Bayesian Analysis (2020) **15**, Number 3, pp. 965–1056

Bayesian Regression Tree Models for Causal Inference: Regularization, Confounding, and Heterogeneous Effects (with Discussion)*

P. Richard Hahn[†], Jared S. Murray[‡], and Carlos M. Carvalho[§]

Abstract. This paper presents a novel nonlinear regression model for estimating heterogeneous treatment effects, geared specifically towards situations with small effect sizes, heterogeneous effects, and strong confounding by observables. Standard nonlinear regression models, which may work quite well for prediction, have two notable weaknesses when used to estimate heterogeneous treatment effects. First, they can yield badly biased estimates of treatment effects when fit to data with strong confounding. The Bayesian causal forest model presented in this paper avoids this problem by directly incorporating an estimate of the propensity function in the specification of the response model, implicitly inducing a covariate-dependent prior on the regression function. Second, standard approaches to response surface modeling do not provide adequate control over the strength of regularization over effect heterogeneity. The Bayesian causal forest model permits treatment effect heterogeneity to be regularized separately from the prognostic effect of control variables, making it possible to informatively "shrink to homogeneity". While we focus on observational data, our methods are equally useful for inferring heterogeneous treatment effects from randomized controlled experiments where careful regularization is somewhat less complicated but no less important. We illustrate these benefits via the reanalysis of an observational study assessing the causal effects of smoking on medical expenditures as well as extensive simulation studies.

Keywords: Bayesian, causal inference, heterogeneous treatment effects, predictor-dependent priors, machine learning, regression trees, regularization, shrinkage.

MSC2020 subject classifications: Primary 62-07, 62J02; secondary 62F15.

1 Introduction

The success of modern predictive modeling is founded on the understanding that flexible predictive models must be carefully regularized in order to achieve good out-of-sample performance (low generalization error). In a causal inference setting, regularization is less straightforward, for (at least) two reasons. One, in the presence of confounding,

*A previous version of the manuscript included a Contributed Discussion by Kolyan Ray, Botond Szabo, and Aad van der Vaart that has been updated after publication to ensure correspondence of the methods used by the authors in their Discussion with the ones used in the main manuscript.

[†]School of Mathematical and Statistical Sciences, Arizona State University, prhahn@asu.edu
[‡]McCombs School of Business, University of Texas at Austin, jared.murray@mccombs.utexas.edu
[§]McCombs School of Business, University of Texas at Austin, carlos.carvalho@mccombs.utexas.edu

 https://doi.org/10.1214/19-BA1195

regularized models originally designed for prediction can bias causal estimates towards some unknown function of high dimensional nuisance parameters (Hahn et al., 2016). Two, when the magnitude of response surface variation due to prognostic effects differs markedly from response surface variation due to treatment effect heterogeneity, simple regularization strategies, which may be adequate for good out-of-sample prediction, provide inadequate control of estimator variance for conditional average treatment effects (leading to large estimation error).

To mitigate these two estimation issues we propose a flexible sum-of-regression-trees — a *forest* — to model a response variable as a function of a binary treatment indicator and a vector of control variables. To address the first issue, we develop a novel prior for the response surface that depends explicitly on estimates of the propensity score as an important 1-dimensional transformation of the covariates (including the treatment assignment). Incorporating this transformation of the covariates is not strictly necessary in response surface modeling in order to obtain consistent estimators, but we show that it can substantially improve treatment effect estimation in the presence of moderate to strong confounding, especially when that confounding is driven by targeted selection — individuals selecting into treatment based on somewhat accurate predictions of the potential outcomes.

To address the second issue, we represent our regression as a sum of two functions: the first models the *prognostic* impact of the control variables (the component of the conditional mean of the response that is unrelated to the treatment effect), while the second represents the treatment effect directly, which itself is a nonlinear function of the observed attributes (capturing possibly heterogeneous effects). We represent each function as a forest. This approach allows the degree of shrinkage on the treatment effect to be modulated *directly* and *separately* of the prognostic effect. In particular, under this parametrization, standard regression tree priors shrink towards homogeneous effects.

In most previous approaches, the prior distribution over treatment effects is induced indirectly, and is therefore difficult to understand and control. Our approach interpolates between two extremes: Modeling the conditional means of treated and control units entirely separately, or including treatment assignment as "just another covariate". The former precludes any borrowing or regularization entirely, while the second can be rather difficult to understand using flexible models. Parametrizing non- and semiparametric models this way is attractive regardless of the specific priors in use.

Comparisons on simulated data show that the new model — which we call the Bayesian causal forest model — performs at least as well as existing approaches for estimating heterogenous treatment effects across a range of plausible data generating processes. More importantly, it performs dramatically better in many cases, especially those with strong confounding, targeted selection, and relatively weak treatment effects, conditions we believe to be common in applied settings.

In Section 7, we demonstrate how our flexible Bayesian model allows us to make rich inferences on heterogeneous treatment effects, including estimates of average and conditional average treatment effects at various levels, in a re-analysis of data from an observational study of the effect of smoking on medical expenditures.

1.1 Relationship to previous literature

As previously noted, the Bayesian causal forest model directly extends ideas from two earlier papers: Hill (2011) and Hahn et al. (2016). Specifically, this paper studies the "regularization-induced confounding" of Hahn et al. (2016) in the context of Bayesian additive regression tree (BART) models as utilized by Hill (2011). In terms of implementation, this paper builds explicitly on the work of Chipman et al. (2010); see also Gramacy and Lee (2008) and Murray (2017). Other notable work on Bayesian treatment effect estimation includes Gustafson and Greenland (2006), Zigler and Dominici (2014), Heckman et al. (2014), Li and Tobias (2014), Roy et al. (2017) and Taddy et al. (2016).

More generally, the intersection between "machine learning" and causal inference is a burgeoning research area.

Papers deploying nonlinear regression (supervised learning) methods in the service of estimation and inference for average treatment effects (ATEs) include targeted maximum likelihood estimation (TMLE) (van der Laan, 2010a,b), double machine learning (Chernozhukov et al., 2016), and generalized boosting (McCaffrey et al., 2004, 2013). These methods all take as inputs regression estimates of either the propensity function or the response surface (or both); in this sense, any advances in predictive modeling have the potential to improve ATE estimation in conjunction with the above approaches. Bayesian causal forests could be used in this capacity as well, although it was designed with conditional average treatment effects (CATEs) in mind.

More recently, attention has turned to CATE estimation. Notable examples include Taddy et al. (2016), who focus on estimating heterogeneous effects from experimental, as opposed to observational data, which is our focus. Su et al. (2012) approach CATE estimation with regression tree ensembles and are in that sense a forerunner of both Bayesian causal forests as well as Wager and Athey (2018),[1] Athey et al. (2019) and Powers et al. (2018). Wager and Athey (2018) is notable for providing the first inferential theory for CATE estimators arising from a random forests representation, based on the infinitesimal jackknife (Efron, 2014; Wager et al., 2014). Friedberg et al. (2018) extend this approach to locally linear forests. Nie and Wager (2017) and Künzel et al. (2019) propose stacking and meta-learning methods, respectively, similar to what TMLE does for the ATE, except tailored to CATE estimation. Shalit et al. (2017) develop a neural network-based estimator of CATEs based on a bound of the generalization error in an approach inspired by domain adaptation (Ganin et al., 2016). Zaidi and Mukherjee (2018) develop a model based on the use of Gaussian processes to directly model the special transformed response (as studied in Athey et al. (2019) and Powers et al. (2018)).

The focus of the present paper is to develop a regularization prior for nonlinear models geared specifically towards situations with small effect sizes, heterogeneous effects,

[1]Note that the Bayesian causal forest model is not the Bayesian analogue of the causal random forest method, as both the motivation and fitting process are quite different; both are tree-based methods for estimating conditional average treatment effects (CATEs), but the similarities end there. Specifically, Chipman et al. (2010) is already substantially different than Breiman (2001), and the ways that BCF modifies BART are simply not analogous to the modifications that causal random forests makes to random forests.

and strong confounding. The research above does not focus specifically on this regime, which is an important one in applied settings.

Finally there are a number of papers that compare and contrast the above methods on real and synthetic data: Wendling et al. (2018), McConnell and Lindner (2019), Dorie and Hill (2017), Dorie et al. (2019), and Hahn et al. (2018). The results of these studies will be discussed in some detail later, but a general trend is that BART-based methods appear to be a strong default choice for heterogeneous effect modeling.

2 Problem statement and notation

Let Y denote a scalar response variable and Z denote a binary treatment indicator variable. Capital Roman letters denote random variables, while realized values appear in lower case, that is, y and z. Let x denote a length d vector of observed control variables. Throughout, we will consider an observed sample of size n independent observations (Y_i, Z_i, x_i), for $i = 1, \ldots n$. When Y or Z (respectively, y or z) are without a subscript, they denote length n column vectors; likewise, $\mathbf{X}$ will denote the $n \times d$ matrix of control variables.

We are interested in estimating various treatment effects. In particular, we are interested in conditional average treatment effects — the amount by which the response Y_i would differ between hypothetical worlds in which the treatment was set to $Z_i = 1$ versus $Z_i = 0$, averaged across subpopulations defined by attributes x. This kind of counterfactual estimand can be formalized in the potential outcomes framework (Imbens and Rubin (2015), chapter 1) by using $Y_i(0)$ and $Y_i(1)$ to denote the outcomes we would have observed if treatment were set to zero or one, respectively. We make the stable unit treatment value assumption (SUTVA) throughout (excluding interference between units and multiple versions of treatment (Imbens and Rubin, 2015)). We observe the potential outcome that corresponds to the realized treatment: $Y_i = Z_i Y_i(1) + (1 - Z_i) Y_i(0)$.

Throughout the paper we will assume that *strong ignorability* holds, which stipulates that

$$Y_i(0), Y_i(1) \perp\!\!\!\perp Z_i \mid \mathbf{X}_i, \tag{2.1}$$

and also that

$$0 < \Pr(Z_i = 1 \mid \mathrm{x}_i) < 1 \tag{2.2}$$

for all $i = 1, \ldots, n$. The first condition assumes we have no unmeasured confounders, and the second condition (overlap) is necessary to estimate treatment effects everywhere in covariate space. Provided that these conditions hold, it follows that $\mathrm{E}(Y_i(z) \mid \mathrm{x}_i) = \mathrm{E}(Y_i \mid \mathrm{x}_i, Z_i = z)$ so our estimand may be expressed as

$$\tau(\mathrm{x}_i) := \mathrm{E}(Y_i \mid \mathrm{x}_i, Z_i = 1) - \mathrm{E}(Y_i \mid \mathrm{x}_i, Z_i = 0). \tag{2.3}$$

For simplicity, we restrict attention to mean-zero additive error representations

$$Y_i = f(\mathrm{x}_i, Z_i) + \epsilon_i, \quad \epsilon_i \sim N(0, \sigma^2) \tag{2.4}$$

so that $\mathrm{E}(Y_i \mid \mathrm{x}_i, Z_i = z_i) = f(\mathrm{x}_i, z_i)$. In this context, (2.1) can be expressed equivalently as $\epsilon_i \perp\!\!\!\perp Z_i \mid \mathrm{x}_i$. The treatment effect of setting $z_i = 1$ versus $z_i = 0$ can therefore be

expressed as

$$\tau(\mathrm{x}_i) := f(\mathrm{x}_i, 1) - f(\mathrm{x}_i, 0).$$

Our contribution in this paper is a careful study of prior specification for f. We propose new prior distributions that improve estimation of the parameter of interest, namely τ. Previous work (Hill, 2011) advocated using a BART prior for $f(\mathrm{x}_i, z_i)$ directly. We instead recommend expressing the response surface as

$$\mathrm{E}(Y_i \mid \mathrm{x}_i, Z_i = z_i) = \mu(\mathrm{x}_i, \hat{\pi}(\mathrm{x}_i)) + \tau(\mathrm{x}_i) z_i, \tag{2.5}$$

where the functions μ and τ are given independent BART priors and $\hat{\pi}(\mathrm{x}_i)$ is an estimate of the propensity score $\pi(\mathrm{x}_i) = \Pr(Z_i = 1 \mid \mathrm{x}_i)$. The following sections motivate this model specification and provide additional context; further modeling details are given in Section 5.

3 Bayesian additive regression trees for heterogeneous treatment effect estimation

Hill (2011) observed that under strong ignorability, treatment effect estimation reduces to response surface estimation. That is, provided that a sufficiently rich collection of control variables are available (to ensure strong ignorability), treatment effect estimation can proceed "merely" by estimating the conditional expectations $\mathrm{E}(Y \mid \mathrm{x}, Z = 1)$ and $\mathrm{E}(Y \mid \mathrm{x}, Z = 0)$. Noting its strong performance in prediction tasks, Hill (2011) advocates the use of the Bayesian additive regression tree model of Chipman et al. (2010) for estimating these conditional expectations.

BART is particularly well-suited to detecting interactions and discontinuities, can be made invariant to monotone transformations of the covariates, and typically requires little parameter tuning. Chipman et al. (2010) provide extensive evidence of BART's excellent predictive performance. BART has also been used successfully for applications in causal inference, for example Green and Kern (2012), Hill et al. (2013), Kern et al. (2016), and Sivaganesan et al. (2017). It has subsequently been demonstrated to successfully infer heterogeneous and average treatment effects in multiple independent simulation studies (Dorie et al., 2019; Wendling et al., 2018), frequently outperforming competitors (and never lagging far behind).

3.1 Specifying the BART prior

The BART prior expresses an unknown function $f(\mathrm{x})$ as a sum of many piecewise constant binary regression trees. (In this section, we suppress z in the notation; implicitly z may be considered as a coordinate of x.) Each tree T_l, $1 \leq l \leq L$, consists of a set of internal decision nodes which define a partition of the covariate space (say $\mathcal{A}_1, \ldots, \mathcal{A}_{B(l)}$), as well as a set of terminal nodes or leaves corresponding to each element of the partition. Further, each element of the partition $\mathcal{A}_b$ is associated a parameter value, m_{lb}. Taken together the partition and the leaf parameters define a piecewise constant function: $g_l(x) = m_{lb}$ if $x \in \mathcal{A}_b$; see Figure 1.

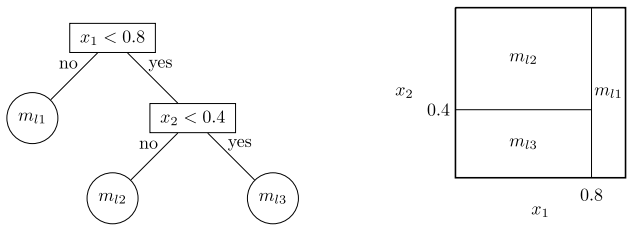


Figure 1: (Left) An example binary tree, with internal nodes labelled by their splitting rules and terminal nodes labelled with the corresponding parameters m_{lb}. (Right) The corresponding partition of the sample space and the step function.

Individual regression trees are then additively combined into a single regression *forest*: $f(\mathrm{x}) = \sum_{l=1}^{L} g_l(\mathrm{x})$. Each of the functions g_l are constrained by their prior to be "weak learners" in the sense that the prior favors small trees and leaf parameters that are near zero. Each tree follows (independently) the prior described in Chipman et al. (1998): the probability that a node at depth h splits is given by $\eta(1+h)^{-\beta}$, $\eta \in (0,1)$, $\beta \in [0,\infty)$.

A variable to split on, as well as a cut-point to split at, are then selected uniformly at random from the available splitting rules. Large, deep trees are given extremely low prior probability by taking $\eta = 0.95$ and $\beta = 2$ as in Chipman et al. (2010). The leaf parameters are assigned independent priors $m_{lb} \sim \mathrm{N}(0, \sigma_m^2)$ where $\sigma_m = \sigma_0/\sqrt{L}$. The induced marginal prior for $f(\mathrm{x})$ is centered at zero and puts approximately 95% of the prior mass within $\pm 2\sigma_0$ (pointwise), and σ_0 can be used to calibrate the plausible range of the regression function. Full details of the BART prior and its implementation are given by Chipman et al. (2010).

In a causal inference context, we are concerned with the impact that the prior over $f(\mathrm{x}, z)$ has on estimating $\tau(\mathrm{x}) = f(\mathrm{x}, 1) - f(\mathrm{x}, 0)$. The choice of BART as a prior over f has particular implications for the induced prior on τ that are difficult to understand: In particular, the induced prior will vary with the dimension of x and the degree of dependence with z. In Section 5 we propose an alternative parameterization that mitigates this problem. But first, the next section develops a more general framework for investigating the influence of prior specification and regularization on treatment effect estimates.

4 The central role of the propensity score in regularized causal modeling

In this section we explore the joint impacts of regularization and confounding on estimation of heterogeneous treatment effects. We find that including an estimate of the propensity score as a covariate reduces the bias of regularized treatment effect estimates in finite samples. We recommend including an estimated propensity score as a covariate as routine practice regardless of the particular models or algorithms used to estimate treatment effects since regularization is necessary to estimate heterogeneous treatment

effects non- or semiparametrically or in high dimensions. To illustrate the potential for biased estimation and motivate our fix, we introduce two key concepts: Regularization induced confounding and targeted selection.

4.1 Regularization-induced confounding

Since treatment effects may be deduced from the conditional expectation function $f(\mathrm{x}_i, z_i)$, a likelihood perspective suggests that the conditional distribution of Y given x and Z is sufficient for estimating treatment effects. While this is true in terms of *identification* of treatment effects, the question of estimation with finite samples is more nuanced. In particular, many functions in the support of the prior will yield approximately equivalent likelihood evaluations, but may imply substantially different treatment effects. This is particularly true in a strong confounding-modest treatment effect regime, where the conditional expectation of Y is largely determined by x rather than Z.

Accordingly, the posterior estimate of the treatment effect is apt to be substantially influenced by the prior distribution over f for realistic sample sizes. This issue was explored by Hahn et al. (2016) in the narrow context of linear regression with continuous treatment and homogenous treatment effect; they call this phenomenon "regularization-induced confounding" (RIC). In the linear regression setting an exact expression for the bias on the treatment effect under standard regularization priors is available in closed form.

Example: RIC in the linear model

Suppose the treatment effect is homogenous and response and treatment model are both linear:

$$\begin{aligned} Y_i &= \tau Z_i + \beta^t \mathrm{x}_i + \varepsilon_i, \\ Z_i &= \gamma^t \mathrm{x}_i + \nu_i; \end{aligned} \tag{4.1}$$

where the error terms are mean zero Gaussian and a multivariate Gaussian prior is placed over all regression coefficients. The Bayes estimator under squared error loss is the posterior mean, so we examine the expression for the bias of $\hat{\tau}_{rr} \equiv \mathrm{E}(\tau \mid Y, \mathrm{z}, \mathbf{X})$. We begin from a standard expression for the bias of the ridge estimator, as given, for example, in Giles and Rayner (1979). Write $\theta = (\tau, \beta^t)^t$, $\tilde{\mathbf{X}} = \begin{pmatrix} \mathrm{z} & \mathbf{X} \end{pmatrix}$ and let $\theta \sim \mathrm{N}(0, \mathbf{M}^{-1})$. Then the bias of the Bayes estimator is

$$\mathrm{bias}(\hat{\theta}_{rr}) = -(\mathbf{M} + \tilde{\mathbf{X}}^t \tilde{\mathbf{X}})^{-1} \mathbf{M} \theta, \tag{4.2}$$

where the bias expectation is taken over Y, conditional on $\mathbf{X}$ and all model parameters.

Consider $M = \left(\begin{smallmatrix} 0 & 0 \\ 0 & \mathrm{I}_p \end{smallmatrix}\right)$, where I_p denotes a p-by-p identity matrix, which corresponds to a ridge prior (with ridge parameter $\lambda = 1$ for simplicity) on the control variables and a non-informative "flat" prior over the first element (τ, the treatment effect). Plugging this into the bias equation (4.2) and noting that

$$(\mathbf{M} + \tilde{\mathbf{X}}^t \tilde{\mathbf{X}})^{-1} = \begin{pmatrix} \mathrm{z}^t \mathrm{z} & \mathrm{z}^t \mathbf{X} \\ \mathbf{X}^t \mathrm{z} & \mathbf{X}^t \mathbf{X} + \mathbf{I}_p \end{pmatrix}^{-1}$$

we obtain

$$\text{bias}(\hat{\tau}_{rr}) = -\left((\mathrm{z}^t\mathrm{z})^{-1}\mathrm{z}^t\mathbf{X}\right)(\mathbf{I} + \mathbf{X}^t(\mathbf{X} - \hat{\mathbf{X}}_\mathrm{z}))^{-1}\beta, \tag{4.3}$$

where $\hat{\mathbf{X}}_\mathrm{z} = \mathrm{z}(\mathrm{z}^t\mathrm{z})^{-1}\mathrm{z}^t\mathbf{X}$. Notice that the leading term $\left((\mathrm{z}^t\mathrm{z})^{-1}\mathrm{z}^t\mathbf{X}\right)$ is a vector of regression coefficients from p univariate regressions predicting X_j given z. With completely randomized treatment assignment these terms will tend to be near zero (and precisely zero in expectation over Z). This ensures that the ridge estimate of τ is nearly unbiased, despite the fact that the middle matrix is generally nonzero. However, in the presence of selection, some of these regression coefficients will be non-zero due to the correlation between Z and the covariates in $\mathbf{X}$. As a result, the bias of $\hat{\tau}_{rr}$ will depend on the form of the design matrix and unknown nuisance parameters β.

The problem here is not simply that $\hat{\tau}_{rr}$ is biased — after all, the insight behind regularization is that some bias can actually improve our average estimation error. Rather, the problem is that the degree of bias is not under the analyst's control (as it depends on unknown nuisance parameters). The use of a naive regularization prior in the presence of confounding can unwittingly induce extreme bias in estimation of the target parameter, *even when all the confounders are measured and the parametric model is correctly specified*.

In more complicated nonparametric regression models with heterogeneous treatment effects a closed-form expression of the bias is not generally available; see Yang et al. (2015) and Chernozhukov et al. (2016) for related results in a partially linear model where effects are homogenous but the $\beta^t\mathrm{x}$ term above is replaced by a nonlinear function. However, note that both of these theoretical results consider *asymptotic* bias in semi- and non-parametric Bayesian and frequentist inference; our attention here to the simple case of the linear model shows that the phenomenon occurs in finite samples even in a parametric model. That said, the RIC phenomenon can be reliably recreated in nonlinear, semiparametric settings. The easiest way to demonstrate this is by considering scenarios where selection into treatment is based on expected outcomes under no treatment, a situation we call *targeted selection*.

4.2 Targeted selection

Targeted selection refers to settings where treatment is assigned based on a prediction of the outcome in the *absence* of treatment, given measured covariates. That is, targeted selection asserts that treatment is being assigned, in part, based on an estimate of the expected potential outcome $\mu(\mathrm{x}) := \mathrm{E}(Y(0) \mid \mathrm{x})$ and that the probability of treatment is generally increasing or decreasing as a function of this estimate. We suspect this selection process is quite common in practice; for example, in medical contexts where risk factors for adverse outcomes are well-understood physicians are more likely to assign treatment to patients with worse expected outcomes in its absence.

Targeted selection implies that there is a particular functional relationship between the propensity score π and the expected outcomes without treatment μ. In particular, suppose for simplicity that there exists a change of variables $\mathrm{x} \to (\mu(\mathrm{x}), \tilde{\mathrm{x}})$ that takes the prognostic function $\mu(\mathrm{x})$ to the first element of the covariate vector. Then targeted selection says that for every $\tilde{\mathrm{x}}$, the propensity function $\mathrm{E}(Z \mid \mathrm{x}) = \pi(\mu, \tilde{\mathrm{x}})$ is (approximately)

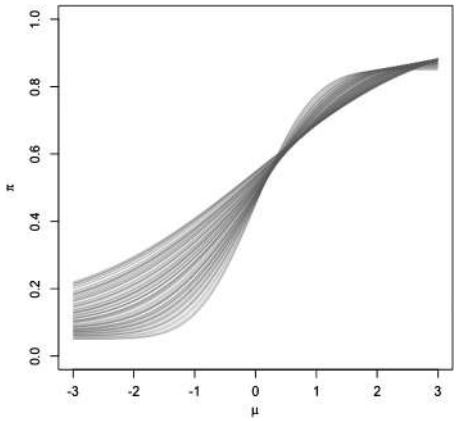


Figure 2: For any value of $\tilde{\mathrm{x}}$, the propensity score $\pi(\mu, \tilde{\mathrm{x}})$ is monotone in the prognostic function μ. Here, many realizations of this function are shown for different values of $\tilde{\mathrm{x}}$.

monotone in μ; see Figure 2 for a visual depiction. If the relationship is strictly monotone so that π is invertible in μ for any $\tilde{x}$, this in turn implies that $\mu(\mathrm{x})$ is a function of $\pi(\mathrm{x})$.

Targeted selection and RIC in the linear model

To help understand how targeted selection leads to RIC, it is helpful to again consider the linear model. There, one can describe RIC in terms of three components: the coefficients defining the propensity function $\mathrm{E}(Z \mid \mathrm{x}) = \gamma\mathrm{x}$, the coefficients defining the prognostic function, $\mathrm{E}(Y \mid Z = 0, \mathrm{x} = x)$, and the strength of the selection as measured by $\mathrm{Var}(Z \mid \mathrm{x}) = \mathrm{Var}(\nu)$. Specifically, note the identity

$$\mathrm{E}(Y \mid \mathrm{x}, Z) = (\tau + b)Z + (\beta - b\gamma)^t\mathrm{x} - b(Z - \gamma^t\mathrm{x}) = \hat{\tau}Z + \hat{\beta}^t\mathrm{x} - \hat{\epsilon}, \tag{4.4}$$

which is true for any value of the scalar parameter b, the bias of $\hat{\tau}$. Intuitively, if neighborhoods of $\hat{\beta} = (\beta - b\gamma)$ have higher prior probability than β and $\mathrm{Var}(\hat{\epsilon}) = b^2\mathrm{Var}(\nu)$ is small on average relative to σ^2, then the posterior distribution for τ is apt to be biased toward $\hat{\tau} = \tau + b$.

The bias will be large precisely when confounding is strong and the selection is targeted: For non-negligible bias the term $b^2\mathrm{Var}(\nu)$ is smallest when $\mathrm{Var}(\nu)$ is small, that is, when selection (hence, confounding) is strong. For priors on β that are centered at zero —which is overwhelmingly the default — the $(\beta - b\gamma)$ term can be made most favorable with respect to the prior when the vector β and γ have the same direction, which corresponds to perfectly targeted selection.

Targeted selection and RIC in nonlinear models

To investigate RIC in more complex regression settings, we start with a simple 2-d example characterized by targeted selection:

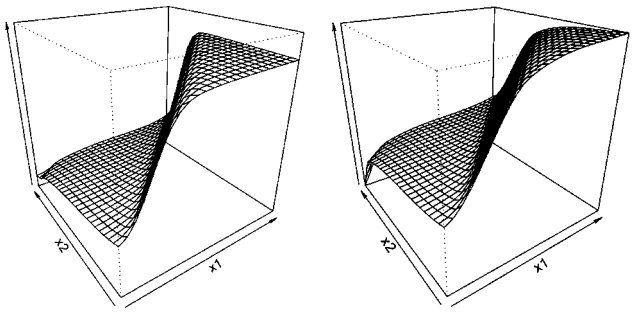


Figure 3: Left panel: The propensity function, π, shown for various values of $\tilde{x}$. The "shelf" at the line $x_1 = x_2$ is a complex shape for many regression methods to represent. Right panel: the analogous plot for the prognostic function μ. Note the similar shapes due to targeted selection; the π function falls between 0 to 1, while the μ function ranges from -3 to 3.

Example 1: $d = 2$, $n = 250$, homogeneous effects

Consider the following data generating process:

$$\begin{aligned} Y_i &= \mu(x_1, x_2) - \tau Z_i + \epsilon_i, \\ \mathrm{E}(Y_i \mid x_{i1}, x_{i2}, Z_i = 1) &= \mu(x_1, x_2), \\ \mathrm{E}(Z_i \mid x_{i1}, x_{i2}) &= \pi(\mu(x_{i1}, x_{i2}), x_{i1}, x_{i2}) \\ &= 0.8\Phi\left(\frac{\mu(x_{i1}, x_{i2})}{0.1(2 - x_{i1} - x_{i2}) + 0.25}\right) + 0.025(x_{i1} + x_{i2}) + 0.05, \\ \epsilon_i &\overset{\text{iid}}{\sim} \mathrm{N}(0, 1), \quad x_{i1}, x_{i2} \overset{\text{iid}}{\sim} \mathrm{Uniform}(0, 1). \end{aligned} \tag{4.5}$$

Suppose that in (4.5) Y is a continuous biometric measure of heart distress, Z is an indicator for having received a heart medication, and x_1 and x_2 are systolic and diastolic blood pressure (in standardized units), respectively. Suppose that it is known that the *difference* between these two measurements is prognostic of high distress levels, with positive levels of $x_1 - x_2$ being a critical threshold. At the same time, suppose that prescribers are targeting the drug towards patients with high levels of diagnostic markers, so the probability of receiving the drug is an increasing function in μ. Figure 3 shows π as a function of x_1 and x_2; Figure 2 shows the relationship between μ and π for various values of $\tilde{x} = x_1 + x_2$.

We simulated 200 datasets of size $n = 250$ according to this data generating process with $\tau = -1$. With only a few covariates, low noise, and a relatively large sample size, we might expect most methods to perform well here. Table 1 shows that standard, unmodified BART exhibits high bias and root mean squared error (RMSE) as well as poor coverage of 95% credible intervals. Our proposed fix (detailed below) improves on both estimation error and coverage, primarily by including an estimate of π as a covariate.

Prior	bias	coverage	rmse
BART	0.27	65%	0.31
BCF	0.14	95%	0.21

Table 1: The standard BART prior exhibits substantial bias in estimating the treatment effect, poor coverage of 95% posterior (quantile-based) credible intervals, and high root mean squared error (rmse). A modified BART prior (denoted BCF) allows splits in an estimated propensity score; it performs markedly better on all three metrics.

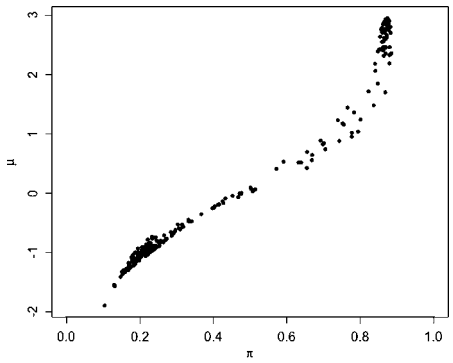


Figure 4: This scatterplot depicts $\mu(\mathrm{x}) = \mathrm{E}(Y \mid Z = 0, \mathrm{x})$ and $\pi(\mathrm{x}) = \mathrm{E}(Z \mid \mathrm{x})$ for a realization from the data generating process from the above example. It shows clear evidence of targeted selection. Such plots, based on estimates $(\hat{\mu}, \hat{\pi})$ can provide evidence of (strong) targeted selection in empirical data.

What explains BART's relatively poor performance on this data generating process (DGP)? First, strong confounding and targeted selection implies that μ is approximately a monotone function of π alone (Figure 4). However, π (and hence μ) is difficult to learn via regression trees — it takes many axis-aligned splits to approximate the "shelf" across the diagonal (see Figure 5), and the BART prior specifically penalizes this kind of complexity. At the same time, due to the strong confounding in this example a single split in Z can stand in for many splits on x_1 and x_2 that would be required to approximate $\mu(\mathrm{x})$. These simpler structures are favored by the BART prior, leading to RIC.

