ROYAL STATISTICAL SOCIETY
DATA | EVIDENCE | DECISIONS

Journal of the Royal Statistical Society
Statistical Methodology
Series B

J. R. Statist. Soc. B (2018)

Bayesian regression tree ensembles that adapt to smoothness and sparsity

Antonio R. Linero

Florida State University, Tallahassee, USA

and Yun Yang

University of Illinois at Urbana–Champaign, Champaign, USA

[Received June 2017. Final revision August 2018]

Summary. Ensembles of decision trees are a useful tool for obtaining flexible estimates of regression functions. Examples of these methods include gradient-boosted decision trees, random forests and Bayesian classification and regression trees. Two potential shortcomings of tree ensembles are their lack of smoothness and their vulnerability to the curse of dimensionality. We show that these issues can be overcome by instead considering sparsity inducing soft decision trees in which the decisions are treated as probabilistic. We implement this in the context of the Bayesian additive regression trees framework and illustrate its promising performance through testing on benchmark data sets. We provide strong theoretical support for our methodology by showing that the posterior distribution concentrates at the minimax rate (up to a logarithmic factor) for sparse functions and functions with additive structures in the high dimensional regime where the dimensionality of the covariate space is allowed to grow nearly exponentially in the sample size. Our method also adapts to the unknown smoothness and sparsity levels, and can be implemented by making minimal modifications to existing Bayesian additive regression tree algorithms.

Keywords: Bayesian additive regression trees; Bayesian non-parametrics; High dimensional regimes; Model averaging; Posterior consistency

1. Introduction

Consider a non-parametric regression model $Y = f_0(X) + \epsilon$ with response Y, $X \in [0,1]^p$ a p-dimensional predictor, f_0 an unknown regressionfunction of interest, and Gaussian noise $\epsilon \sim N(0, \sigma^2)$. Suppose that we observe $\mathcal{D} = ((X_1, Y_1), \ldots, (X_n, Y_n))$ consisting of independent and identically distributed copies of (X, Y). A popular approach to estimating $f_0(x)$ is to form an ensemble of decision trees; common techniques include boosted decision trees (Freund and Schapire, 1999) and random forests (Breiman, 2001). Bayesian tree-based models, such as the Bayesian additive regression tree (BART) model (Chipman *et al.*, 2010), have recently attracted interest from practitioners due to their excellent empirical performance and natural uncertainty quantification; the BART model has been applied in a wide variety of contexts such as non-parametric function estimation with variable selection (Bleich *et al.*, 2014; Linero, 2018), analysis of log-linear models (Murray, 2017) and survival analysis (Sparapani *et al.*, 2016). Additionally, the BART model is consistently among the best performing methodologies in the Atlantic

Address for correspondence: Antonio R. Linero, Department of Statistics, Florida State University, 217 Rogers Building, 117 North Woodward Avenue, Tallahassee, FL 32306, USA.
E-mail: arlinero@stat.fsu.edu

 1369–7412/18/80000

causal inference conference data analysis challenge (Hill, 2011, 2016; Hahn *et al.*, 2017; Dorie *et al.*, 2017).

Despite the recent popularity of Bayesian tree-based models, they suffer from several drawbacks. First, in the regression setting, estimators based on decision trees are not capable of adapting to higher levels of smoothness exhibited in f_0 because of their piecewise constant nature. Second, as illustrated by Linero (2018), they suffer from the curse of dimensionality—their prediction performance deteriorates as the dimensionality p increases. Last, but not least, very little theoretical work has been done for understanding large sample properties of Bayesian tree-based approaches from a frequentist perspective.

In this paper we propose a new method, called the soft Bayesian additive regression tree (SBART) method, which improves both practically and theoretically on existing Bayesian sum-of-trees models. To address the first aforementioned drawback, we employ an ensemble of carefully designed 'soft' decision trees as building blocks in the BART model and show in both empirical studies and theoretical investigation that the resulting Bayesian approach can adapt to the unknown level of smoothness of the true regression function f_0—the corresponding posterior distribution achieves the minimax rate $n^{-\alpha/(2\alpha+p)}$ of contraction up to logarithmic terms (Ghosal *et al.*, 2000) when $f_0 \in C^{\alpha,R}([0,1]^d)$ where $C^{\alpha,R}([0,1]^d)$ denotes a Hölder space with smoothness index α and radius R.

To overcome the curse of dimensionality, we specify sparsity inducing priors (Linero, 2018) for the splitting rule probabilities in the soft decision trees. We show that the SBART method takes advantage of structural sparsity in the true regression function f_0—when f_0 depends on only $d \ll p$ predictors and is α Hölder smooth, the resulting posterior distribution contracts towards the truth at a rate of $n^{-\alpha/(2\alpha+d)} + \sqrt{\{n^{-1}d\log(p)\}}$ up to logarithmic terms, which is nearly minimax optimal even in the high dimensional setting where p grows nearly exponentially fast in n (Yang and Tokdar, 2015). Furthermore, because of the additive nature of sum-of-trees based models, we show that the SBART method can also adapt to low order non-linear interactions: if f_0 can be decomposed into many low dimensional pieces $f_0 = \Sigma_{v=1}^{V} f_{0v}$, where each additive component f_{0v} is d_v sparse and α_v smooth, then the SBART method also achieves a nearly minimax rate of posterior contraction. Compared with the rate for the general sparse case, which allows at most $o\{\log(n)\}$ many active predictors for consistency, the rate for additive structures potentially allows $o(n^{\beta})$ many predictors for some $\beta \in (0,1)$; this partly explains the empirical success of Bayesian sum-of-tree approaches, as many real world phenomena can be explained in terms of a small number of low order interactions.

Our proofs involve a key lemma that links the sum-of-tree type of estimators with kernel-type estimators. Unlike frequentist kernel-type estimators that require prior knowledge on the level of smoothness of f_0 for choosing a smoothness matching kernel, Bayesian sum-of-tree based methods are adaptive, requiring no prior knowledge of the smoothness levels $\{\alpha_v\}$, number of additive components V or degree of lower order interactions d_v, while still attaining nearly minimax rates even under the high dimensional setting. Practically, the SBART method can be implemented by making minimal modifications to existing strategies for fitting Bayesian tree-based models: the sparsity inducing prior uses conditionally conjugate Dirichlet priors which can be easily accommodated during Gibbs sampling, whereas replacing the usual decision trees with soft decision trees requires minor changes to the backfitting algorithm that is typically used with the BART method.

1.1. Related work

There has been a recent surge of interest in the theoretical properties of BART-type models. While our work was under review we learned that, in essentially simultaneous work, Rockova

and van der Pas (2017) had established similar posterior contraction rates for a particularly designed BART prior, using a so-called 'spike-and-tree' prior to enable the ensemble to adapt to sparsity. In particular, they showed that a single deep decision tree can approximate any function with smoothness level $\alpha \leqslant 1$, which is then divided among trees with smaller depth. Our theory instead relies on linking the sum-of-tree type of estimators with kernel-type estimators, which only need shallow trees and motivate the usage of soft decision trees. Practically, the most relevant difference is that our SBART prior enables adaptation to the level of smoothness even when $\alpha > 1$, whereas the use of piecewise constant basis functions in traditional BART models enables only adaptation to functions which are at most Lipschitz smooth ($\alpha \leqslant 1$). An additional difference is that we focus on establishing concentration results for the fractional posterior, which enables less restrictive assumptions about our choice of prior; in our on-line supplementary material, we also provide concentration results for the usual posterior, under more stringent conditions. In even more recent work, Alaa and van der Schaar (2018) have established consistency results for BART-type priors for estimating individual treatment effects in causal inference settings and also noted the limitation of the BART model in adapting to a smoothness order that is higher than $\alpha = 1$.

The soft decision trees that we use are similar in spirit to those used by Irsoy *et al*. (2012), who considered a soft variant of the classification and regression trees algorithm. Our work differs in that our trees are not learned in a greedy fashion, but instead by extending the Bayesian backfitting approach of Chipman *et al*. (2010), we consider an ensemble of soft trees rather than a single tree, we use a different parameterization of the gating function which does not consider oblique decision boundaries and we establish theoretical guarantees for our approach.

The rest of the paper is organized as follows. In Section 2, we develop our SBART prior. In Section 3 we state our theoretical results. In Section 4, we illustrate the methodology on both simulated and real data sets. We finish in Section 5 with a discussion. Proofs are deferred to the appendices. In on-line supplementary material, we provide additional computational details, timing results and additional theoretical results extending our fractional posterior results to the usual posterior.

2. Soft Bayesian sum-of-trees models

2.1. Description of the model

We begin by describing the usual 'hard' decision tree prior that is used in the BART model. We model $f_0(x)$ as the realization of a random function

$$f(x) = \sum_{t=1}^{T} g(x; \mathcal{T}_t, \mathcal{M}_t), \qquad x \in \mathbb{R}^p, \tag{1}$$

where $\mathcal{T}_t$ denotes the topology or splitting rules of the tree, $\mathcal{M}_t = (\mu_{t1}, \ldots, \mu_{tL_t})$ is a collection of parameters for the leaf nodes and L_t denotes the number of leaves. The function $g(x; \mathcal{T}_t, \mathcal{M}_t)$ returns $\Sigma_{l=1}^{L} \mu_{tl} \phi(x; \mathcal{T}_t, l)$ where $\phi(x; \mathcal{T}_t, l)$ is the indicator that x is associated with leaf node l in $\mathcal{T}_t$.

Following Chipman *et al*. (2010), we endow $\mathcal{T}_t$ with a branching process prior. The branching process begins with a root node of depth $k = 0$. For $k = 0, 1, 2, \ldots$, each node at depth k is non-terminal with probability $q(k) = \gamma(1 + k)^{-\beta}$ where $\gamma > 0$ and $\beta > 0$ are hyperparameters controlling the shape of the trees. It is easy to check by using elementary branching process theory that this process terminates almost surely provided that $\beta > 0$ (Athreya and Ney, 2004).

Given the tree topology, each branch node b is given a decision rule of the form $[x_j \leqslant C_b]$, with x going left down the tree if the condition is satisfied and right down the tree otherwise. The predictor j is selected with probability s_j where $s = (s_1, \ldots, s_p)$ is a probability vector. We assume that $C_b \sim \text{uniform}(a, b)$ where a and b are chosen so that the cell of $\mathbb{R}^p$ that is defined by the path to b is split along the jth co-ordinate. The leaf parameters μ_{tl} are assumed independent and identically distributed from an $N(0, \sigma_\mu^2/T)$ distribution. The scaling factor T ensures the stability of the prior on f as the number of trees increases—loosely speaking, the functional central limit theorem implies the convergence of the prior on f to a Gaussian process as $T \to \infty$.

