Interaction Detection with Bayesian Decision Tree Ensembles

Junliang Du **Antonio R. Linero**
Department of Statistics, Florida State University

Abstract

Methods based on Bayesian decision tree ensembles have proven valuable in constructing high-quality predictions, and are particularly attractive in certain settings because they encourage low-order interaction effects. Despite adapting to the presence of low-order interactions for prediction purpose, we show that Bayesian decision tree ensembles are generally anti-conservative for the purpose of conducting interaction detection. We address this problem by introducing Dirichlet process forests (DP-Forests), which leverage the presence of low-order interactions by clustering the trees so that trees within the same cluster focus on detecting a specific interaction. We show on both simulated and benchmark data that DP-Forests perform well relative to existing interaction detection techniques for detecting low-order interactions, attaining very low false-positive and false-negative rates while maintaining the same performance for prediction using a comparable computational budget.

1 INTRODUCTION

In many scientific problems, a primary goal is to discover structures which allow the problem to be described parsimoniously. For example, one may wish to find a small subset of candidate variables that are predictive of a response of interest; this structure is referred to as *sparsity*. Another structure is *interaction* (or *additive*) structure. An extreme case of additive structure is a generalized additive model (see, e.g., Hastie, 2017), where the effects of the predictors combine additively without any interactions. Teasing out additive structures can be valuable because it can substantially simplify the interpretation of a model. For example, if a given predictor does not interact with other predictors then it can be interpreted in isolation without reference to the values of other predictors. When predictors do interact, interpretation of the interactions is typically simplified whenever the interactions are of low-order. We consider the nonparametric regression problem $Y_i = f_0(X_i) + \epsilon_i$, $\epsilon_i \sim \text{Normal}(0, \sigma^2)$, where Y_i is a response of interest and $X_i \in \mathbb{R}^P$ is a vector of predictors, however the methods we develop here can be easily extended to many other settings. The variables x_j and x_k are said to *interact* if $f_0(x)$ cannot be written as $f_0(x) = f_{0\setminus j}(x) + f_{0\setminus k}(x)$ where $f_{0\setminus j}$ and $f_{0\setminus k}$ do not depend on x_j and x_k respectively. One can define higher order interactions similarly: a group of K variables is said to have a K-way interaction if $f_0(x)$ cannot be decomposed as a sum of K or fewer functions, each of which depends on fewer than K of the variables.

Methods which estimate $f_0(x)$ using an ensemble of Bayesian decision trees have proven useful in a number of statistical problems. Beginning with the seminal work of Chipman et al. (2010), Bayesian additive regression trees (BART) have been successfully applied in a diverse range of settings including survival analysis (Sparapani et al., 2016), causal inference (Hahn et al., 2017), variable selection in high dimensional settings (Linero, 2016; Bleich et al., 2014), loglinear models (Murray, 2017), and analysis of functional data (Starling et al., 2018). A key motivating factor for the use of BART is precisely that it is designed to taking advantage of low-order interactions in the data. Indeed, Linero and Yang (2017) and Rockova and van der Pas (2017) illustrate theoretically that the presence of low-order interactions is precisely the type of structure which BART excels at capturing. Hence BART appears to be an ideal tool for extracting low-order and potentially non-linear interactions.

Surprisingly, we show that, despite the ability of BART to capture low-order interactions for *prediction* purposes, it is nonetheless not suitable for conducting fully-Bayesian inference for the *selection* task of

Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics (AISTATS) 2019, Naha, Okinawa, Japan. PMLR: Volume 89.

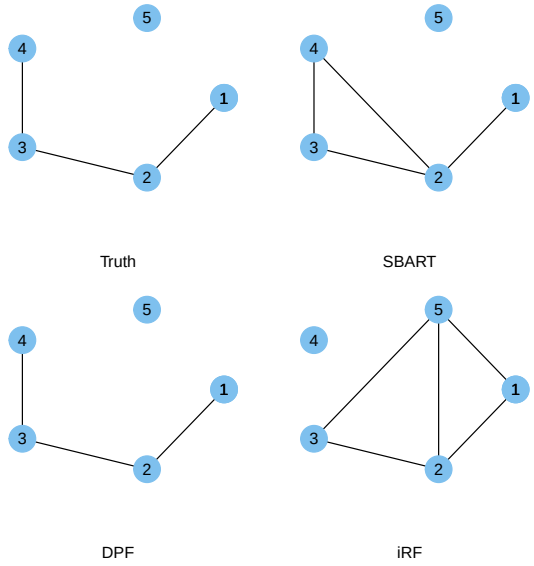

Figure 1: The interaction structure detected in the example from Section 1.1. "Truth" denotes the true interaction structure in the example.

interaction detection. When taken at face value as a Bayesian model, we show empirically that BART generally leads to the detection of spurious interaction effects. This is not contradictory because optimal prediction accuracy is generally *not* sufficient to guarantee consistency in variable selection (see, e.g., Wang et al., 2007).

We discuss the general problem which leads to the detection of spurious interactions; while this development is couched in the BART framework, we believe that the fundamental issues also occur for other decision tree ensembling methods. Specifically, the problem is that there is no penalty associated to including spurious interaction terms in the model. We then introduce a suitable modification to the BART framework which addresses this problem and allows BART detect interactions in a fully-Bayesian fashion. We accomplish this by clustering the trees into non-overlapping groups. Intuitively, the shallow trees comprising each cluster work together to learn a single low-order interaction. To bypass the need to specify the number of clusters, we induce the clustering through a Dirichlet process prior (Ferguson, 1973). We refer to the ensemble constructed in this fashion as a Dirichlet Process Forest (DP-Forest).

