P11

Bayesian additive regression trees and the General BART model

Yaoyuan Vincent Tan and Jason Roy

January 23, 2019

Abstract

Bayesian additive regression trees (BART) is a flexible prediction model/machine learning approach that has gained widespread popularity in recent years. As BART becomes more mainstream, there is an increased need for a paper that walks readers through the details of BART, from what it is to why it works. This tutorial is aimed at providing such a resource. In addition to explaining the different components of BART using simple examples, we also discuss a framework, the General BART model, that unifies some of the recent BART extensions, including semiparametric models, correlated outcomes, statistical matching problems in surveys, and models with weaker distributional assumptions. By showing how these models fit into a single framework, we hope to demonstrate a simple way of applying BART to research problems that go beyond the original independent continuous or binary outcomes framework.

Keywords: semiparametric models; spatial; Dirichlet process mixtures; machine learning; Bayesian nonparametrics

1 Introduction

Owing to its success, Bayesian additive regression trees (BART; Chipman et al., 2010) has gained popularity in the recent years among the research community with numerous applications including biomarker discovery in proteomic studies (Hernández et al., 2015), estimating indoor radon concentrations (Kropat et al., 2015), estimation of causal effects (Leonti et al., 2010; Hill, 2011), genomic studies (Liu et al., 2010), hospital performance evaluation (Liu et al., 2015), prediction of credit risk (Zhang and Härdle, 2010), predicting power outages during hurricane events (Nateghi et al., 2011), prediction of trip durations in transportation (Chipman et al., 2010), and somatic prediction in tumor experiments (Ding et al., 2012). BART has also been extended to survival outcomes (Bonato et al., 2011; Sparapani et al., 2016), multinomial outcomes (Kindo et al., 2016; Agarwal et al., 2013), and semi-continuous outcomes (Linero et al., 2018). In the causal inference literature, notable papers that promote the use of BART include Hill (2011) and Green and Kern (2012). BART has also been consistently among the best performing methods in the Atlantic causal inference data analysis challenge (Hill, 2016; Hahn et al., 2017; Dorie et al., 2017). In addition, BART has been making inroads in the missing data literature. For the imputation of missing covariates, Xu et al. (2016) proposed a way to utilize BART for the sequential imputation of missing covariates, while Kapelner and Bleich (2015) proposed to treat missingness in covariates as a category and set up the splitting criteria so that the eventual likelihood in the Metropolis-Hasting (MH) step of BART is maximized. For the imputation of missing outcomes, Tan et al. (2018a) examined how BART can improve the robustness of existing doubly robust methods in situations where it is likely that both the mean and propensity models could be misspecified. Other more recent attempts to utilize or extend BART include applying BART to quantile regression (Kindo et al., 2016), extending BART to count responses (Murray, 2017), using BART in functional data (Starling et al., 2018), applying BART to recurrent events (Sparapani et al., 2018), identifying subgroups using BART (Sivaganesan et al., 2017; Schnell et al., 2016, 2018), and using BART as a robust model to impute missing principal strata to account for selection bias due to death (Tan et al., 2018).

The widespread use of BART has resulted in many researchers starting to use BART as the reference model for comparison when proposing new statistical or prediction methods which are flexible and/or robust to model misspecification. A few recent examples include Liang et al. (2018), Nalenz and Villani (2018), and Lu et al. (2018). This growing interest for BART raises a need for an in-depth tutorial paper on this topic to help researchers who are interested in using BART better understand the method that they are using and possibly diagnose the likely problems when unexpected results occur. The first portion of this paper is aimed at addressing this.

The second portion of our work revolves around an interesting observation on four works extending BART beyond the original independent continuous or binary outcomes setup. In these papers, they extend BART to semiparametric situations (Zeldow et al., 2018), correlated outcomes (Tan et al., 2018b), survey (Zhang et al., 2007), and robust error assumptions (George et al., 2018). Although these papers were written separately, they surprisingly share a common feature in their framework. In brief, when estimating the posterior distribution, they subtract a latent variable from the outcome and then model this residual as BART. This idea, although simple, is powerful because this can allow researchers to easily extend BART to problems that they may face in their dataset without having to rewrite or re-derive the Monte Carlo Markov Chain (MCMC) procedure for drawing the regression trees in BART. We summarize this idea in a framework unifying these models that we call the General BART model. We suggest how the priors could be set and how the posterior distribution could be estimated. We then show how General BART is related to the these four models. We believe that by presenting our General BART model framework and linking it with the models in these four papers as examples, it will aid researchers who are trying to incorporate and extend BART to solve their research problems.

Our in-depth review of BART in Section 2 focuses on three commonly asked questions regarding BART: What gives BART flexibility? Why is it called a sum of regression trees? What are the mechanics of the BART algorithm? In Section 3, we demonstrate the superior performance of BART compared to the Bayesian linear regression (BLR) when data are generated from a complicated model. We then describe the application of BART to two real-life datasets, one with continuous outcomes and the other with binary outcomes. Section 4 lays out the framework for our General BART model that allows BART to be extended to semiparametric models, correlated outcomes, survey, and situations where a more robust assumption for the error term is needed. We then show how our General BART model is related to these four BART extension models. In each of these description examples, we describe how the prior distributions are set and how the posterior distribution is obtained. We conclude with a discussion in Section 5.

2 Bayesian additive regression trees

We begin our discussion with the independent continuous outcomes BART because this is the most natural way to explain BART. We argue that BART is flexible because it is able to handle non-linear main effects and multi-way interactions without much input from researchers. To demonstrate how BART handles these model features, we explain using a visual example of a regression tree. We then illustrate the concept of a sum of regression trees using a simple example with two regression trees. We next show how a sum of regression trees link with non-linearity. To show how BART determines these non-linear main and multi-way interaction effects automatically, we discuss two perspectives. First, we provide a visual and detailed breakdown of the BART algorithm at work using a simple example, providing intuition for each step along the way. Then, we provide a more rigorous explanation of the BART MCMC algorithm by discussing the prior distribution used for BART and how the posterior distribution is calculated. Finally, we show how these ideas can be extended to independent binary outcomes.

2.1 Continuous outcomes

2.1.1 Formal definition

We begin with the formal definition and notation of BART. Suppose we have a continuous outcome Y and p covariates X for n subjects. The goal is a model that can capture complex relationships between X and Y, with the aim of using it for prediction. BART attempts to estimate $f(x)$ from models of the form $Y = f(X) + \varepsilon_i$, where, for now, $\varepsilon_i \sim N(0, \sigma^2)$, $i = 1, \cdots, n$. To estimate $f(X)$, a sum of regression trees is specified as

$$f(X) = \sum_{j=1}^{m} g(X; T_j, M_j). \tag{1}$$

In Equation (1), T_j is the j^{th} binary tree structure and $M_j = \{\mu_{1j}, \ldots, \mu_{b_j j}\}$ is the vector of terminal node parameters associated with T_j. Note that T_j contains the information of which covariate to split on, the cutoff value in an internal node, as well as where the internal node is located in the binary tree. The constant m is usually fixed at a large number, e.g., 200.

We will next make much more clear what is meant by T_j and M_j, and also how this leads to extremely flexible models. We begin with the simple case of a single regression tree.

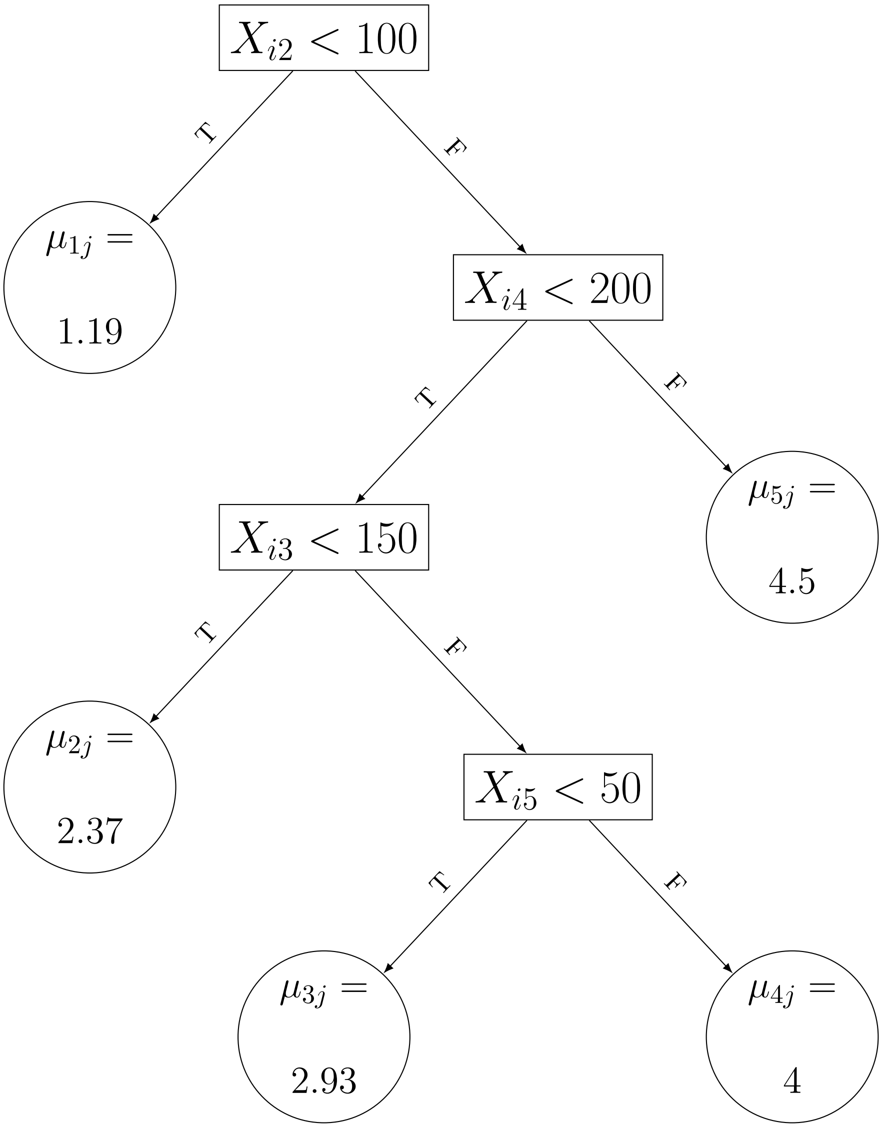
2.1.2 Single regression tree

To understand BART, consider first a single regression tree, rather than a sum of trees. For now we will assume that the tree is known and just focus on how to interpret it and obtain predictions from it. Later, we will describe the priors on these trees since the true tree structure is usually unknown.

Consider the regression tree $g(X; T_j, M_j)$ given in Figure 1. Imagine that we have covariates $X_i = (X_{i1}, \ldots, X_{i5})$ and we would like to know $E(Y_i|X_i)$ for subject i. Each place where there is a split is called a node. At the top node (*root*), there is a condition $X_{i2} < 100$. If $X_{i2} < 100$ is true, then we follow the path to the left, otherwise to the right. Assuming that $X_{i2} < 100$ is true, we see that we arrive at a node which is not split upon. This is called a *terminal node* and the parameter $\mu_{1j} = 1.19$ would be used as the predicted value of Y_i. Suppose instead that $X_{i2} < 100$ is *not* true. Then, moving along the right side, another internal node with condition $X_{i4} < 200$ is encountered. This condition would be checked and, if this condition is true (false), we would follow the path to the left (right). This process continues until we reach a terminal node and the parameter μ_{kj} in that terminal node is assigned as the predicted value of Y_i. Note that μ_{kj} is the mean parameter of the k^{th} node for the j^{th} regression tree. So, for example, a subject i with

$X_{i1} = 30, X_{i2} = 120, X_{i3} = 115, X_{i4} = 191$, and $X_{i5} = 56$ would be assigned a predicted outcome of $\mu_{2j} = 2.37$. The prediction would be exactly the same for another subject i' who instead had covariates $X_{i'1} = 130, X_{i'2} = 135, X_{i'3} = 92, X_{i'4} = 183, X_{i'5} = 10$.

Figure 1: Example of a regression tree $g(X; T_j, M_j)$ where μ_{kj} is the mean parameter of the k^{th} node for the j^{th} regression tree.



In summary, we can view a regression tree as a function that assigns the conditional mean of Y_i to the parameter μ_{kj} i.e. $\mu_{kj} = g(X_i; T_j, M_j) \mapsto E(Y_i|X_i)$. Note that we have not yet discussed how a tree is created and how uncertainty about what to split on and where to split is quantified. We will address that when we introduce priors and algorithms.