We now describe how to convert the hard decision tree that was described above into a soft decision tree. Rather than x following a deterministic path down the tree, x instead follows a probabilistic path, with x going left at branch b with probability

$$\psi(x; \mathcal{T}, b) = \psi\left(\frac{x_j - C_b}{\tau_b}\right),$$

where $\tau_b > 0$ is a bandwidth parameter associated with branch b. Averaging over all possible paths, the probability of going to leaf l is

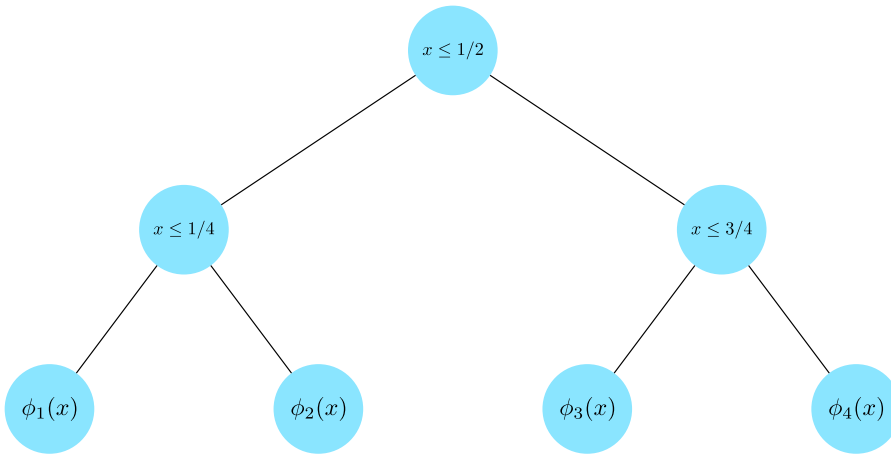
$$\phi(x; \mathcal{T}, l) = \prod_{b \in A(l)} \psi(x; \mathcal{T}, b)^{1-R_b} \{1 - \psi(x; \mathcal{T}, b)\}^{R_b}, \tag{2}$$

where $A(l)$ is the set of ancestor nodes of leaf l and $R_b = 1$ if the path to l goes right at b. The parameter τ_b controls the sharpness of the decision, with the model approaching a hard decision tree as $\tau_b \to 0$, and approaching a constant model as $\tau_b \to \infty$. Unlike hard decision trees where each leaf is constrained to influence the regression function f only locally near its centre $\{C_b\}$, each leaf in the soft decision tree imposes a global effect on f, whose influence as x deviates from the centre depends on the local bandwidths $\{\tau_b\}$. As we shall illustrate, this global effect of local leaves enables the soft tree model to borrow information adaptively across different covariate regions, where the degree of smoothing is determined by the local bandwidth parameters that are learned from the data. This is illustrated in Fig. 1 for a simple univariate soft decision tree. In our illustrations we use the logistic gating function $\psi(x) = \{1 + \exp(-x)\}^{-1}$.

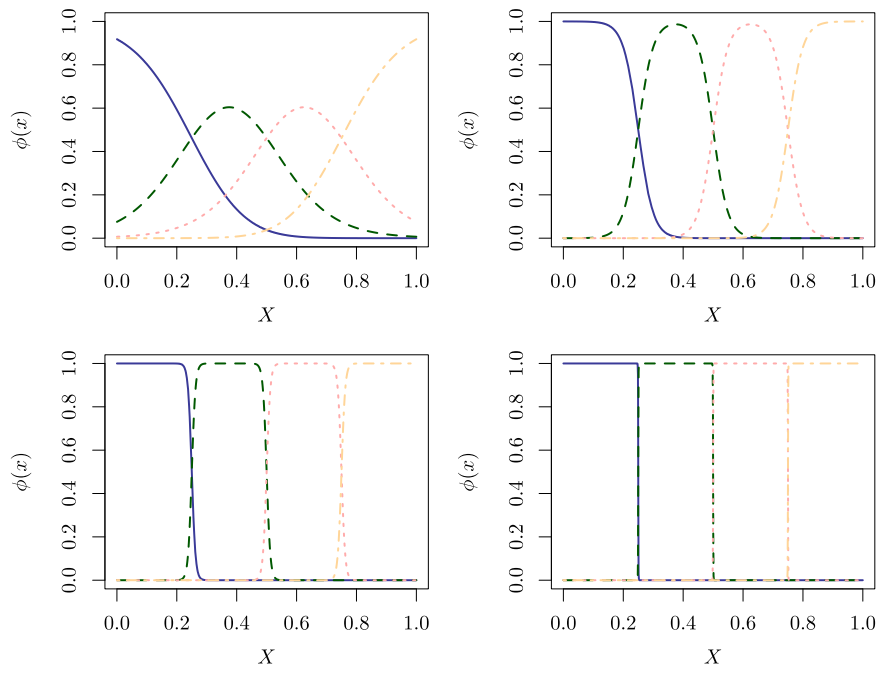
2.2. Smoothness adaptation

A well-known feature of decision trees is their lack of smoothness. Single-tree algorithms such as the classification and regression trees algorithm (Hastie *et al.* (2009), chapter 9.2) result in step function estimates, suggesting that they should not be capable of efficiently estimating smooth functions (Györfi *et al.*, 2006). Methods based on ensembles of decision trees average over many distinct partitions of the predictor space, resulting in some degree of smoothing. Even with this averaging, the estimated regression functions are not smooth. Heuristically, we note that under our BART specification the function f is not differentiable in quadratic mean. Indeed, with trees of depth 1, $p = 1$, and cut points $C_b \sim G$, simple calculations give $E[\{f(x+\delta) - f(x)\}^2] \propto \delta G'(x) + o(\delta)$. Consequently, BART ensembles with a large number of trees resemble nowhere-differentiable continuous functions, and in the limit as $T \to \infty$ the BART prior converges to a nowhere-differentiable Gaussian process. This heuristic argument suggests that the BART model can adapt only to functions with Hölder smoothness level no greater than $\alpha = 1$ (Lipschitz functions).

Fig. 2 compares the fit of the BART to the SBART model with $\tau_b \equiv 0.1$. We see that when $T = 1$ trees are used we require a large number of leaf nodes to model relatively simple functions. At a large scale, we see that the BART fit resembles a nowhere-differentiable continuous function.



(a)



(b)

Fig. 1. (a) Example tree, with cut points at $x = 0.5$, 0.25, 0.75, and (b) weights $\phi_l(x)$ for $l = 1, \ldots, 4$ as functions of x for the values $\tau^{-1} \in \{10, 40, 160, 2560\}$

Although it is an improvement, the estimate from the BART model is still not sufficiently smooth and exhibits large fluctuations.

The fit of the soft decision tree in Fig. 2 by comparison is infinitely differentiable and requires only a small number of parameters. Consequently, we obtain a fit with lower variance and negligible bias. An attractive feature of soft decision trees that is exhibited in Fig. 2 is their

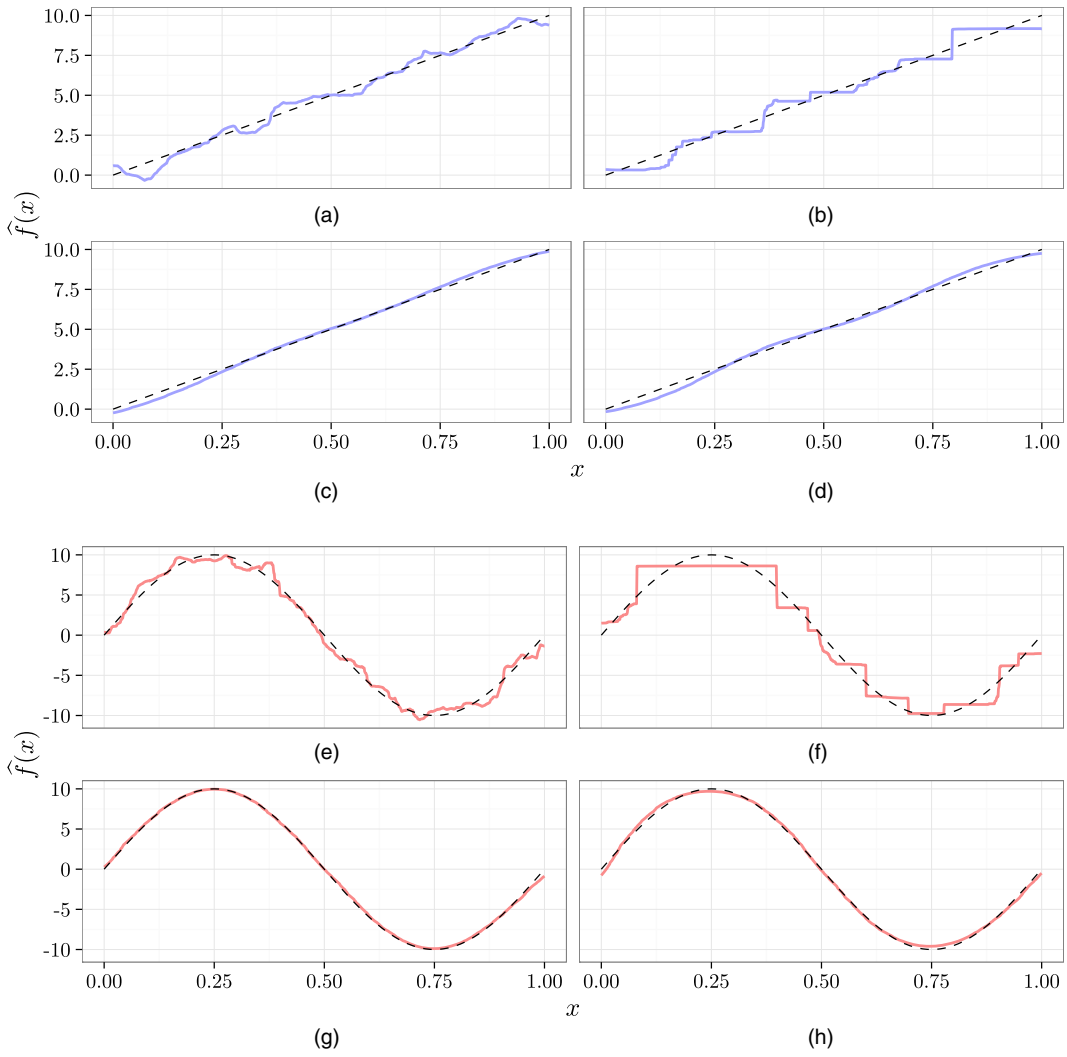


Fig. 2. Posterior means (———) against underlying true regression function (— — —) (the error variance is $\sigma^2 = 2^2$; (a)–(d), $f(x) = 10x_1$; (e)–(h), $f(x) = 10\sin(2\pi x_1)$; BART denotes the BART model with $T = 50$, DT denotes the BART model with $T = 1$, and soft variants are prefixed by Soft): (a), (e) BART, (b), (f) DT; (c), (g) SoftBART; (d), (h) SoftDT

ability to approximate linear relationships. In this case, even when $T = 1$, we recover the smooth functions almost exactly.