1.1 A Simple Example

To motivate the problem, we consider a simulated data example of Vo and Pati (2016). This example takes $P = 100$, $N = 100$, $X_i \sim \text{Normal}(\mathbf{0}, 0.02\,\mathrm{I})$, and $f_0(x) = x_1 + x_2^2 + x_3 + x_4^2 + x_5 + x_1x_2 + x_2x_3 + x_3x_4$. We compare the DP-Forest we propose to a variant of BART referred to as SBART (Linero and Yang, 2017) which can accommodate sparsity in variable selection. We also consider the recently proposed iterative random forests algorithm of Basu et al. (2018), selecting interactions whose stability score is higher than 0.5. In Figure 1 we display the interaction structure detected by each method on this data; while we considered only one iteration of this experiment here, these results are typical of replications of the experiment.

Here, SBART detects a spurious edge between x_2 and x_4. This occurs because BART, despite its fundamentally additive nature, does not include any penalization which discourages unnecessary interactions from being included. On the contrary, BART *expect* interactions to occur between relevant predictions; considering a draw from a BART prior such that x_2 and x_4 are included in the model, an interaction between these variables is a-priori likely. Adapting Bayesian decision tree ensembles to interaction detection then requires a prior which discourages the inclusion of weak interactions. The iRF similarly detects two spurious interactions and misses a relevant interaction between x_3 and x_4.

1.2 Related Work

Recent work has studied the theoretical properties of BART. Linero and Yang (2017) and Rockova and van der Pas (2017) show that certain variants of BART are capable of adaptively attaining near-minimax-optimal rates of posterior concentration when f_0 can be expressed as a sum of low-order interaction terms $f_0(x) = \sum_{v=1}^{V} f_{0v}(x)$ with each $f_{0v}(x)$ depending on a small subset of $\mathcal{S}_v$ of the predictors. In view this, one might conclude that no modification to BART is needed. This is true if one cares only about the mean integrated squared error $\int (f_0(x) - f(x))^2 \, F_0(dx)$ where $X_i \overset{\text{iid}}{\sim} F_0$. Optimal prediction performance, however, does not imply that variable selection and interaction detection are being performed adequately. If $\mathcal{S}_0$ is the true interaction structure of the data $\mathcal{S}$ is an estimate of $\mathcal{S}_0$, then attaining the minimax estimation rate for f_0 in terms of prediction error typically only guarantees that $\mathcal{S}_0 \subseteq \mathcal{S}$ (not $\mathcal{S} \subseteq \mathcal{S}_0$).

Several other methods have been recently proposed in the literature specifically for the task of interaction detection. We offer a non-comprehensive review. For a recent review, see Bien et al. (2013). Lim and Hastie (2015) proposed a hierarchical group-lasso which enforces the constraint that the presence of a given interaction implies the presence of the associated main effects; a similar approach is given by Bien et al. (2013).

A potential shortcoming of these approaches is that they focus on linear models and allow only pairwise interactions. Radchenko and James (2010) propose the VANISH algorithm, which allows for nonlinear effects through the use of basis function expansions, but again limits to pairwise interactions. Several decision-tree based methods have also been proposed. The additive groves procedure of Sorokina et al. (2008) uses an adaptive boosting-type algorithm to sequentially test for the presence of interactions between variables after performing a variable screening step. Basu et al. (2018) propose the iterative random forest (iRF) algorithm which flags "stable" interaction effects as those which appear consistently in many trees in a certain random forest.

2 BAYESIAN TREE ENSEMBLES

2.1 The BART Prior

Our starting point is the Bayesian additive regression trees (BART) framework of Chipman et al. (2010), which treats the function $f_0(\cdot)$ as the realization of a sum of random decision trees

$$f(x) = \sum_{t=1}^{T} g(x; \mathcal{T}_t, \mathcal{M}_t),$$

where $\mathcal{T}_t$ denotes the tree structure (including the decision rules) of the t^{th} tree and $\mathcal{M}_t = \{\mu_{t\ell} : \ell \in \mathcal{L}_t\}$ denotes the parameters associated to the leaf nodes; here, $\mathcal{L}_t$ denotes the collection of leaf nodes of $\mathcal{T}_t$. Let $[x \rightsquigarrow (t, \ell)]$ denote the event that the point x is associated to leaf ℓ in tree t. The function $g(x; \mathcal{T}_t, \mathcal{M}_t)$ then returns $\mu_{t\ell}$ whenever $[x \rightsquigarrow (t, \ell)]$ occurs.

We follow Chipman et al. (2010) and specify a branching process prior for the tree structure $\mathcal{T}_t$. A sample from the prior for $\mathcal{T}_t$ is generated iteratively, starting from a tree with a single node of depth $d = 0$; this is made a branch with two children with probability $q(d) = \gamma/(1+\beta)^d$, and is made a leaf node otherwise. We repeat this process independently for all nodes of depth $d = 1, 2, \ldots$ until all nodes at depth d are leaves. After the structure of the tree is generated, each branch b is associated with a decision rule of the form $[x_j \leq C_b]$. The coordinate j used to construct the decision rule is sampled with probability s_j where $s = (s_1, \ldots, s_P)$ is a probability vector. The splitting proportion s will play a key role later as an avenue for inducing sparsity in the regression function. Finally, we generate $C_b \sim \text{Uniform}(L_j, U_j)$ where $(L_1, U_1) \times \cdots \times (L_P, U_P)$ is the hyper-rectangle corresponding to the values of x that lead to branch b. We remark that this choice for C_b differs from the scheme used by other BART implementations; we adopt it to simplify the full conditionals we derive in Section 3.

