ANNUAL REVIEWS

Annual Review of Statistics and Its Application

Bayesian Additive Regression Trees: A Review and Look Forward

Jennifer Hill,[1] Antonio Linero,[2] and Jared Murray[3]

[1]Department of Applied Statistics, Social Science, and Humanities, New York University, New York, NY 10003, USA; email: jennifer.hill@nyu.edu

[2]Department of Statistics and Data Sciences, University of Texas, Austin, Texas 78712, USA

[3]Department of Information, Risk, and Operations Management, McCombs School of Business, University of Texas, Austin, Texas 78712, USA

Annu. Rev. Stat. Appl. 2020. 7:251–78

First published as a Review in Advance on October 3, 2019

The *Annual Review of Statistics and Its Application* is online at statistics.annualreviews.org

https://doi.org/10.1146/annurev-statistics-031219-041110

Keywords

regression, machine learning, Bayesian nonparametrics, causal inference, regularization

Abstract

Bayesian additive regression trees (BART) provides a flexible approach to fitting a variety of regression models while avoiding strong parametric assumptions. The sum-of-trees model is embedded in a Bayesian inferential framework to support uncertainty quantification and provide a principled approach to regularization through prior specification. This article presents the basic approach and discusses further development of the original algorithm that supports a variety of data structures and assumptions. We describe augmentations of the prior specification to accommodate higher dimensional data and smoother functions. Recent theoretical developments provide justifications for the performance observed in simulations and other settings. Use of BART in causal inference provides an additional avenue for extensions and applications. We discuss software options as well as challenges and future directions.

1. INTRODUCTION

The conditional expectation, or regression, model is one of the most fundamental inferential and predictive models in statistics. The regression of response Y on a vector of predictor variables, $\mathbf{X}$, expressed formally as $\mathrm{E}[Y \mid \mathbf{X}]$, allows us to understand how the mean of that response variable varies with levels of the predictors. The most popular regression model is the linear regression, which assumes a linear additive structure to relate the predictors to the response, as in $\mathrm{E}[Y \mid \mathbf{X}] = \mathbf{X}^\top \beta$, as well as an additive error term such as $Y = \mathrm{E}[Y \mid \mathbf{X}] + \epsilon$, with $\epsilon = N(0, \sigma^2)$. This incarnation of the regression model is computationally straightforward but makes restrictive parametric assumptions regarding linearity and additivity.

Bayesian additive regression trees (BART), an approach introduced by Chipman et al. (2007, 2010), provides an alternative to some of these stringent parametric assumptions. It combines the flexibility of a machine learning algorithm with the formality of likelihood-based inference to create a powerful inferential tool. This article motivates and describes the BART approach to regression modeling and the computation required for posterior inference. We discuss extensions to the original BART formulation to new models for a variety of data types and modifications to the BART prior to accommodate high dimensions and smooth regression functions. Recent theoretical results about BART models are described next. Finally, we discuss at length the application of BART-based methods to causal inference problems. We conclude with a description of available software as well as some challenges and future directions.

2. BAYESIAN ADDITIVE REGRESSION TREES

This section motivates and describes the BART framework.

2.1. Model

For a p-dimensional vector of predictors $\mathbf{x}_i$ and a response Y_i $(1 \leq i \leq n)$ the BART model posits

$$Y_i = f(\mathbf{x}_i) + \epsilon_i, \quad \epsilon_i \overset{\text{iid}}{\sim} N(0, \sigma^2), \tag{1.}$$

where f is represented as the sum of many regression trees. We first briefly describe regression trees. Throughout, we use capital letters to refer to the random variables and lowercase letters to refer to realizations of those variables.

2.1.1. Regression tree. The building block of BART is the regression tree. A regression tree creates a partition of the covariate space into subgroups; the tree fit will be the same for each observation in a given subgroup. Consider the example tree in **Figure 1**, which uses two predictors to split the data into subgroups in the bth tree of an ensemble. Panel *a* displays a set of splitting rules of the form $x_{ij} < C$ attached to interior decisions nodes (boxes) of the tree. Observations are assigned to subgroups by dropping them down the tree, sending them left or right according to the decision rule at each interior node. The bottom circles represent the leaves (also known as terminal nodes) of the tree, one for each subgroup, and each with an associated parameter that is the regression tree's prediction. In traditional applications of regression trees this would simply be the average of the observations in each subgroup, so we refer to these parameters loosely as means.

Figure 1*b* displays the prediction function corresponding to the tree in the panel *a*—in general, a tree and parameter pair (T_b, M_b) parameterizes a step function g that is constant over elements

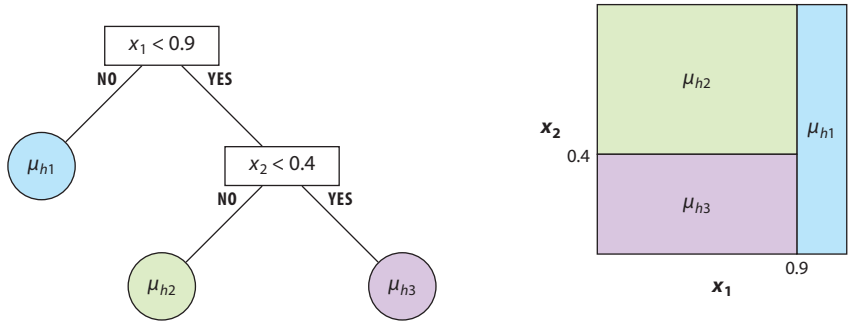

Figure 1

(*a*) An example binary tree, with internal nodes labeled by their splitting rules and leaf nodes labeled with the corresponding parameters μ_{bt}. (*b*) The corresponding partition of the sample space and the step function $g(\mathbf{x}, T_b, M_b)$.

of the partition:

$$g(\mathbf{x}, T_b, M_b) = \mu_{bt} \text{ if } \mathbf{x} \in \mathcal{A}_{bt} \text{ (for } 1 \leq t \leq b_b). \qquad 2.$$

Here $M_b = (\mu_{b1}, \mu_{b2}, \ldots \mu_{bb_b})'$ denotes the collection of parameters for the b_b leaves of the bth tree, T_b.

Traditional applications of regression trees grow the tree greedily to minimize some loss function. BART takes a different approach based on averaging the output of many small trees, similar to boosting.

2.1.2. Sum-of-trees model. The original formulation of the BART sum-of-trees model by Chipman et al. (2007, 2010) starts with an overall fit, f, defined as the sum of the fit of many trees:

$$f(\mathbf{x}) = g(\mathbf{x}, T_1, M_1) + g(\mathbf{x}, T_2, M_2) + \cdots + g(\mathbf{x}, T_m, M_m).$$

Here each (T_b, M_b) corresponds to a single subtree model. Finally, the data are assumed to arise from a model with additive Gaussian errors: $Y = f(\mathbf{x}) + \epsilon$ where $\epsilon \sim N(0, \sigma^2)$.

To avoid overfitting, the trees are encouraged to be small and the parameters are shrunk toward zero, so that each tree-parameter pair constitutes a weak learner. We motivate the intuition behind this approach by a thought experiment. Imagine fitting the first weak-learning tree, $g(\mathbf{x}; T_1, M_1)$, in some reasonable fashion. Since T_1 is constrained to be small, and the elements of M_1 are shrunk toward zero, this first tree explains a small amount of the variation in y but leaves much unexplained. Next, imagine subtracting this fit from the observed response Y to form residuals and fitting the next tree to these residuals. This next tree would explain more of the variation in Y. This process could be performed a total of m times, and each time more of the variation would be explained. We could then consider revisiting earlier trees and updating their fit using the partial residuals based on later trees, iterating to a stable value of some loss function. In the end, we would compose many simple functions like **Figure 1** to obtain a very complicated regression function.

The challenge with this basic approach is avoiding overfitting. For BART, the solution to this problem lies in the prior distribution (as described in the next subsection) and in posterior sampling and averaging rather than loss minimization (as described in Section 2.3).

2.2. Regularization Prior

Similar to boosting, BART allows the number of subtree models, m, to be large. Boosting algorithms can avoid overfitting through tuning parameters. First, the researcher chooses a maximum depth for each tree to constrain each to be a weak learner. Second, the fit from each tree is typically multiplied by a small number to shrink the sum of the fit across trees toward zero. This tuning parameter is generally chosen via cross-validation.

BART provides a related but alternative strategy to avoid overfitting that lets the data speak more naturally. A regularization prior constrains the size and fit of each (T_b, M_b) tree so that each tree contributes only a small part to the overall fit. First we factor the joint distribution of the trees, the means in the leaves, and the residual standard deviation:

$$\begin{aligned} &p((T_1, M_1), (T_2, M_2), \ldots, (T_m, M_m), \sigma) \\ &\quad = p(T_1, T_2, \ldots, T_m) p(M_1, M2, \ldots, M_m \mid T_1, T_2, \ldots, T_m) p(\sigma). \end{aligned}$$

Then these distributions are simplified by a series of independence assumptions:

$$\begin{aligned} p(T_1, T_2, \ldots, T_m) &= \Pi_b p(T_b), \\ p(M_1, M2, \ldots, M_m \mid T_1, T_2, \ldots, T_m) &= \Pi_b p(M_b \mid T_b), \\ p(M_b \mid T_b) &= \Pi_t p(\mu_{bt} \mid T_b). \end{aligned}$$

The prior on the trees strongly favors weak learners: trees that are small with leaf parameters that are close to zero. Each tree is assigned an independent prior (as in Chipman et al. 1998). The probability that a node at depth d splits (is not terminal) is given by

$$\alpha(1 + d)^{-\beta}, \quad \alpha \in (0, 1), \ \beta \in [0, \infty). \tag{3.}$$

The default prior specification proposed by Chipman et al. (2007, 2010) is $\alpha = 0.95$ and $\beta = 2$. This specification strongly favors small trees; for example, it sets the probability of a tree with 1, 2, 3, 4, and over 5 terminal nodes at 0.05, 0.55, 0.28, 0.09, and 0.03, respectively.

In the default specification, the distribution on the splitting variable in each nonterminal node is uniform over the available choices. Given a variable to use in the split rule, the prior distribution on the location for the split is uniform over the range of the selected covariate (or a uniform over a grid of quantiles). For binary or ordinal variables, the cut-points can be defined by the collection of all possible values. Unordered categorical variables with q levels are generally expanded into q binary indicators of these levels.

To further avoid overfitting, Chipman et al. (2007, 2010) carefully calibrate the priors over the leaf parameters μ_{bt}. They begin by scaling the response variable to lie between -0.5 and 0.5, so that we can expect $f(\mathbf{x})$ to lie within this range with high probability. Assume that the prior distribution for each mean follows a normal distribution

$$\mu_{bt} \overset{\text{iid}}{\sim} N(0, \sigma_\mu^2) \quad \text{where } \sigma_\mu = 0.5/(k\sqrt{m}) \tag{4.}$$

(recall that m is the number of trees). Each tree will contribute one μ_{bt} to the overall fit, so the regression function f at any covariate value $\mathbf{x}$ is normally distributed with mean zero and standard deviation equal to $0.5/k$. Then k is the number of prior standard deviations between 0 and 0.5—for instance, if we use the default prior parameterization of Chipman et al. (2007, 2010) of $k = 2$, we assign a 95% prior probability that $\mathrm{E}[Y \mid \mathbf{x}]$ lies between -0.5 and 0.5.

The prior on σ^2 from Equation 1 is specified as an inverse-χ^2 and is also calibrated to the observed data. The choice of the degrees of freedom for this distribution can be motivated by the intuition that if the true model for $\mathrm{E}[Y \mid \mathbf{x}]$ is not linear and additive, a linear regression fit to the data will likely overstate the estimate of the residual standard deviation. Therefore the degrees of freedom for the prior can be specified to represent the probability that the BART residual standard deviation, σ, is less than the estimated residual standard deviation from a linear regression fit to the data, $\hat{\sigma}_{\mathrm{OLS}}$. The default prior sets the degrees of freedom such that this probability, $\Pr(\sigma < \hat{\sigma}_{\mathrm{OLS}})$, equals 0.9. [Chipman et al. (2007, 2010) evaluated this choice in a cross-validation exercise and found it to perform well in practice.]

