Bayesian Regression Trees for High Dimensional Prediction and Variable Selection

Antonio R. Linero
Department of Statistics
Florida State University
arlinero@stat.fsu.edu

Abstract

Decision tree ensembles are an extremely popular tool for obtaining high quality predictions in nonparametric regression problems. Unmodified, however, many commonly used decision tree ensemble methods do not adapt to sparsity in the regime in which the number of predictors is larger than the number of observations. A recent stream of research concerns the construction of decision tree ensembles which are motivated by a generative probabilistic model, the most influential method being the Bayesian additive regression trees (BART) framework. In this paper, we take a Bayesian point of view on this problem and show how to construct priors on decision tree ensembles which are capable of adapting to sparsity in the predictors by placing a sparsity-inducing Dirichlet hyperprior on the splitting proportions of the regression tree prior. We characterize the asymptotic distribution of the number of predictors included in the model and show how this prior can be easily incorporated into existing Markov chain Monte Carlo schemes. We demonstrate that our approach yields useful posterior inclusion probabilities for each predictor and illustrate the usefulness of our approach relative to other decision tree ensemble approaches on both simulated and real datasets.

Keywords. Variable selection, random forests, nonparametric regression, decision trees, Bayesian learning, Bayesian additive regression trees.

Short title. High dimensional Bayesian trees.

1 Introduction

Ensembles of decision trees are a commonly used tool for obtaining high quality predictions for classification and regression tasks; examples include random forests (Breiman, 2001) and boosted decision trees (Freund et al., 1999). Recently, several promising probabilistically motivated methods based on ensembles of decision trees, such as Mondrian Forests (Lakshminarayanan et al., 2014), have been proposed. Similarly, Bayesian approaches which model the unknown function as the realization of a random tree obtain an ensemble of decision trees through posterior averaging (Chipman et al., 1998; Denison et al., 1998).

There are mixed messages in the literature regarding the suitability of these commonly used decision tree procedures when the number of predictors P is large relative to the number of observations N. In practice, tools such as random forests often yield excellent predictions, with reports that they are robust to the presence of irrelevant predictors (see Statnikov et al. 2008, Menze et al. 2011, and the references therein). On the theoretical side, Biau (2012) and Scornet et al. (2015) show that, for fixed P and diverging N, the convergence rate for some types of random forests does not depend on the number of irrelevant predictors. Conversely, Zhu et al. (2015) show that unmodified variants of the random forest algorithm perform suboptimally when P is of comparable order to N.

The focus of this paper is on the development of decision tree ensembles which are suitable when P is of the same magnitude as, or potentially much larger than, N; specifically, we consider the nonparametric regression model

$$Y = f_0(X) + \epsilon_i, \qquad \epsilon_i \sim \mathcal{N}(0, \sigma^2),$$

with X taking values in $\mathbb{R}^P$ and our goal being the recovery of $f_0(x)$. In order for this to be possible, $f_0(x)$ must admit some additional structure. Our working assumption throughout will be the, now standard, sparsity assumption that $f_0(x)$ depends on x only through $(x_q : q \in \mathcal{Q}_0)$

where $\mathcal{Q}_0 \subseteq \{1, \ldots, P\}$ and the size of $\mathcal{Q}_0$, Q_0, is assumed to be much smaller than P. When this sparsity assumption is reasonable, an additional problem of interest is the accurate recovery of $\mathcal{Q}_0$. Examples of other works which assume sparsity in the nonparametric or semiparametric setting include Yang and Tokdar (2015), Storlie et al. (2011), Zhu et al. (2015), and Ravikumar et al. (2009).

As a prelude, to confirm that several commonly-used methods do not generally adapt to sparsity we consider the task of estimating

$$f_0(x) = 10 \sin(\pi x_1 x_2) + 20(x_3 - 0.5)^2 + 10x_4 + 5x_5 \tag{1}$$

given $N = 100$ observations with variance $\sigma^2 = 1$ and $P - 5$ irrelevant predictors. Following Chipman et al. (2010), we consider $P \in \{10, 100, 1000\}$. Results are presented in Figure 1 for the Bayesian additive regression trees (BART) model of Chipman et al. (2010) and the random forests algorithm implemented in the `randomForest` package in `R`. For both methods, predictions are seen to degrade substantially as P increases. At the extreme of $P = 1000$, BART predicts all unseen data to be roughly equal to $N^{-1} \sum_i Y_i$, with a high degree of uncertainty; the situation is similar for random forests, with jackknife confidence intervals (Wager et al., 2014) failing to provide accurate uncertainty quantification. For comparison, we also present results for DART, a proposed modification of BART which is resilient to the presence of large numbers of nuisance predictors, demonstrating that these difficulties are not intrinsic to the problem. The behavior exhibited here is typical of replications of this simulation and are not affected by choice of tuning parameters.

Our primary contribution in this paper is to show that one can attain adaptivity to sparsity within the Bayesian framework in a computationally simple, easy to implement, manner. The Bayesian approach treats the trees within the ensemble as realizations of random trees which are constructed by recursively splitting nodes according to the values of randomly chosen predictor variables. Let $s = (s_1, \ldots, s_P)$ be such that s_j is the probability that predictor j is used to construct a given split. Typically it is assumed that $s_j = P^{-1}$ so that predictors are chosen uniformly at random. We show that by instead choosing s to come from a sparsity-inducing Dirichlet distribution,

$$(s_1, \dots, s_P) \sim \mathcal{D}\left(\frac{\alpha}{P}, \dots, \frac{\alpha}{P}\right), \tag{2}$$

one can obtain adaptivity to sparsity in the function $f_0(x)$. As shown in Figure 2, draws from (2) are nearly sparse when α/P is small; this fact has also been used as a tool for constructing priors with minimax posterior convergence rates in the setting of Bayesian convex aggregation (Yang and Dunson, 2014), variable selection in linear models (Bhattacharya et al., 2015), and anisotropic function estimation (Bhattacharya et al., 2014). Unlike these uses of (2), our usage also induces *exact* sparsity due to the fact predictors with small associated values of s are unlikely to appear in our decision tree ensemble. We give analytic expressions for this preference for sparsity, allowing for subjective knowledge to be incorporated. Additionally, the conjugacy of the Dirichlet prior results in simple, joint, updates for s which can be incorporated into existing Gibbs samplers.