For the prior on $\mathcal{M}_t$ we set $\mu_{t\ell} \overset{\text{iid}}{\sim} \text{Normal}(0, \sigma_\mu^2/T)$ conditional on $\mathcal{T}_t$ and σ_μ^2. By taking the variance to be σ_μ^2/T we ensure that the prior level of signal is constant as T increases. The normal prior is selected for its conjugacy; we note, however, that any prior for $\mu_{t\ell}$ with mean 0 and variance σ_μ/T leads to the approximation $f(x) \sim \text{Normal}(0, \sigma_\mu^2)$ by the central limit theorem. We fix $\beta = 2$ and $\gamma = 0.95$; we refer readers to Linero and Yang (2017) for further details regarding prior specification, and to Chipman et al. (2013) and Linero (2017) for detailed reviews of Bayesian decision tree methods.

2.2 Leveraging Structural Information

Several recent developments have extended the BART methodology to take advantage of structural information. Linero (2016) noted that sparsity in $f_0(x)$ can be accommodated automatically by setting $s \sim \text{Dirichlet}(\alpha/P, \ldots, \alpha/P)$. Recall here that s_j denotes the prior probability that, for a fixed branch, coordinate j will be used to construct a split at a that branch. Hence, if s is nearly-sparse with d non-sparse entries, the prior will encourage realizations from the prior to include only the d predictors with non-sparse entries. Linero and Yang (2017) showed that this prior for s induces highly desirable posterior concentration properties; in particular, the posterior of $f(x)$ concentrates at close to the oracle minimax rate if we had known the relevant predictors beforehand.

Linero and Yang (2017) also introduce the SBART model, which uses soft decision trees (Irsoy et al., 2012) which effectively replace the decision boundaries of BART with smooth sigmoid functions. This allows the SBART model to adapt to the smoothness level of $f(x)$; consequently, if $f_0(x)$ is assumed to be α-Hölder, the posterior for the SBART model concentrates around $f_0(x)$ at close to the oracle minimax rate obtainable when the smoothness level is known a-priori. While the methodology we develop applies to the usual BART models, we will use the SBART model with the sparsity-inducing Dirichlet prior in all of our illustrations.

3 DP-FORESTS

The distribution of $(\mathcal{T}_t, \mathcal{M}_t)$ in the BART model is parameterized by the splitting proportions s, leaf variance σ_μ^2, and tree topology parameters (γ, β). To encourage a small number of low-order interactions, we specify a prior which clusters the trees into non-overlapping groups such that each cluster constructs splits using different subsets of the predictors. A schematic is given in Figure 2 with $T = 4$. In this figure we see that the first two trees are dedicated to

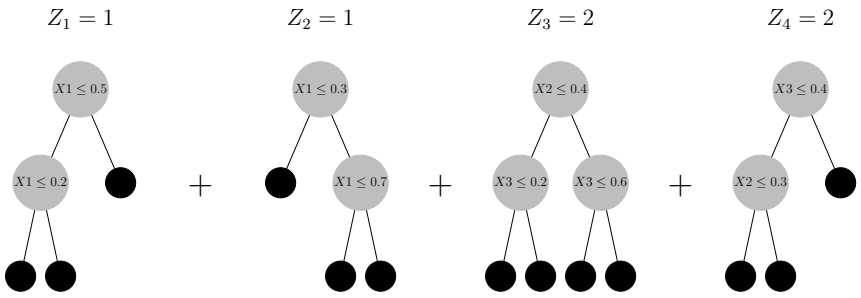

Figure 2: Schematic showing the effect of clustering trees within the ensemble. When $Z_t = 1$ split are constructed with X_1, but when $Z_t = 2$ splits are constructed with (X_2, X_3).

learning a main effect for x_1 while the second two trees are dedicated to learning an interaction between x_2 and x_3.

We induce a clustering by using tree-specific splitting proportions $s^{(t)} \sim G$ and using a Dirichlet process prior on G (Ferguson, 1973). Specifically, we let $s^{(t)} \overset{\text{iid}}{\sim} G$ conditional on G and let $G \sim \text{DP}(\omega G_0)$ where G_0 is a Dirichlet$(\alpha w_1, \ldots, \alpha w_P)$ distribution and ω denotes the precision parameter of the Dirichlet process. Using the latent-cluster interpretation of the Dirichlet process (see, .e.g, Teh et al., 2006) this can be approximated by the following generative model:

1. Draw $\pi \sim \text{Dirichlet}(\omega/K, \ldots, \omega/K)$ for large K.
2. Draw $Z_1, \ldots, Z_T \overset{\text{ind}}{\sim} \text{Categorical}(\pi)$.
3. Draw $s^{(1)}, \ldots s^{(K)} \overset{\text{ind}}{\sim} \text{Dirichlet}(\alpha w_1, \ldots, \alpha w_P)$ where $\sum_{p=1}^{P} w_p = 1, w_p \geq 0$.
4. For $t = 1, \ldots, T$, draw $(\mathcal{T}_t, \mathcal{M}_t)$ as described in Section 2 with $s = s^{(Z_t)}$.

The Z_t's cluster trees such that the trees within each group capture a single low-order interaction. Note that the use of the the sparsity inducing prior in step 3 above ensures that each $s^{(k)}$ will be nearly-sparse, and hence the trees with $Z_t = k$ will split on only a small subset of the predictors. The role played by this weight vector w is to encourage a subset of the predictors to appear in multiple *different* interactions. For example, if there are interactions (X_1, X_2) and (X_2, X_3) we do not want to encourage an additional (X_1, X_3) interaction. A large value of w_2 allows for this by encouraging X_2 to appear in several interactions.