2.3. Computation

Posterior inference in BART models is typically carried out via Markov chain Monte Carlo (MCMC) by adapting tools developed for single-tree models and embedding them into Metropolis-within-Gibbs samplers.

2.3.1. Markov chain Monte Carlo for single-tree models. Chipman et al. (1998), Denison et al. (1998), and Wu et al. (2007) discuss posterior sampling for single-tree classification and regression models. These samplers take a similar blocked Metropolis-Hastings approach to posterior exploration. Roughly, the tree and its parameters are sampled by composition—first drawing from the marginal posterior distribution of the tree T (marginalizing out the leaf parameters) and then sampling the leaf parameters from their full conditional distribution.

We now describe the simplest version of the sampler. To update the current value of the tree, first propose a new tree T^* by sampling from some proposal distribution $q(T^*;\ T)$ (which depends on the current tree T). Compute the Metropolis ratio

$$a := \frac{L(T^*)p(T^*)}{L(T)p(T)} \frac{q(T;\ T^*)}{q(T^*;\ T)}, \qquad 5.$$

where $L(T) = \int \prod_i p(Y_i \mid \mathbf{x}_i, T, M)\, p(M \mid T)\, \mathrm{d}M$ is the marginal likelihood of tree T. The proposed tree is accepted—that is, T is set to T^*—with probability $\min(a, 1)$. Otherwise T remains at its current value. Finally, the parameters M are updated by sampling them from their full conditional distribution $p(M \mid T, \mathcal{D})$ which is typically available in closed form; here, $\mathcal{D} = \{(Y_i, \mathbf{x}_i) : i = 1, \ldots, n\}$ denotes the data. This blocked update avoids delicate issues related to dimension changes in M and also increases the overall efficiency of the algorithm. This basic sampler is readily embedded into larger Metropolis-within-Gibbs MCMC algorithms for more complicated models with additional parameters (which may be conditioned upon above and subsequently updated via additional Gibbs/Metropolis–Hastings steps).

Following Chipman et al. (2007, 2010), most BART implementations use proposal distributions that randomly perturb the current tree by splitting a current leaf into two new leaves (grow), collapsing adjacent leaves back into a single leaf (prune), reassigning the decision rule attached to an interior node (change), or swapping the decision rules assigned to two interior nodes (swap). Pratola et al. (2014) noted that the grow and prune moves are particularly computationally efficient and often yield acceptable MCMC mixing for estimates of the regression function. Additional proposal distributions for MCMC with trees are presented by Wu et al. (2007) and Pratola (2016).

2.3.2. Bayesian backfitting and Markov chain Monte Carlo for BART models. Chipman & McCulloch (2010) introduced an efficient MCMC algorithm for fitting the model in Equation 1.

Each (T_h, M_h) pair is updated in turn, conditioning on σ and the remaining trees and their associated parameters. Formally, their approach constitutes a proper partially collapsed Metropolis-within-Gibbs MCMC sampler, but it is easier to understand as an instance of Bayesian backfitting (Hastie & Tibshirani 2000), as it was originally presented by Chipman et al. (2007, 2010).

Each MCMC update of the pair (T_h, M_h) begins with a clever reparameterization of the response. Simple algebraic manipulations reveal that the likelihood function only depends on (T_h, M_h) through the partial residuals

$$R_{hi} = Y_i - \sum_{l \neq h}^{m} g(\mathbf{x}_i, T_l, M_l), \quad i = 1, \ldots, n. \tag{6.}$$

Since the MCMC update for (T_h, M_h) conditions on all the remaining trees and associated parameters, the model can be (temporarily) reparameterized in terms of these partial residuals. Under the model in Equation 1,

$$R_{hi} \sim N(g(\mathbf{x}_i, T_h, M_h), \sigma^2). \tag{7.}$$

So, to update (T_h, M_h), we can immediately adopt any of the single-tree MCMC updates, treating the partial residuals as the data.

The backfitting analogy breaks down in more complicated models, including many of those in Section 3. However, we can still maintain the efficiency gains of block updates for (T_h, M_h) in these models, provided we can compute the integrated likelihood function

$$L(T_h; T_{(h)}, M_{(h)}, \theta) = \int \left(\prod_{i=1}^{n} p(Y_i \mid T_h, M_h, T_{(h)}, M_{(h)}, \theta) \right) p(M_h \mid T_h, \theta) \mathrm{d}M_h,$$

where θ is a vector of other parameters (such as σ above). Here, $T_{(h)} \equiv \{T_l : 1 \leq l \leq m, l \neq h\}$ is the set of all the trees except T_h, and $M_{(h)}$ is defined similarly. Performing this integral in closed form essentially requires that the prior distribution for M_h be conditionally conjugate to the likelihood function, but there is no need to derive some sort of partial residual with a convenient distribution under the model. Algorithm 1 summarizes one MCMC update of (T_h, M_h) when this integral is available; the MCMC sampler described by Chipman et al. (2007, 2010) is a special case.

Algorithm 1 (One step of a generalized Bayesian backfitting algorithm for updating a single BART function parameterized by $T = \{T_h\}$ and $M = \{M_h\}$ $(1 \leq h \leq m)$).

Input: Data $\mathcal{D}$ and current values for T, M, and other parameters/latent variables (in θ)
Output: New values of T, M
for $1 \leq h \leq m$

1. Propose $T_h^* \sim q(T_h^*;\ T_h)$.
2. Set $a \leftarrow \frac{L(T_h^*;\ T_{(h)}, M_{(h)}, \theta) p(T_h^*)}{L(T_h;\ T_{(h)}, M_{(h)}, \theta) p(T_h)} \frac{q(T_h; T_h^*)}{q(T_h^*; T_h)}$.
3. Set $T_h \leftarrow T_h^*$ with probability $\min(1, a)$.
4. Sample $M_h \sim p(M_h \mid T_h, T_{(h)}, M_{(h)}, \theta, \mathcal{D})$.

end for

Most BART implementations are based on the (generalized) Bayesian backfitting sampler using proposals similar to those described for single tree models. While this algorithm is often effective,

it does not always mix well, and recent work suggests that it can be important to run many chains (as many as 10 or 12) to encourage proper mixing (Carnegie 2019). To further improve mixing we can take advantage of the small trees encouraged by the BART prior. With small trees it becomes feasible to propose entirely new trees rather than perturbing existing trees in the ensemble. Lakshminarayanan et al. (2015) proposed a particle Gibbs sampler using sequential Monte Carlo to generate a proposal. He et al. (2019) introduced a proposal distribution in which the candidate tree is built recursively, stochastically choosing variables to split on and cut-points to split at using integrated likelihoods at each step. This biases the sampler toward more promising tree structures. In both cases these proposal distributions exhibit better mixing when the regression function is complex or the covariate dimension is large.

2.4. Early Empirical Evidence

Chipman et al. (2007) compared the performance of BART relative to several machine learning competitors that were commonly used at that time in the context of 42 different real data sets [these are a subset of those used by Kim et al. (2007), as explained by Chipman et al. (2007, 2010)]. Specifically they created 840 test/training splits in these data sets and compared the performance of boosting, neural nets, and random forests relative to BART with the default parameterization of the prior and BART with hyperparameters chosen by cross-validation. This exercise demonstrated similar performance between BART and the other popular machine learning algorithms when the default prior specification was used for BART (for the other methods, the tuning parameters were chosen via cross-validation). However, BART performance was noticeably better than the rest when cross-validation was used to choose the BART hyperparameters for the regularization prior. We discuss additional evidence of performance below in the context of different inferential goals or extensions of the original model.

3. DEVELOPMENTS IN BART MODELING

In addition to continuous data with Gaussian errors, Chipman et al. (2007, 2010) extended the BART model in Equation 1 to binary classification by taking

$$\Pr(Y_i = 1 \mid \mathbf{x}_i) = \Phi(f(\mathbf{x}_i)),$$

where Φ is the standard normal cumulative distribution function (suppressing an optional fixed offset used to shrink probabilities to a value other than 0.5). Chipman et al. (2007, 2010) used the data augmentation of Albert & Chib (1993) to adapt their Bayesian backfitting sampler:

$$\begin{aligned} Y_i^* &= f(\mathbf{x}_i) + \epsilon_i, \quad \epsilon_i \overset{\text{iid}}{\sim} N(0,1), \\ Y_i &= \mathbf{1}(Y_i^* > 0). \end{aligned} \tag{8.}$$

The probit formulation for binary classification yields a simple MCMC algorithm: First, latent normal response variables Y_i^* are imputed from their truncated normal full conditional distributions. Conditional on these imputed values, the Bayesian backfitting MCMC algorithm can proceed as usual. Performance with this specification has sometimes been less impressive than with continuous responses (Carnegie et al. 2015) but can be improved by using cross-validation to choose hyperparameters rather than using the default prior specification (Dorie et al. 2019).

More recent papers have applied BART prior distributions to functions in a wide variety of different regression models for categorical, count, heteroskedastic, zero-inflated, multivariate, and right-censored survival responses. The following subsections discuss these options.

3.1. Models with Gamma-Poisson Likelihoods

Using the generalized Bayesian backfitting algorithm, functions with BART priors can be embedded in complicated non-Gaussian models. These models all have a common form (possibly after augmenting with Gamma-distributed latent variables) that is amenable to computing the integrated likelihood. Here, we describe the range of models that can be fit using this general approach.

3.1.1. Binary and multinomial logistic regression. Murray (2017) introduced a multinomial logistic model using BART. For a categorical response with $1 \leq j \leq c$ categories, suppose that the probability of observing category j at a given covariate level $\mathbf{x}_i$ is

$$\pi_j(\mathbf{x}_i) = \frac{\exp\left[\sum_{b=1}^{m} g(\mathbf{x}, T_b^{(j)}, M_b^{(j)})\right]}{\sum_{l=1}^{c} \exp\left[\sum_{b=1}^{m} g(\mathbf{x}, T_b^{(l)}, M_b^{(l)})\right]}. \tag{9.}$$

Here, $T^{(j)}$ and $M^{(j)}$ are (potentially) distinct sets of trees and parameters for each outcome category. This model immediately implies that each log-odds function has a BART prior with $2m$ trees: The log-odds in favor of j' over j are given by

$$\log\left(\frac{\pi_{j'}(\mathbf{x}_i)}{\pi_j(\mathbf{x}_i)}\right) = \left[\sum_{b=1}^{m} g(\mathbf{x}, T_b^{(j')}, M_b^{(j')})\right] - \left[\sum_{b=1}^{m} g(\mathbf{x}, T_b^{(j)}, M_b^{(j)})\right] \tag{10.}$$

for any $j \neq j'$. This is useful for prior calibration (discussed in Section 3.1.5).

As written, this model is unidentified. However, proper priors for the trees and parameters yield valid inference over identified quantities like probabilities in Equation 9 or log-odds functions in Equation 10. Working in the unidentified space avoids asymmetries in the prior arising from the arbitrary choice of the reference category; Murray (2017) provided further discussion in the context of BART models.