Before discussing how we reduce RIC, we note that this example is somewhat stylized in that we designed it specifically to be difficult to learn for tree-based models. Other models might suffer less from RIC on this particular example. However, any informative, sparse, or nonparametric prior distribution – any method that imposes meaningful regularization – is susceptible to similar effects, as they prioritize some data-generating processes at the expense of others. Absent prior knowledge of the form of the treatment assignment and outcome models, it is impossible to know *a prior* whether RIC will be an issue. Fortunately it is straightforward to minimize the risk of bias due to RIC.

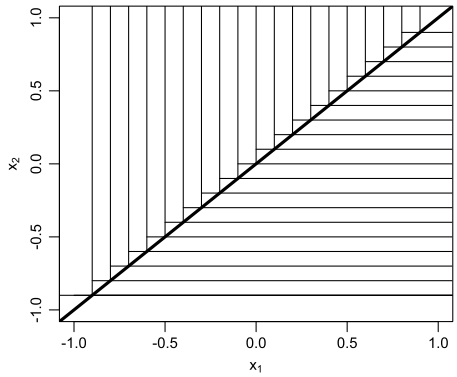


Figure 5: Many axis-aligned splits are required to approximate a step function (or near-step function) along the diagonal in the outcome model, as in Figure 3 (right panel). Since these two regions correspond also to disparate rates of treatment, tree-based regularized regression is apt to overstate the treatment effect.

4.3 Mitigating RIC with covariate-dependent priors

Finally, we arrive at the role of the propensity score in a regularized regression context. The potential for RIC is strongest when $\mu(\mathrm{x})$ is exactly or approximately a function of $\pi(\mathrm{x})$ and when the composition of the two has relatively low prior support. This can lead the model to misattribute the variability of μ, in the direction of π, to Z. A natural solution to this problem would be to include $\pi(\mathrm{x})$ as a covariate, so that it is penalized equitably with changes in the treatment variable Z. That is, when evaluating candidate functions for our estimate of $\mathrm{E}(Y \mid \mathrm{x}, z)$ we want representations involving $\pi(\mathrm{x})$ to be regularized/penalized the same as representations involving z. Of course π is unknown and must be estimated, but this is a straightforward regression problem. Note also that so-called "unlabeled" data can be brought to bear here, meaning π can be estimated from samples of (Z, X) for which the Y value is unobserved, provided the sample is believed to arise from the relevant population.

Mitigating RIC in the linear model

Given an estimate of the propensity function $\hat{z}_i \approx \gamma^t \mathrm{x}_i$, we consider the over-complete regression that includes as regressors both z and $\hat{\mathrm{z}}$. Our design matrix becomes

$$\tilde{\mathbf{X}} = \begin{pmatrix} \mathrm{z} & \hat{\mathrm{z}} & \mathbf{X} \end{pmatrix}.$$

This covariate matrix is degenerate because $\hat{z}$ is in the column span of $\mathbf{X}$ by construction. In a regularized regression problem this degeneracy is no obstacle. Applying the expression for the bias from above, with a flat prior over the coefficient associated with $\hat{\mathrm{z}}$, yields

$$\mathrm{bias}(\hat{\tau}_{rr}) = -\left\{(\tilde{\mathrm{z}}^t\tilde{\mathrm{z}})^{-1}\tilde{\mathrm{z}}^t\mathbf{X}\right\}_1 \left(\mathbf{I} + \mathbf{X}^t(\mathbf{X} - \hat{\mathbf{X}}_{\mathrm{z}})\right)^{-1}\beta = 0,$$

where $\tilde{\mathrm{z}} = (\mathrm{z} \quad \hat{\mathrm{z}})$ and $\left\{(\tilde{\mathrm{z}}^t\tilde{\mathrm{z}})^{-1}\tilde{\mathrm{z}}^t\mathbf{X}\right\}_1$ denotes the top row of $\left\{(\tilde{\mathrm{z}}^t\tilde{\mathrm{z}})^{-1}\tilde{\mathrm{z}}^t\mathbf{X}\right\}$, which corresponds to the regression coefficient associated with z in the two variable regression predicting X_j given $\tilde{z}$. Because $\hat{z}$ captures the observed association between z and x, z is conditionally independent of x given $\hat{z}$, from which we conclude that these regression coefficients will be zero. See Yang et al. (2015) for a similar de-biasing strategy in a partially linear semiparametric context.

Mitigating RIC in nonlinear models

The same strategy also proves effective in the nonlinear setting — simply by including an estimate of the propensity score as a covariate in the BART model, the RIC effect is dramatically mitigated, as can be seen in the second row of Table 1. From a Bayesian perspective, this is simply a judicious variable transformation since our regression model is specified conditional on both Z and x — we are not obliged to consider uncertainty in our estimate of π to obtain valid posterior inference. We obtain another example of a covariate dependent prior, similar to Zellner's g-prior (albeit motivated by very different considerations). See Section 8 for additional discussion of this point. Finally, we believe that including an estimated propensity score will be cheap insurance against RIC when estimating treatment effects using outcome models under other nonparametric priors and using more general nonparametric/machine learning approaches.

To summarize, although it has long been known that the propensity score is a sufficient dimension reduction for estimation of the ATE – and that combining estimates of the response surface and propensity score can improve estimation of average treatment effects (Bang and Robins, 2005), we find that incorporating an estimate of the propensity score into estimation of the response surface can improve estimation of average treatment effects in finite samples. As we will demonstrate in Section 6, these benefits also accrue when estimating (heterogeneous) conditional average treatment effects. Estimating heterogenous effects also calls for careful consideration of regularization applied to the treatment effect function, which we consider in the next section.

5 Regularization for heterogeneous treatment effects: Bayesian causal forests

In much the same way that a direct BART prior on f does not allow careful handling of confounding, it also does not allow separate control over the discovery of heterogeneous effects because there is no explicit control over how f varies in Z. Our solution to this problem is a simple re-parametrization that avoids the indirect specification of the prior over the treatment effects:

$$f(\mathrm{x}_i, z_i) = \mu(\mathrm{x}_i) + \tau(\mathrm{x}_i)z_i. \tag{5.1}$$

This model can be thought of as a linear regression in z with covariate-dependent functions for both the slope and the intercept. Writing the model this way sacrifices

nothing in terms of expressiveness, but permits independent priors to be placed on τ, which is precisely the treatment effect:

$$\mathrm{E}(Y_i \mid \mathrm{x}_i, Z_i = 1) - \mathrm{E}(Y_i \mid \mathrm{x}_i, Z_i = 0) = \{\mu(\mathrm{x}_i) + \tau(\mathrm{x}_i)\} - \mu(\mathrm{x}_i) = \tau(\mathrm{x}_i). \tag{5.2}$$

Under this model, $\mu(\mathrm{x}) = E(Y \mid Z = 0, X = x)$ is a prognostic score in the sense of Hansen (2008), another interpretable quantity, to which we apply a prior distribution independent of τ (as detailed below). A similar idea has been proposed previously for non-tree based models in Imai et al. (2013).

While we will use variants of BART priors for μ and τ (see Section 5.2), this parameterization has many advantages in general, regardless of the specific priors. The most obvious advantage is that the treatment effect is an explicit parameter of the model, $\tau(\mathrm{x})$, and as a result we can specify an appropriate prior on it directly. This parameterization is especially useful when only a subset of the variables in x, say w, are plausible or interesting treatment effect modifiers, since we can directly specify the model as

$$f(\mathrm{x}_i, z_i) = \mu(\mathrm{x}_i) + \tau(\mathrm{w}_i) z_i.$$

This dimension reduction is useful for the usual statistical reasons and also for rendering estimates of heterogeneous treatment effects more interpretable. This has been important in applications of the methods outlined in this paper (Yeager et al., 2019).

Finally, based on the observations of the previous section, we further propose specifying the model as

$$f(\mathrm{x}_i, z_i) = \mu(\mathrm{x}_i, \hat{\pi}_i) + \tau(\mathrm{x}_i) z_i, \tag{5.3}$$

where $\hat{\pi}_i$ is an estimate of the propensity score. Before turning to the details of our model specification, we first contrast this parameterization with two common alternatives.

5.1 Parameterizing regression models of heterogeneous effects

There are two common modeling strategies for estimating heterogeneous effects. The first we discussed above: treat z as "just another covariate" and specify a prior on $f(\mathrm{x}_i, z_i)$, e.g. as in Hill (2011). The second is to fit entirely separate models to the treatment and control data: $(Y \mid Z = z, \mathrm{x}) \sim N(f_z(\mathrm{x}_i), \sigma_z^2)$ with independent priors over the parameters in the $z = 0$ and $z = 1$ models. In this section we argue that neither approach is satisfactory and propose the model in (5.3) as a reasonable interpolation between the two. (See Künzel et al. (2019) for a related discussion comparing these two approaches in a non-model-based setting.)

It is instructive to consider (5.1) as a nonlinear regression analogue of the common strategy of parametrizing contrasts (differences) and aggregates (sums) rather than group-specific location parameters. Specifically, consider a two-group difference-in-means problem:

$$\begin{aligned} Y_{i1} &\overset{\text{iid}}{\sim} \mathrm{N}(\mu_1, \sigma^2), \\ Y_{j2} &\overset{\text{iid}}{\sim} \mathrm{N}(\mu_2, \sigma^2). \end{aligned} \tag{5.4}$$

Although the above parameterization is intuitive, if the estimand of interest is $\mu_1 - \mu_2$, the implied prior over this quantity has variance strictly greater than the variances over

μ_1 or μ_2 individually. This is plainly nonsensical if the analyst has no subject matter knowledge regarding the individual levels of the groups, but has strong prior knowledge that $\mu_1 \approx \mu_2$. This is common in a causal inference setting: If the data come from a randomized experiment where Y_1 constitutes a control sample and Y_2 a treated sample, then subject matter considerations will typically limit the plausible range of treatment effects $\mu_1 - \mu_2$.

The appropriate way to incorporate that kind of knowledge is simply to reparametrize:

$$\begin{aligned} Y_{i1} &\overset{\text{iid}}{\sim} \mathrm{N}(\mu + \tau, \sigma^2), \\ Y_{j2} &\overset{\text{iid}}{\sim} \mathrm{N}(\mu, \sigma^2), \end{aligned} \tag{5.5}$$

whereupon the estimand of interest becomes τ, which can be given an informative prior centered at zero with an appropriate variance. Meanwhile, μ can be given a very vague (perhaps even improper) prior.

While the nonlinear regression context is more complex, the considerations are the same: our goal is simultaneously to let $\mu(\mathrm{x})$ be flexibly learned (to adequately deconfound and obtain more precise inference), while appropriately regularizing $\tau(\mathrm{x})$, which we expect, a priori, to be relatively small in magnitude and "simple" (minimal heterogeneity). Neither of the two more common parametrizations permit this: Independent estimation of $f_0(\mathrm{x})$ and $f_1(\mathrm{x})$ implies a highly vague prior on $\tau(\mathrm{x}) = f_1(\mathrm{x}) - f_0(\mathrm{x})$; i.e. a Gaussian process prior on each would imply a twice-as-variable Gaussian process prior on the difference, as in the simple example above. Estimation based on the single response surface $f(\mathrm{x}, z)$ often does not allow direct point-wise control of $\tau(\mathrm{x}) = f(\mathrm{x}, 1) - f(\mathrm{x}, 0)$ at all. In particular, with a BART prior on f the induced prior on τ depends on incidental features such as the size and distribution of the covariate vector x.

5.2 Prior specification

With the model parameterized as in (5.3), we can specify different BART priors on μ and τ. For μ we use the default suggestions in (Chipman et al., 2010) (200 trees, $\beta = 2$, $\eta = 0.95$), except that we place a half-Cauchy prior over the scale of the leaf parameters with prior median equal to twice the marginal standard deviation of Y (Gelman et al., 2006; Polson et al., 2012). We find that inference over τ is typically insensitive to reasonable deviations from these settings, so long as the prior is not so strong that deconfounding does not take place.

For τ, we prefer stronger regularization. First, we use fewer trees (50 versus 200), as we generally believe that patterns of treatment effect heterogeneity are relatively simple. Second, we set the depth penalty $\beta = 3$ and splitting probability $\eta = 0.25$ (instead of $\beta = 2$ and $\eta = 0.95$) to shrink more strongly toward homogenous effects; the extreme case where none of the trees split at all corresponds to purely homogenous effects. Finally, we replace the half-Cauchy prior over the scale of τ with a half Normal prior, pegging the prior median to the marginal standard deviation of Y. In the absence

of prior information about the plausible range of treatment effects we expect this to be a reasonable upper bound.

5.3 Data-adaptive coding of treatment assignment

A less-desirable feature of the model in (5.3) is that different inferences on $\tau(\mathrm{x})$ can obtain if we code the treatment indicator as zero for treated cases and one for controls, or as $\pm 1/2$ for treated/control units, or any of the other myriad specifications for Z_i that still result in $\tau(\mathrm{x})$ being the treatment effect. When there is a clear reference treatment level one might think of this as a feature, not a bug, but this is often not the case (such as when comparing two active treatments). Because μ and τ alias one another, as under targeted selection, the choice of treatment coding can meaningfully impact posterior inferences, especially when the treated and control response variables have dramatically different marginal variances.

Fortunately, an invariant parameterization is possible, which treats the coding of Z as a parameter to be estimated:

$$\begin{aligned} y_i = \mu(\mathrm{x}_i) + \tilde{\tau}(\mathrm{x}_i) b_{z_i} + \epsilon_i, \quad \epsilon_i \sim N(0, \sigma^2), \\ b_0 \sim N(0, 1/2), \quad b_1 \sim N(0, 1/2). \end{aligned} \tag{5.6}$$

The treatment effect function in this parameter expanded model is

$$\tau(\mathrm{x}_i) = (b_1 - b_0)\tilde{\tau}(\mathrm{x}_i).$$

Noting that $b_1 - b_0 \sim N(0, 1)$ we still obtain a half Normal prior distribution for the scale of the leaf parameters in τ as in the previous subsection, and we can adjust the scale of the half normal prior (e.g. to fix the scale at one marginal standard deviation of Y as above) using a fixed scale in the leaf prior for $\tilde{\tau}$. Posterior inference in this model requires only minor adjustments to Chipman et al. (2010)'s Bayesian backfitting MCMC algorithm. Specifically, note that conditional on τ, μ and σ, updates for b_0 and b_1 follow from standard linear regression updates, with a two-column design matrix with columns $(\tau(\mathrm{x}_i)z_i,\ \tau(\mathrm{x}_i)(1 - z_i))$, no intercept, and "residual" $y_i - \mu(\mathrm{x}_i)$ as the response variable.

Our experiments below all use this parameterization, and it is the default implementation in our software package.

6 Empirical evaluations

In this section, we provide a more extensive look at how BCF compares to various alternatives. In Section 6.1 we compare BCF, generalized random forests (Athey et al., 2019), and a linear model with all three-way interactions as plausible methods for estimating heterogeneous treatment effects with measures of uncertainty. We also consider three specifications of BART: the standard response surface BART that considers the treatment variable as "just another covariate", one where separate BART models are fit to the treatment and control arms of the data, and one where an estimate of the

propensity score is included as a predictor (ps-BART). In Section 6.2 we report on the results of two separate data analysis challenges, where the entire community was invited to submit methods for evaluation on larger synthetic datasets with heterogeneous treatment effects. In both simulation settings we find that BCF performs well under a wide range of scenarios.

In all cases the estimands of interest are either conditional average treatment effects for individual i accounting for all the variables, estimated by the posterior mean treatment effect $\hat{\tau}(\mathrm{x}_i)$, or sample subgroup average treatment effects estimated by $\sum_{i\in\mathcal{S}} \hat{\tau}(\mathrm{x}_i)$, where $\mathcal{S}$ is the subgroup of interest. Credible intervals are computed from Markov chain Monte Carlo (MCMC) output using quantiles.

6.1 Simulation studies

We evaluated three variants of BART, the causal random forest model of Athey et al. (2019) (using the default specification in the `grf` package), and a regularized linear model (LM), using the horseshoe prior (HR) of Carvalho et al. (2010), with up to three way interactions. We consider eight distinct, but closely related, data generating processes, corresponding to the various combinations of toggling three two-level settings: homogeneous versus heterogeneous treatment effects, a linear versus nonlinear conditional expectation function, and two different sample sizes ($n = 250$ and $n = 500$). Five variables comprise x; the first three are continuous, drawn as standard normal random variables, the fourth is a dichotomous variable and the fifth is unordered categorical, taking three levels (denoted 1, 2, 3). The treatment effect is either

$$\tau(\mathrm{x}) = \begin{cases} 3, & \text{homogeneous,} \\ 1 + 2x_2x_5, & \text{heterogeneous,} \end{cases}$$

the prognostic function is either

$$\mu(\mathrm{x}) = \begin{cases} 1 + g(x_4) + x_1x_3, & \text{linear,} \\ -6 + g(x_4) + 6|x_3 - 1|, & \text{nonlinear,} \end{cases}$$

where $g(1) = 2$, $g(2) = -1$ and $g(3) = -4$, and the propensity function is

$$\pi(\mathrm{x}_i) = 0.8\Phi(3\mu(\mathrm{x}_i)/s - 0.5x_1) + 0.05 + u_i/10,$$

where s is the standard deviation of μ taken over the observed sample and $u_i \sim \mathrm{Uniform}(0, 1)$.

To evaluate each method we consider three criteria, applied to two different estimands. First, we consider how each method does at estimating the (sample) average treatment effect (ATE) according to root mean square error, coverage, and average interval length. Then, we consider the same criteria, except applied to estimates of the conditional average treatment effect (CATE), averaged over the sample. Results are based on 200 independent replications for each DGP. Results are reported in Tables 2 (for the linear DGP) and 3 (for the nonlinear DGP). The important trends are as follows:

- BCF and ps-BART benefit dramatically by explicitly protecting against RIC;
- BART-(f_0, f_1) and causal random forests both exhibit subpar performance;
- all methods improve with a larger sample;

		Homogeneous effect						Heterogeneous effects					
n	Method	ATE			CATE			ATE			CATE		
		rmse	cover	len	rmse	cover	len	rmse	cover	len	rmse	cover	len
250	BCF	0.21	0.92	0.91	0.48	0.96	2.0	0.27	0.84	0.99	1.09	0.91	3.3
	ps-BART	0.22	0.94	0.97	0.44	0.99	2.3	0.31	0.90	1.13	1.30	0.89	3.5
	BART	0.34	0.73	0.94	0.54	0.95	2.3	0.45	0.65	1.10	1.36	0.87	3.4
	BART (f_0, f_1)	0.56	0.41	0.99	0.92	0.93	3.4	0.61	0.44	1.14	1.47	0.90	4.5
	Causal RF	0.34	0.73	0.98	0.47	0.84	1.3	0.49	0.68	1.25	1.58	0.68	2.4
	LM + HS	0.14	0.96	0.83	0.26	0.99	1.7	0.17	0.94	0.89	0.33	0.99	1.9
500	BCF	0.16	0.88	0.60	0.38	0.95	1.4	0.16	0.90	0.64	0.79	0.89	2.4
	ps-BART	0.18	0.86	0.63	0.35	0.99	1.8	0.16	0.90	0.69	0.86	0.95	2.8
	BART	0.27	0.61	0.61	0.42	0.95	1.8	0.25	0.76	0.67	0.88	0.94	2.8
	BART (f_0, f_1)	0.47	0.21	0.66	0.80	0.93	3.1	0.42	0.42	0.75	1.16	0.92	3.9
	Causal RF	0.36	0.47	0.69	0.52	0.75	1.2	0.40	0.59	0.88	1.30	0.71	2.1
	LM + HS	0.11	0.96	0.54	0.18	0.99	1.0	0.12	0.93	0.59	0.22	0.98	1.2

Table 2: Simulation study results when the true DGP is a linear model with third order interactions. Root mean square estimation error (rmse), coverage (cover) and average interval length (len) are reported for both the average treatment effect (ATE) estimates and the conditional average treatment effect estimates (CATE).

		Homogeneous effect						Heterogeneous effects					
n	Method	ATE			CATE			ATE			CATE		
		rmse	cover	len	rmse	cover	len	rmse	cover	len	rmse	cover	len
250	BCF	0.26	0.945	1.3	0.63	0.94	2.5	0.30	0.930	1.4	1.3	0.93	4.5
	ps-BART	0.54	0.780	1.6	1.00	0.96	4.3	0.56	0.805	1.7	1.7	0.91	5.4
	BART	0.84	0.425	1.5	1.20	0.90	4.1	0.84	0.430	1.6	1.8	0.87	5.2
	BART (f_0, f_1)	1.48	0.035	1.5	2.42	0.80	6.4	1.44	0.085	1.6	2.6	0.83	7.1
	Causal RF	0.81	0.425	1.5	0.84	0.70	2.0	1.10	0.305	1.8	1.8	0.66	3.4
	LM + HS	1.77	0.015	1.8	2.13	0.54	4.4	1.65	0.085	1.9	2.2	0.62	4.8
500	BCF	0.20	0.945	0.97	0.47	0.94	1.9	0.23	0.910	0.97	1.0	0.92	3.4
	ps-BART	0.24	0.910	1.07	0.62	0.99	3.3	0.26	0.890	1.06	1.1	0.95	4.1
	BART	0.31	0.790	1.00	0.63	0.98	3.0	0.33	0.760	1.00	1.1	0.94	3.9
	BART (f_0, f_1)	1.11	0.035	1.18	2.11	0.81	5.8	1.09	0.065	1.17	2.3	0.82	6.2
	Causal RF	0.39	0.650	1.00	0.54	0.87	1.7	0.59	0.515	1.18	1.5	0.73	2.8
	LM + HS	1.76	0.005	1.34	2.19	0.40	3.5	1.71	0.000	1.34	2.2	0.45	3.7

Table 3: Simulation study results when the true DGP is nonlinear. Root mean square estimation error (rmse), coverage (cover) and average interval length (len) are reported for both the average treatment effect (ATE) estimates and the conditional average treatment effect estimates (CATE).

- BCF priors are extra helpful at smaller sample sizes, when estimation is difficult;
- the linear model dominates when correct, but fares extremely poorly when wrong;
- BCF's improvements over ps-BART are more pronounced in the nonlinear DGP;
- BCF's average interval length is notably smaller than the ps-BART interval, usually (but not always) with comparable coverage.

6.2 Atlantic causal inference conference data analysis challenges

The Atlantic Causal Inference Conference (ACIC) has featured a data analysis challenge since 2016. Participants are given a large number of synthetic datasets and invited to submit their estimates of treatment effects along with confidence or credible intervals where available. Specifically, participants were asked to produce estimates and uncertainty intervals for the sample average treatment effect on the treated, as well as conditional average treatment effects for each unit. Methods were evaluated based on a range of criteria including estimation error and coverage of uncertainty intervals. The datasets and ground truths are publicly available, so while BCF was not entered into either the 2016 or 2017 competitions we can benchmark its performance against a suite of methods that we did not choose, design, or implement.

ACIC 2016 competition

The 2016 contest design, submitted methods, and results are summarized in Dorie et al. (2019). Based on an early draft of our manuscript Dorie et al. (2019) also evaluated a version of BART that included an estimate of the propensity score, which was one of the top methods on bias and RMSE for estimating the sample average treatment effect among the treated (SATT). BART with the propensity score outperformed BART without the propensity score on bias, RMSE, and coverage for the SATT, and was a leading method overall.

Therefore, rather than include results for all 30 methods here we simply include BART and ps-BART as leading contenders for estimating heterogeneous treatment effects in this setting. Using the publicly-available competition datasets (Dorie and Hill, 2017) we implemented two additional methods: BCF and causal random forests as implemented in the R package `grf` (Athey et al., 2019), using 4,000 trees to obtain confidence intervals for conditional average treatment effects and a doubly robust estimator for the SATT (as suggested in the package documentation).

Table 4 collects the results of our methods (ps-BART and BCF) as well as BART and causal random forests. Causal random forests performed notably worse than BART-based methods on every metric. BCF performed best in terms of estimation error for CATE and SATT, as measured by bias and absolute bias. While the differences in the various metrics are relatively small compared to their standard deviation across the 7,700 simulated datasets, nearly all the pairwise differences between BCF and the other methods are statistically significant as measured by a permutation test (Table 5). The sole exception is the test for a difference in bias between ps-BART and BCF, suggesting the presence of RIC in at least some of the simulated datasets. This is especially notable since the datasets were not intentionally simulated to include targeted selection.

Dorie et al. (2019) note that all submitted methods were "somewhat disappointing" in inference for the SATT (i.e., few methods had coverage near the nominal rate with reasonably sized intervals). However, ps-BART did relatively well, 88% coverage of a 95% credible interval and one of the smallest interval lengths. ps-BART had slightly better coverage than BCF (88% versus 82%), with an average interval length that was 45% larger than BCF. Vanilla BART and BCF had similar coverage rates, but BART's interval length was about 55% larger than BCF. Dorie et al. (2019) found that TMLE-based adjustments could improve the coverage of BART-based estimates of the SATT; we expect that similar benefits would accrue using BCF with a TMLE adjustment, but obtaining valid confidence intervals for SATT is not our focus so we did not pursue this further.

	Coverage	Int. Len.	Bias	(SD)	|Bias|	(SD)	PEHE	(SD)
BCF	0.82	0.026	−0.0009	(0.01)	0.008	0.010	0.33	0.18
ps-BART	0.88	0.038	−0.0011	(0.01)	0.010	0.011	0.34	0.16
BART	0.81	0.040	−0.0016	(0.02)	0.012	0.013	0.36	0.19
Causal RF	0.58	0.055	−0.0155	(0.04)	0.029	0.027	0.45	0.21

Table 4: Abbreviated ACIC 2016 contest results. Coverage and average interval length are reported for nominal 95% uncertainty intervals. Bias and |Bias| are average bias and average absolute bias, respectively, over the. PEHE denotes the average "precision in estimating heterogeneous effects": the average root mean squared error of CATE estimates for each unit in a dataset (Hill, 2011).

	Diff Bias	p	Diff |Bias|	p	Diff PEHE	p
ps-BART	−0.00020	0.146	0.0011	$< 1e^{-4}$	0.010	$< 1e^{-4}$
BART	−0.00070	$< 1e^{-4}$	0.0031	$< 1e^{-4}$	0.037	$< 1e^{-4}$
Causal RF	−0.01453	$< 1e^{-4}$	0.0204	$< 1e^{-4}$	0.125	$< 1e^{-4}$

Table 5: Tests and estimates for differences between BCF and other methods in the ACIC 2016 competition. The p-values are from bootstrapp permutation tests with 100,000 replicates.

ACIC 2017 competition

The ACIC 2017 competition was designed to have average treatment effects that were smaller, with heterogenous treatment effects that were less variable, relative to the 2016 datasets. Arguably, the 2016 competition included many datasets with unrealistically large average treatment effects and similarly unrealistic degrees of heterogeneity.[2] Additionally, the 2017 competition explicitly incorporated targeted selection (unlike the 2016 datasets). The ACIC 2017 competition design and results are summarized completely in Hahn et al. (2018); here we report selected results for the datasets with independent additive errors.

[2]Across the 2016 competition datasets, the interquartile range of the SATT was 0.57 to 0.79 in standard deviations of Y, with a median value of 0.68. The standard deviation of the conditional average treatment effects for the sample units had an interquartile range of 0.24 to 0.93, again in units of standard deviations of Y. A significant fraction of the variability in Y was explained by heterogeneous treatment effects in a large number of the simulated datasets.

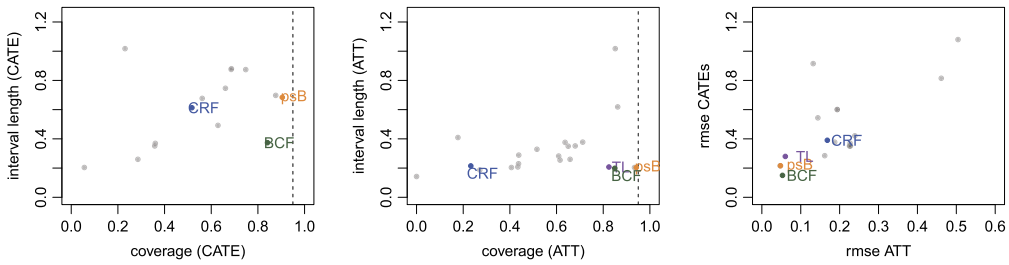


Figure 6: Each data point represents one method. ps-BART (psB, in orange) was submitted by a group independent of the authors based on a draft of this manuscript. TL (purple) is a TMLE-based submission that performed will for estimating SATT, but did not furnish estimates of conditional average treatment effects. BCF (green) and causal random forests (CRF, blue) were not part of the original contest. For descriptions of the other methods refer to Hahn et al. (2018).

Figure 6 contains the results of the 2017 competition. The patterns here are largely similar to the 2016 competition, despite some stark differences in the generation of synthetic datasets. ps-BART and BCF have the lowest estimation error for CATE and SATE. The closest competitor on estimation error was a TMLE-based approach. We also see that ps-BART edges BCF slightly in terms of coverage once again, although BCF has much shorter intervals. Causal random forests does not perform well, with coverage for SATT and CATE far below the nominal rate.

7 The effect of smoking on medical expenditures

7.1 Background and data

As an empirical demonstration of the Bayesian causal forest model, we consider the question of how smoking affects medical expenditures. This question is of interest as it relates to lawsuits against the tobacco industry. The lack of experimental data speaking to this question motivates the reliance on observational data. This question has been studied in several previous papers; see Zeger et al. (2000) and references therein. Here, we follow Imai and Van Dyk (2004) in analyzing data extracted from the 1987 National Medical Expenditure Survey (NMES) by Johnson et al. (2003). The NMES records many subject-level covariates and boasts third-party-verified medical expenses. Specifically, our regression includes the following ten patient attributes:

- `age`: age in years at the time of the survey
- `smoke_age`: age in years when the individual started smoking
- `gender`: male or female
- `race`: other, black or white
- `marriage_status`: married, widowed, divorced, separated, never married
- `education_level`: college graduate, some college, high school graduate, other

- `census_region`: geographic location, Northeast, Midwest, South, West
- `poverty_status`: poor, near poor, low income, middle income, high income
- `seat_belt`: does patient regularly use a seat belt when in a car
- `years_quit`: how many years since the individual quit smoking.

The response variable is the natural logarithm of annual medical expenditures, which makes the normality of the errors more plausible. Under this transformation, the treatment effect corresponds to a multiplicative effect on medical expenditure. Following Imai and Van Dyk (2004), we restrict our analysis to smokers who had non-zero medical expenditure. Our treatment variable is an indicator of heavy lifetime smoking, which we define to be greater than 17 *pack-years*, the equivalent of 17 years of pack-a-day smoking. See again Imai and Van Dyk (2004) for more discussion of this variable. We scrutinize the overlap assumption and exclude individuals younger than 28 on the grounds that it is improbable for someone that young to have achieved this level of exposure. After making these restrictions, our sample consists of $n = 6,798$ individuals.

7.2 Results

Here, we highlight the differences that arise when analyzing this data using standard BART versus using BCF. First, the estimated expected responses from the two models have correlation of 0.98, so that the two models concur on the nonlinear prediction problem. This suggests that, as was intended, BCF will inherit BART's outstanding predictive capabilities. By contrast, the estimated individual treatment effects are only correlated 0.70. The most notable differences between these CATE estimates is that the BCF estimates exhibit a strong trend in the age variable, as shown in Figure 9; the BCF estimates suggest that smoking has a pronounced impact on the health expenditures of younger people.