3.1 Properties of the Prior

The degree of sparsity within each cluster of trees, as well as the overall number of clusters used, are determined by the hyperparameters α and ω. These hyperparameters are key in determining the interaction structures that the prior favors. To help anchor intuition we first consider several special cases of the DP-Forests model. First, we consider the behavior of the prior as $\alpha \to 0$ with ω fixed. In this case, with high probability each $s^{(t)}$ will have only one non-sparse entry. Consequently, each tree in the ensemble will split on at most one predictor. Because the trees are composed additively, this implies that none of the variables interact, and hence the prior concentrates on a sparse generalized additive model (SPAM, Ravikumar et al., 2007). On the other hand, as $\alpha \to \infty$ we see that $s^{(t)} \to (w_1, \ldots, w_P)$ so that the prior reverts to original BART model with splitting proportions given by $(w_1, \ldots, w_P)$ described by Bleich et al. (2014).

We can conduct a similar analysis with α fixed and ω with $K \to \infty$. As $\omega \to \infty$, each tree will be associated to a unique $s^{(t)}$. As $\omega \to 0$, on the other hand, all of the trees share the same $s^{(t)}$ so that the model collapses to the Dirichlet additive regression trees model described by Linero (2016).

The key difference between BART and a DP-Forest is that, once two variables are included, BART does not penalize interactions. Let A_i and A_j denote the event that variable i and j are included in the model, let A_{ij} denote the event that variables i and j interact, and let $\Pi_{\alpha,\omega}$ denote the joint prior distribution for $\mathcal{T}_1, \ldots, \mathcal{T}_T$. We study the prior on the interaction structure by examining the probabilities $\Lambda(\alpha, \omega) = \Pi_{\alpha,\omega}(A_{ij} \mid A_i \cap A_j)$, and $\Xi(\alpha, \omega) = \Pi_{\alpha,\omega}(A_{ik} \mid A_{ij} \cap A_{kj})$. In words, Λ is the probability that (i, j) interact given that both variables are relevant, while Ξ represents the probability that (i, k) interact given that (i, j) and (k, j) interact. Additionally, we examine the relationship between the average number of two-way interactions included in the model and the number of variables included.

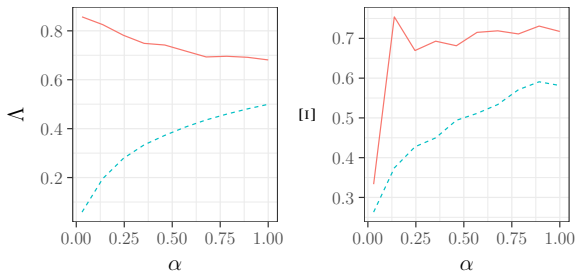

Figure 3: Plots of various quantities for $\omega = 0$ (solid, corresponding to SBART) and $\omega = 1$ (dashed) with $P = 5$ and $T = 50$. Left: plot of α against Λ. Middle: Plot of α against Ξ. Right: plot of the number of variables included in the model against the number of interactions. Values are computed approximately by sampling from the prior distribution.

Figure 3 shows several relationships between these quantities as α varies for both SBART and DP-Forests. We see that Λ is quite large for all values of α with SBART, implying that the prior expects any variables included in the model to interact; the trend is decreasing in α only because a larger number of predictors will be included in the model, causing variables to compete for branches in the ensemble. DP-Forests do not encourage the inclusion of interactions, particularly when α is small. Next, we see that Ξ is also uniformly large for SBART. This implies that the prior does not encourage interaction structures like the truth from Figure 1, while a DP-Forest with a small choice of α does.

3.2 Default Prior Settings

A benefit of the BART framework is the existence of default priors which require minimal tuning from users. Where applicable, we do not stray from the defaults recommended in Section 2. Specific to DP-Forests, the key parameter controlling the behavior of the model is α. On the basis of Figure 3 we recommend choosing α to be small; we have found setting $\alpha \sim$ Exponential with mean 0.1 to work well. Conversely, in our illustrations the results for the DP-Forest model do not depend strongly on ω, and we set $\omega \sim$ Exponential(1). This leaves the weight vector $w = (w_1, \ldots, w_P)$ to be specified. In our illustrations, we first run a screening step which removes irrelevant predictors. In principle any method can be used for screening; in our illustrations, we use SBART to screen variables which have posterior inclusion probability below 50%, and set $w_j \propto I(j \text{ is not screened})$. A more principled alternative is to use another sparsity-inducing prior on w. As suggested by the reviewers, we also considered a hierarchical prior in which $\alpha w_j = \xi_j \overset{\text{ind}}{\sim} \text{Gamma}(a/P, b)$. This is equivalent to setting $\alpha \sim \text{Gamma}(a, b)$ and $w_j \sim \text{Dirichlet}(a/P, \ldots, a/P)$. We found that this did not perform as well in experiments as using variable screening and omit the results.

3.3 Computation and Inference

Inference for the DP-Forest model can be carried out using a Gibbs sampler with the Bayesian backfitting approach of Chipman et al. (2010). The Gibbs sampler operates on the state space $(\{\mathcal{T}_t, \mathcal{M}_t, Z_t\}_{t=1}^T, \{s^{(k)}, \pi_k\}_{k=1}^K, \alpha, \omega, \sigma_\mu^2, \sigma^2)$. We use standard Metropolis-within-Gibbs proposals to update $\mathcal{T}_t$ and $\mathcal{M}_t$; see Kapelner and Bleich (2016) and Pratola (2016) for details. The parameters α, ω, σ_μ^2, and σ^2 can all be updated easily using the slice sampling algorithm of Neal (2003). Finally, $Z_t, s^{(k)}$, and π all have conjugate full-conditional distributions:

Full conditional for π: Note that π is conditionally independent of all parameters given (ω, Z). By conjugacy of the Dirichlet distribution to multinomial sampling we have the full conditional $\pi \sim \text{Dirichlet}(\omega/K + m_1, \ldots, \omega/K + m_K)$ where $m_k = \sum_t I(Z_t = k)$.