Regression tree as an analysis of variance (ANOVA) model. Another way to think of the regression tree in Figure 1 is to view it as the following analysis of variance (ANOVA) model:

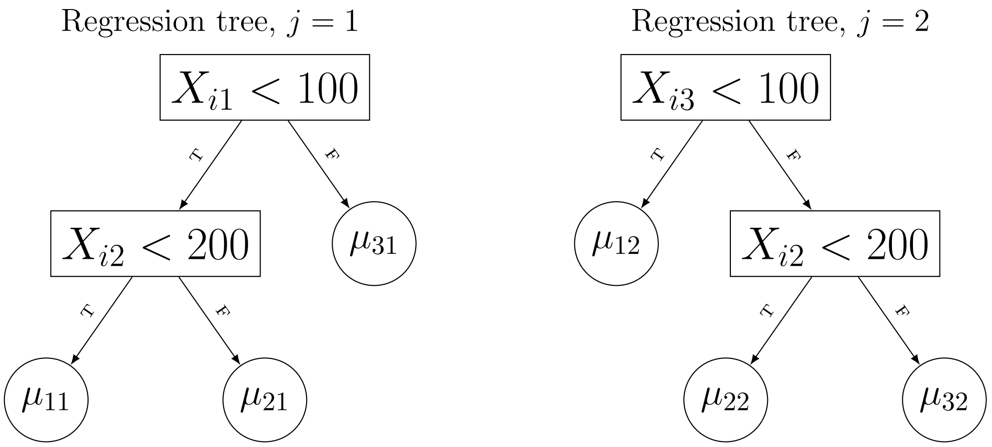
$$\begin{aligned} Y_i = {} & \mu_{1j} I\{X_{i2} < 100\} + \mu_{2j} I\{X_{i2} \geq 100\} I\{X_{i4} < 200\} I\{X_{i3} < 150\} \\ & + \mu_{3j} I\{X_{i2} \geq 100\} I\{X_{i4} < 200\} I\{X_{i3} \geq 150\} I\{X_{i5} < 50\} \\ & + \mu_{4j} I\{X_{i2} \geq 100\} I\{X_{i4} < 200\} I\{X_{i3} \geq 150\} I\{X_{i5} \geq 50\} \\ & + \mu_{5j} I\{X_{i2} \geq 100\} I\{X_{i4} \geq 200\} + \varepsilon_i, \end{aligned}$$

where $I\{.\}$ is the indicator function and $\varepsilon_i \sim N(0, \sigma^2)$. We can see that the term $\mu_{1j} I\{X_{i2} < 100\}$ corresponds to the terminal node on the top left corner of Figure 1, $\mu_{2j} I\{X_{i2} \geq 100\} I\{X_{i4} < 200\} I\{X_{i3} < 150\}$ correspond to the terminal node on the middle right of Figure 1, and so on. We can think of $\mu_{1j} I\{X_{i2} < 100\}$ as a main effect, because it only involves the second variable X_{i2}, while $\mu_{2j} I\{X_{i2} \geq 100\} I\{X_{i4} < 200\} I\{X_{i3} < 150\}$ is a three way interaction effect involving the second (X_{i2}), fourth (X_{i4}), and third variable (X_{i3}). By viewing a regression tree as an ANOVA model, we can easily see why a regression tree and hence, BART, is able to handle main and multi-way interaction effects.

2.1.3 Sum of regression trees

We next consider a sum of regression trees. To illustrate the main idea, we focus on an example with $m = 2$ trees and $p = 3$ covariates. Suppose we were given the two trees in Figure 2.

Figure 2: Illustrating the sum of regression trees using a simple two regression tree example.



The resulting conditional mean of Y given X is $\sum_{j=1}^{2} g(X; T_j, M_j)$. Consider the hypothetical data from $n = 10$ subjects given in Table 1. We can see that the quantity that is being 'summed' and eventually allocated to $E(Y_i|X_i)$ is not the regression tree or tree structure, but, the value that each j^{th} tree structure assigns to subject i. This is one way to think of a sum of regression trees. It allocates a sum of parameters μ_{kj} to subject i. Note that contrary to initial intuition, it is the sum of μ_{kj} that is allocated to rather than the mean of the μ_{kj}'s. This is mainly because BART calculates each posterior draw of the regression tree function $g(X; T_j, M_j)$ using a leave-one-out concept, which we shall elaborate on shortly.

Table 1: The values of $\sum_{j=1}^{2} g(X; T_j, M_j)$ from the regression trees in Figure 2.

i	Y	X_1	X_2	X_3	$g(X;T_1,M_1)$	$g(X;T_2,M_2)$	$\sum_{j=1}^{2} g(X;T_j,M_j)$
1	Y_1	-182	235	-333	μ_{21}	μ_{12}	$\mu_{21}+\mu_{12}$
2	Y_2	54	339	244	μ_{21}	μ_{22}	$\mu_{21}+\mu_{22}$
3	Y_3	-106	-50	-682	μ_{11}	μ_{12}	$\mu_{11}+\mu_{12}$
4	Y_4	-80	-62	-320	μ_{11}	μ_{12}	$\mu_{11}+\mu_{12}$
5	Y_5	-123	198	-77	μ_{11}	μ_{12}	$\mu_{11}+\mu_{12}$
6	Y_6	175	108	-46	μ_{31}	μ_{12}	$\mu_{31}+\mu_{12}$
7	Y_7	-44	11	136	μ_{11}	μ_{22}	$\mu_{11}+\mu_{22}$
8	Y_8	-131	-10	-70	μ_{11}	μ_{12}	$\mu_{11}+\mu_{12}$
9	Y_9	-56	68	257	μ_{11}	μ_{22}	$\mu_{11}+\mu_{22}$
10	Y_{10}	7	324	282	μ_{21}	μ_{32}	$\mu_{21}+\mu_{32}$

Another way to view the concept of a sum of regression trees is to think of the regression trees in Figure 2 as ANOVA models. Then, the sum of trees is the following ANOVA model:

$$
\begin{aligned}
Y_i &= g(X;T_1,M_1) + g(X;T_2,M_2) + \varepsilon_i \\
&= \mu_{11} I\{X_{i1} < 100\} I\{X_{i1} < 200\} + \mu_{21} I\{X_{i1} < 100\} I\{X_{i1} \geq 200\} + \mu_{31} I\{X_{i1} \geq 100\} \\
&+ \mu_{12} I\{X_{i3} < 100\} + \mu_{22} I\{X_{i3} \geq 100\} I\{X_{i2} < 200\} \\
&+ \mu_{32} I\{X_{i3} \geq 100\} I\{X_{i2} \geq 200\} + \varepsilon_i .
\end{aligned}
$$

Non-linearity of BART. From this simple example, we can see how BART handles non-linearity. Each single regression tree is a simple step-wise function or ANOVA model. When we sum regression trees together, we are actually summing together these ANOVA models or step-wise functions, and, as a result, we eventually obtain a more complicated step-wise function which can approximate the non-linearities in the main and multiple-way interactions. It is this ability to handle non-linear main and multiple-way interaction effects that makes BART a flexible model. But unlike many flexible models, BART does not require the researcher to specify the main and multi-way interaction effects.

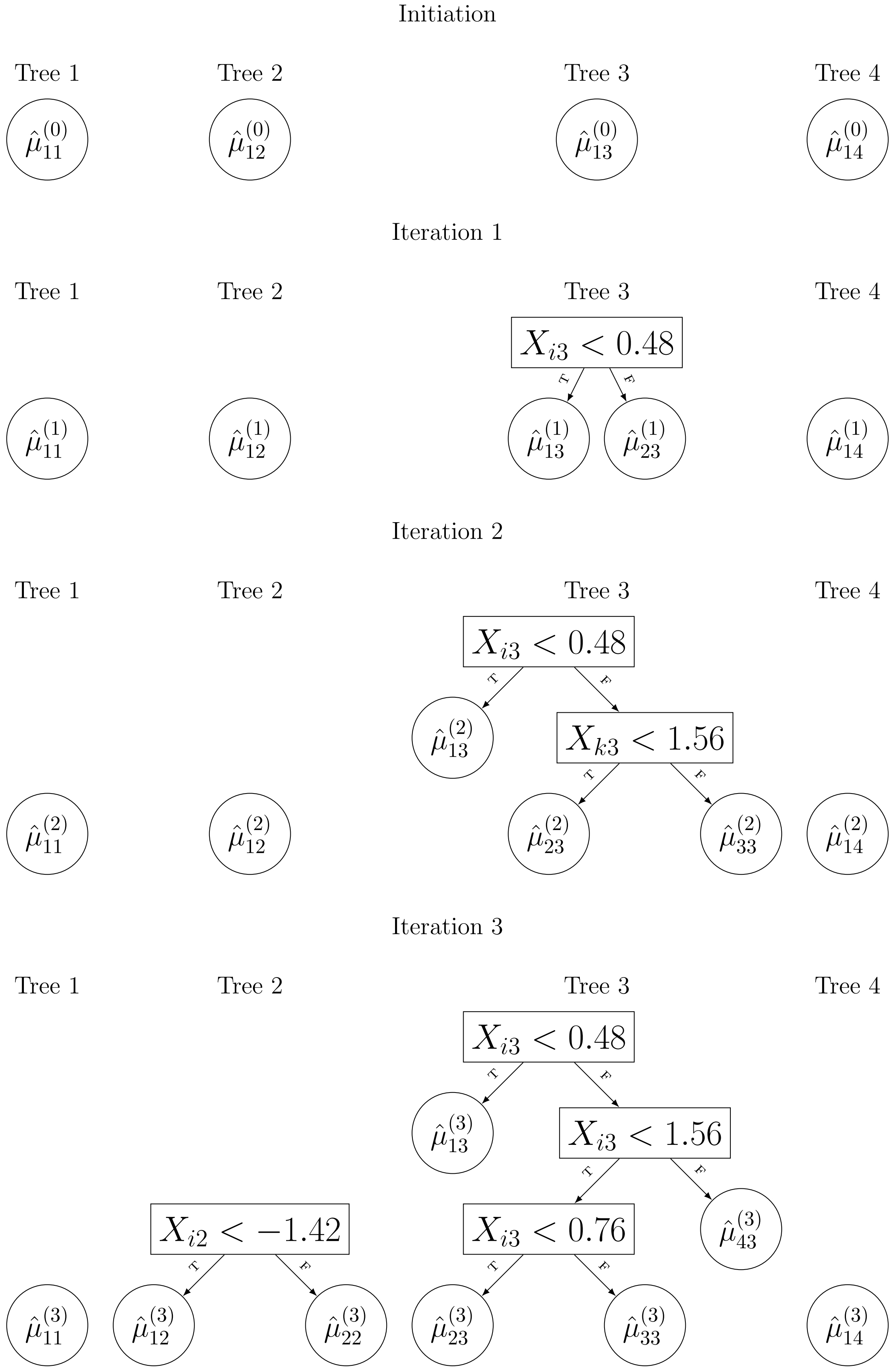
Prior distributions. In the examples above, we have taken the trees as a given, including which variables to split on, the splitting values, and the mean parameters at each terminal node. In practice, each $g(X; T_j, M_j)$ is unknown. We therefore need prior distributions for these functions. Thus, we can also think of BART as a Bayesian model where the mean function itself is unknown. A major advantage of this approach is that uncertainty about both the functional form and the parameters will be accounted for in the posterior predictive distribution of Y.

Before getting into the details of the prior distributions and MCMC algorithm, we will first walk through a simple example to build the intuition.

2.1.4 BART machinery: a visual perspective

In our simple example, we have three covariates $X = (X_1, X_2, X_3)$ and a continuous outcome Y. We run the BART MCMC with four regression trees for 5 iterations on this dataset and at each iteration, we present the regression tree structures to illustrate how the BART machinery works as it goes through each MCMC step. When Y and X are provided to BART, BART first initializes the four regression trees to single root nodes (See "Initiation" in Figure 3). Since all four regression trees are single root nodes, the parameters initialized for these nodes would be $\mu_{ij} = \frac{\bar{Y}}{m} = \frac{\bar{Y}}{4}$.

Figure 3: Intuition of BART to iteration 3 of the MCMC steps within BART with $m = 4$.



