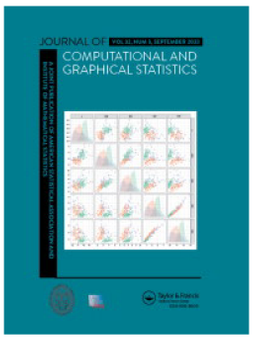

Taylor & Francis
Taylor & Francis Group

Journal of Computational and Graphical Statistics

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/ucgs20

Gibbs Priors for Bayesian Nonparametric Variable Selection with Weak Learners

Antonio R. Linero & Junliang Du

To cite this article: Antonio R. Linero & Junliang Du (2023) Gibbs Priors for Bayesian Nonparametric Variable Selection with Weak Learners, Journal of Computational and Graphical Statistics, 32:3, 1046-1059, DOI: 10.1080/10618600.2022.2142594

To link to this article: https://doi.org/10.1080/10618600.2022.2142594

JOURNAL OF COMPUTATIONAL AND GRAPHICAL STATISTICS
2023, VOL. 32, NO. 3, 1046–1059
https://doi.org/10.1080/10618600.2022.2142594

Gibbs Priors for Bayesian Nonparametric Variable Selection with Weak Learners

Antonio R. Linero[a] and Junliang Du[b]

[a]Department of Statistics and Data Sciences, University of Texas at Austin, Austin, TX; [b]Department of Statistics, Florida State University, Tallahassee, FL

ABSTRACT

We consider the problem of high-dimensional Bayesian nonparametric variable selection using an aggregation of so-called "weak learners." The most popular variant of this is the Bayesian additive regression trees (BART) model, which is the natural Bayesian analog to boosting decision trees. In this article, we use Gibbs distributions on random partitions to induce sparsity in ensembles of weak learners. Looking at BART as a special case, we show that the class of Gibbs priors includes two recently proposed models—the Dirichlet additive regression trees (DART) model and the spike-and-forest model—as extremal cases, and we show that certain Gibbs priors are capable of achieving the benefits of both the DART and spike-and-forest models while avoiding some of their key drawbacks. We then show the promising performance of Gibbs priors for other classes of weak learners, such as tensor products of spline basis functions. A Pólya Urn scheme is developed for efficient computations. Supplementary materials for this article are available online.

ARTICLE HISTORY
Received February 2022
Accepted October 2022

KEYWORDS
Bayesian nonparametrics; Machine learning; Model selection/variable selection; Nonparametric regression

1. Introduction

Bayesian machine learning methods that aggregate many "weak learners" have seen increased interest in recent years due to their ability to combine the principled uncertainty quantification of Bayesian inference with the predictive accuracy of modern machine learning (Chipman, George, and McCulloch 2010; Awaya and Ma 2021). The strategy is to approximate a function of interest as $r_0(x) \approx \sum_{m=1}^{M} \beta_m\, b(x;\gamma_m)$ where the $b(x;\gamma_m)$'s are relatively simple functions of a multivariate x determined by γ_m (such as a step function or spline) and β_m is a regression coefficient. The most popular implementation of this idea is the *Bayesian additive regression trees* (BART) framework, which builds an ensemble of shallow decision trees. BART was inspired by, and is generally competitive with, decision tree boosting (Freund, Schapire, and Abe 1999; Friedman 2001), but also gives immediate uncertainty quantification (Chipman, George, and McCulloch 2010). BART has been successfully used in many settings, including applications to survival analysis (Sparapani et al. 2016; Linero et al. 2021; Basak et al. 2021), log-linear models (Murray 2021), heteroscedastic regression (Pratola et al. 2020), causal inference (Hill 2011; Dorie et al. 2019; Hahn, Murray, and Carvalho 2020), and density regression (Li, Linero, and Murray 2022), among others. Rather than estimating the function by using greedy stagewise learning to "boost" the model, Bayesian ensembles are fit using a "Bayesian backfitting" algorithm (Hastie and Tibshirani 2000), which iteratively refines the learners using Markov chain Monte Carlo (MCMC).

Ensembles of weak learners are often applied in high-dimensional settings where some form of variable selection is desirable for the purpose of both regularization and interpretability. The learners $b(x;\gamma)$ are typically chosen to be "sparse" in the sense that $b(x;\gamma)$ depends on $x = (x_1,\ldots,x_P)$ only through a small number of the x_j's; because of this, one might guess that ensembles of weak learners naturally adapt to sparsity. Theoretical results for boosting notwithstanding (Bühlmann 2006), Linero (2018) shows that this is not the case in the Bayesian setting and that penalization beyond the inherent sparsity of the ensemble is required—the essential difficulty is that combining many weak learners together provides many opportunities for irrelevant variables to sneak into the model. Developing simple, effective, and computationally efficient ways of imposing sparsity in general Bayesian ensembles is therefore of great interest.

Our goal is to propose a general technique for imposing sparsity in Bayesian ensembles of weak learners. For simplicity we will focus on BART models; we emphasize, however, that this is not at all required by our approach, and we show in Section 5 that other choices of weak learners can also perform remarkably well. Our primary reason for focusing on BART is practical, as (i) it is highly accurate for many tasks of interest (Chipman, George, and McCulloch 2010; Dorie et al. 2019), (ii) it is the most popular Bayesian ensemble of weak learners, and (iii) it already has sparsity-inducing variants which we can contrast our approach with (Linero 2018; Rockova and van der Pas 2020). An abbreviated summary of this work is given in Linero and Du (2021).

Our main strategy is to build sparsity into the model using a Gibbs-type random partition process (Pitman 2002; Gnedin and Pitman 2006; De Blasi et al. 2013; Miller and Harrison 2018). Standard applications of random partitions include clustering, species sampling, and constructing random measures for

CONTACT Antonio R. Linero antonio.linero@austin.utexas.edu Department of Statistics and Data Sciences, University of Texas at Austin, Austin, TX.
Supplementary materials for this article are available online. Please go to www.tandfonline.com/r/JCGS.

Bayesian nonparametric models. We show here that random partition processes can also be remarkably useful for performing Bayesian nonparametric variable selection. At a high level, the prior we propose (a) randomly partitions the decision rules in a decision tree ensemble (or the variable selection indicators for a generic weak learner) into clusters and then (b) assigns a unique predictor to each cluster. We consider the class of *Gibbs priors* (De Blasi et al. 2013) for the random partition, which is rich enough to allow for efficient computations, intuitive prior specifications, and simple analytic properties.

We show that, when applied to BART, the class of Gibbs priors includes two recently proposed variable selection strategies as special cases: the DART model of Linero (2018) and the spike-and-forest model of Rockova and van der Pas (2020). These models correspond to extreme Gibbs priors that are suboptimal for different reasons. An advantage of the DART prior is that it is nearly trivial to implement, requiring only an additional step in standard Gibbs samplers; on the theoretical side, however, the variable selection properties of DART are not well understood, and empirically we find that the posterior will tend to overstate the number of relevant predictors. Conversely, spike-and-forest priors are very difficult to implement due to their need for rather complex MCMC schemes; to the best of our knowledge, the only variant of this model that is computationally tractable uses an approximate Bayesian computation (ABC, Liu, Ročková, and Wang 2021) algorithm that targets a surrogate distribution other than the posterior. Spike-and-forest priors are easier to analyze, however, as they allow for the size of the model to be controlled directly. In addition to being simpler to study, priors which directly control the size of the model are arguably more intuitive, particularly for users who are already experienced with spike-and-slab (George and McCulloch 1993) priors. When placed in this context, Gibbs priors carry the strengths of both models: we obtain the efficient computations and ease of implementation of the DART model while retaining the intuitive appeal and theoretical properties of the spike-and-forest prior.

We provide numerical evidence illustrating the advantages of Gibbs priors for BART models; specifically, MCMC schemes for Gibbs priors mix much better than spike-and-forest priors while also performing better at variable selection than DART due to a reduction in false positives. To make computations efficient, we take advantage of a characterization of Gibbs priors in terms of a Pólya urn scheme: if a predictor is used in a given decision rule, this directly increases the probability of it being used in other decision rules. We use this Pólya urn scheme to construct an efficient Metropolis-Hastings algorithm for updating the decision rules in the ensemble; a side benefit of this scheme is that it also provides a simple Metropolis-Hastings algorithm for spike-and-forest priors. We then illustrate the use of our methodology on four benchmark datasets, where we show that Gibbs priors generally lead to both sparser models and improved predictive performance.

Moving away from BART, we also argue that Gibbs priors are highly effective as a variable selection mechanism for a wide class of weak learners including: polynomials, tensor products of splines, and radial basis functions. We illustrate this in particular for Bayesian multivariate adaptive regression splines (MARS, Friedman 1991; Denison, Mallick, and Smith 1998) where the Gibbs prior performs extremely well on a commonly used benchmark simulation. Of independent interest, we find that a weak-learner-ensembling variant of Bayesian MARS performs much better than the usual (non-Bayesian) MARS algorithm, providing another data point in the literature showing the power of Bayesian ensembling of weak learners.

In Section 2 we review BART, introduce the Gibbs prior, and establish its fundamental properties. In Section 3 we construct a Metropolis-Hastings algorithm for updating the decision trees. In Section 4 we illustrate our methodology on both simulated and benchmark datasets. In Section 5 we show how to apply Gibbs priors to other classes of weak learners, and show that Gibbs priors work extremely well with multivariate adaptive regression splines. We conclude in Section 6 with a discussion. An implementation of our methodology is available publicly at *www.github.com/theodds/SoftBart*.

2. Model Description

2.1. Review of BART

The Bayesian additive regression trees (BART, Chipman, George, and McCulloch 2010) framework models an unknown function of interest as $r(x) = \sum_{t=1}^{T} g(x; \mathcal{T}_t, \mathcal{M}_t)$ where $g(x; \mathcal{T}_t, \mathcal{M}_t)$ is a *regression tree* with decision tree $\mathcal{T}_t$ and leaf node parameters $\mathcal{M}_t$. Together $\mathcal{T}_t$ and $\mathcal{M}_t$ define the step function $g(x; \mathcal{T}_t, \mathcal{M}_t)$; a schematic illustrating this is given in Figure 1. We let $\mathcal{L}(\mathcal{T})$ denote the leaf nodes (i.e., nodes with no children) of $\mathcal{T}$ and let $\mathcal{B}(\mathcal{T})$ denote the branch nodes (i.e., nodes which are not leaves). We use the notation $x \rightsquigarrow n$ for some $n \in \mathcal{L}(\mathcal{T}) \cup \mathcal{B}(\mathcal{T})$ to mean that x passes through the node

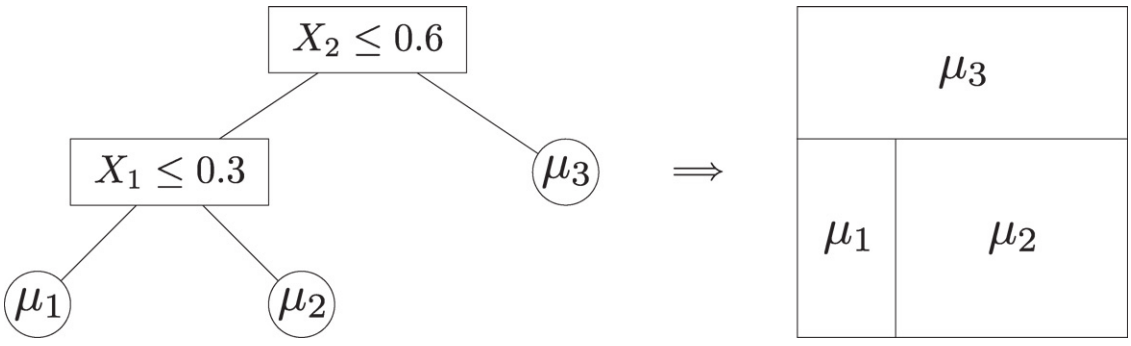

Figure 1. Schematic showing how a decision tree and leaf node parameters (left) maps to a step function of the predictors (right).

n at some point in its path to a leaf node; because of how the tree is constructed, each n is associated with a hyperrectangle of the form $H = \prod_{p=1}^{P}[A_p, B_p]$ such that $x \rightsquigarrow n$ if and only if $x \in H$. The prediction associated to leaf ℓ in tree t is given by $\mu_{t\ell}$ so that $\mathcal{M}_t = \{\mu_{t\ell} : \ell \in \mathcal{L}(\mathcal{T}_t)\}$. Associated to each $b \in \mathcal{B}(\mathcal{T})$ is a splitting rule of the form $[x_j \le C_b]$, with x associated to the left (right) node if the rule is true (false).