Full conditional for $s^{(k)}$: The conjugacy of the Dirichlet prior to multinomial sampling implies a Dirichlet full-conditional when a single s is used. To account for the clustering, we only consider the branches associated to trees with $Z_t = k$, giving the full conditional $s^{(k)} \sim \text{Dirichlet}(\alpha w_1 + c_1^{(k)}, \ldots, \alpha w_P + c_P^{(k)})$ where $c_j^{(k)}$ is the number of branches associated to cluster k which split on predictor j.

Full conditional for Z_t: Let $p(k)$ denote the full conditional for Z_t. The term $[Z_t = k]$ comes in only through the factors π_k (the prior probability of $Z_t = k$) and $\prod_{j=1}^P s_j^{(k)c_{tj}}$ where c_{tj} is the number of branches of tree t which split on predictor j (the likelihood of tree t having split on the predictors that it has, give $Z_t = k$). Hence $p(k) \propto \pi_k \prod_{j=1}^P s_j^{(k)c_{tj}}$.

Putting these pieces together, we arrive at Algorithm 1, which describes a single iteration of the Gibbs sampler.

4 EXPERIMENTS

We now compare DP-Forests to existing methods on a number of synthetic datasets. We consider the following methods in addition to DP-Forests and SBART.

Additive groves: The additive groves procedure of Sorokina et al. (2008). Because tuning of the additive groves algorithm is compute-intensive, we ran several pilot studies to choose appropriate tuning parameters which perform well for the given simulation settings.

Algorithm 1 Bayesian backfitting algorithm

for $t = 1, \ldots, T$ **do**
  Update $(\mathcal{T}_t, \mathcal{M}_t)$ via Metropolis-Hastings.
  Sample $Z_t \sim p(k), k = 1, \ldots, K$ where $p(k) \propto \pi_k \prod_{j=1}^{P} s_j^{(k)c_{tj}}$ and c_{tj} is the number of branches associated to tree t which split on predictor j.
end for
for $k = 1, \ldots, K$ **do**
  Sample $s^{(k)} \sim \text{Dirichlet}(\alpha w_1 + c_1^{(k)}, \ldots, \alpha w_P + c_P^{(k)})$ where $c_j^{(k)}$ is the number of branches associated to cluster k which split on predictor j.
end for
Sample $\pi \sim \text{Dirichlet}(\omega/K + m_1, \ldots, \omega/K + m_K)$ where $m_k = \sum_{t=1}^{T} I(Z_t = k)$.
Sample $(\sigma, \sigma_\mu, \alpha, \omega)$ using slice sampling.

Hierarchical group lasso: The hierarchical group lasso proposed by Lim and Hastie (2015) for interaction detection; we abbreviate this method by HL. This procedure was designed with linearity of $f_0(x)$ in mind. Tuning parameters are selected by cross-validation.

Hierarchical group lasso, least squares: HL is used to *select* the interactions and main effects, while the coefficients are estimated by least squares; we abbreviate this method by HL-LS. Tuning parameters are selected by cross validation.

Iterative random forests: The iterative random forests (iRF) procedure proposed by Basu et al. (2018) as implemented in the `iRF` package on `CRAN`. We use the default $T = 500$ trees and 10 iterations of the iRF algorithm.

Our simulation settings are borrowed from several existing works; we do not compare our methods to these other works due to a lack of publicly available software.

(S1) (Radchenko and James, 2010) We generate $X_i \sim \text{Uniform}([0,1]^P)$ where $P = 50$, $N = 300$, and $\sigma^2 = 1$. We let $f_0(x)$ be

$$\sqrt{0.5}\left[\sum_{v=1}^{V} f_v(x) + f_1(x)f_2(x) + f_1(x)f_3(x)\right]$$

where $f_1(x) = x_1$, $f_2(x) = (1 + x_2)^{-1}$, $f_3 = \sin(x_3)$, $f_4(x) = e^{x_4}$, and $f_5(x) = x_5^2$. Each $f_v(x)$ is further centered and scaled so that $E(f_v(X_i)) = 0$ and $\text{Var}(f_v(X_i)) = 1$.

(S2) (Vo and Pati, 2016) We generate $X_i \sim \text{Normal}(\mathbf{0}, \mathrm{I})$ with $N = 100$, $P = 100$, and $\sigma = 0.14$. We let $f_0(x) = x_1 + x_2^2 + x_3 + x_4^2 + x_5 + x_1x_2 + x_2x_3 + x_3x_4$.

(S3) Same as (S2), but without the interaction effects.