Despite a wider range of values in the CATE estimates (due largely to the inferred trend in the age variable), the ATE estimate of BCF is notably lower than that of BART, the posterior 95% credible intervals being translated by 0.05, $(0.00, 0.20)$ for BCF vs $(0.05, 0.25)$ for BART. The higher estimate of BART is possibly a result of RIC. Figure 7 shows a LOWESS trend between the estimated propensity and prognostic scores (from BCF); the monotone trend is suggestive of targeted selection (high medical expenses are predictive of heavy smoking) and hints at the possibility of RIC-type inflation of the BART ATE estimate (compare to Figures 2 and 4).

Although the vast majority of individual treatment effect estimates are statistically uncertain, as reflected in posterior 95% credible intervals that contain zero (Figure 8), the evidence for subgroup heterogeneity is relatively strong, as uncovered by the following posterior exploration strategy. First, we grow a parsimonious regression tree to the point estimates of the individual treatment effects (using the `rpart` package in `R`); see the left panel of Figure 10. Then, based on the candidate subgroups revealed by the regression summary tree, we plot a posterior histogram of the difference between any two covariate-defined subgroups. The right panel of Figure 10 shows the posterior distribution of the difference between men younger than 46 and women over 66; virtually all of the posterior mass is above zero, suggesting that the treatment effect of heavy

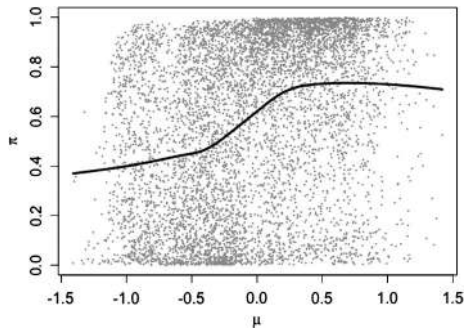


Figure 7: Each gray dot depicts the estimated propensity and prognostic scores for an individual. The solid bold line depicts a locally estimated scatterplot smoothed (LOWESS) trend fit to these points; the monotonicity is suggestive of targeted selection. Compare to Figures 2 and 4.

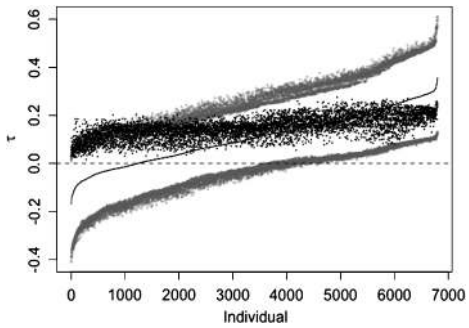


Figure 8: Point estimates of individual treatment effects are shown in black. The smooth line depicts the estimates from BCF, which are ordered from smallest to largest. The unordered points represent the corresponding individual treatment effect (ITE) estimates from BART. Note that the BART estimates seem to be higher, on average, than the BCF estimates. The upper and lower gray dots correspond to the posterior 95% credible interval end points associated with the BCF estimates; most ITE intervals contain zero, especially those with smaller (even negative) point estimates.

smoking is discernibly different for these two groups, with young men having a substantially higher estimated subgroup ATE. This approach, although somewhat informal, is a method of exploring the *posterior distribution* and, as such, any resulting summaries are still valid Bayesian inferences. Moreover, such Bayesian "fit-the-fit" posterior summarization strategies can be formalized from a decision theoretic perspective (Sivaganesan et al., 2017; Hahn and Carvalho, 2015); we do not explore this possibility further here.

From the above we conclude that how a model treats the age variable would seem to have an outsized impact on the way that predictive patterns are decomposed into treatment effect estimates based on this data, as age plausibly has prognostic, propensity

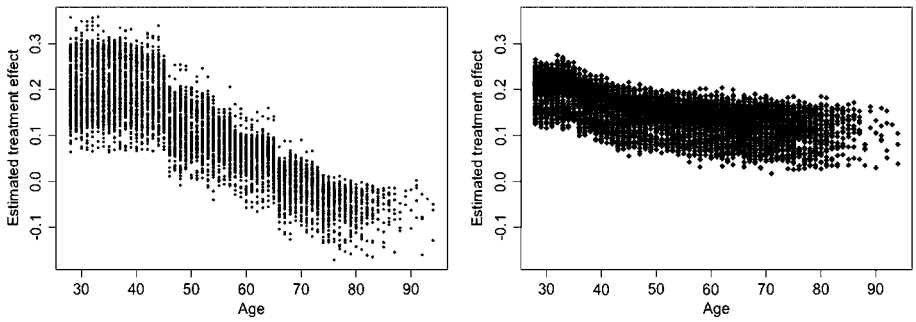


Figure 9: Each point depicts the estimated treatment effect for an individual. The BCF model (left panel) detects pronounced heterogeneity moderated by the age variable, whereas BART (right panel) does not.

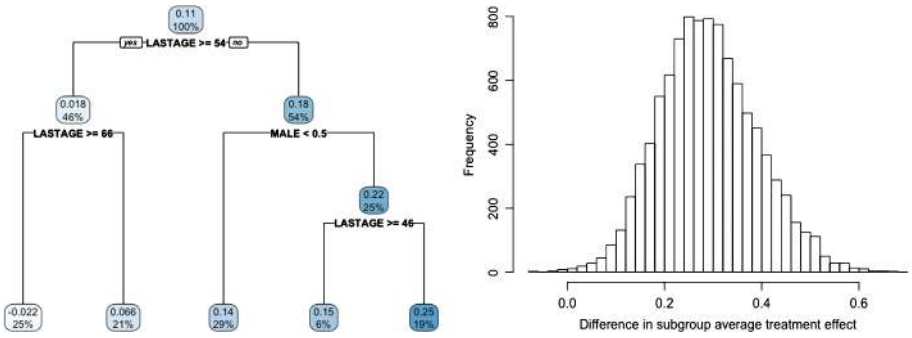


Figure 10: Left panel: a summarizing regression tree fit to posterior point estimates of individual treatment effects. The top number in each box is the average subgroup treatment effect in that partition of the population, the lower number shows the percentage of the total sample constituting that subgroup. Age and gender are flagged as important moderating variables. Right panel: based on the tree in the left panel, we consider the difference in treatment effects between men younger than 46 and women older than 66; a posterior histogram of this difference shows that nearly all of the posterior mass is above zero, indicating that these two subgroups are discernibly different, with young men having substantially higher subgroup average treatment effect.

and moderating roles simultaneously. Although it is difficult to trace the exact mechanism by which it happens, the BART model clearly de-emphasizes the moderating role, whereas the BCF model is designed specifically to capture such trends. Possible explanations for the age heterogeneity could be a mixed additive-multiplicative effect combined with higher baseline expenditures for older individuals or possibly survivor bias (as also mentioned in Imai and Van Dyk (2004)), but further speculation is beyond the scope of this analysis.

8 Discussion

We conclude by drawing out themes relating the Bayesian causal forest model to earlier work and by explicitly addressing common questions we have received while presenting the work in conferences and seminars.

8.1 Zellner priors for non- and semiparametric Bayesian causal inference

In Section 4 we showed that the current gold standard in nonparametric Bayesian regression models for causal inference (BART) is susceptible to regression induced confounding as described by Hahn et al. (2016). The solution we propose is to include an estimate of the propensity score as a covariate in the outcome model. This induces a prior distribution that treats Z_i and $\hat{\pi}_i$ equitably, discouraging the outcome model from erroneously attributing the effect of confounders to the treatment variable. Here we justify and collect arguments in favor of this approach. We discuss an argument against, namely that it does not incorporate uncertainty in the propensity score, in a later subsection.

Conditioning on an estimate of the propensity score is readily justified: Because our regression model is conditional on Z and $\mathbf{X}$, it is perfectly legitimate to condition our prior on them as well. This approach is widely used in linear regression, the most common example being Zellner's g-prior (Zellner, 1986) which parametrizes the prior covariance of a vector of regression coefficients in terms of a plug-in estimate of the predictor variables' covariance matrix. Nodding to this heritage, we propose to call general predictor-dependent priors "Zellner priors".

In the Bayesian causal forest model, we specify a prior over f by applying an independent BART prior that includes $\hat{\pi}(\mathrm{x}_i)$ as one of its splitting dimensions. That is, because $\hat{\pi}(\mathrm{x}_i)$ is a fixed function of x_i, f is still a function $f : (\mathcal{X}, \mathcal{Z}) \mapsto \mathbb{R}$; the inclusion of $\hat{\pi}(\mathrm{x}_i)$ among the splitting dimensions does not materially change the support of the prior, but it does alter which functions are deemed more likely. Therefore, although writing $f(\mathrm{x}_i, z_i, \hat{\pi}(\mathrm{x}_i))$ is suggestive of how the prior is implemented in practice, we prefer notation such as

$$\begin{aligned} Y_i &= f(\mathrm{x}_i, z_i) + \epsilon_i, \quad \epsilon_i \overset{\text{iid}}{\sim} \mathrm{N}(0, \sigma^2), \\ f &\sim \mathrm{BART}(\mathbf{X}, Z, \hat{\pi}), \end{aligned} \tag{8.1}$$

where $\hat{\pi}$ is itself a function of $(\mathbf{X}, Z)$. Viewing BART as a prior in this way highlights the fact that various transformations of the data could be computed beforehand, prior to fitting the data with the default BART priors; the choice of transformations will control the nature of the regularization that is imposed. In conventional predictive modeling there is often no particular knowledge of which transformations of the covariates might be helpful. However, in the treatment effect context the propensity score is a natural and, in fact, critical choice.

Finally, some have argued that a committed subjective Bayesian is specifically enjoined from encoding prior dependence on the propensity score in the outcome model

based on philosophical considerations (Robins and Ritov, 1997). We disagree; to the extent that phenomena like targeted selection are plausible, variation in treatment assignment is informative about variation in outcomes under control, and it would be inadvisable for a Bayesian – committed subjective or otherwise – to ignore it.

8.2 Why not use only the propensity score? vs. Why use the propensity score at all?

It has long been recognized that regression on the propensity score is a useful dimension reduction tactic (Rosenbaum and Rubin, 1983). For the purpose of estimating average treatment effects, a regression model on the one-dimensional propensity score is sufficient for the task, allowing one to side-step estimating high dimensional nuisance parameters. In our notation, if π is assumed known, then one need only infer $f(\pi)$. That said, there are several reasons one should include the control vector x_i in its entirety (in addition to the propensity score).

The first reason is pragmatic: If one wants to identify heterogeneous effects, one needs to include any potential effect moderating variables anyway, precluding any dimension reduction at the outset.

Second, if we are to take a conditionally-iid Bayesian regression approach to inference and we do not in fact believe the response to depend on $\mathbf{X}$ strictly through the propensity score, we simply must include the covariates themselves and model the conditional distribution $p(Y \mid Z, \mathbf{X})$ (otherwise the error distribution is highly dependent, integrated across $\mathbf{X}$). The justification for making inference about average treatment effects using regression or stratification on the propensity score alone is entirely frequentist; this approach is not without its merits, and we do not intend to argue frequency calibration is not desirable, but a fully Bayesian approach has its own appeal.

Third, if our propensity score model is inadequate (misspecified or otherwise poorly estimated), including the full predictor vector allows for the possibility that the response surface model remains correctly specified.

The converse question, *Why bother with the propensity score if one is doing a high dimensional regression anyway?*, has been answered in the main body of this paper. Incorporating the propensity score (or another balancing score) yields a prior that can more readily adapt to complex patterns of confounding. In fact, in the context of response surface modeling for causal effects, failing to include an estimate of the propensity score (or another balancing score) can lead to additional bias in treatment effect estimates, as shown by the simple, low-dimensional example in Section 4.

8.3 Why not joint response-treatment modeling and what about uncertainty in the propensity score?

Using a presumptive model for Z to obtain $\hat{\pi}$ invites the suggestion of fitting a joint model for (Y, Z). Indeed, this is the approach taken in Hahn et al. (2016) as well as earlier papers, including Rosenbaum and Rubin (1983), Robins et al. (1992), McCandless

et al. (2009), Wang et al. (2012), and Zigler and Dominici (2014). While this approach is certainly reasonable, the Zellner prior approach would seem to afford all the same benefits while avoiding the distorted inferences that would result from a joint model if the propensity score model is misspecified (Zigler and Dominici, 2014).

One might argue that our Zellner prior approach gives under-dispersed posterior inference in the sense that it fails to account for the fact that $\hat{\pi}$ is simply a point estimate (and perhaps a bad one). However, this objection is somewhat misguided. First, as discussed elsewhere (e.g. Hill (2011)), inference on individual or subgroup treatment effects follows directly from the conditional distribution $(Y \mid Z, \mathbf{X})$. To continue our analogy with the more familiar Zellner g-prior, to model $(Y \mid Z, \mathbf{X})$ we are no more obligated to consider uncertainty in $\hat{\pi}$ than we are to consider uncertainty in $(\mathbf{X}'\mathbf{X})^{-1}$ when using a g-prior for on the coefficients of a linear model. Second, $\hat{\pi}$ appears in the model *along with* the full predictor vector x: it is provided as a hint, not as a certainty. This model is at least as capable of estimating a complex response surface as the corresponding model without $\hat{\pi}$, and the increase in variance incurred by the addition of one additional "covariate" can be more than offset by the bias reduction in the estimation of treatment effects.

On the other hand, we readily acknowledge that one might be interested in what inferences would obtain if we used different $\hat{\pi}$ estimates. One might consider fitting a series of BCF models with different estimates of $\hat{\pi}$, perhaps from alternative models or other procedures. This is a natural form of sensitivity analysis in light of the fact that the adjustments proposed in this paper only work if $\hat{\pi}$ accurately approximates π. However, it is worth noting that the available (z, x) data speak to this question: a host of empirically proven prediction methods (i.e. neural networks, support vector machines, random forests, boosting, or any ensemble method) can be used to construct candidate $\hat{\pi}$ and cross-validation may be used to gauge their accuracy. Only if a "tie" in generalization error (predicting Z) is encountered must one turn to sensitivity analysis.

8.4 Connections to doubly robust estimation

Our combination of propensity score estimation and outcome modeling is superficially reminiscent of doubly robust estimation (Bang and Robins, 2005), where propensity score and outcome regression models are combined to yield consistent estimates of finite dimensional treatment effects, provided at least one model is correctly specified. We do not claim our approach is doubly robust, however, and in all of our examples above we use the natural Bayesian estimates of (conditional) average treatment effects rather than doubly robust versions. Empirically these seem to have good frequency properties. The motivation behind our approach (the parameterization and including an estimate of the propensity score) is in fact quite different than that behind doubly robust estimation, as we are not focused on consistency under partial misspecification, nor obtaining parametric rates of convergence for the ATE/ATT, but rather on regularizing in a way that avoids RIC. To our knowledge, none of the literature on doubly robust estimators explicitly considers bias-variance trade-off ideas in this way.

Although it was not our focus here we do expect that BCF would perform well in the context of doubly robust estimation. For example, a TMLE-based approach using

SuperLearner with BART as a component of the ensemble was a top performer in the 2017 ACIC contest. BCF and ps-BART generally improve on vanilla BART, and should be useful in that context. As another example, Nie and Wager (2017) showed that using ps-BART (motivated by an early draft of this paper) as a component of a stacked estimator of heterogeneous treatment effects fitted using the *R*-learner yielded improved performance over the individual heterogeneous treatment effect estimators. We hope that researchers will continue to see the promise of BCF and related methods as components of estimators derived from frequentist considerations.

8.5 The role of theory versus simulations in methodological comparisons

Recent theory on posterior consistency and rates of posterior concentration for Bayesian tree models in prediction contexts (Linero and Yang, 2018; Rocková and van der Pas, 2017; Rocková and Saha, 2019) should apply to the BCF parametrization with some adaptation. However, the existing results require significant modifications to the BART prior that may make them unreliable guides to practical performance. Likewise, recent results demonstrate the consistency of recursive approximations to single-tree Bayesian regression models (He, 2019) in the setting of generalized additive models and these results are possibly applicable to BCF type parametrizations.

Despite only nascent theory — and none speaking to frequentist coverage of Bayesian credible intervals — BCF should be of interest to anyone seeking reliable estimates of heterogeneous treatment effects, for two reasons. The first is that many of the existing approaches to fusing machine learning and causal inference make use of first-stage regression estimates for which no dedicated theory is strictly necessary, for instance Künzel et al. (2019) and Nie and Wager (2017). In this context, BCF can be regarded as another supervised learning algorithm, alongside neural networks, support vector machines, random, forests, etc. and would be of special interest insofar as it obtains better first-stage estimates than these other methods.

The second, and more important reason, that BCF is an important development is its performance in simulation studies by us and others. Unlike many simulation studies, designed with the goal of showcasing a method's strengths, our simulation studies were designed prior to the model's development, with an eye towards realism. Specifically, our simulation protocol was created specifically to correct perceived weaknesses in previous synthetic data sets in the causal machine learning literature: no or very weak confounding, implausibly large treatment effects, and unrealistically large variation in treatment effects (including sign variation). By contrast, our data generating processes reflect our assumptions about real data for which heterogeneous treatment effects are commonly sought: strong confounding, and modest treatment effects and treatment effect heterogeneity (relative to the magnitude of unmeasured sources of variation). It was these considerations that led us to the concept of targeted selection (Section 4.2), for example.

By utilizing realistic, rather than convenient or favorable, data generating processes, our simulations are a principled approach to assessing the finite sample operating characteristics of various methods. Not only did this process reassure us that BCF has good

frequentist properties in the regimes we examined, but it also alerted us to what cold comfort asymptotic theory can be in actual finite samples, as methods with available theory did not perform as well as the theory would suggest. Finally, we would also note that other simulation studies carefully designed by other researchers reach similar conclusions (e.g. ACIC 2016, described above, and Wendling et al. (2018); McConnell and Lindner (2019)).

References

Athey, S., Tibshirani, J., Wager, S., et al. (2019). "Generalized random forests." *The Annals of Statistics*, 47(2): 1148–1178. MR3909963. doi: https://doi.org/10.1214/18-AOS1709. 967, 981, 984

Bang, H. and Robins, J. M. (2005). "Doubly robust estimation in missing data and causal inference models." *Biometrics*, 61(4): 962–973. MR2216189. doi: https://doi.org/10.1111/j.1541-0420.2005.00377.x. 978, 992

Breiman, L. (2001). "Random forests." *Machine learning*, 45(1): 5–32. MR3874153. 967

Carvalho, C. M., Polson, N. G, and Scott, J. G. (2010). "The horseshoe estimator for sparse signals." *Biometrika*, 97(2): 465–480. Oxford University Press. 981

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., et al. (2016). "Double machine learning for treatment and causal parameters." *arXiv preprint* arXiv:1608.00060. MR3769544. doi: https://doi.org/10.1111/ectj.12097. 967, 972

Chipman, H., George, E., and McCulloch, R. (1998). "Bayesian CART model search." *Journal of the American Statistical Association*, 93(443): 935–948. 970

Chipman, H. A., George, E. I., and McCulloch, R. E. (2010). "BART: Bayesian additive regression trees." *The Annals of Applied Statistics*, 266–298. MR2758172. doi: https://doi.org/10.1214/09-AOAS285. 967, 969, 970, 980, 981

Dorie, V. and Hill, J. (2017). *aciccomp2016: Atlantic Causal Inference Conference Competition 2016 Simulation*. R package version 0.1-0. 968, 984

Dorie, V., Hill, J., Shalit, U., Scott, M., Cervone, D., et al. (2019). "Automated versus do-it-yourself methods for causal inference: Lessons learned from a data analysis competition." *Statistical Science*, 34(1): 43–68. MR3938963. doi: https://doi.org/10.1214/18-STS667. 968, 969, 984

Efron, B. (2014). "Estimation and accuracy after model selection." *Journal of the American Statistical Association*, 109(507): 991–1007. MR3265671. doi: https://doi.org/10.1080/01621459.2013.823775. 967

Friedberg, R., Tibshirani, J., Athey, S., and Wager, S. (2018). "Local linear forests." *arXiv preprint* arXiv:1807.11408. MR3909963. doi: https://doi.org/10.1214/18-AOS1709. 967

Ganin, Y., Ustinova, E., Ajakan, H., Germain, P., Larochelle, H., Laviolette, F., Marchand, M., and Lempitsky, V. (2016). "Domain-adversarial training of neural net-

works." *The Journal of Machine Learning Research*, 17(1): 2096–2030. MR3504619. 967

Gelman, A. et al. (2006). "Prior distributions for variance parameters in hierarchical models (comment on article by Browne and Draper)." *Bayesian Analysis*, 1(3): 515–534. MR2221284. doi: https://doi.org/10.1214/06-BA117A. 980

Giles, D. and Rayner, A. (1979). "The mean squared errors of the maximum likelihood and natural-conjugate Bayes regression estimators." *Journal of Econometrics*, 11(2): 319–334. MR0555633. doi: https://doi.org/10.1016/0304-4076(79)90043-5. 971

Gramacy, R. B. and Lee, H. K. (2008). "Bayesian treed Gaussian process models with an application to computer modeling." *Journal of the American Statistical Association*, 103(483). MR2528830. doi: https://doi.org/10.1198/016214508000000689. 967

Green, D. P. and Kern, H. L. (2012). "Modeling heterogeneous treatment effects in survey experiments with Bayesian additive regression trees." *Public opinion quarterly*, nfs036. 969

Gustafson, P. and Greenland, S. (2006). "Curious phenomena in Bayesian adjustment for exposure misclassification." *Statistics in Medicine*, 25(1): 87–103. MR2222076. doi: https://doi.org/10.1002/sim.2341. 967

Hahn, P. R. and Carvalho, C. M. (2015). "Decoupling shrinkage and selection in Bayesian linear models: a posterior summary perspective." *Journal of the American Statistical Association*, 110(509): 435–448. MR3338514. doi: https://doi.org/10.1080/01621459.2014.993077. 988

Hahn, P. R., Dorie, V., and Murray, J. S. (2018). *Atlantic Causal Inference Conference (ACIC) Data Analysis Challenge 2017*. 968, 985, 986

Hahn, P. R., Puelz, D., He, J., and Carvalho, C. M. (2016). "Regularization and confounding in linear regression for treatment effect estimation." *Bayesian Analysis*. MR3737947. doi: https://doi.org/10.1214/16-BA1044. 966, 967, 971, 989, 991

Hansen, B. B. (2008). "The prognostic analogue of the propensity score." *Biometrika*, 95(2): 481–488. MR2521594. doi: https://doi.org/10.1093/biomet/asn004. 978

He, J. (2019). "Stochastic tree ensembles for regularized supervised learning." Technical report, University of Chicago Booth School of Business. 993

Heckman, J. J., Lopes, H. F., and Piatek, R. (2014). "Treatment effects: A Bayesian perspective." *Econometric reviews*, 33(1-4): 36–67. MR3170840. doi: https://doi.org/10.1080/07474938.2013.807103. 967

Hill, J., Su, Y.-S., et al. (2013). "Assessing lack of common support in causal inference using Bayesian nonparametrics: Implications for evaluating the effect of breastfeeding on children's cognitive outcomes." *The Annals of Applied Statistics*, 7(3): 1386–1420. MR3127952. doi: https://doi.org/10.1214/13-AOAS630. 969

Hill, J. L. (2011). "Bayesian nonparametric modeling for causal inference." *Journal of Computational and Graphical Statistics*, 20(1). MR2816546. doi: https://doi.org/10.1198/jcgs.2010.08162. 967, 969, 979, 985, 991

Imai, K., Ratkovic, M., et al. (2013). "Estimating treatment effect heterogeneity in randomized program evaluation." *The Annals of Applied Statistics*, 7(1): 443–470. MR3086426. doi: https://doi.org/10.1214/12-AOAS593. 978

Imai, K. and Van Dyk, D. A. (2004). "Causal inference with general treatment regimes: Generalizing the propensity score." *Journal of the American Statistical Association*, 99(467): 854–866. MR2090918. doi: https://doi.org/10.1198/016214504000001187. 986, 989

Imbens, G. W. and Rubin, D. B. (2015). *Causal Inference in Statistics, Social, and Biomedical Sciences*. Cambridge University Press. MR3309951. doi: https://doi.org/10.1017/CBO9781139025751. 968

Johnson, E., Dominici, F., Griswold, M., and Zeger, S. L. (2003). "Disease cases and their medical costs attributable to smoking: an analysis of the national medical expenditure survey." *Journal of Econometrics*, 112(1): 135–151. MR1963235. doi: https://doi.org/10.1016/S0304-4076(02)00157-4. 986

Kern, H. L., Stuart, E. A., Hill, J., and Green, D. P. (2016). "Assessing methods for generalizing experimental impact estimates to target populations." *Journal of Research on Educational Effectiveness*, 9(1): 103–127. 969

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. 967, 979, 993

Li, M. and Tobias, J. L. (2014). "Bayesian analysis of treatment effect models." In Jeliazkov, I. and Yang, X.-S. (eds.), *Bayesian Inference in the Social Sciences*, chapter 3, 63–90. Wiley. MR3307626. 967

Linero, A. R. and Yang, Y. (2018). "Bayesian regression tree ensembles that adapt to smoothness and sparsity." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 80(5): 1087–1110. MR3874311. doi: https://doi.org/10.1111/rssb.12293. 993

McCaffrey, D. F., Griffin, B. A., Almirall, D., Slaughter, M. E., Ramchand, R., and Burgette, L. F. (2013). "A tutorial on propensity score estimation for multiple treatments using generalized boosted models." *Statistics in Medicine*, 32(19): 3388–3414. MR3074364. doi: https://doi.org/10.1002/sim.5753. 967

McCaffrey, D. F., Ridgeway, G., and Morral, A. R. (2004). "Propensity score estimation with boosted regression for evaluating causal effects in observational studies." *Psychological Methods*, 9(4): 403. 967

McCandless, L. C., Gustafson, P., and Austin, P. C. (2009). "Bayesian propensity score analysis for observational data." *Statistics in Medicine*, 28(1): 94–112. MR2655553. doi: https://doi.org/10.1002/sim.3460. 991

McConnell, K. J. and Lindner, S. (2019). "Estimating treatment effects with machine learning." *Health services research*. 968, 993

Murray, J. S. (2017). "Log-Linear Bayesian Additive Regression Trees for Multinomial Logistic and Count Regression Models." *arXiv preprint* arXiv:1701.01503. MR3484083. doi: https://doi.org/10.1002/sta4.110. 967

Nie, X. and Wager, S. (2017). "Quasi-oracle estimation of heterogeneous treatment effects." *arXiv preprint* arXiv:1712.04912. 967, 992, 993

Polson, N. G., Scott, J. G., et al. (2012). "On the half-Cauchy prior for a global scale parameter." *Bayesian Analysis*, 7(4): 887–902. MR3000018. doi: https://doi.org/10.1214/12-BA730. 980

Powers, S., Qian, J., Jung, K., Schuler, A., Shah, N. H., Hastie, T., and Tibshirani, R. (2018). "Some methods for heterogeneous treatment effect estimation in high dimensions." *Statistics in medicine*, 37(11): 1767–1787. MR3799840. doi: https://doi.org/10.1002/sim.7623. 967

Robins, J. M., Mark, S. D., and Newey, W. K. (1992). "Estimating exposure effects by modelling the expectation of exposure conditional on confounders." *Biometrics*, 479–495. MR1173493. doi: https://doi.org/10.2307/2532304. 991

Robins, J. M. and Ritov, Y. (1997). "Toward a curse of dimensionality appropriate (CODA) asymptotic theory for semi-parametric models." *Statistics in medicine*, 16(3): 285–319. 990

Rocková, V. and Saha, E. (2019). "On Theory for BART." In *The 22nd International Conference on Artificial Intelligence and Statistics*, 2839–2848. 993

Rocková, V. and van der Pas, S. (2017). "Posterior concentration for Bayesian regression trees and forests." *Annals of Statistics (In Revision)*, 1–40. 993

Rosenbaum, P. R. and Rubin, D. B. (1983). "The central role of the propensity score in observational studies for causal effects." *Biometrika*, 41–55. MR0742974. doi: https://doi.org/10.1093/biomet/70.1.41. 990, 991

Roy, J., Lum, K. J., Zeldow, B., Dworkin, J. D., Re III, V. L., and Daniels, M. J. (2017). "Bayesian nonparametric generative models for causal inference with missing at random covariates." *Biometrics*. MR3908137. 967

Shalit, U., Johansson, F. D., and Sontag, D. (2017). "Estimating individual treatment effect: generalization bounds and algorithms." In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, 3076–3085. JMLR.org. 967

Sivaganesan, S., Müller, P., and Huang, B. (2017). "Subgroup finding via Bayesian additive regression trees." *Statistics in Medicine*. MR3660139. doi: https://doi.org/10.1002/sim.7276. 969, 988

Su, X., Kang, J., Fan, J., Levine, R. A., and Yan, X. (2012). "Facilitating score and causal inference trees for large observational studies." *Journal of Machine Learning Research*, 13(Oct): 2955–2994. MR2997717. 967

Taddy, M., Gardner, M., Chen, L., and Draper, D. (2016). "A Nonparametric Bayesian Analysis of Heterogenous Treatment Effects in Digital Experimentation." *Journal of Business & Economic Statistics*, 34(4): 661–672. MR3548002. doi: https://doi.org/10.1080/07350015.2016.1172013. 967

van der Laan, M. J. (2010a). "Targeted maximum likelihood based causal inference: Part I." *The International Journal of Biostatistics*, 6(2). MR2595112. doi: https://doi.org/10.2202/1557-4679.1241. 967

van der Laan, M. J. (2010b). "Targeted maximum likelihood based causal inference: Part II." *The International Journal of Biostatistics*, 6(2). 967

Wager, S. and Athey, S. (2018). "Estimation and inference of heterogeneous treatment effects using random forests." *Journal of the American Statistical Association*, 113(523): 1228–1242. MR3862353. doi: https://doi.org/10.1080/01621459.2017.1319839. 967

Wager, S., Hastie, T., and Efron, B. (2014). "Confidence intervals for random forests: The jackknife and the infinitesimal jackknife." *The Journal of Machine Learning Research*, 15(1): 1625–1651. MR3225243. 967

Wang, C., Parmigiani, G., and Dominici, F. (2012). "Bayesian effect estimation accounting for adjustment uncertainty." *Biometrics*, 68(3): 661–671. MR3055168. doi: https://doi.org/10.1111/j.1541-0420.2011.01731.x. 991

Wendling, T., Jung, K., Callahan, A., Schuler, A., Shah, N., and Gallego, B. (2018). "Comparing methods for estimation of heterogeneous treatment effects using observational data from health care databases." *Statistics in medicine*, 37(23): 3309–3324. MR3856345. doi: https://doi.org/10.1002/sim.7820. 968, 969, 993

Yang, Y., Cheng, G., and Dunson, D. B. (2015). "Semiparametric Bernstein-von Mises Theorem: Second Order Studies." *arXiv preprint* arXiv:1503.04493. MR3020419. doi: https://doi.org/10.1214/13-EJS768. 972, 977

Yeager, D. S., Hanselman, P., Walton, G. M., Murray, J. S., Crosnoe, R., Muller, C., Tipton, E., Schneider, B., Hulleman, C. S., Hinojosa, C. P., Paunesku, D., Romero, C., Flint, K., Roberts, A., Trott, J., Iachan, R., Buontempo, J., Yang, S. M., Carvalho, C. M., Hahn, P. R., Gopalan, M., Mhatre, P., Ferguson, R., Duckworth, A. L., and Dweck, C. S. (2019). "A national experiment reveals where a growth mindset improves achievement." *Nature*, 573(7774): 364–369. URL https://doi.org/10.1038/s41586-019-1466-y. 978

Zaidi, A. and Mukherjee, S. (2018). "Gaussian Process Mixtures for Estimating Heterogeneous Treatment Effects." *arXiv preprint* arXiv:1812.07153. 967

Zeger, S. L., Wyant, T., Miller, L. S., and Samet, J. (2000). "Statistical testimony on damages in Minnesota v. Tobacco Industry." In *Statistical Science in the Courtroom*, 303–320. Springer. 986

Zellner, A. (1986). "On assessing prior distributions and Bayesian regression analysis with g-prior distributions." *Bayesian inference and decision techniques: Essays in Honor of Bruno De Finetti*, 6: 233–243. MR0881437. 990

Zigler, C. M. and Dominici, F. (2014). "Uncertainty in propensity score estimation: Bayesian methods for variable selection and model-averaged causal effects." *Journal of the American Statistical Association*, 109(505): 95–107. MR3180549. doi: https://doi.org/10.1080/01621459.2013.869498. 967, 991

Invited Discussion

Arman Oganisian* and Jason A. Roy†

1 Introduction

We congratulate the authors for this important contribution to the causal inference literature. Even beyond the specific method that was proposed, we found their treatment and presentation of ideas stimulating and highly relevant. Overall, we think the Bayesian causal forest (BCF) is a compelling addition to the Bayesian causal inference toolkit.