For concreteness, we will consider the semiparametric regression model

$$Y_i = r(X_i) + \epsilon_i, \qquad \epsilon_i \overset{\text{iid}}{\sim} \text{Normal}(0, \sigma^2), \tag{1}$$

with the understanding that all of our developements extend to all other types of settings for which BART models exist. To specify a BART prior we need (i) a prior $\pi_{\mathcal{T}}(\cdot)$ on the shape of the trees and (ii) a prior $\pi_{\mathcal{M}}(\cdot \mid \mathcal{T})$ on the values of the leaf node parameters given the trees. We then specify independent priors for the decision trees and leaf node parameters $(\mathcal{T}_t, \mathcal{M}_t) \overset{\text{iid}}{\sim} \pi_{\mathcal{T}}(\mathcal{T})\,\pi_{\mathcal{M}}(\mathcal{M} \mid \mathcal{T})$, and write $r \sim \text{BART}_T(\pi_{\mathcal{T}}, \pi_{\mathcal{M}})$ to denote that $r(\cdot)$ has a BART prior with T trees. Throughout this work we use default BART priors described by Chipman, George, and McCulloch (2010). Under this prior, a tree $\mathcal{T}$ can be generated from the prior by iterating the following steps:

G1. Initialize the $\mathcal{T}$ to consist of a single node of depth $d = 0$.
G2. For each node of depth d, make that node a branch node with two children with probability $\rho(d) = \gamma(1 + d)^{-\beta}$. Otherwise, the node becomes a leaf.
G3. If there are any branch nodes at depth d, set $d \leftarrow d+1$ and return to Step 2. Otherwise, continue.
G4. Assign each branch node b a splitting rule of the form $[x_{j_b} \le C_b]$ by sampling $j_b \sim \text{Categorical}(s)$ and sampling $C_b \sim \text{Uniform}(A_{j_b}, B_{j_b})$ where $\prod_p[A_p, B_p]$ is the hyperrectangle associated to b.
G5. Assign each leaf node a prediction $\mu_\ell \sim \text{Normal}(0, \sigma_\mu^2)$.

We emphasize that iterating through the above steps constitutes a draw from the *prior* distribution $\pi_{\mathcal{T}}$ rather than the posterior, which is approximated via the Bayesian backfitting algorithm.

Remark 1. Throughout this work, we will make a minor modification to the BART model to improve performance by using the *soft Bayesian additive regression trees* (SBART) prior; we defer interested readers to Linero and Yang (2018) for details. SBART replaces the trees $\mathcal{T}_t$ in the ensemble with *soft decision trees* (Irsoy, Yildiz, and Alpaydin 2012). For our purposes, the distinction between decision trees and soft decision trees does not impact the methodology we propose.

Remark 2. To make the connection between decision trees and learners of the form $r(x) = \sum_{m=1}^{M} \beta_m\, b(x; \gamma_m)$ more explicit, note that we can write the decision tree in Figure 1 as $g(x; \mathcal{T}, \mathcal{M}) = \mu_1\, I(x_1 \le 0.3)\, I(x_2 \le 0.6) + \mu_2\, I(x_1 > 0.3)\, I(x_2 \le 0.6) + \mu_3\, I(x_2 > 0.6)$ where $I(A)$ denotes an *indicator function* that is 1 when A is true and 0 otherwise. More generally, a decision tree ensemble can be written as $r(x) = \sum_{t=1}^{T}\sum_{\ell \in \mathcal{L}(\mathcal{T}_t)} \mu_{t\ell}\, I(x \rightsquigarrow \ell) = \sum_{m=1}^{M} \beta_m\, b(x; \gamma_m)$ where the β_m's correspond to the $\mu_{t\ell}$'s, the $b(x; \gamma_m)$'s correspond to the $I(x \rightsquigarrow \ell)$'s, and M is equal to the number of leaves in the ensemble. Hence, a decision tree corresponds to taking $r(x) = \sum_{m=1}^{M} \beta_m\, b(x; \gamma_m)$ where M is learned adaptively.

Remark 3. We do not place any restrictions on the number of observations per leaf; this is needed for our Bayesian backfitting algorithm, as our derivations require each node to have the possibility of splitting on any of the predictors. This does not hurt performance in practice because (i) the trees are already very shallow, making empty leaf nodes uncommon and (ii) the heavy use of regularization prevents empty leaf nodes from exerting much influence on the fit.

2.2. Sparsity Inducing Priors

Sparsity in the generative scheme described in Section 2.1 can be encoded at Step G4 via the probability s_j of using predictor j to construct a decision rule (Bleich et al. 2014). The variable used to construct a splitting rule is categorically distributed according to $s = (s_1, \ldots, s_P) \in \mathbb{S}_{P-1}$ where $\mathbb{S}_{P-1} = \{s : s_j \ge 0, \sum_j s_j = 1\}$ is a simplex; for example, in Figure 1 the prior probability that the root is associated with variable X_2 is given by s_2. The default value of the s_j's is P^{-1} in most software implementations so that all predictors are equally likely to be used for each rule, but there is no reason that this must be the case. By taking $s_j \ll 1/P$, for example, we can effectively eliminate x_j from the model.

Linero (2018) recommends taking $s \sim \text{Dirichlet}(\alpha/P, \ldots, \alpha/P)$ to perform variable selection in high dimensions. We refer to this as the Dirichlet additive regression trees, or *DART*, prior. When $\alpha \ll P$, this prior ensures that s will be nearly sparse in the sense that most entries of s will be very close to 0; because of this, the prior favors using only a small number $Q \ll P$ of predictors in the model. This preference can be sharply quantified, with Linero (2018) showing that, given that there are B decision rules in the ensemble, we have $Q - 1 \overset{\bullet}{\sim} \text{Poisson}\{\alpha \sum_{j=0}^{B-1}(\alpha + j)^{-1}\}$; more precisely, if Q_B is the number of variables included in an ensemble with B decision rules and s is given a $\text{Dirichlet}(\alpha_B/P, \ldots, \alpha_B/P)$ prior such that $\alpha_B \sum_{j=0}^{B-1}(\alpha_B + j)^{-1} \to \theta$ as $B, P \to \infty$, then $Q_B - 1 \to \text{Poisson}(\theta)$ in distribution. A further prior on α can be specified to allow for finer control of Q. Additionally, the Dirichlet prior is (provided that the prior $\pi_{\mathcal{T}}$ is specified so that all predictors can be used at every decision rule) conditionally conjugate, with the full conditional being $s \sim \text{Dirichlet}(\alpha/P + m_1, \ldots, \alpha/P + m_P)$ where m_j is the number of times variable j is included in the ensemble.

Rockova and van der Pas (2020) propose an alternative sparsity-inducing prior which more directly controls the number of relevant predictors. Modifying their setup slightly, their *spike-and-forest* prior works in terms of the set $\mathcal{S}$, taking

$$D \sim \pi_D \qquad [\mathcal{S} \mid D] \sim \binom{P}{D}^{-1} I(|\mathcal{S}| = D) \qquad \text{and}$$

$$s_j = \frac{I(j \in \mathcal{S})}{D}.$$

That is, we first sample the total number D of variables which will be allowed to split on, then we sample a subset of variables of size D from the available P variables. Given that a variable is included in $\mathcal{S}$ it will have $s_j = 1/D$; otherwise, $s_j = 0$ so that variable j will not appear in any splitting rule. The variable D is distinct from the total number of variables which appear Q, as there is a possibility that variable j is never used to construct a splitting rule even if $j \in \mathcal{S}$.

There are several substantive differences between the spike-and-forest and Dirichlet priors. Much of the benefit of the DART prior is computational. To the best of our knowledge, there has not been any practical algorithm for implementing the spike-and-forest prior; one is proposed by Liu, Ročková, and Wang (2021) but this is seen to perform worse than their ABC-Forests algorithm. We will show that part of the issue is that, because $s_j \equiv 1/D$, it is difficult for any Bayesian backfitting algorithm to add or remove predictors from the model. If there are B branches in the ensemble then adding a new variable at a single location must contend with the fact that $1/B \ll 1/D$ so that a-priori the new variable is appearing in far fewer branches than it should. Conversely, removing an irrelevant variable is difficult because the posterior is encouraged to have an existing variable appear in B/D branches, but modifying a tree in a local fashion can only reduce the number of appearances by 1. By contrast, DART has a simple implementation which only requires sampling s from its full conditional. It also avoids these mixing problems because it allows for (and, indeed, encourages) entries of $s_j > 0$ which are smaller than $1/D$.

A minor additional benefit of the DART prior is that it allows for different variable importances among the relevant variables; for example, the model can adapt to the need for more splits on (say) x_1 than x_2 by having $s_1 > s_2$. This is not possible for the spike-and-forest prior, as $s_1 = s_2 = 1/D$ when both x_1 and x_2 are relevant. Linero (2017) links this property of DART to anisotropic Gaussian processes, which are appropriate if $r_0(x)$ varies more in some directions than others.

Benefits of spike-and-forest priors over DART include the relative transparency of the prior specification and their known theoretical properties. Rather than inducing a prior on Q through a prior on s, the spike-and-forest prior directly works in terms of variable inclusion indicators $\gamma_j = I(j \in \mathcal{S})$, a strategy which is familiar to most applied Bayesian researchers. On the theoretical side, Liu, Ročková, and Wang (2021) prove variable selection consistency results for spike-and-forest priors. It is not clear whether similar results could be proven for DART. One challenge is that the Dirichlet prior may not sufficiently penalize a variable appearing (say) only once in the ensemble, causing the DART posterior to overestimate the number of relevant variables on average; a similar phenomenon was noted by Miller and Harrison (2013) to cause Dirichlet process mixture models to be inconsistent for the number of mixture components in infinite Gaussian mixture models.

Remark 4. In this work, we select variable x_j for final inclusion into the model using the *posterior inclusion probability*, which is given by the posterior probability that x_j is used in at least one decision rule. The final selected model is given by the *median probability model*, which includes variable x_j if its posterior inclusion probability is at least 0.5 (Barbieri and Berger 2004).

2.3. Gibbs Priors

When applied to Bayesian decision tree ensembles, the Gibbs priors we introduce combine the practical benefits of DART priors with the conceptual and theoretical benefits of spike-and-forest priors. Our starting point, which corresponds to a particular Gibbs prior, is a simple merger of the two priors together by adding a Dirichlet hyperprior into the spike-and-forest hierarchy:

$$D \sim \pi_D, \qquad [\mathcal{S} \mid D] \sim \binom{P}{D}^{-1} I(|\mathcal{S}| = D), \quad \text{and}$$
$$[s \mid \mathcal{S}] \sim \text{Dirichlet}(\alpha\,\gamma_1, \ldots, \alpha\,\gamma_P) \tag{2}$$

where $\gamma_j = I(j \in \mathcal{S})$. We adopt the convention that if the shape parameter associated to s_j is zero then $s_j \equiv 0$; for example, we interpret $s \sim \text{Dirichlet}(1, 0, 2)$ to mean that $(s_1, s_3) \sim \text{Dirichlet}(1, 2)$ and $s_2 \equiv 0$. Model (2) generalizes both the spike-and-forest and DART priors: we obtain the DART model when $D \equiv P$ (with α/P in the role of α) and the spike-and-forest prior when $\alpha \to \infty$.