Methodologically, we focus on incorporating (2) into the BART framework; for ease of presentation, we label the modification of BART with the Dirichlet splitting rule prior as as DART, standing for Dirichlet additive regression trees. The gains from using (2) are most obvious in the case of BART, as BART ensembles generally include a large number of branches and, as a result, typically include a large number of splits on spurious predictors when $f_0(x)$ is sparse. When incorporated into the BART framework, the Dirichlet splitting rule prior possess a number of practical advantages. In the context of variable selection, existing BART-based approaches are forced to abandon the fully-Bayesian approach, and instead focus on qualitative measures of variable importance (Chipman et al., 2010) or utilize non-Bayesian permutation-based techniques (Bleich et al., 2014). Moreover, they are forced to use a small number of trees when conducting variable selection, resulting in the need to treat variable selection and prediction very differently. By contrast, because our methodology encourages sparsity, we are able to take a fully-Bayesian approach to variable selection, and do not need to place restrictions on the number of trees used in the model.

The outline of the paper is as follows. In Section 2 we review the BART framework. In

Section 3 we describe the sparsity-inducing Dirichlet prior, characterize the induced prior on $\mathcal{Q}_0$, and describe how to conduct variable selection. In Section 4 we provide computational details. In Section 5 we conduct a detailed simulation study and apply the methodology to some real datasets. We close in Section 6 with a discussion. All proofs, as well as results of additional simulation studies and computational details, are deferred to the supplementary material.

2 Review of Bayesian additive regression trees

Motivated by the success of boosting, and building on previous work on Bayesian classification and regression tree algorithms (Chipman et al., 1998; Denison et al., 1998), Chipman et al. (2010) developed the Bayesian additive regression trees, or BART, model. Focusing on the regression setting, $f_0(x)$ is modeled as the realization of a random sum of decision trees $f(x)$ given by

$$f(x) = \sum_{t=1}^{T} \mathcal{T}_t(x), \qquad x \in \mathbb{R}^P.$$

Each regression tree $\mathcal{T}_t(x)$ is determined by a binary tree structure S_t consisting of the topology and splitting rules of the tree, and a vector μ_t of parameters associated to the terminal nodes of S_t such that $\mathcal{T}_t(x) = \mu_{t\ell}$ if x is associated with terminal node ℓ of tree t. This model was shown to have substantial promise as a general purpose regression technique and has been extended to provide techniques for variable selection (Bleich et al., 2014), regression with heteroskedastic errors (Bleich and Kapelner, 2014), and causal inference (Hill, 2011), among others.

The models we consider are based on the following prior for $f(x)$. Each binary tree structure S_t is generated independently in the following manner. Let $q(d) : \mathbb{N} \to [0, 1]$. We initialize S_t as a tree with a single node at depth $d = 0$. This node is then given two child nodes of depth $d + 1$ with probability $q(d)$ and is terminal otherwise. This process iterates for the nodes at depths $d = 1, 2, \ldots$

until all nodes are terminal. A common choice for $q(d)$ (Chipman et al., 1998, 2010) is

$$q(d) = \frac{\gamma}{(1+d)^{\beta}}, \qquad \gamma \in (0,1), \quad \beta \in [0,\infty). \tag{3}$$

To each internal node we assign a splitting rule of the form $[x_j \le C]$. Each x associated to this internal node is then associated to one of its children according to whether x satisfies the splitting rule or not. The predictor used to construct a splitting rule is chosen according to the probability vector $s = (s_1, \ldots, s_P)$. There are several possibilities for the distribution of C given that predictor j is chosen to construct the splitting rule. Most implementations of BART use the following data-dependent prior (Chipman et al., 1998). We call a splitting rule *trivial* if it contradicts a splitting rule higher in the tree.

Assumption 2.1. Given that predictor j is selected, draw C uniformly from the collection of observed values $X_{1j}, \ldots, X_{nj}$ which lead to nontrivial splitting rules. If no such rule exists, draw a new predictor according to s and try again. The node becomes terminal if it is impossible to construct a nontrivial splitting rule.

We will also make use of a slightly modified version of Assumption 2.1 which greatly simplifies the analytic properties of the prior.

Assumption 2.2. Given that predictor j is selected, draw C uniformly from the collection of observed values $X_{1j}, \ldots, X_{nj}$ which lead to nontrivial splitting rules. If no such rule exists, record that a split on predictor j was attempted and increment the depth of the current node.

These two assumptions differ only in what happens when no further splitting is possible on a chosen predictor. Because trees constructed based on (3) with typical values of (γ, β) are very shallow, these two assumptions only differ substantively when some predictors only have a small number of unique sample values. In both our real and simulated examples, the X_j's are continuous, so that in practice there is effectively no difference between the two assumptions. When this is not the case, Assumption 2.2 effectively results in slightly smaller trees on average than Assumption 2.1.

Finally, associated to each terminal node ℓ in the tree, we draw a mean parameter $\mu_{t\ell} \sim \mathcal{N}(0, \sigma_\mu^2/T)$ where T is the number of trees in the ensemble. The normalization by T is chosen to ensure that the process of adding trees does not cause $f(x)$ to either "blow up" or tend to 0; instead, f can be shown to tend towards a Gaussian process as $T \to \infty$ under mild conditions. This correspondence with Gaussian processes gives some insight into how BART is capable of avoiding overfitting despite including a massive number of parameters.

Not all choices of $q(d)$ lead to finite trees almost-surely. Let D_{xt} denote the depth of the terminal node associated with a fixed x and observe that the D_{xt}'s are identically distributed. The following propositions, which follow from elementary branching process theory (Athreya and Ney, 2012), give sufficient conditions for S_t to be finite almost-surely.

Proposition 2.3. *Let $|S_t|$ denote the number of nodes in tree t and let $p(d)$ denote the mass function of D_{xt}. Let $G_p(\cdot)$ denote the probability generating function of $p(d)$. Then,*

$$E(|S_t|) = \sum_{d=0}^{\infty} 2^d \Pr(D_{xt} \geq d) = 2G_p(2) - 1.$$

In particular, S_t is finite almost-surely if $G_p(2) < \infty$.

Proposition 2.4. *For $\beta > 0$ and any $\gamma \in (0, 1)$, tree structures generated according to (3) are finite almost-surely . For $\beta = 0$, tree structures are finite with probability*$\min\{1, (1-\gamma)/\gamma\}$.

The above mechanism for sampling S_t is not the only one proposed in the literature. An alternative to specifying $q(d)$ is to place a prior directly on $|S_t|$ and a conditionally-uniform prior on the space of tree topologies (Denison et al., 1998). A similar alternative is to instead split existing nodes uniformly at random until some criteria is met (Biau et al., 2008). Finally, the Mondrian process (Roy and Teh, 2009) has also been proposed for drawing random tree structures (Lakshminarayanan et al., 2014).