2.3. Prior specification and implementation

Following Chipman *et al*. (2010), in this section we develop a 'default' SBART prior. The goal is to develop a prior which can be used routinely, without requiring the user to specify any hyperparameters; although the choices below may appear *ad hoc*, they have been found to work remarkably well across a wide range of data sets. After adopting the following default prior, users may wish to tune the number of trees T and the parameter r in the prior for τ_b further, or

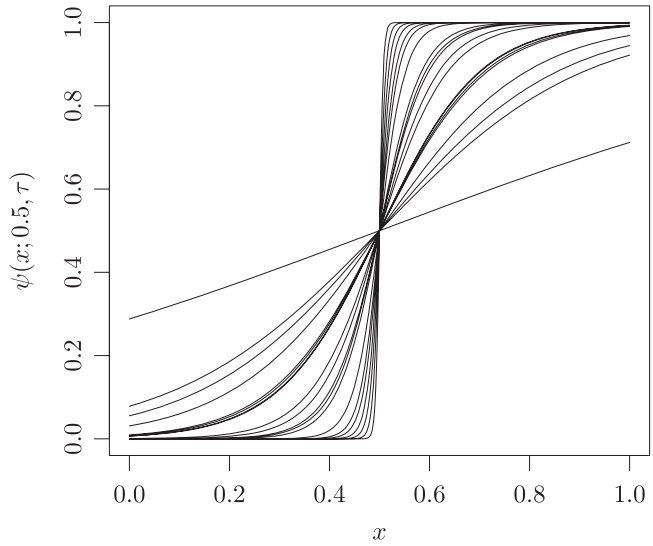


Fig. 3. Draws of the gating function $\psi(x; \mathcal{T}, b)$ when $\tau_b \sim \text{Exp}(0.1)$ and $C_b = 0.5$

to use additional information regarding the targeted level of sparsity. We stress, however, that a reasonable baseline level of performance is obtained without the need to do any further tuning.

Following Chipman *et al*. (2010), we recommend scaling Y so that most or all of the responses fall in the interval $[-0.5, 0.5]$. We also preprocess X_j so that $X_j \sim \text{uniform}(0, 1)$ approximately by applying a quantile normalization in which each X_{ij} is mapped to its rank, with $\min X_{ij} = 1$ and $\max X_{ij} = n$. We then apply a linear transformation so that the values of X_{ij} are in $[0, 1]$. The goal of this preprocessing of X is to make the prior invariant under monotone transformations of X, which is a highly desirable property of the original default BART model.

We now describe our default prior for the bandwidths τ_b and the splitting proportions $s = (s_1, \ldots, s_p)$. We use a sparsity inducing Dirichlet prior:

$$s \sim \mathcal{D}(a/p^{\xi}, \ldots, a/p^{\xi}), \qquad \xi \geqslant 1. \tag{3}$$

Our theoretical results require $\xi > 1$; however, in practice we find that setting $\xi = 1$ works adequately. This Dirichlet prior for s was introduced by Linero (2018); throughout, we refer to the BART model with prior (3) as the Dirichlet additive regression tree (DART) model to contrast it with the BART model when no such sparsity inducing prior is used. The parameter a controls the expected amount of sparsity in f. Conditionally on there being B branches in the ensemble, the number of predictors has a $1 + \text{Poisson}(\theta)$ distribution (i.e. it is equal in distribution to $1 + Z$, where $Z \sim \text{Poisson}(\theta)$) with $\theta = a\Sigma_{i=1}^{B-1}(a + i)^{-1}$ (Linero, 2018) when $\xi = 1$. When prior information is available on the sparsity of f_0, we can choose a to match the targeted amount of sparsity. By default we use a compound gamma prior; $a/(a + \lambda_a) \sim \text{Be}(a_a, b_a)$, with $a_a = 0.5, b_a = 1$ and $\lambda_a = p$. This prior attempts to strike a balance between the sparse and non-sparse settings by having an infinite density at 0, median $\alpha = p/4$ and an infinite mean.

There are several possibilities for choosing the bandwidth τ_b. In preliminary work, using tree-specific τ_ts shared across branches in a fixed tree worked well, with $\tau_t \sim \text{Exp}(r)$ where $E(\tau_t) = r$. Our illustrations use $r = 0.1$, which, as shown in Fig. 3, gives a wide range of possible gating functions. An interesting feature of the sampled gating functions is that both approximate step functions and approximately linear functions are supported.

We give $\sigma = \text{var}(\epsilon)^{1/2}$ a half-Cauchy prior, $\sigma \sim \text{Cauchy}_+(0, \hat{\sigma})$. Again following Chipman *et al.* (2010), $\hat{\sigma}$ is an estimate of σ based on the data. We use an estimate $\hat{\sigma}_{\text{lasso}}$ of σ obtained by fitting the lasso using the `glmnet` package in R.

The model has hyperparameters $(\sigma_\mu^2, \gamma, \beta, T)$. In preliminary work, we did not have success in placing priors on γ and β, and instead we fixed $\gamma = 0.95$ and $\beta = 2$ (Chipman *et al.*, 2010). We give σ_μ a half-Cauchy prior, $\sigma_\mu \sim \text{Cauchy}_+(0, 0.25)$, where 0.25 is chosen so that σ_μ has median equal to the default value that was recommended by Chipman *et al.* (2010).

An important remaining specification is the number of trees T to include in the ensemble. The theoretical results that we establish in Section 3 make use of a prior distribution on T; however, our attempts to incorporate a prior on T by using reversible jump methods (Green, 1995) resulted in poor mixing of the associated Markov chain Monte Carlo algorithms. Generally, we have found that fixing T at a default value of $T = 50$ or $T = 200$ is sufficient to attain good performance on most data sets. Tuning T further often provides a modest increase in performance but may be worth the effort on some data sets (see Section 4.3).

There are various possible options for tuning T, such as approximate leave-one-out (LOO) cross-validation using Pareto-smoothed importance sampling (Vehtari *et al.*, 2017), maximizing an approximate marginal likelihood obtained by using (say) the widely applicable Bayesian information criterion (WBIC) (Watanabe, 2013), or K-fold cross-validation as recommended by Chipman *et al.* (2010). The advantage of the WBIC and Pareto-smoothed importance sampling leave-one-out (PSISLOO) cross-validation is that they require fitting the model only once for each value of T. In practice, we have found that approximations such as the WBIC and PSISLOO cross-validation are unreliable, with PSISLOO cross-validation prone to overfitting and the WBIC requiring potentially very long chains to estimate. Fig. 4 displays the values of PSISLOO cross-validation, a WBIC approximation of the negative marginal likelihood of T (Watanabe, 2013) and fivefold cross-validation, when used to select T for a replicate of the illustration in Section 4.1 with $p = 100$ predictors. Both the WBIC and cross-validation select $T = 10$, which also minimizes the root-mean-squared error $\int \{f_0(x) - \hat{f}(x)\}^2 \mathrm{d}x$. Resource permitting, we have found that K-fold cross-validation is the most reliable method for selecting T.

As a default we use the following priors throughout the paper.

$$
\begin{gathered}
s \sim \mathcal{D}(a/p, \ldots, a/p), \\
a/(a+p) \sim \text{Be}(0.5, 1), \\
\tau_t \overset{\text{indep}}{\sim} \text{Exp}(0.1), \\
\sigma_\mu \sim \text{Cauchy}_+(0, 0.25), \\
\sigma \sim \text{Cauchy}_+(0, \hat{\sigma}_{\text{lasso}}), \\
\gamma = 0.95, \\
\beta = 2.
\end{gathered}
$$

2.4. Variable grouping prior

The sparsity inducing prior (3) can be extended to allow penalization of groups of predictors simultaneously, in a manner similar to the group lasso (Yuan and Lin, 2006). Suppose that the predictors can be divided into M groups of size P_m. We set

$$
\begin{gathered}
s_{mk} = u_m v_{mk}, \\
u \sim \mathcal{D}(a/M, \ldots, a/M), \\
v_m \sim \mathcal{D}(\omega/P_m, \ldots, \omega/P_m).
\end{gathered}
$$

We primarily use the grouping prior to allow for the inclusion of categorical predictors through the inclusion of dummy variables. This is an extension of the approach that is used by the

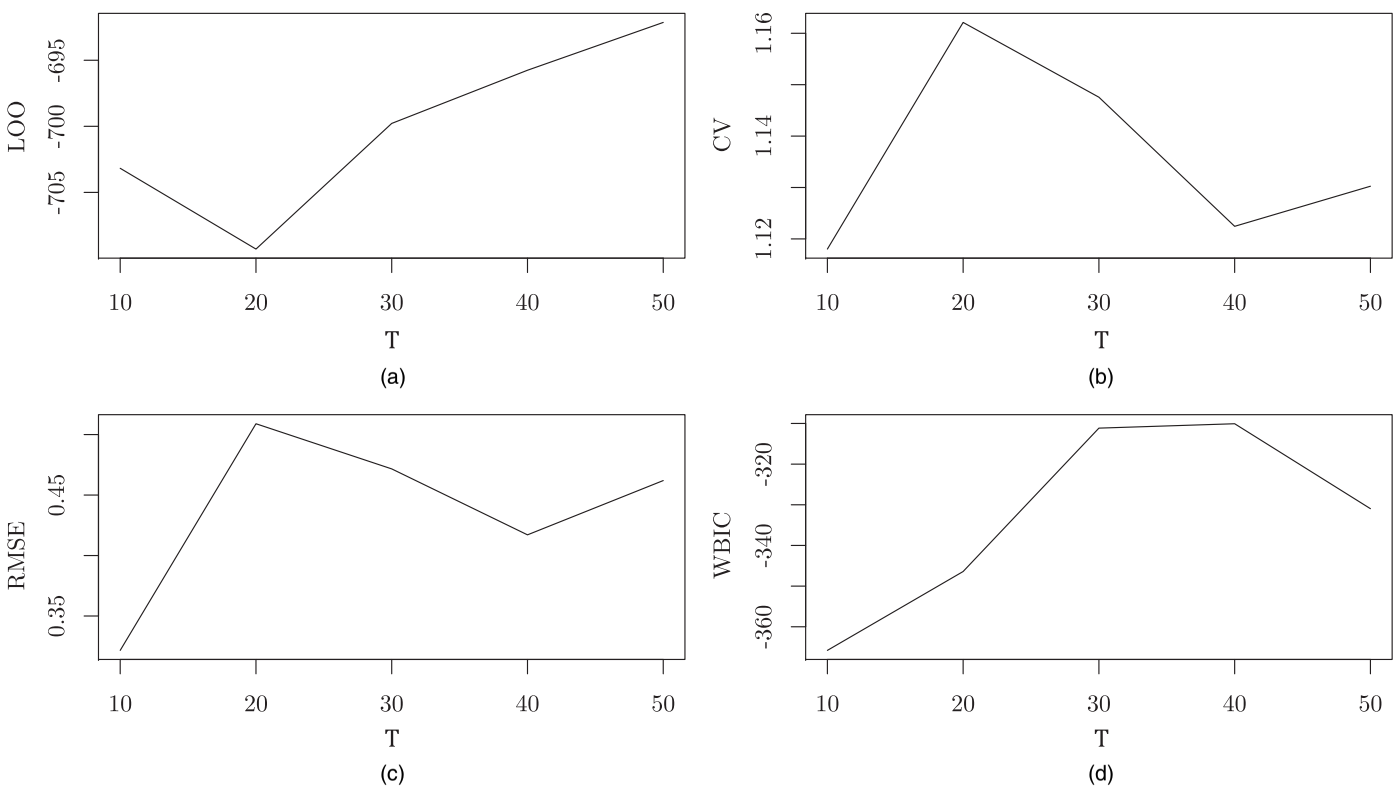


Fig. 4. Selecting T by using (a) LOO cross-validation, (b) cross-validation and (d) the WBIC, with (c) the population root-mean-squared error for f

Table 1. Algorithm 1: Bayesian backfitting algorithm

1, *for* $t = 1, \ldots, T$ *do*
2, set $Y_i^* \leftarrow Y_i - \Sigma_{k \neq t} g(X; \mathcal{T}_k, \mathcal{M}_k)$ for $i = 1, \ldots, N$
3, sample $\mathcal{T}_t \sim \text{Metrop}_{\mathcal{T}}(Y^*, X, \tau_t, h)$
4, sample $\tau_t \sim \text{Metrop}_{\tau}(Y^*, X, \mathcal{T}_t, h)$
5, sample $\mathcal{M}_t \sim N(\hat{\mu}_t, \Omega_t)$ with $(\hat{\mu}_t, \Omega_t)$ described as in the on-line supplementary material
6, *end for*
7, sample $s \sim \mathcal{D}(a/p^{\xi} + c_1, \ldots, a/p^{\xi} + c_p)$ where $c_j = \#\{b : \text{branch } b \text{ splits on predictor } j\}$
8, sample (σ, σ_μ, a) as described in the supplementary materials

`bartMachine` package in R. An alternative approach to the inclusion of categorical predictors, which is used in the `BayesTree` package, is to construct decision rules based on a dummy variable $Z_j = I(X_j \in A_b)$ where A_b is a random subset of the possible values of predictor j. In our illustrations, we let $\omega \to \infty$ so that $v_{mk} = P_m^{-1}$ and set $a/(a+M) \sim \text{Be}(0.5, 1)$.