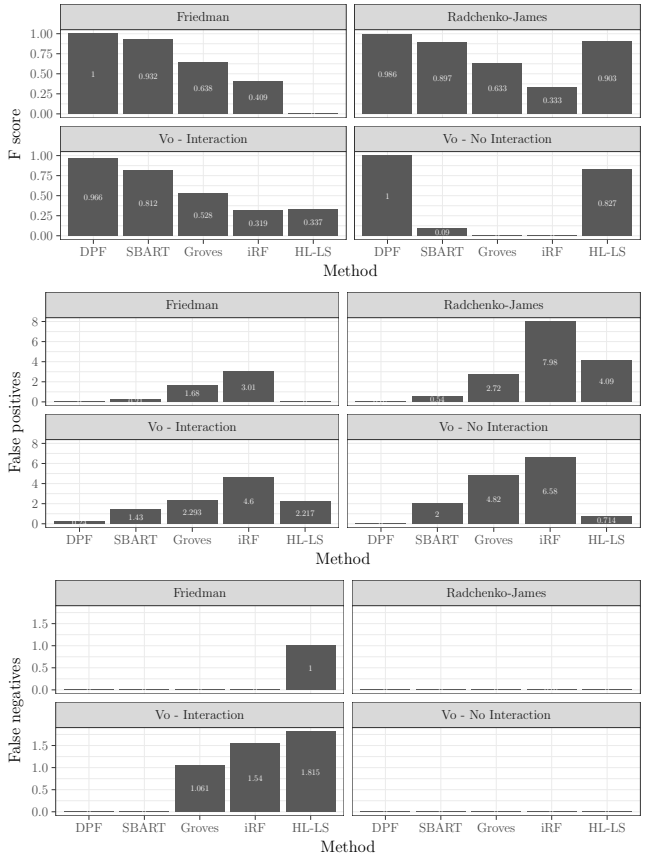

Figure 4: Barplot of results for interaction detection. The top row gives the average F_1 score for each method for detecting interactions. The second row gives the average number of false positive interactions detected. The bottom row gives the average number of false negatives detected. The average for each method is given on each bar.

(S4) (Friedman, 1991) A common test case for BART, we generate $X_i \sim \text{Uniform}([0,1]^P)$ with $P = 250, N = 250$, and $\sigma^2 = 1$. We set $f_0(x) = 10\sin(x_1x_2) + 20(x_3 - 0.5)^2 + 10x_4 + 5x_5$.

Each of these scenarios was replicated 100 times. We evaluate each method according to the average number of false positives (FPs), false negatives (FNs), F_1 score, and integrated root-mean squared error $\|f_0 - \widehat{f}\|_2$. The F_1 score is a commonly used measure of overall accuracy that balances false positives against false negatives in variable selection tasks; see, for example, Zhang and Yang (2015).

Results for interaction detection are given in Figure 4. We omit the results for HL because HL-LS performs uniformly better. Under all simulation settings, DP-Forests perform better than all other methods according to F_1 score. SBART is also competitive with other

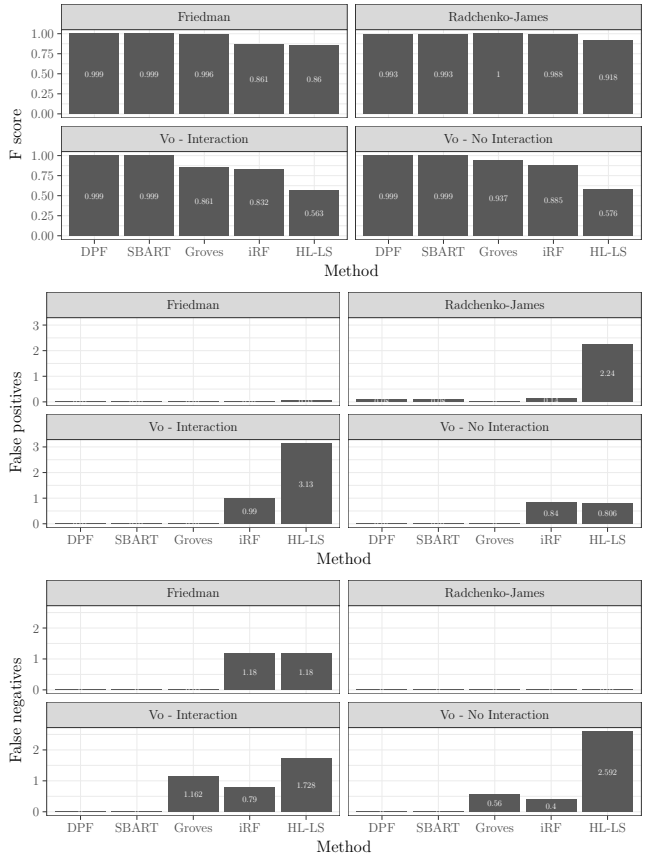

Figure 5: Barplot of results for detecting main effects.

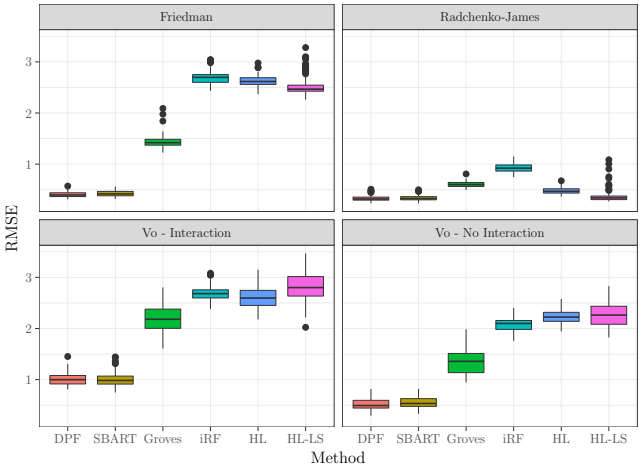

Figure 6: Boxplots given the distribution of integrated root mean-squared error for each method for each simulation setting.

procedures on many of the datasets. As expected, the primary problem with SBART is that it has a relatively large number of false positives, i.e. it is susceptible to detecting spurious interactions. This issue is most pronounced on (S2) and (S3), with SBART detecting between 1.5 and 2 spurious interactions.