3 Dirichlet splitting rule priors

3.1 Basic properties

Recall that s_j represents the probability that, at a given internal node, predictor j is chosen to construct the split. By default, existing Bayesian decision trees typically fix $s_j = P^{-1}$, one argument being that this provides a non-informative prior on $\mathcal{S}_t$. Another non-informative possibility is the uniform prior on s, corresponding to $s \sim \mathcal{D}(1, \ldots, 1)$.

In high dimensional settings, however, it is generally impossible to construct priors which are "non-informative", with seemingly non-informative choices of priors actually conveying dogmatic information. We argue that this is the case if one takes either $s_j = P^{-1}$ or $s \sim \mathcal{D}(1, \ldots, 1)$. For simplicity, we implicitly condition on the tree topologies $\mathcal{S}_1, \ldots, \mathcal{S}_T$ so that r, the number of splitting rules in the ensemble, is known. Troubling behavior occurs when P is large and r is held fixed.

Proposition 3.1. *Let Q denote the number of predictors used in constructing $f(x)$. Then, under either Assumption 2.1 or Assumption 2.2, we have $E(Q) = r + O_r(P^{-1})$ and $\operatorname{Var}(Q) = O_r(P^{-1})$ for both $s_j \equiv P^{-1}$ and $s \sim \mathcal{D}(1, \ldots, 1)$.*

The content of Proposition 3.1 is that, as $P \to \infty$, the prior quickly concentrates on models in which r predictors are included, the maximal number possible. Rather than expressing ignorance about the ensemble, these priors express preference for models which are highly non-sparse, with each included predictor accounting for as small a proportion of the signal as possible.

While seemingly a minor modification of the typical prior on Bayesian decision trees, the prior (2) induces drastically behavior than fixing $s_j = P^{-1}$ or setting $s \sim \mathcal{D}(1, \ldots, 1)$. In particular, the preference for non-sparse models with low-signal predictors is removed.

Proposition 3.2. *Under Assumption 2.2 with $s \sim \mathcal{D}(\alpha/P, \ldots, \alpha/P)$, conditional on r, the probability that a given predictor is associated to at least one internal node of the ensemble $\mathcal{T}_1, \ldots, \mathcal{T}_T$*

is

$$\Pr(\text{variable } j \text{ is included}) = 1 - \frac{\left\{\alpha(1-P^{-1})\right\}^{(r)}}{\alpha^{(r)}} = \frac{\alpha}{P}\left\{\psi(\alpha+r) - \psi(\alpha)\right\} + O_{\alpha,r}(P^{-2}),$$

where $\alpha^{(r)} = (\alpha + r - 1)(\alpha + r - 2)\cdots\alpha$ *and* $\psi(x) = d/dx \log \Gamma(x)$ *is the digamma function. Hence, the expected number of variables included in the ensemble is*

$$E(Q) = P \times \left[1 - \frac{\left\{\alpha(1-P^{-1})\right\}^{(r)}}{\alpha^{(r)}}\right] = \alpha\left\{\psi(\alpha+r) - \psi(\alpha)\right\} + O_{\alpha,r}(P^{-1}). \tag{4}$$

To understand the implication of this result, consider the artificial scenario with $P = \infty$. In this case the number of predictors included in the ensemble has mean $\alpha\left\{\psi(\alpha+r) - \psi(\alpha)\right\} \sim \alpha \log(1 + r/\alpha)$. In contrast to Proposition 3.1, this prior favors using a much smaller number of predictors. The number of predictors included in the ensemble is impacted by r, but at a tolerable logarithmic rate.

While this is an improvement, the dependence on r is unsatisfying. The next theorem addresses this and additionally provides an asymptotic description of the prior on Q.

Theorem 3.3. *Set* $\alpha = \theta/\log r$ *with* θ *fixed. Then*$[Q-1] \to \text{Poisson}(\theta)$ *in distribution as* $P, r \to \infty$ *at arbitrary rates.*

As shown in Figure 3, the Poisson approximation to $Q - 1$ in Theorem 3.3 is very accurate, provided that one uses either $E(Q - 1)$ or $\alpha\{\psi(\alpha + r) - \psi(\alpha)\} - 1$ in place of $\theta = \alpha \log r$. This is essential, as the convergence of $E(Q - 1)$ to θ occurs at a logarithmic rate in r.

3.2 Fully Bayesian variable selection

We mention two approaches to conducting variable selection using DART. The first is to use the posterior probability that a variable appears in a splitting rule at least one time in the ensemble, i.e., the probability that a variable exerts some influence on the response. While existing BART implementations do not use these probabilities, we note that these are appropriate quantities to use from a decision-theoretic perspective; for example, selecting the predictors with at least a 50% posterior probability of appearing in at least one split gives the median probability model (Barbieri and Berger, 2004). These probabilities can be estimated from the output of the Markov chain Monte Carlo algorithms typically used to fit these models.

An alternative is to regard the s_j's as measures of variable importance, with the idea that variables are important if they are used for many splitting rules. This approach is not as straightforward to interpret, but allows one to assess the relative importance of variables which are included in the model.

The above methods are particularly attractive as they requires neither additional computational overhead nor artificial restrictions on T. By contrast, existing BART-based approaches (Bleich et al., 2014) compare how often a predictor appears in a splitting rule relative to a null distribution obtained by fitting the model repeatedly to permutations of the data. This requires refitting the model many times, a process which is already computationally intensive. Figure 4 compares the fully-Bayesian variable selection properties of BART to DART when the true regression function is (1) with $Q = 5$ active predictors and $P = 95$ nuisance predictors and $\sigma^2 = 1$. When $T = 200$, variable selection using the posterior inclusion probability with the Dirichlet prior performs very well, while BART produces inadequate results in the sense that all irrelevant predictors are included in the model with high probability. Setting $T = 20$ closes the gap somewhat, at the cost of using a model with weaker predictions.

3.3 Choice of α

The choice of α is highly important. As illustrated by (4), α plays an central role in determining the degree of sparsity the model expects. One approach is to choose α to correspond to some targeted level of sparsity, potentially informed by subject-matter considerations. Another approach is to place a prior on α. This allows for the data to determine an appropriate degree of sparsity. We consider priors of the form

$$\frac{\alpha}{\alpha + \rho} \sim \text{Beta}(a, b) \tag{5}$$

for some hyperparameters (a, b, ρ). When $a = b = 1$, this corresponds to the prior density $\rho/(\alpha+\rho)^2$ for α, which has Cauchy-like tails and median ρ. The heavy tails here allow for large values of α so that the prior can revert to the BART prior when $f_0(x)$ is not sparse. We consider $b = 1$ and $a \in \{0.5, 1\}$, with $a = 0.5$ giving additional preference for sparsity in the prior. In our simulation and real-data examples, we use $a = 0.5$, $b = 1$, and $\rho = P$; the most important selection here is ρ, and smaller values of ρ than P are perhaps more appropriate when one has strong a priori reason to expect that f_0 is sparse. In Section 5 we show that use of a prior is competitive with optimal values of α.