A major focus of the paper is the development of sensible priors for estimating the conditional outcome mean, $f(x_i, z_i)$. For instance, we could place a single nonparametric prior over f or estimate $f(x_i, Z_i = 0)$ and $f(x_i, Z_i = 1)$ separately. Instead, the authors make a strong case for partitioning f into a prognostic score and treatment effect component: $f(x_i, z_i) = \mu(x_i) + \tau(x_i)z_i$. This allows placement of separate priors on $\mu()$ and $\tau()$ – which both have direct causal interpretations. In practice, each of these objects may require different degrees of regularization. As the authors suggest, the treatment effect is likely to be simple relative to the prognostic function, and we would therefore want to express a prior for $\tau()$ that "shrinks more strongly towards homogeneous effects" while keeping the prior for $\mu()$ relatively non-informative.

At first glance, estimation of conditional average treatment effects (CATEs) using f appears to be merely a prediction problem that can be tackled with any of the available array of flexible machine learning methods. After all, we are simply contrasting predictions under varying interventions on Z. However, these prediction methods often regularize heavily to penalize overfitting and, therefore, attain better out-of-sample predictions. This class of models includes BART – which is regularized to favor shallow trees. This paper and the earlier paper by Hahn et al. (2018) demonstrate that direct application of these regularized methods to causal estimation can lead to regularization-induced confounding (RIC). Drawing attention to this phenomenon is a major contribution of the paper, especially because the use of regularized models are so widespread in causal inference.

Hahn et al. (2018) motivate this regularization from the perspective of "the set of candidate control variables is often quite large relative to the available sample size." That is, even if we know that the true model is linear, a very high-dimensional x will necessitate regularization. On the other hand, if x is low-dimensional, leaving the form unspecified will require some regularization. In the former, we regularize to constrain the large covariate space while in the latter we regularize to constrain the very large function space. Bayesian nonparametric (BNP) models such as BART are highly parameterized and require regularization in either of these settings.

*Department of Biostatistics, Epidemiology, and Informatics, University of Pennsylvania, Philadelphia, PA, USA, aoganisi@upenn.edu

†Department of Biostatistics and Epidemiology, Rutgers University, Piscataway, NJ, USA, jason.roy@rutgers.edu

The authors also emphasize that "targeted selection" can exacerbate RIC. Suppose Z is highly correlated with a function of the X that also happens to be highly correlated with the prognostic score (mean of outcome under no treatment). Essentially, this introduces a collinearity problem. Further, suppose $\mu()$ is a complex function. Regularization may prevent our estimate of the prognostic score from adequately describing the complexity of $\mu()$. At the same time, collinearity with Z may lead to biased posterior inference. The authors propose including the propensity score as one of the predictors in the $\mu()$, which they argue should improve the performance of the models when treatment is targeted towards improving mean potential outcome under no treatment.

In the following sections, we discuss in more depth some aspects of the work and some different considerations. We accept that specification of the function $f(x_i, z_i)$ as a sum of a prognostic function and treatment effect function is reasonable. We then consider other questions: Do we expect that targeted selection is common in practice? What other selection mechanisms are likely? How might this approach fare when there are positivity violations? Finally, what are implications for using BART as a prior rather than other priors for functions? We hope that discussion around these topics will provide readers with deeper insights into BCF.

2 Targeted Selection

The authors defined targeted selection as the situation where "treatment is assigned based on a prediction of the outcome in the absence of treatment, given measured covariates." Under targeted selection the treatment selection mechanism (i.e. the propensity score) is a function of the prognostic score. Now, consider including the propensity score in the $\mu()$ function, $f(x_i, z_i) = \mu(x_i, \widehat{\pi}(x_i)) + \tau(x_i)z_i$. There are two things to note. First, $\mu()$ should be less complex than if we had not included the propensity score. Essentially, $\hat{\pi}$ is itself a complex function of the x that should be predictive of the outcome under no treatment, giving us a helpful "hint" (authors' term). In absence of $\hat{\pi}$, BART would need to work harder (i.e. introduce more splits) to capture this complexity. Second, note that the propensity score is a balancing score, i.e., $X \perp\!\!\!\perp Z|\pi(x)$. Conditioning on the propensity score should therefore reduce collinearity between Z and the X in $\mu()$. So it seems there is some intuitive reason to believe that, under targeted selection, including the propensity score will be beneficial. Indeed, the authors demonstrated that with several simulation studies.

This does beg the question of whether targeted selection is likely in practice. It could be the case, for example, that a clinician might be more likely to assign someone to the treatment group if they think the outlook in absence of treatment, $E(Y(0)|X)$, is not favorable. In their nonlinear simulation scenario in Section 6.1, there is a very strong monotone relationship between the prognostic score and propensity score (Figure 1). Bayesian Causal Forest (BCF) seems to perform well in these situations with strong targeted selection.

However, treatment decisions are ideally made based on trying to predict $E(Y(1) - Y(0)|X)$. For simplicity, suppose X is a measure of patient frailty, Z is surgery, and Y

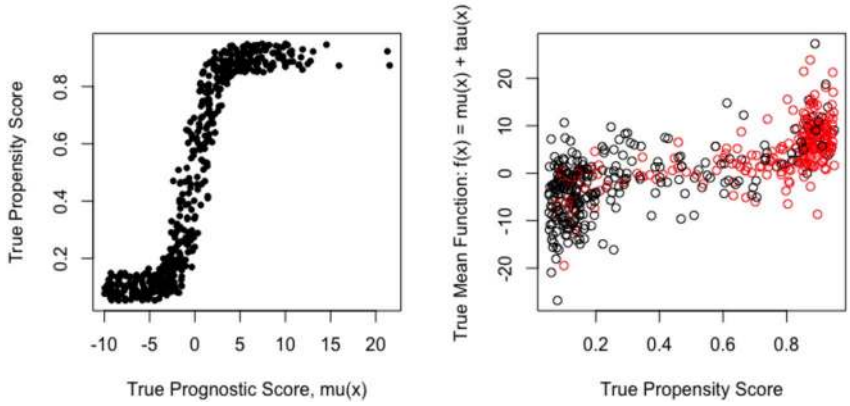


Figure 1: Summary of true functions for nonlinear simulation scenario of Section 6.1 of Hahn et al. (2020). We generated 500 observations from the assumed model. The plot on the left are realizations of the true propensity score versus the true prognostic score. The plot on the right is the true propensity score versus the true mean function. Red points are subjects in the treated group.

is survival. We might expect to see a prognostic score that is indeed monotone (worse prognosis as X increases), but a propensity score that is parabolic in frailty – increasing at first and then decreases when frailty gets too large. This is because after some level of frailty, risk of surgery overwhelms the benefit. In the extreme case, it is true that expected survival without surgery is quite low but a very frail patient could die during surgery. Another example is cancer treatment. Treatment decisions ideally target expected *differences* in quality adjusted life years under treatment or no treatment, rather than expectation under no treatment. When there are many confounders involved, it seems even less likely to us that there would be a clean story of targeted selection. In the next section we consider some issues that could arise under different selection mechanisms. In particular, we consider selection mechanisms that are driven by protocol.

3 BCF Performance in Non-Overlap Regions

Much of the simulation work in Hahn et al. (2020) assumes a targeted selection mechanism where the propensity score $\pi(x)$ is a (approximately) monotone function of the prognostic score $\mu(x) = E[Y(0) \mid x]$. However, highly irregular selection mechanisms can arise in the real-world settings. As a biomedical example, consider that cancer decisions are often made using standardized treatment protocols based on a handful of patient characteristics such as cancer stage and age. This type of selection mechanism can lead to positivity violations. For instance, the physician may hesitate to defy the protocol for some subgroup of patients – even if they believe $\mu(x)$ to be very poor. This creates regions of non-overlap, $\mathcal{X}^{NO} \subset \mathcal{X}$, where $P(Z = z \mid X = x) \approx 1$ for $x \in \mathcal{X}^{NO}$ for

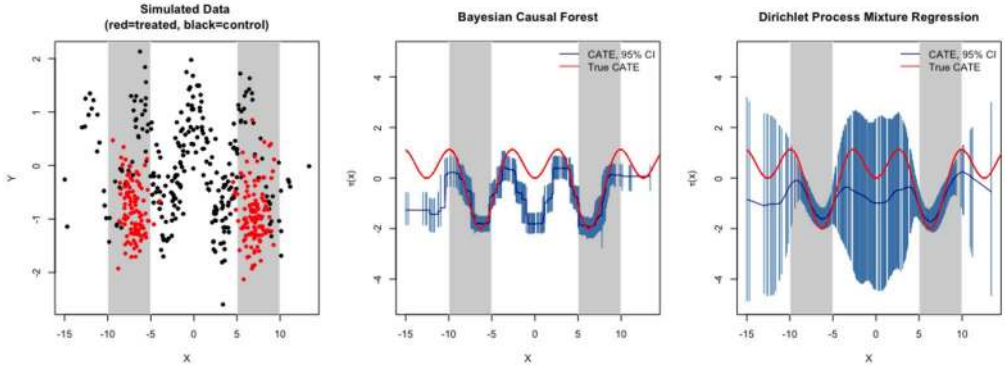


Figure 2: (left) Data simulated with positivity violations with overlap regions shown in gray. (center) BCF estimate of $\tau(x)$ – the conditional average treatment effect – is just as uncertain in overlap and non-overlap regions. The step-function is seen most clearly at the edges and near $X = 0$, where BART predicts the same, constant outcome mean. (right) Alternatively, DPM regression allows for higher uncertainty in the non-overlap regions while being smoother at the edges and near $X = 0$.

some treatment value z. Since there are (nearly) only subjects of one treatment group in these regions, BCF is forced to extrapolate the conditional mean outcome under the other treatment in this region. Incorrect extrapolation can lead to poor point and interval estimates. This problem is aggravated by the tension between satisfying ignorability and positivity (Cole and Hernán, 2008). That is, we often aim to condition on as many covariates, X, as possible to reduce bias from ignorability violations. However, the more covariates we condition on the more likely non-overlap regions will exist in some strata of the very high-dimensional X – leading to bias and variance from positivity violations. For all these reasons it is imperative to assess how BCF behaves in the presence of non-overlap.

Figure 2 (left) shows simulated data with some extreme positivity violations, for illustrative purposes. On the left panel, we see that the treated and untreated subjects overlap in the gray regions. The non-overlap regions are the complement of the gray regions, where there are mostly subjects of only one treatment group. Since BCF is BART-based, it inherits the (1) step-function behavior and (2) constant error variance of BART. Due to (1), BART favors extrapolating in these regions with a constant value across even substantially different x. This is undesirable if, for instance, we believe that patients with event slightly different x values have slightly different outcome means – which we think is a reasonable belief. The center panel of Figure 2 shows the BCF estimate of the CATE, $\tau(x)$, across x. We see that at the edges ($X < -10$ and $X > 10$) and near $X = 0$, BART is more or less predicting a constant value. Due in part to (2), $Var[\epsilon] = \sigma^2$ is constant with respect to X. This prevents model uncertainty from varying within $\mathcal{X}^{NO}$. However, intuitively, uncertainty about the treatment effect should be greater the "deeper" x_i is in the non-overlap region. The center panel of Figure 2 shows that, despite having no data on the conditional mean outcome among treated

subjects in the non-overlap regions, BCF posterior intervals remain about as narrow as in the overlap regions in this example.

For comparison, the right panel of Figure 2 shows the estimate of $\tau(x)$ from a Dirichlet process mixture (DPM) regression – which has become popular in the causal literature (e.g. Xu et al., 2018; Oganisian et al., 2020). Since it is a mixture model, the conditional mean function of the DPM is a *smooth* function of the covariates, not a step function. Moreover, it makes no homoskedasticity assumption. Both of these traits address points (1) and (2). Even in the edges and near $X = 0$, we allow for modest differences in $\tau(x)$ for modest differences x. Moreover, the posterior intervals in the non-overlap regions are substantially wider – reflecting increased uncertainty due to extrapolation. Note that both models seem to perform about as well in the overlap regions – but much work is needed to rigorously compare BART with other models in these settings.

We speculate that BCF performance can be improved in these settings with only minor adjustments to the underlying BART models. For instance, Nethery et al. (2019) augment BART with a spline in order to better extrapolate in non-overlap regions. The use of the spline mitigates the step-function behavior. Recent work by George et al. (2018) proposes relaxing the homoskedasticity assumption of BART. They propose the regression $E[Y \mid X] = f(X) + \epsilon$ with $f \sim BART$ and $\epsilon \sim G$. However, instead of setting $G = N(0, \sigma^2)$, they place a Dirichlet Process prior on G, $G \sim DP(\alpha, G_0)$ – yielding more flexible uncertainty estimates. The "soft" decision trees of Linero and Yang (2018), which enforce smoothness, is another promising modification to the standard BART algorithm that could be worth exploring.

A non-modeling alternative may be to find the non-overlap regions and trim accordingly. However, trimmed estimators target a causal effect only among those in the overlap regions, who may not be representative of the target population. Moreover, it can be practically difficult to determine these non-overlap regions or even diagnose positivity violations. In high-dimensional settings, visual inspection as in Figure 2 is not feasible. Histograms of the estimated propensity scores, being 1-dimensional summary of the covariates, can be used instead. In this approach, a histogram of the scores for treated subjects is overlaid on the histogram of the scores for untreated subjects. Subjects are included in the analysis if their propensity scores are in the region where these two histograms overlap. However, this method crucially depends on the estimate of the propensity score model and, more importantly, non-overlap regions in the covariate space can be difficult to see after dimension-reducing via the propensity score model. This is especially true if the treatment/exposure is rare. For these reasons, it may be ideal to instead have a model that performs reasonably in these regions by default. Finally, one might claim that positivity is not really a central issue since it is meaningless to even speak of the causal effect in non-overlap regions – the effect is undefined. While this argument perhaps works for *structural* positivity violations, it does not address random violations of positivity. For instance, due to sampling difficulties we may sample too few treated Native American subjects. However, this does not imply the CATE for Native Americans is undefined or does not exist.

4 Some Alternatives to BART

As alluded to in the introduction, if we are interested in $f(x_i, z_i) = \mu(x_i) + \tau(x_i)z_i$, we could specify other priors appropriate for functions to $\mu()$ and $\tau()$. While the BART prior has been shown to perform well in a variety of practical settings, it does have some drawbacks. In particular, it does not easily accommodate prior beliefs about the function.

Alternatively, a Gaussian process (GP) prior is both quite flexible and allows more intricate prior specifications. For instance, we could center the GP prior for μ around a linear, additive function $\mu_0(x_i) = x_i'\beta$. The posterior can depart from this prior as demanded by the complexity of the data – but can shrink back towards this parametric model in regions where the data is sparse. Additionally, different choices of GP covariance can accommodate different prior beliefs about smoothness. For example, the squared exponential specification allows for differences in outcome means to be proportional to squared differences in covariate vectors. Similarly a GP prior for τ could be centered on a constant function $\tau_0(x_i) = \theta$. This expresses the prior belief in a homogeneous treatment effect, while allowing for posterior deviations away from this prior towards more flexible functional forms.

Another BNP prior worth mentioning is a DP mixture mentioned in the previous section. In the next few paragraphs, we propose a particular version of a DP mixture that would allow for incorporating a (local) propensity score. Consider the following joint model for the outcome and treatment, $p(Y_i, Z_i \mid X_i)$, expressed hierarchically as

$$\begin{aligned} Y_i \mid Z_i, X_i, \omega_i \sim & p(Y|Z, X, \omega_i) \\ Z_i \mid X_i, \theta_i \sim & \pi(Z \mid X, \theta_i) \\ (\omega_i, \theta_i)_{1:n} \mid G \sim & G. \end{aligned} \tag{4.1}$$

Above we specified a "local" conditional outcome model, $p(Y \mid Z, X, \omega_i)$, as a function of treatment and covariates, followed by a propensity score model, $\pi(Z \mid X, \theta_i) = P(Z = 1 \mid X, \theta_i)$. For instance, the outcome model could be a Gaussian with a linear regression and unknown variance. The local propensity score model could be a logistic regression for instance. This is reminiscent of the two-stage approach in BCF that estimates both a propensity score and outcome model. It is important to note that the local outcome model does not explicitly condition on the propensity score which may alleviate or avoid feedback issues (Zigler et al., 2013).

In Equation (4.1) we have saturated the model with more parameters than observations: subject-specific parameter vectors govern the outcome model, ω_i, and propensity score model, θ_i. Now say these parameters live in the space $(\omega, \theta) \in \Omega \times \Theta$ and G is some unknown probability distribution over this space. We consider the Enriched Dirichlet Process (EDP) prior (Wade et al., 2011) on G. EDP-based models have been employed in Bayesian nonparametric causal modeling with some success (Roy et al., 2018), but here we use it to model the propensity score. Specifically, we assume

$$G \sim EDP(\alpha_\omega, \alpha_\theta, G_0).$$

The EDP is a stochastic process with each realization being a random joint distribution, G, over $\Omega \times \Theta$, centered around a base distribution G_0. The base distribution along with concentration parameters α_ω and α_θ are hyperparameters. We refer the reader to Wade et al. (2011) and Wade et al. (2014) for a detailed discussion of these parameters, as they are not central to our presentation. What is important is that this is a highly flexible prior and the two concentration parameters allow us to control the degree of regularization on the outcome model (via α_ω) and propensity score model (via α_θ), separately. Moreover, marginalizing over the parameters yields the induced conditional outcome regression

$$E[Y \mid Z, X, G] = \int \frac{\pi(Z \mid X, \theta)}{\int \pi(Z \mid X, \theta) dG(\omega, \theta)} \cdot E\Big[Y \mid Z, X, \omega\Big] \; dG(\omega, \theta). \tag{4.2}$$

The induced regression is a flexible, nonlinear function of covariates. Specifically, it is a mixture of "local" outcome regressions with propensity score-dependent mixture weights, $\frac{\pi(Z|X,\theta)}{\int \pi(Z|X,\theta) dG(\omega,\theta)}$. Consider a parameter pair (ω, θ). When estimating mean outcome under treatment $Z = z$, local outcome models governed by ω are given more weight if the probability of treatment z is higher under the corresponding θ parameter, relative to other potential θs.

Notice that nowhere did we condition on the propensity score. Rather, it appears naturally in the resulting outcome regression. Uncertainty about both the propensity score and outcome models is captured in full by the posterior. This is in contrast to the BCF approach where the propensity score estimate is treated as fixed. The authors argue that this estimate, $\hat{\pi}(x_i)$, is fully a function of $\{Z_i, X_i\}_{1:n}$ – which is treated as fixed in the model. Therefore uncertainty in the propensity score need not be considered. However, suppose π was estimated using the same data via the same logistic regression. Even in this case, differences in prior uncertainty about the regression parameters impact uncertainty about $\hat{\pi}$. This, in turn, may lead to different degrees of uncertainty about $f(x_i, z_i, \hat{\pi}(x))$. We agree that in practice this is likely not a big issue, but it does seem unsatisfying from a Bayesian perspective. We speculate that perhaps other BNP priors such as the one above are more naturally suited to this task.

To conclude, different nonparametric priors have different advantages and we hope consideration of some other BNP priors will provide readers with relevant context around BCF – which also has desirable properties. For instance, one limitation of the proposed EDP approach is the difficulty in specifying a prior on the treatment effect apart from the other components of the conditional outcome model. However, this is precisely one of the strengths of BCF. Similarly, while Gaussian processes may induce smoothness in the regression, it could be argued BART-based models are easier to implement in practice and work well off-the-shelf with minimal tuning. Future work rigorously comparing such BNP priors head-to-head in a variety of causal settings would a useful addition to the literature.

References

Cole, S. R. and Hernán, M. A. (2008). "Constructing inverse probability weights for marginal structural models." *American Journal of Epidemiology*, 168(6): 656–664.

URL https://doi.org/10.1093/aje/kwn164 1001

George, E., Laud, P., Logan, B., McCulloch, R., and Sparapani, R. (2018). "Fully Nonparametric Bayesian Additive Regression Trees." 1002

Hahn, P. R., Carvalho, C. M., Puelz, D., and He, J. (2018). "Regularization and confounding in linear regression for treatment effect estimation." *Bayesian Anal.*, 13(1): 163–182. URL https://doi.org/10.1214/16-BA1044 MR3737947. doi: https://doi.org/10.1214/16-BA1044. 998

Hahn, P. R., Murray, J. S., and Carvalho, C. M. (2020). "Bayesian regression tree models for causal inference: regularization, confounding, and heterogeneous effects." Advance publication. URL https://doi.org/10.1214/19-BA1195 1000

Linero, A. R. and Yang, Y. (2018). "Bayesian regression tree ensembles that adapt to smoothness and sparsity." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 80(5): 1087–1110. URL https://rss.onlinelibrary.wiley.com/doi/abs/10.1111/rssb.12293 MR3874311. doi: https://doi.org/10.1111/rssb.12293. 1002

Nethery, R. C., Mealli, F., and Dominici, F. (2019). "Estimating population average causal effects in the presence of non-overlap: The effect of natural gas compressor station exposure on cancer mortality." *Ann. Appl. Stat.*, 13(2): 1242–1267. URL https://doi.org/10.1214/18-AOAS1231 MR3963570. doi: https://doi.org/10.1214/18-AOAS1231. 1002

Oganisian, A., Mitra, N., and Roy, J. A. (2020). "A Bayesian nonparametric model for zero-inflated outcomes: prediction, clustering, and causal estimation." *Biometrics. (to appear)*. URL https://doi.org/10.1111/biom.13244 1002

Roy, J., Lum, K. J., Zeldow, B., Dworkin, J. D., Re III, V. L., and Daniels, M. J. (2018). "Bayesian nonparametric generative models for causal inference with missing at random covariates." *Biometrics*, 74(4): 1193–1202. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/biom.12875 MR3908137. doi: https://doi.org/10.1111/biom.12875. 1003

Wade, S., Dunson, D. B., Petrone, S., and Trippa, L. (2014). "Improving prediction from Dirichlet process mixtures via enrichment." *Journal of Machine Learning Research*, 15(30): 1041–1071. URL http://jmlr.org/papers/v15/wade14a.html MR3195338. 1004

Wade, S., Mongelluzzo, S., and Petrone, S. (2011). "An enriched conjugate prior for Bayesian nonparametric inference." *Bayesian Anal.*, 6(3): 359–385. URL https://doi.org/10.1214/11-BA614 MR2843536. doi: https://doi.org/10.1214/ba/1339616468. 1003, 1004

Xu, D., Daniels, M., and Winterstein, A. (2018). "A Bayesian nonparametric approach to causal inference on quantiles." *Biometrics*, 74: 986–996. MR3860719. doi: https://doi.org/10.1111/biom.12863. 1002

Zigler, C. M., Watts, K., Yeh, R. W., Wang, Y., Coull, B. A., and Dominici, F. (2013). "Model feedback in Bayesian propensity score estimation." *Biometrics*, 69(1): 263–273. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1541-0420.2012.01830.x MR3058073. doi: https://doi.org/10.1111/j.1541-0420.2012.01830.x. 1003

Invited Discussion

Georgia Papadogeorgou* and Fan Li†

We congratulate Hahn, Murray and Carvalho (henceforth referred to as HMC) for their important contribution to the growing field of *Bayesian causal inference*. The authors tackle the problem of estimating treatment effect heterogeneity through conditional average treatment effects (CATE), a hard problem that can be framed within the context of high dimensional causal inference and multiple hypothesis testing. HMC adopt a flexible outcome model and they discuss the importance of prior specifications within the Bayesian framework that achieve a two-fold goal: appropriate confounding adjustment for unbiased effect estimation, and sufficient flexibility for estimation of heterogeneity. We regard the separation of these two components within the modeling framework as the most attractive feature of the proposed approach. First, the authors stress the importance of including the estimated propensity scores in the outcome model for more accurate confounding adjustment. Even though the use of propensity scores in Bayesian causal inference is subject to debate (see Section 1), we believe it is important to stress the central role of the propensity score in causal inference, irrespective of the mode of inference. Second, we believe that estimation of CATE is rightfully placed within the Bayesian framework in which the model formulation allows for heterogeneity along any covariate while shrinking small signals towards homogeneity. By viewing estimation of CATE within the scope of multiple testing, the Bayesian framework provides an inherent and automatic way for multiplicity control (Scott and Berger, 2010), while acknowledging and documenting uncertainty in estimating heterogeneity along all covariates.

With our discussion, we hope to shed light to the following aspects of this approach: (1) What is the role of the propensity score in Bayesian causal inference? (2) How does shrinkage towards homogeneity drive estimation of CATE? (3) What is the interplay between the choice of non-parametric prior distribution and limited covariate overlap in terms of uncertainty quantification in the estimation of CATE?

1 The role of propensity score in Bayesian causal inference and double-robustness

There has been a long debate of the role of propensity score in Bayesian causal inference (Sims, 2006; McCandless et al., 2009; Robins and Wasserman, 2012; Zigler et al., 2013; Robins et al., 2015). On one hand, under the assumption of ignorable assignment mechanism (and the parameters of the propensity score model and outcome model are *a priori* independent and distinct) (Ding and Li, 2018), the propensity score drops out from the likelihood of the outcomes, and therefore in principle does not matter in

*Assistant Professor, Department of Statistics, University of Florida, Gainesville, FL 32611, gpapadogeorgou@ufl.edu; url: https://gpapadogeorgou.netlify.app

†Associate Professor, Department of Statistical Science, Duke University, Durham, NC 27708, fl35@duke.edu; url: http://www2.stat.duke.edu/ fl35

Bayesian causal inference. On the other hand, vast empirical evidences (including HMC) suggest that proper inclusion of the propensity score substantially improves Bayesian causal inference. To understand the reason of adding the propensity score as a predictor in the outcome model, it is important to first clarify this seemingly paradoxical phenomenon. First, Robins et al. (2015) pointed out that the propensity score, as a one-dimensional summary of the covariates, is crucial for dimension reduction in causal inference with high dimensional data, irrespective of the mode of inference. A second important insight comes from the frequentist's perspective. Belloni et al. (2014) show that good performance in predicting either the observed outcome or the propensity score alone does not necessarily translate into good performance in estimating the causal effects. Chernozhukov et al. (2018) further pointed out that it is necessary to combine high-dimensional (e.g. machine learning) models for estimating the propensity score and outcome models in order to achieve $\sqrt{N}$ consistency in estimating the average treatment effect. These insights speak to the necessity of combining propensity score and outcome models for estimating causal effects in high-dimensional settings. This is closely related to the class of double-robust (DR) estimators (Scharfstein et al., 1999; Lunceford and Davidian, 2004; Bang and Robins, 2005). An estimator is DR if it is consistent as long as either the propensity score model or the outcome model, but not necessarily both, is correctly specified. Though the concept of DR was originally developed in the form of inverse probability weighting, many different forms have since been proposed, the essence of which is to augment a nonparametric propensity score estimator (e.g. weighting, stratification, or matching) by an outcome model or vice versa.

We view HMC's proposal as a Bayesian analogue of DR. A similar frequentist proposal, due to Rod Little and co-authors, uses the propensity score (in the form of penalized splines) as an additional predictor in the outcome model (e.g. Little and An, 2004; Zhou et al., 2019). In a sense, both HMC and Little's methods can be viewed as a continuous version of the mixed approach of combining propensity score stratification and outcome modeling (Gutman and Rubin, 2013): conducting an outcome regression within the stratum defined by a specific propensity score value. While the theoretical and empirical advantages of the DR estimators over non augmented-estimators have been well established in the frequentist's paradigm (e.g., Funk et al., 2011), HMC are among the first to do so in the Bayesian paradigm (another example is Antonelli et al., 2018), which also naturally expands the domain of DR from estimating average treatment effects to CATEs. Moving forward, we believe it would be worthwhile to rigorously define the Bayesian analog of DR (a definition of which is given by Antonelli et al., 2018) and prove the DR property of the specific prior. This will allow analysts to generalize from the BART prior to other Bayesian nonparametric priors (see Section 3).

2 Homogeneity-induced bias: The bias-variance trade-off of compromising between separate and simultaneous modeling

BART has a number of advantages, e.g. its robustness to the choice of hyperparameters, making it an attractive tool for causal effect estimation. There is substantial merit in

understanding *how BART can be used best* for estimating CATEs. In this section, we dive into the implications of imposing a prior distribution that shrinks towards homogeneity of treatment effects.

For simplicity, we consider a single covariate X. We focus on the CATE $\tau(x) = E[Y_i(1) - Y_i(0)|X_i = x]$ and on three ways to use BART: (1) adopt independent BART priors to model Y given X within the treated and control groups (refered to as separate model), (2) specify $E[Y|Z, X]$ using a BART prior (refered to as simultaneous model), and (3) the proposed approach by HMC. The separate model provides full flexibility for estimating heterogeneous treatment effects since the mean potential outcome under each treatment arm is modeled separately and without any sharing of information between treated and control groups. In contrast, the simultaneous model uses *the same* tree structure for estimating the two potential outcomes, and treatment effect heterogeneity is described via tree branches that split *both* on the treatment *and* the covariate. We view the approach of HMC as a "compromise" between separate and simultaneous modeling. By adopting two BART priors, the authors allow for flexible modeling of $E[Y(0)|X]$ and penalize deviations of $E[Y(1)|X]$ towards homogeneity of treatment effects. In that sense, their model formulation can be thought of as including a tree structure that forces the first split along Z. Therefore, the proposed approach shares information among treated and control groups, but ensures that shrinkage of estimates of $\tau(x)$ are towards homogeneity.