2.5. Posterior computation

We use the Bayesian backfitting approach that was described by Chipman *et al.* (2010) to construct a Markov chain Monte Carlo algorithm to sample approximately from the posterior.

Within algorithm 1 (Table 1), $\mathcal{T}_t$ is updated by using a Metropolis–Hastings proposal. Proposals consist of one of three possible moves: birth, which turns a leaf node into a branch node, death, which turns a branch node into a leaf node, and change, which changes the decision rule of a branch b. A detailed description of these moves, and their associated transition probabilities, is given in the on-line supplementary materials.

Constructing efficient updates for $\mathcal{T}_t$ and τ_t requires marginalizing over $\mathcal{M}_t$. Because the errors are assumed Gaussian, this marginalization can be carried out in closed form. The main computational drawback of the SBART relative to the BART method lies in this marginalization, as the SBART method requires computing a likelihood contribution for each leaf–observation pair, whereas the BART method requires only a single likelihood contribution for each tree. Hence, if the trees are deep, the BART algorithm will be substantially faster. By the construction of the prior, trees generally are not sufficiently deep for this difference to be prohibitive.

The Dirichlet prior $s \sim \mathcal{D}(a/p^{\xi}, \ldots, a/p^{\xi})$ enables a straightforward Gibbs sampling update, with the full conditional given by $s \sim \mathcal{D}(a/p^{\xi} + c_1, \ldots, a/p^{\xi} + c_p)$, where $c_j = \#\{b : \text{branch } b \text{ splits on predictor } j\}$. When the grouping prior is used we also obtain simple Gibbs sampling updates, with $u \sim \mathcal{D}(a/M + z_1, \ldots, a/M + z_M)$ and $v_m \sim \mathcal{D}(\omega/P_m + c_{m1}, \ldots, \omega/P_m + c_{mP_m})$, where $z_m = \#\{b : \text{branch } b \text{ splits on a predictor in group } m\}$ and $c_{mk} = \#\{b : \text{branch } b \text{ splits on predictor } mk\}$.

3. Theoretical results

We study the theoretical properties of the SBART procedure from a frequentist perspective by assuming that $(Y_1, Y_2, \ldots, Y_n)$ are generated from the model $Y_i = f_0(X_i) + \epsilon_i$ with some true unknown regression function f_0. We assume that f_0 is a function over $[0, 1]^p$. We prove posterior consistency results when f_0 is a member of certain Hölder spaces. Let $\mathcal{C}^{\alpha}([0, 1]^p)$ denote the Hölder space with smoothness index α, i.e. the space of functions on $[0, 1]^p$ with bounded partial derivatives up to order β, where β is the largest integer strictly less than α and such that the partial derivatives of order β are Hölder continuous of order $\alpha - \beta$. Let $\mathcal{C}^{\alpha, R}([0, 1]^p) = \{f \in$

$\mathcal{C}^\alpha([0,1]^p) : \|f\|_\alpha \leqslant R\}$ denote the Hölder ball of radius R with respect to the Hölder norm $\|f\|_\alpha$ (see Ghosal and van der Vaart (2017), appendix C).

We consider the posterior convergence of the Bayesian fractional posterior obtained by raising the likelihood function by a factor $\eta \in (0, 1]$ in the Bayes formula

$$\Pi_{n,\eta}(A) = \frac{\int_A \prod_{i=1}^n p_f(Y_i|X_i)^\eta \, \Pi(\mathrm{d}f)}{\int \prod_{i=1}^n p_f(Y_i|X_i)^\eta \, \Pi(\mathrm{d}f)}, \tag{4}$$

where Π denotes the prior probability measure over $\mathcal{L}^2([0,1]^p)$: the $\mathcal{L}^2$-space over $[0,1]^p$. Fractional posteriors have gained renewed attention in Bayesian statistics because of their robustness to model misspecification (Grünwald, 2012; Miller and Dunson, 2018). According to Walker and Hjort (2001), the fractional posterior can be viewed as combining the original likelihood function with a data-dependent prior that is divided by a portion of the likelihood. This data-dependent reweighting in the prior helps to prevent possible inconsistencies by reducing the weights of those parameter values that 'track the data too closely'. Additionally, the fractional posterior with $\eta < 1$ permits much simpler theoretical analyses. Note that $\eta = 1$ corresponds to the usual posterior distribution. Abusing the notation slightly, we shall also use Π to denote the prior probability measure over the parameters $(\mathcal{T}_t, \mathcal{M}_t)$ and any hyperparameters in the model. Our goal is to find a sequence $\{\varepsilon_n : n \geqslant 1\}$ such that, for a sufficiently large constant M and fixed η,

$$\Pi_{n,\eta}[\|f - f_0\|_n \geqslant M\varepsilon_n] \to 0, \qquad \text{in probability as } n, p \to \infty, \tag{5}$$

where $\|\cdot\|_n$ denotes the $\mathcal{L}^2(\mathbb{P}_n)$ norm on the function space $\mathcal{L}^2([0,1]^p)$ defined by

$$\|f - g\|_n^2 = n^{-1} \sum_{i=1}^n \{f(X_i) - g(X_i)\}^2.$$

The sequence ε_n is then an upper bound on the posterior contraction rate. The norm $\|\cdot\|_n$ is a commonly adopted discrepancy metric in function estimation problems.

In this section, we focus on establishing result (5) for $\eta < 1$. The benefit of considering $\eta < 1$ is that this allows us to bypass verifying technical conditions regarding the effective support of the prior and the existence of a certain sieve (Ghosal *et al.*, 2000; Ghosal and van der Vaart, 2007a), which enables result (5) to be established under very weak conditions. In the online supplementary material we establish posterior consistency for $\eta = 1$ under more stringent conditions on the prior.

The main condition governing the posterior contraction rate is that the prior Π is sufficiently 'thick' at f_0, in the sense that there is a $C > 0$ such that

$$\Pi\{B_{\varepsilon_n}(f_0)\} \geqslant \exp(-Cn\varepsilon_n^2), \tag{6}$$

where $B_\varepsilon(f_0)$ denotes an ε-Kullback–Leibler neighbourhood of the truth

$$B_\epsilon(f_0) = \left\{ f : n^{-1} \sum_{i=1}^n \int p_{f_0}^{(i)} \log\left(\frac{p_{f_0}^{(i)}}{p_f^{(i)}}\right) \mathrm{d}y \leqslant \varepsilon^2 \right\} \bigcap \left\{ f : n^{-1} \sum_{i=1}^n \int p_{f_0}^{(i)} \log^2\left(\frac{p_{f_0}^{(i)}}{p_f^{(i)}}\right) \mathrm{d}y \leqslant \varepsilon^2 \right\},$$

where $p_f^{(i)}$ denotes the ith Gaussian density with mean $f(X_i)$ and variance σ^2. For convenience, we adopt the customary practice of assuming that σ is fixed and known when studying the posterior contraction rate. In the regression setting, it is straightforward to verify that the Kullback–

Leibler neighbourhood $B_\varepsilon(f_0)$ contains the $\mathcal{L}^2(\mathbb{P}_n)$ neighbourhood $\{\|f - f_0\|_n \leqslant 2\sigma\varepsilon\}$. Therefore, to establish condition (6), it suffices to find ε_n such that

$$\Pi(\|f - f_0\|_\infty \leqslant 2\sigma\varepsilon_n) \geqslant \exp(-Cn\varepsilon_n^2)$$

holds, where $\|g\|_\infty = \sup_{x\in[0,1]^p} |g(x)|$ denotes the supremum norm of a function g in $\mathcal{L}^2([0,1]^p)$.

We establish condition (6) for a wide class of tree-based models by deriving sharp small ball probabilities in the $\|\cdot\|_\infty$ norm around the true regression function f_0. To be general, we consider any gating function $\psi : \mathbb{R} \to \mathbb{R}$ satisfying the following assumption.

Assumption 1 (gating function). Let $K = \psi(1-\psi)$ be an 'effective' kernel function associated with gating function ψ such that $\sup_{x\in\mathbb{R}} |\psi'(x)| < \infty$.

(a) $\int_{-\infty}^{\infty} K(x)\mathrm{d}x > 0$ and, for any positive integer m, $\int_{-\infty}^{\infty} |x|^m |K(x)|\mathrm{d}x < \infty$.
(b) The function K can be extended to a uniformly bounded analytic function on the strip $\mathcal{S}(\rho) = \{z = x + \sqrt{(-1)}\, y \in \mathbb{C} : (x, y) \in \mathbb{R}^2, |y| \leqslant \rho\}$ in the complex plane for some constant $\rho > 0$.

Recall that μ_{tl} is the value that is assigned to leaf l of tree t, for $l = 1, 2, \ldots, L_t$ and $t = 1, \ldots, T$, and τ_b is the bandwidth parameter that is associated with branch b. Our first result shows that any smooth function can be approximated by a sum of soft decision trees taking form (1) in a way such that the number of trees T and the approximation error are optimally balanced. This lemma is interesting in its own right since it indicates that any d-dimensional smooth function can be approximated within error ε by using at most $\mathrm{poly}(\varepsilon^{-1})$ many properly rescaled logistic activation functions.

Lemma 1 (approximation by sum of soft decision trees). Suppose that the gating assumption 1 holds for the gating function ψ. For any function $f_0 \in \mathcal{C}^{\alpha,R}([0,1]^d)$, any $\epsilon > 0$ and $\tau > 0$, there is a sum of soft decision trees with a single bandwidth $\tau_b \equiv \tau$ for all branches,

$$\tilde{f}(x) = \sum_{t=1}^{T} g(x; \tilde{\mathcal{T}}_t, \tilde{\mathcal{M}}_t), \qquad x \in \mathbb{R}^p,$$

where each tree $\tilde{\mathcal{T}}_t$ has at most $2d$ branches, $T \leqslant C_1\tau^{-d}\log^d(1/\epsilon)$, $\Sigma_{t,l}|\tilde{\mu}_{tl}| \leqslant C_1\tau^{-d}\|f_0\|_\infty$ and

$$\|\tilde{f} - f_0\|_\infty \leqslant D_1 R(\tau^\alpha + \varepsilon\tau^{-d}),$$

where C_1 and D_1 are constants independent of (ε, τ).