A third option is to treat α as a tuning parameter and select it by cross-validation. This approach performs well and avoids the delicate issue of prior specification. The most serious downside to this is the additional computational expense of performing cross-validation.

4 Computational details

Existing implementations of BART conduct inference through Markov chain Monte Carlo (MCMC), with the $\mathcal{T}_t$'s being iteratively updated through Bayesian backfitting (Hastie et al., 2000). The critical component of these schemes is the construction of good Metropolis-Hastings steps for updating the topology $\mathcal{T}_t$ (Pratola, 2016; Lakshminarayanan et al., 2015).

Assuming this MCMC framework is already in place, the Metropolis-Hastings steps can be easily modified to account for s (Kapelner and Bleich, 2013). The only remaining detail is how to construct a valid update for s. When all predictors are candidates for splitting at each internal node of the ensemble, the Dirichlet prior gives a conjugate Gibbs-sampling update for s,

$$s \sim \mathcal{D}\left(\frac{\alpha}{P} + m_1, \ldots, \frac{\alpha}{P} + m_P\right), \tag{6}$$

where m_j denotes the number of attempted splits on predictor j. This allows for a simple, joint, update of s. Under Assumption 2.2, this update is always valid. Under Assumption 2.1, this update is only valid when all internal nodes can split on all predictors. We outline two strategies in the supplemental material for addressing Assumption 2.1. The first is to simply use (6) as a proposal distribution in an independence Metropolis-Hastings sampler. The second achieves an update of the form (6) by augmenting the trees with a latent history of proposed splits at nodes which have ineligible predictors. We note that in all examples we consider these modifications were not needed as all predictors had many unique values.

In addition to being a difficult problem statistically, Bayesian variable selection is also fraught with computational issues even in linear models (Ročková and George, 2014). Use of variable specific shrinkage often results in posteriors which are multimodal. As such, naively placing the update (6) into an existing algorithm may not succeed. To address this issue, we consider two techniques. The first is to use a tempering strategy, instead using the prior $s \sim \mathcal{D}(\alpha_t, \ldots, \alpha_t)$ at iteration t during some initial warmup phase, with α_t chosen so that $\alpha_t \downarrow \alpha/P$. This takes advantage of the fact that the default BART algorithm, which corresponds to $\alpha_t = \infty$, mixes well and allows us to initialize the chain in a high-quality mode of the posterior. Our second technique is to simply initialize the chain from a realization of the BART posterior, i.e., we "turn on" the update (6) after a some number of warmup iterations. When α is given a prior we have found that this technique is sufficient to obtain good performance in our examples without requiring the tempering strategy.

While computation time per iteration for DART is essentially the same as BART, the mixing of the chain may take more time. In all examples we consider, no more than 5000 warmup iterations and 5000 sampling iterations were used. Another potential issue is the Markov chain becoming trapped in a mode of the posterior, especially in cases where the predictors are highly correlated. This is a common occurrence for Bayesian variable selection when priors which are not log-concave are used. In these cases, it may be wise to run multiple chains. We show in the simulation studies of Section 5 that DART can be surprisingly resilient to the presence of highly correlated nuisance predictors.

5 Applications

5.1 Simulation study

We first evaluate DART under a variety of simulation settings. We consider four scenarios; the first is the example of Friedman (1991), the second and third are borrowed from Zhu et al. (2015), and in the last we draw regression functions from the prior.

Friedman The benchmark regression function (1) due to Friedman (1991), with various settings of σ^2, N, and P.

Checkerboard We let $X_i \overset{\text{iid}}{\sim} \mathcal{N}(0, \Sigma)$ where $\Sigma_{jk} = 0.9^{|j-k|}$, and let $Y_i = 2X_{i,50}X_{i,100} + 2X_{i,150}X_{i,200} + \epsilon_i$. Zhu et al. (2015) refer to this as a "checkerboard-like model with strong correlation," whom we follow in setting $N = 300$ and $\sigma^2 = 1$.

Linear A linear model, with $X_i \overset{\text{iid}}{\sim} \mathcal{N}(0, \Sigma)$. We set $\Sigma_{jk} = 0.5^{|j-k|} + 0.2I(i \neq j)$ and $Y_i = 2X_{i,50} + 2X_{i,100} + 4X_{i,150} + \epsilon_i$. We consider $N = 200$ and $\sigma^2 = 1$ for this model.

Tree Regression functions are generated by drawing from the BART/DART prior. We consider $N = P = 200$, $\sigma^2 = 1$, $\sigma_\mu^2 = 3$, and $\Sigma_{jk} = \rho^{-|j-k|}$ with $\rho \in \{0.2, 0.8\}$.

Methods are compared by root mean squared error, RMSE $= [\sum_{i=1}^{1000}\{f_0(X_i^\star) - \widehat{f}(X_i^\star)\}^2/1000]^{1/2}$, where the $X_i^\star$'s form a held-out test set.

For the Friedman, Checkerboard, and Linear scenarios we set $\alpha = 1$ across all settings. In limited post-hoc simulations we found this performs slightly worse than selecting α by cross-validation. Under ideal conditions α would have been chosen by cross-validation for each simulated dataset; however, this strategy would not have been practical for a simulation study. We use the Tree scenario to study the effect of choice of α on results.

The Friedman scenario will be considered in substantial detail. First we examined how the root mean square for various algorithms are impacted by the number of predictors. When fitting BART, we used 5-fold cross-validation to tune T and σ_μ^2, while the default settings of Chipman et al. (2010) were used with DART. In addition to BART and DART, we considered Gaussian process regression (GP) with an automatic relevance determination prior (Rasmussen and Williams, 2005) implemented in the `matlab` package `GPStuff` (Vanhatalo et al., 2013), the MARS algorithm (Friedman, 1991), and random forests (Breiman, 2001). Results of this simulation are presented in Figure 5. A striking feature of Figure 5 is that, in the low-noise setting, DART appears completely insensitive to the number of irrelevant predictors, producing results roughly equivalent to the BART prior under the oracle model with $P = 5$. In addition to DART, MARS and Gaussian process regression also attempt to adapt to sparsity, and we see that MARS is also robust to P. The GP method obtains the best performance for small values of P under the low-noise setting and is initially also robust, but becomes unreliable for larger values of P. Results for GP are worse when $\sigma^2 = 10$ and are omitted. We suspect these problems arise because of difficulties initializing the shrinkage parameters of the Gaussian process, and multiple initializations were used to improve results. GP also faces steep computational issues, taking several hours to fit at the $P = 1000$ setting, and so is omitted from further consideration.