Figure 2 compares the DART, spike-and-forest, and (2) in terms of their induced prior on s. In this figure, the center of the simplex corresponds to the point $(1/3, 1/3, 1/3)$, the edges correspond to 2-sparse models, and the vertices to 1-sparse models. Like the spike-and-forest model, the Gibbs prior assigns mass to values of s on the edges and vertices of the simplex while not forcing the values of s to be any particular value. This provides a path through the state-space of s which connects models with differing levels of sparsity, which heuristically should help MCMC schemes that make local changes to the tree structures; by contrast, good samplers for the spike-and-forest prior must "jump" across edges and vertices, which require large, simultaneous, modifications to the tree topologies.

We also see from Figure 2 that the DART prior tends to favor imbalance in s, particularly for small values of α. For example, if $s_{(1)} < s_{(2)} < \cdots < s_{(P)}$ are the order statistics of s,

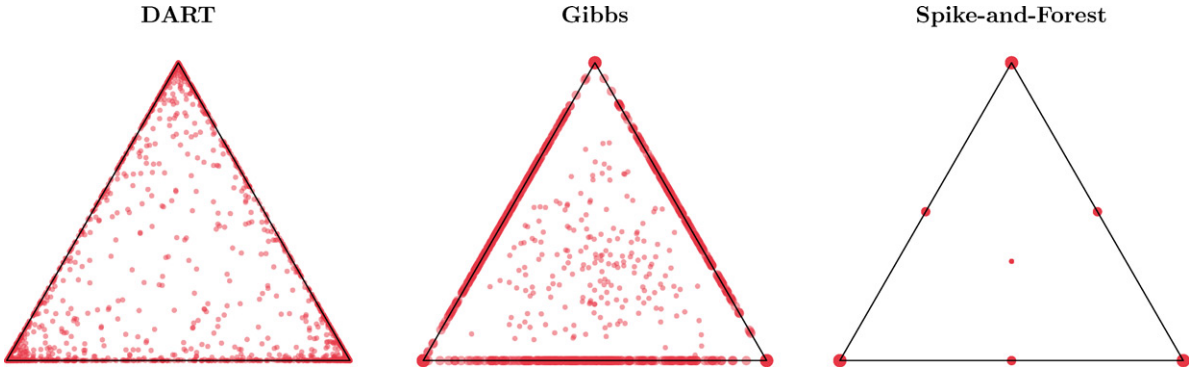

Figure 2. Samples of $s = (s_1, s_2, s_3)$ under the DART, Gibbs, and spike-and-forest priors for $P = 3$. The parameters for each prior are chosen so that the expected model size is roughly 1.5; the value $\alpha = 2.5$ is used for the Gibbs prior.

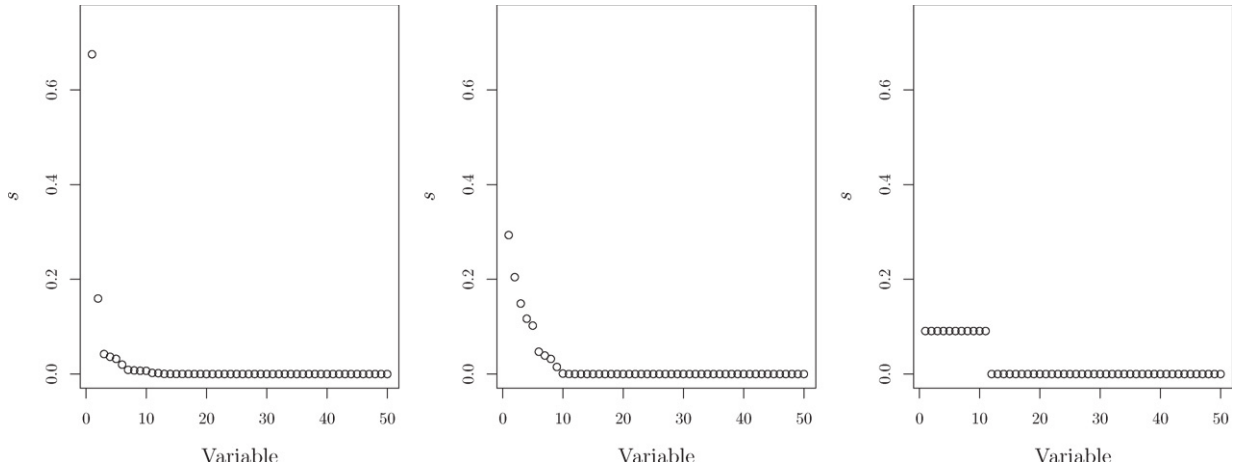

Figure 3. Samples of s from the DART (left), Gibbs (middle), and spike-and-forest (right) priors when each prior is tuned to have 11 out of 50 relevant variables on average. We see that, even when multiple variables are relevant, the DART prior loads most of the weight onto a single predictor, whereas the Gibbs and spike-and-forest priors are more equitable.

then the sparsity-inducing Dirichlet prior causes exponential decay in $s_{(j)}$ as j decreases. This behavior causes most of the weight to be concentrated in a small number of variables even among the *relevant* predictors. This is illustrated in Figure 3, which shows how the s vector decays for a single sample from each of the priors tuned to have 11 variables on average. This property of sparsity-inducing Dirichlet priors has been noted in other contexts as well, such as mixture modeling using Dirichlet processes (Miller and Harrison 2018, sec. 5.3). By contrast, taking $\alpha = 1$ in (2) distributes the weight *uniformly* across the relevant predictors. Another conceptual flaw of the DART prior is that the number of variables included in the model is sensitive to the number of decision rules in the ensemble: because s_j is nonzero for all j, this implies that all P predictors will be included in the model if we allow the number of decision rules in the ensemble B to tend to ∞. While this effect is mild (Linero 2018 shows that the number of predictors included grows logarithmically in B) we would prefer for the number of predictors in the model to be decoupled from the number of trees.

Conceptually, the prior (2) has the same benefits as the spike-and-forest prior: it directly penalizes the model size and performs variable selection through an interpretable spike-and-slab. Like the DART model, it also allows for differential weighting of the relevant variables.

It is not clear yet, however, that the prior (2) possesses the computational tractability of DART; indeed, it seems like we have made the problem more complicated by introducing the additional parameters $\gamma_1, \ldots, \gamma_P$ and hyperprior π_D into the model for s. We can make progress by instead casting (2) in terms of *random partition processes*. The following definition is taken from (Pitman 2002, chap. 2).

Definition 1. A random partition process on $\{1, 2, \ldots, \}$ is a sequence of partitions $\mathcal{C}_1, \mathcal{C}_2, \ldots$ such that $\mathcal{C}_k$ is a partition of $[k] = \{1, \ldots, k\}$ and such that (for all $k' < k$) $\mathcal{C}_k$ and $\mathcal{C}_{k'}$ are *consistent* in the sense that the restriction of $\mathcal{C}_k$ to $[k']$ is equal to $\mathcal{C}_{k'}$. We write $k \sim k'$ if k and k' are members of the same equivalence class in $\mathcal{C}_k$. A random partition process is called *exchangeable* if the mass function $\pi(\mathcal{C}_k)$ depends only on the sizes and number of the equivalence classes of $\mathcal{C}_k$, that is, if we can write the distribution of $\mathcal{C}_k$ in terms of an *exchangeable partition probability function (EPPF)* $\pi(\mathcal{C}_k) = \mathfrak{q}(|c_1|, \ldots, |c_j|)$such that $\mathfrak{q}$ is symmetric, $\sum_j |c_j| = k$, and $\mathfrak{q}(n_1, \ldots, n_j) = \sum_{i=1}^{j} \mathfrak{q}(n_1, \ldots, n_i + 1, \ldots, n_k) + \mathfrak{q}(n_1, \ldots, n_j, 1)$.

In the case of BART priors, a partition naturally arises on the set of branches of the ensemble. Let $\mathcal{B}_k = \{b_1, \ldots, b_k\}$ denote the collection of the first k branches of the ensemble and let $\mathcal{B} = \mathcal{B}_B$ be the collection of all B branches. Then the branches can be partitioned according to whether they split on the same predictor as follows: we let $\mathcal{C}_k$ be a partition of $\{1, \ldots, k\}$ such that $n \sim n'$ if-and-only-if $j_n = j_{n'}$ (here, $n \sim n'$ means that n and n' are members of the same equivalence class of the partition); hence, two elements $b_n, b_{n'}$ of $\mathcal{B}_k$ split on the same predictor if-and-only-if n and n' are members of the same equivalence class of $\mathcal{C}_k$.

The advantage of taking this perspective is that it allows us to avoid working with the γ_j's and s, working instead with the induced prior distribution on $\mathcal{C}_B \equiv \mathcal{C}$ where B is the number of branches. While this may seem even more complicated, it turns out this is actually rather easy provided that the $\mathcal{C}_k$'s are a *Gibbs type* random partition process.

Definition 2. An exchangeable random partition process $\mathcal{C}_1, \mathcal{C}_2, \ldots$ is of *Gibbs type* (Gnedin and Pitman 2006) if $\mathcal{C}_k$ has mass function

$$\pi_k(\mathcal{C}_k) = V_k(|\mathcal{C}_k|) \prod_{c \in \mathcal{C}_k} \frac{\Gamma(\alpha + |c|)}{\Gamma(\alpha)} \tag{3}$$

for some $\alpha > 0$ and positive weights $V_k(q)$ for $q > 0$. We say the process is *P-finite* if $|\mathcal{C}_k| \leq P$ almost-surely for all k, in which case we let $D = \lim_{k \to \infty} |\mathcal{C}_k|$ denote the (random) number of equivalence classes in the limit and let $\pi_D(d)$ denote the induced mass function of D.

We write $\mathcal{C} \sim \text{Gibbs}(V_B, \alpha)$ to denote that $\mathcal{C}$ has the associated Gibbs prior. Exchangeability in the above definition refers to the fact that the probabilities are invariant under relabeling of the natural numbers, for example, the probability of partitioning

$\{1, 2, 3, 4, 5\}$ as $\{\{1, 2\}, \{3, 4, 5\}\}$ is the same as the probability of partitioning $\{2, 3, 5, 6, 7\}$ as $\{\{2, 5\}, \{3, 6, 7\}\}$.

The Proposition 1, which follows from Theorem 3.1 of Miller and Harrison (2018), states that the hierarchical specification (2) is associated to a P-finite Gibbs type prior provided that $\pi_D(d)$ is supported on $\{1, \ldots, P\}$. We use this specification as a default, as it allows us to specify a prior directly on D.

Proposition 1. The prior (2) implies that $\mathcal{C} \sim \text{Gibbs}(V_B, \alpha)$ given B where

$$V_B(t) = \sum_{d=t}^{P} \frac{d!}{(d-t)!} \frac{\Gamma(\alpha\, d)}{\Gamma(\alpha\, d + B)} \pi_D(d). \tag{4}$$

As a special case, the DART prior $s \sim \text{Dirichlet}(\alpha/P, \ldots, \alpha/P)$ corresponds to taking $V_B(t) = \frac{P!}{(P-t)!} \frac{\Gamma(\alpha)}{\Gamma(\alpha+B)}$.

We note that the variable D in Definition 2 *is not* identical to the number of relevant variables Q, as it is possible for not all D variables to be used in the ensemble. Instead we have $Q \to D$ as $B \to \infty$. For the DART prior we have $\pi_D(P) = 1$ (in the sense of Definition 2), another instantiation of the fact that the DART prior will include all predictors as $B \to \infty$.

Once the decision rules have been partitioned, we simply randomly assign each equivalence class a unique predictor.