With the help of lemma 1, we establish result (6) as a direct consequence of the following result, where we make the following assumptions on the prior distribution.

Assumption 2 (prior conditions).

(a) There are some constants (C_1, C_2) such that the prior distribution on the number of trees T satisfies $\Pi(T = t) \geqslant C_1\exp(-C_2 t)$ for $t = 0, 1, 2, \ldots$.
(b) The prior density π_τ of tree-specific bandwidth parameters τ_t satisfies $\pi_\tau(\tau) \geqslant a_1\tau^{a_2}$ for some constants $a_1, a_2 > 0$ for all sufficiently small τ.
(c) The prior on the splitting proportion vector s is $\mathcal{D}(a/p^\xi, \ldots, a/p^\xi)$ for some $\xi > 1$ and $a > 0$.
(d) The leaf coefficients μ_{tl} are independent and identically distributed with density π_μ where $\pi_\mu(\mu) \geqslant B_1\exp(-B_2|\mu|)$ for all μ and some positive constants B_1 and B_2.
(e) $\Pi(D_t = k) > 0$ for $k = 0, 1, \ldots, 2d$, where D_t denotes the depth of tree t and d is as in the following theorem 1.

Remark 1. Condition 2, part (a), is very weak and is satisfied, for example, by setting $T \sim$ geometric(π_T). Similarly, part (b) is satisfied by our choice of $\tau_t \sim \text{Exp}(r)$. Condition 2, part (d), which assumes that the μ_{tl}s have sufficiently heavy tails, is adopted for the simplicity of the assumption of independent and identically distributed coefficients but can be weakened to allow for the hierarchical model in which $\mu_{tl} \sim N(0, \sigma_\mu^2/T)$ with $\sigma_\mu \sim \text{Cauchy}_+(0, \sigma_\sigma)$.

Remark 2. In the on-line supplementary material we show that, under extra technical conditions on the prior, the usual posterior (fractional posterior with $\eta = 1$) can attain the same rate of convergence as in theorem 2 below. These extra conditions are needed to control the size of the effective support of the prior and show the existence of a certain sieve (Ghosal *et al.*, 2000). In particular, assumption 2 needs only certain lower bounds on the prior density (mass) functions, whereas assumption SP in the supplementary material requires some upper bound on the tail prior probability of various parameters in the model.

Theorem 1 (prior concentration for sparse function). Suppose that assumptions 1 and 2 are satisfied. Let $f_0 \in \mathcal{C}^{\alpha, R}([0,1]^p)$ be a bounded regression function that depends on at most d covariates. Then there are constants A and C independent of (n, p) such that, for all sufficiently large n, the prior Π over regression function f satisfies

$$\Pi(\|f - f_0\|_\infty \leqslant A\varepsilon_n) \geqslant \exp(-Cn\varepsilon_n^2),$$

where $\varepsilon_n = n^{-\alpha/(2\alpha+d)} \log(n)^t + \sqrt{\{n^{-1} d \log(p)\}}$ for any $t \geqslant \alpha(d+1)/(2\alpha+d)$.

The following posterior concentration rate for sparse functions follows immediately from theorem 1 and theorem 3.2 in Bhattacharya *et al.* (2016) (see also section 4.1 therein).

Theorem 2 (posterior convergence rate for sparse truth). Suppose that assumptions 1 and 2 are satisfied. Let $f_0 \in \mathcal{C}^{\alpha, R}([0,1]^p)$ be a bounded regression function that depends on at most only d covariates. If $n\varepsilon_n^2 \to \infty$ and $\varepsilon_n \to 0$ as $n, p \to \infty$, then, for all sufficiently large constants $M > 0$, we have

$$\Pi_{n,\eta}(\|f - f_0\|_n \geqslant M\varepsilon_n) \to 0, \qquad \text{in probability as } n, p \to \infty,$$

where $\varepsilon_n = n^{-\alpha/(2\alpha+d)} \log^t(n) + \sqrt{\{n^{-1} d \log(p)\}}$ for any $t \geqslant \alpha(d+1)/(2\alpha+d)$.

This result shows a salient feature of our sum of soft decision trees model—by introducing the soft thresholding, the resulting posterior contraction rate adapts to the unknown smoothness level α of the truth f_0, attaining a nearly minimax rate (Yang and Tokdar, 2015) without the need to know α in advance. Our next result shows that, if the truth admits a sparse additive structure $f_0 = \Sigma_{v=1}^V f_{0,v}(x)$, where each additive component $f_{0,v}(x)$ is sparse and depends only on d_v covariates for $v = 1, \ldots, V$, then the posterior contraction rate also adaptively (with respect to both the additive structure and unknown smoothness of each additive component) attains a nearly minimax rate (Yang and Tokdar, 2015) up to $\log(n)$-terms, which leads to a second salient feature of the sum of soft decision tree model—it also adaptively learns any unknown lower order non-linear interactions between the covariates.

Theorem 3 (posterior convergence rate for additive sparse truth). Suppose that assumptions 1 and 2 are satisfied. Let $f_0 = \Sigma_{v=1}^V f_{0,v}$, where the vth additive component $f_{0,v}$ belongs to $\mathcal{C}^{\alpha_v, R}([0,1]^p)$ and is bounded and depends on at most only d_v covariates for $v = 1, \ldots, V$. If $n\varepsilon_n^2 \to \infty$ and $\varepsilon_n \to 0$ as $n, p \to \infty$, then, for all sufficiently large constant $M > 0$, we have

$$\Pi_{n,\eta}(\|f - f_0\|_n \geqslant M\varepsilon_n) \to 0, \qquad \text{in probability as } n, p \to \infty,$$

where

$$\varepsilon_n = \sum_{v=1}^{V} n^{-\alpha_v/(2\alpha_v+d_v)} \log^t(n) + \sum_{v=1}^{V} \sqrt{\{n^{-1} d_v \log(p)\}}$$

for any $t \geqslant \max_v \alpha_v(d_v+1)/(2\alpha_v+d_v)$.

4. Illustrations

4.1. Friedman's example

A standard test case, which was initially proposed by Friedman (1991) (see also Chipman *et al.* (2010)), sets

$$f_0(x) = 10\sin(\pi x_1 x_2) + 20(x_3 - 0.5)^2 + 10x_4 + 5x_5. \tag{7}$$

This $f_0(x)$ features two non-linear terms and two linear terms, with a non-linear interaction.

In this experiment, we consider $n = 250$ observations, $\sigma^2 \in \{1, 10\}$ and p from 5 to 1000 along an evenly spaced grid on the scale of $\log(p)$. We compare the SBART method with the BART, DART, gradient-boosted decision trees (`xgboost`), the lasso (`glmnet`) and random forests (`randomForest`) methods. A similar experiment was conducted by Linero (2018), who showed that the sparsity inducing prior that is used by the DART method resulted in substantial gains in performance over the BART method. The purpose of this experiment is to demonstrate the further gains which are possible when the smoothness of function (9) is also leveraged.

Methods are compared by root-mean-squared error, $\text{RMSE} = \{\int \{f(x) - \hat{f}(x)\}^2 \, \mathrm{d}x\}^{1/2}$, which is approximated by Monte Carlo integration. For the Bayesian procedures, we take $\hat{f}$ to be the pointwise posterior mean of f. The DART and SBART models use their respective default priors and were fitted by using 2500 warm-up iterations and 2500 sampling iterations, whereas cross-validation was used to tune the hyperparameters for the BART model. The non-Bayesian methods were tuned by using cross-validation for each replication of the experiment.

Results are given in Fig. 5. Among the methods considered, the SBART method performs the best, obtaining a sizable improvement over the DART method in both the low noise and the high noise settings. Because of the use of a sparsity inducing prior, both the DART and the SBART methods are largely invariant to the number of nuisance predictors, whereas random forests, the BART–CV method and boosting have errors increasing in $\log(p)$. The lasso also has a stable, albeit poor, performance as p increases.

We now compare the SBART with the DART method for the task of variable selection (see Linero (2018) for a detailed comparison of the DART, BART, random forests and lasso methods, which found that the DART method performed best among these methods). Our goal is to assess whether leveraging smoothness can improve on the good variable selection properties of the DART method. We modify Friedman's function, taking instead

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

where λ is a tuning parameter for the simulation. A variable is included if its posterior inclusion probability exceeds 50%. We consider $\lambda \in [0.1, 1]$. As measures of accuracy, we consider precision TP/(TP + FP), recall TP/(TP + FN) and F_1-score (harmonic mean of precision and recall), where TP, FP and FN denote the number of true positive, false positive and false negative results respectively.

Results for 20 replications and $\sigma^2 = 1$ are given in Fig. 6, along with the average RMSE. First, we see that both the DART and the SBART results have a precision which is roughly constant in

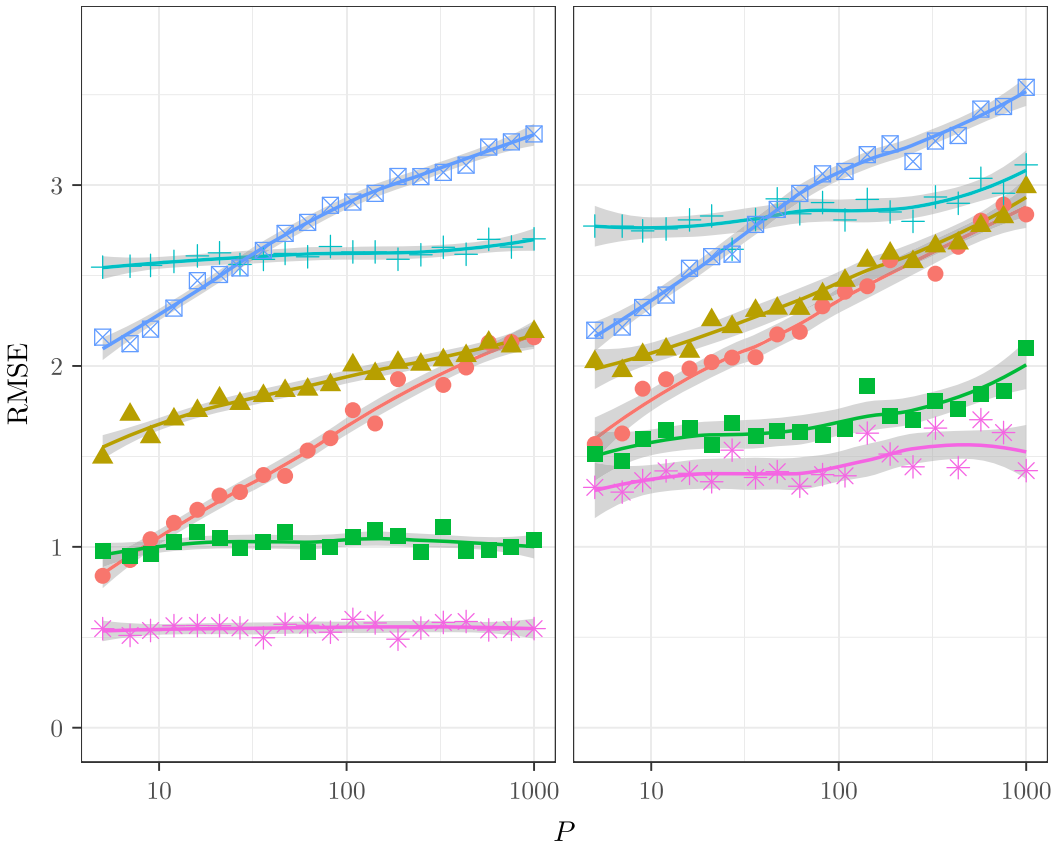


Fig. 5. Average root-mean-squared error of various methods, as a function of the dimension P of the predictor space (to aid visualization, we also give a LOESS smooth with Monte Carlo standard error); ●, BART–CV; ▲, boosting; ■, DART; +, lasso; ⊠, random forests; ✳, SoftBART): (a) $\sigma^2 = 1$; (b) $\sigma^2 = 10$

λ, with the SBART method performing uniformly better. This makes intuitive sense, as varying λ should have little influence on whether irrelevant predictors are selected. The precision of both methods is heavily dependent on λ, and we see that the SBART method is generally capable of detecting smaller signal levels; at its largest, the difference in recall is about 10%. Once the signal level is sufficiently high, both methods detect all relevant predictors consistently. The F_1-score reflects a mixture of these two behaviours. Perhaps most interesting is the influence of λ on RMSE. As λ increases, the performance of the DART method degrades whereas the SBART performance remains roughly constant. Intuitively this is because, as λ increases, the DART algorithm must use an increasing number of branches to capture the additional signal in the data, whereas the SBART method is capable of representing the effects corresponding to (x_4, x_5) with fewer parameters.