To gain more insight into the relationship between the number of trees in the ensemble and the variable selection properties of DART, we also considered the Friedman scenario in a factorial design, taking $P \in \{200, 500, 1000\}$, $N \in \{100, 250, 500\}$, and $\sigma^2 \in \{10, 25\}$. For BART and DART, the number of trees were varied with $T \in \{20, 50, 200\}$. Each setting was replicated 200 times.

To conduct variable selection with the BART algorithm, we use the GSE method of Bleich et al. (2014). In addition to BART and DART, we considered using recursive feature elimination with the random forests algorithm implemented in the `caret` package and Bayesian variable selection with a spike-and-slab prior implemented in the `spikeslab` package. For evaluation metrics, we considered the precision, recall, and F_1 scores, given by prec = TP/(TP+FP), rec = TP/(TP+FN), and $F_1 = 2 \cdot \text{prec} \cdot \text{rec}/(\text{prec}+\text{rec})$ respectively, where TP denotes the number of predictors correctly flagged as influential, FP denotes the number of predictors incorrectly flagged as influential, and FN denotes the number of predictors incorrectly flagged as non-influential. The F_1 score is often used as an overall summary which balances precision and recall.

Results for the variable selection simulation for $N = 100$ are given in Table 1, with the settings $N = 250$ and $N = 500$ deferred to the supplementary material. The DART procedure obtains the best performance in terms of F_1 score, with the lone exception being the spike-and-slab prior attaining better performance at the $P = 200, \sigma^2 = 25$ setting. Direct comparisons between BART and DART are not easy to make, as the simulation study reveals that the technique of Bleich et al. (2014) prioritizes precision over recall, while this is less true of DART. It is possible to bring DART closer BART by using a cutoff for DART other than 50%; in results which are omitted, a 95% cutoff brings DART closer to BART in terms of precision while still obtaining a higher F_1 score. Both DART and BART are more conservative in allowing predictors into the model when compared to the spike-and-slab and random forest methods, both of which have higher recall but lower precision in the high-noise setting.

Also of note is the role played by the number of trees in the ensemble. Following Bleich et al. (2014) we only considered variable selection for $T = 20$ for BART. For DART, increasing the number of trees increased the recall at the expense of precision; as the number of nuisance predictors increases this results in better variable selection as measured by F_1 for the smaller values of T. For larger values of P we also see that $T = 50$ generally performs better than $T = 200$, though this effect is more pronounced for BART than DART.

Aggregate results for the Friedman, Checkerboard, and Linear scenarios are given in Figure 6. In addition to the random forest and spike-and-slab competitors, we also consider gradient boosting implemented in the `gbm` package, the LASSO implemented in the `glmnet` package, support vector regression implemented in the `e1071` package, the MARS algorithm implemented in the `earth` package, and the reinforcement learning trees (RLT) algorithm of Zhu et al. (2015) implemented in the `RLT` package. The results of 300 independent replications of the simulation are summarized in boxplots.

In the Checkerboard scenario the DART procedure is clearly best at $P = 200, 500$. At $P = 1000$, DART still performs best on average, but the distribution of RMSE is wider than other procedures. This occurs when the DART algorithm fails to focus on the active predictors, we suspect because MCMC fails to find the best mode of the posterior. Overall, DART performs quite well considering the extreme correlation in the predictors. The second best procedure in the Checkerboard scenario is the RLT procedure. In the Linear scenario, the LASSO and spike-and-slab methods perform the best, primarily because they correctly assume a linear model holds. Among methods which do not assume a linear model, DART and MARS perform best.

We use the Tree scenario to assess the robustness of the DART procedure to settings which are not sparse. We additionally aim to determine the impact of the choice of α or prior on α. Regression functions were generated with $\alpha = 3$ and $\alpha = \infty$, the latter corresponding to the BART procedure. We consider $\alpha \in \{4^0, 4^1, \dots, 4^5\}$. Results are given in Figure 7 and are based on 300 replications of the experiment. The choice of α here is highly influential. The main messages of this experiment are first that even when the underlying truth is not sparse DART is capable of performing at least as well as BART, and second that when the underlying truth is sparse, DART performs substantially better. We argue, then, that including the Dirichlet splitting-rule prior in BART , from a performance perspective, has little practical downside.

Another message of Figure 7 is that the model with prior (5) on α attains near-optimal performance. We used (5) with $a = 0.5, b = 1, \rho = P$; this favors non-sparse underlying functions, accounting for the slightly suboptimal behavior when the true function is sparse. In further simulations we found that 10-fold cross-validation reliably selected $\alpha = 1$ or $\alpha = 4$ when the DART prior held, but occasionally did not select large values of α when the BART prior held.

We conclude that DART is a highly competitive procedure under a variety of settings in which sparsity holds. It also appears to be robust to high correlation in the data. Moreover, as shown in the Tree scenario, the DART procedure does not break down when the true regression is non-sparse.

5.2 Applications to datasets

We illustrate the proposed methodology on three datasets, each of which illustrates different possible behaviors of DART; the first dataset has a sparse truth, the second is such that both sparse and non-sparse methods perform well, and the third is such that methods with sparse solutions perform poorly. The first dataset, `WIPP`, consists of data from a computer model for two-phase fluid flow. This dataset was analyzed previous by Storlie et al. (2011) to illustrate their ACOSSO technique. The `WIPP` dataset consists of $P = 31$ predictors and $N = 300$ observations. The second dataset, `triazines`, contains data on $N = 186$ molecular compounds with the goal of predicting biological activity from $P = 60$ features. This dataset is available from the UCI Machine Learning repository. The last dataset, `bbb`, contains data on $N = 208$ drugs with the goal of predicting the drugs associated brain-blood partition ratio (Mente and Lombardo, 2005). Predictions are based on $P = 134$ molecular descriptors. This dataset is available in the `caret` package in R.

Methods were evaluated by a 5-fold cross-validation estimate of root mean squared error $\text{RMSE} = [E\{(\hat{f}(X) - Y)^2\}]^{1/2}$. In addition to the additive regression tree models we also consider gradient boosted decision trees, support vector machine regression, the LASSO, random forests, and MARS. Hyperparameters for all methods but DART were optimized for each of the folds separately via a second layer of cross-validation. For BART, we applied both the default prior (BART-default) and a BART model which cross-validated over T and σ_μ^2 (BART-CV). For DART we used the default prior and gave α the prior (5) with $a = 0.5, b = 1, \rho = P$. The cross-validation was replicated 20 times and the results were averaged over.