3.1.2. Poisson and negative binomial regression, with or without covariate-dependent zero inflation. Murray (2017) also described a range of count regression models in this class. For count responses, we begin with Poisson or negative binomial models with mean function $\mathrm{E}(Y_i \mid \mathbf{x}_i) = \mu_{0i} f(\mathbf{x}_i)$. Here μ_{0i} is a fixed offset such as an adjustment for unit-level exposure, or we may take $\mu_{i0} \equiv \mu_0$ to center the prior for the regression function at μ_0. We induce a log-linear model for the mean function by assuming

$$\log[f(\mathbf{x})] = \sum_{b=1}^{m} g(\mathbf{x}, T_b, M_b).$$

The Poisson model is completely specified by the mean function. The negative binomial regression model has an additional parameter κ, which controls the degree of overdispersion relative to

the Poisson. Under the negative binomial model,

$$\mathrm{Var}(Y_i \mid \mathbf{x}_i) = \mathrm{E}(Y_i \mid \mathbf{x}_i)\left(1 + \frac{\mathrm{E}(Y_i \mid \mathbf{x}_i)}{\kappa}\right).$$

As $\kappa \to \infty$, the negative binomial model converges to the Poisson. The probability mass function under the Poisson model is

$$p_{\mathrm{P}}(y \mid \mathbf{x}, f) = \frac{\exp[-\mu_{0i} f(\mathbf{x})][\mu_{0i} f(\mathbf{x})]^y}{y!}.$$

For the negative binomial model we have

$$p_{\mathrm{NB}}(y \mid \mathbf{x}, f, \kappa) = \frac{\Gamma(\kappa + y)}{\Gamma(\kappa) y!}\left(\frac{\kappa}{\kappa + \mu_{0i} f(\mathbf{x}_i)}\right)^{\kappa}\left(\frac{\mu_{0i} f(\mathbf{x}_i)}{\kappa + \mu_{0i} f(\mathbf{x}_i)}\right)^{y_i}.$$

Many data sets exhibit an excess of zero values. Zero-inflated variants of Poisson or negative binomial regression models accommodate the extra zeros by adding a point mass component (Lambert 1992, Greene 1994):

$$\Pr(Y_i = y \mid \mathbf{x}_i) = \begin{cases} (1 - \omega(\mathbf{x}_i)) + \omega(\mathbf{x}_i) p(y \mid \mathbf{x}_i, f, \kappa) & \text{if } y = 0 \\ \omega(\mathbf{x}_i) p(y \mid \mathbf{x}_i, f, \kappa) & \text{if } y > 0, \end{cases}$$

where $p(y \mid \mathbf{x}_i, f, \kappa)$ is the probability mass function of a Poisson or negative binomial with mean $\mu_{0i} f(\mathbf{x})$ and dispersion κ, and $1 - \omega(\mathbf{x}_i)$ is the probability that a zero is due to the point mass component. In log-linear BART variants of zero-inflated models, f is assigned a log-linear prior and $\omega(\mathbf{x}_i)$ is assigned a logistic BART prior similar to Equation 9 (Murray 2017).

3.1.3. Heteroscedastic regression for continuous responses. We can allow both the mean and variance functions to depend on covariates (maintaining Gaussian errors):

$$y_i = f(\mathbf{x}_i) + \sigma(\mathbf{x}_i)\epsilon_i, \quad \epsilon_i \overset{\text{iid}}{\sim} N(0, \sigma_0^2), \tag{11.}$$

where $\sigma^2(\cdot)$ is given a log-linear BART prior,

$$\log[\sigma^2(\mathbf{x})] = \sum_{b=1}^{m} g(\mathbf{x}, T_b^{\sigma}, M_b^{\sigma}).$$

This model is described in detail by Pratola et al. (2017) (see also Murray 2017, Linero et al. 2018). Note that this model implies a BART prior on the precision and standard deviation functions as well. Non-Gaussian but independent and identically distributed (iid) errors are considered in George et al. (2018).

3.1.4. Gamma/inverse gamma regression. Linero et al. (2018) described Gamma regression models for strictly positive data. For example, consider a gamma response model $Y_i \sim \mathrm{Gam}(\alpha, \alpha f(\mathbf{X}_i))$ with density

$$p(y \mid \mathbf{x}) = \frac{\alpha^{\alpha} f(\mathbf{x})^{\alpha}}{\Gamma(\alpha)} y^{\alpha - 1} e^{-\alpha f(\mathbf{x}) y}.$$

In this case, the mean is $E(Y_i \mid \mathbf{X}_i = \mathbf{x}) = f(\mathbf{x})^{-1}$, and $\log f(\mathbf{x}) = \sum_{b=1}^{m} g(\mathbf{x}, T_b, M_b)$ is given a BART prior. Linero et al. (2018) also considered hurdle versions of these models, where the response includes a spike at zero and a continuous component bounded away from zero, with the covariate-dependent probability of an observed zero modeled via probit BART.

3.1.5. The common Gamma-Poisson likelihood and prior specification for non-Gaussian models. All the models in this subsection share a common Gamma-Poisson likelihood form, possibly after augmentation with latent variables. Considering the likelihood as a function of a single regression function f at a time, we can write the density of $\mathbf{Y} = (Y_1, \ldots, Y_n)$ as

$$p(\mathbf{y} \mid \mathbf{x}, \theta) = \prod_{i=1}^{n} w_i f(\mathbf{x}_i)^{u_i} \exp[v_i f(\mathbf{x}_i)], \qquad 12.$$

where θ collects additional model parameters (like the overdispersion parameter κ in negative binomial models, or the variance offset σ_0^2 in the heteroskedastic regression model), latent variables from data augmentation (such as the Gamma latent variables introduced in the multinomial logistic and zero-inflated/negative binomial regression models), and other regression functions (such as the mean regression function in the heteroscedastic regression model, when we consider the likelihood in terms of the variance function). Here w_i, u_i, and v_i are some functions of θ and y_i that will vary depending on the model under consideration.

Many, but not all, models in the exponential family are amenable to this treatment. For example, the logistic regression model described in Section 3.1.1 does not have a likelihood of this form (even imposing identifying restrictions). Neither do the zero-inflated count regression models. However, Murray (2017) described simple data augmentation schemes based on untruncated latent Gamma random variables that cause the augmented likelihood to factor into terms of the form in Equation 12.

In order to analytically compute the necessary integrated likelihoods in the generalized Bayesian backfitting Algorithm 1, the prior on (exponentiated) leaf parameters $\lambda_{bt} = \exp(\mu_{bt})$ should be conjugate to a Gamma-Poisson likelihood of form similar to Equation 12 [see Murray (2017) for a detailed derivation]. Gamma priors on λ_{bt} are a natural choice but induce slightly asymmetric priors on μ_{bt}. Murray (2017) proposed an equal probability mixture of Gamma and inverse Gamma distributions to maintain conjugacy and yield a symmetric prior distribution on μ_{bt}. When the number of trees m is large, the difference is likely immaterial.

Whatever the prior on leaf parameters, sensibly calibrating it is important for obtaining good performance. A reasonable strategy is to choose parameters such that $\mathrm{E}(\mu_{bt}) = 0$ and $\mathrm{Var}(\mu_{bt}) = \sigma_\mu^2/m$, so that $\mathrm{E}(\log[f(\mathbf{x})]) = 0$ (which can then be shifted using offsets) and $\mathrm{Var}(\log[f(\mathbf{x})]) = \sigma_\mu^2$, so σ_μ can be calibrated using the data and/or prior expectations about the plausible range of the regression function as in the original BART model. This can be accomplished by moment matching; for the special case of Gamma priors, Murray (2017) showed that the correct parameters are well-approximated by a shape parameter of $m/\sigma_\mu^2 + 0.5$ and rate parameter of m/σ_μ^2.

3.2. Shared Versus Distinct Trees and Multivariate Leaf Parameters

An interesting issue arises in the models above that include multiple BART functions. For example, as described above, the multinomial regression model includes m trees for each outcome category.[1] When the number of categories is large or the outcomes are rare, pooling information across categories by sharing their tree structure can be beneficial. A similar issue arises in heteroscedastic

[1] Murray (2017) suggested using either 200 total trees or 100 trees per category by default.

regression models, where the variance function is just intrinsically more difficult to learn than the mean function. Linero et al. (2018) demonstrate the benefit of using a common set of trees in a hurdle and heteroscedastic regression model applied to data from the Medicare Expenditure Panel Survey.

In general, of course, it is entirely possible that the mean and variance function involve distinct sets of covariates or are structurally dissimilar. Murray (2017) gives an example where a substantively important interaction is present in the excess zero probability of a zero-inflated negative binomial regression model but absent in the mean function. A shared tree model may shrink away this effect, or induce noise in the simpler regression function, or require more trees to accommodate this structure. While the correct answer about whether to share trees is probably data set–specific, these trade-offs deserve further investigation.

We can construe shared tree models as regular BART priors with a vector of leaf parameters—that is, replacing μ_{bt} with a vector $(\mu_{bt1}, \mu_{bt2}, \ldots, \mu_{btp})$, where the jth regression function is parameterized by μ_{btj}. Including dependence between the elements of this vector provides another interesting avenue for shrinkage. A similar approach was taken by Starling et al. (2019), who replaced the scalar leaf parameters with functional parameters drawn from a Gaussian process prior.

3.3. Survival Models

The BART framework can also be applied to survival data for both semiparametric and nonparametric models. Bonato et al. (2010) use a very general latent-variable strategy for constructing BART models based on essentially any parametric model. For example, consider a generic hierarchical model

$$\begin{aligned} Y_i &\sim p(y \mid \omega_i), \\ \omega_i &\sim \mathrm{N}(f(\mathbf{x}_i), \sigma^2), \end{aligned}$$

where, as above, $f(\mathbf{x})$ is a sum of trees. Because the random effect ω_i separates the response Y_i from $f(\mathbf{x}_i)$, regardless of the parametric form of $p(y \mid \omega)$, the Bayesian backfitting algorithm based on the nonparametric regression model can be applied to the model $\omega_i = N(f(\mathbf{x}_i), \sigma^2)$. Bonato et al. (2010) consider a Weibull regression model

$$p(y \mid \omega, \tau) = \tau y^{\tau - 1} \exp\left\{\omega - \exp(\omega) y^{\tau}\right\}$$

and a log-normal model with $\log(Y_i) \sim N(\omega_i, \sigma^2)$. The random effects ω_i can then be sampled from their full conditionals using a Metropolis–Hastings update. This approach model can also be applied to the proportional hazards model

$$\lambda(y, \mathbf{x}) = \lambda_0(y) \exp(f(\mathbf{x}))$$

using a gamma process prior for the cumulative hazard function $\Lambda_0(y) = \int_0^y \lambda_0(s)\, \mathrm{d}s$. This is essentially a Bayesian version of Cox's proportional hazards model (Cox 1972, Ibrahim et al. 2005) in which the usual linear effect $\mathbf{x}^\top \beta$ is replaced with the sum of trees $f(\mathbf{x})$.

A fully nonparametric survival model is given by Sparapani et al. (2016) by assuming a discrete-time survival model in which the observed failure and censoring times $t_1, \ldots, t_n$ are assumed to account for the entire support of the response. They then introduce a sequence of binary indicators Z_{ij} such that $Z_{ij} = 1$ if $Y_i = t_j$, and $Z_{ij} = 0$ otherwise. This essentially converts the

survival problem into a series of binary regression problems: We model $Y_{ij} \sim \text{Bernoulli}(\Phi(f(t_j, \mathbf{x}_i)))$ and fit the model using the BART probit regression model (Equation 8).

4. DEVELOPMENTS IN THE BART PRIOR SPECIFICATION

The default Chipman et al. (2007, 2010) BART prior specification described in Section 2.2 has been found to perform well across a wide variety of settings. In some circumstances, however, the default prior specification either performs suboptimally or breaks down entirely. Fortunately, with only modest modifications, we can recover the performance of BART. We describe two settings in which modifications to the BART prior can be used to greatly improve the performance of BART. The first is under the high-dimensional, ultrasparse, needles-in-a-haystack regime. The second is the setting where the underlying regression function $f(\mathbf{x})$ is, in reality, a highly smooth function.

4.1. BART in High Dimensions

When the number of covariates, p, is of modest size relative to the number of samples, BART models are quite resilient to the inclusion of spurious predictors. When there are a large number of potentially irrelevant predictors that we wish to control for, however, BART can break down. To make the point, consider the needles-in-a-haystack regime in which X is of dimension $p \gg n$. **Figure 2** illustrates this point using the Friedman test function

$$f(\mathbf{x}) = 10 \sin(\pi x_1 x_2) + 20(x_3 - 0.5)^2 + 10x_4 + 5x_5,$$

which defines the mean structure even as the number of irrelevant predictors, $p - 5$, increases. In each panel, a BART model is fit to a sample of $N = 100$ points from the model $Y_i = f(\mathbf{X}_i) + \epsilon_i$, with $\mathbf{X}_i$s uniformly distributed on $[0, 1]$ and $\epsilon_i \sim \text{N}(0, 1)$, and performance is evaluated on a held-out test set of 100 additional samples of $\mathbf{X}_i$. The true values of $f(\mathbf{X}_i)$ for the held-out points are given on the x-axis, while the predictions from the model are given on the y-axis, with a credible interval for the prediction given around each point. The predictions from a perfectly fit model should lie on the 45° line.