4.2. Approximation of non-smooth and locally smooth functions

A potential concern with the use of soft decision trees is that they may not be able to capture fine scale variability in the underlying regression function. An extreme example of this is when f is a step function. We consider the regression function $f(x) = 2 - 4I(x_1 < 0.5)$. In this case, one

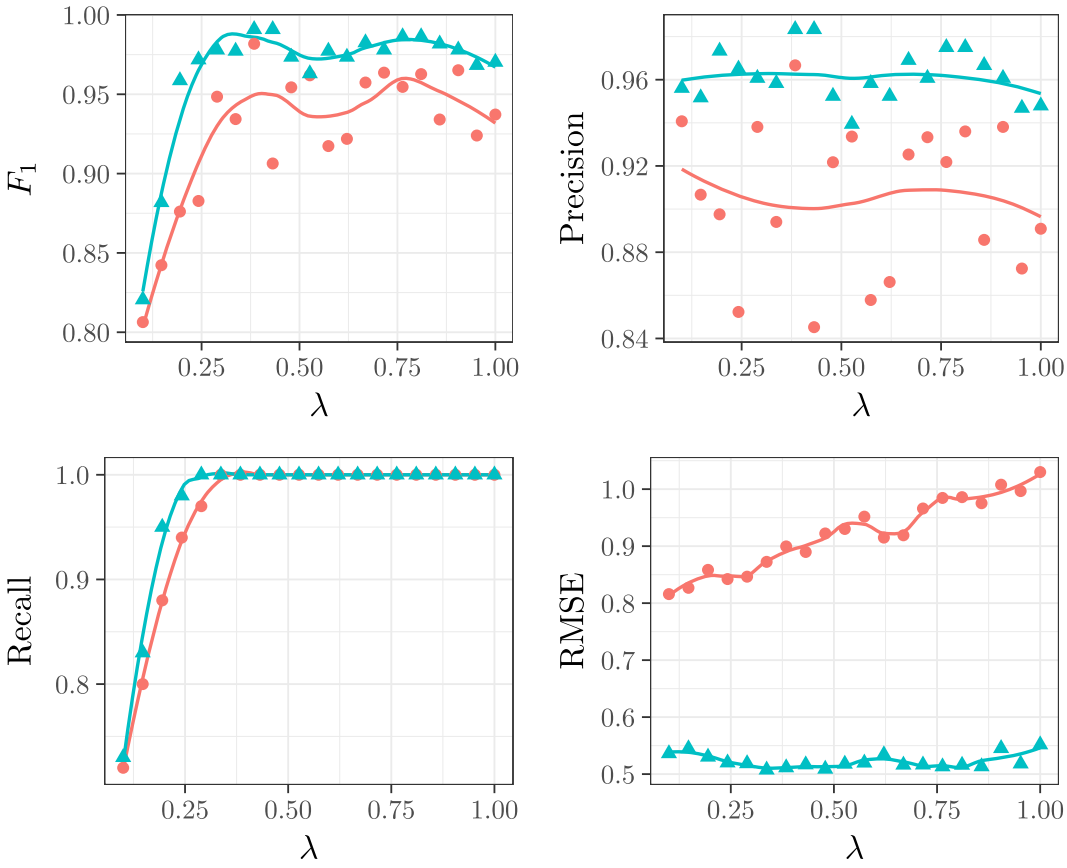


Fig. 6. Results for variable selection, with a LOESS smooth to aid visualization: ●, DART; ▲, SoftBART

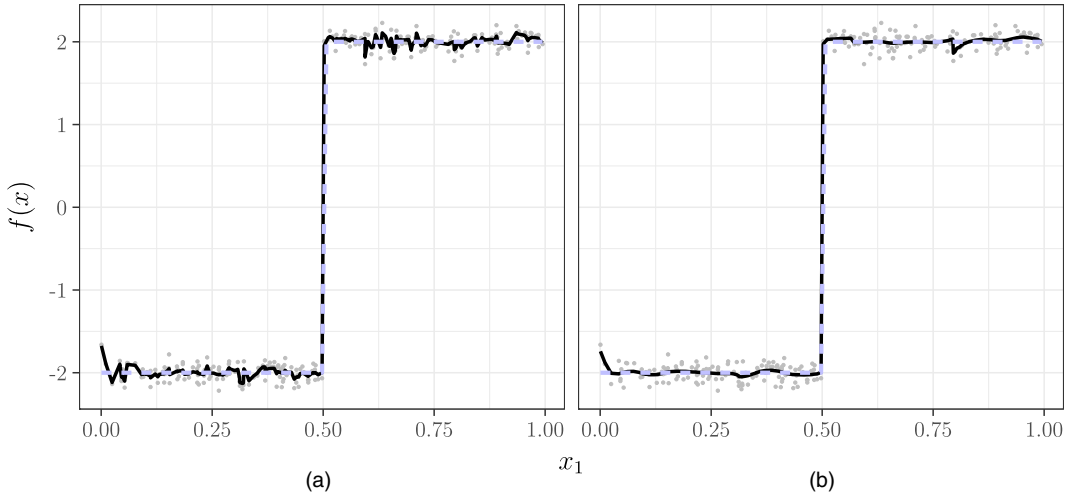


Fig. 7. Estimate of $f(x) = 2 - 4I(x_1 < 0.5)$ by using the posterior mean under (a) the BART and (b) SBART priors: – – –, true mean, ———, fit; ●, observed data

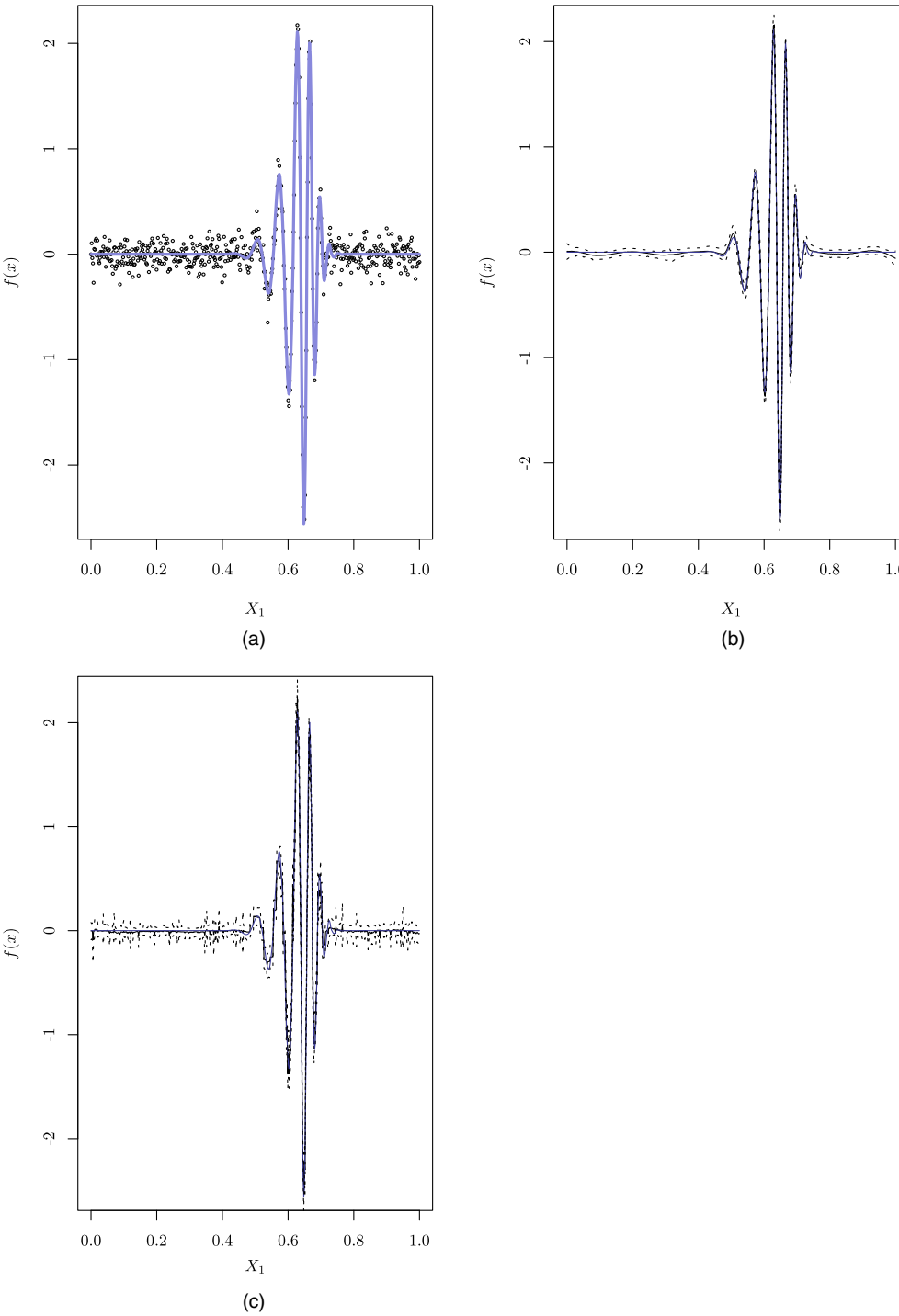


Fig. 8. (a) Raw data, consisting of observations drawn with the Daubechies wavelet as the mean function (———), (b) fit of the SBART model to the raw data, with pointwise 95% posterior credible bands, and (c) fit of the BART model to the raw data, with pointwise 95% posterior credible bands

might expect soft decision trees to perform suboptimally relative to hard decision trees because a soft decision tree must model the jump at 0 in a continuous fashion.