Results are given in Table 2. To facilitate comparisons, we normalize the RMSE of each method by the RMSE of DART. We also consider the average proportion of the predictors which are used by each method; we do not consider the method of Bleich et al. (2014) here, as we are interested in the relationship between performance and the number of predictors used to form predictions, as opposed to performing variable selection.

On the WIPP dataset we see the advantage of the DART, which performed best by a wide margin. To predict the response only a small subset of the predictors is needed, which DART takes advantage of. On the triazines data set, DART performs essentially the same as BART, but is able to do so using only a fraction of the predictors. Gradient boosting and random forests, however, performed best on this dataset. The results on the bbb dataset are very interesting; note that there is a negative association between sparsity of each method and predictive performance. Here, DART gives essentially the same performance as BART but, unlike on the triazines dataset, almost all variables are used. This is because the response in bbb does not depend on a small number of predictors, and to account for this the posterior of α is concentrated on $\alpha > P$; at $\alpha = \infty$, DART is equivalent to BART, so the two methods are essentially the same when α is large. This behavior is encouraged by our heavy-tailed prior for α.

We conclude that the Dirichlet prior gives a nearly free improvement for BART in practice. When warranted, it discards a large number of predictors and potentially gains a substantial increase in performance. On the other hand, by using a prior on α, we can shield ourselves from an erroneous assumption that many predictors are irrelevant.

6 Discussion

In this paper we have demonstrated the utility of the Dirichlet splitting probability prior for Bayesian tree-based models, with a focus on the BART framework, in both prediction and variable selection problems. Additionally, we have empirically seen that tree-based methods such as random forests and boosting do not naturally adapt to sparsity in the $P \gg N$ regime. We note that our approach also extends to categorical response variables via data augmentation strategies (Chipman et al., 2010; Kindo et al., 2016).

We outline some interesting areas for future work. First, nothing has been said about the statistical theory underlying the posterior of the DART model, and it would be desirable to formalize the benefits of this model in terms of the convergence rate of the posterior in $Q \ll P$ settings. More generally, to our knowledge there has been no systematic development of theory for Bayesian regression tree models, even in the $T = 1$ setting. This situation is not uncommon for decision tree ensembles with tree structures that adapt to the data; for example, while progress is being made (Scornet et al., 2015; Biau et al., 2008), little underlying theory exists for the random forest algorithms used in practice.

An interesting possibility suggested by the success of the Dirichlet prior is the usage of penalization within algorithms for constructing trees. One idea is to apply the penalty implied by marginalizing over the Dirichlet prior when determining whether to split an internal node on a given predictor.

It is also of interest to consider priors which make more use of available prior information. This might include information about which variables are likely to be important, or information about which predictors are a-priori likely to occur together in groups (Ročková and George, 2014). Additionally, it would be interesting to develop priors which actively seek to find interaction effects.

Acknowledgements

The author thanks Fred Huffer for helpful discussions, as well as two anonymous reviewers whose comments helped improve this manuscript. This research was supported by the Office of the Secretary of Defense, Directorate of Operational Test and Evaluation under the Science of Test research program, #SOT-FSU-FATs-06.

References

Athreya, K. B. and Ney, P. E. (2012). *Branching processes*, volume 196. Springer Science & Business Media.

Barbieri, M. M. and Berger, J. O. (2004). Optimal predictive model selection. *Annals of Statistics*, pages 870–897.

Bhattacharya, A., Pati, D., and Dunson, D. (2014). Anisotropic function estimation using multi-bandwidth Gaussian processes. *Annals of statistics*, 42(1):352–381.

Bhattacharya, A., Pati, D., Pillai, N. S., and Dunson, D. B. (2015). Dirichlet-Laplace priors for optimal shrinkage. *Journal of the American Statistical Association, forthcoming*.

Biau, G. (2012). Analysis of a random forests model. *Journal of Machine Learning Research*, 13(Apr):1063–1095.

Biau, G., Devroye, L., and Lugosi, G. (2008). Consistency of random forests and other averaging classifiers. *Journal of Machine Learning Research*, 9(Sep):2015–2033.

Bleich, J. and Kapelner, A. (2014). Bayesian additive regression trees with parametric models of heteroskedasticity. *arXiv preprint arXiv:1402.5397*.

Bleich, J., Kapelner, A., George, E. I., Jensen, S. T., et al. (2014). Variable selection for BART: An application to gene regulation. *Annals of Applied Statistics*, 8(3):1750–1781.

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

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

Chipman, H. A., George, E. I., and McCulloch, R. E. (2010). BART: Bayesian additive regression trees. *Annals of Applied Statistics*, 4(1):266–298.

Denison, D. G., Mallick, B. K., and Smith, A. F. (1998). A Bayesian CART algorithm. *Biometrika*, 85(2):363–377.

Freund, Y., Schapire, R., and Abe, N. (1999). A short introduction to boosting. *Journal of the Japanese Society for Artificial Intelligence*, 14(5):771–780.

Friedman, J. H. (1991). Multivariate adaptive regression splines. *Annals of Statistics*, pages 1–67.

Hastie, T., Tibshirani, R., et al. (2000). Bayesian backfitting (with comments and a rejoinder by the authors. *Statistical Science*, 15(3):196–223.

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

Kapelner, A. and Bleich, J. (2013). bartmachine: Machine learning with bayesian additive regression trees. *arXiv preprint arXiv:1312.2171*.

Kindo, B. P., Wang, H., and Peña, E. A. (2016). Multinomial probit bayesian additive regression trees. *Stat*, 5(1):119–131.

Lakshminarayanan, B., Roy, D. M., and Teh, Y. W. (2014). Mondrian forests: Efficient online random forests. In *Advances in Neural Information Processing Systems*, pages 3140–3148.

Lakshminarayanan, B., Roy, D. M., and Teh, Y. W. (2015). Particle gibbs for bayesian additive regression trees. In *AISTATS*.

Mente, S. and Lombardo, F. (2005). A recursive-partitioning model for blood–brain barrier permeation. *Journal of Computer-Aided Molecular Design*, 19(7):465–481.

Menze, B. H., Kelm, B. M., Splitthoff, D. N., Koethe, U., and Hamprecht, F. A. (2011). On oblique random forests. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 453–469. Springer.

Pratola, M. T. (2016). Efficient metropolis–hastings proposal mechanisms for Bayesian regression tree models. *Bayesian Analysis*.