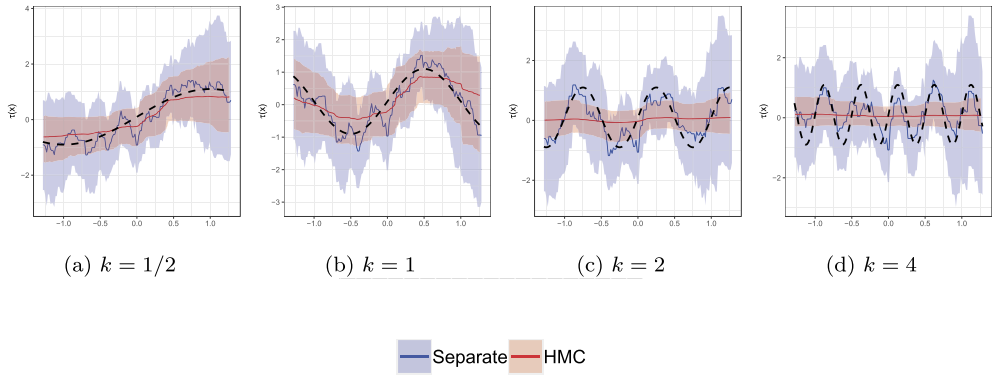
We now use a simple simulation example to compare the flexibility of the separate model versus the HMC method in estimating the CATE. We generate the covariate X from $X \sim N(0, 1)$, and generate the treatment assignment and the outcome from

$$Z \mid X \sim \text{Bernoulli}(p(X)),\ \text{logit}(p(X)) = 2X$$
$$Y \mid Z, X \sim \text{N}(0.1Z + X^2 + Z\sin(kX\pi), 1).$$

In this case, the true CATE $\tau(x) = 0.1+\sin(kx\pi)$. We consider values of $k \in \{1/2, 1, 2, 4\}$ to describe increasingly complex CATE. The dashed black lines in Figure 1 show the true values of $\tau(x)$. We estimate $\tau(x)$ based on a sample of size 300, and for values of x for which covariate overlap exists (within the 0.1 and 0.9 quantiles of the standard normal distribution). We use the standard implementation of BART in the `BayesTree` R package for the separate model, and the R package `bcf` by HMC with the true propensity score.

Figure 1 shows estimates of $\tau(x)$ and pointwise 95% credible intervals for the separate and HMC models. When $\tau(x)$ varies smoothly as a function of x ($k \in \{1/2, 1\}$), the HMC model is more efficient in estimating the CATE than the separate model. However, when $\tau(x)$ varies quickly with x ($k \in \{2, 4\}$), HMC's model mistakenly shrinks effect estimates towards homogeneity. These results illustrate that, compared to the separate model, the HMC model has lower variance, but has the potential to lead to *homogeneity-induced bias*. Even though this bias is expected to be eliminated as sample size increases, this simple example suggests that the homogeneity-inducing BART prior may potentially lead to large biases in the presence of a complex heterogeneity structure.

In most applications, complex heterogeneity across a *single* covariate like the one in Figure 1 for $k \in \{2, 4\}$ is unlikely to exist. However, these scenarios are representative



(a) $k = 1/2$ (b) $k = 1$ (c) $k = 2$ (d) $k = 4$

Figure 1: Estimates of heterogeneous treatment effects $\tau(x)$ and pointwise 95% credible intervals for the separate model (blue) and the model of HMC (red) based on a sample of size 300 and increasing complex heterogeneity structure. The dashed black lines show the true value of $\tau(x)$.

of situations with *many* covariates for which heterogeneity along each covariate separately might be small, but the true heterogeneity structure includes interactions among multiple covariates. In these situations, the homogeneity-inducing prior is expected to fail to identify the true complexity of HTE. In that sense, if the "subgroup" for which treatment is most effective is defined in terms of multiple subject characteristics, we suspect that the homogeneity-inducing prior would lead to estimates of $\tau(x)$ that fail to identify it. This bias-variance trade-off between CATE complexity and efficient estimation will exacerbate when there is lack of covariate overlap and/or limited sample size.

3 Choice of Bayesian nonparametric prior distribution

BART is a member of the general class of Bayesian nonparametric priors. A natural question would be "what about other priors?" In the context of CATE estimation, a particular relevant issue is (covariate) overlap and uncertainty quantification. Specifically, HMC demonstrated that the advantages in terms of point estimates of their BART prior over several other methods vary according to the degree of confounding, which is commonly referred to as overlap in the standard terminology of causal inference. Heuristically speaking, overlap means the similarity of the covariate distributions between the treatment and control groups. Overlap is a key concept that separates causal inference from traditional two-sample inference. In the region with good overlap (or low confounding in HMC's terminology), different specification of the outcome model usually leads to similar causal estimates. In contrast, in the region with poor overlap, the uncertainty of causal estimation is much higher and the results are sensitive to the specification of the outcome model. Thus, besides point estimation, an important criterion in choosing the outcome model for CATE estimation is proper uncertainty quantification.

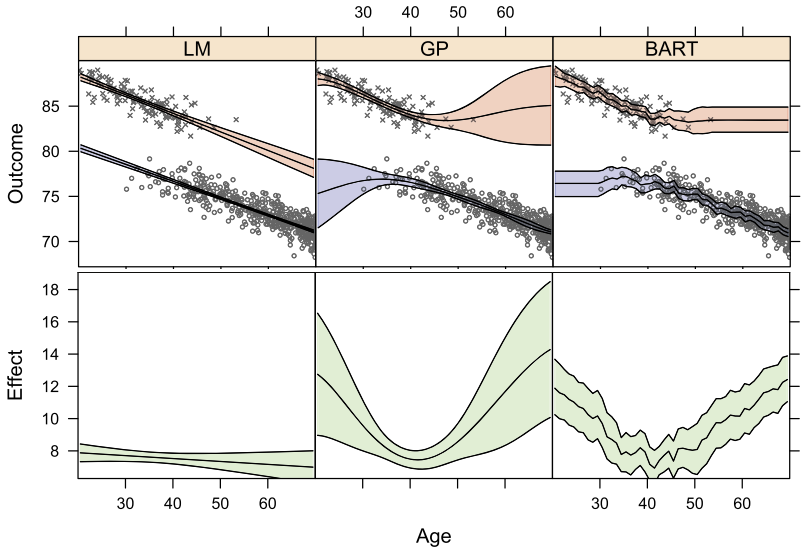


Figure 2: Estimates of missing counterfactuals (upper panel) and CATE (lower panel) and corresponding uncertainty band as a function of the single covariate 'Age' by three different models: linear model (LM); Gaussian Process (GP); BART, in the example of *poor overlap (or high confounding)* in Section 3. ×: treated units; ○: control units.

Given its underlying tree structure, intuitively BART may not have the flexibility to capture the additional uncertainty in regions of poor overlap, whereas some other "smoother" Bayesian nonparametric models such as the Gaussian Process may fare better. We demonstrate this through a simple example due to Surya Tokdar (who grants us the permission). Consider a scenario where a single covariate 'age' influences both treatment assignment and a continuous outcome; younger people are more likely to receive the treatment and higher outcome scores. Specifically, we generate a sample with 200 treated units and 300 control units. The only covariate (X_i) follows a Gamma distribution with mean 60 and 35 in the control and treatment group, respectively, and with standard deviation 8 in both groups. The true outcome model is as follows:

$$Y_i = 90 \cdot \mathbf{1}\{Z_i = 1\} + 82 \cdot \mathbf{1}\{1 - Z_i\} - 0.2X_i + \epsilon_i,\ \epsilon_i \sim N(0, 1).$$

As shown in the data cloud by group in Figure 2, the above data generating process creates a case where the degree of overlap varies significantly across the range of the single covariate: there is good overlap around age of 40, but lack of overlap increases as we move to the two tails of the range of age.

To estimate CATE, we fit the following outcome model separately in the treatment and control groups: $Y_i = f(X_i) + \epsilon_i$, $\epsilon_i \overset{iid}{\sim} N(0, \sigma^2)$. We choose three prior specifications for $f(x)$: (1) a BART prior similarly to Hill (2011) and HMC but without the propensity score; (2) a linear model with Gaussian prior: $f(x) \sim N(\beta x, \delta^2)$; (3) a GP prior (Rasmussen, 2003) with the covariance function specified using a Gaus-

sian kernel with signal-to-noise ratio parameter ρ and inverse-bandwidth parameter λ: $(f(x_1), f(x_2), \ldots, f(x_n))^T \sim N(0, \Sigma)$ where $\Sigma_{ij} = \sigma^2\rho^2 \exp\{-\lambda^2\|x_i - x_j\|^2\}$). For each unit, we predict the missing counterfactual outcome by plugging the covariate into the fitted model of the opposite group, and then estimate the individual treatment effect by the difference between observed outcome and predicted counterfactual outcome. Figure 2 shows the predicted counterfactuals and CATE, as well as the associated uncertainty band as a function of age. The true effects curve is deliberately omitted to focus on issue of uncertainty quantification. In the region of good overlap, all three models lead to similar point and interval estimates of CATE. However, marked difference emerges in the region of poor overlap. Here the linear model appears overconfident in predicting counterfactuals and thus estimating CATE. GP trades potential bias with increased uncertainty bands as overlap decreases and produces a more adaptive uncertainty quantification. BART produces shorter error bars than GP (but wider than the linear model), but width of the uncertainty band remains similar regardless of the degree of overlap, which is clearly over confident in the region of poor overlap. This pattern is not surprising given how GP and BART are constructed, on which we do not elaborate here.

Of course, it is prudent not to overly generalize the message from a single example. Nonetheless, we feel that deeper investigation is warranted on how BART-based and other Bayesian nonparametric priors perform in terms of uncertainty quantification of CATE under different degree of overlap, which is a central problem in causal inference with high dimensional data and causal estimands.

References

Antonelli, J., Papadogeorgou, G., and Dominici, F. (2018). "Causal Inference in high dimensions: A marriage between Bayesian modeling and good frequentist properties." *arXiv preprint arXiv:1805.04899*. 1008

Bang, H. and Robins, J. (2005). "Doubly robust estimation in missing data and causal inference models." *Biometrics*, 61: 962–972. MR2216189. doi: https://doi.org/10.1111/j.1541-0420.2005.00377.x. 1008

Belloni, A., Chernozhukov, V., and Hansen, C. (2014). "Inference on treatment effects after selection among high-dimensional controls." *The Review of Economic Studies*, 81(2): 608–650. MR3207983. doi: https://doi.org/10.1093/restud/rdt044. 1008

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., and Robins, J. (2018). "Double/debiased machine learning for treatment and structural parameters." *The Econometrics Journal*, 21(1): C1–C68. MR3769544. doi: https://doi.org/10.1111/ectj.12097. 1008

Ding, P. and Li, F. (2018). "Causal inference: A missing data perspective." *Statistical Science*, 33(2): 214–237. MR3797711. doi: https://doi.org/10.1214/18-STS645. 1007

Funk, M. J., Westreich, D., Wiesen, C., Stürmer, T., Brookhart, M. A., and Davidian, M. (2011). "Doubly robust estimation of causal effects." *American Journal of Epidemiology*, 173(7): 761–767. 1008

Gutman, R. and Rubin, D. B. (2013). "Robust estimation of causal effects of binary treatments in unconfounded studies with dichotomous outcomes." *Statistics in Medicine*, 32(11): 1795–1814. MR3067363. doi: https://doi.org/10.1002/sim.5627. 1008

Hill, J. L. (2011). "Bayesian nonparametric modeling for causal inference." *Journal of Computational and Graphical Statistics*, 20(1): 217–240. MR2816546. doi: https://doi.org/10.1198/jcgs.2010.08162. 1011

Little, R. and An, H. (2004). "Robust likelihood-based analysis of multivariate data with missing values." *Statistica Sinica*, 949–968. MR2089342. 1008

Lunceford, J. and Davidian, M. (2004). "Stratification and weighting via the propensity score in estimation of causal treatment effects: A comparative study." *Statistics in Medicine*, 23: 2937–2960. 1008

McCandless, L. C., Gustafson, P., and Austin, P. C. (2009). "Bayesian propensity score analysis for observational data." *Statistics in Medicine*, 28(1): 94–112. MR2655553. doi: https://doi.org/10.1002/sim.3460. 1007

Rasmussen, C. E. (2003). "Gaussian processes in machine learning." In *Summer School on Machine Learning*, 63–71. Springer. 1011

Robins, J. and Wasserman, L. (2012). "Robins and Wasserman Respond to a Nobel Prize Winner." URL https://normaldeviate.wordpress.com/2012/08/28/robins-and-wasserman-respond-to-a-nobel-prize-winner/ 1007

Robins, J. M., Hernán, M. A., and Wasserman, L. (2015). "On Bayesian estimation of marginal structural models." *Biometrics*, 71(2): 296. MR3366233. doi: https://doi.org/10.1111/biom.12273. 1007, 1008

Scharfstein, D., Rotnitzky, A., and Robins, J. (1999). "Adjusting for nonignorable drop-out using semiparametric nonresponse models (with discussion)." *Journal of the American Statistical Association*, 94: 1096–1146. MR1731478. doi: https://doi.org/10.2307/2669923. 1008

Scott, J. G. and Berger, J. O. (2010). "Bayes and empirical-Bayes multiplicity adjustment in the variable-selection problem." *The Annals of Statistics*, 2587–2619. MR2722450. doi: https://doi.org/10.1214/10-AOS792. 1007

Sims, C. (2006). "On an example of Larry Wasserman." *Online manuscript, available from* http://sims.princeton.edu/yftp/WassermanExmpl/WassermanComment.pdf, 2(10). 1007

Zhou, T., Elliott, M. R., and Little, R. J. (2019). "Penalized spline of propensity methods for treatment comparison." *Journal of the American Statistical Association*, 114(525): 1–19. MR3941229. doi: https://doi.org/10.1080/01621459.2018.1518234. 1008

Zigler, C. M., Watts, K., Yeh, R. W., Wang, Y., Coull, B. A., and Dominici, F. (2013). "Model feedback in Bayesian propensity score estimation." *Biometrics*, 69(1): 263–273. MR3058073. doi: https://doi.org/10.1111/j.1541-0420.2012.01830.x. 1007

Invited Discussion*

Stefan Wager†

The literature on heterogeneous treatment effect estimation has been extremely active over the past few years, and the paper by Hahn, Murray, and Carvalho (2020) is a major addition to it. Hahn et al. (2020) show how to design priors for heterogeneous treatment effects that are robust to what the authors call "regularization-induced confounding" and "targeted selection" and, as such, open the door to considerably more robust and reliable Bayesian inference of treatment heterogeneity under unconfoundedness. The authors convincingly show that their innovations add considerable value over a simpler Bayesian forest approach following, e.g., early work from Hill (2011).

The publication of this paper provides a nice opportunity to reflect on just how fast this area has developed over the past years. When Hahn et al. (2020) released the first draft of their manuscript on arXiv in 2017, there were still major questions about how best to approach the problem of heterogeneous treatment effect estimation as evidenced by, e.g., discussions at that year's Atlantic Causal Inference Conference. By now, in contrast, there appears to be fairly widespread consensus on conceptual ideas that underpin good estimators of treatment heterogeneity. This comment offers one take on 3 ideas which, I believe, have played a major role in pushing the field forward. Of course, these ideas manifest themselves differently depending on methodological context (e.g., frequentist vs. Bayesian methods, trees vs. lasso), but overall they seem to have broad applicability.

Dedicated regularization for treatment effects Following notation from Hahn et al. (2020), we want to estimate the effect of a binary treatment $Z_i \in \{0, 1\}$ on an outcome $Y_i \in \mathbb{R}$ as a function of covariates $X_i \in \mathcal{X}$. Following the Neyman–Rubin causal model (Imbens and Rubin, 2015), we posit potential outcomes $\{Y_i(0), Y_i(1)\}$ such that $Y_i = Y_i(Z_i)$, and seek to estimate the conditional average treatment effect function $\tau(x) = \mathbb{E}\left[Y_i(1) - Y_i(0) \,\middle|\, X_i = x\right]$. For purposes of identification, we assume unconfoundedness (Rosenbaum and Rubin, 1983), i.e., that treatment is as good as random conditionally on X_i: $\{Y_i(0), Y_i(1)\} \perp\!\!\!\perp Z_i \,\middle|\, X_i$.

One can readily check that, under unconfoundedness, we have $\tau(x) = \mu(x; 1) - \mu(x; 0)$ with $\mu(x; z) = \mathbb{E}\left[Y_i \,\middle|\, X_i = x, Z_i = z\right]$. This suggests a simple strategy for nonparametric estimation of the conditional average treatment effect function $\tau(x)$: First learn $\hat{\mu}(x; 0)$ and $\hat{\mu}(x; 1)$ by fitting separate predictive models to the control and treated samples respectively, and then set $\hat{\tau}(x) = \hat{\mu}(x; 1) - \hat{\mu}(x; 0)$. This may, however, lead to problems. In general, modern machine learning methods operate via some type of representation learning; and, if we use different representations to express $\hat{\mu}(x; 0)$ and $\hat{\mu}(x; 1)$, then $\hat{\tau}(x)$ may be excessively noisy or biased. As a simple example, consider the case where $\hat{\mu}(x; 0)$ and $\hat{\mu}(x; 1)$ are both decision trees—but with different splits. In this

*This work was supported by National Science Foundation grant DMS-1916163.
†Stanford University, swager@stanford.edu

case, $\hat{\tau}(x) = \hat{\mu}(x; 1) - \hat{\mu}(x; 0)$ would be quite unstable, and in particular would have a more complicated shape than either $\hat{\mu}(x; 0)$ or $\hat{\mu}(x; 1)$ on its own. Künzel, Sekhon, Bickel, and Yu (2019) provide further examples of this issue.

A first major advance in the literature on treatment heterogeneity was the realization that, in a high-dimensional or non-parametric setting, it's important to use dedicated regularizers that directly push $\hat{\tau}(x)$ to have a simple form. A simple application of this idea arises in the case of the lasso (Hastie, Tibshirani, and Wainwright, 2015). Assume that $\mu(x; z) = x \cdot \beta_{(z)}$ for some high-dimensional vector z, so that $\tau(x) = x \cdot (\beta_{(1)} - \beta_{(0)})$. A naïve analysis might fit separate lasso regressions on the treated and control units, but such an approach may learn different sparsity patterns for $\beta_{(0)}$ and $\beta_{(1)}$, thus resulting in an unstable $\tau(x)$ estimate. A better approach is to reparametrize $\mu(x; z) = x \cdot b + (2z - 1)\, x \cdot \delta$, where $b = (\beta_{(0)} + \beta_{(1)})/2$ and $\delta = \beta_{(1)} - \beta_{(0)}$; then, we can apply separate sparsity penalties on b and δ. This is a simple idea but, by directly pushing the treatment effect parameter δ towards sparsity, it often improves performance considerably.

Hahn et al. (2020) show how to adapt it to Bayesian prior design, while Athey and Imbens (2016) discuss a modification of regression trees that directly target $\tau(x)$. More subtle ideas for algorithmically regularizing the treatment effect function include the X-learner of Künzel, Sekhon, Bickel, and Yu (2019), and refitting predictions from a first step analysis as in the Virtual Twins method of Foster, Taylor, and Ruberg (2011).

The propensity score as a covariate Once we've dealt with egregious instability of $\hat{\tau}(x)$ by using appropriate regularizers, a next concern is whether confounding effects may bleed into treatment effect estimates due to finite sample effects. As is well explained in the section on "targeted selection" in Hahn et al. (2020), this concern arises whenever the propensity score $\pi(x) = \mathbb{P}\left[Z_i = 1 \,\middle|\, X_i = x\right]$ is associated with the baseline effect $\mu(x; 0)$. If $\mu(x; 0)$ takes on a complicated non-parametric specification that cannot be perfectly captured in finite samples and we use a method that underfits the baseline function such that $\mu(x; 0) - \hat{\mu}(x; 0)$ is positively (or negatively) correlated with $\pi(x)$, then we may easily have this baseline error push us to over- (or under-) estimate $\tau(x)$.

Considerations of this type have attracted considerable attention in the causal inference community for several decades (Robins and Ritov, 1997), and play a key role in discussions of how best to do variable selection when estimating global causal parameters (e.g., Belloni, Chernozhukov, and Hansen, 2014). The general message is that, in order to be robust to associations between $\pi(x)$ and $\mu(x; 0)$, one needs to fit the propensity score $\hat{\pi}(x)$ and adjust for it in the final modeling step. Hahn et al. (2020) propose the simple idea of using a propensity estimate as a feature when estimating the baseline effect, i.e., they fit the baseline as $\hat{\mu}(x, \hat{\pi}(x); 0)$, and find this to work well empirically.

Treatment-focused loss functions A final idea that unlocks a general suite of tools for heterogeneous treatment effect estimation is the use of loss functions that directly isolate the treatment effect function $\tau(x)$. The idea of using target-specific loss functions has a

long history in causal inference, going back at least to van der Laan and Dudoit (2003). In the context heterogeneous treatment effect estimation, a simple instance of this idea is the "transformed outcome" method, which starts from the following observation. Under unconfoundedness, we can check that $\mathbb{E}\left[\Delta_i \,\middle|\, X_i = x\right] = \tau(x)$, where $\Delta_i = Z_i Y_i / \pi(X_i) - (1 - Z_i) Y_i / (1 - \pi(X_i))$. Thus, in a randomized trial where propensity scores $\pi(x)$ are known a-priori, we can estimate $\tau(x)$ by first forming the modified outcomes Δ_i and then running a non-parametric regression of Δ_i on X_i (Tian, Alizadeh, Gentles, and Tibshirani, 2014). Baseline effects $\mu(x; 0)$ never even appear in this specification, which largely obviates concerns related to regularizing or under-fitting of this term.

One limitation of the transformed outcome method is that it is not robust to errors in estimating the propensity score when $\pi(x)$ is not known a-priori, and several "robust" loss functions for treatment effect estimation that remedy this issue have recently been proposed. The R-learner (Nie and Wager, 2020) builds on the partially linear model estimator of Robinson (1988) to develop a loss function that is first-order robust to errors in estimating $\pi(x)$ and baseline effects; see also Athey, Tibshirani, and Wager (2019) and Zhao, Small, and Ertefaie (2017) for variants of this idea applied to random forests and the lasso specifically. Meanwhile, the DR-learner (Fan, Hsu, Lieli, and Zhang, 2019; Kennedy, 2020; Zimmert and Lechner, 2019) estimates $\tau(x)$ by regressing the augmented inverse-propensity weighted scores of Robins, Rotnitzky, and Zhao (1994) against X_i. Overall, the promise of such robust loss functions is that they enable accurate estimation of $\tau(x)$ even when $\pi(x)$ and $\mu(x; 0)$ may be difficult to estimate, thus generalizing well known results on semiparametric inference for "global" targets like the average treatment effect; see Chernozhukov, Chetverikov, Demirer, Duflo, Hansen, Newey, and Robins (2018), and references therein.

Closing thoughts Recent advances in methods for treatment heterogeneity have enabled a large toolkit of practical methods available to applied researchers. As a community, we now appear to be at a point where we can quickly remix these ideas to develop new methods for treatment heterogeneity that can address different application-specific challenges, and have a formal understanding of how dedicated methods are able to accurately target $\tau(x)$. Several open questions remain, however. In particular, while reading the paper of Hahn, Murray, and Carvalho (2020), I was left wondering about the following two:

- When using treatment-focused loss functions, it's possible to show that (under appropriate conditions), the accuracy with which we can estimate $\tau(x)$ is insensitive to our rate of convergence on the nuisance components $\pi(x)$ and $\mu(x; z)$, even when $\hat{\pi}(x)$ and $\hat{\mu}(x; z)$ may converge an order of magnitude slower than $\hat{\tau}(x)$ (Kennedy, 2020; Nie and Wager, 2020). Do analogous results hold for the approach of Hahn et al. (2020), where $\hat{\pi}(x)$ is used as a covariate when fitting $\hat{\mu}(\cdot)$? In the context of estimating an average treatment effect, Hirano, Imbens, and Ridder (2003) showed that using an estimated propensity score for weighting could sometimes be enough to achieve efficiency. Does any intuition of this type carry over to the problem heterogeneous treatment effect estimation?

- In Hahn et al. (2020), the propensity score is only used as a covariate to make us robust to potential confounding effects. In some applications, however, there may also be interest in the propensity score as an effect modifier. For example, when studying the returns to college education, Brand and Xie (2010) argue that students who are least likely to attend college a-priori may be the ones who benefit the most from attendance; and, in applications like these, it may be of interest to allow $\tau(x)$ explicitly depend on $\pi(x)$. Of particular interest here would be to understand how $\tau(x)$ varies with the true propensity score $\pi(x)$, and not just the estimate $\hat{\pi}(x)$.

Finally, I want to thank Hahn, Murray, and Carvalho (2020) for preparing this very nice paper, and look forward to seeing how the community builds on their results in the future.

References

Susan Athey and Guido W. Imbens. Recursive partitioning for heterogeneous causal effects. *Proceedings of the National Academy of Sciences*, 113(27):7353–7360, 2016. MR3531135. doi: https://doi.org/10.1073/pnas.1510489113. 1015

Susan Athey, Julie Tibshirani, and Stefan Wager. Generalized random forests. *The Annals of Statistics*, 47(2):1148–1178, 2019. MR3909963. doi: https://doi.org/10.1214/18-AOS1709. 1016

Alexandre Belloni, Victor Chernozhukov, and Christian Hansen. Inference on treatment effects after selection among high-dimensional controls. *The Review of Economic Studies*, 81(2):608–650, 2014. MR3207983. doi: https://doi.org/10.1093/restud/rdt044. 1015

Jennie E. Brand and Yu Xie. Who benefits most from college? Evidence for negative selection in heterogeneous economic returns to higher education. *American Sociological Review*, 75(2):273–302, 2010. doi: https://doi.org/10.1177/0003122410363567. 1017

Victor Chernozhukov, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins. Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, 21(1):1–68, 2018. MR3769544. doi: https://doi.org/10.1111/ectj.12097. 1016

Qingliang Fan, Yu-Chin Hsu, Robert P. Lieli, and Yichong Zhang. Estimation of conditional average treatment effects with high-dimensional data. *arXiv preprint arXiv:1908.02399*, 2019. 1016

Jared C. Foster, Jeremy M. G. Taylor, and Stephen J. Ruberg. Subgroup identification from randomized clinical trial data. *Statistics in Medicine*, 30(24):2867–2880, 2011. MR2844689. doi: https://doi.org/10.1002/sim.4322. 1015

P. Richard Hahn, Jared S. Murray, and Carlos M. Carvalho. Bayesian regression tree models for causal inference: regularization, confounding, and heterogeneous effects. *Bayesian Analysis*, forthcoming, 2020. 1014, 1015, 1016, 1017

Trevor Hastie, Robert Tibshirani, and Martin Wainwright. *Statistical Learning with Sparsity: The Lasso and Generalizations*. CRC Press, 2015. MR3616141. doi: https://doi.org/10.1201/b18401. 1015

Jennifer L. Hill. Bayesian nonparametric modeling for causal inference. *Journal of Computational and Graphical Statistics*, 20(1):217–240, 2011. MR2816546. doi: https://doi.org/10.1198/jcgs.2010.08162. 1014

Keisuke Hirano, Guido W. Imbens, and Geert Ridder. Efficient estimation of average treatment effects using the estimated propensity score. *Econometrica*, 71(4):1161–1189, 2003. MR1995826. doi: https://doi.org/10.1111/1468-0262.00442. 1016

Guido W. Imbens and Donald B. Rubin. *Causal Inference in Statistics, Social, and Biomedical Sciences*. Cambridge University Press, 2015. MR3309951. doi: https://doi.org/10.1017/CBO9781139025751. 1014

Edward H. Kennedy. Optimal doubly robust estimation of heterogeneous causal effects. *arXiv preprint arXiv:2004.14497*, 2020. 1016

Sören R. Künzel, Jasjeet S. Sekhon, Peter J. Bickel, and Bin Yu. Metalearners for estimating heterogeneous treatment effects using machine learning. *Proceedings of the National Academy of Sciences*, 116(10):4156–4165, 2019. doi: https://doi.org/10.1073/pnas.1804597116. 1015

Xinkun Nie and Stefan Wager. Quasi-oracle estimation of heterogeneous treatment effects. *Biometrika*, forthcoming, 2020. 1016

James M. Robins and Ya'acov Ritov. Toward a curse of dimensionality appropriate (CODA) asymptotic theory for semi-parametric models. *Statistics in Medicine*, 16(3):285–319, 1997. 1015

James M. Robins, Andrea Rotnitzky, and Lue Ping Zhao. Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association*, 89(427):846–866, 1994. MR1294730. doi: https://doi.org/10.1080/01621459.1994.10476818. 1016

Peter M. Robinson. Root-n-consistent semiparametric regression. *Econometrica*, 56(4):931–954, 1988. MR0951762. doi: https://doi.org/10.2307/1912705. 1016

Paul R. Rosenbaum and Donald B. Rubin. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55, 1983. MR0742974. doi: https://doi.org/10.1093/biomet/70.1.41. 1014

Lu Tian, Ash A. Alizadeh, Andrew J. Gentles, and Robert Tibshirani. A simple method for estimating interactions between a treatment and a large number of covariates. *Journal of the American Statistical Association*, 109(508):1517–1532, 2014. MR3293607. doi: https://doi.org/10.1080/01621459.2014.951443. 1016

Mark J. van der Laan and Sandrine Dudoit. Unified cross-validation methodology for selection among estimators and a general cross-validated adaptive epsilon-net estimator: Finite sample oracle inequalities and examples. Technical report, UC Berkeley Division of Biostatistics, Berkeley CA, 2003. 1016

Qingyuan Zhao, Dylan S. Small, and Ashkan Ertefaie. Selective inference for effect modification via the lasso. *arXiv preprint arXiv:1705.08020*, 2017. 1016

Michael Zimmert and Michael Lechner. Nonparametric estimation of causal heterogeneity under high-dimensional confounding. *arXiv preprint arXiv:1908.08779*, 2019. 1016

Contributed Discussion*

Liangyuan Hu†

Congratulations to Hahn, Murray and Carvalho on a nice contribution. The authors propose the Bayesian causal forest (BCF) model. It reduces the bias and improves frequentist coverage of Bayesian credible intervals of treatment effect estimates in the presence of strong confounding and treatment effect heterogeneity. The BCF model builds upon the popular and empirically proven prediction method, Bayesian Additive Regression Trees (BART) (Chipman et al., 2010). BCF reformulates the response surface model as the sum of two functions, one modeling the prognostic impact of the control variables and one representing the treatment effect. This formulation directly incorporates estimates of the propensity score (PS) which induces a covariate-dependent prior on the regression function and regularizes treatment effect heterogeneity separately from the prognostic effect of control variables. It is impressive that the proposed method, in many settings, has better bias reduction, more consistent 95% coverage probability and shorter uncertainty intervals compared to the vanilla BART, which boasts better performance in a host of modern causal inference studies, including Hill (2011), Wendling et al. (2018), Dorie et al. (2019) and Hu et al. (2020), to name a few.

This paper offers an extensive study to explicate and evaluate the performance of the BCF model in different settings and provides a detailed discussion about its utility in causal inference. It is a welcomed addition to the causal machine learning literature.

I will emphasize the contribution of the BCF model to the field of causal inference through discussions on two topics: 1) the difference between the PS in the BCF model and the Bayesian PS in a Bayesian updating approach, 2) an alternative exposition of the role of the PS in outcome modeling based methods for the estimation of causal effects. I will conclude with comments on avenues for future research involving BCF that will be important and much needed in the era of Big data.