Additive groves and iterative random forests generally perform worse than SBART. In addition to having a larger false positives rate, these procedures are also prone to false negatives under simulation (S2). With the exception of (S1), the hierarchical group-lasso (HL-LS) performs worse than the other methods. Under (S1), HL-LS has reasonable performance as each component of $f_0(x)$ can be reasonably well-approximated by the assumed linear model. HL-LS also appears to perform well under (S3); this, however, is due to the fact that HL-LS typically misses several main effects, which is a substantially worse outcome than detecting a spurious interaction. The nonlinearities under (S2) and (S4) also create problems for HL-LS.

All methods perform better for detecting the main effects. SBART and DP-Forests give identical results for the main effects due to the use of SBART in screening for DP-Forests. (S1) is the easiest setting, with all methods having very few false negatives and HL-LS the only method having non-negligible false-positives. Under (S2), the non-Bayesian procedures all have non-negligible false negatives, and iRF and HL-LS are additionally prone to false positives; the story is similar under (S3), with HL-LS performing better in terms of false positives but worse in terms of false negatives. All methods perform well in terms of false positives under (S4), however iRF and HL-LS also suffer from many false negatives.

Results for assessing prediction performance in terms of integrated root mean-squared error (RMSE) are given in Figure 6. SBART and DP-Forests perform very similarly in terms of RMSE. All other methods perform substantially worse under all settings. This is likely due to a multitude of factors. First, any false negatives will contribute to poor predictive performance. Second, SBART and DP-Forests are able to take advantage of underlying smoothness in the response function which additive groves and iterative random forests cannot, while HL and HL-LS suffer from an incorrect model specification.

SBART and DP-Forests are competitive in terms of runtime. For example, on a single replicate of (S4), SBART and DP-Forests took 118 seconds and 241 seconds respectively to obtain 40,000 samples from the posterior. By comparison, iRF took 279 second, HL-LS took 91 seconds, and additive groves took 4966 seconds. Additive groves was by far the slowest procedure, due to the fact that recursive feature elimination is used. We conclude that, under these settings, DP-Forests outperform all competitors are a competitive computational budget.

Method	RMSE
DP-Forests	1.00
iRF	1.22
HL	1.18
Additive Groves	1.16

Table 1: Cross-validation estimate of root mean-squared prediction error on the Boston housing dataset normalized by the RMSE of the DP-Forest.

We also consider the publicly available Boston housing dataset of Harrison and Rubinfeld (1978). Analysis of the interaction structures present in this dataset was previously undertaken by Radchenko and James (2010) and Vo and Pati (2016). This dataset consists of $P = 13$ predictors and $N = 506$ neighborhoods, and a continuous response corresponding to the median house value in a given neighborhood.

We compare the methods in terms of goodness-of-fit, which is evaluated using a 5-fold cross validated estimate of root mean squared prediction error. Results are given in Table 1. For prediction, the DP-Forest and SBART outperform the competing methods.

The DP-Forest includes most of the predictors in the model. This can be contrasted with the fit of a sparse additive model (SPAM) Ravikumar et al. (2007) and the fit of the VANISH model reported by Radchenko and James (2010), which include only a small number of predictors. Like the VANISH algorithm, the DP-Forest selects one interaction: there is strong evidence of an interaction between `DIS` (distance to an employment center in Boston) and `LSTAT` (the proportion of individuals in a neighborhood who are lower-status). This interaction was highly stable, and was selected by every fit to the data during cross-validation; additionally, this interaction was selected by additive groves in 4 out of 5 folds during cross-validation. Interestingly, this interaction was reportedly *not* selected by VANISH, which instead selects an interaction between the variables `NOX` (nitrus-oxide concentration) and `LSTAT`.

Figure 7 gives a visualization of the `LSTAT`-`DIS` interaction. To summarize the interaction we use a "fit-the-fit" strategy and fit a generalized additive model to the fitted-values of the DP-Forest with a thin plate spline term for the interaction (Wood, 2003). The plot then displays the `LSTAT`-specific effect of `DIS` for the $10^{\text{th}}, 20^{\text{th}}, \ldots, 90^{\text{th}}$ quantiles of `LSTAT`. This GAM nearly reproduces the fitted values from the DP-Forest and is easier to visualize. We see in Figure 7 a clear interaction between `DIS` and `LSTAT`. Intuitively, one expects that the closer a neighborhood is to an industry center the more expensive the housing will be. This is correct for areas with fewer lower-status individuals; however, this trend does not hold when there is a higher percentage of lower-status individuals. We remark also that the data is well supported near 0 for all values of `LSTAT`, so that this behavior is unlikely to be due to extrapolation, though extrapolation may be an issue for large values of both `LSTAT` and `DIS`.

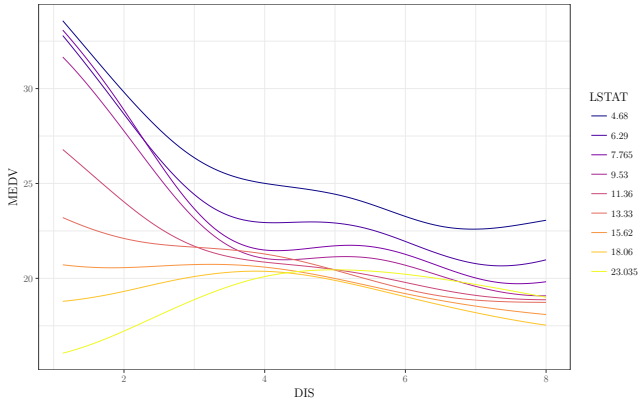

Figure 7: Graphical summary of the effect of distance `DIS` on `MEDV` for various values of `LSTAT`.