Definition 3. Let $j^\star = (j_1^\star, \ldots, j_Q^\star)$ denote the unique splitting variables used in the ensemble. We say that $(j_1, \ldots, j_B)$ has a *finite Gibbs prior* if

$$\mathcal{C} \sim \text{Gibbs}(V_B, \alpha) \qquad \text{and} \qquad \pi(j^\star \mid \mathcal{C}) = \frac{(P-Q)!}{P!}$$

for $1 \le j_n^\star \le P$ and $j_n^\star \ne j_{n'}^\star$.

The following result, which is established in Appendix A, characterizes the Gibbs prior in terms of a Pólya urn scheme.

Proposition 2. Let $\mathcal{J}_b = \{j_{b'} : b' \ne b\}$ denote the collection of branches not including the bth. Then we have

$$\pi(j_b = j \mid \mathcal{J}_b) = \begin{cases} \frac{V_B(Q_b)}{V_{B-1}(Q_b)} (\alpha + m_j^{(-b)}) & \text{if } m_j^{(-b)} \ne 0, \\ \frac{V_B(Q_b+1)}{(P-Q_b)\, V_{B-1}(Q_b)} \alpha & \text{otherwise,} \end{cases} \tag{5}$$

where $m_j^{(-b)}$ denotes the number of times the jth variable is used in a splitting rule excluding the bth branch and Q_b is the number of unique variables in $\mathcal{J}_b$. In the special case of the spike-and-forest prior, we have

$$\pi(j_b = j \mid \mathcal{J}_b) = \begin{cases} \frac{V'_B(Q_b)}{V'_{B-1}(Q_b)} & \text{if } m_j^{(-b)} \ne 0, \\ \frac{V'_B(Q_b+1)}{(P-Q_b)\, V'_{B-1}(Q_b)} & \text{otherwise,} \end{cases} \tag{6}$$

where $V'_B(t) = \sum_{d=t}^{P} \frac{d!}{(d-t)!\, d^B} \pi_D(d)$.

As noted by Miller and Harrison (2018), the mechanism for partitioning the j_b's in this fashion is analogous to the Pólya urn scheme of Blackwell and MacQueen (1973) that gives rise to the Dirichlet process. Using the Pólya urn representation, we can alternately describe the generative mechanism for the trees by modifying step G4 in Section 2.1 to state the following:

G4A. Assign each branch node b a splitting rule of the form $[x_{j_b} \le C_b]$ by sampling j_b according to Proposition 2 with $\mathcal{J}_b$ consisting of all previously sampled splitting rules. Then sample C_b as in G4.

An interesting aspect of this characterization is that it provides an alternate explanation of the sparsity-inducing properties of the prior—the scheme for generating splitting rules reinforces the coordinates which have been used in the past, with V_B and α determining how strongly previously chosen variables are reinforced. Because they are special cases, the same characterization in terms of reinforcement holds for the DART and spike-and-forests priors.

Importantly, because Gibbs type priors are exchangeable, the probabilities in Proposition 2 remain the same regardless of the order in which the j_b's are generated.

Proposition 3. To sample splitting rules from the Gibbs prior it suffices to iteratively sample rules $j_1, \ldots, j_B$ successively from (5) with $\mathcal{J}_b$ replaced by $\mathcal{J}_{b_n} = \{j_{b_{n'}} : n' < n\}$ and B replaced by n.

To implement the Gibbs prior we need to compute $V_B(t)$ for potentially many different values of (t, B). Fortunately, if we use (2) with a fixed choice of π_D then these values only need to be computed once. In our implementations, whenever we require $V_B(t)$ we first check if it has already been computed; if it has not, then we compute it using (4) and store the result in a hash table, while if it has then we simply retrieve the value from the table. This ensures that computation of $V_B(t)$ requires negligible overhead relative to sampling the model via MCMC. A less efficient, but still practical, solution is to compute $V_B(t)$ for all plausible values of (t, B) prior to running the chain, as done by Miller and Harrison (2018).

2.4. Default Priors

To specify a Gibbs prior we need to select (π_D, α, V_B). As a default we recommend using the Gibbs prior associated with (2) so that V_B is chosen according to (4). We then specify a uniform prior on the selected variables by taking $\alpha = 1$.

An advantage of the Gibbs prior is that the choice of $\pi_D(d)$ is effectively arbitrary and makes little difference in terms of computations. By default, we encode a preference for sparsity by using a *truncated zeta distribution* $\pi_D(d) = d^{-\zeta} / \sum_{p=1}^{P} p^{-\zeta}$. For $\zeta > 0$ the truncated zeta distribution encodes a preference for sparsity while when $\zeta = 0$ it induces a uniform prior on $\{1, \ldots, P\}$. In our illustrations we set $\zeta = 1$.

For $\zeta \le 2$ the zeta prior has the feature that, as $P \to \infty$, the prior expectation of D diverges, while it does not for $\zeta > 2$. One can, for example, choose ζ specifically to target a desired prior mean for D. Alternatively, one can use a complexity penalizing prior that more heavily penalizes the number of predictors included in the model. For example, Rockova and van der Pas (2020) study geometric priors of the form $\pi_D(d) \propto c^{-d}\, P^{-a\, d}$ for some positive constants (a, c).

For the remaining parameters in the model $(\gamma, \beta, \sigma^2, \sigma_\mu^2)$ we use the default settings described by Linero and Yang (2018) after scaling the outcome Y_i to lie in $[-0.5, 0.5]$. For the sake of self-containment, these choices are $\gamma = 0.95$, $\beta = 2$, $\sigma_\mu =$

$0.5/(k\sqrt{T})$ where $k = 2$, and $\sigma \sim \text{Cauchy}_+(0, \widehat{\sigma})$ where $\widehat{\sigma}$ is a pilot estimate of σ obtained by fitting the lasso to the data.

3. Markov Chain Monte Carlo for Gibbs Priors

In this section we develop a Metropolis-Hastings algorithm for fitting BART models that incorporate the Gibbs prior. Fortunately, the Pólya urn characterization in Proposition 2 makes it easy to derive the Metropolis-Hastings acceptance ratios.

BART models are typically updated using a two-stage Metropolis-within-Gibbs algorithm. We define $\mathcal{T}_{-t} = \{\mathcal{T}_m : m \neq t\}$ and $\mathcal{M}_{-t}$ similarly. We let $q(\mathcal{T}' \mid \mathcal{T})$ denote a yet-to-be-specified Markov transition function on the state space of trees. The following steps are used to update the pair $(\mathcal{T}_t, \mathcal{M}_t)$.

1. Propose a new tree structure $\mathcal{T}' \sim q(\mathcal{T}' \mid \mathcal{T}_t)$ then set $\mathcal{T}_t \leftarrow \mathcal{T}$ with probability

$$A = 1 \wedge \frac{\pi_{\mathcal{T}}(\mathcal{T}')\,\Lambda(\mathcal{T}' \mid \mathcal{T}_{-t}, \mathcal{M}_{-t}, \boldsymbol{X}, \boldsymbol{Y})\, q(\mathcal{T}_t \mid \mathcal{T}')}{\pi_{\mathcal{T}}(\mathcal{T}_t)\,\Lambda(\mathcal{T}_t \mid \mathcal{T}_{-t}, \mathcal{M}_{-t}, \boldsymbol{X}, \boldsymbol{Y})\, q(\mathcal{T}' \mid \mathcal{T}_t)},$$

   where Λ is the integrated likelihood function

$$\Lambda(\mathcal{T} \mid \mathcal{T}_{-t}, \mathcal{M}_{-t}, \boldsymbol{X}, \boldsymbol{Y}) = \prod_{\ell \in \mathcal{L}(\mathcal{T})} \int \text{Normal}(\mu \mid 0, \sigma_\mu^2) \prod_{i: X_i \rightsquigarrow \ell} \text{Normal}(Y_i \mid \lambda_i + \mu, \sigma^2)\, d\mu$$

   and $\lambda_i = \sum_{m \neq t} g(X_i; \mathcal{T}_m, \mathcal{M}_m)$.
2. Draw $\mathcal{M}_t$ from its full conditional distribution.

The particular formulas for Λ and the full conditional of $\mathcal{M}_t$ are not of relevance to us, however, and we refer interested readers to Kapelner and Bleich (2016) for details. To lighten notation, we will write $\Lambda(\mathcal{T}) = \Lambda(\mathcal{T} \mid \mathcal{T}_{-t}, \mathcal{M}_{-t}, \boldsymbol{X}, \boldsymbol{Y})$, with the relevant index t inferred by the reader from context. Both steps rely critically on the conjugacy of the normal prior to the normal likelihood. Hill, Linero, and Murray (2019) show more generally that this can be carried out for other conjugate likelihood/prior pairs such as log-linear (Murray 2021) and gamma (Linero, Sinha, and Lipsitz 2020) BART models.

The proposal distribution $q(\mathcal{T} \mid \mathcal{T}')$ we use is a mixture of local modifications to $\mathcal{T}$. We consider the following steps.

BIRTH Convert a randomly chosen leaf node ℓ of depth d into a branch node with two children ℓ_L and ℓ_R by randomly selecting a predictor j with probability (5)—or (6) for the spike-and-forest prior—and randomly selecting a cutpoint $C_\ell \sim \text{Uniform}(A_j, B_j)$ where $\prod_{p=1}^P [A_p, B_p]$ is the hyperrectangle associated to $\{x : x \rightsquigarrow \ell\}$.

DEATH Convert a randomly chosen branch b with exactly two child nodes of depth d into a leaf node by deleting its children.

PRIOR Sample $\mathcal{T}'$ from its prior distribution conditional on $(\mathcal{T}_{-t}, \mathcal{M}_{-t})$ using (5)—or (6) for the spike-and-forest prior—to sample splitting coordinates conditional on all other existing splits in the ensemble.

The following proposition gives the associated Metropolis-Hastings acceptance probabilities. Recall that $\rho(d) = \gamma(1 + d)^{-\beta}$ is the prior probability that a node of depth d is a branch.

Proposition 4. For the BIRTH, DEATH, and PRIOR moves, respectively, a valid acceptance probability is given by $A = 1 \wedge R$ where

$$\begin{aligned} R_{\texttt{BIRTH}} &= \frac{\rho(d)\,\{1-\rho(d+1)\}^2}{1-\rho(d)} \times \frac{\Lambda(\mathcal{T}')}{\Lambda(\mathcal{T})} \times \frac{q_{\texttt{DEATH}}(\mathcal{T}')\,|\mathcal{L}(\mathcal{T})|}{q_{\texttt{BIRTH}}(\mathcal{T})\,|\text{NOG}(\mathcal{T}')|} \\ R_{\texttt{DEATH}} &= \frac{1-\rho(d)}{\rho(d)\,\{1-\rho(d+1)\}^2} \times \frac{\Lambda(\mathcal{T}')}{\Lambda(\mathcal{T})} \times \frac{q_{\texttt{BIRTH}}(\mathcal{T}')\,|\text{NOG}(\mathcal{T})|}{q_{\texttt{DEATH}}(\mathcal{T})\,|\mathcal{L}(\mathcal{T}')|} \\ R_{\texttt{PRIOR}} &= \frac{\Lambda(\mathcal{T}')}{\Lambda(\mathcal{T})}, \end{aligned}$$

and $\text{NOG}(\mathcal{T})$ denotes the set of branches with exactly two children which are leaves, $q_{\texttt{DEATH}}(\mathcal{T})$ denotes the probability of proposing a DEATH to modify $\mathcal{T}$, $q_{\texttt{BIRTH}}(\mathcal{T})$ denotes the probability of proposing a BIRTH to modify $\mathcal{T}$, and $|A|$ denotes the number of elements in the set A.