If we fit the Chipman et al. (2007, 2010) BART model, we see that the bias increases substantially as p increases; accordingly, the credible intervals widen as well. We can correct this undesirable behavior by making a simple modification to the prior. Recall that in Section 2.2 we stated that, in sampling a tree from the prior, that the predictors were chosen uniformly at random from the collection of possible predictors. We modify this assumption by introducing a vector of splitting proportions $s = (s_1, \ldots, s_p)$ such that the probability that predictor j is used to construct a split is given by s_j. Observe that if (say) $s_1 = 0$ then we know that x_1 will not appear in the model. But we do not actually need $s_j = 0$ to achieve this either; if, for example, $s_1 < 10^{-10}$, then x_1 is still extremely unlikely to appear in the model, even if the number of splits is large.

Motivated by this observation, Linero (2018) proposed a sparsity-inducing Dirichlet prior $s \sim \text{Dirichlet}(\xi/p, \ldots, \xi/p)$ to achieve a prior that encourages most components of s to be extremely small, effectively filtering out most of the variables in the model. The second row of **Figure 2**, labeled DART (indicating Dirichlet additive regression trees), demonstrates this behavior. We see that, even for large values of p, the predictive performance of the model is quite strong. A related approach is described by Rockova & van der Pas (2017). They propose a spike-and-tree prior, which sets $s_j = \delta_j / \sum_{k=1}^{p} \delta_k$ where $(\delta_1, \ldots, \delta_p)$ is a vector of binary indicator variables. Equivalently, we select a group of D variables to be included in the model and set $s_j = D^{-1}$ if x_j is included in the model. Liu et al. (2019) show that certain spike-and-tree priors have very

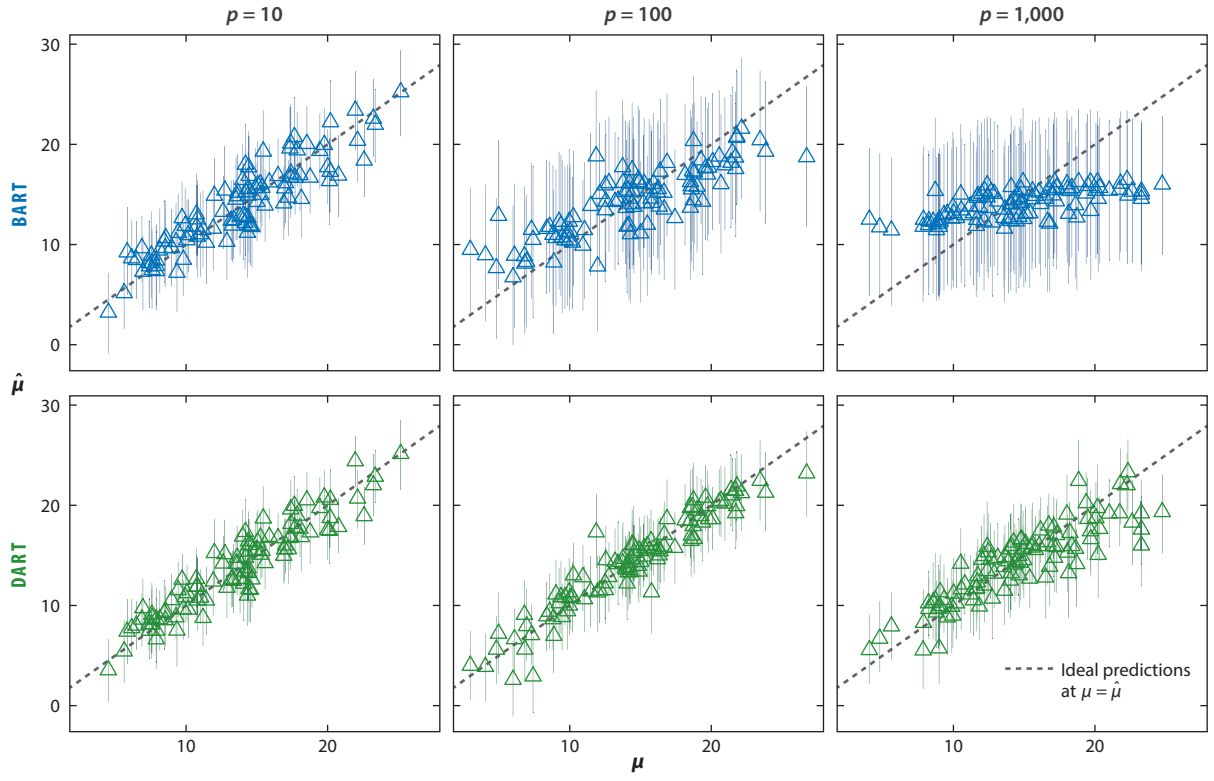

Figure 2

Plot of the regression function $f(\mathbf{x}) = 10\sin(\pi x_1 x_2) + 20(x_3 - 0.5)^2 + 10x_4 + 5x_5$ against the predicted value from the model for 100 held-out observations, with 95% credible intervals around each prediction. BART denotes the Bayesian additive regression tree prior, DART denotes the same model but using a Dirichlet prior on the splitting proportions, μ denotes the true value of $f(\mathbf{x})$, $\widehat{\mu}$ denotes the predicted value of $f(\mathbf{x})$, and P denotes the number of predictors in total. Error variance is $\sigma^2 = 1$. Figure adapted with permission from Linero (2018).

attractive variable selection properties. A drawback of these models is that they are not easy to implement in practice.

4.2. Soft BART Models

A problem with procedures such as classification and regression trees (CART) for growing decision trees is that the resulting estimates of $f(\mathbf{x})$ are step functions, which are not well suited to approximating continuous or differentiable functions. The BART model obtains some level of smoothing from the fact that the estimate $\widehat{f}(\mathbf{x})$ is obtained by averaging over the posterior distribution; depending on the prior used for the tree structures, the estimate of $\widehat{f}(\mathbf{x})$ may even be continuous. Nevertheless, if the underlying $f(\mathbf{x})$ is differentiable, then the performance of BART will be poor relative to alternatives that can take advantage of this additional smoothness. This point is illustrated in **Figure 3**. The BART estimates resemble continuous, nowhere differentiable, functions. The right panels show estimates using a soft version of BART, described next.

A regression tree can be represented as $g(\mathbf{x}; T, M) = \sum_t \phi_t(\mathbf{x})\,\mu_t$ where $\phi_t(\mathbf{x}) = 1$ or 0 according to whether $\mathbf{x}$ is associated to leaf t or not. We can express $\phi_t(\mathbf{x})$ in terms of the rules r of the

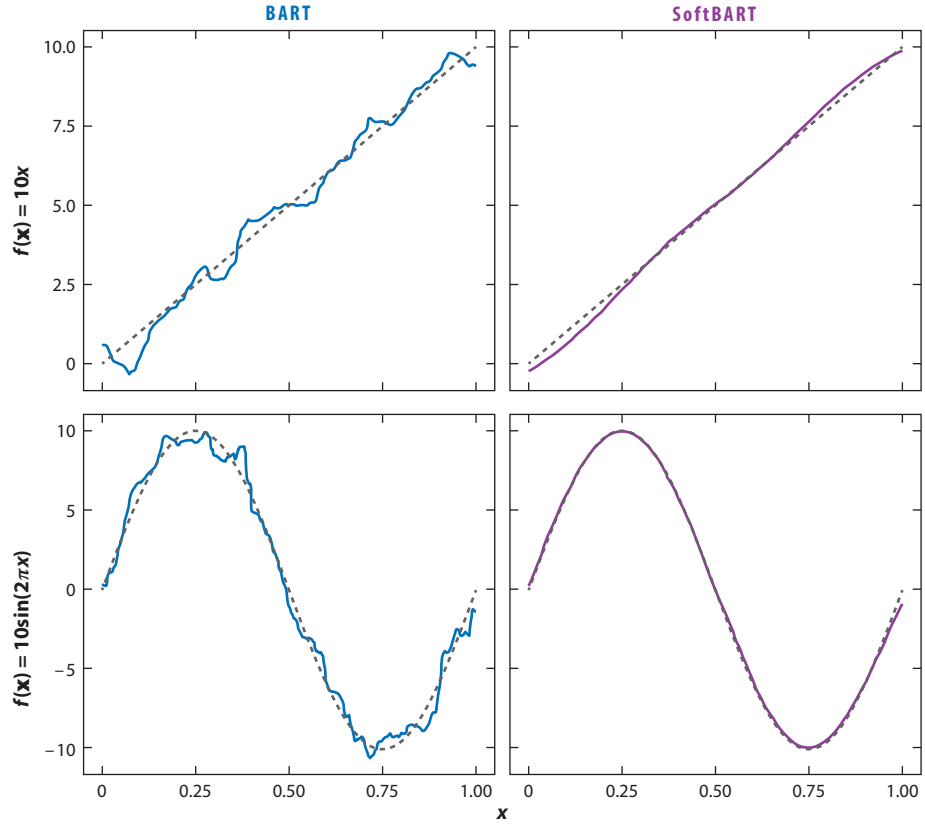

Figure 3

Posterior means (*solid lines*) against underlying true regression function (*dashed lines*). The left panels display the estimate of $f(\mathbf{x})$ when BART is used to estimate (*top*) $f(\mathbf{x}) = 10x$ and (*bottom*) $f(\mathbf{x}) = 10\sin(2\pi x)$. Error variance is $\sigma^2 = 2^2$. The right panels show estimates using a soft version of BART. Adapted with permission from Linero & Yang (2018).

tree as

$$\phi_t(\mathbf{x}) = \prod_{r \in A(t)} I(x_{j(r)} \leq C_r)^{L_{rt}} I(x_{j(r)} > C_r)^{1-L_{rt}}, \tag{13.}$$

where $A(t)$ is the collection of nodes that are on the path from the root to leaf t, $j(r)$ denotes the predictor used to construct the split for the rule r, and $L_{rt} = 1$ if the path from the root to t goes left at r ($L_{rt} = 0$ otherwise). One way to induce smoothness in decision trees is to regard the decisions made at each node as random rather than deterministic. That is, rather than going left if $x_j \leq C_r$ and right otherwise, we instead go right with probability $\psi(x_j; C_r, \tau)$ where $\psi(\cdot; C, \tau)$ is some cumulative distribution function. For example, we might set

$$\psi(x; C, \tau) = (1 + e^{-(x-C)/\tau})^{-1},$$

so that values of x that are much smaller than C will have a high probability of going left, while values of x that are much larger than C will have a high probability of going right. Note that the original BART model is recovered in the limit as $\tau \to 0$. Using this idea, we obtain a new formula

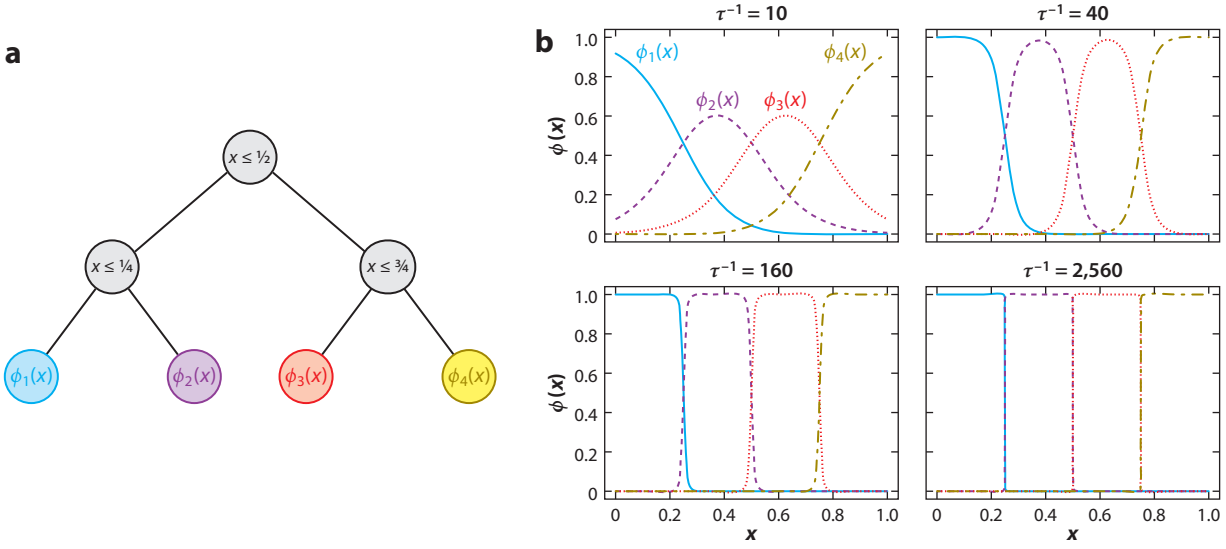

Figure 4

(*a*) A tree with cut-points $C_r = 0.5, 0.25$, and 0.75 on the predictor x_1. (*b*) The weights $\phi_t(\mathbf{x})$ defined in Equation 14 using $\tau^{-1} \in \{10,\ 40,\ 160,\ 2{,}560\}$, such that we predict $\sum_t \mu_t \phi_t(\mathbf{x})$. We see that, as $\tau \to 0$, the weight functions converge to step functions, which corresponds to the usual decision tree. Adapted with permission from Linero & Yang (2018).

for leaf node weights:

$$\phi_t(\mathbf{x}) = \prod_{r \in A(t)} [1 - \psi(x_{j(r)}; C_r, \tau)]^{L_{rt}}\ \psi(x_{j(r)}; C_r, \tau)^{1-L_{rt}}. \qquad 14.$$

For example, consider the tree in **Figure 4**, which for simplicity depends only on a univariate predictor x. The weight $\phi_2(x)$ is obtained as the probability of going left at the root $[1 - \psi(x; 0.5, \tau)]$ and then right $[\psi(x; 0.25, \tau)]$.