5 DISCUSSION

We have introduced Dirichlet process forests (DP-Forests) and applied them to the problem of interaction detection. We demonstrated on both synthetic and real data that DP-Forests lead to improved interaction detection. Additionally, we demonstrated that DP-Forests are highly competitive with commonly used machine learning techniques for detecting low-order interactions.

There are a number of modifications one might make to improve performance further. One possibility is to allow σ_μ to also vary by mixture component. This would allow different mixture components to have different signal levels; for example, under simulation (S4), we would expect that a smaller value of σ_μ^2 is appropriate for the mixture component responsible for x_5 relative to x_4. The proposed DP-Forests model captures this feature only indirectly through the number of trees assigned to each mixture component.

Additionally, it would be interesting to quantify the improvement in performance of DP-Forests over SBART theoretically. It is unknown whether SBART is variable-selection consistent, and establishing theoretically that DP-Forests are consistent for interaction detection while SBART is not remains an open problem.

References

Basu, S., Kumbier, K., Brown, J. B., and Yu, B. (2018). iterative random forests to discover predictive and stable high-order interactions. *Proceedings of the National Academy of Sciences*, page 201711236.

Bien, J., Taylor, J., and Tibshirani, R. (2013). A lasso for hierarchical interactions. *Annals of statistics*, 41(3):1111.

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

Chipman, H., George, E. I., Gramacy, R. B., and McCulloch, R. (2013). Bayesian treed response surface models. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 3(4):298–305.

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

Ferguson, T. S. (1973). A Bayesian analysis of some nonparametric problems. *The Annals of Statistics*, 1:209–230.

Friedman, J. H. (1991). Multivariate adaptive regression splines. *The Annals of Statistics*, 19(1):1–67.

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

Harrison, D. and Rubinfeld, D. L. (1978). Hedonic prices and demand for clean air. *Journal of Environmental Economics and Management*, 5:81–102.

Hastie, T. J. (2017). Generalized additive models. In *Statistical models in S*, pages 249–307. Routledge.

Irsoy, O., Yildiz, O. T., and Alpaydin, E. (2012). Soft decision trees. In *Proceedings of the International Conference on Pattern Recognition*.

Kapelner, A. and Bleich, J. (2016). bartMachine: Machine learning with Bayesian additive regression trees. *Journal of Statistical Software*, 70(4):1–40.

Lim, M. and Hastie, T. (2015). Learning interactions via hierarchical group-lasso regularization. *Journal of Computational and Graphical Statistics*, 24(3):627–654.

Linero, A. R. (2016). Bayesian regression trees for high dimensional prediction and variable selection. *Journal of the American Statistical Association*. To appear.

Linero, A. R. (2017). A review of tree-based Bayesian methods. *Communications for Statistical Applications and Methods*, 24(6):543–559.

Linero, A. R. and Yang, Y. (2017). Bayesian regression tree ensembles that adapt to smoothness and sparsity. *arXiv preprint arXiv:1707.09461*.

Murray, J. S. (2017). Log-linear Bayesian additive regression trees for categorical and count responses. *arXiv preprint arXiv:1701.01503*.

Neal, R. M. (2003). Slice sampling. *The Annals of Statistics*, 31:705–767.

Pratola, M. (2016). Efficient Metropolis-Hastings proposal mechanisms for Bayesian regression tree models. *Bayesian Analysis*, 11(3):885–911.

Radchenko, P. and James, G. M. (2010). Variable selection using adaptive nonlinear interaction structures in high dimensions. *Journal of the American Statistical Association*, 105(492):1541–1553.

Ravikumar, P., Liu, H., Lafferty, J., and Wasserman, L. (2007). SPAM: Sparse additive models. In *Proceedings of the 20th International Conference on Neural Information Processing Systems*, pages 1201–1208.

Rockova, V. and van der Pas, S. (2017). Posterior concentration for Bayesian regression trees and their ensembles. *arXiv preprint arXiv:1078.08734*.

Sorokina, D., Caruana, R., Riedewald, M., and Fink, D. (2008). Detecting statistical interactions with additive groves of trees. In *Proceedings of the 25th international conference on Machine learning*, pages 1000–1007. ACM.

Sparapani, R. A., Logan, B. R., McCulloch, R. E., and Laud, P. W. (2016). Nonparametric survival analysis using Bayesian additive regression trees (BART). *Statistics in medicine*.

Starling, J. E., Murray, J. S., Carvalho, C. M., Bukowski, R., and Scott, J. G. (2018). Functional response regression with funbart: an analysis of patient-specific stillbirth risk. *arXiv preprint arXiv:1805.07656*.

Teh, Y. W., Jordan, M. I., Beal, M. J., and Blei, D. M. (2006). Hierarchical Dirichlet processes. *Journal of the American Statistical Association*, 101(476):1566–1581.

Vo, G. and Pati, D. (2016). Sparse additive Gaussian process with soft interactions. *arXiv preprint arXiv:1607.02670*.

Wang, H., Li, R., and Tsai, C.-L. (2007). Tuning parameter selectors for the smoothly clipped absolute deviation method. *Biometrika*, 94(3):553–568.

Wood, S. N. (2003). Thin plate regression splines. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 65(1):95–114.

Zhang, Y. and Yang, Y. (2015). Cross-validation for selecting a model selection procedure. *Journal of Econometrics*, 187(1):95–112.