Surprisingly, ensembles of soft decision trees can outperform ensembles of hard decision trees even in this case. Fig. 7 shows fits of the BART and SBART models to $n = 250$ data points and a high signal of $\sigma = 0.1$. We see that both methods can capture the large jump discontinuity at $x_1 = 0.5$. The SBART algorithm performs better away from the discontinuity, however, because the level of smoothness is allowed to vary at different points in the covariate space. The trees that are responsible for the jump discontinuity have small τ_ts to replicate a step function effectively, whereas elsewhere the trees have large τ_ts to allow the function essentially to be constant.

The ability to select different τ_ts enables the SBART method to obtain locally adaptive behaviour. To illustrate this, Fig. 8 gives the fit of the BART and SBART models when $f(x)$ is a highly localized Daubechies wavelet of smoothness order 10. We see that the SBART method is capable of adapting both to the constant regions outside the support of the wavelet and the fast oscillatory behaviour within the support of the wavelet. The fit of the BART model, by contrast, has many artefacts outside the support of the wavelet and has generally wider credible bands.

4.3. Benchmark data sets

We compare the SBART method with various tree-based and non-tree-based methods on several benchmark data sets. We consider the BART, DART, lasso (`glmnet`), random forests (`randomForest`) and gradient-boosted decision trees (`xgboost`) methods. The parameters for the non-Bayesian procedures were chosen, separately for each fit, by using the `caret` package. Default priors (with $T = 50$) for the SBART and DART models were used; additionally, we consider selecting the hyperparameters of the SBART and BART models by cross-validation.

10 data sets are considered. Aside from bbb and wipp, the data sets are a subset of those considered by Kim *et al.* (2007). Although we consider only a subset of these data sets, no data sets that were considered for this experiment were omitted. Attributes of these data sets are presented in Table 2. The response in each data set was transformed to be approximately Gaussian. The bbb, triazines and wipp data sets were also considered by Linero (2018) to illustrate features of the sparsity inducing priors for decision tree methods.

Results of the experiment are given in Table 2. Methods are compared by an estimate of their root-mean predictive error obtained by using fivefold cross-validation, with the results averaged over 20 replications of the cross-validation. For each experiment, the root-mean predictive error for each method is normalized by the root mean predictive error for the SBART–CV method, so that scores higher than 1.00 correspond to worse performance than that of the SBART algorithm and scores lower than 1.00 correspond to better performance.

The SBART or SBART–CV method is seen to perform very well in practice, attaining the best performance on eight out of the 10 data sets. The results here are consistent with the general observation of Chipman *et al.* (2010) that the BART method outperforms gradient boosting and random forests in aggregate over many data sets. Two data sets stand out as particularly interesting. First, for the tecator data set, the SBART method outperforms all other methods by a very wide margin, indicating that leveraging smoothness for this data set is essential for attaining good performance. Second, the only data set for which the SBART–CV method substantially outperforms the SBART method is the hatco data set, where tuning the number of trees is required to attain optimal performance. This indicates that, for most data sets, the default SBART procedure works very well, but that if we want to be absolutely sure of optimal performance we should tune T.

Table 2. Results of the experiment described in Section 4.3†

Data	*Results for the following methods:*						
	BART–CV	*DART*	*SBART*	*RF*	*XGB*	*Lasso*	*SBART–CV*
ais	*1.00* (1)	*1.00* (1)	*1.00* (1)	1.01 (5)	1.03 (6)	1.04 (7)	*1.00* (1)
abalone	1.03 (4)	1.03 (4)	*1.00* (1)	1.02 (3)	1.03 (4)	1.12 (7)	*1.00* (1)
bbb	1.07 (6)	1.04 (4)	*0.99* (1)	1.01 (3)	1.05 (5)	1.10 (7)	1.00 (2)
cpu	0.98 (2)	1.01 (5)	1.01 (4)	*0.97* (1)	1.02 (6)	1.31 (7)	1.00 (3)
diamonds	1.15 (4)	1.07 (3)	1.01 (2)	2.29 (6)	1.43 (5)	3.53 (7)	*1.00* (1)
hatco	1.14 (3)	1.15 (4)	1.10 (2)	1.39 (6)	1.20 (5)	1.44 (7)	*1.00* (1)
servo	1.02 (3)	1.02 (3)	*0.99* (1)	1.17 (6)	1.06 (5)	1.75 (7)	1.00 (2)
tecator	1.87 (4)	1.63 (4)	*0.98* (1)	1.95 (7)	1.56 (3)	1.85 (5)	1.00 (2)
triazines	0.98 (3)	0.99 (4)	0.99 (4)	*0.92* (1)	0.94 (2)	1.13 (7)	1.00 (6)
wipp	1.19 (4)	1.14 (3)	1.03 (2)	1.43 (7)	1.28 (5)	1.41 (6)	*1.00* (1)
Average RMPE	1.14 (4)	1.11 (3)	1.01 (2)	1.32 (6)	1.16 (5)	1.57 (7)	*1.00* (1)
Average rank	3.4 (3)	3.5 (4)	*1.9* (1)	4.5 (5)	4.6 (6)	6.7 (7)	2 (2)

†The columns associated with the methods give their root mean predictive error RMPE, normalized by the root mean predictive error of the SBART–CV method. In parentheses, we give the rank of the method among the five approaches. The best-ranked method for each data set is given in italics.

5. Discussion

We have introduced a novel Bayesian sum-of-trees framework and demonstrated that it can attain a meaningful improvement over existing methods both in simulated experiments and in practice. This was accomplished by incorporating soft decision trees and sparsity inducing priors. We also provided theoretical support in the form of nearly optimal results for posterior concentration, adaptively over smoothness classes, when $f_0(x)$ is a sparse, or additive, function.

Although this paper has focused only on the case of non-parametric regression, the methodology proposed extends in a straightforward manner to other settings. For example, the case of binary classification can be addressed in the usual way via a probit link and data augmentation.

Our theoretical results concern the rate of convergence of the posterior. Another relevant question is whether the model can consistently estimate the model support, i.e. we can ask under what conditions $\Pi(S = S_0|\mathcal{D}) \to 1$ as $n \to \infty$, where $S = \{p\colon$ predictor p appears in the ensemble$\}$ and $S_0 = \{p : f_0$ depends on $p\}$. This is an interesting area for future research.

Software which implements the SBART method is available on line from `https://github.com/theodds/SoftBART` and is undergoing active development. Our code is based on the implementation of the BART model in the `BayesTree` package. Given enough optimization, we hope that our implementation could reach speeds within a modest factor of existing highly optimized implementations of the BART algorithm (Kapelner and Bleich, 2016).

Acknowledgements

This work was partially supported by National Science Foundation grant DMS-1712870 and Department of Defense grant SOT-FSU-FATs-16-06.

Appendix A: Proof of lemma 1

Let $K_\tau^{(d)}(x_1, x_2, \ldots, x_d) = \tau^{-d}\Pi_{j=1}^d K(x_j/\tau)$ denote a d-dimensional tensor product of the rescaled one-

dimensional kernel K in assumption 1, where recall that τ is the bandwidth parameter in the gating function. Let $C_K := \int K(x)\mathrm{d}x$ denote the normalization constant of K, so that we can write $K = C_K\tilde{K}$, and the rescaled kernel function $\tilde{K}$ has a unit normalization constant. Also write $\tilde{K}_\tau^{(d)} = K_\tau^{(d)}/C_K^d$. It is easy to verify that $\tilde{K}$ also satisfies the two conditions in assumption 1, though it may not be associated with any $\tilde{\psi}$.

Our proof is composed of three steps. First, we provide error bound estimates of approximating any α-smooth function by a convolution $K_\tau^{(d)} * g$ with some carefully constructed function g for any $\tau > 0$. Second, we show that any continuous convolution $K_\tau^{(d)} * g$ can be approximated by a discrete sum $\Sigma_{t=1}^T \mu_t K_\tau^{(d)}(\cdot - x_t)$ with at most $O(\tau^{-d})$ atoms. Lastly, we provide an error bound estimate on approximating this sum of kernels with a sum of soft decision trees by identifying each kernel component $K_\tau^{(d)}(\cdot - x_t)$ as one particular leaf in the tth soft decision tree $g(x; \mathcal{T}_t, \mathcal{M}_t)$ whose depth is at most $2d$ via splitting at most $2d$ times, for $t = 1, \ldots, T$.

Step 1: this step follows as a direct result of the following lemma, which is adapted from lemma 3.4 of De Jonge *et al.* (2010).

Lemma 2. Under assumption 1, for any $f_0 \in \mathcal{C}^{\alpha,R}([0,1]^d)$, there are some constants (M_1, M_2) independent of τ, and a function $T_{b,\tau} f_0$ satisfying $\|T_{b,\tau} f_0\|_\infty \leqslant M_1 R$, such that

$$\|\tilde{K}_\tau^{(d)} * (T_{b,\tau} f_0) - f_0\|_\infty \leqslant M_2 R \tau^\alpha.$$

From lemma 2, we immediately have

$$\|K_\tau^{(d)} * g - f_0\|_\infty \leqslant M_2 R \tau^\alpha,$$

where $g = C_K^{-d} T_{b,\tau} f_0$ satisfies $\|g\|_\infty \leqslant M_1' R$, with $M_1' = C_K^{-d} M_1$ independent of τ.

Step 2: this step generalizes the theory of approximating a continuous one-dimensional density function by a mixture of Gaussian distributions that was developed in Ghosal and van der Vaart (2007b) by a location mixture of any kernel K satisfying assumption 1. We also extend their result from density estimation to general function estimation as demanded in our regression setting, where the target function f may not integrate to 1 and can take negative values. First, we state an extension of lemma 3.1 of Ghosal and van der Vaart (2001) from dimension 1 to dimension d, and from the Gaussian kernel to any kernel K satisfying assumption 1.

Lemma 3. Under assumption 1, for any probability density function p_0 on $[0,1]^d$, any $\epsilon > 0$ and $\tau \in (0,1)$, there is a discrete measure $P_\tau = \Sigma_{t=1}^T r_t \delta_{x_t}$ with $T \leqslant C_1 \tau^{-d} \log^d(1/\tau)$ support points such that $\Sigma_{t=1}^T r_t = 1$ and

$$\|K_\tau^{(d)} * p_0 - \sum_{t=1}^T r_t K_\tau^{(d)}(\cdot - x_t)\|_\infty \leqslant D_1 \epsilon / \tau^d,$$

where (C_1, D_1) are independent of τ and K.

Proof. We sketch only the key difference in the proof from lemma 3.1 of Ghosal and van der Vaart (2001) in the one-dimensional case, and a proof for extending the result from the one-dimensional case to the multi-dimensional case follows similar lines to those in the proof of theorem 7 in Shen *et al.* (2013) (by replacing the Gaussian kernel with the kernel K).