Linero & Yang (2018) refer to trees constructed using these randomized decisions as soft trees and the BART variant using soft decision trees as SBART. Whereas Equation 13 is not a smooth function of $\mathbf{x}$, Equation 14 is. As illustrated in **Figure 4**, the parameter τ controls how sharp the decisions in the tree are, with the limiting case $\tau \to 0$ corresponding to the usual BART model. Although it is a seemingly minor modification, Linero & Yang (2018) showed that using soft decision trees carries substantial theoretical and practical benefits. The primary drawback behind this modification is the need to compute $\phi_t(\mathbf{x})$ for every leaf t, rather than just at a single leaf node like BART; consequently, fitting SBART is somewhat slower than fitting BART.

5. DEVELOPMENTS IN BART THEORY

A recent stream of work has attempted to characterize the theoretical properties of using BART ensembles. The ultimate goal is to identify the properties that explain the excellent practical performance of BART. Several articles have made encouraging progress toward this goal.

Formally, we regard the sum of trees f as having a prior distribution $f \sim \Pi$, and we let $\Pi_{\mathcal{D}_n}(\cdot)$ denote the posterior distribution given data $\mathcal{D}_n$. We suppose there is a true underlying f_0 that we wish to recover, although we do not assume that f_0 is itself a sum-of-trees. For simplicity, we

consider the nonparametric regression problem $Y_i \sim N(f_0(\mathbf{X}_i), \sigma^2)$ with the covariate $\mathbf{X}_i \in [0,1]^p$ assumed random.

We can quantify the theoretical performance of a Bayesian method through its posterior concentration rate. We say that the posterior concentration rate is (at least) ϵ_n if $\Pi_{\mathcal{D}_n}(\|f - f_0\|_n > K\epsilon_n) \to 0$ in probability, for some constant $K > 0$, where $\|f - f_0\|_n^2 = n^{-1}\sum_{i=1}^n (f(\mathbf{X}_i) - f_0(\mathbf{X}_i))^2$. Different rates are possible depending on what assumptions one makes about f_0. For example, the optimal rate of convergence when f_0 is a twice-differentiable function of p variables is $\epsilon_n = n^{-2/(4+p)}$. For a generic α-times continuously differentiable function f_0 of p variables, the optimal rate of contraction is $n^{-\alpha/(2\alpha+p)}$. More precisely, this is the rate attainable for $f_0 \in C^\alpha([0,1]^p)$ of α-Hölder smooth functions (see, for example, Van Der Vaart & Wellner 1996).

Rockova & van der Pas (2017) and Linero & Yang (2018) considered the properties of BART priors in this framework and showed that certain BART priors are capable of attaining the rate $n^{-\alpha/(2\alpha+p)}(\log n)^c$ for $\alpha \leq 1$, while the SBART procedure described in Section 4.2 attains this rate for all $\alpha > 0$. Up to the logarithmic term, BART and SBART attain the best rate possible without prior knowledge of α. We note here the theoretical benefit of smoothness—if $f_0(\mathbf{x})$ is just twice differentiable, then the convergence rate of BART can be improved via smoothing (Linero & Yang 2018).

BART attaining a near-optimal convergence rate for a generic p-dimensional function does not explain the superior practical performance of BART when p is large. Note that, due to the curse of dimensionality, the rate degrades exponentially in p. By imposing more structure on f_0, however, we can improve the rate of convergence. Using the sparsity-inducing priors described in Section 4.1, certain BART priors are also capable of filtering out irrelevant predictors. In particular, if only $D \ll p$ of the predictors are relevant, we instead obtain the rate

$$n^{-\alpha/(2\alpha+D)}(\log n)^c + (D \log p/n)^{1/2}.$$

For large p, this is substantially faster than the naive rate $n^{-\alpha/(2\alpha+p)}$ (Linero & Yang 2018).

Sparsity alone still falls short of explaining the benefits of BART. Intuitively, each shallow tree in the sum $\sum_{b=1}^m g(\mathbf{x}, T_b, M_b)$ can be thought of as a low-order interaction. Hence, we expect BART to perform particularly well when the true $f_0(\mathbf{x})$ also consists primarily of low-order interactions. To formalize this, we assume that $f_0(\mathbf{x})$ is composed additively of V interactions

$$f_0(\mathbf{x}) = \sum_{v=1}^V f_{0v}(\mathbf{x}),$$

where f_{0v} depends only on $D_v \ll p$ of the predictors. For example, in a generalized additive model (Hastie & Tibshirani 1987), we would have $D_v = 1$ for all v. Linero & Yang (2018) and Rockova & van der Pas (2017) also study this setting and establish that BART and SBART attain close to the optimal rate of convergence for additive functions. For example, Linero & Yang (2018) establishes a rate of convergence of

$$\epsilon_n = \sum_{v=1}^V n^{-\alpha_v/(2\alpha_v+D_v)}(\log n)^c + (n^{-1}D_v \log p)^{1/2}$$

for SBART, without prior knowledge of the smoothness levels α_v, number of terms V, or interaction orders D_v. This is close to the optimal rate for this problem given by Yang & Tokdar (2015). A gap remains between the particular conditions and priors (which we have omitted) needed to

obtain these convergence rates and the form BART typically uses in practice. The most oppressive condition, required by both Rockova & van der Pas (2017) and Linero & Yang (2018) to obtain adaption to additive functions, is the use of a prior on the number of trees m. In practice, T is often set to a default value of $m = 200$, or else is selected by cross-validation. It is likely that selecting m by cross-validation also recovers these optimal convergence rates, but this has not been established theoretically.

6. BART FOR CAUSAL INFERENCE

One of the most popular applications of BART has been in the field of causal inference. Causal inference requires comparisons between an observed outcome and a counterfactual outcome that would have manifested under a different treatment regime.[2] Therefore estimating causal effects requires making inferences about data (counterfactual outcomes) that the researcher does not get to directly observe. In the absence of a randomized experiment, this requires both strong assumptions about the ability of our covariates to capture information about these counterfactuals and modeling strategies that are reliable across a variety of assumptions about how those outcomes relate to the covariates. BART can be useful for the latter goal.

6.1. Causal Inference Assumptions for Average Treatment Effects

We can formalize the idea of counterfactual outcomes though the potential outcome framework (Rubin 1978). In particular we denote the potential outcome under treatment as $Y(Z = 1) \equiv Y(1)$ and the potential outcome under control as $Y(Z = 0) \equiv Y(0)$, where Z denotes a binary treatment assignment variable. Estimation of average treatment effects such as $\mathrm{E}(Y(1) - Y(0))$ is straightforward given a pristine randomized experiment that yields data satisfying $Y(0), Y(1) \perp Z$. However most causal research questions are not afforded this luxury and rely instead on observational data where the treatment groups are likely to have different distributions for their potential outcomes. In other words, in observational studies, it is likely that observations exposed to treatment or control differ with regard to pretreatment characteristics that also affect their outcomes (confounders). We can make progress if these differences are captured in observed covariates and the distribution of these covariates is sufficiently similar across groups.

Expressed formally, identifying treatment effects requires two primary assumptions. The most important assumption, which is also the most difficult to satisfy, is the ignorability assumption, formalized as $Y(0), Y(1) \perp Z \mid \mathbf{x}$. In essence, this assumption says that if we had two groups of people with identical values on all observed covariates, $\mathbf{x}$, and then we assume that these groups had the same chance of receiving the treatment. Equivalently for this problem, these groups have the same distribution of potential outcomes.

Typically researchers invoke an assumption even stronger than ignorability referred to as strong ignorability. This combines the standard ignorability assumption with an assumption of overlap or common support. Formally this requires that $0 < \Pr(Z = 1 \mid \mathbf{x}) < 1$. If this fails to hold, then we may have neighborhoods of the confounder space where there are treated units but no control units or vice versa. That is, empirical counterfactuals (Hill & Su 2013) may not exist for all observations. Since most causal inference methods rely on some sort of modeling of the response surface, failure to satisfy this assumption forces stronger reliance on the parametric assumptions of that model.

[2]While this is the dominant perspective among methodological researchers in causal inference we note that dissenting opinions exist (for example Dawid 2000).

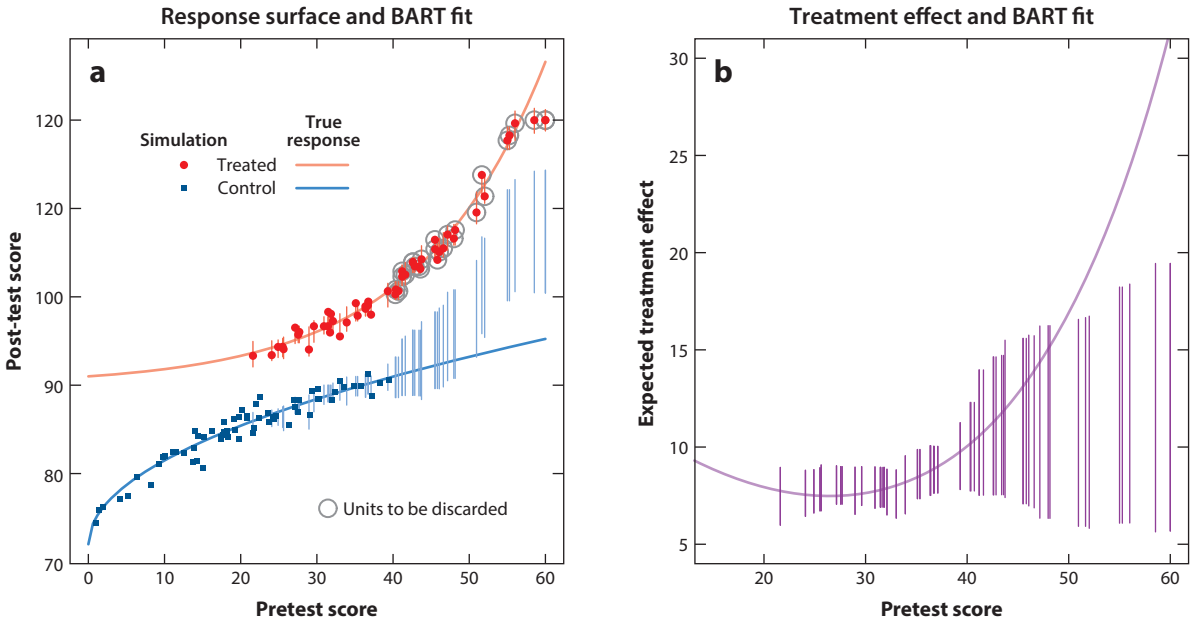

Figure 5

(*a*) Simulated data (*points*) and true response surface curves for treated (*red*) and control (*blue*) conditions. The red upper curve and points that follow it correspond to the treatment condition; the blue lower curve and points that follow it correspond to the control condition. BART inference for each treated observation is displayed as a 95% posterior interval for predicted $Y(1)$ and $Y(0)$. (*b*) Solid curve represents the expected treatment effect as it varies with X. BART inference is displayed as 95% posterior intervals (*vertical lines*) for the treatment effect for each treated unit. Abbreviation: BART, Bayesian additive regression trees.