Rasmussen, C. E. and Williams, C. K. I. (2005). *Gaussian Processes for Machine Learning (Adaptive Computation and Machine Learning)*. The MIT Press.

Ravikumar, P., Lafferty, J., Liu, H., and Wasserman, L. (2009). Sparse additive models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 71(5):1009–1030.

Ročková, V. and George, E. I. (2014). EMVS: The EM approach to Bayesian variable selection. *Journal of the American Statistical Association*, 109(506):828–846.

Roy, D. M. and Teh, Y. W. (2009). The mondrian process. In *Advances in neural information processing systems*, pages 1377–1384.

Scornet, E., Biau, G., Vert, J.-P., et al. (2015). Consistency of random forests. *The Annals of Statistics*, 43(4):1716–1741.

Statnikov, A., Wang, L., and Aliferis, C. F. (2008). A comprehensive comparison of random forests and support vector machines for microarray-based cancer classification. *BMC bioinformatics*, 9(1):1.

Storlie, C. B., Bondell, H. D., Reich, B. J., and Zhang, H. H. (2011). Surface estimation, variable selection, and the nonparametric oracle property. *Statistica Sinica*, 21:679–705.

Vanhatalo, J., Riihimäki, J., Hartikainen, J., Jylänki, P., Tolvanen, V., and Vehtari, A. (2013). Gpstuff: Bayesian modeling with gaussian processes. *Journal of Machine Learning Research*, 14(Apr):1175–1179.

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

Yang, Y. and Dunson, D. B. (2014). Minimax optimal bayesian aggregation. *arXiv preprint arXiv:1403.1345*.

Yang, Y. and Tokdar, S. T. (2015). Minimax-optimal nonparametric regression in high dimensions. *Annals of Statistics*, 43(2):652–674.

Zhu, R., Zeng, D., and Kosorok, M. R. (2015). Reinforcement learning trees. *Journal of the American Statistical Association*, 110(512):1770–1784.

	P = 50				P = 200				P = 1000			
	RMSE	Recall	Precision	F_1	RMSE	Recall	Precision	F_1	RMSE	Recall	Precision	F_1
σ^2 = 10												
BART-20	3.057	0.601	**0.994**	0.732	3.564	0.530	**0.975**	0.671	4.291	0.464	**0.840**	0.581
BART-50	3.040	—	—	—	3.410	—	—	—	3.896	—	—	—
BART-200	3.230	—	—	—	3.868	—	—	—	4.462	—	—	—
DART-20	2.635	0.795	0.926	0.842	2.852	0.681	0.927	0.770	3.317	0.538	0.835	**0.631**
DART-50	2.464	0.830	0.932	0.867	**2.707**	**0.713**	0.945	**0.800**	**3.268**	0.522	0.835	0.620
DART-200	**2.438**	**0.852**	0.921	**0.878**	2.729	0.710	0.898	0.782	3.416	0.481	0.750	0.567
Spike and slab	2.847	0.798	0.669	0.708	3.245	0.676	0.847	0.730	4.465	0.198	0.784	0.275
Random forest	3.604	0.778	0.794	0.755	4.001	0.694	0.637	0.621	4.323	**0.609**	0.469	0.474
Standard Error	0.026	0.013	0.017	0.011	0.034	0.013	0.018	0.012	0.045	0.014	0.024	0.016
σ^2 = 25												
BART-20	3.637	0.332	**0.988**	0.466	4.194	0.300	**0.931**	0.427	4.764	0.250	0.708	0.344
BART-50	3.577	—	—	—	4.000	—	—	—	4.474	—	—	—
BART-200	3.758	—	—	—	4.255	—	—	—	4.665	—	—	—
DART-20	3.373	0.548	0.922	0.669	3.789	0.400	0.870	0.520	4.314	0.271	0.717	**0.363**
DART-50	**3.318**	0.587	0.885	**0.690**	**3.783**	0.405	0.826	0.520	**4.284**	0.267	0.685	0.356
DART-200	3.346	0.614	0.813	0.688	3.850	0.421	0.681	0.506	4.510	0.225	0.499	0.292
Spike and Slab	3.331	0.722	0.664	0.665	3.924	**0.493**	0.741	**0.553**	4.738	0.102	**0.822**	0.150
Random forest	3.843	**0.780**	0.543	0.573	4.252	0.725	0.328	0.369	4.560	**0.685**	0.186	0.213
Standard Error	0.035	0.014	0.020	0.016	0.041	0.015	0.018	0.016	0.048	0.020	0.024	0.016

Table 1: Results of the variable selection simulation study for the Friedman example for $N = 100$. The maximal standard error for each measure is also given. The best result for each category is in bold. Results for $N = 250$ and $N = 500$ are given in the supplementary material.

	Dataset					
Method	WIPP		Triazines		BBB	
DART	1.00*	(0.29)*	1.00	(0.11)*	1.00	(0.99)
BART-default	1.14	(1.00)	1.01	(0.96)	0.99*	(0.99)
BART-CV	1.10	(0.91)	0.98	(0.43)	1.01	(0.86)
LASSO	1.34	(0.87)	1.14	(0.26)	1.17	(0.27)
Random forests	1.44	(0.61)	0.93*	(0.79)	1.05	(0.95)
Support vector regression	1.39	(1.00)	1.07	(1.00)	1.03	(1.00)
Boosting	1.20	(0.97)	0.96	(0.44)	1.08	(0.77)
MARS	1.28	(0.35)	1.27	(0.16)	1.30	(0.07)*

Table 2: Performance on datasets, as measured by RMSE normalized by the RMSE of DART. Values in parenthesis are estimates of the proportion of predictors used by the different methods, with the Bayesian methods using the median probability model. Lowest values for each dataset are given asterisks.

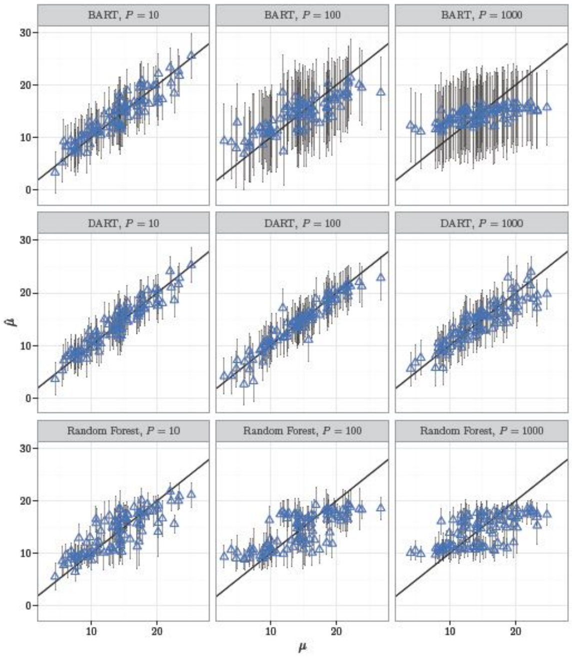

Figure 1: Plot of the the true mean $f_0(x) = \mu$ against the estimated mean $\hat{f}(x) = \hat{\mu}$ given 100 training examples on a held-out set of 100 observations, with 95% intervals for the mean response. The error variance is set to $\sigma^2 = 1$.