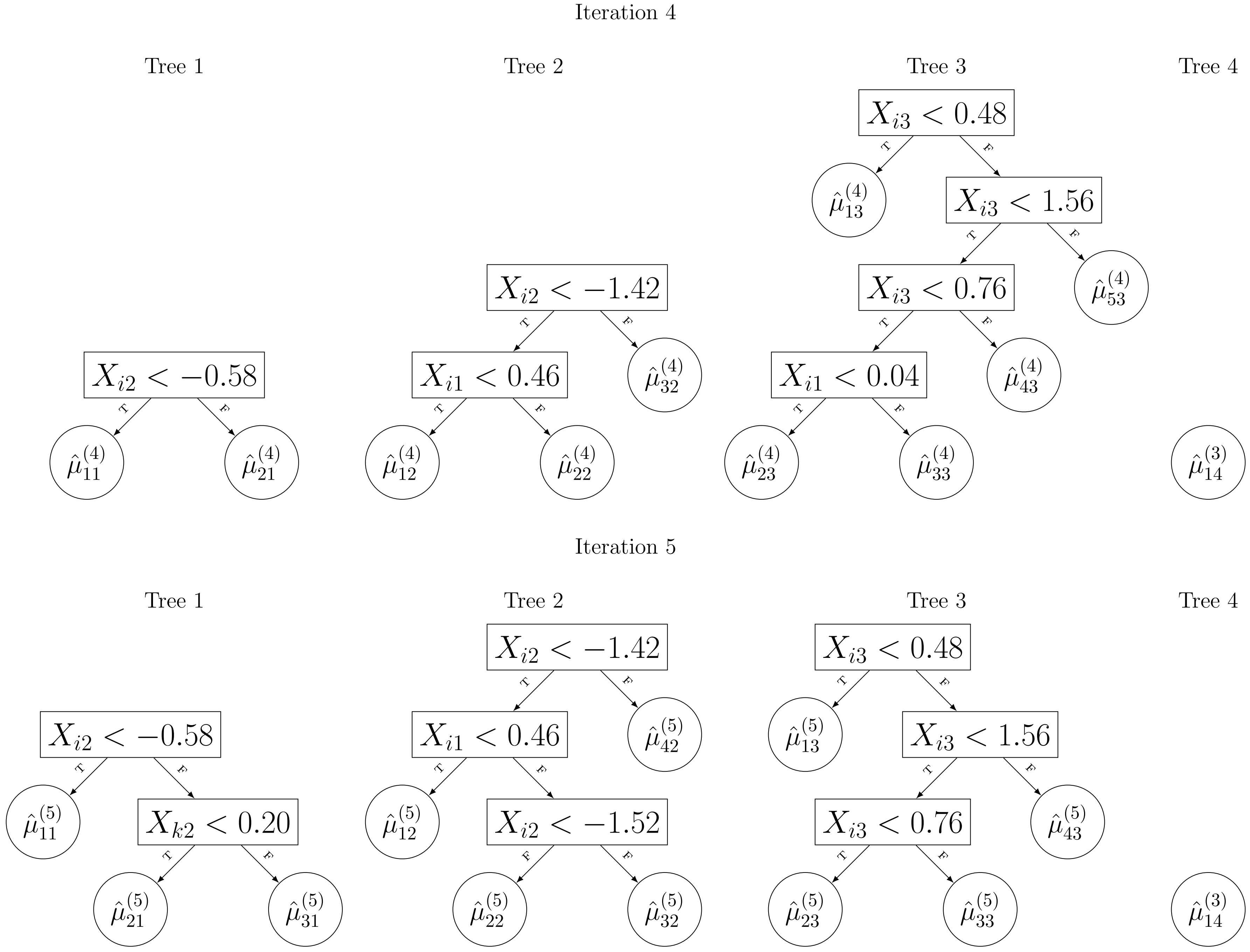
Initiation

Iteration 1

Iteration 2

Iteration 3

Figure 4: Iterations 4 and 5 of the MCMC steps within BART with $m = 4$.



With the initializations in place, BART starts to draw the tree structures for each regression tree in the first MCMC iteration. Without loss of generality, let us start with determining (T_1, M_1), the first regression tree. This is possible because the ordering of regression tree calculation does not matter. We first calculate $R_1 = Y - \sum_{j \neq 1} g(X, T_j, M_j) = Y - [g(X, T_2, M_2) + g(X, T_3, M_3) + g(X, T_4, M_4)] = Y - 3 \times \frac{\bar{Y}}{4}$. Then a MH algorithm is used to determine the posterior draw of the tree structure, T_1 for this iteration. The basic idea of MH is to propose a new tree structure from T_1, call this T_1^*, and then calculate the probability of whether T_1^* should be accepted, taking into consideration: $R_1|T_1^*$ (the likelihood of the residual given the new tree structure), $R_1|T_1$ (the likelihood of the residual given the previous tree structure), the probability of observing T_1^*, the probability of observing T_1, the probability of moving from T_1^* to T_1, and the probability of moving from T_1 to T_1^*. We describe the different types of moves from T_1 to T_1^* in detail in the next subsection. If T_1^* is accepted, T_1 is updated to become T_1^* i.e. $T_1 = T_1^*$. Else, nothing would be changed for T_1. From Figure 3, we can see that T_1^* was not accepted in the first iteration so the tree structure remains as a single root node. The algorithm then updates M_1 based on the new updated regression structure for T_1 and moves on to determine (T_2, M_2).

To determine (T_2, M_2), again the algorithm calculates $R_2 = Y - \sum_{j \neq 2} g(X, T_j, M_j) = Y - [g(X, T_1, M_1) + g(X, T_3, M_3) + g(X, T_4, M_4)] = Y - [\hat{\mu}_{11}^{(1)} + 2 \times \frac{\bar{Y}}{4}]$, where $\hat{\mu}_{11}^{(1)}$ is the updated parameter for regression tree 1. Similarly, MH is used to propose a new T_2^* and R_2 is used to calculate the acceptance probability to decide whether T_2^* should be accepted. Again, we see from Figure 3 that T_2^* was not accepted and hence a single parameter $\hat{\mu}_{12}^{(1)}$, drawn from $M_2|T_2, R_2, \sigma$, is used for $g(X, T_2, M_2)$. For (T_3, M_3), the MH iteration result is more interesting because the newly proposed T_3^* was accepted and we can see from Figure

3 that a new tree structure was used for T_3 in Iteration 1. As a result, when calculating R_4, this becomes $R_4 = Y - [\hat{\mu}_{11}^{(1)} + \hat{\mu}_{21}^{(1)} + \hat{\mu}_{13}^{(1)} I\{X_3 < 0.48\} + \hat{\mu}_{23}^{(1)} I\{X_3 \geq 0.48\} + \hat{\mu}_{14}^{(1)}]$. T_4^* was not accepted and a single node T_4 was updated as the tree structure for (T_4, M_4). Once the regression tree draws are complete, the BART then proceeds to draw the rest of the parameters in the BART model. More details in the next subsection.

Figures 3 and 4 give the full iterations from initiation to iteration 5. From these figures we can see how the four regression trees grow and change from one iteration of the MCMC to another. This iterative process runs for a burn-in period (typically 100 to 1000 iterations), before those draws are discarded, and then run for as long as needed to obtain a sufficient number of draws from the posterior distribution of $\sum_{j=1}^{m} g(X, T_j, M_j)$. After any full iteration in the MCMC algorithm, we have a full set of trees. We can therefore obtain a predicted value of Y for any X of interest (simply by summing the terminal node μ's). By obtaining predictions across many iterations, we also can easily obtain a 95% prediction interval. Another point to note is how shallow the regression trees are in Figures 3 and 4 with a maximum depth of 3. This is because the regression trees are heavily penalized (via the prior) to reduce the likelihood for a single tree to grow very deep. This concept is borrowed from the idea that many weak models combined together performs much better than utilizing a very strong model which requires careful tweaking in order for the model to perform well.

2.1.5 A rigorous perspective on the BART algorithm

Now that we have a visual understanding of how the BART algorithm works, we shall give a more rigorous explanation of BART. First, we start with the prior distributions for BART.

The prior distribution for Equation (1) is $P(T_1, M_1, \ldots, T_m, M_m, \sigma)$. The usual prior specification is that $\{(T_1, M_1), \ldots, (T_m, M_m)\}$ and σ are independent and that $(T_1, M_1), \ldots, (T_m, M_m)$ are independent of each other. Then the prior distribution can be written as

$$\begin{aligned} P(T_1, M_1, \ldots, T_m, M_m, \sigma) &= P(T_1, M_1, \ldots, T_m, M_m)P(\sigma) \\ &= [\prod_j^m P(T_j, M_j)]P(\sigma) \\ &= [\prod_j^m P(M_j|T_j)P(T_j)]P(\sigma) \\ &= [\prod_j^m \{\prod_k^{b_j} P(\mu_{kj}|T_j)\}P(T_j)]P(\sigma). \end{aligned} \tag{2}$$

For the third to fourth line in Equation (2), recall that $M_j = \{\mu_{1j}, \ldots, \mu_{b_j j}\}$ is the vector of terminal node parameters associated with T_j and each node parameter μ_{kj} is usually assumed to be independent of each other. Equation (2) implies that we need to set distributions for the priors $\mu_{kj}|T_j$, σ, and T_j. The priors for $\mu_{kj}|T_j$ and σ are usually given as $\mu_{kj}|T_j \sim N(\mu_\mu, \sigma_\mu^2)$ and $\sigma^2 \sim IG(\frac{\nu}{2}, \frac{\nu\lambda}{2})$ respectively, where $IG(\alpha, \beta)$ is the inverse gamma distribution with shape parameter α and rate parameter β.

The prior for $P(T_j)$ is more interesting and can be specified using three aspects:

1. The probability that a node at depth $d = 0, 1, \ldots$ would split, which is given by $\frac{\alpha}{(1+d)^\beta}$. The parameter $\alpha \in \{0, 1\}$ controls how likely a node would split, with larger values increasing the likelihood of a split. The number of terminal nodes is controlled by parameter $\beta > 0$, with larger values of β reducing the number of terminal nodes. This aspect is important as this is the penalizing feature of BART which prevents BART from overfitting and allowing convergence of BART to the target function $f(X)$ (Ročková and Saha, 2018). As mentioned in the previous subsection, this aspect also

allows many shallow (weak) regression trees to be fit and eventually summed together to obtain a stronger model.

2. The distribution used to select the covariate to split upon in an internal node. The default suggested distribution is the uniform distribution. Recent work (Ročková and van der Pas, 2017; Linero, 2018) have argued that the uniform distribution does not promote variable selection and should be replaced if variable selection is desired.

3. The distribution used to select the cutoff point in an internal node once the covariate is selected. The default suggested distribution is the uniform distribution.

The setting of the hyper-parameters for the BART priors is rather technical so we refer interested readers to our Appendix for how this can be done.

The prior distribution would induce the posterior distribution

$$\begin{aligned} P[(T_1, M_1), \ldots, (T_m, M_m), \sigma | Y] &\propto P(Y | (T_1, M_1), \ldots, (T_m, M_m), \sigma) \\ &\times P((T_1, M_1), \ldots, (T_m, M_m), \sigma) \end{aligned}$$

which can be simplified into two major posterior draws using Gibbs sampling. First, draw m successive

$$P[(T_j, M_j) | T_{(j)}, M_{(j)}, Y, \sigma] \tag{3}$$

for $j = 1, \ldots, m$, where $T_{(j)}$ and $M_{(j)}$ consist of all the tree structures and terminal nodes except for the j^{th} tree structure and terminal node; then, draw

$$P[\sigma | (T_1, M_1), \ldots, (T_m, M_m), Y] \tag{4}$$

from $IG(\frac{\nu+n}{2}, \frac{\nu\lambda + \sum_{i=1}^{n}(Y_i - \sum_{j=1}^{m} g(X_i, T_j, M_j))^2}{2})$.

To obtain a draw from (3), note that this distribution depends on $(T_{(j)}, M_{(j)}, Y, \sigma)$ through

$$R_j = Y - \sum_{w \neq j} g(X, T_w, M_w), \tag{5}$$

the residuals of the $m-1$ regression sum of trees fit excluding the j^{th} tree (Recall our visual example in the previous subsection). Thus (3) is equivalent to the posterior draw from a single regression tree $R_{ij} = g(X_i, T_j, M_j) + \varepsilon_i$ or

$$P[(T_j, M_j)|R_j, \sigma]. \tag{6}$$

We can obtain a draw from (6) by first integrating out M_j to obtain $P(T_j|R_j, \sigma)$. This is possible since a conjugate Normal prior on μ_{kj} was employed. We draw $P(T_j|R_j, \sigma)$ using a MH algorithm where first, we generate a candidate tree T_j^* for the j^{th} tree with probability distribution $q(T_j, T_j^*)$ and then, we accept T_j^* with probability

$$\alpha(T_j, T_j^*) = \min\left\{1, \frac{q(T_j^*, T_j)}{q(T_j, T_j^*)} \frac{P(R_j|X, T_j^*, M_j)}{P(R_j|X, T_j, M_j)} \frac{P(T_j^*)}{P(T_j)}\right\}. \tag{7}$$

$\frac{q(T_j^*, T_j)}{q(T_j, T_j^*)}$ is the ratio of the probability of how the previous tree moves to the new tree against the probability of how the new tree moves to the previous tree, $\frac{P(R_j|X, T_j^*, M_j)}{P(R_j|X, T_j, M_j)}$ is the likelihood ratio of the new tree against the previous tree, and $\frac{P(T_j^*)}{P(T_j)}$ is the ratio of the probability of the new tree against the previous tree.