6.2. BART and Parametric Assumptions

Consider estimation of a conditional average treatment effect (CATE), defined as

$$\tau(\mathbf{x}) \equiv \mathrm{E}[Y(1) - Y(0) \mid \mathbf{x}]. \qquad 15.$$

CATEs may be interesting in their own right, perhaps as estimates of individual treatment effects, or averaged against different distributions of $\mathbf{x}$ to estimate various population, sample, or subgroup average treatment effects. Taken together, the assumptions in the previous subsection imply that $\mathrm{E}[Y(z) \mid \mathbf{x}] = \mathrm{E}[Y \mid Z = z, \mathbf{x}]$. In this case we can express the CATE in terms of the observed (rather than potential) outcomes: $\tau(\mathbf{x}) = \mathrm{E}[Y \mid Z = 1, \mathbf{x}] - \mathrm{E}[Y \mid Z = 0, \mathbf{x}]$. These conditional expectations can be estimated using regression models—including BART—that model $\mathrm{E}[Y \mid Z = z, \mathbf{x}] \equiv f(\mathbf{x}, z)$ directly.

Taking this response surface approach to causal modeling demands a flexible model for f. Imagine trying to estimate an average treatment effect for all or some subset of the observations in **Figure 5*a*** [this example is taken from Hill (2011)]. For this exercise we assume that ignorability holds given a scalar pretest score, denoted by x. This Figure displays $\mathrm{E}[Y(0) \mid x]$ and $\mathrm{E}[Y(1) \mid x]$ as well as observed data points for treated and control observations.

In **Figure 5*a***, the upper red curve represents $E[Y(1) \mid x]$ and the lower blue one $E[Y(0) \mid x]$. The red circles close to the upper curve are the treated and the blue squares close to the lower curve are the untreated (ignore the circled points for now). Since there is only one confounding covariate, pretest, the difference between the treatment and control curves in the response surface at any level of pretest represents the treatment effect for observations with that value of the pretest.

This figure reveals several important features of the data. First, the treatment and control surfaces are nonlinear and not parallel. As a consequence, the response surface may be tricky to estimate well with run-of-the-mill statistical models, particularly if there are many covariates in the set of confounders. Second, the overlap across the two groups with respect to the only confounder, pretest, is not strong. That means that we do not have any data to inform $Y(0)$ for observations with high pretest scores or to inform $Y(1)$ for observations with low pretest scores.

Here, the true sample average treatment effect for the treated (SATT) computed among the n_0 treated units is

$$\frac{1}{n_0} \sum_{i:Z_i=1} Y_i(1) - Y_i(0) = 12.2.$$

A linear regression fit (without an interaction term) to the data yields a substantial underestimate, 7.1 (standard error .62), of the SATT. Even with an interaction included, this fit performs poorly due to the nonlinearity.

In contrast, **Figure 5*a*** displays the BART fit to the response surface. Each vertical line segment corresponds to posterior uncertainty intervals for either $E[Y_i(0) \mid x_i]$ or $E[Y_i(1) \mid x_i]$ for each treated observation. The fit is quite good until we try to predict $E[Y_i(0) \mid x_i]$ beyond the support of the data.

Figure 5*b* shows the results of using the BART fit to estimate $\tau(\mathbf{x}_i) = E[Y_i(1) - Y_i(0) \mid x_i]$ for the treated units. The vertical lines display 95% posterior intervals of $\tau(\mathbf{x}_i)$ for each treated observation. The curve is the true treatment effect $\tau(\mathbf{x})$ as it varies with levels of the pretest $\mathbf{x}$. The SATT can be estimated using the posterior distribution of $\frac{1}{n_0} \sum_{i:Z_i=1} \tau(\mathbf{x}_i)$, readily computed from MCMC output. The BART-based estimate of the SATT is 9.5 with 95% posterior interval (7.7, 11.8). We obtain a somewhat better estimate and wider uncertainty intervals than the linear regression, but we still underestimate the true SATT because of the failure of overlap in this data set. We return to this issue in Section 6.4.

In principle, any method that flexibly estimates f could be used to model these conditional expectations, though the strong performance of BART in the absence of cross-validation is a useful feature (Chipman et al. 2007, 2010). Hill (2011) was the first paper to describe the advantages of using BART for causal inference estimation over common alternatives in the causal inference literature. Hill et al. (2011), Green & Kern (2012), Kern et al. (2016), Hahn et al. (2017), and Wendling et al. (2018) also support the usefulness of BART in this setting. The results from the 2016 Causal Inference Data Analysis Challenge (Dorie et al. 2019) provide further support.[3]

6.3. Reparameterizations and Regularization-Induced Confounding in Causal Models

Hahn et al. (2017) proposed some modifications to the BART model to tailor it to causal modeling. Hill (2011) originally used BART to estimate causal effects by appending the treatment indicator z as a covariate and fitting the model

$$y_i = f(\mathbf{x}_i, z_i) + \epsilon_i. \qquad 16.$$

Substantial empirical evidence points to the effectiveness of this approach. However, it has some limitations. This model does not allow for direct regularization of treatment effects, the prior

[3]The 2017, 2018, and 2019 incarnations of these challenges also provide support for the superior performance of BART for causal inference. The 2017 results are available as a technical report (Hahn et al. 2019); the 2018 and 2019 results are currently unpublished but were announced at the Atlantic Causal Inference Conferences with which they were associated.

distribution on $\tau(\mathbf{x})$ is induced only indirectly, and it is impossible to adjust the prior distribution over heterogeneous treatment effects independently of the prognostic or direct effects of covariates on the response. Furthermore, every covariate in the model is necessarily a potential confounder and effect modifier. In practice, in an observational study, we may have a large set of potential confounders to control for and a more limited set of variables that are substantively interesting effect modifiers. Similarly, with data from a randomized controlled trial, we may want to include many variables as controls to reduce residual variation while including a more limited set of variables driving treatment effect heterogeneity.

Hahn et al. (2017) propose parameterizing the model directly in terms of the heterogeneous treatment effects:

$$y_i = \mu(\mathbf{x}_i) + \tau(\tilde{\mathbf{x}}_i)z_i + \epsilon_i \qquad 17.$$

[i.e., setting $f(\mathbf{x}_i, z_i) = \mu(\mathbf{x}_i) + \tau(\tilde{\mathbf{x}}_i)z_i$ in Equation 16], where $\tilde{\mathbf{x}}_i$ could be the same as $\mathbf{x}$ or include only a subset of the variables in $\mathbf{x}$. This model is exactly as expressive as Equation 16 but includes the treatment effect function $\tau(\tilde{x}_i)$ as a distinct parameter that can be regularized independently of μ with an independent prior distribution. Hahn et al. (2017) model both μ and τ with BART, tweaking the prior distribution on trees to apply stronger shrinkage to the treatment effect function τ by default. Zeldow et al. (2018) propose a model parameterized similarly to Equation 17, replacing τ with a parametric (linear) function of the treatment and effect modifiers while the rest of the confounders enter through the BART function $\mu(\mathbf{x}_i)$.

Finally, as a general suggestion in settings with moderate to strong confounding, Hahn et al. (2017) also suggest including an estimate of the propensity score among the covariates in μ (or in f if using the original BART formulation, a modification they term propensity score BART). They suggest this specifically as a solution to regularization-induced confounding, a phenomenon where estimates of average treatment effects and CATEs estimated using response surface methods can exhibit substantial bias when regularization is applied during estimation—even when all the confounders are measured and included.

6.4. Extensions for Overlap

One vulnerability of BART identified by Hill (2011) and captured in **Figure 5** is that there is nothing to prevent the model from extrapolating over areas of the covariate space where common support does not exist. This problem is not unique to BART; it is shared by all causal modeling strategies that do not first discard (or severely downweight) units in these areas. Such extrapolations can lead to biased inferences because of the lack of information available to identify either $\mathrm{E}[Y(0) \mid \mathbf{x}]$ or $\mathrm{E}[Y(1) \mid \mathbf{x}]$ in these regions. Hill & Su (2013) present a BART-based solution to this problem that has advantages over propensity-score-based approaches.

To understand how it works, we return to **Figure 5**. Notice how the uncertainty bounds grow much wider in the range where there is no overlap across treatment groups (pretest > 40). The individual-specific treatment effect intervals nicely cover the true conditional treatment effects until we start to leave this neighborhood. However, inference in this region is based on extrapolation. Hill & Su (2013) present two solutions that capitalize on the differential in posterior standard deviations for the factual outcome [for example, $Y(1)$ for a treated unit] relative to the counterfactual outcome [$Y(0)$ for that same treated unit]. Roughly speaking, if the latter is very large relative to the former (as it is for many of the treated observations with pretest scores above 40 in the above plot), then the unit is considered to have insufficient overlap and is flagged for discarding. The units flagged by the BART approach for discarding in **Figure 5*a*** are circled; the corresponding intervals in panel *b* are displayed with lines.

This strategy has important advantages over propensity score matching approaches to identifying overlap. Propensity score strategies weight most strongly the variables most predictive of the treatment variable. They ignore information about overlap embedded in the response variable. However, it is only important to enforce overlap with respect to covariates that are associated with both the treatment assignment and the outcome. Thus, propensity score strategies may be too conservative and inappropriately discard or downweight observations (Hill & Su 2013).

In contrast, the BART approach naturally and coherently incorporates information in the outcome. Hill & Su (2013) demonstrate this through an illustrative example as well as a simulation study that compares the BART approaches to propensity-score based approaches across several different assumptions about the data generating process.

6.5. Extensions for Violations of Ignorability

The primary motivation behind using BART (or propensity scores, etc.) for causal inference is to avoid the bias incurred through misspecification of the response surface: $E[Y(0) \mid \mathbf{x}]$ and $E[Y(1) \mid \mathbf{x}]$. However, the more difficult assumption to relax (in the absence of a controlled randomized or natural experiment) is the ignorability assumption, in large part because this assumption is untestable.

One approach to dealing with the challenge is to evaluate the sensitivity of a study to potential unmeasured confounding across a variety of assumptions about the potential strength of that confounding. This allows the researcher to understand what level of confounding would be needed to alter their conclusions about the effect (magnitude, sign, or significance). Dorie et al. (2016) propose a simulation-based, two-parameter sensitivity analysis strategy that uses BART to fit the model for the response. This extends earlier work by Carnegie et al. (2016) that imposed a strict parametric model. This extension results in an easily interpretable framework for evaluating the impact of an unmeasured confounder that also limits the number of modeling assumptions. Dorie et al. (2016) evaluated this approach in a large-scale simulation setting.

The usefulness of this sensitivity analysis approach was also demonstrated with high blood pressure data taken from the third National Health and Nutrition Examination Survey. **Figure 6** shows the results of the sensitivity analysis for the effect of taking beta blockers and diuretics on systolic blood pressure. Each point on the plot represents a combination of sensitivity parameters reflecting the strength of association between the unobserved confounder and the treatment (x-axis) and the strength of the unobserved confounder and the outcome (y-axis). Contour lines reflect the set of such points (combinations of sensitivity parameters) that would result in a particular (standardized) estimate of the treatment effect (displayed on the line).

While the linear sensitivity analysis and the semiparametric sensitivity analysis produce similar naive treatment effect estimates (-0.16 and -0.15, respectively), the results of the latter are more sensitive to unobserved confounding than the former. For instance, an unobserved confounder with a coefficient in the treatment model of -0.5 and coefficient of 0.25 for the outcome model (refer to the corresponding locations in **Figure 6**) would not change the statistical significance of the results using the linear sensitivity analysis, while when using a semiparametric sensitivity analysis, the naive treatment effect of -0.15 is already not statistically significant. A comparison between the two panels in this Figure emphasizes that our inference about sensitivity of the treatment effects to unmeasured confounding can substantively change in important ways, depending on how the response surface is predicted. This method is currently available in the `treatSens` package in R (Carnegie et al. 2015).