1 Distinction from the Bayesian propensity scores

It is necessary to make a distinction between incorporating the estimated PSs as a covariate in the BCF model and combining so called "Bayesian propensity scores" and Bayesian inference for the estimation of causal effects in a single Bayesian updating approach (Zigler et al., 2013; Zigler and Dominici, 2014). The Bayesian PS has received recent attention in the literature (Kaplan and Chen, 2012; Zigler, 2016; Liao and Zigler, 2020). In essence, a series of work have demonstrated that the model feedback, i.e., the propagation of information from the outcome model to the PS model, would distort inferences about the causal effect. In the BCF model, the independent BART prior is placed over f, $f \sim \text{BART}(X, Z, \hat{\pi})$, where $\hat{\pi}$ is the estimated PS and included as one of

*Main article DOI: 10.1214/19-BA1195.

†Department of Population Health Science and Policy, Icahn School of Medicine at Mount Sinai, liangyuan.hu@mssm.edu

the splitting dimensions. The $\hat{\pi}$ is included in the BCF model as an additional covariate for splitting and is not updated or contaminated by the outcome information, thereby setting BCF free from the inference issue caused by the Bayesian model feedback.

2 The role of the propensity score: connections to the confounding function

I now turn to exposing the role of the PS in reducing the bias in the estimates of causal effects under targeted selection. The inclusion of the PS as a covariate in the response model is closely connected to the confounding function approach (Robins, 1999), designed for removing the bias due to unmeasured confounding from the treatment effect estimates. The targeted selection described in Hahn et al. (2020) suggests that the treatment assignment probability $\pi(x)$ depends on $\mu(x) = E(Y|Z = 0, x)$. Using Figure 4 as an example to illustrate, individuals who would have higher outcomes without treatment are more likely to be treated. In another word, treated individuals would have higher (potential) outcomes than untreated individuals to no treatment. This is a violation of the ignorability assumption. To see why, define a confounding function as $c(z, x) = E(Y(z)|Z = 1, X = x) - E(Y(z)|Z = 0, X = x)$, $z \in \{0, 1\}$. When ignorability holds, $c(z, x) = 0$ for $z = 0$ and $z = 1$. When ignorability is violated, such as in the presence of targeted selection, $c(0, x) > 0$. The violation of ignorability gives rise to biased estimates of the causal effects. An unbiased effect estimate can be obtained by "correcting" the *observed* outcome for unmeasured confounding, namely, $Y - [E(Y(z)|Z = z, x) - E(Y(z)|x)]$ (Robins, 1999; Brumback et al., 2004). Applying the law of total expectation to $E(Y(z)|x)$ yields a "corrected" outcome (see a short proof in Hu, 2020),

$$Y^C = Y - \{(1 - \pi(x))z - \pi(x)(1 - z)\}c(z, x),\ z \in \{0, 1\}. \tag{2.1}$$

Differencing the conditional expectations of Y^C between the treated and untreated individuals gives an unbiased effect estimate, $E(Y^C|X = x, Z = 1) - E(Y^C|X = x, Z = 0)$. In equation (2.1), $c(z, x)$ is a user-supplied prior distribution (a range of values in frequentist approaches) representing our beliefs about the degree of ignorability violation, or, targeted selection (Hogan et al., 2014). We see that the PS, $\pi(x)$, is an integral part in the outcome for causal modeling when there is a need to remove the bias attributable to targeted selection. Relating to the inclusion of $\hat{\pi}$ in the BCF model, the $c(z, x)$ can be deemed as characterizing how big of a role the PS plays in the estimation of causal effects. With strong targeted selection, $c(z, x)$ would deviate from zero substantially, then $\hat{\pi}$ is important for bias reduction. In the absence of targeted selection, $c(z, x)$ would be close to zero, and the role of $\hat{\pi}$ is diminished.

3 Final thought on possible extensions

My final thought is about extending the BCF model to meet the emerging methodological needs, particularly in the Biostatistical research. First, a common causal estimand

of interest is the average treatment effect on the treated (ATT). The BCF model does not seem to be readily implementable for the ATT estimation. Although, the idea of ps-BART can be easily applied to estimate the ATT effect. Second, in the era of Big data, given the wealth of information captured in large-scale data, it is rare that treatment regimens are defined in terms of two treatments only. Refined causal inference approaches are in great demand for the multiple treatment settings. Hu et al. (2020) investigated the operating characteristics of several machine learning based causal inference techniques in the multiple treatment setting, and found that BART-based methods generally had the best performance. It would be a useful addition to the causal machine learning literature if the BCF model can be extended to simultaneously compare more than two active treatments.

Supplementary Material

Appendix for Comment on Article by Hahn, Murray and Carvalho
(DOI: 10.1214/20-BA1195CT1SUPP; .pdf).

References

Brumback, B. A., Hernán, M. A., Haneuse, S. J., and Robins, J. M. (2004). "Sensitivity analyses for unmeasured confounding assuming a marginal structural model for repeated measures." *Statistics in Medicine*, 23(5): 749–767. 1021

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. MR2758172. doi: https://doi.org/10.1214/09-AOAS285. 1020

Dorie, V., Hill, J., Shalit, U., Scott, M., Cervone, D., et al. (2019). "Automated versus do-it-yourself methods for causal inference: Lessons learned from a data analysis competition." *Statistical Science*, 34(1): 43–68. MR3938963. doi: https://doi.org/10.1214/18-STS667. 1020

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*. 1021

Hill, J. L. (2011). "Bayesian nonparametric modeling for causal inference." *Journal of Computational and Graphical Statistics*, 20(1): 217–240. MR2816546. doi: https://doi.org/10.1198/jcgs.2010.08162. 1020

Hogan, J., Daniels, M., and Hu, L. (2014). "A Bayesian perspective on assessing sensitivity to assumptions about unobserved data." In Molenberghs, G., Fitzmaurice, G., Kenward, M., Tsiatis, A., and Verbeke, G. (eds.), *Handbook of Missing Data Methodology*, 405–434. New York: Boca Raton, FL, USA: Chapman & Hall/CRC Press. MR3380562. 1021

Hu, L. (2020). "Supplementary Material of "Contributed Discussion"." *Bayesian Analysis*. doi: https://doi.org/10.1214/20-BA1195CT1SUPP. 1021

Hu, L., Gu, C., Lopez, M., Ji, J., and Wisnivesky, J. (2020). "Estimation of Causal Effects of Multiple Treatments in Observational Studies with a Binary Outcome." *Statistical Methods in Medical Research*, In Press: 0962280220921909. 1020, 1022

Kaplan, D. and Chen, J. (2012). "A two-step Bayesian approach for propensity score analysis: Simulations and case study." *Psychometrika*, 77(3): 581–609. MR2943114. doi: https://doi.org/10.1007/s11336-012-9262-8. 1020

Liao, S. X. and Zigler, C. M. (2020). "Uncertainty in the design stage of two-stage Bayesian propensity score analysis." *Statistics in Medicine*. 1020

Robins, J. M. (1999). "Association, causation, and marginal structural models." *Synthese*, 151–179. MR1766776. doi: https://doi.org/10.1023/A:1005285815569. 1021

Wendling, T., Jung, K., Callahan, A., Schuler, A., Shah, N., and Gallego, B. (2018). "Comparing methods for estimation of heterogeneous treatment effects using observational data from health care databases." *Statistics in Medicine*, 37(23): 3309–3324. MR3856345. doi: https://doi.org/10.1002/sim.7820. 1020

Zigler, C. M. (2016). "The central role of Bayes' theorem for joint estimation of causal effects and propensity scores." *The American Statistician*, 70(1): 47–54. MR3480670. doi: https://doi.org/10.1080/00031305.2015.1111260. 1020

Zigler, C. M. and Dominici, F. (2014). "Uncertainty in propensity score estimation: Bayesian methods for variable selection and model-averaged causal effects." *Journal of the American Statistical Association*, 109(505): 95–107. MR3180549. doi: https://doi.org/10.1080/01621459.2013.869498. 1020

Zigler, C. M., Watts, K., Yeh, R. W., Wang, Y., Coull, B. A., and Dominici, F. (2013). "Model feedback in Bayesian propensity score estimation." *Biometrics*, 69(1): 263–273. MR3058073. doi: https://doi.org/10.1111/j.1541-0420.2012.01830.x. 1020

Contributed Discussion

Jingyu He* and Nikolay Krantsevich†

The authors show that by expressing both the prognostic and treatment terms as BART models, the Bayesian causal forest model (BCF) inherits the accuracy and hyperparameter robustness of BART. However, BCF similarly inherits some of BART limitations, specifically slow run-times and/or slow mixing, leading to artificially short posterior credible intervals. Recent work (He et al., 2019; He and Hahn, 2020) has investigated stochastic heuristics for fitting BART-like models (XBART) and proposed to use those model fits as seed values for independent Markov chain Monte Carlo runs of the original BART random walk algorithm. Here, we report some preliminary work which applies this technique to the Bayesian causal forests model, resulting in faster point estimates and improved coverage of posterior credible intervals. These findings should be of interest to readers concerned with these important practical aspects of the BCF model. The procedure is:

1. Fit an "accelerated BCF" (XBCF) model on the data, performing s sweeps through the data with the first m burn-in. Store the last $s - m$ sweeps. Usually s is orders of magnitude smaller than a typical MCMC run; we use $s = 40$ in the simulation results reported below.
2. Run $s-m$ independent BCF Markov chains, initialized at each of the $s-m$ forests obtained from XBCF. Because the BCF models are seeded with good starting values, each chain is run for significantly fewer sweeps.
3. Lastly, collate the $s - m$ BCF sample and construct posterior estimates and summaries.

To compare this modified approach to the one described in the paper, we repeat the simulation studies from section 6.1. We run BCF for 4000 iterations, with 2000 discarded as burn-in. For the XBCF approach we perform 40 sweeps total, with 15 burn-in iterations. For the "warm-start" BCF approach, each of the 25 XBCF draws are used to initialize a BCF MCMC run, from which 110 total draws are obtained, with just 10 burn-in iterations being discarded. The results are presented in the table below. The main findings are:

- The root mean square error (RMSE) of the average treatment effect (ATE) is similar for all three methods, with XBCF runtime being 4 times faster on average.
- XBCF provides the best conditional average treatment effect (CATE) estimation for the homogeneous treatment effect case.
- Warm-start BCF always performs better than regular BCF in CATE estimation, in terms of both RMSE and coverage. In general for both ATE and CATE, warm-start BCF provides the best coverage among all three methods.

*College of Business, City University of Hong Kong, Hong Kong SAR
†School of Mathematical and Statistical Sciences, Arizona State University, Tempe, Arizona, USA, krantsevich@asu.edu

The superiority of XBCF in the case of estimating CATE with homogeneous treatment is likely a fortuitous benefit of under-exploring the posterior, but more investigation is required. The improved coverage is substantial, which is unsurprising in some regards; it is well-known among BART users that collating multiple chains can improve posterior exploration and hence coverage. Importantly, XBCF is able to find good seeding points fast, which then requires essentially no burn-in. Note that the runtime for warm-start BCF can be reduced by running individual BCF models in parallel. The reported time is without any parallelization, and a more practical number would be at least four times less than the reported duration (plus the initial time to fit XBCF). A detailed description of both XBCF and warm-start BCF will be presented as a separate paper, and we hope both of them will find their use along with the originally proposed BCF model.

Data Generating Process			Results							
Prognostic Term	Treatment	Sample Size	Method	RMSE		Coverage		I.L.		Time
				ATE	CATE	ATE	CATE	ATE	CATE	
Linear	Homogeneous	500	XBCF	0.17	0.21	0.85	0.91	0.67	0.94	3.51
			warm	0.17	0.36	0.93	0.99	0.80	2.19	17.51
			BCF	0.17	0.55	0.92	0.84	0.77	1.58	14.47
Linear	Homogeneous	250	XBCF	0.23	0.28	0.87	0.92	0.93	1.28	2.15
			warm	0.23	0.46	0.95	1.00	1.13	2.89	9.80
			BCF	0.23	0.72	0.92	0.83	1.08	2.11	7.69
Linear	Heterogeneous	500	XBCF	0.15	0.98	0.84	0.77	0.61	2.23	3.69
			warm	0.16	0.74	0.94	0.97	0.75	3.10	17.94
			BCF	0.17	0.85	0.92	0.87	0.73	2.48	15.30
Linear	Heterogeneous	250	XBCF	0.25	1.27	0.82	0.73	0.91	2.66	2.11
			warm	0.24	0.89	0.92	0.98	1.06	4.07	9.19
			BCF	0.25	1.05	0.88	0.91	1.03	3.44	7.91
Nonlinear	Homogeneous	500	XBCF	0.12	0.16	0.79	0.89	0.40	0.58	3.66
			warm	0.11	0.28	0.90	0.99	0.48	1.50	18.13
			BCF	0.11	0.36	0.90	0.90	0.46	1.18	14.85
Nonlinear	Homogeneous	250	XBCF	0.18	0.22	0.78	0.86	0.57	0.81	2.26
			warm	0.16	0.35	0.91	0.99	0.68	1.96	9.72
			BCF	0.15	0.46	0.90	0.91	0.66	1.60	7.78
Nonlinear	Heterogeneous	500	XBCF	0.12	0.69	0.75	0.77	0.38	1.59	4.23
			warm	0.11	0.56	0.88	0.96	0.47	2.27	19.98
			BCF	0.12	0.65	0.87	0.82	0.46	1.70	16.52
Nonlinear	Heterogeneous	250	XBCF	0.20	0.89	0.72	0.76	0.56	2.02	2.38
			warm	0.17	0.68	0.90	0.97	0.69	2.92	9.68
			BCF	0.18	0.78	0.88	0.88	0.67	2.39	8.15

Table 1: Results of root mean squared error (RMSE), interval length (I.L.) and interval coverage (Coverage) for ATE and CATE estimator under different data generating processes. The column Time is running time in seconds.

References

He, J. and Hahn, P. R. (2020). "Stochastic tree ensembles for regularized nonlinear regression." *Technical report*. 1024

He, J., Yalov, S., and Hahn, P. R. (2019). "XBART: Accelerated Bayesian Additive Regression Trees." In *The 22nd International Conference on Artificial Intelligence and Statistics*, 1130–1138. 1024

Contributed Discussion

Kolyan Ray*, Botond Szabó†, and Aad van der Vaart‡

Estimating a causal effect has a long history and it is delightful to see increasing attention in the Bayesian literature. We thank the authors for their insightful contribution. In our recent work, we have also found it profitable to include an estimate of the propensity score in the prior modelling, and found a particular way that helps to reduce bias when estimating an average treatment effect.

Under the potential outcomes model given by formula (2.1) in the paper, we have $Y^z|\,X \sim Y|\,X, Z = z$, whence the density of the observed data (Y, Z, X) on a single individual factorizes as

$$(y, z, x) \mapsto p_{Y^z|X}(y|\,x)\,\pi(x)^z\big(1-\pi(x)\big)^{1-z}\,p_X(x), \qquad (z \in \{0,1\}).$$

The first term is the conditional density of the counterfactual Y^z given X. In the paper under discussion, these are fixed as $p_{Y^z|X}(y|\,x) = \phi_\sigma\big(y - f(x,z)\big)$ by equation (2.4), which we interpret as $Y^z = f(X,z) + \varepsilon$ with $X \perp\!\!\!\perp \varepsilon \sim N(0,\sigma^2)$. The function $\pi(x) = \mathrm{P}(Z = 1|\,X = x)$ is the propensity score.

The causal effect at level x is $\tau(x) = \int y\big(p_{Y^1|X}(y|\,x) - p_{Y^0|X}(y|\,x)\big)\,dy$. The most common approaches to estimate an average of this (such as (C)ATE) is to write $\mathrm{E}Y^z$ either by "standardization" as $\mathrm{E}\mathrm{E}(Y|\,Z=z, X)$ or by "inverse weighting" as $\mathrm{E}\big(YZ/\pi(X)\big)$. The first leads to regression type estimators, the second to the Horvitz-Thompson estimator. It is notable that the second, unlike the first, will use an estimate of the propensity score. It is well known that interpolations between the two methods are useful. See for instance Hernan and Robins (2019).

A Bayesian would put priors on the unknown parameters $p_{Y^z|X}$, π and p_X, and derive the induced posterior distribution of (C)ATE. The question is which priors to use. It seems to be an old reflex to mirror the factorization in the likelihood by prior independence of the factors. However, this will make the posterior of $p_{Y^z|X}$ functionally free of the prior on π, and this will be transferred to the posterior of the (C)ATE. Both the paper under discussion and the extensive literature on frequentist methods suggest that this is not good.

With this in mind, in the papers Ray and van der Vaart (2018) and Ray and Szabó (2019), we proposed a class of dependent priors. The one that is easiest to implement uses an estimator $\hat\pi$, as in the paper under discussion. However, information theory for semiparametric models suggests that $\hat\pi$ should be inserted in a different way. Our simulation studies appear to confirm our theoretical results. For W a Gaussian process independent of $\lambda \sim N(0,\nu^2)$, we modelled the regression function as

$$f(x,z) = W(x,z) + \lambda\left(\frac{z}{\hat\pi(x)} - \frac{1-z}{1-\hat\pi(x)}\right).$$

*Imperial College London, kolyan.ray@ic.ac.uk
†Leiden University, b.t.szabo@math.leidenuniv.nl
‡Leiden University, avdvaart@math.leidenuniv.nl

Method	$n = 250$			$n = 500$		
	RMSE	width CI ± sd	Coverage	RMSE	width CI ± sd	Coverage
BART	0.257	0.992 ± 0.088	0.94	0.155	0.617 ± 0.055	0.96
BCF	0.419	1.229 ± 0.200	0.88	0.239	0.802 ± 0.093	0.94
BART-ps	0.295	1.196 ± 0.130	0.95	0.182	0.783 ± 0.078	0.98
GP PS	0.251	1.178 ± 0.217	0.97	0.170	0.755 ± 0.113	0.98

Table 1: Estimating the ATE over 200 simulation runs.

We used the square-exponential process with hyper parameters fixed by empirical Bayes. The variance λ was empirically selected by a Lepski type procedure and $\hat{\pi}$ fitted using truncated logistic regression. For estimation of the ATE, a prior on P_X is needed; we used the Dirichlet process.

The improvement can be explained as a correction for bias. In the language of the paper under discussion, one could call this correction for "regularization induced confounding". However, it seems that different from what this name suggests, the general principle concerns bias that may result from estimating a high-dimensional parameter and is not special to causality or confounding. In the same way, it does not seem that "targeted selection" is a precondition for the usefulness of bias selection. Rather, not using independent prior modeling seems one of these novel properties of high-dimensional modeling that remain to be further explored.

To examine the performance of these methods, we provide some simulations, see Ray and Szabó (2019) for further simulations. Consider covariates $x_i \sim^{iid} N_p(0, I_p)$, $i = 1, \ldots, n$, with $p = 10$ and $n = 250$ or $n = 500$. We take the propensity score and response surface to be

$$\pi(x_i) = \Phi\big(0.15x_{i,1} - 0.4x_{i,2} - 0.5x_{i,5} - e^{2x_{i,6}} + e^{2x_{i,4}}\big)$$
$$f(x_i, z_i) = 0.8x_{i,1} + 0.4x_{i,2}^3 + 0.25e^{|x_{i,2}|} + 0.8x_{i,5}^2 - 1.5\sin(x_{i,5}) + z_i(1 + 2x_{i,2}x_{i,5}),$$

which has heterogeneous treatment effect $\tau(x_i) = 1 + 2x_{i,2}x_{i,5}$ and induces observations $Y_i = f(x_i, z_i) + \varepsilon_i$, $\varepsilon_i \sim^{iid} N(0, 1)$. We compare the performance of our method (GP PS) with BART, BART-ps (in both cases using the "bartCause" package with default settings), and BCF ("bcf" package with default settings, $\hat{\pi}$ fitted using logistic regression). The results are in Table 1.

We have considered multiple experiments, and find that BART-ps generally provides the most robust method with typically (close to) the best performance. Our method performs the best for more difficult problems, such as high-dimensional covariates and/or complex propensity scores or regression functions, where correcting for bias becomes more important. Picking a prior for W that allows for sparsity or variable selection, such as BART, may further improve the performance of our method in such settings. We note that our method is specifically targeted at estimating the (C)ATE rather than the entire response surface f, as for BART, BART-ps and BCF.

References

Hernan, M.A. and Robins, J.M. (2019). "Causal Inference." Chapman & Hall/CRC Monographs on Statistics & Applied Probability. Taylor & Francis, https://books.google.nl/books?id=_KnHIAAACAAJ, 9781420076165. 1026

Ray, K. and van der Vaart, A. (2018). "Semiparametric Bayesian causal inference." To appear in *Annals of Statistics*, arXiv:1808.04246 [math.ST]. 1026

Ray, K. and Szabó, B. (2019). "Debiased Bayesian inference for average treatment effects." *Advances in Neural Information Processing Systems*, 11952–11962. 1026, 1027

Contributed Discussion

Estevão B. Prado[†,‡,§,*], Eoghan O'Neill[¶], Belinda Hernández[‖], Andrew C. Parnell[†,‡,§], and Rafael A. Moral[†,‡]

1 Discussion

We congratulate the authors for their stimulating and excellent work on applying Bayesian trees to causal inference modelling. In this discussion, we extend the authors' work by evaluating the models on higher dimensional data sets. We remark in passing that it seems odd that the paper only contains model performance metrics on training data, but to allow for valid comparisons, we follow their approach. In particular we show that, for higher dimensions, some of the existing (non-causal) models have equivalent or superior performance to BCF for the simulations used in the paper.

We carried out a small simulation study to investigate the performance of BCF, ps-BART, Causal RF and other methods. We extend the simulations carried out in the paper by setting $n = 250$, $p = (5, 50, 100, 500)$ and consider the structure for $\tau(\mathrm{x})$, $\mu(\mathrm{x})$ and $\pi(\mathrm{x})$ presented in Section 6.1. By varying the number of covariates, we aim to see how BCF and ps-BART behave as well as motivate other algorithms that are designed to deal with large p. For instance, Hernández et al. (2018) propose a BART-based algorithm suitable for high-dimensional data (in particular when $p > 10,000$), named BART-BMA, that uses Bayesian model averaging and does not utilise an MCMC algorithm. Via simulation studies, they show that BART-BMA outperforms the standard BART when the number of covariates is large. In this context, Linero (2018) introduces Dirichlet BART (DART) that modifies the variable selection in BART by updating the probability of a predictor being selected as a split variable via a Dirichlet distribution. With this change, DART tends to produce more accurate predictions than BART in situations where p is large. We also explored MOTR-BART (Prado et al., 2020), which is an algorithm that generalises BART by generating the predictions based on piece-wise linear functions rather than terminal node constants. Here, we are specially interested in seeing how MOTR-BART (with 10 trees) performs when $\mu(\mathrm{x})$ is linear, as the regularised linear regression with the horseshoe prior presented the best results in Table 2 of Section 6.1.

In Figure 1, we present the RMSE obtained from 50 Monte Carlo simulations for the Conditional Average Treatment Effect (CATE), and measured the methods' performance on the training sets considering an estimate of the propensity score as a covariate for the non-causal algorithms. When $p = 500$, it was not possible to run BCF

*Corresponding author: estevao.prado@mu.ie.
[†]Hamilton Institute, Maynooth University, Ireland
[‡]Department of Mathematics and Statistics, Maynooth University, Ireland
[§]Insight: The National Centre for Data Analytics, Maynooth University, Ireland
[¶]Econometric Institute, Erasmus University Rotterdam, Netherlands
[‖]School of Medicine, Trinity College Dublin, Ireland

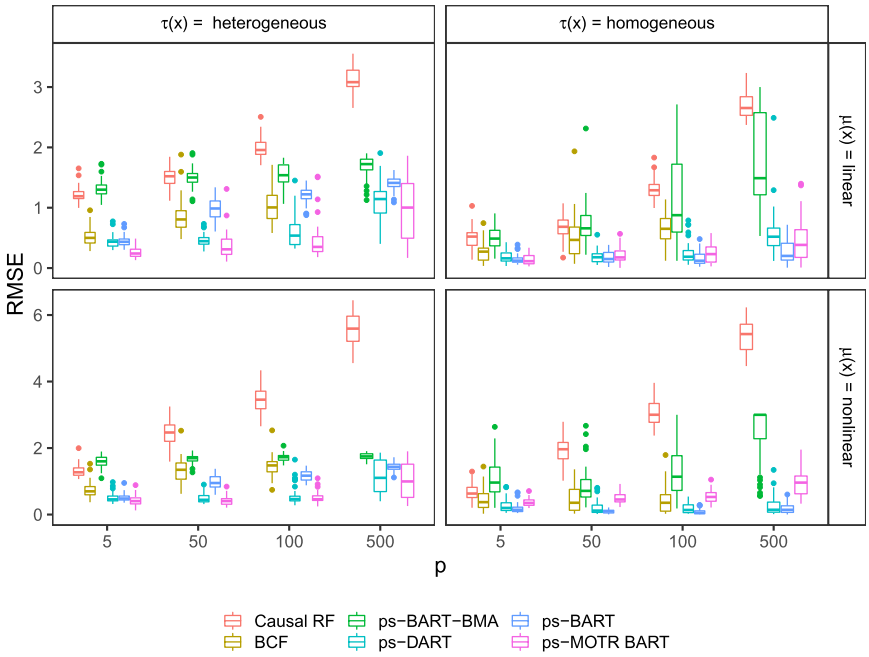


Figure 1: Simulation study results of RMSE for Conditional Average Treatment Effect (CATE).

due to numerical errors. Firstly, we see that the Causal RF algorithm is highly sensitive to the number of covariates. For all combinations of $\tau(\mathrm{x})$, $\mu(\mathrm{x})$ and estimands, the RMSE values for Causal RF tend to increase as p gets bigger. When $\tau(\mathrm{x})$ is heterogeneous, we see that ps-MOTR-BART is competitive even when the structure of $\mu(\mathrm{x})$ is nonlinear. In addition, ps-DART presents lower RMSE values than ps-BART, which might suggest that ps-DART could be an alternative in situations where p is large (Santos and Lopes, 2018). On the right-hand side of Figure 1, the results of RMSE for CATE are shown when $\tau(\mathrm{x})$ is homogeneous, which we believe is unrealistic in practice. Here, we see that for $p < 500$ and linear $\mu(\mathrm{x})$ that ps-BART, ps-DART and ps-MOTR-BART present excellent levels of accuracy. For instance, when $p = (5, 50$ and $100)$, they produce similar results with the three generating more accurate estimates than BCF. With $\tau(\mathrm{x})$ homogeneous and a nonlinear structure for $\mu(\mathrm{x})$, however, we note that MOTR-BART does not produce as accurate estimates as ps-BART and ps-DART.

In Figure 2, we see that the results for the Average Treatment Effect (ATE) are similar. That is, ps-BART and ps-DART perform well across all simulations and ps-MOTR-BART performs particularly well in data with heterogeneous effects and linear $\mu(\mathrm{x})$. Also, BCF does not give better results than ps-BART or ps-DART in any setting, although it outperforms Causal RF and ps-BART-BMA.

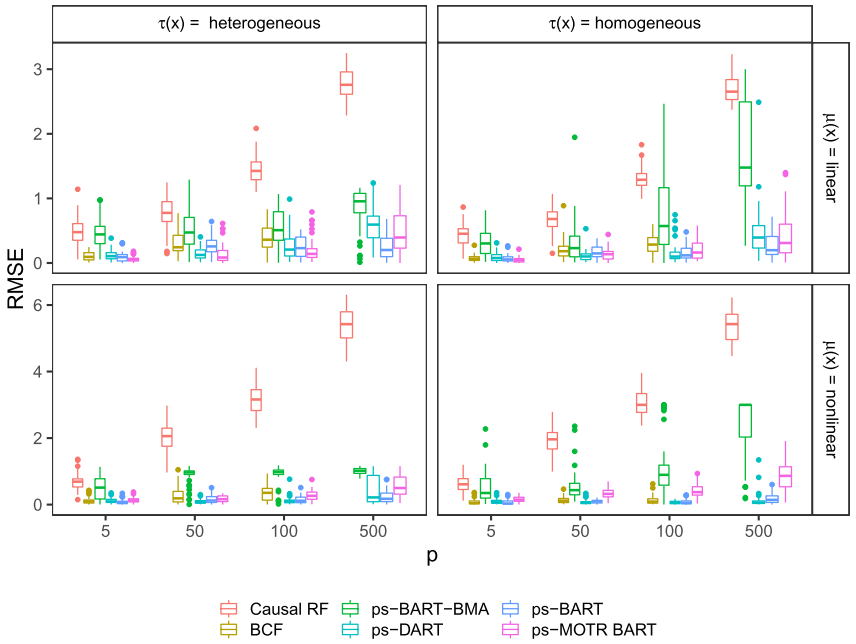


Figure 2: Simulation study results of RMSE for Average Treatment Effect (ATE).

Although not shown, we also explored the methods' performance on test data. Perhaps due to the non-stochastic nature of the simulation equations, we did not observe large differences between the results in training versus test performance. The results presented here can be reproduced by using the R scripts available at https://github.com/ebprado/BCF-discussion-paper.

References

Hernández, B., Raftery, A. E., Pennington, S. R., and Parnell, A. C. (2018). "Bayesian additive regression trees using Bayesian model averaging." *Statistics and Computing*, 28(4): 869–890. MR3766048. doi: https://doi.org/10.1007/s11222-017-9767-1. 1029

Linero, A. R. (2018). "Bayesian regression trees for high-dimensional prediction and variable selection." *Journal of the American Statistical Association*, 113(522): 626–636. MR3832214. doi: https://doi.org/10.1080/01621459.2016.1264957. 1029

Prado, E. B., Moral, R. A., and Parnell, A. C. (2020). "Bayesian additive regression trees with model trees." *arXiv preprint arXiv:2006.07493*. MR3858513. doi: https://doi.org/10.4310/SII.2018.v11.n4.a1. 1029

Santos, P. H. F. d. and Lopes, H. F. (2018). "Tree-based bayesian treatment effect analysis." *arXiv preprint arXiv:1808.09507*. 1030

Contributed Discussion

Harrison Zhu*, Xing Liu†, Alberto Caron‡, Ioanna Manolopoulou§,
Seth Flaxman¶, and François-Xavier Briol‖

In the Neyman-Rubin causal model, patient i is represented through a triplet (X_i, Z_i, Y_i), where X_i denotes covariates, Z_i denotes whether the patient received a treatment ($Z_i = 1$) or not ($Z_i = 0$), and Y_i represents the outcome (in particular $Y_i(1)$ is when the patient received the treatment and $Y_i(0)$ if they did not). Given such data for n patients, we would like to answer questions about the effect of the treatment on the outcome variables. These questions can be answered by considering certain statistics of interest (Hill, 2011). These include the (population) average treatment effect (ATE), given by $\mathbb{E}[Y(1) - Y(0)]$, the sample average treatment effect (SATE), given by $\frac{1}{n}\sum_{i=1}^{n}(Y_i(1) - Y_i(0))$, the population average effect of the treatment on the treated (PATT), given by $\mathbb{E}[Y(1) - Y(0)|Z = 1]$, and the sample average treatment effect of the treatment on the treated (SATT), given by $\frac{1}{n}\sum_{i:Z_i=1}(Y_i(1) - Y_i(0))$.

An interesting point is that each of these quantities can be expressed as an integral of the conditional average treatment effect (CATE), $\tau(x) = \mathbb{E}[Y(1) - Y(0)|X = x]$, over some distribution on covariates. As discussed in Hahn et al. (2020), this can be estimated by integrating a model $\hat{\tau}$ of τ, such as a Bayesian posterior mean. This remark, although seemingly trivial, is particularly interesting since it opens up connections with the field of probabilistic numerics and especially Bayesian probabilistic numerical integration (BPNI) (Diaconis, 1988; O'Hagan, 1991; Rasmussen and Ghahramani, 2002; Briol et al., 2019).