A new tree T_j^* can be proposed given the previous tree T_j using four local steps: (i) grow, where a terminal node is split into two new child nodes; (ii) prune, where two terminal child nodes immediately under the same non-terminal node are combined together such that their parent non-terminal node becomes a terminal node; (iii) swap, the splitting criteria of two non-terminal nodes are swapped; (iv) change, the splitting criteria of a single non-terminal node is changed. Once we have the draw of $P(T_j|R_j, \sigma)$, we then draw $P(\mu_{kj}|T_j, R_j, \sigma) \sim$

$N(\frac{\sigma_\mu^2 \sum_{k=1}^{n_k} R_{kj}}{n_k\sigma_\mu^2+\sigma^2}, \frac{\sigma^2\sigma_\mu^2}{n_k\sigma_\mu^2+\sigma^2})$, where R_{kj} is the subset of elements in R_j allocated to the terminal node parameter μ_{kj} and n_k is the number of R_{kj}s allocated to μ_{kj}. We derive $P(\mu_{kj}|T_j, R_j, \sigma)$, Equation (4), and Equation (7) for the grow and prune steps as an example in our Appendix.

2.2 Binary outcomes

For binary outcomes, BART can be extended using a probit model. Specifically,

$$P[Y_i = 1|X_i, (T_1, M_1), \ldots, (T_m, M_m)] = \Phi[\sum_{j=1}^{m} g(X_i; T_j, M_j)]$$

where $\Phi[.]$ is the cumulative distribution function of a standard normal distribution and i indexes the subjects $i = 1, \ldots, n$. With such a setup, only priors for $(T_1, M_1), \ldots, (T_m, M_m)$ are needed. The same decomposition in Equation (2) without σ can be employed and the similar prior specifications for $\mu_{kj}|T_j$ and T_j can be used. The setup of the hyper-parameters are slightly different from that of continuous outcomes and we describe this in the Appendix.

To estimate the posterior distribution, data augmentation (Albert and Chib, 1993) can be used. In essence, we first draw a latent variable $Z = \{Z_1, \ldots, Z_n\}$ as follows:

$$Z_i \sim N_{(-\infty,0)}[\sum_{j=1}^{m} g(X_i; T_j, M_j), 1] \quad \text{if } Y_i = 0,$$
$$Z_i \sim N_{(0,\infty)}[\sum_{j=1}^{m} g(X_i; T_j, M_j), 1] \quad \text{if } Y_i = 1$$

where $N_{(a,b)}[\mu, \sigma^2]$ is a truncated normal distribution with mean μ and variance σ^2 truncated at (a, b). Next, we can treat Z as the continuous outcome for a BART model with

$$Z = \sum_{j=1}^{m} g(X; T_j, M_j) + \varepsilon \tag{8}$$

where $\varepsilon \sim N(0, 1)$ because we employed a probit link. The usual posterior estimation for a continuous outcome BART with $\sigma \equiv 1$ can now be employed on Equation (8) for one iteration in the MCMC. The updated $\sum_{j=1}^{m} g(X; T_j, M_j)$ can then be used to draw a new Z and this new Z can be used to draw another iteration of $\sum_{j=1}^{m} g(X; T_j, M_j)$. The process can then be repeated till convergence.

3 Illustrating the performance for BART

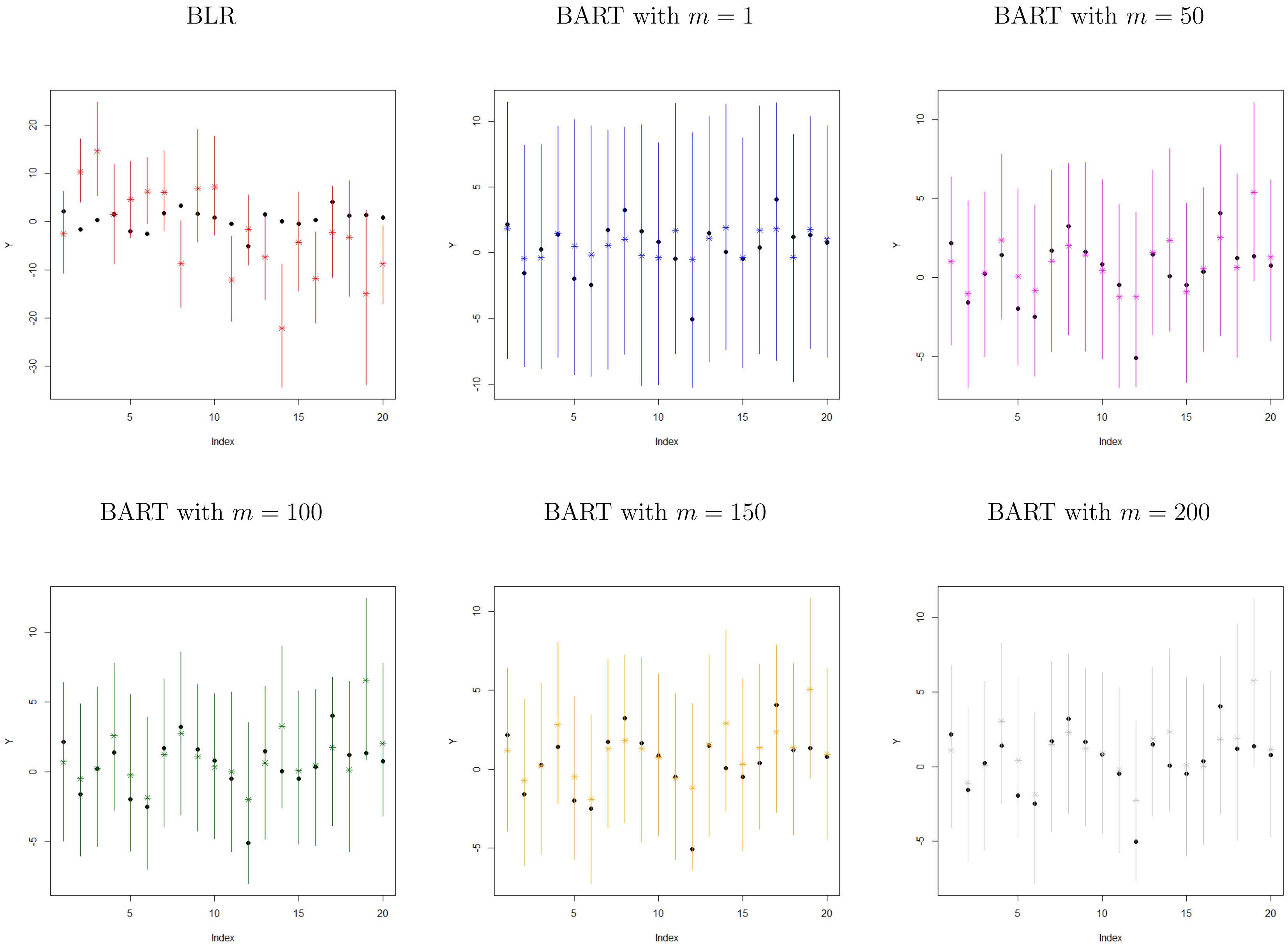
3.1 Posterior performance via synthetic data

We generated a synthetic data set with $p = 3$, $n = 1,000$ and, the true model for Y_i is

$$\begin{aligned} Y_i = {} & 0.5 + 0.1X_{i1} + 0.3X_{i2}^2 + 0.7\sin(X_{i3}) + 0.2X_{i1}X_{i2} + 0.9\sqrt{|X_{i1}X_{i3}|} \\ & + 0.4\exp(X_{i2}X_{i3}) + 0.8\log(|X_{i1}X_{i2}X_{i3}|) + \varepsilon_i \end{aligned}$$

with $X_{ip} \sim N(0, 1)$ and $\varepsilon_i \sim N(0, 2)$. The goal is to demonstrate that BART can predict Y's effectively even in complex, non-linear models, and also properly accounts for prediction uncertainty compared to a parametric BLR model. To this end, we randomly selected 880 samples as the training set and then use the remaining 20 samples as the testing set. We also varied the number of trees used by BART to illustrate how varying m affects the performance of BART. We plotted the point estimate and 95% credible interval of the 20 randomly selected testing data points and compared them with their true values in Figure 5. The codes to implement this simulation will be made available on https://github.com/yaoyuanvincent.

Figure 5: Posterior mean and 95% credible interval of Bayesian linear regression (BLR) and BART with $m = 1, 50, 100, 150, 200$ for 20 randomly selected testing set outcomes. $n = 1,000$, Black=true value, colored=model estimates.



BLR

BART with $m = 1$

BART with $m = 50$

BART with $m = 100$

BART with $m = 150$

BART with $m = 200$

We can see from Figure 5 that most of the point estimates of BLR were far away from their true values and many of the true values were not covered by the 95% credible interval. For BART with a single tree, although the true values were mostly covered by the 95% credible interval, the point estimates were far from their true values. When we increased the number of trees to 50 in BART, we see a significant improvement in terms of bias (closeness to the true values) compared to both BLR and BART with $m = 1$. In addition, we see a narrowing of the 95% intervals. We see that as we increase the number of trees, the point estimate and 95% intervals stabilize. In other words, we might see a big difference between $m = 1$ and $m = 50$, and virtually no difference between $m = 200$ and $m = 20,000$. In practice, the idea is to choose a large enough value for m (default is often 200) so that it very well approximates the results that would have been obtained if more trees were used. One way to determine an m that is sufficiently large is with cross validation (Chipman et al., 2010).

3.2 Predicting the Standardized Hospitalization Ratio from the 2013 Centers for Medicare and Medicaid Services Dialysis Facility Compare dataset

We next present an example to demonstrate how BART can be applied to a dataset to improve prediction over the usual multiple linear regression model. The 2013 Centers for Medicare and Medicaid Services Dialysis Facility Compare dataset contains information regarding 105 quality measures and 21 facility characteristics of all dialysis facilities in the US, including US territories. This dataset is available publicly at https://data.medicare.gov/data/archives/dialysis-facility-compare. The codes and data to implement this analysis will be made available on https://github.com/yaoyuanvincent. We are interested in finding a model that can better predict the standardized hospitalization ratio (SHR). This quantity is important because a large portion of dialysis cost for End Stage Renal Disease (ESRD) patients can be attributed to patient hospitalizations.

Table 2: Descriptive statistics of dialysis facility characteristics and quality measures (n=5,774).

Parameters (% missing)	Mean (s.d)/Frequency (%)	Parameters (% missing)	Mean (s.d)/Frequency (%)
Arterial Venous Fistula (3)	63.27 (11.22)	Number of stations	18.18 (8.27)
Avg. Hemoglobin<10.0 g/dL (5)	12.86 (10.32)	Serum P. 3.5-4.5mg/dL** (2)	28.52 (5.13)
Chain name:		Shift after 5pm?	
Davita	1,812 (31)	Yes	1,097 (19)
FMC	1,760 (30)	No	4,677 (81)
Independent	820 (14)	SHR	1.00 (0.31)
Medium	740 (13)	SMR (2)	1.02 (0.29)
Small	642 (12)	STR (7)	1.01 (0.54)
Patient volume*	100.07 (60.93)	Type:	
Facility Age (years)	14.47 (9.81)	All (HD, Home HD, & PD)	1,443 (25)
For profit?		HD & PD	1,897 (33)
Yes	4,967 (86)	HD & Home HD	103 (2)
No	806 (14)	HD alone	2,331 (40)
HD≥1.2 Kt/V (4)	88.52 (9.85)	URR≥65% (7)	98.77 (3.04)
Hypercalcemia (3)	2.37 (3.20)	Vas. Catheter>90 days (3)	10.74 (6.66)