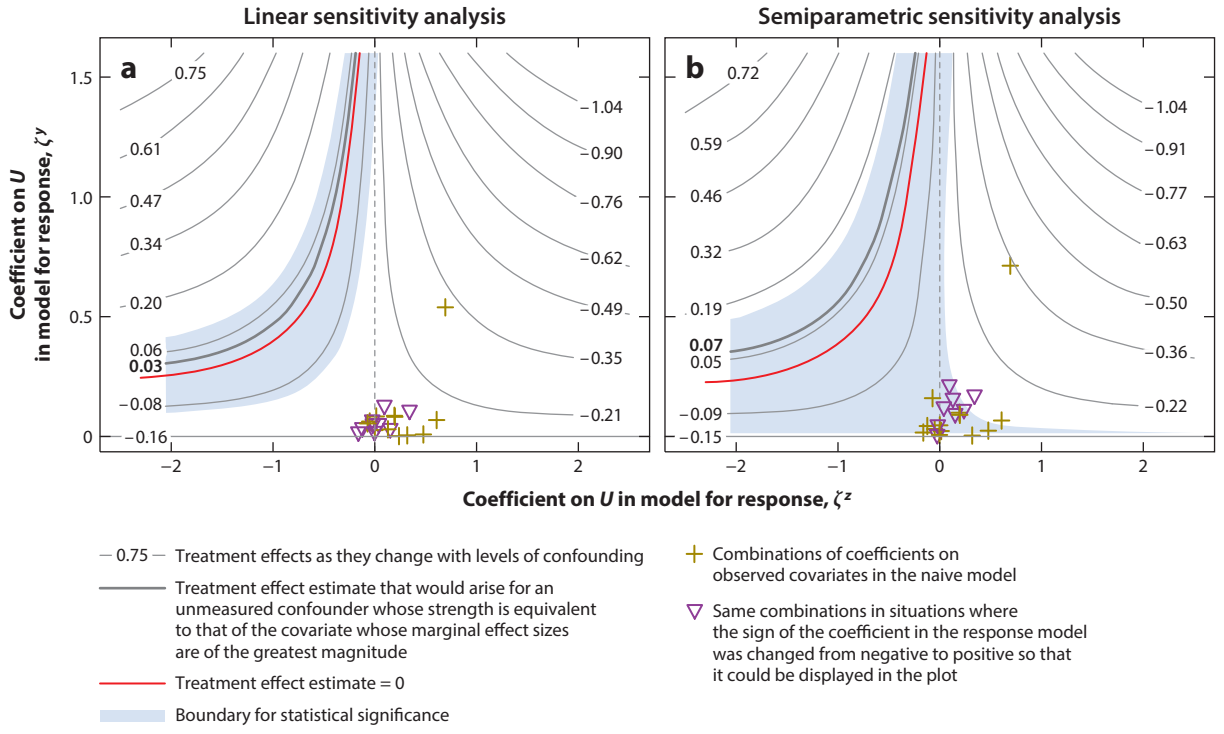

Figure 6

Sensitivity analysis results for the effects of taking beta blockers and diuretics on systolic blood pressure using the data from the third National Health and Nutrition Examination Survey. (*a*) The results of the linear sensitivity analysis in which the response surface is fitted with the linear combination of covariates. (*b*) The results of the semiparametric sensitivity analysis. Gray contour lines each correspond to a (standardized) estimates of the treatment effect corrected for bias indicated by the associated confounding levels shown on the line. Abbreviation: *U*, unobserved confounder.

6.6. Summarizing Causal Models and Making Decisions

One of the benefits of using rich Bayesian models like BART for estimating causal effects is that the posterior provides simultaneous inference on CATEs, sample average treatment effects, and everything in between. Moreover, it provides a wealth of opportunities for summarizing the information in the posteriors.

A simple but useful summary is a waterfall plot of point estimates and uncertainty intervals from posterior distributions for the CATE of every sample unit. **Figure 5*b*** displays a simple example of this, with the *x*-axis ordered by the single covariate, pretest. In examples with multiple covariates, this axis could be ordered instead by the magnitude of the treatment effect estimates (i.e., posterior means) for each person.

Green & Kern (2012) present partial dependence plots (Friedman 2001), which are generated by estimating and averaging counterfactual treatment effect functions, which in turn are generated by manipulating potential effect moderators (instead of counterfactual predictions as in typical partial dependence plots). This can help us understand how the CATEs vary with covariates, although they are subject to similar limitations as partial dependence plots for predictive models (see, for example, Goldstein et al. 2015, for extensive discussion of these).

Hahn et al. (2017) summarize the posterior distribution of $\tau(\mathbf{x})$ by fitting a CART tree to the posterior mean of $\tau(\mathbf{x})$ and estimating the posterior distribution of subgroup average treatment

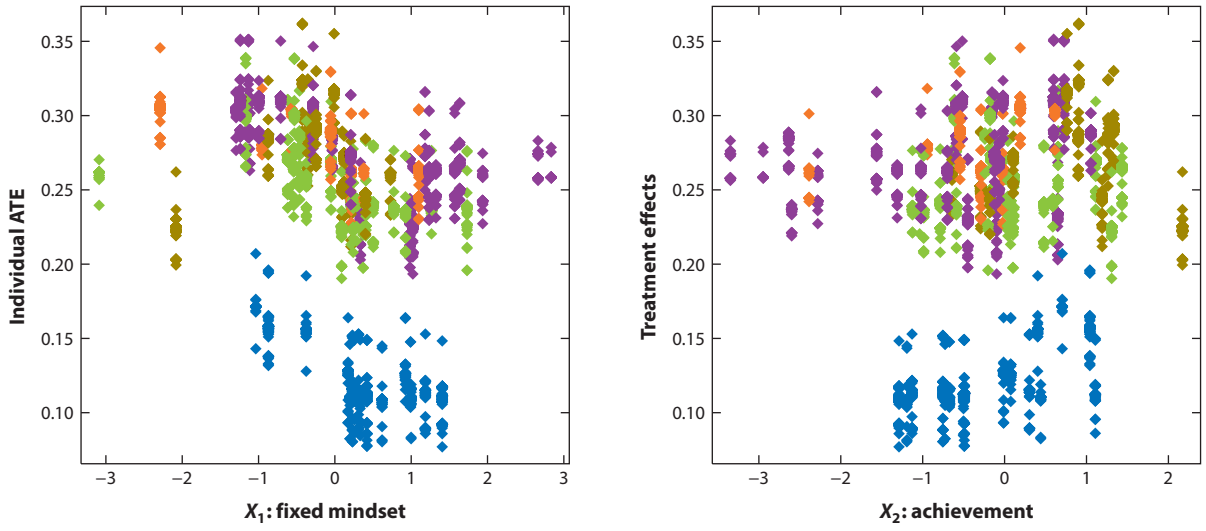

Figure 7

Scatter plots of school-specific CATEs with respect to two potential modifiers (levels of fixed mindset and achievement) instead reveals unexpected effect modification by levels of urbanicity, as displayed by the different colors of the treatment effects. Abbreviations: ATE, average treatment effect; CATE, conditional average treatment effect. Adapted with permission from Carnegie et al. (2019).

effects for each leaf, as well as the posterior distribution of the difference in subgroup average treatment effects, which is often more illuminating. Sivaganesan et al. (2017) present a more formal decision-theoretic approach to subgroup finding that is in much the same spirit.

Carnegie et al. (2019) use a variety of graphical summaries to explore treatment effect modification in an education example. **Figure 7** is reproduced from that paper and displays scatter plots of school-specified CATEs plotted versus each of two key covariates—a measure of fixed mindset beliefs and a measure of school-level achievement—revealing a cluster of schools with substantially different posterior treatment effects. A simple regression tree fit to these means with the covariates as predictors revealed that the modifier creating this clustering was a measure of a variable that the researchers were told reflects a characteristic referred to as "urbanicity" of the schools. The observations in **Figure 7** are colored according to the five levels of this variable.

Causal BART models also yield full posterior distributions over the outcomes under treatment and control. With these posterior distributions in hand, one can provide a formal decision theoretic treatment of optimal treatment selection for individuals by selecting the treatment rule that maximizes individual posterior expected utility. Logan et al. (2017) use BART in this framework to estimate individualized treatment rules.

6.7. Lessons Learned About BART from the 2016 Causal Inference Data Analysis Challenge

In an attempt to better understand the landscape of causal inference methods currently available, a Kaggle-style causal inference data analysis challenge was launched in 2016 in conjunction with that year's Atlantic Causal Inference Conference. The competition was motivated by the fact that although dozens of causal inference methods for observational studies are available to researchers, typical methods papers compare the relative efficacy of just two or three methods at a time. Moreover, these papers commonly are written by researchers who, however well meaning,

are interested in showcasing their own method. Thus, it is unclear that such comparisons are entirely fair to the alternative methods considered. The goal of the competition was to facilitate a broader comparison of methods since each submission was intended to show the method in its best light. A detailed description of the challenge motivation, development, and results is provided by Dorie et al. (2019).

In the primary competition track (black-box methods), 14 (distinct) black-box methods were submitted and the performance of these was evaluated across 7,700 data sets that varied the level of difficulty in estimating causal effects by altering features of the data such as overlap, balance, nonlinearity of the assignment mechanism and response surface, level of treatment effect heterogeneity, and rank of the confounder space (relative to that of the full set of covariates). These methods covered a wide variety of causal inference techniques, such as propensity score matching, inverse probability weighting, and regression adjustment. The four winners of the black-box competition were: (*a*) BART, (*b*) super learner with a targeted minimum loss-based estimation adjustment (TMLE) (Polley et al. 2016), (*c*) calCause (a proprietary IBM ensemble algorithm), and (*d*) `h2O`, an ensemble approach available in R (LeDell 2016). All of these flexibly fit the response surface. All of the non-BART approaches relied on ensembles of methods.

While many of the automated algorithms submitted to this competition performed well with regard to root mean squared error and bias, performance varied more widely with regard to interval coverage and length. All of the submitted methods were somewhat disappointing in this regard.

This poor performance, combined with curiosity about the potential effectiveness of combinations of features of the top performing methods, led the organizers to create a suite of new methods. Changes included adding a TMLE adjustment to BART, using symmetric (rather than percentile) intervals to summarize BART posterior distributions, relying on cross-validation to choose BART hyperparameters, including the propensity score as a an inverse probability of treatment weight or as a covariate in the BART model for the response surface [as proposed by Hahn et al. (2017)], and running multiple MCMC chains for BART. In addition, the super learner team took advantage of a one-month extension to the submission period after initial results were reported to resubmit their algorithm with BART included in the ensemble library. All of these augmentations improved performance. These features have been included in the `bartCause` software (discussed in the next section).

7. SOFTWARE

There are now a large number of R packages that make the application of BART models routine. Fast implementations of BART for regression and classification are given by the `dbarts` and `bartMachine` packages. The `dbarts` is a drop-in extension of the original `BayesTree` package and is being actively developed. Some `dbarts` features of note are that (*a*) it uses C++ with efficient data structures and thus is much faster than `BayesTree`, (*b*) it is natively parallelized within and across chains, (*c*) the sampler state can be updated/can embed in larger model, (*d*) it incorporates parallel cross-validation as an option, (*e*) weights can be included, (*f*) it can use model fit to predict for another data set, and (*g*) the default prior has been adjusted for better fit with binary response variables. Based on recent work by Carnegie (2019) demonstrating problems with mixing across chains, the default number of chains is now ten. `dbarts` is available on the Comprehensive R Archive Network (CRAN) (**https://cran.r-project.org/**).

Another option is the `BART` package, which (in addition to regression and classification) fits the survival model of Sparapani et al. (2016) and allows for the use of sparsity-inducing Dirichlet priors for high-dimensional problems described in Section 4.1. `BART` is also available on CRAN.

The SBART model described in Section 4.2 can be fit with the `SoftBart` package, which is currently available on GitHub (**https://github.com/**), and handles regression and classification problems.