In BPNI, the goal is to tackle challenging problems in numerical analysis, such as the computation of an intractable integral, using tools from Bayesian nonparametrics. The motivation is that the Bayesian framework can be used to quantify uncertainty over the value of the integral. This is done through three steps: (i) a prior is placed over the integrand, (ii) this prior is conditioned on values of the function to obtain a posterior on the integrand, (iii) this posterior on the integrand implies a (one-dimensional) posterior on the value of the integral. Different prior choices allow us to encode properties of the integrand, such as smoothness or periodicity, in a straightforward manner, leading to algorithms which respect these properties. The most common model is a Gaussian process (GP); although more recent work also considers alternatives such as Bayesian additive regression trees (BART) (Zhu et al., 2020) or multi-output Gaussian processes (Xi et al., 2018; Gessner et al., 2019).

*Department of Mathematics, Imperial College London, hbz15@ic.ac.uk
†Department of Pure Mathematics and Mathematical Statistics, University of Cambridge, liuxing971015@outlook.com
‡Department of Statistical Science, University College London, alberto.caron.19@ucl.ac.uk
§Department of Statistical Science, University College London, i.manolopoulou@ucl.ac.uk
¶Department of Mathematics, Imperial College London, s.flaxman@imperial.ac.uk
‖Department of Statistical Science, University College London and The Alan Turing Institute, f.briol@ucl.ac.uk

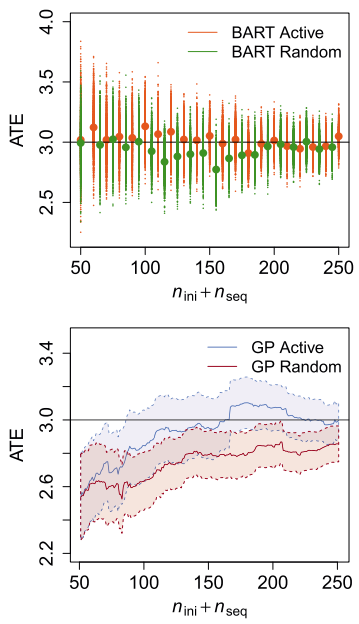


Figure 1: Estimates of the ATE for active learning and random sampling.

We can hence see the computation of ATE, PATT, SATE and SATT through integrals of the CATE as applications of BPNI to causal inference. This leads to several remarks:

1. There are a number of consistency results for BPNI methods, including refined convergence rates in a variety of scenarios depending on the model used, the domain of integration, the data generating process and the smoothness of the integrand. These could provide strong theoretical guarantees for the estimation of the ATE, SATE, SATT and PATT in a wide variety of settings; see Briol et al. (2019); Kanagawa and Hennig (2019); Kanagawa et al. (2020); Wynne et al. (2020).

2. **Active Learning:** The field of BPNI has derived a variety of experimental design schemes targeting directly the efficient approximation of integrals (rather than approximation of functions); see Osborne et al. (2012); Gunter et al. (2014); Briol et al. (2015); Jiang et al. (2019) in the case of Gaussian process models, and Zhu et al. (2020) for BART. Again, these could lead to more efficient estimates of the quantities of interest in causal inference.

To highlight the potential benefits of active learning schemes for causal inference with GPs and BART, we considered the synthetic example in Section 6.1 of Hahn et al. (2020) with homogeneous treatment effects and linear prognostic function.[1] We assume

[1]The code is available at https://github.com/ImperialCollegeLondon/BART-Int.

the model $Y_i = f(X_i, Z_i) + \epsilon_i$, where $\epsilon_i \sim N(0, 0.1^2)$ and the covariates X_i are i.i.d. with a known distribution Π. We consider the computation of the ATE estimated as $\Pi_m[\hat{f}(X,1) - \hat{f}(X,0)]$, where $\hat{f}$ is the fitted posterior of the Bayesian model (e.g. BART or GP) on f, and Π_m is an empirical distribution formed by samples $\{\tilde{x}_i\}_{i=1}^m$ representative of Π (and potentially different from the observed covariates $\{x_i\}_{i=1}^n$).

For the active learning algorithms (see Zhu et al., 2020 for full details), we use a candidate set of size $m = 2000$, then begin with $n_{\text{ini}} = 50$ initial random design points and acquire $n_{\text{seq}} = 200$ additional points. We use a sequential design with new point selected one at a time through the following objective: At iteration n, we select x_{n+1} and z_{n+1} as follows $\text{argmax}_{c=(x,z)} \mathbb{V}[f(x,z)\pi(x)e(z)|\{(x_i, z_i, y_i)\}_{i=1}^n]$, where π is the density of Π and e is the propensity function. We can see that active learning helped both models to obtain improved estimates of the true value of ATE (ATE = 3).

Overall, we conclude that the fields of causal inference and BPNI could benefit from further interactions. In this note, we pointed out how recent advances in BPNI could lead to further practical and theoretical advances in causal inference (including through a small synthetic example), but it is also clear that applications in causal inference could provide motivation for the development of novel BPNI algorithms.

References

Briol, F.-X., Oates, C. J., Girolami, M., and Osborne, M. A. (2015). "Frank-Wolfe Bayesian quadrature: Probabilistic integration with theoretical guarantees." In *Neural Information Processing Systems*, 1162–1170. 1033

Briol, F.-X., Oates, C. J., Girolami, M., Osborne, M. A., and Sejdinovic, D. (2019). "Probabilistic integration: A role in statistical computation? (with discussion)." *Statistical Science*, 34(1): 1–22. MR3938958. doi: https://doi.org/10.1214/18-STS660. 1032, 1033

Diaconis, P. (1988). "Bayesian numerical analysis." *Statistical Decision Theory and Related Topics IV*, 163–175. MR0927099. 1032

Gessner, A., Gonzalez, J., and Mahsereci, M. (2019). "Active multi-information source Bayesian quadrature." In *Uncertainty in Artificial Intelligence*. 1032

Gunter, T., Garnett, R., Osborne, M., Hennig, P., and Roberts, S. (2014). "Sampling for inference in probabilistic models with fast Bayesian quadrature." In *Advances in Neural Information Processing Systems*, 2789–2797. 1033

Hahn, P. R., Murray, J. S., and Carvalho, C. M. (2020). "Bayesian regression tree models for causal inference: regularization, confounding, and heterogeneous effects." *Bayesian Analysis*. 1032, 1033

Hill, J. L. (2011). "Bayesian nonparametric modeling for causal inference." *Journal of Computational and Graphical Statistics*, 20(1): 217–240. MR2816546. doi: https://doi.org/10.1198/jcgs.2010.08162. 1032

Jiang, S., Chai, H., Gonzalez, J., and Garnett, R. (2019). "BINOCULARS for efficient, nonmyopic sequential experimental design." *arXiv:1909.04568*. 1033

Kanagawa, M. and Hennig, P. (2019). "Convergence guarantees for adaptive Bayesian quadrature methods." In *Neural Information Processing Systems*, 6237–6248. 1033

Kanagawa, M., Sriperumbudur, B. K., and Fukumizu, K. (2020). "Convergence analysis of deterministic kernel-based quadrature rules in misspecified settings." *Foundations of Computational Mathematics*, 20: 155–194. MR4056928. doi: https://doi.org/10.1007/s10208-018-09407-7. 1033

O'Hagan, A. (1991). "Bayes–Hermite quadrature." *Journal of Statistical Planning and Inference*, 29: 245–260. MR1144171. doi: https://doi.org/10.1016/0378-3758(91)90002-V. 1032

Osborne, M. A., Duvenaud, D., Garnett, R., Rasmussen, C. E., Roberts, S., and Ghahramani, Z. (2012). "Active learning of model evidence using Bayesian quadrature." In *Advances in Neural Information Processing Systems*, 46–54. 1033

Rasmussen, C. and Ghahramani, Z. (2002). "Bayesian Monte Carlo." In *Advances in Neural Information Processing Systems*, 489–496. 1032

Wynne, G., Briol, F.-X., and Girolami, M. (2020). "Convergence guarantees for Gaussian process means with misspecified likelihoods and smoothness." *arXiv:2001.10818*. 1033

Xi, X., Briol, F.-X., and Girolami, M. (2018). "Bayesian quadrature for multiple related integrals." In *International Conference on Machine Learning*, 8533–8564. MR3577382. doi: https://doi.org/10.1214/16-BA1017A. 1032

Zhu, H., Liu, X., Kang, R., Shen, Z., Flaxman, S., and Briol, F.-X. (2020). "Bayesian probabilistic numerical integration with tree-based models." *arXiv:2006.05371*. 1032, 1033, 1034

Contributed Discussion

Daria Bystrova*, Julyan Arbel*, and Thibaud Rahier†

In this discussion we elaborate on the specific scenario when the selection into treatment depends on the outcome under no treatment. We reproduce Example 1 in the paper with $p = 2$ control variables and $n = 250$ observations under the data generating process

$$Y_i = \mu(x_1, x_2) - \tau Z_i + \epsilon_i, \quad i = 1, \dots, n,$$
$$\epsilon_i \sim \mathrm{N}(0,1), \quad x_{i1}, x_{i2} \sim \mathrm{Uniform}(0,1),$$

where Y is a measure of heart distress, Z the treatment indicator, and x_1 and x_2 are two control variables. The treatment effect τ is supposed to be homogeneous and set to $\tau = 1$ here. The prognostic function is set to $\mu(x_1, x_2) := \mathrm{E}(Y \mid x_1, x_2, Z = 0) = \mu(x_1, x_2) = 6\Phi(2(x_1 - x_2)) - 3$.

The authors argue that, under strong confounding, *the estimation of average treatment effect (ATE) using BART can exhibit severe bias.* The BCF model is designed to solve this issue, by including the propensity score $\pi(x)$ (or its estimate) in the set of predictors when learning ATE. In this contributed comment, we have chosen to study the extent to which BCF exhibited smaller bias than BART for different levels of confounding. The confounding 'amount' is controlled by a scalar $\alpha \in [0,1]$ hyperparameter in our experiments:

$$\pi_\alpha(\mu(x_1,x_2),x_1,x_2) := \alpha \left(0.8\Phi\left(\frac{\mu(x_1,x_2)}{0.1(2-x_1-x_2)+0.25)}\right)+0.025(x_1+x_2)\right) + 0.05\beta_\alpha,$$
$$\beta_\alpha := \frac{1}{2}(19 - 17\alpha).$$

For $\alpha = 0$ the propensity score is constant and we are in a randomized controlled trial situation (no confounding); for $\alpha = 1$ the propensity score is roughly equivalent to the one used in Example 1 presented in the paper (high confounding). The rationale behind the choice of β_α is that π_α is required to stay roughly constant for average values of the control variables $x_1 = x_2 = 1/2$; see Figure 1. The bias$(\hat{\tau}) = \mathrm{E}[\hat{\tau} - \tau]$ is computed as an empirical average over 50 datasets.[1]

We observe from Figure 1 that for α close to 0, BART and BCF exhibit similar bias, which is expected since BCF has no regularization-induced confounding (RIC) to correct in that case. Moreover, the more α increases, the larger the bias exhibited by BART compared to BCF. This is consistent with the fact that the bias increase implied by the increase of α is mainly due to the increase in confounding, which BCF was designed to handle better than BART.

*Univ. Grenoble Alpes, Inria, CNRS, LJK, 38000 Grenoble, France, daria.bystrova@inria.fr, julyan.arbel@inria.fr

†Criteo AI Lab, 38000 Grenoble, France, t.rahier@criteo.com

[1]Code is available on Github https://github.com/dbystrova/bcf_discussion.

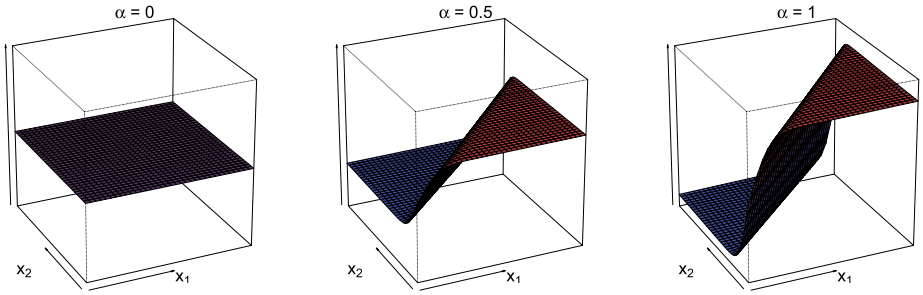


Figure 1: Propensity function $\pi(x_1, x_2)$ for different values of α in $\{0, 0.5, 1\}$.

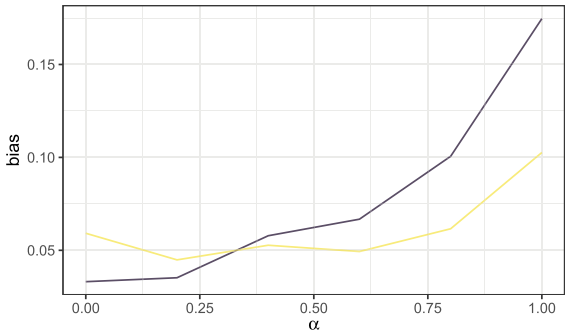


Figure 2: Bias for BART (purple) and BCF (yellow) models in Example 1 of the paper for varying confounding 'amount' $\alpha \in [0, 1]$.

The results of our exploratory study are mainly consistent with what was expected upon reading the paper. However, there remains some unanswered questions. (i) Our experiments show that for $\alpha = 0$, BART exhibits slightly lower bias than BCF: this is intriguing as in the case of constant propensity score, both BCF and BART are expected to behave the same in terms of ATE estimation. (ii) We notice that the bias exhibited by BCF increases when α increases (even though it does less so than BART's bias). From our understanding, the only additional source of bias the models could suffer from when $\alpha > 0$ compared to $\alpha = 0$ is the result of RIC, which we expect to be fully handled by BCF. These results suggest either that the increase of α induces an other form of bias, or that BCF is not entirely solving the RIC problem.

Note that both of these open questions could be explained by sub-optimal tuning of BART and BCF models in our experiments. We are conscious that our study only scratches the surface of the complicated problem of causal inference and regularization-induced confounding. More experiments are needed to fully understand the extent to which BCF better handles RIC than BART.

Rejoinder

P. Richard Hahn*, Jared S. Murray†, and Carlos M. Carvalho‡

We would like to thank the discussants for sharing their perspectives on our paper, and for suggesting numerous interesting avenues for improvements and extensions. Before responding to specific points raised in the discussions, we would like to take the opportunity to provide readers with updates about software, applications, and recent methodological developments.

First, the BCF R package is now in version 2.0[1] with improvements that include:

- Parallel computing (within and between chains).
- Saving tree "traces", to generate predictions and treatment effects at new locations without refitting the model.
- Automated MCMC diagnostics via the `coda` package for a handful of key parameters (the error variance, ATE, etc.).
- New vignettes to make the package more accessible, and to illustrate how to construct posterior summaries like the subgroup-finding trees presented in this paper.

The bulk of this is due to hard work from our collaborators at Mathematica, initiated by Mariel Finucane and her team including Peter Mariani, Constance Delannoy, and Lauren Forrow. Many thanks to them!

Indeed, we are happy to report that BCF is being used by Mathematica (Ghosh et al., 2020) and various other researchers with interesting and challenging applied problems (e.g. King et al. (2019); Bryan et al. (2019); Bail et al. (2020)). On our end, we deployed BCF in the National Study of Learning Mindsets (NSLM) (Yeager et al., 2019), a large-scale randomized study specifically designed to investigate the heterogeneous effects of a light-touch growth mindset intervention in first effect heterogeneity.

The initial findings from the NSLM reported in Yeager et al. (2019) include a BCF-based analysis of the treatment effect heterogeneity. Using BCF we identified similar patterns of heterogeneity to what was estimated using painstakingly pre-registered linear models, both by examining treatment effects in pre-registered subgroups and using posterior summarization tools like the CART-based subgroup search procedure presented in our paper. Although this was a randomized experiment, the BCF parameterization was vital here – we needed different sets of variables in μ and τ, as we had to incorporate

*School of Mathematical and Statistical Sciences, Arizona State University, prhahn@asu.edu
†McCombs School of Business, University of Texas at Austin, jared.murray@mccombs.utexas.edu
‡McCombs School of Business, University of Texas at Austin, carlos.carvalho@mccombs.utexas.edu
[1] `https://github.com/jaredsmurray/bcf`.

numerous baseline variables (in μ) to reduce residual standard deviation while focusing on only a handful of key treatment effect moderators (in τ) backed by substantive theory and previous studies.

Recent methodological developments related to BCF include BCF with multilevel modeling for dependent observations (Yeager et al., 2019), extensions to heterogeneous linear effects of continuous treatments (Woody et al., 2020a), priors on τ that induce partial smoothness (Starling et al., 2019), and new approaches to posterior summarization, building on earlier work in Hahn and Carvalho (2015) and Woody et al. (2020b). Specifically, (Woody et al., 2020a) re-examine Levitt's famous abortion and crime dataset (Levitt and Dubner, 2010) using an extension of BCF for continuous treatments. They corroborate many of the original findings while uncovering new evidence for heterogeneous effects. The use of BCF here is notable because the estimates and uncertainty intervals of treatment effects are known to be sensitive to *how* control variables appear in their model – that is, the choice of nonlinearities and interactions. Woody et al. (2020a) address the model specification dilemma by fitting BCF. (See also Bryan et al. (2019) and Dweck and Yeager (2019), who also use BCF to avoid specification search). Further, the posterior summaries of treatment effect heterogeneity in Woody et al. (2020a) take some pains to estimate *partial* effects of treatment effect moderators, which is important for understanding treatment effect moderation when the moderators are not independent.

Although they do not refer to it as such, in their discussion Zhu et al. consider using BCF for addressing "transportability" of causal effects. Transportability refers to the problem of generalizing estimated treatment effects to a new population. Under some identification conditions, estimating the ATE in the new population involves integrating $\tau(\mathrm{x})$ over a new population distribution of x, say P'_X. A simple Monte Carlo approach to this problem would sample directly from P'_X. However, this naive approach may require a large number of samples to achieve an accurate calculation of the population average treatment effect (PATE), and collecting x measurements in the target population is often costly (e.g. by conducting a survey). As such, active learning and other adaptive experimental designs have been proposed whereby one can accurately estimate the PATE with fewer samples and function evaluations of $\tau(\mathrm{x})$. BCF is useful here to the extent that it furnishes good estimates of $\tau(\mathrm{x})$, and the ability to control the set of possible effect moderators (as opposed to requiring that each variable be a potential confounder and moderator) can reduce the dimensionality in a critical way.

Finally, recent computational developments promise to expand the applicability and accessibility of BCF even further. Extending a computational approach developed by He (2019); He et al. (2019), He and Krantsevich's discussion reports on some recent advances on approximating and simulating from the BCF posterior, aiming to address both scalability (in n) and the inadequate mixing that sometimes occurs using the typical Metropolis-Hastings MCMC algorithm for updating trees and their parameters. Empirically, this poor mixing tends manifests as credible intervals that are artificially short. Running multiple independent MCMC chains is known to ameliorate or at least diagnose poor mixing in Bayesian tree models (Carnegie, 2019), but realizing these benefits can require long burn-in periods since it is difficult to identify "good" but

dispersed staring points for each chain. The clever stochastic search He and Krantsevich propose is quite promising, evidently yielding sets of chains that converge rapidly to their stationary distribution.

Turning now to the remaining discussions, we have organized our rejoinder around several common themes that appeared across the discussion papers.

1 On the central role of the propensity score

Professor Wager provides a thorough summary of recent (and some not-so-recent) research on heterogeneous treatment effect estimation, to which he has been a prominent contributor. As a part of this review, he gives a nice overview of the variety of different ways that propensity scores and their analogues have proven important in estimating treatment effects. It is curious to us that the importance of using an (estimated) propensity score (PS) to estimate causal effects does not seem to have been controversial for frequentists, but has been the topic of much controversy among Bayesians.

1.1 Uncertainty in π and uncertainty in $\hat{\pi}$

Some of our discussants are still uncomfortable using a plug-in $\hat{\pi}$. We might never reach consensus here, but we think the discussion has helped us understand the point of divergence. Under the identifying assumptions in the paper, we tend to approach the problem of inferring treatment effects from a (Bayesian) outcome modeling perspective (rather than modeling treatment assignment and outcomes) since with ignorable treatment assignment we are not *obligated* to model treatment assignment and doing so invites new complications. By parameterizing the model in terms of the treatment effects and including $\hat{\pi}$ as a covariate in μ, *we are not attempting to model treatment assignment or introducing the true propensity score as a parameter*. Rather, we are inducing a prior over μ that is less likely to regularize away fluctuations μ along an important direction – a direction that is likely to be misattributed to Z.

Organisian and Roy express the common Bayesian perspective that

> Differences in prior uncertainty about the regression parameters [of the fitted propensity function] impacts uncertainty about [sic] $\hat{\pi}$. This, in turn, may lead to different degrees of uncertainty about $f(x_i, z_i, \hat{\pi}(x_i))$.

The error in this argument is a confusion between π and $\hat{\pi}$. There is no uncertainty about $\hat{\pi}$; it is a choice we make in our prior. Our priors, models, and ultimate inference for treatment effects are all conditional on realized treatment assignments and covariates, and $\hat{\pi}$ is a deterministic function of these. Fixing $\hat{\pi}$ merely represents a choice of how to represent (x, z) in constructing a prior over μ.[2]

Zooming out from causal inference, we make similar choices every time we decide to combine or transform covariates in a regression model, for example. These choices can

[2]While it may not be obvious to those unfamiliar with tree models, choosing different coordinate systems for their inputs generally leads to meaningfully different priors.

be important and consequential, as is how we define $\hat{\pi}$, but this does not mean that they *require* further prior distributions to yield valid Bayesian inference. We discuss this at some length in the paper, where we suggest comparing or ensembling different choices for $\hat{\pi}$ based on their consistency with observed treatment assignments and covariates (ignorant of the outcomes), and conducting sensitivity analysis to different choices for $\hat{\pi}$ the same way we would for other important components of a prior distribution.

Of course, there *is* uncertainty about π, and we would have to mind that uncertainty if we specified our model as $f(x_i, z_i, \pi(x))$, or if we were jointly modeling treatment assignment and outcomes in order to induce a particular outcome model (as Organisian and Roy propose). But that is not our proposal.

To revisit the analogy from section 8.4 of our paper, the popular Zellner prior in linear regression is illustrative here: when Zellner specifies his prior in terms of $\mathbf{X}^t\mathbf{X}$, no one insists that his prior is inadequate for failing to account for estimation uncertainty regarding the true covariance structure of X. One *could* specify the Zellner prior in terms of $\text{cov}(X)$ and then integrate over the associated uncertainty, but doing so would require specifying a model for X, with unclear benefits and significant additional complexity (Hahn et al., 2013). The position that every choice we make when specifying a prior distribution requires its own prior distribution leads quickly to an infinite regress of uncertainty and the collapse of Bayesian inference.

As for Organisian and Roy's proposal of inducing dependence on π via a joint model of response and treatment assignment, that approach poses its own risks. For example, the extent to which the propensity score actually features in their proposed outcome model (after conditioning on Z) depends in a complex way on how the base measures/mixture kernels are specified, not just on prior parameters in the EDP, since the dependence arises through mixture component assignment and the number of components realized will depend on the form of the base measure. And even if the induced form of the outcome model is appealing, estimating the parameters of that conditional model by fitting a joint model does *not* generally yield the best-fitting parameters for the conditional model alone (Hahn et al., 2013; Hahn, 2019)

1.2 The propensity score as effect moderator

In addition to its primary deconfounding role, Wager asks whether the propensity score might also be considered an effect moderator:

> In some applications, however, there may also be interest in the propensity score as an effect modifier.

In the notation of BCF, this question becomes: Does the propensity score go in τ (in addition to μ)? Quite possibly, yes, although we do not do so as a default. See our response to Organisian and Roy in the next section for one example of when that idea would be advisable. Whether or not the estimated propensity score appears as an argument for τ, one may *summarize* treatment effects as a function of $\hat{\pi}$, for example by computing the posterior of subgroup average treatment effects across quantile buckets of the propensity score.

Along similar lines, Ray, Szabo, and van der Vaart propose a regression model with an explicit adjustment that implies that the propensity score is *always* an effect moderator (absent some very specific cancellations in their W), but in a very particular way. The particular way that $\hat{\pi}$ enters into the model is specifically designed to debias inference for the ATE, not necessarily to improve estimation of CATEs. The resulting model is, in our opinion, prima facia implausible. That is, the form of the proposed model is not the sort of thing a subject matter expert would ever be likely to propose, but rather an instance of "frequentist pursuit" (Robins et al., 2015) that predictably does relatively well when evaluated against frequentist criteria.[3] We tend to think of ourselves as modelers first, so our preferred mode of optimizing for frequentist operating characteristics would be to include estimates from BCF or another model as inputs to a doubly robust estimation procedure; see e.g. Dorie et al. (2019) where (ps)BART with a TMLE adjustment did better in repeated sampling than BART or TMLE without BART in estimating ATT. That is, we would rather specify a plausible model and then make an adjustment to model-based estimates rather than making a frequency-motivated adjustment to the model itself. This approach is under-explored in the literature and under-utilized in practice.

1.3 True versus estimated propensity scores

Wager poses a hypothetical question that is, in a sense, the opposite of the proposition that uncertainty about π is somehow essential: what if, instead of having to account for uncertainty in our propensity score estimate, we might be better off using the estimate (as opposed to the true propensity function)?

> In the context of estimating an average treatment effect, Hirano, Imbens, and Ridder [2003] showed that using an estimated propensity score for weighting could sometimes be enough to achieve efficiency.

This is an interesting question. Hirano et al. are often read as implying that all we need is the estimated PS, and in fact if someone handed us the true PS we would not need or want to use it at all. But they were considering a limited universe of estimation procedures (IPW using either the true or estimated PS). Herren and Hahn (2020) show that in fact one can improve inferences using the true propensity score considering a slightly larger set of estimation procedures. In the context of ps-BART/BCF, in simulations we do generally observe modest gains using the true propensity score versus an estimate, although the gap narrows predictably when sample sizes are large or the propensity score function is easier to estimate.

1.4 ps-BART and BCF are complements, not competitors

To conclude this section, we wish to note for the record that the idea of using a propensity score estimate as a covariate in BART (or in other Bayesian nonparametric outcome

[3]We suspect Wasserman and Robins might accuse us of frequentist pursuit too, for regressing on an estimated propensity score. Our consciences are clear as we were primarily in search of better Bayesian inference, and – more seriously – we feel our approach is well-motivated on substantive grounds, and per the discussion in Section 1.1.

models for treatment effects) had, to our knowledge, not been proposed prior to our paper. In response to an early draft of this paper, ps-BART was added to the methods implemented in the ACIC 2016 data challenge Dorie et al. (2019) where it significantly improved on BART, and eventually became the default in the R package `bartCause`. Although using the PS as a covariate within BART seems like an obvious idea in retrospect (and had been done in other contexts previously, as mentioned by Wager), it was not standard practice in 2016 or 2017. Thus, we are inclined to view "ps-BART" as one half of our contribution here (the other being explicitly parameterizing the model in terms of τ so that it is possible control its prior) rather than as a competing methodology. Simply including an estimated propensity score does much to mitigate the potential for RIC, and is, well, simple to implement. RIC is not unique to Bayesian tree models and we expect that other Bayesian nonparametric methods would benefit from this simple adaptation.

2 Does targeted selection exist, and does it matter?

In the paper, we motivated our suggestion to incorporate the propensity score as a predictor with the idea of what we called "targeted selection" – where $\pi(x)$ is a monotone function of $\mathrm{E}(Y \mid x, Z = 0)$, or approximately so, the intuition being that individuals select into treatment based on predicted outcomes under control. Some of the discussants, such as Papadogeorgou and Li, Wager, and Bystrova et al. do not consider this idea controversial. Others, namely Organisian and Roy and Ray et al. question if targeted selection is a legitimate concern, or driving the improvements realized by incorporating the estimated propensity score.

Ray et al. write:

> However, it seems that different from what this name suggests, the general principle concerns bias that may result from estimating a high-dimensional parameter and is not special to causality or confounding. In the same way, it does not seem that "targeted selection" is a precondition for the usefulness of bias selection. Rather, not using independent prior modeling seems one of these novel properties of high-dimensional modeling that remain to be further explored.

We never meant to imply that targeted selection is the *only* mechanism by which regularization-induced confounding may arise, merely that it is one way such bias can occur, and it is especially pernicious, since it leads $\mu(\mathrm{x})$ to vary in the direction of $\mathrm{E}(Z \mid \mathrm{x}) = \pi(\mathrm{x})$. Furthermore it is a plausible one in the context of causal inference, meaning we might reasonably expect it to show up in some applications. It is certainly true that high-dimensional inference is hard for a number of reasons, but note that we can reliably induce RIC using targeted selection in low dimensions ($p = 2$) and using linear models. The issue is not merely one of high-dimensional modeling either.

Meanwhile, Organisian and Roy question targeted selection on substantive grounds, stating that:

> ... Treatment decisions are ideally made based on trying to predict $\mathrm{E}(Y(1) - Y(0) \mid X)$.

This is certainly a distinct selection mechanism that we considered but did not report on in our paper – the possibility that π is a function of (an estimate of) τ. Supposing selection works this way requires imagining that individuals have detailed knowledge of the heterogeneous effects of treatments. This may be true in some cases, but in many other cases it seems more likely that we would have detailed information about risk factors for adverse outcomes absent treatment, and a general sense that available treatments tend to be effective (based on the ATE or perhaps subgroup effects in one or more RCTs, for example). This is consistent with our own experiences in doctors' offices.

However, we see no reason that both of these mechanisms couldn't be at play. In fact, even with detailed knowledge of treatment effect heterogeneity we would be foolish to select into or out of treatment without reference to expected outcomes in the absence of treatment. If a drug is expected to lower a patient's blood pressure by 100 units (so $\tau(x) = -100$), the doctor desperately needs to know $\mathrm{E}(Y \mid x, z = 0)$ as well; if the baseline blood pressure is 220, then the drug is indicated, if it is 120, the drug would be deadly! So we take issue with the assertion that "treatment decisions are ideally made" using predictions of $\tau(x)$. Ideally one would make treatment decisions based on predictions of (or probability distributions over) outcomes under *both* treatment decisions.

2.1 The connection between targeted selection and RIC

The idea of targeted selection arose from our experience trying to generate data that exhibited significant confounding that was difficult to remedy with a simple (nonparametric) regression adjustment – what we call *regularization induced confounding* (RIC), that is, bias that persists in finite samples even when one has all measured confounders.

This is easy to do with linear models and homogenous effects, which we knew from earlier work (Hahn et al., 2016b). In a traditional linear regression model for causal inference, the mere existence of confounders – variables appearing both in $\pi(x)$ and $\mu(x)$ – results in "aliasing", by which we mean that in finite samples it is unclear what variation in the response is due to μ and what is due to τ and Z. While μ and τ are identified, in practice they can be hard to tell apart. In Bayesian terminology, the likelihood for the treatment effect coefficient is liable to be multi-modal and/or has flat regions, even when conditioning on a sufficient set of controls. RIC arises because the prior over regression coefficients might strongly prefer one of those regions solely because of the prior on the *control variable coefficients*, rather than beliefs about treatment effects. We provide some calculations showing this in our paper, and extensive evidence in Hahn et al. (2016b).