The expressions above are particularly simple because we have incorporated Proposition 2 into the proposal distribution, causing the entire Gibbs prior structure to drop out of the acceptance ratios; in fact, Proposition 4 gives the same acceptance ratios given by Kapelner and Bleich (2016). Hence, implementing the Gibbs prior only requires modifying the proposal distribution $q(\mathcal{T} \mid \mathcal{T}')$ in the BIRTH, DEATH, and PRIOR moves of existing BART implementations.

4. Illustrations

4.1. Simulation Experiment

We conduct a simulation experiment to illustrate the salient features of the Gibbs prior. We consider the data generating mechanism $Y_i \sim \text{Normal}\{r_0(X_i), \sigma^2\}$ where

$$r_0(x) = 10 \sin(\pi\, x_1\, x_2) + 20(x_3 - 0.5)^2 + 10\, x_4 + 5\, x_5. \quad (7)$$

This regression function possess nonlinearities, interactions, and linear terms. This simulation scenario was considered first by Friedman (1991) and has since been used many times in the BART literature as a benchmark. We allow for $P > 5$ so that x_j is irrelevant for all $j > 5$.

Each of the methods under consideration (DART, spike-and-forest, and Gibbs priors) are fit with $T \in \{50, 200, 500\}$ trees and 10,000 samples from the Markov chain, with the first 5000 discarded to burn-in. Both the spike-and-forest and Gibbs priors use the truncated zeta distribution with $\zeta = 1$ for the model size, while the DART prior uses the hyperprior $\alpha/(\alpha + P) \sim \text{Beta}(0.5, 1)$. We use the Gibbs prior implied by (2) with $\alpha \equiv 1$ or, equivalently, we let the nonzero components of s be uniformly distributed on the simplex $\mathbb{S}_{D-1}$.

We first fit the spike-and-forest and Gibbs prior models to the data with $N = 250$, $P \in \{7, 50, 150, 400, 1000\}$ and $\sigma \in \{3, 5\}$. Some results of a single fit are given in Figure 4 when $\sigma = 3$ and $P = 1000$. From the posterior inclusion probability plot we see that both the DART and the Gibbs prior are effective at removing irrelevant predictors, whereas the spike-and-forest prior does not choose the correct model (using a posterior inclusion probability cutoff of 0.5). The middle and bottom rows of Figure 4 suggest that this is partially due to poor mixing for the spike-and-forest prior: the mixing of the model size is substantially worse for the spike-and-forest prior due to the fact that our MCMC scheme has trouble eliminating predictors. This

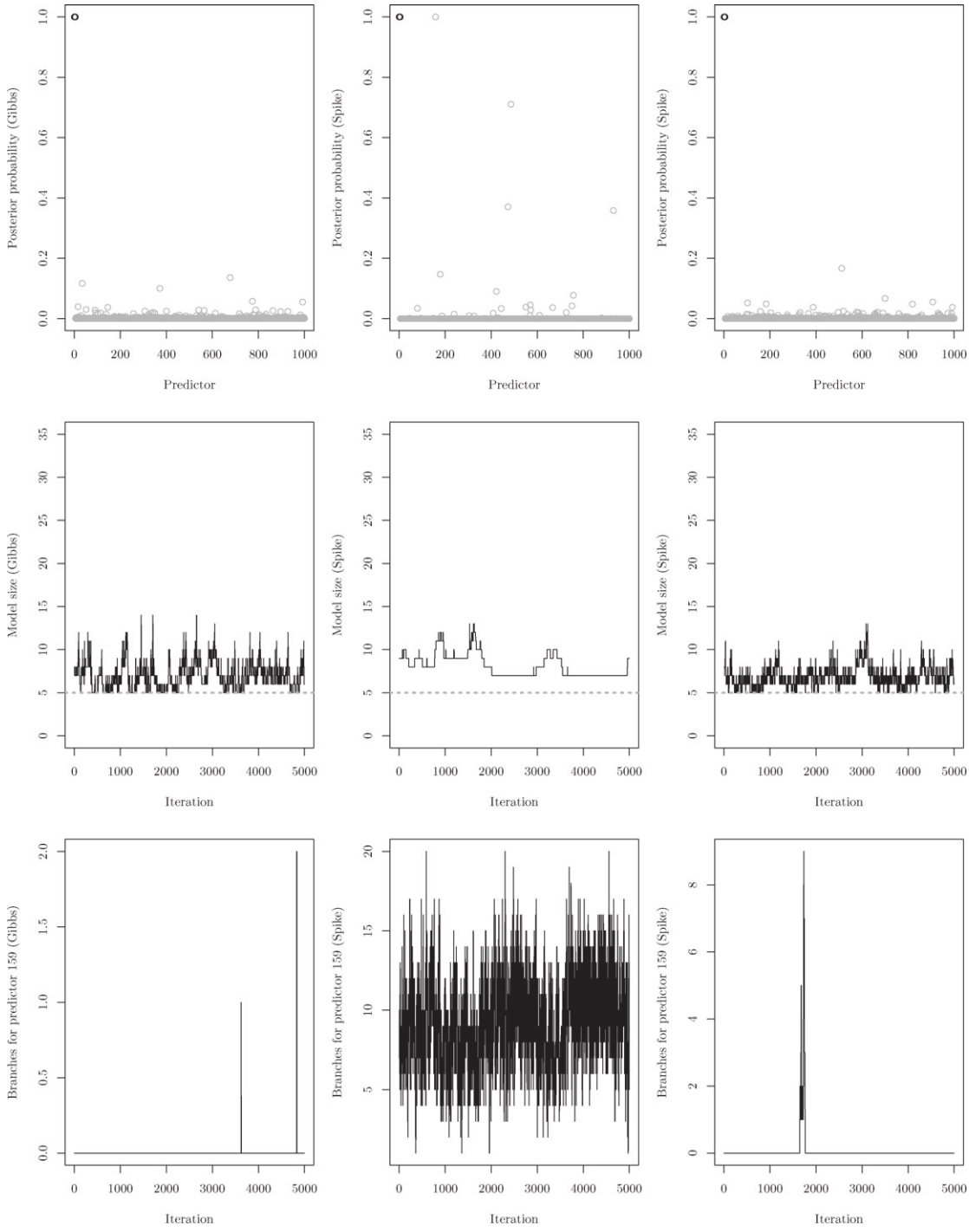

Figure 4. Results of the simulation study of Section 4.1 for the Gibbs (left), spike-and-forest (middle), and DART (right) for a single run. Top: posterior inclusion probabilities of the variables, with spurious predictors in gray. Middle: traceplot of the model size. Bottom: traceplot of the number of times variable x_{159} is included, a spurious variable which was included by the spike-and-forest prior but not the Gibbs prior.

is apparent in the bottom row, where we look at the mixing of the spuriously selected variable x_{159}: we see that the the DART and Gibbs priors are capable of including, using, and then subsequently removing this variable, whereas the MCMC scheme for the spike-and-forest model cannot remove this variable once it has been included in the model.

We replicated the simulation 200 times for each combination of P and σ and computed the following performance metrics. Below, the number of "positives" is the number of predictors which are included in the median probability model (Barbieri and Berger 2004) and the number of "true positives" is the number of predictors included in the median probability model which are relevant (we similarly define the "false positives," "negatives," and so forth).

Precision. The ratio of true positives to total positives.

Recall. The ratio of true positives to the total number of relevant predictors.

F_1 Score. The harmonic mean of the precision and recall, used as an omnibus measure of variable selection accuracy.

Table 1. Results of the simulation study; standard errors for all quantities are less than 0.01.

			$T = 50$			$T = 200$			$T = 500$		
P	σ	Method	Prec	Rec	F_1	Prec	Rec	F_1	Prec	Rec	F_1
7	3	DART	0.79	1.00	0.88	0.75	1.00	0.85	0.72	1.00	0.84
		Gibbs	***0.98***	1.00	***0.99***	***0.98***	1.00	***0.99***	***0.98***	1.00	***0.99***
		SAD	0.95	1.00	0.97	0.71	1.00	0.83	0.71	1.00	0.83
	5	DART	0.72	1.00	0.83	0.71	1.00	0.83	0.71	1.00	0.83
		Gibbs	***0.95***	1.00	***0.97***	**0.91**	1.00	**0.95**	**0.92**	1.00	**0.95**
		SAD	0.88	0.99	0.93	0.71	1.00	0.83	0.71	1.00	0.83
50	3	DART	0.95	1.00	0.97	0.92	1.00	0.96	0.91	1.00	0.95
		Gibbs	***0.98***	1.00	***0.99***	***0.97***	1.00	**0.98**	**0.97**	1.00	**0.98**
		SAD	0.88	1.00	0.93	0.23	1.00	0.38	0.10	1.00	0.19
	5	DART	0.85	**0.96**	0.89	0.77	0.96	0.84	0.72	0.97	0.80
		Gibbs	***0.92***	0.94	***0.92***	**0.91**	0.94	***0.92***	***0.92***	0.93	***0.92***
		SAD	0.82	0.95	0.87	0.20	**0.97**	0.34	0.10	***1.00***	0.19
150	3	DART	0.95	1.00	0.97	0.94	1.00	0.97	0.93	1.00	0.96
		Gibbs	***0.98***	1.00	***0.99***	**0.97**	1.00	**0.98**	**0.97**	1.00	**0.98**
		SAD	0.83	1.00	0.91	0.23	1.00	0.37	0.08	1.00	0.16
	5	DART	0.88	0.91	0.89	0.87	0.92	0.88	0.82	0.92	0.86
		Gibbs	***0.93***	0.90	***0.90***	**0.92**	0.90	**0.90**	**0.91**	0.90	**0.89**
		SAD	0.82	**0.92**	0.86	0.22	**0.94**	0.36	0.07	***0.96***	0.14
400	3	DART	0.94	1.00	0.97	0.92	1.00	0.96	0.92	1.00	0.96
		Gibbs	***0.98***	1.00	***0.99***	**0.95**	1.00	**0.97**	**0.94**	1.00	**0.96**
		SAD	0.81	1.00	0.89	0.22	1.00	0.37	0.08	0.99	0.16
	5	DART	0.89	**0.86**	0.86	0.86	0.88	0.85	0.84	0.88	0.84
		Gibbs	***0.92***	0.85	***0.87***	**0.91**	0.85	***0.87***	**0.89**	0.84	**0.85**
		SAD	0.80	**0.86**	0.82	0.25	***0.90***	0.38	0.10	***0.90***	0.18
1000	3	DART	0.98	**0.99**	***0.98***	**0.97**	**0.99**	**0.97**	0.88	***1.00***	0.93
		Gibbs	***0.99***	0.98	***0.98***	**0.97**	0.98	**0.97**	**0.94**	0.99	**0.96**
		SAD	0.80	**0.99**	0.88	0.22	0.98	0.36	0.09	0.97	0.16
	5	DART	***0.98***	0.74	***0.84***	**0.95**	0.74	0.82	0.84	0.81	**0.81**
		Gibbs	0.93	0.78	***0.84***	0.90	0.79	**0.83**	**0.86**	0.79	**0.81**
		SAD	0.80	**0.80**	0.79	0.30	**0.84**	0.43	0.14	***0.85***	0.24

NOTE: The precision, recall, and F_1 scores represent averages of these quantities over 200 replications of the experiment. Bold text denotes best performance for a given value of T, while bold italic text denotes best performance cross *all* values of T.

Results are given in Table 1.

We note at the outset that the spike-and-forest prior, as implemented using the same MCMC scheme as the Gibbs prior, performs poorly across all settings (especially so when $T = 200$ or $T = 500$). We attribute this primarily to the poor mixing of the local sampler for the spike-and-forest prior seen in Figure 4 rather than to any fundamental problem with the spike-and-forest model itself; at a minimum, the mixing issues are severe enough that we cannot rule out that this is what is causing the poor performance.

