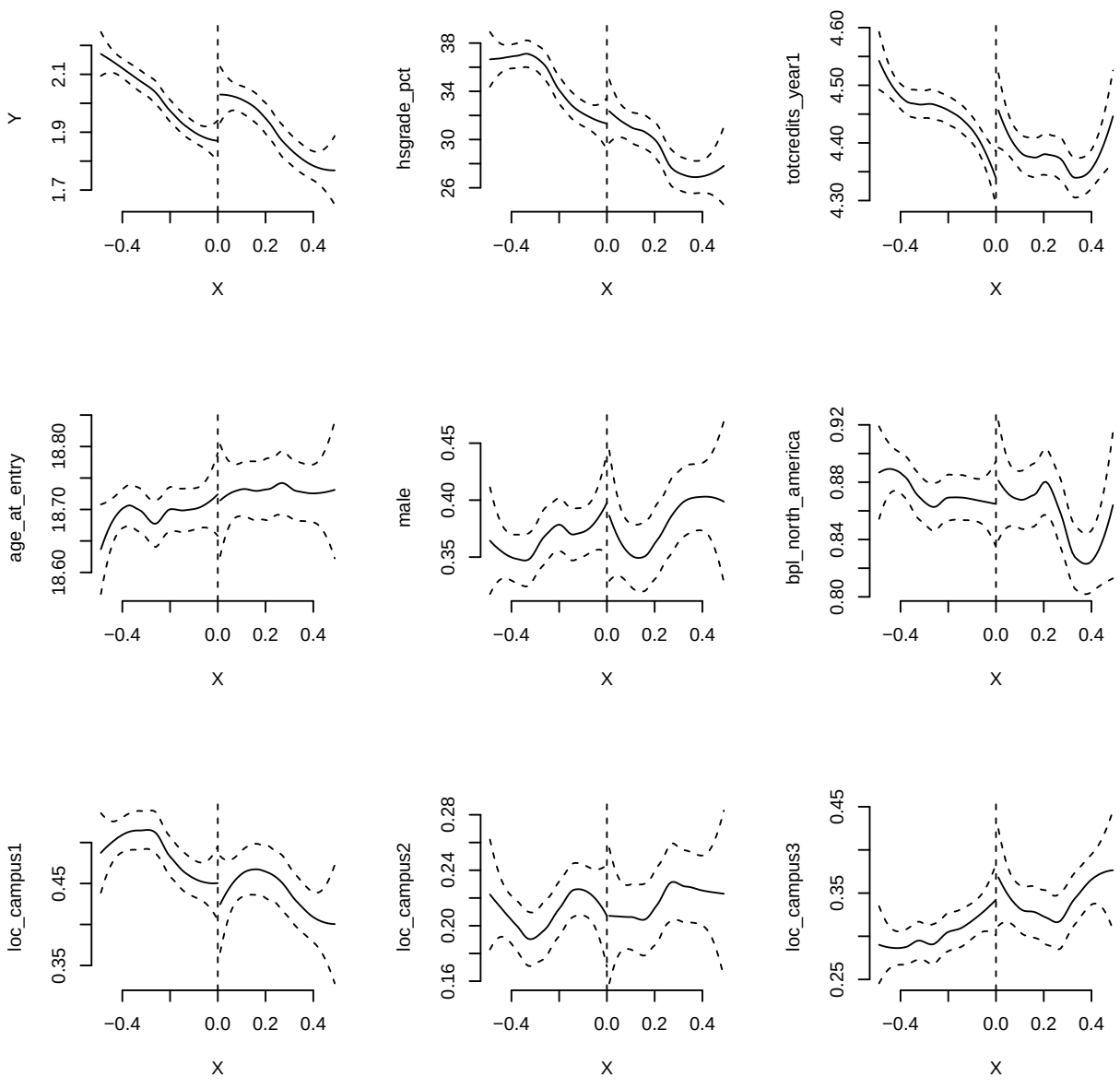
Application: effect of academic probation on education