In nonlinear models, this aliasing is less likely to arise. Nonlinear prognostic effects (partial effects of control variables) do not mask their own confounding influence over the entire domain the way hyperplanes do. So as long as the prognostic impact of a control variable differs in form from its confounding impact *somewhere* along its domain, we can spot it. We concede that this explanation is not as clear as it might be, but hopefully it is at least suggestive; we have not yet figured out how to tackle this explanation with much mathematical rigor.

What we do know is that if you generate data from, say, two arbitrary high degree polynomials, one governing the response surface and the other governing the treatment selection mechanism, regularization-induced confounding is not likely to be much of an issue. This raised an important question for us: does this mean that simply fitting a nonlinear model solves the causal inference problem? Or does it mean that the putative data generating process is not realistic? After considerable reflection, we concluded that targeted selection was a plausible explanation for how one could get appreciable aliasing (that is, the kind of problematic confounding that leads to RIC) and that, therefore, RIC was something that needed to be remedied. Finally, we note that aliasing in nonlinear models can indeed occur "by accident", and we observe this in the 2016 ACIC data sets (which did not employ a targeted selection data generating process). However, in practice, targeted selection is a reliable, plausible, and context-appropriate way to induce aliasing, and therefore RIC.

3 Maybe something works better than BCF

Our discussants put forth a bevy of alternatives to and modifications of BCF, some actually implementing the methods and assessing them in simulations as well. This is great! Obviously the current paper would not exist if we hadn't had a similar impulse, and we would encourage these authors to evaluate their proposals carefully against good-faith implementations of BART, ps-BART, and BCF. We would also plug one more time the ACIC data challenges as a useful testbed that come with a host of results from different methods for free. In subsequent sections we have more to say about the particular case of how BCF and other methods perform under poor or no overlap, and designing empirical studies, but a few comments on implementing BCF are in order here.

3.1 One nice thing about explicit priors: you can change them

We arrived at our default prior for $\tau(\mathrm{x})$ in BCF through careful consideration and extensive simulation exercises. Our choices reflect a belief that a default prior should be skeptical of heterogeneity by default, and shrink toward constant effects. We operationalize these beliefs in our default prior by (1) penalizing depth more strongly than the original BART prior, since the trees in τ already represent interactions (with the treatment indicator), (2) reducing the prior splitting probability that a tree splits at all, since the presence of more root-only trees shrinks more strongly toward homogenous effects, and (3) using relatively few trees in τ (50 by default), which amplifies the shrinkage implied by (1) and (2).

But as the title above suggests, all priors are subject to revision. In our own applications of BCF we routinely use non-default priors. For example, when prior knowledge strongly suggests *some* treatment effect heterogeneity, we might increase the number of trees and increase the probability of the first split for each tree, but continue to strongly penalize tree depth in our prior. By favoring trees with a single split *a priori* we shrink toward additive rather than homogeneous (constant) treatment effect functions.

Prior choices are consequential. As Papadogeorgou and Li discuss, there are times when strongly shrinking towards homogeneity may be unwise and in those cases ps-BART – or even the f_0, f_1 parameterization absent a propensity score estimate – may outperform default BCF. But so may non-default BCF! In our opinion, many simulation exercises feature implausibly large treatment effects, treatment effect variation, and low levels of noise that will tend to favor other priors. But we remain convinced that the additional shrinkage of τ provided by BCF is a sensible default in many applied problems, and the BCF parameterization – unlike BART or ps-BART – furnishes a knob to directly modulate the prior over treatment effect heterogeneity.

4 (Lack of) overlap, and what to do about it

Organisian and Roy and Papadogeorgou and Li both raise the issue of overlap and how BART/BCF behave under violations of overlap. This seems to be a fairly common concern; we hear it raised in workshops and as questions in our own talks. Overlap, like ignorability, is a strong identifying assumption that deserves more attention than it gets. The same is true of practical violations of overlap – when the true $\pi(\mathrm{x})$ is bounded away from zero or one, but we still have few or no treated or control cases in some portion of covariate space due to small sample size or bad luck. Practical overlap violations mean that some CATEs may be inestimable in our particular sample, even though they are formally identified.

Some of our discussants seem troubled by how BART/BCF behave under *actual* violations of overlap. We are not. In our view, it is simply too much to expect a model to generically do "well" when we are trying to make inference far from the support of the observed data, unless we are bringing very well-informed model and prior specifications to the table. While the best nonparametric priors will include some (typically weak) information about the size, variability, and smoothness of treatment effects, they do not usually furnish the kinds of detailed prior information that would make them reliable interpolators or extrapolators. In other contexts we expect this posture would be less controversial, but for some reason it takes on a different valence when talking specifically about overlap in causal inference.

For instance, in their example Papadogeorgou and Li indicate that "BART is clearly overconfident in the region of poor overlap". But examining their Figure 2, to our eye the BART intervals are actually at least as wide as the Gaussian process (GP) intervals everywhere we have treated and control data! We would argue that all three methods are overconfident in their extrapolations beyond that range, where the treatment effects and counterfactual predictions are unidentified (or at least inestimable), although we would certainly agree that BART is *more* overconfident. Their example also illustrates the potential danger of blindly extrapolating under very flexible models: Note how the GP-estimated expected outcomes tick up or down outside the range of the data based on a handful of observations at the extremes, as opposed to BART and the linear model which extrapolate in predictable ways. (To be clear, Papadogeorgou and Li do not themselves suggest that we blindly extrapolate!) So while we agree with Organisian and Roy's sentiment that "it may be ideal to instead have a model that performs reasonably

in these regions by default", we are not optimistic that a default procedure will solve our overlap problems for us – especially not one based on very flexible models – and are disinclined to trust models alone to estimate unidentifiable or inestimable quantities.

In our view, the best way to deal with poor overlap or violations of overlap is to try to identify regions of poor overlap, and (1) restrict inferences to regions of sufficient overlap, however we might define that, and/or (2) conduct a careful sensitivity analysis under different models and assumptions for how CATEs in the poor overlap region relate to the CATEs where we have overlap. Obviously this is much easier said than done – particularly (2) – but we have applied this mode of analysis profitably in other situations where selection and missing data led to a lack of point identification (Hahn et al., 2016a); see also Linero and Daniels (2015, 2018); Franks et al. (2019) for similar approaches to Bayesian inference without point identification. Hill et al. (2013) give some diagnostics for assessing overlap using BART output; their approach is adaptable to any Bayesian nonparametric model, including BCF, and there are many other strategies for addressing (1).

Organisian and Roy make some other specific prescriptions for how to handle lack of overlap. These include

- Heteroskedastic errors. Organisian and Roy state that in BART/BCF "var(ϵ_i) is constant with respect to x. This prevents model uncertainty from varying within $\mathcal{X}_{NO}$ (the region of covariate space with no overlap)". This is obviously false (evident from their own figures, as well as Figure 2 in Papadogeorgou and Li, and this increased model uncertainty features prominently in Hill et al. (2013)). The statement confuses uncertainty about future observations with uncertainty about model parameters. Further, one suggested fix (George et al., 2019) doesn't actually feature heteroskedastic errors, and a heteroskedastic BART fit (Pratola et al., 2020; Murray, 2020) to e.g. Papadogeorgou and Li's example in Figure 2 will estimate an approximately constant variance function $\sigma^2(\text{x})$ and furnish essentially identical inferences as a homoskedastic model, since the true error variance is evidently constant. Heteroskedasticity may be desirable in its own right, particularly across treatment arms, but it is not a solution for representing uncertainty due to lack of overlap.

- Soft-BART (Linero and Yang, 2018). This prior smooths over decision boundaries, but results in the same constant extrapolations as BART. In Papadogeorgou and Li's example we would get smoother curves over the support of the data, but essentially the same inferences at high and low ages. Its smooth *interpolations* could perhaps improve estimation error when overlap violations occur in the interior of covariate space, but this unlikely to yield substantially different inferences or wider intervals than a similar BART prior since the difference between the two priors is essentially averaging sigmoids versus step functions. Compare this to the nearly unrestricted functional form of GP or DPM interpolations, which is what yields their large credible intervals.

- Nethery et al. (2019)'s "fit-the-fit" spline-extrapolation method is closest to what we propose above, and to what we have done before in cases of un- or partial

identification (i.e. obtain a posterior over identified quantities, and model the unidentified quantities conditional on those). Without inspection, using higher-degree additive splines fit to a BART fit to fill in effects in the non-overlap region makes us nervous for many of the same reasons as extrapolating from a GP fit. But an additive spline "extrapolator" fit to the BART fit is at least relatively easy to inspect, diagnose, and refine as part of a sensitivity analysis (see Woody et al. (2020b) for some related discussion).

Finally, Papadogeorgou and Li suggest further investigation of how different methods behave under varying *degrees* of overlap. We heartily endorse this! A good starting point might be revisiting previous empirical evaluations like the ACIC data challenges, which feature varying degrees of selection and confounding, and computing performance metrics within subgroups defined by the true propensity score.

5 Designing and interpreting empirical studies

Several of our discussants include small repeated sampling studies of BART, BCF, and other methods. These kinds of empirical studies are close to our hearts. In this section we react to a few of these after outlining our thinking regarding the design and interpretation of empirical studies (see also Carvalho et al. (2019), where we apply these principles while also calibrating against a real dataset).

5.1 Specifying the data generating processes (DGPs)

Simulation studies in machine learning and statistics papers often tell you more about the data generating process than they tell you about the real-world performance of the models/estimators/algorithms being compared. Data generating processes should be designed with one of two main goals in mind. One, they might seek to be representative of the intended use case, in which case the performance comparisons can be suggestive about what sort of real-world performance we might expect. Two, they might seek to probe the regions of "parameter space" where various methods break down or where certain modeling assumptions are violated. Our simulation studies here were created with real-world fidelity top of mind, although the probing use was also important to us. We specifically considered the following when designing our DGPs:

- *How big are the treatment effects?* The magnitude of the ITE's – which we can examine in a simulation study – should be realistic with respect to the magnitude of the corresponding levels of $Y(1)$ and $Y(0)$. This is easy to inspect by plotting simulated values of $Y_i(1) - Y_i(0)$ or $E(Y(1) - Y(0) \mid \mathrm{x})$ against $Y_i(0)$ or $E(Y(1) \mid \mathrm{x})$, or by plotting histograms of their ratios.
- *How heterogeneous are the treatment effects?* Here, we can plot a histogram of $\tau_i = Y_i(1) - Y_i(0)$ and compare it to histograms of $Y_i(1)$ and $Y_i(0)$. In many "naive" simulations we observe that the variance of one potential outcome is dramatically larger than the other, meaning that the vast majority of variation in

the observed Y is due to treatment. This makes for an unrealistically easy causal problem, because even if confounding is present, unless it is incredibly strong, the bias of the omitted variables is a tiny fraction of most of the CATEs.

- *How hard is the estimation problem (even with no confounding)?* We also need to pay attention to error terms. It is very easy to accidentally generate additive error terms that are huge or miniscule relative to the ITEs. Neither situation will be informative about which methods are best since they are either "too hard" or "too easy" for minor methodological differences to matter. In all of our simulation protocols we set the scale parameter of our error terms as some percentage of the sample variance of $\mathrm{E}(Y)$. This is an important hedge against testing against signal to noise ratios either too close to zero or orders of magnitude larger than is realistic.

- *How strong is the confounding?* This problem is the most challenging, and trying to get control of it is part of what led us to describing targeted selection. If $\pi(\mathrm{x})$ is a deterministic monotone function of only $\mu(\mathrm{x})$ then we can parametrize strength of confounding in terms of, say, a logistic coefficient: $\pi(\mu) = 1/(1 + \exp(-\beta\mu)$. Whatever parametrization you choose, it is a good idea to monitor how difficult confounding is making the problem. This can be done by comparing inferences that include only sufficient predictors $\pi(\mathrm{x})$ and $\tau(\mathrm{x})$ (which you know as lord of the simulation) versus inferences based on the raw controls/moderators x. And it is always important to check the range of the true propensity score for overlap; an easy way to deal with a misbehaving selection model is to rescale the true propensity scores to $(\epsilon, 1 - \epsilon)$.

Turning to some of the DGPs presented by our discussants:

- Ray et al.'s simulation protocol provides a nice case study of how we interrogate DGPs. In Figure 1 we see: Huge treatment effects and massive treatment effect heterogeneity – including positive and negative treatment effects. The level of noise is very low relative to the variability of the treatment effects and outcomes. The selection is extremely strong, to the point of very likely violating overlap (and almost surely yielding practical violations of overlap when $n = 250$ or even 500). This DGP also features two incredibly strong instruments: X_4 and X_6 account for the majority of the extreme variation in the propensity scores, but don't appear in the outcome model and are not even necessary for deconfounding; see Figure 2.

- Bystrova, Arbel and Rahier report on a small simulation study that compares the bias of BART and BCF in a data generating process parametrized by strength of confounding. They show that both models exhibit some bias as confounding increases, but that BCF increases at a substantially slower rate than BART. Their finding is consistent with our own experience and we wish we had included this plot in our original paper!

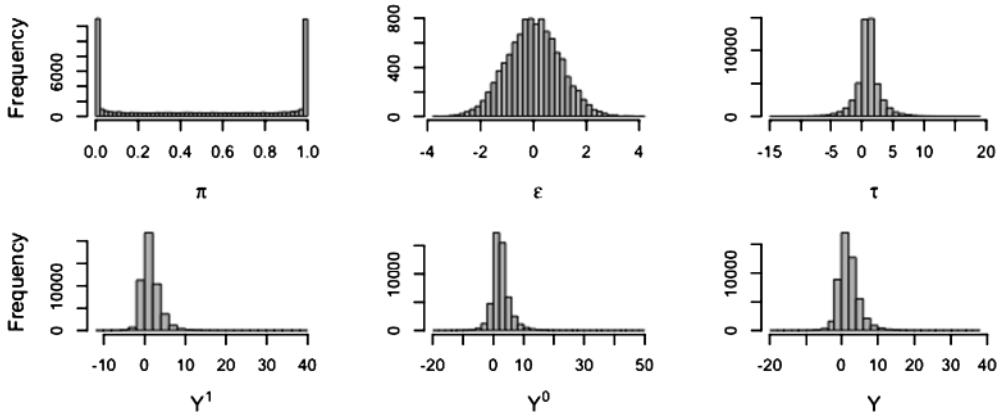


Figure 1: Diagnostic plots from the RSV DGP reveal excessively large treatment effects, tremendous treatment effect heterogeneity (including sign differences), relatively puny noise terms, and possible problems with overlap due to a distribution of propensity scores concentrated at the boundaries of the interval.

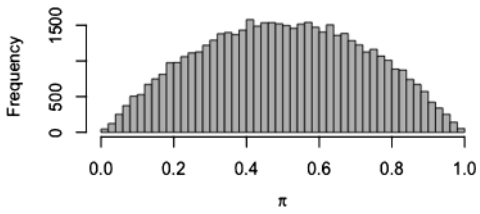


Figure 2: Histogram of the propensity scores from their DGP after removing the two instruments. Compare to the upper left panel of Figure 1.

- Prado et al. made the curious choice to omit noise in the outcome model entirely; the only randomness in their simulations is randomly generating X and Z, after which Y is a deterministic function. We are surprised that they were able to get the Bayesian tree models to mix, as these algorithms are notoriously slow mixing when σ is small (Pratola, 2016).

5.2 Are we comparing priors, models, methods, implementations, or defaults?

Once we've settled on a set of DGPs, as good experimentalists we need to decide on the factor or factors we want to vary in our experiments and do our best to control everything else. This is harder than it sounds! There are several plausible factors we might want to vary, and while the lines between each item below are not so clear-cut, we think it's useful to distinguish between a few different comparisons we might be interested in:

- Comparing priors or models. For example, assuming a common likelihood, we might compare different tree priors for τ in the BCF framework, or compare GP versus BART priors for τ. To the extent that other components of the prior/model are less interesting (e.g. the standard independent prior on σ^2, or how the propensity score is estimated in methods that incorporate it) we should try to fix them across methods.

- Comparing methods. Here we use "method" as shorthand for a procedure that turns data into estimates. For Bayesian methods, this includes the computational procedure for approximating or sampling from the posterior distribution, but wouldn't include details that depend on the specific dataset at hand like the necessary length of an MCMC chain to obtain appropriate mixing.

- Comparing implementations (of methods). The implementation of a method might include e.g. the number and length of chains in an MCMC algorithm, or how covariates are transformed.

- Comparing default implementations. Downloading someone's software, and running it using the author's defaults.

Any of these comparisons may be useful and informative if they are executed well. However, we should always be clear about what we are comparing. For example, Ray et al. compare three different methods that incorporate an estimated propensity score in one fashion or another. But they use a different propensity score estimate for each method, and all are misspecified to one degree or another. Further, only their own approach uses a truncated logistic regression (presumably to tame extreme estimates given the true propensity scores in Figure 1). This is a missed opportunity, as that would be an easy factor to fix across methods – and using a flexible estimator for the propensity score, or at least one that isn't misspecified, would shed more light on the value of including a PS estimate. As it stands, in their simulation, variability across methods is aliased with variability due to how the propensity score is estimated.

As another example, Prado et al. compare several "ps-models" with BCF as p gets large relative to n. They note a numerical error in BCF for $p > n$. The reason for this is that we adopted Chipman et al. (2010)'s method for calibrating the prior on σ, which incorporates a guess at σ. By default this is an estimate of the residual standard deviation from a least squares fit, which obviously fails when $p > n$. Many other methods (e.g. ps-BART) use the same prior, but for those methods the authors chose to replace this guess with the marginal standard deviation of y (for both $p \leq n$ and $p > n$). This is a dramatically different prior, especially as p increases, and the prior on the error variance can be rather consequential in practice. Again, this is a missed opportunity – the BCF R package also allows for an alternate guess at σ – since variability between BCF and other methods in their experiment is aliased with variability across choices for $p(\sigma^2)$.[4]

[4]Their choice to set the true $\sigma = 0$ makes the prior especially consequential – all the priors are bounded away from the true value, but priors that encourage larger values of σ may actually have an advantage since MCMC for Bayesian tree models is notoriously unreliable when σ is low (Pratola, 2016).

Finally, several of the discussants proposed Gaussian process models with limited discussion of the covariance function and how its parameters are set or inferred. The covariance function is often pivotal to their success. Unsurprisingly, the squared exponential covariance function performs splendidly on very smooth response surfaces, but what happens when this strong assumption is violated? By contrast, BART has a long track record of adapting successfully to a wide variety of unknown covariance structures and this robustness is why we chose to design BCF around BART priors.[5]

We don't mean to be hypercritical of our discussants, or to discourage empirical comparisons of any kind. Designing these comparisons is incredibly difficult! But the payoff is great – simulation studies can greatly augment our theoretical understanding, so long as we hold our investigations to high standards and design them with intent. We should strive to articulate what questions a simulation study means to address, and provide an argument for why the specified DGP(s) will help answer them. Finally, we note that the ACIC data challenges are uniquely valuable in the degree of rigor they afford. On the one hand, they provide comparisons of implementations that have been constructed by experts in the methods who have no control over the data generating processes. On the other hand, each year new data generating processes can be designed to probe the methods in specific ways. We feel that such open challenges should be a routine part of methods development in the future, in causal inference and elsewhere (see Heaton et al. (2019) for a recent high-profile example in spatial statistics).

5.3 Estimands and loss functions

We also need to choose target estimands and loss functions. In our paper, we only considered sample ATE/ATT or CATEs at realized covariate values as estimands. We focused on SATE/SATT since estimating the population ATE/ATT requires some understanding of how the sample relates to the population of interest, which is of secondary interest (and largely absent in many observational studies). Like most other empirical evaluations of methods for estimating heterogeneous treatment effects (e.g. the ACIC competitions) we focused on CATEs at realized covariate values because this is already a rather ambitious undertaking, and considering other covariate values risks adventuring into regions of poor overlap and again requires some notion of a broader population where we might want to estimate treatment effects (Prado et al. are curious about our choice to evaluate CATEs over the "training" data, but since we are estimating treatment effects – not making predictions – there is no risk of overfitting here).

The choice of estimand should be fixed for all methods. Ray et al. compared estimators and intervals targeting different estimands, with BCF and BART estimates targeting the SATE and GP-PS targeting PATE. It would have been easy to use the same Bayesian bootstrap estimate for $p(x)$ to adjust BART/BCF intervals to target

[5]Although not widely appreciated, BART actually *is* a Gaussian process, conditional on the trees (integrating over Gaussian priors over the leaf parameters). Specifically, the trees define a covariance function where the correlation between points x and x' are a function of the proportion of trees in the forest in which the two points occupy the same leaf. As the number of trees is increased, this covariance function becomes increasingly smooth, although it is singular and nonstationary for a finite number of trees.

the PATE, and failing to do so means that we are comparing apples-and-oranges when assessing coverage.

In the paper, we reported standard loss functions and performance metrics: RMSE/PEHE, coverage, interval length. In retrospect, we wish we would have also reported more tailored loss functions. For example, when estimating CATEs we might attach lower losses to errors in the direction of the ATE, representing a desire for conservative inferences about treatment effect heterogeneity.

6 Conclusion and acknowledgements

We thank the editorial board of Bayesian Analysis for the opportunity to present our work as a discussion paper, and to all the participants in this discussion for devoting their valuable time to considering the ideas in our paper and offering their own. We hope that applied data analysts in a variety of fields will continue to embrace BCF as a way to extract causal insights from their observational data, and are excited to see how the statistics and machine learning communities build on these ideas in the future.

References

Bail, C. A., Guay, B., Maloney, E., Combs, A., Hillygus, D. S., Merhout, F., Freelon, D., and Volfovsky, A. (2020). "Assessing the Russian Internet Research Agency's impact on the political attitudes and behaviors of American Twitter users in late 2017." *Proceedings of the National Academy of Sciences of the United States of America*, 117(1): 243–250. 1038

Bryan, C. J., Yeager, D. S., and O'Brien, J. M. (2019). "Replicator degrees of freedom allow publication of misleading failures to replicate." *Proceedings of the National Academy of Sciences*, 116(51): 25535–25545. URL https://www.pnas.org/content/116/51/25535. 1038, 1039

Carnegie, N. (2019). "Contributions of Model Features to BART Causal Inference Performance Using ACIC 2016 Competition Data." *Statistical Science*, 34(1): 90–93. MR3938969. doi: https://doi.org/10.1214/18-STS682. 1039

Carvalho, C., Feller, A., Murray, J. S., Woody, S., and Yeager, D. (2019). "Assessing Treatment Effect Variation in Observational Studies: Results from a Data Challenge." *Observational Studies*, 5: 21–35. URL https://obsstudies.org/277-2/. 1048

Chipman, H. A., George, E. I., and McCulloch, R. E. (2010). "BART: Bayesian additive regression trees." *The Annals of Applied Statistics*, 266–298. MR2758172. doi: https://doi.org/10.1214/09-AOAS285. 1051

Dorie, V., Hill, J., Shalit, U., Scott, M., Cervone, D., et al. (2019). "Automated versus do-it-yourself methods for causal inference: Lessons learned from a data analysis competition." *Statistical Science*, 34(1): 43–68. MR3938963. doi: https://doi.org/10.1214/18-STS667. 1042, 1043

Dweck, C. and Yeager, D. (2019). "A Simple Re-Analysis Overturns a "Failure to Replicate" and Highlights an Opportunity to Improve Scientific Practice: Commentary on Li and Bates." URL https://www.researchgate.net/publication/337856605. 1039

Franks, A., D'Amour, A., and Feller, A. (2019). "Flexible Sensitivity Analysis for Observational Studies Without Observable Implications." *Journal of the American Statistical Association*, 1–33. 1047

George, E., Laud, P., Logan, B., McCulloch, R., and Sparapani, R. (2019). "Fully Nonparametric Bayesian Additive Regression Trees." In *Topics in Identification, Limited Dependent Variables, Partial Observability, Experimentation, and Flexible Modeling: Part B*, volume 40B of *Advances in Econometrics*, 89–110. Emerald Publishing Limited. 1047

Ghosh, A., Orzol, S., Dale, S., Laird, J., Fu, N., Singh, P., Kim, M.-Y., Markovitz, A., Swankoski, K., Duda, N., Machta, R., Urato, C., and, M. F. (2020). "Independent Evaluation of Comprehensive Primary Care Plus (CPC+) Second Annual Report: Appendices to the Supplemental Volume." 1038

Hahn, P. R. (2019). "An illustration of the risk of borrowing information via a shared likelihood." *arXiv preprint 1905.09715*. 1041

Hahn, P. R. and Carvalho, C. M. (2015). "Decoupling shrinkage and selection in Bayesian linear models: a posterior summary perspective." *Journal of the American Statistical Association*, 110(509): 435–448. MR3338514. doi: https://doi.org/10.1080/01621459.2014.993077. 1039

Hahn, P. R., Carvalho, C. M., and Mukherjee, S. (2013). "Partial Factor Modeling: Predictor-Dependent Shrinkage for Linear Regression." *Journal of the American Statistical Association*, 108(503): 999–1008. MR3174679. doi: https://doi.org/10.1080/01621459.2013.779843. 1041

Hahn, P. R., Murray, J. S., and Manolopoulou, I. (2016a). "A Bayesian Partial Identification Approach to Inferring the Prevalence of Accounting Misconduct." *Journal of the American Statistical Association*, 111(513): 14–26. MR3494635. doi: https://doi.org/10.1080/01621459.2015.1084307. 1047

Hahn, P. R., Puelz, D., He, J., and Carvalho, C. M. (2016b). "Regularization and confounding in linear regression for treatment effect estimation." *Bayesian Analysis*. MR3737947. doi: https://doi.org/10.1214/16-BA1044. 1044

He, J. (2019). "Stochastic tree ensembles for regularized supervised learning." Technical report, University of Chicago Booth School of Business. 1039

He, J., Yalov, S., and Hahn, P. R. (2019). "XBART: Accelerated Bayesian Additive Regression Trees." In *The 22nd International Conference on Artificial Intelligence and Statistics*, 1130–1138. 1039

Heaton, M. J., Datta, A., Finley, A. O., Furrer, R., Guinness, J., Guhaniyogi, R., Gerber, F., Gramacy, R. B., Hammerling, D., Katzfuss, M., Lindgren, F., Nychka, D. W., Sun,

F., and Zammit-Mangion, A. (2019). "A Case Study Competition Among Methods for Analyzing Large Spatial Data." *Journal of agricultural, biological, and environmental statistics*, 24(3): 398–425. 1052

Herren, A. and Hahn, P. R. (2020). "Semi-supervised learning and the question of true versus estimated propensity scores." *arXiv preprint arXiv:2009.06472*. 1042

Hill, J., Su, Y.-S., et al. (2013). "Assessing lack of common support in causal inference using Bayesian nonparametrics: Implications for evaluating the effect of breastfeeding on children's cognitive outcomes." *The Annals of Applied Statistics*, 7(3): 1386–1420. MR3127952. doi: https://doi.org/10.1214/13-AOAS630. 1047

King, C. R., Escallier, K. E., Ju, Y.-E. S., Lin, N., Palanca, B. J., McKinnon, S. L., and Avidan, M. S. (2019). "Obstructive sleep apnoea, positive airway pressure treatment and postoperative delirium: protocol for a retrospective observational study." *BMJ open*, 9(8): e026649. 1038

Levitt, S. D. and Dubner, S. J. (2010). *Freakonomics*, volume 61. Sperling & Kupfer editori. 1039

Linero, A. R. and Daniels, M. J. (2015). "A Flexible Bayesian Approach to Monotone Missing Data in Longitudinal Studies with Informative Dropout with Application to a Schizophrenia Clinical Trial." *Journal of the American Statistical Association*, 110(1): 45–55. MR3338485. doi: https://doi.org/10.1080/01621459.2014.969424. 1047

Linero, A. R. and Daniels, M. J. (2018). "Bayesian Approaches for Missing Not at Random Outcome Data: The Role of Identifying Restrictions." *Statist. Sci.*, 33(2): 198–213. MR3797710. doi: https://doi.org/10.1214/17-STS630. 1047

Linero, A. R. and Yang, Y. (2018). "Bayesian regression tree ensembles that adapt to smoothness and sparsity." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 80(5): 1087–1110. MR3874311. doi: https://doi.org/10.1111/rssb.12293. 1047

Murray, J. S. (2020). "Log-Linear Bayesian Additive Regression Trees for Multinomial Logistic and Count Regression Models." *Journal of the American Statistical Association*, 0(ja): 1–35. doi: https://doi.org/10.1080/01621459.2020.1813587. 1047

Nethery, R. C., Mealli, F., and Dominici, F. (2019). "Estimating population average causal effects in the presence of non-overlap: The effect of natural gas compressor station exposure on cancer mortality." *Ann. Appl. Stat.*, 13(2): 1242–1267. MR3963570. doi: https://doi.org/10.1214/18-AOAS1231. 1047

Pratola, M. T. (2016). "Efficient Metropolis-Hastings Proposal Mechanisms for Bayesian Regression Tree Models." *Bayesian Anal.*, 11(3): 885–911. MR3543912. doi: https://doi.org/10.1214/16-BA999. 1050, 1051

Pratola, M. T., Chipman, H. A., George, E. I., and McCulloch, R. E. (2020). "Heteroscedastic BART via Multiplicative Regression Trees." *Journal of computational and graphical statistics: a joint publication of American Statistical Association, Institute of Mathematical Statistics, Interface Foundation of North America*, 29(2): 405–417. MR4116052. doi: https://doi.org/10.1080/10618600.2019.1677243. 1047

Robins, J. M., Hernán, M. A., and Wasserman, L. (2015). "Discussion of 'On Bayesian estimation of marginal structural models'." *Biometrics*, 71(2): 296–299. MR3366233. doi: https://doi.org/10.1111/biom.12273. 1042

Starling, J. E., Murray, J. S., Lohr, P. A., Aiken, A. R. A., Carvalho, C. M., and Scott, J. G. (2019). "Targeted Smooth Bayesian Causal Forests: An analysis of heterogeneous treatment effects for simultaneous versus interval medical abortion regimens over gestation." *arXiv preprint 1905.09405*. 1039

Woody, S., Carvalho, C. M., Hahn, P. R., and Murray, J. S. (2020a). "Estimating heterogeneous effects of continuous exposures using Bayesian tree ensembles: revisiting the impact of abortion rates on crime." *arXiv preprint 2007.09845*. 1039

Woody, S., Carvalho, C. M., and Murray, J. S. (2020b). "Model interpretation through lower-dimensional posterior summarization." *Journal of Computational and Graphical Statistics*, 0(ja): 1–34. doi: https://doi.org/10.1080/10618600.2020.1796684. 1039, 1048

Yeager, D. S., Hanselman, P., Walton, G. M., Murray, J. S., Crosnoe, R., Muller, C., Tipton, E., Schneider, B., Hulleman, C. S., Hinojosa, C. P., Paunesku, D., Romero, C., Flint, K., Roberts, A., Trott, J., Iachan, R., Buontempo, J., Yang, S. M., Carvalho, C. M., Hahn, P. R., Gopalan, M., Mhatre, P., Ferguson, R., Duckworth, A. L., and Dweck, C. S. (2019). "A national experiment reveals where a growth mindset improves achievement." *Nature*, 573(7774): 364–369. 1038, 1039