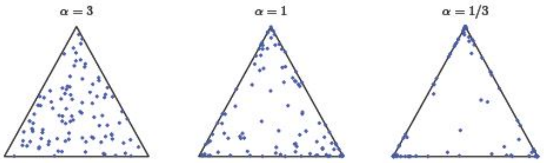

Figure 2: Draws from $\mathcal{D}(\alpha/3, \alpha/3, \alpha/3)$ priors on the simplex for differing values of α. Vertices of the simplex correspond to one-sparse probability vectors, edges to two-sparse vectors, and interior points to dense vectors.

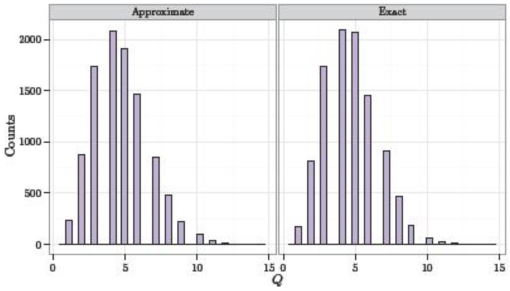

Figure 3: A histogram of draws of Q from both the approximate $1 + \text{Poisson}\{E(Q - 1)\}$ distribution and the exact distribution for $r = 300$, $\theta = 4$, and $P = 100$.

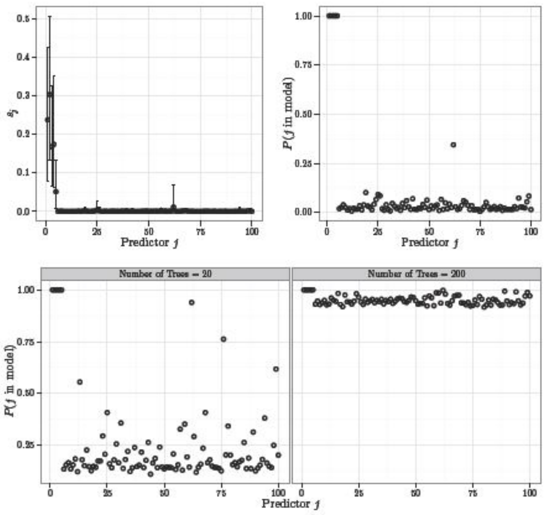

Figure 4: Top: Fully-Bayesian variable selection for DART with $T = 200$; the top-left plot gives the posterior mean and 95% credible intervals for the s_j's and top-right gives the posterior probability of inclusion in the model. Bottom: Posterior variable inclusion probabilities for BART $T = 20$ and $T = 200$.

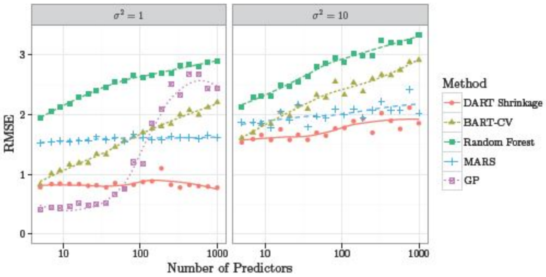

Figure 5: Graph of the RMSE for different procedures as a function of P on the log scale. Estimates are based on 5 replications of the simulation at each value of P, with a smoothing spline added for each algorithm to ease visualization.

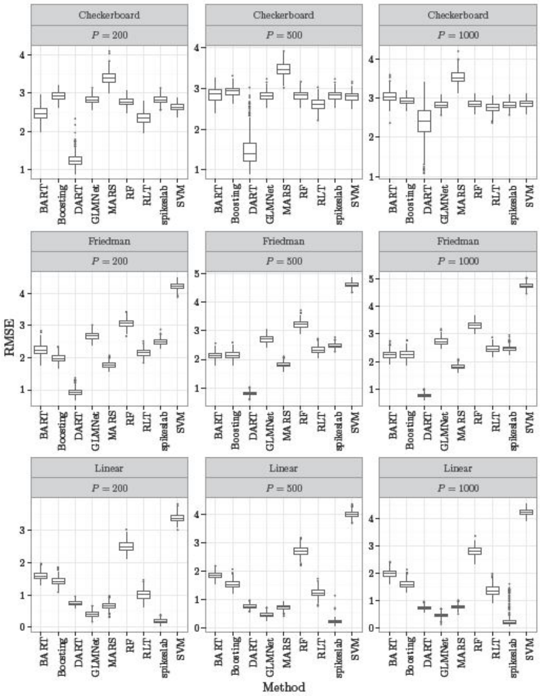

Figure 6: Results of simulation study for the Friedman, Checkerboard, and Linear scenarios. Boxplots give the quartiles and median, with whiskers extending to 1.5 times the interquartile range, and points beyond the whiskers displayed as outliers.

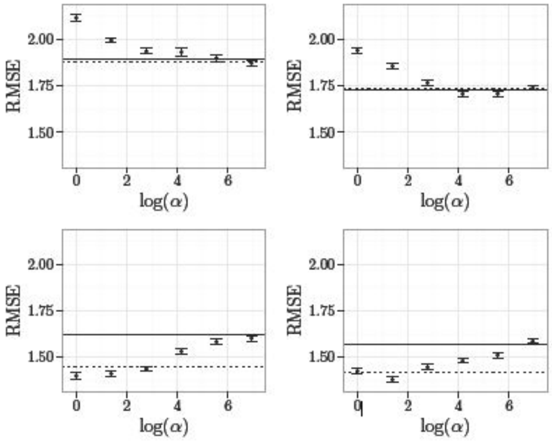

Figure 7: Mean loss and error bars for various settings of α in the Tree scenario, with 95% confidence intervals for the mean. Solid lines indicate the performance of BART and dashed lines indicate the performance of DART with a prior on α. Top panels are based on the BART prior, bottom panels are based on the DART prior, left panels use $\rho = 0.2$, and right panels use $\rho = 0.8$.