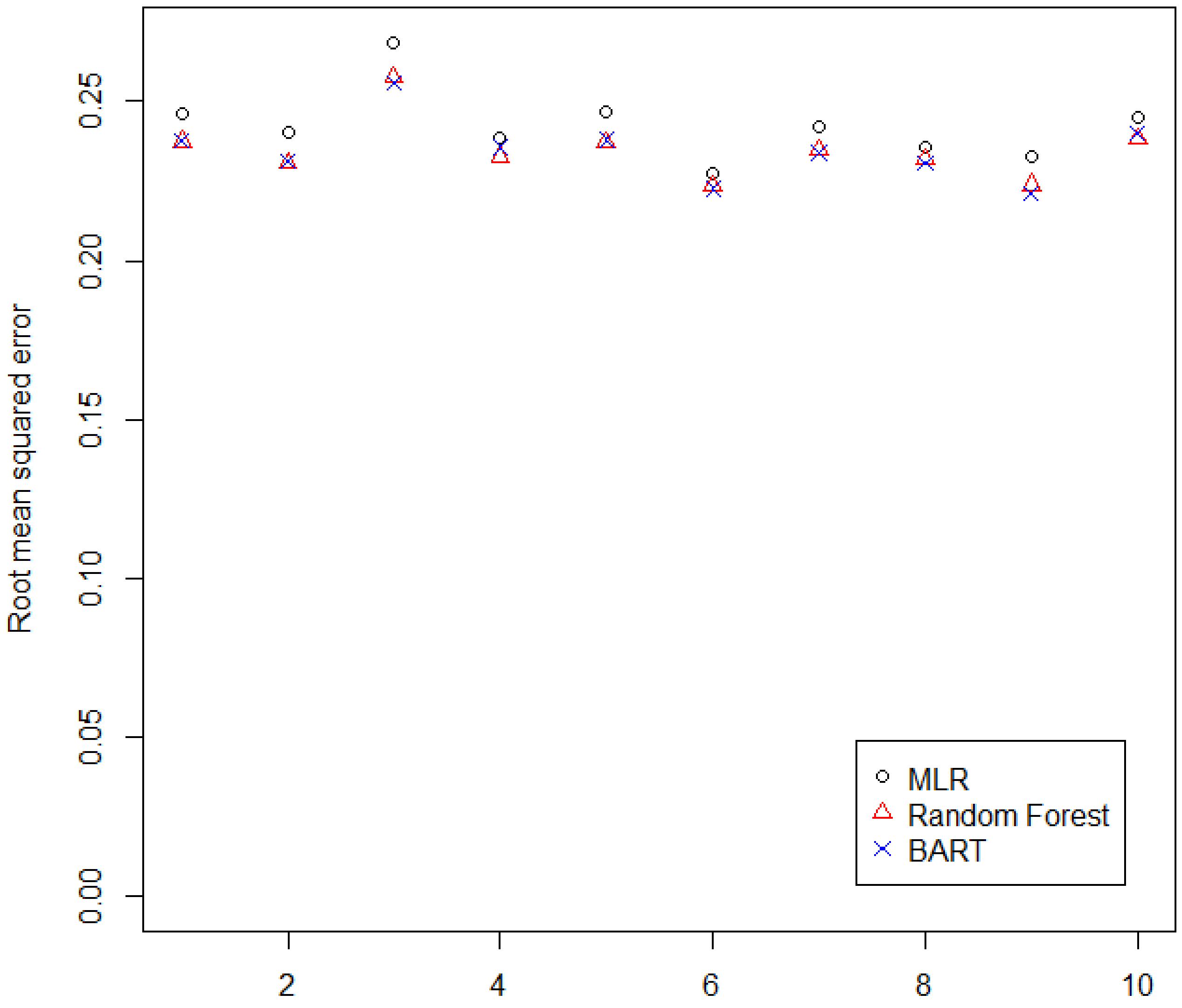
*Estimated. **Normal range.

Table 2 shows some descriptive statistics for this dataset. SHR was adjusted for a patients age, sex, duration of ESRD, comorbidities, and body mass index at ESRD incidence.

We removed 463 facilities (7%) with missing SHR values because of small patient number. We also removed peritoneal dialysis (PD) removal greater than 1.7 Kt/V because of the high proportion of missingness (80%). We combined pediatric hemodialysis (HD) removal greater than 1.2 Kt/V with adult HD removal greater than 1.2 Kt/V because most facilities (92%) do not provide pediatric HD. We re-categorized the chain names to "Davita," "Fresenius Medical Care (FMC)," "Independent," "Medium," and "Small." "Medium" consists of chains with 100-500 facilities while "Small" are chains with less than 100 facilities. To estimate patient volume, we used the maximum of the number of patients reported by each quality measure group: Urea Reduction Ratio (URR), HD, PD, Hemoglobin (HGB), Vascular Access, SHR, SMR, STR, Hypercalcemia (HCAL), and Serum phosphorus (SP). We also logarithm-transformed (log) SHR, SMR, and STR so that the theoretical range for these log standardized measures will be $-\infty$ to ∞.

For our analysis, we used the log-transformed SHR as the outcome and the variables in Table 2 as the predictors. We used the root mean squared error (RMSE) of a 10-fold cross-validation to compare the prediction performance from multiple linear regression (MLR), Random Forest (RF), and BART. For RF and BART, we used the default settings from the *R* packages *randomForest* and *BayesTree* respectively. The 10 RMSEs produced by each method from the 10-fold cross validation is provided in Figure 6. It is clear from this figure that BART and RF produce very similar prediction performances and is better compared to MLR. The mean of these 10 values also suggested a similar picture with MLR producing a mean of 0.24 while RF and BART produced a mean of 0.23.

Figure 6: Root mean squared error for the 10-fold cross-validation of multiple linear regression (MLR), random forest, and Bayesian additive regression trees of log transformed standardized hospitalization ratio (SHR)



3.3 Predicting left turn stops at an intersection

We next present another example where BART showed improvement in the prediction performance of a binary outcome. In Tan et al. (2017), the authors were interested in predicting whether a human driven vehicle would stop at an intersection before making a left turn. Left turns are important in countries with right side driving because most vehicle conflicts including crashes at intersections occur during left turns. Knowledge of whether a human driven vehicle would stop before executing a left turn would help driverless vehicles be better able to make decisions at an intersection. More details about this dataset can be found in Tan et al. (2017). In brief, the data comes from the Integrated Vehicle Based Safety System (IVBSS) study conducted by Sayer et al. (2011). This study collected driving data from 108 licensed drivers in Michigan between April 2009 and April 2010. Each driver drove one of the sixteen research vehicles fitted with various recording devices to capture the vehicle dynamics while the subject is driving on public roads for 12 days. In particular, Tan et al. (2017) focused on the vehicle speeds of all left turns at an intersection starting from 100 meters away from the center of an intersection to the center of an intersection. They then transformed the vehicle speed time series to a distance series. Having the vehicle speed at each distance as the columns and each turn as the rows, they then performed principal components analysis (PCA) on these vehicle speeds using moving windows of 6 meters from 94 meters away to 1 meter away from the center of an intersection. This implies that at each meter, a PCA analysis was conducted using 6 meters of vehicle speeds, i.e. at 94 meters, 94 to 100 meters away was used, at 93 meters, 93 to 99 meters away was used and so on until 1 meter away. They used a 6 meter moving window because they found that longer windows did not improve prediction performance and a 6 meter moving window provided the best prediction performance. At each meter, the first three principal components (PCs) from the corresponding 6 meter moving window PCAs were then used to determine the prediction model with the outcome as whether the vehicle stopped (vehicle speed $< 1m/s$) in the future with stopped coded as 1 and not stop coded as 0. Only the first three PCs were used because these three PCs explained nearly 99% of the variance in the 6 meter moving window distance series of vehicle speeds as well as provided the best prediction performance. This setup resulted in 94 models corresponding with 94 datasets for each meter. In order to keep our presentation concise, we focus on the dataset halfway through the turn maneuver (50 meters away from the center of an intersection) which is made up of the first 3 PCs of the PCA on vehicle speed from 50 to 56 meters and the outcome of whether the vehicle stopped in the future from 49 meters to the center of an intersection. We ran a 10-fold cross-validation on this dataset and compared the binary prediction results of logistic regression, RF, and BART. Since the outcome of interest for this dataset was binary, we used the area under the receiver operating curve (AUC) to determine the prediction performance instead of the RMSE, which is more suited for continuous outcomes.

Figure 7: Area under the receiver operating characteristic curve for the 10-fold cross-validation of logistic regression, random forest, and Bayesian additive regression trees of left turn stop probabilities at an intersection.

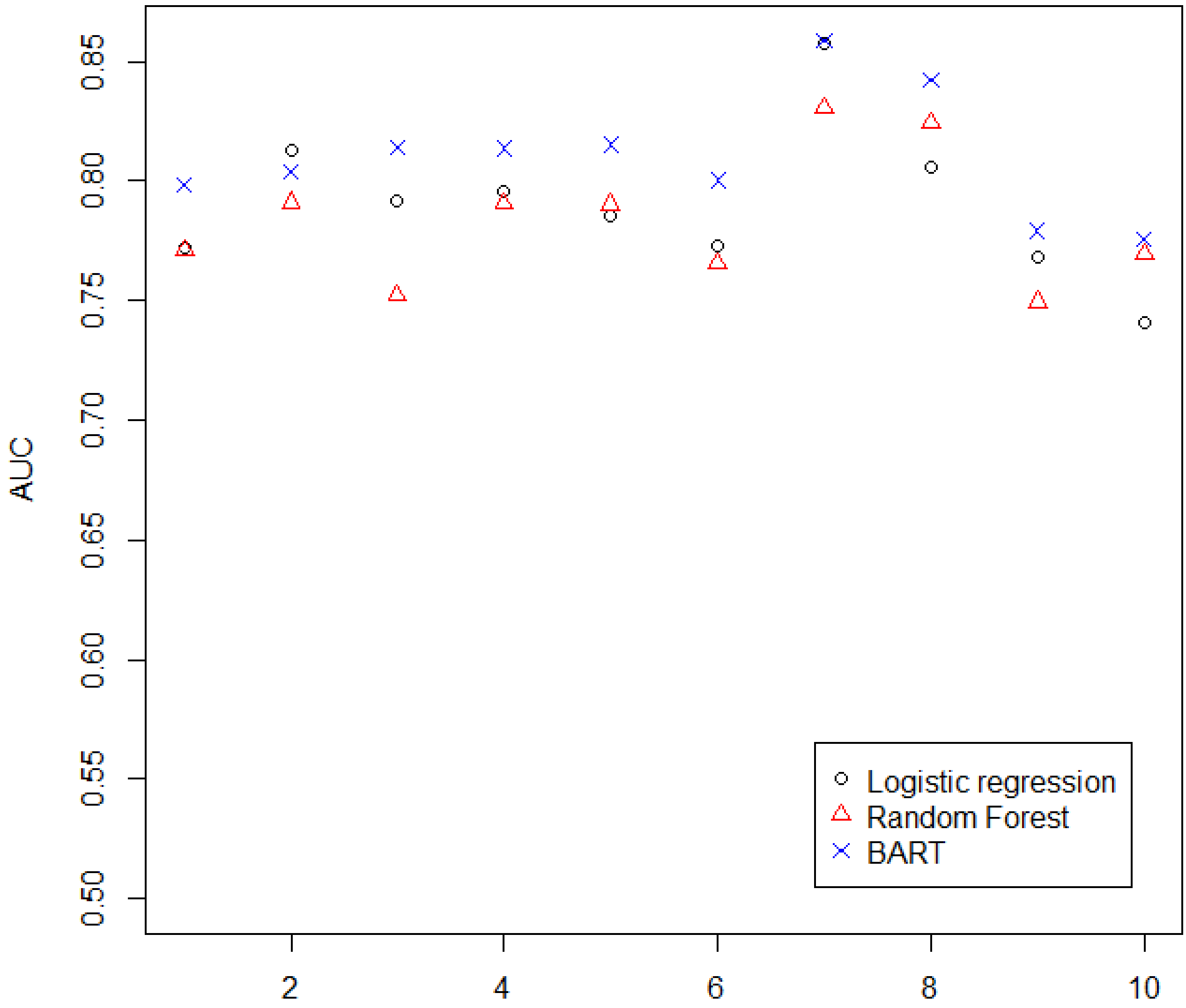


Figure 7 shows the results of the 10 AUCs produced by each method from the 10-fold cross validation. It is clear that BART performed extremely well in predicting whether the human-driven vehicle would stop in the future at an intersection before making a left turn, much better than either logistic regression or RF. This is also evident from the mean of the 10 cross validation AUC values produced by each method. BART produced a mean of 0.81 compared to 0.79 from logistic regression and 0.78 from RF.

4 General BART model

Recently researchers have extended or generalized BART to a wider variety of settings, including clustered data, spatial data, semi-parametric models, and to situations where more flexible distributions for the error term is needed. Here we describe a more general BART framework that includes all of these cases and more. An important feature of this general BART model is they can be fitted without very extensive re-derivation for the MCMC draws of the regression trees described in Section 2. That is, the MCMC algorithm we described previously only needs small adjustments to handle this more general setting.