Several packages implement the causal inference methodology described above. The `bcf` package (available on CRAN) implements the causal forest methodology described in Section 6.3. The `treatSens` package (available on CRAN) implements the sensitivity analysis strategy of Dorie et al. (2016) to understand the possible influence of inferences to unmeasured confounding. Finally, the `bartCause` package (available on GitHub) implements the BART causal model described by Hill (2011) and overlap checks described by Hill & Su (2013), as well as a range of augmentations described in Section 6.7 that were developed based on lessons learned from the 2016 Causal Inference Challenge (doubly robust strategies, TMLE, multiple chains). `bartCause` now implements the causal forest methodology implemented in `bcf` as well.

8. CHALLENGES AND FUTURE DIRECTIONS

BART has been extended in a variety of directions since it was first introduced a little more than a decade ago. We are aware of researchers pursuing avenues to make BART useful in a still wider range of settings. These include extensions to accommodate multilevel data with covariates at each level, time-varying data, and multiple outcomes.

Posterior computation has improved since the initial implementation of BART, but room for further improvement remains. Most BART implementations can handle hundreds of covariates and tens of thousands of observations, although mixing of the MCMC algorithm tends to degrade as either the sample size or dimension gets larger. Scaling to larger data sets (both in terms of the number of observations and the number of predictors) would naturally be quite useful. In all likelihood this will be more than an engineering exercise, and more efficient algorithms for posterior inference will be necessary.

While recent theoretical developments have been encouraging, the theory surrounding BART lags considerably behind related Bayesian nonparametric approaches such as Gaussian process regression. Goals for BART in this direction include establishing conditions under which the posterior credible intervals have frequentist validity and establishing semiparametric Bernstein–von Mises theorems (Bickel & Kleijn 2012) for the regression function and the causal effects that BART is often used to estimate.

9. CONCLUSION

BART capitalizes on the strengths of both machine learning and Bayesian inference. It uses the nonparametric sum-of-trees model to allow for flexible fit of the mean structure of a regression. But it also reaps the benefits of a Bayesian inferential framework with regard to uncertainty quantification and regularization through data-calibrated priors.

While initially the usefulness of the algorithm was limited to continuous and binary outcomes, the structure has been generalized since its introduction by Chipman et al. (2007, 2010) to accommodate a far wider variety of data structures. Advances in the prior specification provide further generalization to support high dimensions and better approximate smoother (continuous or differentiable) functions. Applications and extensions in the field of causal inference are promising and currently constitute a vibrant research area.

Theoretical developments provide justification for the strong performance evidenced in a growing body of simulation-based evaluations of its efficacy. A range of software options make these tools available to a wide range of researchers. We anticipate a good deal more development of theory, methods, and software in the decade to come.

DISCLOSURE STATEMENT

The authors are not aware of any affiliations, memberships, funding, or financial holdings that might be perceived as affecting the objectivity of this review.

ACKNOWLEDGMENTS

This research was supported by the Office of the Secretary of Defense, Directorate of Operational Test and Evaluation, and the Test Resource Management Center under the Science of Test research program. It was also supported by the grants from the National Science Foundation (NSF-DMS 1712870 and NSF SES-1631970), the Institute of Education Sciences (R305D110037), and the Office of Naval Research (N000141712141).

LITERATURE CITED

Albert JH, Chib S. 1993. Bayesian analysis of binary and polychotomous response data. *J. Am. Stat. Assoc.* 88:669–79

Bickel P, Kleijn B. 2012. The semiparametric Bernstein–von Mises theorem. *Ann. Stat.* 40:206–37

Bonato V, Baladandayuthapani V, Broom BM, Sulman EP, Aldape KD, Do KA. 2010. Bayesian ensemble methods for survival prediction in gene expression data. *Bioinformatics* 27:359–67

Carnegie N. 2019. Contributions of model features to BART causal inference performance using ACIC 2016 competition data. *Stat. Sci.* 34:90–93

Carnegie N, Dorie V, Hill J. 2019. Examining treatment effect heterogeneity using BART. *Obs. Stud.* In press

Carnegie NB, Harada M, Dorie V, Hill J. 2015. `treatSens`: sensitivity analysis for causal inference. *R package*. **https://rdrr.io/cran/treatSens/**

Carnegie NB, Harada M, Hill J. 2016. Assessing sensitivity to unmeasured confounding using a simulated potential confounder. *J. Res. Educ. Eff.* 9:395–420

Chipman HA, George EI, McCulloch RE. 1998. Bayesian CART model search. *J. Am. Stat. Assoc.* 93:935–48

Chipman HA, George EI, McCulloch R. 2007. Bayesian ensemble learning. In *Advances in Neural Information Processing Systems 19*, ed. B Schölkopf, J Platt, T Hoffman, pp. 265–72. Cambridge, MA: MIT Press

Chipman HA, George EI, McCulloch RE. 2010. BART: Bayesian additive regression trees. *Ann. Appl. Stat.* 4:266–98

Chipman HA, McCulloch R. 2010. `BayesTree`: Bayesian methods for tree based models. *R package*. **https://cran.r-project.org/web/packages/BayesTree/BayesTree.pdf**

Cox DR. 1972. Regression models and life-tables. *J. R. Stat. Soc. B* 34:187–202

Dawid AP. 2000. Causal inference without counterfactuals. *J. Am. Stat. Assoc.* 95:407–24

Denison DGT, Mallick BK, Smith AFM. 1998. A Bayesian CART algorithm. *Biometrika* 85:363–77

Dorie V, Carnegie NB, Harada M, Hill J. 2016. A flexible, interpretable framework for assessing sensitivity to unmeasured confounding. *Stat. Med.* 35:3453–70

Dorie V, Hill J, Shalit U, Scott M, Cervone D. 2019. Automated versus do-it-yourself methods for causal inference: lessons learned from a data analysis competition. *Stat. Sci.* 34:43–68

Friedman JH. 2001. Greedy function approximation: a gradient boosting machine. *Ann. Stat.* 29:1189–232

George E, Laud P, Logan B, McCulloch R, Sparapani R. 2018. Fully nonparametric Bayesian additive regression trees. arXiv:1807.00068 [stat.ML]

Goldstein A, Kapelner A, Bleich J, Pitkin E. 2015. Peeking inside the black box: visualizing statistical learning with plots of individual conditional expectation. *J. Comput. Graph. Stat.* 24:44–65

Green DP, Kern HL. 2012. Modeling heterogeneous treatment effects in survey experiments with Bayesian additive regression trees. *Public Opin. Q.* 76:491–511

Greene WH. 1994. *Accounting for excess zeros and sample selection in Poisson and negative binomial regression models*. Work. Pap. EC-94-10, New York Univ.

Hahn PR, Dorie V, Murray JS. 2019. Atlantic Causal Inference Conference (ACIC) data analysis challenge 2017. arXiv:1905.09515 [stat.ME]

Hahn PR, Murray JS, Carvalho C. 2017. Bayesian regression tree models for causal inference: regularization, confounding, and heterogeneous effects. arXiv:1706.09523 [stat.ME]
Hastie T, Tibshirani R. 1987. Generalized additive models: some applications. *J. Am. Stat. Assoc.* 82:371–86
Hastie T, Tibshirani R. 2000. Bayesian backfitting (with comments and a rejoinder by the authors). *Stat. Sci.* 15:196–223
He J, Yalov S, Hahn PR. 2019. XBART: Accelerated Bayesian additive regression trees. *PMLR* 89:1130–38
Hill J. 2011. Bayesian nonparametric modeling for causal inference. *J. Comput. Graph. Stat.* 20:217–40
Hill J, Su YS. 2013. Assessing lack of common support in causal inference using Bayesian nonparametrics: implications for evaluating the effect of breastfeeding on children's cognitive outcomes. *Ann. Appl. Stat.* 7:1386–420
Hill J, Weiss C, Zhai F. 2011. Challenges with propensity score strategies in a high-dimensional setting and a potential alternative. *Multivar. Behav. Res.* 46:477–513
Ibrahim JG, Chen MH, Sinha D. 2005. Bayesian survival analysis. In *Wiley StatsRef: Statistics Reference Online*, ed. N Balakrishnan, T Colton, B Everitt, W Piegorsch, F Ruggeri, JL Teugels. **https://onlinelibrary.wiley.com/doi/abs/10.1002/9781118445112.stat06003**
Kern HL, Stuart EA, Hill JL, Green DP. 2016. Assessing methods for generalizing experimental impact estimates to target samples. *J. Res. Educ. Eff.* 9:103–27
Kim H, Loh WY, Shih YS, Chaudhuri P. 2007. Visualizable and interpretable regression models with good prediction power. *IIE Trans.* 39:565–79
Lakshminarayanan B, Roy DM, Teh YW, Unit G. 2015. Particle Gibbs for Bayesian additive regression trees. *PMLR* 38:553–61
Lambert D. 1992. Zero-inflated Poisson regression, with an application to defects in manufacturing. *Technometrics* 34:1–14
LeDell E. 2016. `h2oensemble`: H2o ensemble learning. *R package*. **https://github.com/h2oai/h2o-3/tree/master/h2o-r/ensemble**
Linero AR. 2018. Bayesian regression trees for high-dimensional prediction and variable selection. *J. Am. Stat. Assoc.* 113:626–36
Linero AR, Sinha D, Lipsitz SR. 2018. Semiparametric mixed-scale models using shared Bayesian forests. arXiv:1809.08521 [stat.ME]
Linero AR, Yang Y. 2018. Bayesian regression tree ensembles that adapt to smoothness and sparsity. *J. R. Stat. Soc. B* 80:1087–110
Liu Y, Ročková V, Wang Y. 2019. Variable selection with ABC Bayesian forests. arXiv:1806.02304 [stat.ME]
Logan BR, Sparapani R, McCulloch RE, Laud PW. 2017. Decision making and uncertainty quantification for individualized treatments using Bayesian additive regression trees. *Stat. Methods Med. Res.* 28:1079–93
Murray JS. 2017. Log-linear Bayesian additive regression trees for categorical and count responses. arXiv:1701.01503 [stat.ME]
Polley E, LeDell E, Kennedy C, van der Laan M. 2016. `SuperLearner`: super learner prediction. *R package*. **https://CRAN.R-project.org/package=SuperLearner**
Pratola MT. 2016. Efficient Metropolis-Hastings proposal mechanisms for Bayesian regression tree models. *Bayesian Anal.* 11:885–911
Pratola MT, Chipman HA, Gattiker JR, Higdon DM, McCulloch R, Rust WN. 2014. Parallel Bayesian additive regression trees. *J. Comput. Graph. Stat.* 23:830–52
Pratola MT, Chipman HA, George EI, McCulloch RE. 2017. Heteroscedastic BART using multiplicative regression trees. arXiv:1709.07542 [stat.ME]
Rockova V, van der Pas S. 2017. Posterior concentration for Bayesian regression trees and their ensembles. arXiv:1708.08734 [math.ST]
Rubin DB. 1978. Bayesian inference for causal effects: the role of randomization. *Ann. Stat.* 6:34–58
Sivaganesan S, Müller P, Huang B. 2017. Subgroup finding via Bayesian additive regression trees. *Stat. Med.* 36:2391–403
Sparapani RA, Logan BR, McCulloch RE, Laud PW. 2016. Nonparametric survival analysis using Bayesian additive regression trees (BART). *Stat. Med.* 35(16):2741–53

Starling JE, Murray JS, Carvalho CM, Bukowski R, Scott JG. 2019. BART with targeted smoothing: an analysis of patient-specific stillbirth risk. arXiv:1805.07656 [stat.ME]

Van Der Vaart AW, Wellner JA. 1996. *Weak Convergence*. New York: Springer

Wendling T, Jung K, Callahan A, Schuler A, Shah N, Gallego B. 2018. Comparing methods for estimation of heterogeneous treatment effects using observational data from health care databases. *Stat. Med.* 37:3309–24

Wu Y, Tjelmeland H, West M. 2007. Bayesian CART: prior specification and posterior simulation. *J. Comput. Graph. Stat.* 16:44–66

Yang Y, Tokdar ST. 2015. Minimax-optimal nonparametric regression in high dimensions. *Ann. Stat.* 43:652–74

Zeldow B, Re VL III, Roy J. 2018. A semiparametric modeling approach using Bayesian additive regression trees with an application to evaluate heterogeneous treatment effects. arXiv:1806.04200 [stat.AP]