The results for $P = 7$ reveal one of the benefits of the Gibbs prior relative to DART: the default DART prior recommended by Linero (2018) tends to be ineffective at filtering out irrelevant predictors when P is small. This is because DART provides only indirect control on the number of variables included in the model, and is more tolerant of variables which have miniscule impact on the outcome than the spike-and-forest or Gibbs priors. Conversely, the Gibbs prior (which explicitly penalizes the inclusion of unnecessary predictors) performs well in this setting.

The number of trees T included in the model has a relatively minor impact on the Gibbs and DART priors, with these models performing slightly better when using fewer trees; the finding that using fewer trees may result in better variable selection is consistent with other results in the literature (Bleich et al. 2014). Speaking generally, however, the recall of the DART and Gibbs priors are fairly robust to the choice of T, while the precision of the Gibbs prior tends to decay more slowly as T is increased (for example, the mean precision of the Gibbs prior for $(P, \sigma) = (50, 5)$ is the same for $T = 50$ and $T = 500$, while the precision of DART decreases from 85% to 72%.) We also see that for larger values of P (400 and 1000) the recall for all methods drops. This is due to systematic omissions of the two lowest-signal variables (x_3 and x_5) for all methods, which is caused both by a lack of signal and the large number of noise variables making the variable selection task intrinsically more difficult.

Summarizing the main points: DART, being more tolerant of spurious variables, naturally performs worse in terms of precision and better in terms of recall than the Gibbs prior. The Gibbs prior has the advantage of performing well across the board when P is small, and also has the advantage of having a precision which decays only mildly as T increases, making it more robust to the number of trees in the ensemble than DART.

Computational Cost. In Table 2 we give a brief comparison of the time to collect 10,000 samples from the approximate

Table 2. Time to fit the DART and Gibbs priors in seconds for various T and P; Ratio denotes the time of Gibbs divided by the time of DART.

T	P	DART	Gibbs	Ratio
50	100	73	86	1.18
50	1000	73	88	1.21
200	100	279	340	1.22
200	1000	279	356	1.28

posterior distribution for the DART and Gibbs priors under our simulation scenario with $N = 250$ and $\sigma = 1$. We generally expect for Gibbs to be more expensive due to both the cost of computing $V_B(t)$ and the more involved sampling of the candidate predictors for each split. From Table 2 we see that the Gibbs prior is only slightly slower (20%–30%) computationally than DART.

4.2. Analysis of Benchmark Datasets

We compare the performance of Gibbs priors to DART on four publicly available benchmark datasets: the `Boston` dataset available in the `MASS` package, the `Hitters` dataset available in the `ISLR` package, the `WIPP` dataset provided by Storlie et al. (2011), and the `Tecator` dataset available from the datasets archive of StatLib at *http://lib.stat.cmu.edu/datasets/*. We do not include the spike-and-forest prior due to its aforementioned mixing problems. We evaluate the priors using 5-fold cross-validation replicated four times (so that each model was fit 20 times in total). Results for the different datasets are given in Table 3. We measure accuracy in terms of the heldout root mean-squared error $\sqrt{N^{-1}\sum_i(\widehat{Y}_i - Y_i)^2}$ (RMSE) where $\widehat{Y}_i$ is the predicted value of Y_i computed on the fold with observation i held out. We also report the average size of the median probability model across the 20 fits (Model Size) and the proportion of the 20 train/test splits that the Gibbs prior outperforms DART and vice-versa (%Superior); for example, if %Superior is 100% for Gibbs then Gibbs outperformed DART across all 20 splits. Each fit used $T = 50$ trees and the chains were run for 15,000 iterations with the first 5000 iterations discarded to burn-in.

The results in Table 3 show that the Gibbs prior obtains better results than DART using fewer predictors on average. For all datasets considered, the Gibbs prior made use of fewer variables in the median probability model while also obtaining a lower RMSE on average over all train/test splits. The Gibbs prior also obtained lower RMSE on substantially more of the splits than DART, with the only marginal case occurring on the `WIPP` dataset. When comparing across the partitions into 5-folds (as opposed to comparing across both the partitions and the folds themselves), Gibbs outperformed DART on 100% of the partitions. Lastly, we note that the Gibbs prior with the truncated zeta choice for $\pi_D(d)$ is flexible enough to allow for dense models when the underlying data generating process is dense; we see this for the `Boston` dataset, where both the Gibbs and DART priors include the majority of the variables.

Table 3. Results of the 5-fold cross-validation experiment on the `Boston`, `Hitters`, `Tecator`, and `WIPP` datasets.

Dataset	Method	RMSE	Model size	%Superior
`Boston`	DART	1.09	10.7	15%
	Gibbs	**1.00**	**10.3**	**85%**
`Hitters`	DART	1.12	12.7	15%
	Gibbs	**1.00**	**7.6**	**85%**
`Tecator`	DART	1.09	15.9	20%
	Gibbs	**1.00**	**8.2**	**80%**
`WIPP`	DART	1.06	11.7	40%
	Gibbs	**1.00**	**8.6**	**60%**

NOTE: Best results per dataset are given in bold.

We now consider the `Hitters` dataset in detail. This dataset is described fully in the `ISLR` package (James et al. 2021) and "…is part of the data that was used in the 1988 ASA Graphics Section Poster Session," where an inferential goal was to determine which variables are predictive of a baseball player's salary. The dataset contains measurements of the salaries of Major League Baseball (MLB) players during the 1986–1987 season, along with some candidate predictors of salary: number of years playing in MLB, total number of hits both during the current season and up-to the current season, and so forth. In total there are 19 predictors. We find that, on the `Hitters` dataset, there is evidence that the Gibbs prior outperforms DART: the heldout RMSE is lower both on average and consistently over many splits into train/test sets. Additionally, it uses fewer of the predictors. Posterior inclusion probabilities for the DART and Gibbs prior are given in Figure 5. Examining the predictors used, we find that the Gibbs prior focuses on measures of past performance—the cumulative hits, runs, runs-batted-in, at-bats, and years in the league *prior* to the current season. DART instead uses features which measure player performance both in past seasons and the current season; while metrics of performance in the current season are possibly useful from a predictive perspective (the Gibbs prior prefers number of walks in the current season to the cumulative number of walks) they cannot causally determine the salary since the salary is set before the season. Given the redundancy in these features, we might expect that it is unnecessary to use both current and past performance, with a preference for past rather than current performance. This preference is reflected in the results for the Gibbs prior, but not for the DART prior.

Next we consider the `WIPP` dataset described in Storlie et al. (2011). This dataset comes from a two phase fluid flow computer model for a Waste Isolation Pilot Plant; the goal considered by Storlie et al. (2011) is to predict the cumulative brine flow into the plant at 10,000 years assuming a drilling intrusion at 1000 years. There are 31 input variables, but not all of them are known to be relevant for prediction. Posterior inclusion probabilities for this model for the Gibbs and DART priors are given in Figure 6. We see again that the Gibbs and DART priors agree on the most relevant variables, but that the Gibbs prior more aggressively removes the less useful variables.

In the supplementary materials we also provide traceplots of the fit to the `WIPP` and `Hitters` datasets. There we see that the Gibbs prior tends to mix better than the DART prior—this is somewhat expected due to the fact that the Gibbs prior makes use of a collapsed Gibbs sampler, that is, we have integrated out s).

5. Other Classes of Weak Learners

In this section we show how to apply Gibbs priors to classes of weak learners beyond the decision tree models considered here. We consider a class of weak learners with the same variable inclusion structure as decision trees: the function $b(x;\gamma)$ consists of some number of variables $\{j_1,\ldots,j_K\}$ where the variables j_b are allowed to repeat. We list some possible examples below:

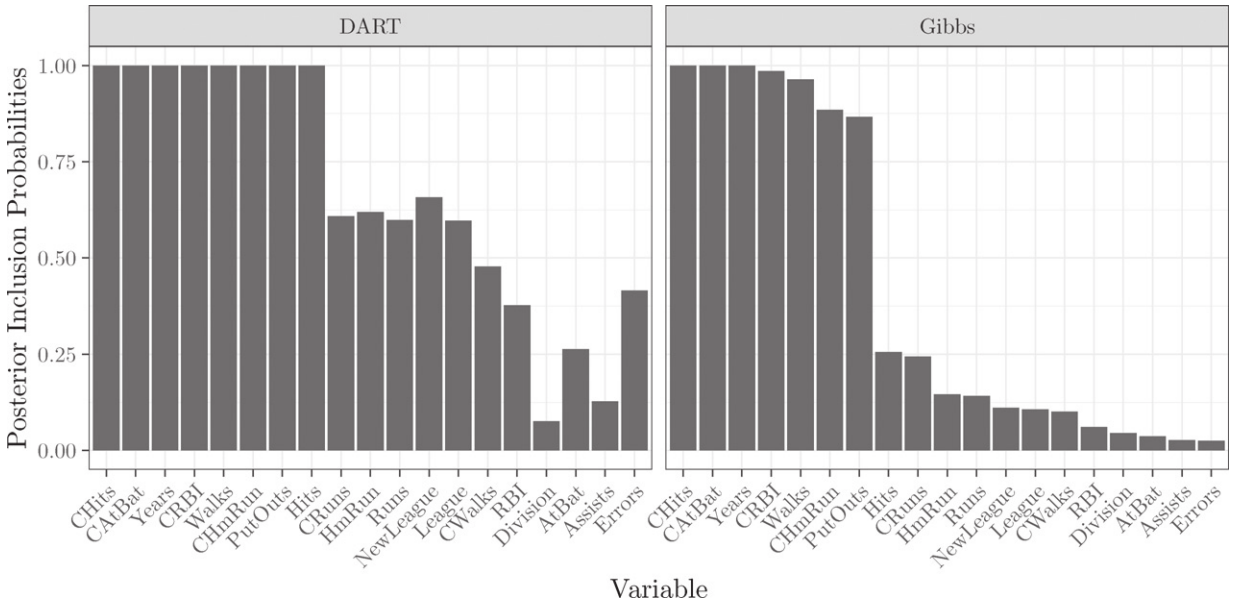

Figure 5. Posterior inclusion probabilities for the variables in the `Hitters` dataset for the DART and Gibbs priors. A description of these variables is given in the `ISLR` package in `R`. Variables are ordered by their posterior inclusion probability under the Gibbs prior.

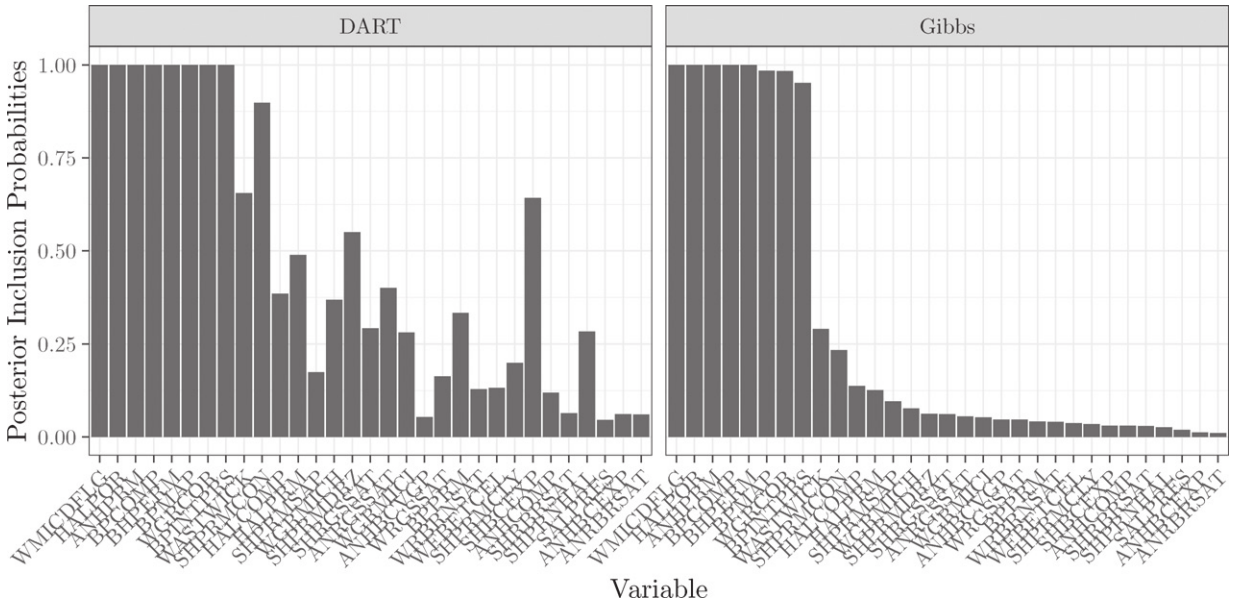

Figure 6. Posterior inclusion probabilities for the variables in the `WIPP` dataset for the DART and Gibbs priors. A description of these variables is given in Storlie et al. (2011). Variables are ordered by their posterior inclusion probabilities under the Gibbs prior.