To set up our General BART model, suppose once again that we have a continuous outcome Y and p covariates $X = \{X_1, \dots, X_p\}$. Suppose also that we have another set of q covariates $W = \{W_1, \dots, W_q\}$ such that no two columns in X and W are the same. Then, we can extend Equation (1) as follows:

$$Y = G[X, (T, M)] + H(W, \Theta) + \varepsilon \tag{9}$$

where $G[X, (T, M)] = \sum_{j=1}^{m} g(X; T_j, M_j)$ with $(T, M) = [(T_1, M_1), \dots, (T_m, M_m)]$, $H(.)$ is a function that works on W using parameter Θ, and $\varepsilon \sim G(\Sigma)$ can be any distribution with parameter Σ.

Assuming that (T, M), Θ, and Σ are independent, the prior distribution for Equation (9) is $P(T, M)P(\Theta)P(\Sigma)$. Assuming again that the (T_j, M_j)'s within (T, M) are independent of each other, $P(T, M)$ can be decomposed into $\prod_j^m\{\prod_{k=1}^{b_j} P(\mu_{kj}|T_j)\}P(T_j)$. The priors needed are thus $P(\mu_{kj}|T_j)$, $P(T_j)$, $P(\Theta)$, and $P(\Sigma)$. Note that it is possible to model Θ and Σ jointly so that the prior distribution becomes $\prod_j^m\{\prod_{k=1}^{b_j} P(\mu_{kj}|T_j)\}P(T_j)P(\Theta, \Sigma)$ instead. We shall see this in Example 4.4.

To obtain the posterior distribution of $P[(T, M), \Theta, \Sigma|Y, X, W]$, Gibbs sampling can be used. For $P[(T, M)|\Theta, \Sigma, Y, X, W]$, this can be seen as drawing from the following model

$$\tilde{Y} = G[X, (T, M)] + \varepsilon \tag{10}$$

where $\tilde{Y} = Y - H(W, \Theta)$ which is just a BART model with a modified outcome $\tilde{Y}$. Hence, the BART algorithm presented in Section 2 can be used to draw (T, M), the regression trees. Similarly, $P[\Theta|(T, M), \Sigma, Y, X, W]$ can be obtained by drawing from the model

$$Y' = H(W, \Theta) + \varepsilon \tag{11}$$

where $Y' = Y - G[X, (T, M)]$. This posterior draw depends on the function $H(.)$ being used as well as the prior distribution specified for Θ. As there are many possibilities where we can set up $H(.)$ and Θ, we shall not discuss the specifics here. The examples we present in the subsequent subsections will highlight a few of these possibilities we have seen in the literature thus far. Finally, drawing from $P[\Sigma|(T, M), \Theta, Y, X, W]$ is just drawing from the model

$$R = \varepsilon \tag{12}$$

where $R = Y - G[X, (T, M)] - H(W, \Theta)$. Again, many possibilities are available for setting up the prior distribution for Σ and hence the distributional assumption for ε. The default is usually $e_i \sim N(0, \sigma^2)$ where $\Sigma = \sigma^2 \sim IG(\frac{\nu}{2}, \frac{\nu\lambda}{2})$. Example 4.4 shows an alternative distributional assumption for ε and hence, Σ. Iterating through the above three Gibbs steps will give us the posterior draw of $P[(T, M), \Theta, \Sigma | Y, X, W]$.

For binary outcomes, the probit link can once again be used where

$$P[Y_i = 1 | X, (T_1, M_1), \ldots, (T_m, M_m)] = \Phi[G\{X, (T, M)\} + H(W, \Theta)].$$

Under this framework, we will only need priors for $P(T, M)$ and $P(\Theta)$. $P(T, M)$ can be decomposed once again into $\prod_j^m \{\prod_{k=1}^{b_j} P(\mu_{kj} | T_j)\} P(T_j)$ if we are willing to assume that the m trees are independent of one another, and data augmentation (Albert and Chib, 1996) can be used obtain the posterior distribution. We can draw

$$Z_i \sim N_{(-\infty,0)}[G[X, (T, M)] + H(W, \Theta), 1] \quad \text{if } Y_i = 0,$$

$$Z_i \sim N_{(0,\infty)}[G[X, (T, M)] + H(W, \Theta), 1] \quad \text{if } Y_i = 1$$

and then treat Z as the outcome for the model in Equation (9). This would imply that $\varepsilon_k \sim N(0, 1)$ in Equation (9) and we can apply the Gibbs sampling procedure we described for continuous outcomes using Z instead of Y with $\Sigma = \sigma = 1$. Iterating through the latent draws and Gibbs steps will produce the posterior distribution that we require.

With the general framework and model for BART in place, we are now equipped to consider how Zeldow et al. (2018), Tan et al. (2018b), (Zhang et al., 2007), and George et al. (2018) extended BART to solve their research problems in the next fours subsections.

4.1 Semiparametric BART

The semiparametric BART was first presented by Zeldow et al. (2018). Their idea was to have a model where the effects of interest are modeled linearly with at most simple interactions to keep the associated parameters interpretable while having the nuisance or confounder variables be modeled as flexibly as possible. In its simplest form, we have under the framework of Equation (9) that

$$H(W,\Theta) = \theta_0 + \theta_1 W_1 + \ldots + \theta_q W_q$$

where $W = \{W_1, \ldots, W_q\}$, $\Theta = \{\theta_0, \ldots, \theta_q\}$, and $\varepsilon_i \sim N(0, \sigma^2)$ with $\Sigma = \sigma$. Prior distributions for $\mu_{kj}|T_j$, T_j, and σ^2 follow the usual distributions we use for BART while $\Theta \sim MVN(\beta, \Omega)$, possibly. Posterior estimation follows the procedure we described in Section 4 using Gibbs Sampling. For Equations (10) and (12), since they suggested using the default BART priors, the usual BART mechanisms can be applied to obtain the posterior draws. For Equation (11), $\Theta \sim MVN(\beta, \Omega)$ implies that we can treat this as the usual BLR and standard Bayesian methods could be used to obtain the posterior draw for Θ. The framework for binary outcomes follows easily using the data augmentation step we describe in Section 4.

4.2 Random intercept BART for correlated outcomes

Random intercept BART (riBART) was proposed by Tan et al. (2018b) as a method to handle correlated continuous or binary outcomes with correlated binary outcomes as the main focus.

Under the framework of Equation (9), we have $H(W, \Theta) = Wa$, where $\Theta = (a, \tau)$ and

$$W = \begin{bmatrix} 1 & 0 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 1 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \dots & 1 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & 1 \end{bmatrix}$$

i.e. W is a matrix with 1 repeated n_1 times in the first column and 0s for the rest of the column, 0 repeated n_1 times in the second column followed by 1 repeated n_2 times and then 0s for the rest of the columns, and so on until for the last column we have 0 repeated $\sum_{l=1}^{L-1} n_l$ times and then 1 repeated n_L times, $a = \{a_1, \dots, a_L\}$ with $a_l|\tau^2 \sim N(0, \tau^2)$. l indexes the subject. Once again, $\varepsilon \sim N(0, \sigma^2)$ with $\Sigma = \sigma$ and the usual BART priors for σ, $\mu_{kj}|T_j$, and T_j can be employed. a_l and ε are assumed to be independent. A simple prior of $\tau^2 \sim IG(1, 1)$ could be used although more robust or complicated priors are possible. Posterior estimation and binary outcomes then follow the procedure described in Section 4 easily.

4.3 Spatial BART for a statistical matched problem

The Spatial BART approach of Zhang et al. (2007) was proposed to handle statistical matched problems (Rässler, 2002) that occur in surveys. In statistical matched problems, inference is desired for the relationship between two different variables collected by two different datasets on the same subject. For example, survey A may collect information on income but survey B collects information on blood pressure. Both surveys A and B contain subjects that overlap. The relationship between income and blood pressure is then desired. To solve this problem, Spatial BART essentially uses a framework similar to that of riBART described in Example 4.2 with a more complicated prior distribution for Θ. The specification for W and a is the same but the distribution placed on a is instead the conditionally autoregressive prior which can be specified as

$$a|\rho, \delta^2 \sim N(0, \delta^2(H - \rho C)^{-1}) \tag{13}$$

where $C = c_{il}$ is a $I \times I$ adjacency matrix, $l = 1, \ldots, I$, with $c_{il} = 1$ if group i and group l are (spatial) neighbors for $i \neq l$; $c_{il} = 0$ otherwise; and $c_{il} = 0$ if $i = l$. H is a diagonal $I \times I$ matrix with diagonals $h_i = \sum_{l=1}^{I} c_{il}$, ρ is a parameter with range $(-1, 1)$, and δ^2 is the variance component for Equation (13). ρ and δ^2 are hyperparameters that is prespecified. Finally, $\varepsilon \sim N(0, \sigma^2)$ and Equation (9) is completed by placing the usual BART priors for σ, $\mu_{kj}|T_j$, and T_j. Posterior draws again follow the procedures we outlines in Section 4.

4.4 Dirichlet Process Mixture BART

The Dirichlet Process Mixture (DPM) BART was proposed by George et al. (2018) to enhance the robustness of distributional assumption for ε in Equation (1). To do this they focused on a different specification for ε by assuming that

$$\begin{aligned} \varepsilon_i &\sim N(a_i, \sigma_i^2), \\ (a_i, \sigma_i^2) &\sim D, \\ D &\sim DP(D_0, \alpha) \end{aligned}$$

where D denote a random discrete distribution and DP denotes the Dirichlet process with parameters D_0 and $\alpha > 0$. The atoms of D can be seen as iid draws from D_0. α on the other hand determines weight allocated to atom of discrete D. Higher values of α imply that the weights would be spread out among the atoms. Lower values of α imply that weights would be concentrated on only a few atoms. Although the assumption of $\varepsilon_i \sim N(a_i, \sigma_i^2)$ suggests that each subject will have their own mean and variance for the error term, the placement of a Dirichlet process on D restricts the number of unique components for (a_i, σ_i^2) to $K < n$, which ensures that this model would still be identifiable. Viewing DPMBART as a form of Equation (9), we have $H(W, \Theta) = Wa$ where W and a have the same structure as riBART and $P(\Theta, \Sigma) = (a_i, \sigma_i^2)$. Note here that we are no longer assuming that a_i and ε_i are independent unlike in some of our previous examples.

The priors for DPMBART are D_0 and α. For D_0, the commonly employed form is $P(\mu, \sigma|\nu, \lambda, \mu_0, k_0) = P(\sigma|\nu, \lambda)P(\mu|\sigma, \mu_0, k_0)$. George et al. (2018) specified their priors as

$$\sigma^2 \sim \frac{\nu\lambda}{\chi_\nu^2}; \quad \mu|\sigma \sim N(\mu_0, \frac{\sigma^2}{k_0}).$$

ν is set at 10 to make the spread of error for a single component k tighter. λ is chosen using the idea from how λ is determined in BART with the quantile set at 0.95 instead of 0.9 (See Appendix A for how λ is determined in BART). For μ_0, because DPMBART subtracts $\bar{Y}$ from Y, $\mu_0 = 0$. For k_0, the residuals of a multiple linear regression fit is used to place μ

into the range of these residuals, r. The marginal distribution of is $\mu \sim \frac{\sqrt{\lambda}}{\sqrt{k_0}} t_\nu$, where t_ν is a t distribution with ν degrees of freedom. Let k_s be the scaling for μ. Given $k_s = 10$, k_0 can be chosen by solving

$$\max |r_k| = k_s \frac{\sqrt{\lambda}}{\sqrt{k_0}}.$$

For α, the prior used by DPMBART is the same as in Section 2.5 of Rossi (2014) where the idea is to relate α to the number of unique components in (a_i, σ_i^2).

The posterior draw for DPMBART follows most of the ideas discussed in General BART where first, the idea of Equation (10) is used to draw $(T, M)|a_i, \sigma_i^2$. The slight difference is to view this as a weighted BART draw with $\varepsilon \sim N(0, w_i\sigma^2)$. The second draw, $(a_i, \sigma^2)|(T, M)$ follows Equation (11) which can be solved by using draws (a) and (b) of the algorithm in Section 1.3.3 of Escobar and West in Dey et al. (1998). The final draw is $\alpha|(a_i, \sigma^2)$. This is obtained by putting α on a grid and the using Bayes' theorem with $P(\alpha|(a_i, \sigma_i^2)) = P(\alpha|K) \propto P(K|\alpha)P(\alpha)$ where K is the number of unique (a_i, σ_i^2)'s.

5 Discussion

In this tutorial, we walked through the BART model and algorithm in detail, and presented a generalized model based on recent extensions. We believe this is important because of the growing use of BART in research applications as well as being used as a competitor model for new modeling or prediction methods. By clarifying the various components of BART, we hope that researchers will be more comfortable using BART in practice.