- 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: fit-the-fit

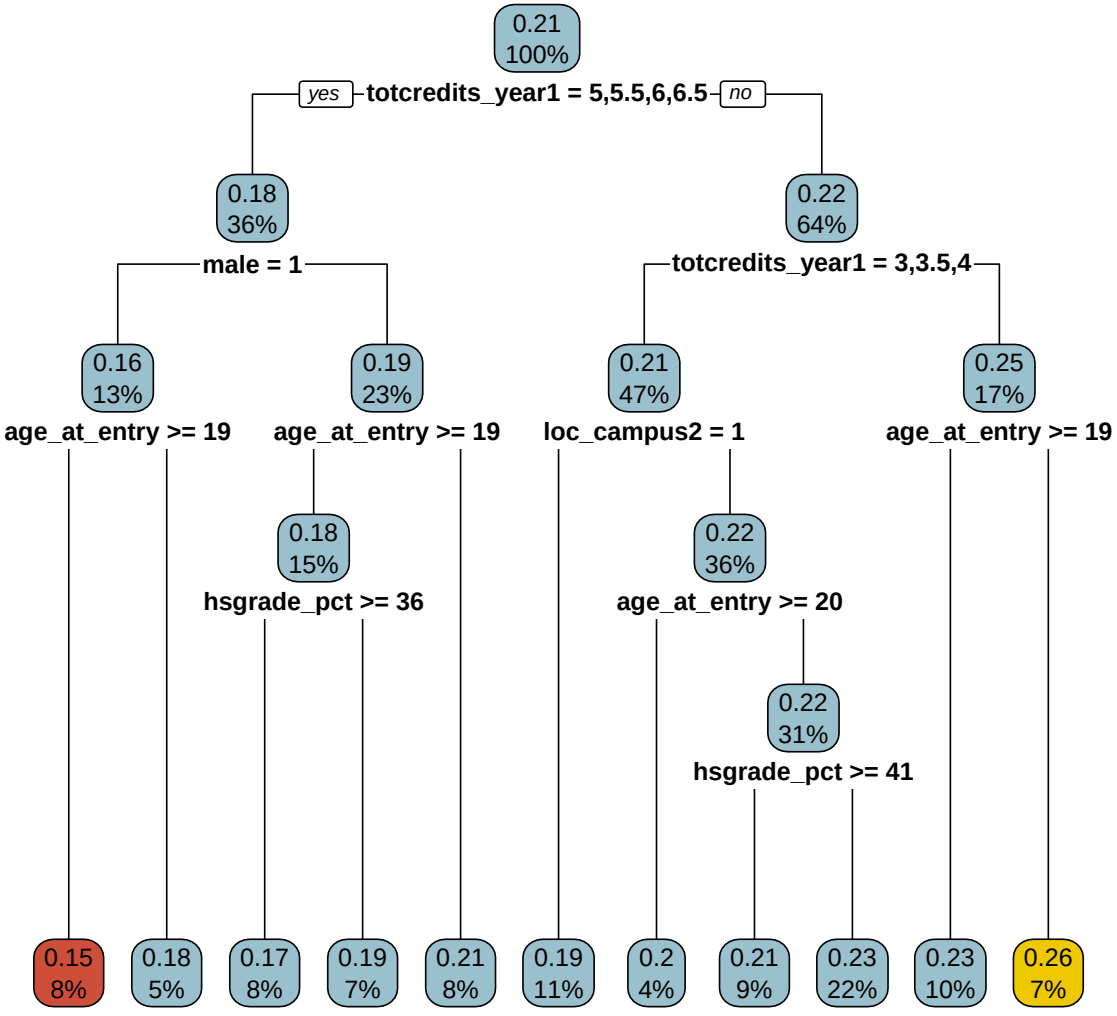
- As in Hahn et al. (2020), we explore the individual effect estimates – the posterior mean of the individual effects – by fitting a CART tree to these estimates based on the covariate set ('fit-the-fit')

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

Application: fit-the-fit I

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 full sample summary flags high-school performance, birth place, gender, campus location and credits in first year as important moderators; the separate campus fits indicate heterogeneity between the campuses: for campus 1, high-school performance, credits attempted and birth place are flagged as important moderators, while for campus 2, high-school performance and gender are important and, for campus 3, gender, birth place and age at entry are the important moderators

Application: fit-the-fit II



Application: fit-the-fit

Consider the two groups of students at opposite ends of the treatment effect range discovered by the effect moderation tree:

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

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

Application: fit-the-fit

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

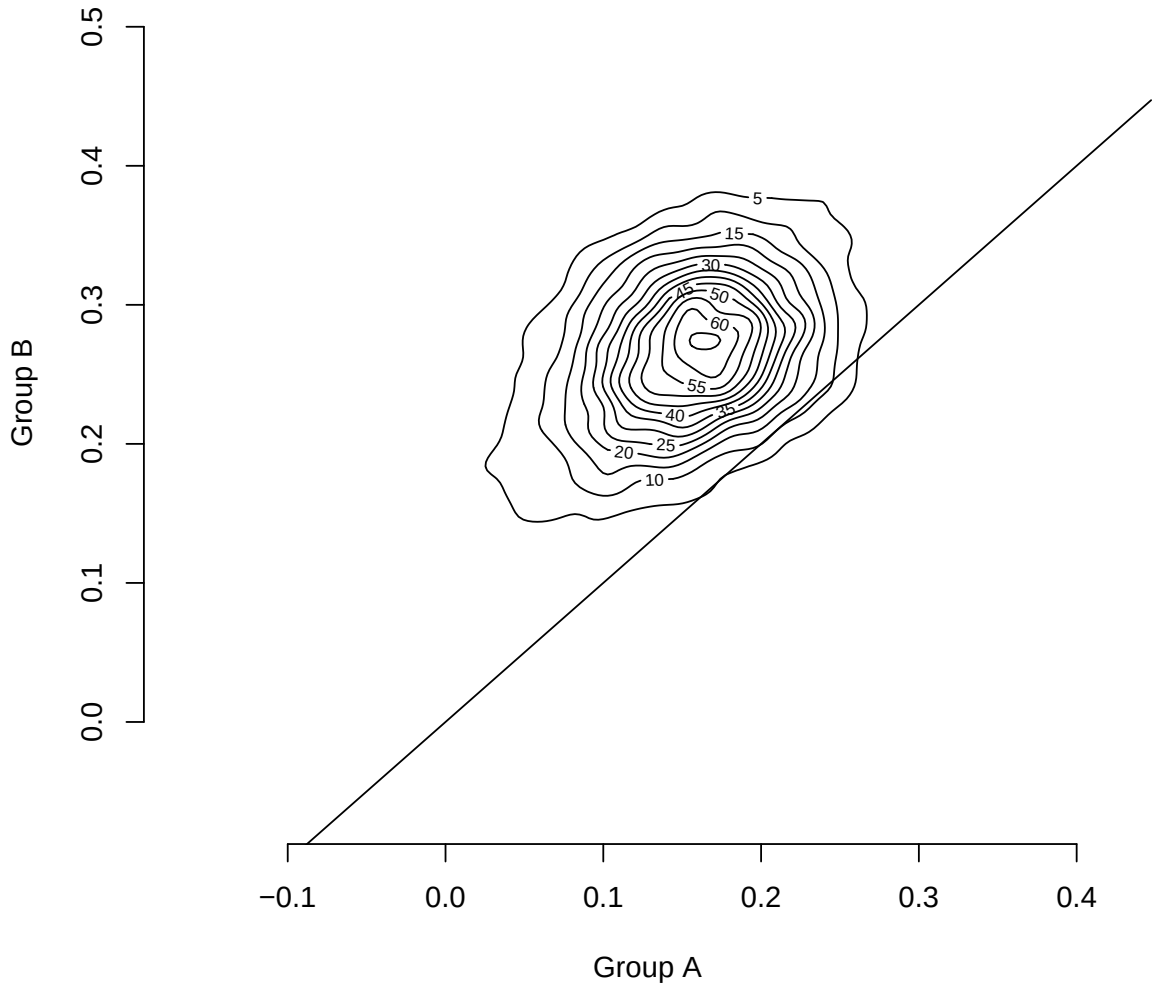
$$\bar{\tau}_A^{(h)} = \frac{1}{n_A} \sum_{i:w_i} \tau^{(h)}(0, w_i), \tag{17}$$

where h indexes a posterior draw and n_A denotes the number of individuals in the group A.

Application - Posterior Comparisons I

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, 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

Application - Posterior Comparisons II



Conclusion

- **Main contributions:** incorporating RDD assumptions into the BART framework and producing reliable CATE estimates
- **Results:** BARDDT presents lower errors and more stable performance compared to other commonly used estimators based on parametric specifications, and to off-the-shelf BART implementations
- **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., and Titiunik, R. (2015). rdrobust: An r package for robust nonparametric inference in regression-discontinuity designs. *R J.*, 7(1):38.

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.

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

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*.

Final references II

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

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*.

Final references III

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.