- Polynomial learners of the form $b(x;\gamma) = \prod_{k=1}^{K} x_{j_k}$.
- Multivariate adaptive regression splines (MARS) of the form $b(x;\gamma) = \prod_{k=1}^{K} \max\{0, Z_k(x_{j_k} - c_k)\}$ where $Z_k \in \{-1, 1\}$ (Friedman 1991; Denison, Mallick, and Smith 1998).
- Radial basis functions of the form $b(x;\gamma) = \exp\{-\rho \sum_{j=1}^{P} Z_j(x_j - \mu_j)^2\}$ where Z_j is the number of times variable j is selected.

In the case of MARS, for example, γ_m consists of $\{Z_k^{(m)}, c_k^{(m)}, j_k^{(m)} : k = 1, \ldots, K_m\}$. We can then apply the Gibbs prior to the collection $\{j_k^{(m)} : 1 \le k \le K_m, 1 \le m \le M\}$ of coordinates, that is, we randomly partition the $j_k^{(m)}$'s and then assign a unique predictor to each equivalence class.

To illustrate, we fit Bayesian MARS to $N = 250$ samples generated from the model (7) with $P = 500$ predictors and $\sigma = 1$. We used $M = 50$ basis functions. We set $\zeta = 1$ and chose K_m to take the values 0, 1, and 2 with prior probability 10%, 50%, and 40%. For the model parameters we set $\beta_m \sim \text{Normal}(0, \sigma_m^2)$ and $Z_k = \pm 1$ with equal probability. We specify half-Cauchy priors $\sigma \sim \text{Cauchy}_+(0, \widehat{\sigma})$ and $\sigma_\beta \sim \text{Cauchy}_+(0, \widehat{\sigma}_\beta)$ where $\widehat{\sigma}$ is a pilot estimate of the standard deviation obtained from fitting a lasso to the scaled data and $\widehat{\sigma}_\beta = M^{-1/2}$. In the supplementary materials we give a simple Bayesian backfitting algorithm for fitting this model.

Results are given in Figure 7 for a single fit of the model using (i) the Gibbs prior and (ii) a prior which selects each covariate in the basis function uniformly at random from the possible

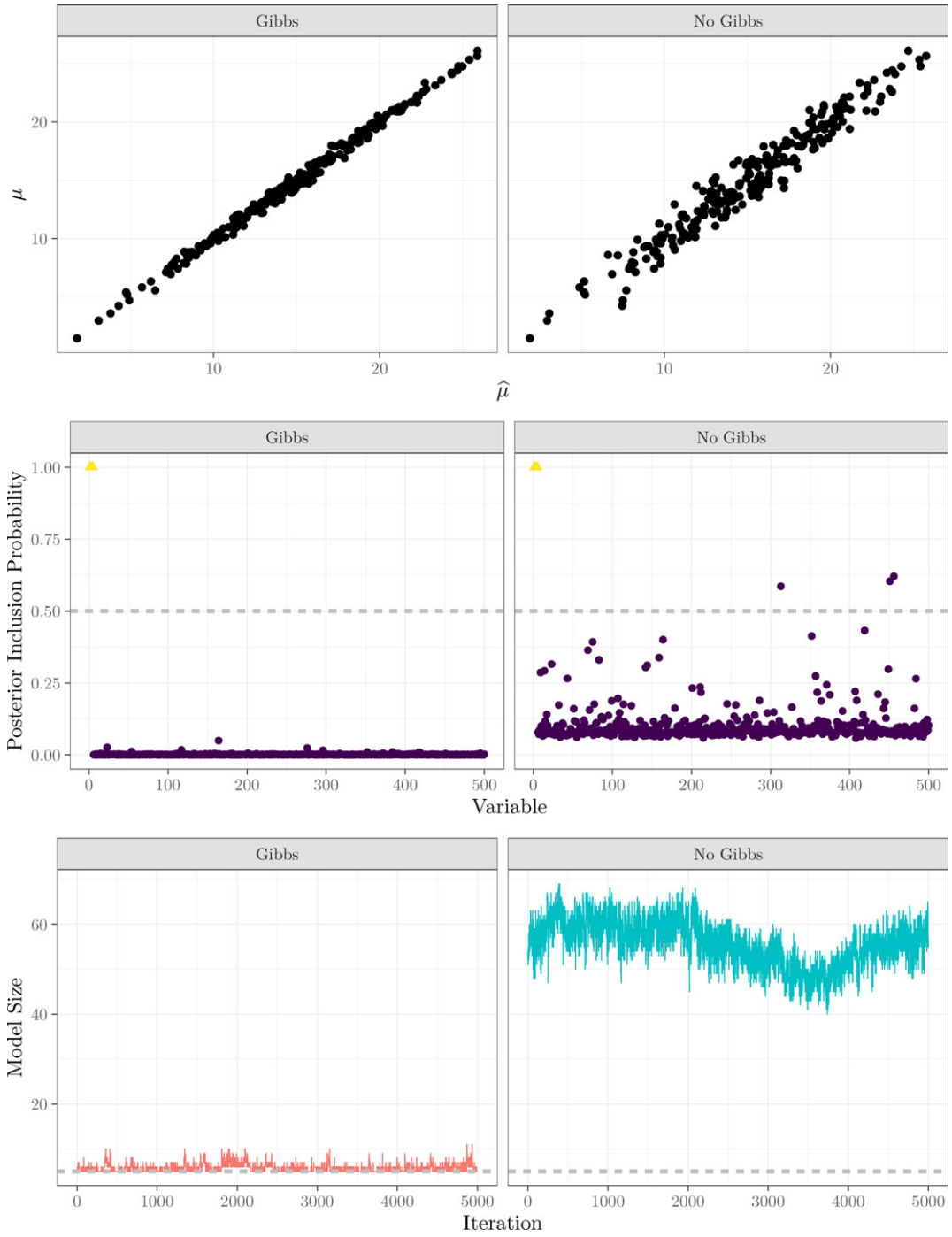

Figure 7. Top: predictions from the MARS models plotted against the true values of the function on a test set of 250 held out X_i's. Middle: posterior inclusion probabilities for the MARS models with and without the Gibbs prior; relevant predictors are given by yellow triangles and the gray line at 0.5 is the cutoff for inclusion in the median probability model. Bottom: traceplot of the total number of predictors used on a given sample from the posterior, obtained from the Bayesian backfitting algorithm with and without the Gibbs prior; the gray line is the true model size (5).

covariates; the behavior we see is representative of replications of the experiment. We see from Figure 7 that, when combined with the Gibbs prior, the MARS basis function expansion performs extremely well; we have found that it outperforms BART and is generally competitive with the SBART algorithm of Linero and Yang (2018) on this problem. The Gibbs prior is capable of accurately filtering out irrelevant variables, and only selects the five relevant predictors for inclusion in the model. The Gibbs prior also prefers parsimonious models, with 61% of the samples from the posterior containing only the relevant predictors.

By contrast, the predictive performance of Bayesian MARS without the Gibbs prior is rather poor, and it performs substantially worse than DART on this problem (see the results of Linero 2018). Without the Gibbs prior we find that Bayesian MARS includes several spurious covariates, both in the median probability model and in the individual samples of $r(x)$ from the posterior; to the extent that it eliminates most of the irrelevant variables from the median probability model, the bottom row of Figure 7 shows that this is simply due to the fact that the model can only accommodate so many variables with $M = 50$

basis functions. On average, the model without the Gibbs prior includes 55 variables, although which irrelevant predictors are included changes from sample to sample. In predictive terms, the root mean-squared error on a held out set of covariates $(X_1^\star, \ldots, X_N^\star)$ given by $\sqrt{N^{-1}\sum_i (\widehat{r}(X_i^\star) - r_0(X_i^\star))^2}$ is roughly one third as large when using the Gibbs prior (RMSE $= 0.36$) relative to not using the Gibbs prior (RMSE = 1.06).

We conclude from this example that random basis function expansions, such as Bayesian MARS, can perform extremely well with the Gibbs prior for variable selection. We note that, remarkably, Bayesian MARS substantially outperforms the MARS algorithm as implemented in the `earth` package in R (RMSE $= 1.26$ under the constraint $K_m \leq 2$), an observation which agrees with the simulation results of Linero (2018). Like our Gibbs prior, the MARS algorithm also allows for penalizing the inclusion of additional variables to induce sparsity; tuning of this parameter by 10-fold cross-validation resulted in a test-set RMSE of 1.22, and hence still performs far worse than the Bayesian variant. The same is true when one compares BART to its Frequentist variant of boosted decision trees (Linero 2018; Linero and Yang 2018). One might conjecture on this basis that the main strength of BART is not its use of decision trees, but rather that ensembling of weak learners in the Bayesian framework is powerful in general. Given that implementing BART is a much more arduous task than implementing our Bayesian MARS model, we believe that using weak learners other than decision trees may be a better starting point for researchers.

6. Discussion

Much of our work is inspired by Bayesian nonparametric methods, which have long used Gibbs distributions as priors on exchangeable random partitions for mixture modeling (De Blasi et al. 2013); in a sense, replacing a DART prior with a Gibbs prior for us is analogous to the move from Dirichlet process mixture models to the mixtures-of-finite-mixtures models of Miller and Harrison (2018). By borrowing the notion of a Gibbs prior from the theory of random partitions, we have introduced a method for performing nonparametric variable selection using Bayesian decision tree ensembles. We argue that it shares the computational benefits of the DART prior while also being more intuitive. Our method works by partitioning weak learners (in this case, step functions) in such a way that different groups of learners focus on a limited set of variables.

We have not attempted a rigorous analysis of the theoretical properties of BART models using Gibbs priors along the lines of Linero and Yang (2018), Rockova and van der Pas (2020), Liu, Ročková, and Wang (2021). While we have not done so here, it is easy to adapt the proof of Theorem 3 of Linero and Yang (2018) to allow for the use of Gibbs priors; this result establishes near-minimax predictive accuracy of the SBART model adaptively over smoothness and sparsity levels. Less immediate are the variable selection consistency results of Liu, Ročková, and Wang (2021); we conjecture that there are no obstructions to obtaining similar results, at least with α fixed as $(N, P) \to \infty$, provided that an appropriate complexity-penalizing prior $\pi_D(d)$ is used. We leave establishing variable selection consistency to future work.

It would be interesting to develop Gibbs prior extensions of the Dirichlet process forest prior (Du and Linero 2019b), which is useful for performing interaction detection, or the overlapping group Dirichlet prior (Du and Linero 2019a), which is used to perform bi-level selection using BART. Both of these approaches make use of *hierarchical* structure; for example, the Dirichlet process forest constructions mirrors the structure of the hierarchical Dirichlet process (Teh et al. 2006). Extending Gibbs priors to accommodate hierarchical structure is nontrivial, but may be possible using the hierarchical species sampling framework of Bassetti, Casarin, and Rossini (2020). We defer this to future work.