Despite the success of BART, there has been a growing number of papers that point out limitations of BART and propose modifications. One issue is the inability of BART to do variable selection due to the use of the uniform prior to select the covariate to be split upon in the internal nodes. One simple solution is to allow researchers to place different prior probabilities on each covariate (Kapelner and Bleich, 2016). Other solutions include using a Dirichlet Process Prior for selecting covariates (Linero, 2018) or using a spike-and-slab prior (Liu et al., 2018). Another commonly addressed issue is the computation speed of BART. Due to the many MH steps that BART require, computation speed of BART can often be slow, especially when the sample size n and/or the number of covriates p is large. One direction is to parallelize the computational steps in BART, which was proposed by Pratola et al. (2014) and Kapelner and Bleich (2016). The other direction is to improve the efficiency of the MH steps which leads to the reduction in the number of trees needed. Notable examples include Lakshminarayanan et al. (2015), where particle Gibbs sampling was used to propose the tree structure T_j's; Entezari et al. (2018), where likelihood inflated sampling was used to calculate the MH steps, and more recently He et al. (2018), where they proposed to use a different tree-growing algorithm which grows the tree from scratch (root node) at each iteration. Other less discussed issues with BART include the problem of under estimation of the uncertainty of BART caused by inefficient mixing when the true variation is small (Pratola, 2016), inability of BART to handle smooth functions (Linero and Yang, 2018), and inclusion of many spurious interactions when the number of covariates is large (Du and Linero, 2018). Finally, the posterior concentration properties of BART have also been discussed recently by Ročková and van der Pas (2017), Ročková and Saha (2018), and Linero and Yang (2018). These works provide theoretical proof of why BART has been successful in many data applications we have seen thus far.

A second component we focused on was how we can extend BART using a very simple idea without having to re-write the whole MCMC algorithm to draw the regression trees. We term this framework General BART. This framework has already been used by various authors to extend BART to semiparamteric situations where a portion of the model was desired to be linear and more interpretable, correlated outcomes, solve the statistical matching problem in survey, and improve the robustness assumption of the error term in BART. By unifying these methods under a single framework and showing how these methods are related to the General BART model, we hope to provide researchers a guide and inspiration of how to possibly extend BART to their research work where the use of the simple independent continuous or binary BART model is insufficient. For example, researchers working with longitudinal data may want a more flexible modeling portion for the random effects and hence may want to model $H(W, \Theta)$ as BART. Another possibility is to combine the ideas in Examples 4.1, 4.2, and 4.4, i.e. correlated outcomes with an interpretable linear model portion and robust error assumptions. Such are the possibilities for our proposed General BART framework.

We do note that the critical component of our General BART framework is re-writing the model in such a way that the MCMC draw of the regression trees can be done separately from the rest of the model. In situations where this is not possible, re-writing of the MCMC procedure for the regression trees may be needed. An example of this would occur if, rather than mapping the outcome to a parameter at the terminal node of a regression tree, it is mapped to a regression model. However, we feel that the general BART model is flexible enough to handle many of the extensions that might be of interest to researchers.

References

Agarwal, R., Ranjan, P., and Chipman, H. (2013). A new Bayesian ensemble of trees approach for land cover classification of satellite imagery. *Canadian Journal of Remote Sensing* **39,** 507–520.

Albert, J. and Chib, S. (1993). Bayesian Analysis of Binary and Polychotomous Response Data. *Journal of the American Statistical Association* **88,** 669–679.

Albert, J. and Chib, S. (1996). *Bayesian modeling of binary repeated measures data with application to crossover trials.* In Bayesian Biostatistics, D. A. Berry and D. K. Stangl, eds. New York: Marcel Dekker.

Bonato, V., Baladandayuthapani, V., Broom, B., Sulman, E., Aldape, K., and Do, K. (2011). Bayesian ensemble methods for survival prediction in gene expression data. *Bioinformatics* **27,** 359–367.

Chipman, H., George, E., Lemp, L., and McCulloch, R. (2010). Bayesian flexible modeling of trip durations. *Transportation Research Part B* **44,** 686–698.

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

Dey, D., Müller, P., and Sinha, D. (1998). *Practical Nonparametric and Semiparametric Bayesian Statistics.* Springer-Verlag, New York, New York.

Ding, J., Bashashati, A., Roth, A., Oloumi, A., Tse, K., Zeng, T., Haffari, G., Hirst, M., Marra, M., Condon, A., Aparicio, S., and Shah, S. (2012). Feature based classifiers for somatic mutation detection in tumour-normal paired sequencing data. *Bioinformatics* **28,** 167–175.

Dorie, V., Hill, J., Shalit, U., Scott, M., and Cervone, D. (2017). Automated versus do-it-yourself methods for causal inference: lessons learned from a data analysis competition. *arXiv* page 1707.02641.

Du, J. and Linero, A. (2018). Interaction Detection with Bayesian Decision Tree Ensembles. *arXiv* page 1809.08524.

Entezari, R., Craiu, R., and Rosenthal, J. (2018). Likelihood inflating sampling algorithm. *The Canadian Journal of Statistics* **46,** 147–175.

George, E., Laud, P., Logan, B., McCulloch, R., and Sparapani, R. (2018). Fully Nonparametric Bayesian Additive Regression Trees. *arXiv* page 1807.00068.

Green, D. and Kern, H. (2012). Modeling Heterogeneous Treatment Effects in Survey Experiments with Bayesian Additive Regression Trees. *Public Opinion Quarterly* **76,** 491–511.

Hahn, P., Murray, J., and Carvalho, C. M. (2017). Bayesian regression tree models for causal inference: regularization, confounding, and heterogeneous effects. *arXiv* page 1706.09523.

He, J., Yalov, S., and Hahn, R. (2018). Accelerated Bayesian Additive Regression Trees. *arXiv* page 1810.02215.

Hernández, B., Pennington, S., and Parnell, A. (2015). Bayesian methods for proteomic biomarker development. *EuPA Open Proteomics* **9,** 54–64.

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

Hill, J. (2016). Atlantic Causal Inference Conference Competition results. New York University, New York. (Available from http://jenniferhill7.wixsite.com/acic-2016/competition).

Kapelner, A. and Bleich, J. (2015). Prediction with missing data via Bayesian additive regression trees. *The Canadian Journal of Statistics* **43,** 224–239.

Kapelner, A. and Bleich, J. (2016). bartMachine: Machine Learning with Bayesian Additive Regression Trees. *Journal of Statistical Software* **70,** 1–40.

Kindo, B., Wang, H., Hanson, T., and Pena, E. (2016). Bayesian quantile additive regression trees. *arXiv* page 1607.02676.

Kindo, B., Wang, H., and Pena, E. (2016). Multinomial probit Bayesian additive regression trees. *Stat* **5,** 119–131.

Kropat, G., Bochud, F., Jaboyedoff, M., Laedermann, J., Murith, C., Palacios (Gruson), M., and Baechler, S. (2015). Improved predictive mapping of indoor radon concentrations using ensemble regression trees based on automatic clustering of geological units. *Journal of Environmental Radioactivity* **147,** 51–62.

Lakshminarayanan, B., Roy, D., and Teh, Y. (2015). Particle Gibbs for Bayesian Additive Regression Trees. *arXiv* page 1502.04622.

Leonti, M., Cabras, S., Weckerle, C., Solinas, M., and Casu, L. (2010). The causal dependence of present plant knowledge on herbals – Contemporary medicinal plant use in Campania (Italy) compared to Matthioli (1568). *Journal of Ethnopharmacology* **130,** 379–391.

Liang, F., Li, Q., and Zhou, L. (2018). Bayesian Neural Networks for Selection of Drug Sensitive Genes. *Journal of the American Statistical Association* **113,** 955–972.

Linero, A. (2018). Bayesian Regression Trees for High-Dimensional Prediction and Variable Selection. *Journal of the American Statistical Association* **113,** 626–636.

Linero, A., Sinhay, D., and Lipsitzz, S. (2018). Semiparametric Mixed-Scale Models Using Shared Bayesian Forests. *arXiv* page 1809.08521.

Linero, A. and Yang, Y. (2018). Bayesian regression tree ensembles that adapt to smoothness and sparsity. *Journal of the Royal Statistical Society - Series B* **80,** 1087–1110.

Liu, Y., , Ročková, V., and Wang, Y. (2018). ABC Variable Selection with Bayesian Forests. *arXiv* page 1806.02304.

Liu, Y., Shao, Z., and Yuan, G. (2010). Prediction of Polycomb target genes in mouse embryonic stem cells. *Genomics* **96,** 17–26.

Liu, Y., Traskin, M., Lorch, S., George, E., and Small, D. (2015). Ensemble of trees approaches to risk adjustment for evaluating a hospital's performance. *Health Care Management Science* **18,** 58–66.

Lu, M., Sadiq, S., Feaster, D., and Ishwarana, H. (2018). Estimating Individual Treatment Effect in Observational Data Using Random Forest Methods. *Journal of Computational and Graphical Statistics* **27,** 209–219.

Murray, J. (2017). Log-Linear Bayesian Additive Regression Trees for Categorical and Count Responses. *arXiv* page 1701.01503.

Nalenz, M. and Villani, M. (2018). Tree ensembles with rule structured horseshoe regularization. *Annals of Applied Statistics* **12,** 2379–2408.

Nateghi, R., Guikema, S., and Quiring, S. (2011). Comparison and validation of statistical methods for predicting power outage durations in the event of hurricanes. *Risk analysis* **31,** 1897–1906.

Pratola, M. (2016). Efficient MetropolisHastings Proposal Mechanisms for Bayesian Regression Tree Models. *Bayesian Analysis* **11,** 885–911.

Pratola, M., Chipman, H., Gattiker, J., Higdon, D., McCulloch, R., and Rust, W. (2014). Parallel Bayesian Additive Regression Trees. *Journal of Computational and Graphical Statistics* **23,** 830–852.

Rässler, S. (2002). *Statistical matching: A frequentist theory, practical applications and alternative bayesian approaches* . Lecture Notes in Statistics, Springer Verlag, New York.

Rossi, P. (2014). *Bayesian Non- and Semi-parametric Methods and Applications* . Princeton University Press, Princeton, New Jersey.

Ročková, V. and Saha, E. (2018). On Theory for BART. *arXiv* page 1810.00787.

Ročková, V. and van der Pas, S. (2017). Posterior Concentration for Bayesian Regression Trees and their Ensembles. *arXiv* page 1708.08734.

Sayer, J., Bogard, S., Buonarosa, M., LeBlanc, D., Funkhouser, D., Bao, S., Blankespoor, A., and Winkler, C. (2011). Integrated Vehicle-Based Safety Systems Light-Vehicle Field Operational Test Key Findings Report DOT HS 811 416. National Center for Statistics and Analysis, NHTSA, U.S. Department of Transportation, Washington, DC. (Available from `http://www.nhtsa.gov/DOT/NHTSA/NVS/Crash%20Avoidance/Technical%20Publications/2011/811416.pdf`).

Schnell, P., Muller, P., Tang, Q., and Carlin, B. (2018). Multiplicity-adjusted semiparametric benefiting subgroup identification in clinical trials. *Clinical Trials* **15,** 75–86.

Schnell, P., Tang, Q., Offen, W., and Carlin, B. (2016). A Bayesian Credible Subgroups Approach to Identifying Patient Subgroups with Positive Treatment Effects. *Biometrics* **72,** 1026–1036.

Sivaganesan, S., Muller, P., and Huang, B. (2017). Subgroup finding via Bayesian additive regression trees. *Statistics in Medicine* **36,** 2391–2403.

Sparapani, R., Logan, B., McCulloch, R., and Laud, P. (2016). Nonparametric survival analysis using Bayesian Additive Regression Trees (BART). *Statistics in Medicine* **35,** 2741–2753.

Sparapani, R., Rein, L., Tarima, S., Jackson, T., and Meurer, J. (2018). Non-parametric recurrent events analysis with BART and an application to the hospital admissions of patients with diabetes. *Biostatistics* page Ahead of print.

Starling, J., Murray, J., Carvalho, C., Bukowski, R., and Scott, J. (2018). BART with Targeted Smoothing: An analysis of patient-specific stillbirth risk. *arXiv* page 1805.07656.

Tan, Y., Elliott, M., and Flannagan, C. (2017). Development of a real-time prediction model of driver behavior at intersections using kinematic time series data. *Accident Analysis & Prevention* **106,** 428–436.

Tan, Y., Flannagan, A., and Elliott, M. (2018a). "Robust-squared" Imputation Models Using BART. *arXiv* page 1801.03147.

Tan, Y., Flannagan, C., and Elliott, M. (2018b). Predicting human-driving behavior to help driverless vehicles drive: random intercept Bayesian additive regression trees. *Statistics and its Interface* **11,** 557–572.

Tan, Y., Flannagan, C., Pool, L., and Elliott, M. (2018). Accounting for selection bias due to death in estimating the effect of wealth shock on cognition for the Health and Retirement Study. *arXiv* page 1812.08855.

Xu, D., Daniels, M., and Winterstein, A. (2016). Sequential BART for imputation of missing covariates. *Biostatistics* **17,** 589–602.

Zeldow, B., Lo Re, V. r., and Roy, J. (2018). A semiparametric modeling approach using Bayesian Additive Regression Trees with an application to evaluate heterogeneous treatment effects. *arXiv* page 1806.04200.

Zhang, J. and Härdle, W. (2010). The Bayesian Additive Classification Tree applied to credit risk modelling. *Computational Statistics and Data Analysis* **54,** 1197–1205.

Zhang, S., Shih, Y., and Müller, P. (2007). A Spatially-adjusted Bayesian Additive Regression Tree Model to Merge Two Datasets. *Bayesian Analysis* **2,** 611–634.

A Hyperparameters for BART

The hyperparameters for continuous outcomes BART that needs to be set are: α, β, μ_μ, σ_μ, ν, and λ. These hyperparameters are constructed as a mix of apriori fixed and data-driven. For α and β, the default values of $\alpha = 0.95$ and $\beta = 2$ provide a balanced penalizing effect for the probability of a node splitting (Chipman et al., 2010). For μ_μ and σ_μ, they are set such that $E[Y|X] \sim N(m\mu_\mu, m\sigma_\mu^2)$ assigns high probability to the interval $(\min(Y), \max(Y))$.

This can be achieved by defining v such that $\min(Y) = m\mu_\mu - v\sqrt{m}\sigma_\mu$ and $\max(Y) = m\mu_\mu + v\sqrt{m}\sigma_\mu$. For ease of posterior distribution calculation, Y is transformed to become $\tilde{Y} = \frac{Y - \frac{\min(Y)+\max(Y)}{2}}{\max(Y)-\min(Y)}$. This results in $\tilde{Y} \in (-0.5, 0.5)$ where $\min(Y) = -0.5$ and $\max(Y) = 0.5$. This has the effect of allowing the hyperparamter μ_μ to be set as 0 and σ_μ to be determined as $\sigma_\mu = \frac{0.5}{v\sqrt{m}}$ where v is to be chosen. For $v = 2$, $N(m\mu_\mu, m\sigma_\mu^2)$ assigns a prior probability of 0.95 to the interval $(\min(Y), \max(Y))$ and is the default value. Finally for ν and λ, the default value for ν is 3 and λ is the value such that $P(\sigma^2 < s^2; \nu, \lambda) = 0.9$ where s^2 is the estimated variance of the residuals from the multiple linear regression with Y as the outcomes and X as the covariates.

For binary outcomes, the α and β hyperparameters are the same but the μ_μ and σ_μ hyperparameters are specified differently from continuous outcomes BART. To set the hyperparameters for μ_μ and σ_μ, we set $\mu_\mu = 0$ and $\sigma_\mu = \frac{3}{v\sqrt{m}}$ where $v = 2$ would result in an approximate 95% probability that draws of $\sum_{j=1}^m g(X; T_j, M_j)$ will be within $(-3, 3)$. No transformation of the latent variable Z would be needed.

B Posterior distributions for μ_{kj} and σ^2 in BART

B.1 $P(\mu_{kj}|T_j, \sigma, R_j)$

Let $R_{kj} = (R_{kj1}, \ldots, R_{kjn_k})^T$ be a subset from R_j where n_k is the number of R_{kjh}s allocated to the terminal node with parameter μ_{kj} and h indexes the subjects allocated to the terminal node with parameter μ_{kj}. We note that $R_{kjh}|g(X_{kjh}, T_j, M_j), \sigma \sim N(\mu_{kj}, \sigma^2)$ and $\mu_{kj}|T_j \sim$

$N(\mu_\mu, \sigma_\mu^2)$. Then the posterior distribution of μ_{kj} is given by

$$\begin{aligned}
P(\mu_{kj}|T_j, \sigma, R_j) &\propto P(R_{kj}|T_j, \mu_{kj}, \sigma)P(\mu_{kj}|T_j) \\
&\propto \exp[-\frac{\sum_h (R_{kjh} - \mu_{kj})^2}{2\sigma^2}] \exp[-\frac{(\mu_{kj} - \mu_\mu)^2}{2\sigma_\mu^2}] \\
&\propto \exp[-\frac{(n_k\sigma_\mu^2 + \sigma^2)\mu_{kj}^2 - 2(\sigma_\mu^2 \sum_h R_{kjh} + \sigma^2\mu_\mu)\mu_{kj}}{2\sigma^2\sigma_\mu^2}] \\
&\propto \exp[-\frac{(\mu_{kj} - \frac{\sigma_\mu^2 \sum_h R_{kjh} + \sigma^2\mu_\mu}{n_k\sigma_\mu^2 + \sigma^2})^2}{2\frac{\sigma^2\sigma_\mu^2}{n_k\sigma_\mu^2 + \sigma^2}}]
\end{aligned}$$

where $\sum_h (R_{kjh} - \mu_{kj})^2$ is the summation of the squared difference between the parameter μ_{kj} and the R_{kjh}s allocated to the terminal node with parameter μ_{kj}.

B.2 $P(\sigma^2|(T_1, M_1), \ldots, (T_m, M_m), Y)$

Let $Y = (Y_1, \ldots, Y_n)^T$ and i index the subjects $i = 1, \ldots, n$. With $\sigma^2 \sim IG(\frac{\nu}{2}, \frac{\nu\lambda}{2})$, we obtain the posterior draw of σ as follows

$$\begin{aligned}
P(\sigma^2|(T_1, M_1), \ldots, (T_m, M_m), Y) &\propto P(Y|(T_1, M_1), \ldots, (T_m, M_m), \sigma)P(\sigma^2) \\
&= P(Y|\sum_{j=1}^m g(X, T_j, M_j), \sigma)P(\sigma^2) \\
&= \{\prod_{i=1}^n (\sigma^2)^{-\frac{1}{2}} \exp[-\frac{(Y_i - \sum_{j=1}^m g(X_i, T_j, M_j))^2}{2\sigma^2}]\} \\
&\quad (\sigma^2)^{-(\frac{\nu}{2}+1)} \exp(-\frac{\nu\lambda}{2\sigma^2}) \\
&= (\sigma^2)^{-(\frac{\nu+n}{2}+1)} \\
&\quad \exp[-\frac{\nu\lambda + \sum_{i=1}^n (Y_i - \sum_{j=1}^m g(X_i, T_j, M_j))^2}{2\sigma^2}]
\end{aligned}$$

where $\sum_j^m g(X_i, T_j, M_j)$ is the predicted value of BART assigned to observed outcome Y_i.

C Metropolis-Hastings ratio for the grow and prune step

This section is modified from Appendix A of Kapelner and Bleich (2016). Note that

$$\alpha(T_j, T_j^*) = \min\{1, \frac{q(T_j^*, T_j)}{q(T_j, T_j^*)} \frac{P(R_j|X, T_j^*, M_j)}{P(R_j|X, T_j, M_j)} \frac{P(T_j^*)}{P(T_j)}\}.$$

where $\frac{q(T_j^*,T_j)}{q(T_j,T_j^*)}$ is the transition ratio, $\frac{P(R_j|X,T_j^*,M_j)}{P(R_j|X,T_j,M_j)}$ is the likelihood ratio, and $\frac{P(T_j^*)}{P(T_j)}$ is the tree structure ratio of Kapelner and Bleich, Appendix A. We now present the explicit formula for each ratio under the grow and prune proposal.

C.1 Grow proposal

C.1.1 Transition ratio

$q(T_j^*, T_j)$ indicates the probability of moving from T_j to T_j^* i.e. selecting and terminal node and growing two children from T_j. Hence,

$$\begin{aligned} P(T_j^*|T_j) = P(grow)&P(\text{selecting terminal node to grow from})\times \\ &P(\text{selecting covariate to split from})\times \\ &P(\text{selecting value to split on}) \\ = P(grow)&\frac{1}{b_j}\frac{1}{p}\frac{1}{\eta}. \end{aligned}$$

In the above equation, $P(grow)$ can be decided by the researcher although the default provided is 0.25, b_j is the number of available terminal nodes to split on in T_j, p is the number of variables left in the partition of the chosen terminal node, and η is the number of unique values left in the chosen variable after adjusting for the parents' splits.

$q(T_j, T_j^*)$ on the other hand indicates a pruning move which involves the probability of selecting the correct internal node to prune on such T_j^* becomes T_j. This is given as

$$\begin{aligned} P(T_j|T_j^*) &= P(prune)P(\text{selecting the correct internal node to prune}) \\ &= P(prune)\frac{1}{w_2^*} \end{aligned}$$

where w_2^* denotes the number of internal nodes which have only two children terminal nodes.

This gives a transition ratio of

$$\frac{q(T_j^*, T_j)}{q(T_j, T_j^*)} = \frac{P(T_j^*|T_j)}{P(T_j|T_j^*)} = \frac{P(prune)}{P(grow)}\frac{b_j p\eta}{w_2^*}.$$

If there are no variables with two or more unique values, this transition ratio will be set to 0.

C.1.2 Likelihood ratio

Since the rest of the tree structure will be the same between T_j^* and T_j except for the terminal node where the two children are grown, we need only concentrate on this terminal node. Let l be the selected node and l_L and l_R be the two children of the grow step. Then

$$\begin{aligned} \frac{P(R_j|X, T_j^*, M_j)}{P(R_j|X, T_j, M_j)} &= \frac{P(R_{l_{(L,1)},j}, \ldots, R_{l_{(L,n_L)},j}|\sigma^2)P(R_{l_{(R,1)},j}, \ldots, R_{l_{(R,n_R)},j}|\sigma^2)}{P(R_{1,j}, \ldots, R_{n_l,j}|\sigma^2)} \\ &= \sqrt{\frac{\sigma^2(\sigma^2 + n_l\sigma_\mu^2)}{(\sigma^2 + n_L\sigma_\mu^2)(\sigma^2 + n_R\sigma_\mu^2)}}\exp[\frac{\sigma_\mu^2}{2\sigma^2}(\frac{(\sum_{k=1}^{n_L} R_{l_{(L,k)},j})^2}{\sigma^2 + n_L\sigma_\mu^2} \\ &\quad + \frac{(\sum_{k=1}^{n_R} R_{l_{(R,k)},j})^2}{\sigma^2 + n_R\sigma_\mu^2} - \frac{(\sum_{k=1}^{n_l} R_{l_{(l,k)},j})^2}{\sigma^2 + n_l\sigma_\mu^2})]. \end{aligned}$$

C.1.3 Tree structure ratio

Because the T_j can be specified using three aspects, we let $P_{SPLIT}(\theta)$ denote the probability that a selected node θ will split and $P_{RULE}(\theta)$ denote the probability that which variable and value is selected. Then based on $P_{SPLIT}(\theta) \propto \frac{\alpha}{(1+d_\theta)^\beta}$ and because T_j and T_j^* only differs at the children nodes, we have

$$\begin{aligned}
\frac{P(T_j^*)}{P(T_j)} &= \frac{\prod_{\theta \in H^*_{terminals}}(1 - P_{SPLIT}(\theta)) \prod_{\theta \in H^*_{internals}} P_{SPLIT}(\theta) \prod_{\theta \in H^*_{internals}} P_{RULE}(\theta)}{\prod_{\theta \in H_{terminals}}(1 - P_{SPLIT}(\theta)) \prod_{\theta \in H_{internals}} P_{SPLIT}(\theta) \prod_{\theta \in H_{internals}} P_{RULE}(\theta)} \\
&= \frac{[1 - P_{SPLIT}(\theta_L)][1 - P_{SPLIT}(\theta_R)] P_{SPLIT}(\theta) P_{RULE}(\theta)}{1 - P_{SPLIT}(\theta)} \\
&= \frac{(1 - \frac{\alpha}{(1+d_{\theta_L})^\beta})(1 - \frac{\alpha}{(1+d_{\theta_R})^\beta})\frac{\alpha}{(1+d_\theta)^\beta}\frac{1}{p}\frac{1}{\eta}}{\frac{\alpha}{(1+d_\theta)^\beta}} \\
&= \alpha \frac{(1 - \frac{\alpha}{(2+d_\theta)^\beta})^2}{[(1+d_\theta)^\beta - \alpha]p\eta}
\end{aligned}$$

because $d_{\theta_L} = d_{\theta_R} = d_\theta + 1$.

C.2 Prune proposal

Since prune is the direct opposite of the grow proposal, the explicit formula of $\alpha(T_j, T_j^*)$ will just be the inverse of the grow proposal.