We have also found some evidence that the performance of BART may be due to the general power of ensembling weak learners, rather than to the choice of decision trees as weak learners. Decision trees have some computational benefits, but MARS weak learners are also computationally convenient due to their nonhierarchical structure and their sparsity. Precisely characterizing the theoretical properties of generic Bayesian weak learner ensembles is another potentially fruitful direction for research.

Appendix A. Proof of Proposition 2

Let $\mathcal{C}$ denote the partition on $\{1, \ldots, B\}$ induced by taking $n \sim n'$ if $j_n = j_{n'}$ where $\mathcal{J} = \{j_1, \ldots, j_B\}$ is a list of all the splitting variables in the ensemble. Let $\mathcal{C}_b$ denote the partition associated with $\mathcal{J}_b = \{j_n : n \neq b\}$ instead. Note that the conditional distribution of $\mathcal{J}_b$ on either $\mathcal{C}_b$ or $\mathcal{C}$ uniformly assigns variables to each equivalence class, that is, $\pi(\mathcal{J}_b \mid \mathcal{C}_b) = \frac{(P-Q_b)!}{P!}$. If $m_j^{(-b)} \neq 0$ then $\mathcal{C}$ is obtained by adding b to the equivalence class associated with j, so that the desired probability is

$$\pi(\mathcal{C} \mid \mathcal{J}_b) = \frac{\pi(\mathcal{J}_b, \mathcal{C})}{\pi(\mathcal{J}_b)} = \frac{\pi(\mathcal{J}_b \mid \mathcal{C})\,\pi(\mathcal{C})}{\pi(\mathcal{J}_b \mid \mathcal{C}_b)\,\pi(\mathcal{C}_b)} = \frac{\pi(\mathcal{C})}{\pi(\mathcal{C}_b)}$$

$$= \frac{V_B(Q_b)}{V_{B-1}(Q_b)}(\alpha + m_j^{(-b)})$$

by (3) and canceling the common terms (the second equality follows from the fact that $\pi(\mathcal{J}_b) = \pi(\mathcal{J}_b, \mathcal{C}_b)$).

If $m_j^{(-b)} = 0$ then setting $j = j_b$ corresponds to adding a new equivalence class. The desired probability is given by

$$\pi(\mathcal{J} \mid \mathcal{J}_b) = \pi(\mathcal{J} \mid \mathcal{C}, \mathcal{J}_b)\,\pi(\mathcal{C} \mid \mathcal{J}_b) = \frac{\pi(\mathcal{C} \mid \mathcal{J}_b)}{P - Q_b}$$

where the second equality follows from the fact that $j_b = j$ is chosen uniformly from the $(P - Q_b)$ variables which have not been split on. Next, we have

$$\pi(\mathcal{C} \mid \mathcal{J}_b) = \frac{\pi(\mathcal{J}_b \mid \mathcal{C})\,\pi(\mathcal{C})}{\pi(\mathcal{J}_b \mid \mathcal{C}_b)\,\pi(\mathcal{C}_b)} = \frac{\pi(\mathcal{C})}{\pi(\mathcal{C}_b)} = \frac{V_B(Q_b+1)}{V_{B-1}(Q_b)}\alpha$$

by (3) and canceling common terms.

The result for the spike-and-forest prior follows from the above computations and Proposition 1 by multiplying the numerator and denominator by α^{B-1} and taking $\alpha \to \infty$ after observing that $\frac{d!}{(d-t)!}\frac{\alpha^B\,\Gamma(\alpha\, d)}{\Gamma(\alpha\, d+B)} \to \frac{d!}{(d-t)!\,d^B}$.

Supplementary Materials

The Supplementary Material contains additional algorithmic details for the Bayesian MARS model, proofs of propositions, and MCMC diagnostics.

Acknowledgments

The authors are grateful to associate editor and two anonymous referees for their helpful feedback.

Disclosure Statement

The authors report that there are no competing interests to declare.

Funding

This material is based upon work supported by the National Science Foundation under grant DMS-2144933.

References

Awaya, N., and Ma, L. (2021), "Tree Boosting for Learning Probability Measures," arXiv preprint arXiv:2101.11083. [1046]

Barbieri, M. M., and Berger, J. O. (2004), "Optimal Predictive Model Selection," *The Annals of Statistics*, 32, 870–897. [1049,1053]

Basak, P., Linero, A., Sinha, D., and Lipsitz, S. (2021), "Semiparametric Analysis of Clustered Interval-Censored Survival Data Using Soft Bayesian Additive Regression Trees (SBART)," *Biometrics*, 78, 880–893. [1046]

Bassetti, F., Casarin, R., and Rossini, L. (2020), "Hierarchical Species Sampling Models," *Bayesian Analysis*, 15, 809–838. [1058]

Bühlmann, P. (2006), "Boosting for High-Dimensional Linear Models," *The Annals of Statistics*, 34, 559–583. [1046]

Blackwell, D., and MacQueen, J. B. (1973), "Ferguson Distributions via Pólya Urn Schemes," *The Annals of Statistics*, 1, 353–355. [1051]

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, 1750–1781. [1048,1054]

Chipman, H. A., George, E. I., and McCulloch, R. E. (2010), "BART: Bayesian Additive Regression Trees," *The Annals of Applied Statistics*, 4, 266–298. [1046,1047,1048]

De Blasi, P., Favaro, S., Lijoi, A., Mena, R. H., Prünster, I., and Ruggiero, M. (2013), "Are Gibbs-Type Priors the Most Natural Generalization of the Dirichlet Process?" *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37, 212–229. [1046,1047,1058]

Denison, D., Mallick, B., and Smith, A. (1998), "Bayesian MARS," *Statistics and Computing*, 8, 337–346. [1047,1056]

Dorie, V., Hill, J., Shalit, U., Scott, M., and Cervone, D. (2019), "Automated versus Do-it-yourself Methods for Causal Inference: Lessons Learned from a Data Analysis Competition," *Statistical Science*, 34, 43–68. [1046]

Du, J., and Linero, A. R. (2019a), "Incorporating Grouping Information into Bayesian Decision Tree Ensembles," in *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of Proceedings of Machine Learning Research, pp. 1686–1695. PMLR. [1058]

——— (2019b), "Interaction Detection with Bayesian Decision Tree Ensembles," in *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, volume 89 of *Proceedings of Machine Learning Research*, pp. 108–117. PMLR. [1058]

Freund, Y., Schapire, R., and Abe, N. (1999), "A Short Introduction to Boosting," *Journal of Japanese Society For Artificial Intelligence*, 4, 771–780. [1046]

Friedman, J. H. (1991), "Multivariate Adaptive Regression Splines," *The Annals of Statistics*, 19, 1–67. [1047,1052,1056]

——— (2001), "Greedy Function Approximation: A Gradient Boosting Machine," *The Annals of Statistics*, 29, 1189–1232. [1046]

George, E. I., and McCulloch, R. E. (1993), "Variable Selection via Gibbs Sampling," *Journal of the American Statistical Association*, 88, 881–889. [1047]

Gnedin, A., and Pitman, J. (2006), "Exchangeable Gibbs Partitions and Stirling Triangles," *Journal of Mathematical Sciences*, 138, 5674–5685. [1046,1050]

Hahn, P. R., Murray, J. S., and Carvalho, C. M. (2020), "Bayesian Regression Tree Models for Causal Inference: Regularization, Confounding, and Heterogeneous Effects," (with Discussion), *Bayesian Analysis*, 15, 965–1056. [1046]

Hastie, T., and Tibshirani, R. (2000), "Bayesian Backfitting," *Statistical Science*, 15, 196–223. [1046]

Hill, J., Linero, A. R., and Murray, J. (2019), "Bayesian Additive Regression Trees: A Review and Look Forward," *Annual Review of Statistics and Its Application*, 7, 251–278. [1052]

Hill, J. L. (2011), "Bayesian Nonparametric Modeling for Causal Inference," *Journal of Computational and Graphical Statistics*, 20, 217–240. [1046]

Irsoy, O., Yildiz, O. T., and Alpaydin, E. (2012), "Soft Decision Trees," in *Proceedings of the 21st International Conference on Pattern Recognition (ICPR2012)*, pp. 1819–1822. [1048]

James, G., Witten, D., Hastie, T., and Tibshirani, R. (2021), *ISLR: Data for an Introduction to Statistical Learning with Applications in R*. R package version 1.4. [1055]

Kapelner, A., and Bleich, J. (2016), "bartMachine: Machine Learning with Bayesian Additive Regression Trees," *Journal of Statistical Software*, 70, 1–40. [1052]

Li, Y., Linero, A. R., and Murray, J. S. (2022), "Adaptive Conditional Distribution Estimation with Bayesian Decision Tree Ensembles," *Journal of the American Statistical Association*. DOI:10.1080/01621459.2022.2037431 [1046]

Linero, A. R. (2017), "A Review of Tree-based Bayesian Methods," *Communications for Statistical Applications and Methods*, 24, 543–559. [1049]

——— (2018), "Bayesian Regression Trees for High-Dimensional Prediction and Variable Selection," *Journal of the American Statistical Association*, 113, 626–636. [1046,1047,1048,1050,1054,1057,1058]

Linero, A. R., Basak, P., Li, Y., and Sinha, D. (2021), "Bayesian Survival Tree Ensembles with Submodel Shrinkage," *Bayesian Analysis*, 17, 997–1020. [1046]

Linero, A. R., and Du, J. (2021), "Variable Selection for Bayesian Decision Tree Ensembles," in *Handbook of Bayesian Variable Selection*, eds. M. G. Tadesse and M. Vannuci, pp. 415–440, Boca Raton, FL: Chapman and Hall/CRC. [1046]

Linero, A. R., Sinha, D., and Lipsitz, S. R. (2020), "Semiparametric Mixed-Scale Models Using Shared Bayesian forests," *Biometrics*, 76, 131–144. [1052]

Linero, A. R., and Yang, Y. (2018), "Bayesian Regression Tree Ensembles that Adapt to Smoothness and Sparsity," *Journal of the Royal Statistical Society*, Series B, 80, 1087–1110. [1048,1051,1057,1058]

Liu, Y., Ročková, V., and Wang, Y. (2021), "Variable Selection with ABC Bayesian Forests," *Journal of the Royal Statistical Society*, Series B, 83, 453–481. [1047,1049,1058]

Miller, J. W., and Harrison, M. T. (2013), "A Simple Example of Dirichlet Process Mixture Inconsistency for the Number of Components," in *Advances in Neural Information Processing Systems*, pp. 199–206. [1049]

——— (2018), "Mixture Models with a Prior on the Number of Components," *Journal of the American Statistical Association*, 113, 340–356. [1046,1050,1051,1058]

Murray, J. S. (2021), "Log-Linear Bayesian Additive Regression Trees for Multinomial Logistic and Count Regression Models," *Journal of the American Statistical Association*, 116, 756–769. [1046,1052]

Pitman, J. (2002), "Combinatorial Stochastic Processes," Technical Report 621, Department of Statistics, University of California, Berkeley. [1046,1050]

Pratola, M., Chipman, H., George, E., and McCulloch, R. (2020), "Heteroscedastic BART via Multiplicative Regression Trees," *Journal of Computational and Graphical Statistics*, 29, 405–417. [1046]

Rockova, V., and van der Pas, S. (2020), "Posterior Concentration for Bayesian Regression Trees and Forests," *The Annals of Statistics*, 48, 2108–2131. [1046,1047,1048,1051,1058]

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*, 35, 2741–2753. [1046]

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. [1055,1056]

Teh, Y. W., Jordan, M. I., Beal, M. J., and Blei, D. M. (2006), "Hierarchical Dirichlet Processes," *Journal of the American Statistical Association*, 101, 1566–1581. [1058]