The only key property of the Gaussian kernel that was used in the proof of lemma 3.1 of Ghosal and van der Vaart (2001) is in bounding the remainder term in the kth-order Taylor series expansion in their equation (3.11), where they used the fact that, for any $k \geqslant 1$, the kth-order derivative of the standard Gaussian density function $\phi(x) = (2\pi)^{-1/2}\exp(-x^2/2)$ at the origin $x = 0$ satisfies the bound

$$|\phi^{(k)}(0)/k!| \leqslant C_1 \exp(-C_2 k),$$

for some sufficiently large constant $C_2 > 0$ (since we focus on the approximation error only over the unit interval $[0,1]$, we do not need to include the additional $\log(1/\varepsilon)$ term in equation (3.11) therein). Therefore, it suffices to verify a similar exponential decay bound for the kth-order derivative of function $K_\kappa := \tau^{-1} K(\cdot/\kappa)$ for some sufficiently large number $\kappa > 0$ depending on C. In fact, under assumption 1, $K(\cdot)$ can be analytically extended to the strip $\mathcal{S}(\rho)$ in the complex plane (for simplicity, we use the same notation K to denote this extension), which implies by applying Cauchy's integral formula that

$$\frac{K_\kappa^{(k)}(0)}{k!} = \frac{1}{2\pi\sqrt{(-1)}} \oint_{\Gamma_\kappa} \frac{K_\kappa(z)}{z^{k+1}} \mathrm{d}z,$$

where the closed path Γ_κ is chosen as a counterclockwise circle centred at the origin with radius $\kappa\rho$. Since K is uniformly bounded on the path Γ_κ by assumption 1, we can further deduce that

$$\left| \frac{K_\kappa^{(k)}(0)}{k!} \right| \leqslant \frac{D}{\kappa^{k+2}\rho^{k+1}} \leqslant D\exp(-C_2 k)$$

holds as long as $\kappa \geqslant \rho^{-1}\exp(C_2)$, where D is some constant depending only on K, which completes the proof. □

With lemma 3 on the density function approximation as our preparation, we now return to the problem of approximating any general bounded function g over $[0,1]^d$. Note that we always have the decomposition $g = g_+ - g_-$ where $g_+ = \max\{0, g(x)\}$ and $g_-(x) = \max\{0, -g(x)\}$ are the positive parts and negative parts of g respectively, and both of them are non-negative and bounded over $[0,1]^d$. Let $A_+ = \int_{[0,1]^d} g_+(x)\mathrm{d}x \leqslant \|g\|_\infty$ and $A_- = \int_{[0,1]^d} g_-(x)\mathrm{d}x \leqslant \|g\|_\infty$. It is obvious that g_+/A_+ and g_-/A_- are two legitimate probability density functions over $[0,1]^d$. By applying lemma 3, we can find two discrete measures $P_+ = \Sigma_{t=1}^{T_+} r_t^+ \delta_{x_t^+}$ and $P_- = \Sigma_{t=1}^{T_-} r_t^- \delta_{x_t^-}$ such that

$$|A_+^{-1} K_\tau^{(d)} * g_+(x) - \sum_{t=1}^{T_+} r_t^+ K_\tau^{(d)}(x - x_t^+)| \leqslant D\varepsilon/\tau^d,$$
$$|A_-^{-1} K_\tau^{(d)} * g_-(x) - \sum_{t=1}^{T_-} r_t^- K_\tau^{(d)}(x - x_t^-)| \leqslant D\varepsilon/\tau^d,$$

for any $x \in [0,1]^d$ and $\max\{T_+, T_-\} \leqslant C\tau^{-d}\log^d(1/\varepsilon)$. Now we combine these two discrete measures into a new discrete signed measure

$$P_0 = \sum_{t=1}^{T_+} A_+ r_t^+ K_\tau^{(d)}(x - x_t^+) + \sum_{t=1}^{T_-} (-A_- r_t^-) K_\tau^{(d)}(x - x_t^-),$$

which will be denoted as $\Sigma_{t=1}^{T} \mu_t K_\tau^{(d)}(\cdot - x_t)$. Then $T \leqslant T_- + T_+ \leqslant 2C\tau^{-d}\log^d(1/\varepsilon)$ and

$$|K_\tau^{(d)} * g(x) - \sum_{t=1}^{T} \mu_t K_\tau^{(d)}(x - x_t)| \leqslant (A_+ + A_-) D\varepsilon/\tau^d \leqslant 2D\|g\|_\infty \varepsilon/\tau^d,$$

for all $x \in [0,1]^d$. Moreover, we have $\Sigma_{t=1}^{T} |\mu_t| \leqslant A_+ \Sigma_{t=1}^{T_+} r_t^+ + A_- \Sigma_{t=1}^{T_-} r_t^- \leqslant 2\|g\|_\infty$.

Step 3: in the last step, for each component $\mu_t K_\tau^{(d)}(\cdot - x_t)$ in the sum, we construct a soft decision tree $\tilde{\mathcal{T}}_t$ and its associated leaf values $\tilde{\mathcal{M}}_t$ in a way such that

(a) the tree splits exactly $2d$ times and
(b) the weight function $\phi(x; \mathcal{T}_t, l_t)$ that is specified in equation (2) associated with one particular leaf l_t equals $\tau^d K_\tau^{(d)}(\cdot - x_t)$,

so that the existence of the sum of soft decision trees follows by setting the values $\tilde{\mu}_{tl}$ that are associated with other leaves $l \neq l_t$ in this trees to be 0, and the value of this leaf as $\tilde{\mu}_{tl_t} = \tau^{-d}\mu_t$. In fact, for any $y = (y_1, \ldots, y_d) \in [0,1]^d$, we have the decomposition $K_\tau^{(d)}(y) = \Pi_{j=1}^d \tau^{-d}\psi(y_j/\tau)\{1 - \psi(y_j/\tau)\}$. Consequently, we can construct the tree $\mathcal{T}_t$ by sequentially splitting twice along each co-ordinate $x_{t,j}$ $(j = 1, 2, \ldots, d)$ of the centre $x_t = (x_{t,1}, \ldots, x_{t,d})$ in $\mu_t K_\tau^{(d)}(\cdot - x_t)$, so that the particular leaf at the end point of the path that goes once left and once right at the two branches associated with $x_{t,j}$, for $j = 1, \ldots, d$, receives weight

$$\phi(\cdot; \mathcal{T}_t, l_t) = \prod_{j=1}^{d} \psi\left(\frac{\cdot - x_{t,j}}{\tau}\right)\left\{1 - \psi\left(\frac{\cdot - x_{t,j}}{\tau}\right)\right\} = \tau^d K_\tau^{(d)}(x_t),$$

implying that, for any x, $g(x; \tilde{\mathcal{T}}_t, \tilde{\mathcal{M}}_t) = \tilde{\mu}_{tl_t}\phi(x; \tilde{\mathcal{T}}_t, l_t) = \mu_t K_\tau^{(d)}(x - x_t)$. Since this construction is valid for any $t = 1, \ldots, T$, we have $\Sigma_{t=1}^{T} \mu_t K_\tau^{(d)}(x - x_t) = \Sigma_{t=1}^{T} g(x; \tilde{\mathcal{T}}_t, \tilde{\mathcal{M}}_t)$.

Finally, a combination of steps 1–3 together yields a proof of lemma 3.

Appendix B: Proof of theorem 1

For convenience, we use the same notation C to denote some constant independent of (n, p), whose value may change from line to line. Without loss of generality, we may assume that f_0 depends only on its first d co-ordinates. Applying lemma 1, we obtain that, for some parameters τ and ε to be determined later, there exists some $\tilde{f} = \sum_{t=1}^{\tilde{T}} g(x; \tilde{\mathcal{T}}_t, \tilde{\mathcal{M}}_t)$ such that $\tilde{T} \leqslant C\tau^{-d}\log^d(\varepsilon^{-1})$, $\|\tilde{f} - f_0\|_\infty \leqslant C(\tau^\alpha + \varepsilon\tau^{-d})$ and the total number of splits (all are along the first d co-ordinates) across all trees are at most $2d\tilde{T}$ ($2d\tilde{T}$ many leaves in total).

Recall that our prior over the sum of soft decision tree function f is specified in a hierarchical manner: first, we specify the number T of trees and the tree topology $\mathcal{T} = \{\mathcal{T}_1, \ldots, \mathcal{T}_T\}$; second, conditionally on these we decide the co-ordinates in all splits across all the decision trees; third, we sample the independent splitting locations along all the selected co-ordinates; last, we sample bandwidth parameters τ_t that are associated with each tree and parameters μ associated with all leaves across the trees. We denote by $\tilde{T}$ and $\tilde{\mathcal{T}}$ the corresponding number of trees and the tree topology of $\tilde{f}$.

We denote all the splitting co-ordinates of f given T and the tree topology $\mathcal{T}$ by $S \in \{1, \ldots, p\}^N$, where $N = \sum_{t=1}^{T}(L_t - 1) \leqslant 2dT$ and recall that L_t denotes the number of leaves in the tth tree, and denote by $\tilde{S}$ the corresponding vector that is associated with $\tilde{f}$. We also denote the set of all splitting locations (along the selected splitting co-ordinates) and bandwidths as $x = (x_1, x_2, \ldots, x_N) \in \mathbb{R}^N$ and $\tau_S = (\tau_1, \tau_2, \ldots, \tau_T) \in \mathbb{R}_+^T$ respectively, and the set of all leaf values as $\mu = (\mu_1, \ldots, \mu_{N+T}) \in \mathbb{R}^{N+T}$. We also define $\tilde{x}^N$ and $\tilde{\mu}$ in a similar way. By construction, it is easy to check that, if f shares the same T, tree topology $\mathcal{T}$ and splitting co-ordinates S as $\tilde{f}$, then if $\{x, \tau_S, \mu\}$ are sufficiently close to $\{\tilde{x}, \tau, \tilde{\mu}\}$ in the sense that, for any $\delta > 0$,

$$\max_{u=1,\ldots,N} |x_u - \tilde{x}_u| \leqslant C\tau^{2d}\delta,$$

$$\max_{u=1,\ldots,T} |\tau_u - \tau| \leqslant C\tau^{d+1}\delta$$

and

$$\max_{u=1,\ldots,N+T} |\mu_u - \tilde{\mu}_u| \leqslant CT^{-1}\tau^d\delta,$$

we have the following perturbation error bound by applying the triangle inequality:

$$\left| \sum_{t=1}^{T} g(x; \mathcal{T}_t, \mathcal{M}_t) - \sum_{t=1}^{\tilde{T}} g(x; \tilde{\mathcal{T}}_t, \tilde{\mathcal{M}}_t) \right| \leqslant C\delta, \qquad \text{for all } x \in [0,1]^p. \tag{8}$$

Now we apply theorem 2.1 in Yang and Dunson (2014) on the prior concentration probability for a high dimensional Dirichlet distribution and assumption 2, part (c), to obtain that the splitting proportion vector $s = (s_1, \ldots, s_p)$ satisfies

$$\Pi\left[s_j \geqslant (2d)^{-1} \text{ for } j = 1, \ldots, d, \text{ and } \sum_{j=d+1}^{p} s_j \leqslant d^{-1} \right] \geqslant \exp\{-Cd\log(p)\}. \tag{9}$$

This combined with the fact that each tree has depth at most $2d$ and assumption 2, part (e), implies that the prior probability of $\mathcal{T} = \tilde{\mathcal{T}}$ given $T = \tilde{T}$ can be lower bounded by

$$\Pi(\mathcal{T} = \tilde{\mathcal{T}} | T = \tilde{T}) \geqslant Cd^{-N} \geqslant \exp\{-C\tau^{-d}\log^d(\varepsilon^{-1})\},$$

where we have used the fact that $N \leqslant C\tau^{-d}\log^d(\varepsilon^{-1})$ in the last step. The perturbation error bound (8) implies that

$$\begin{aligned}\Pi(\|f - \tilde{f}\|_\infty \leqslant C\delta | \mathcal{T} = \tilde{\mathcal{T}}, T = \tilde{T}) &\geqslant \Pi\Bigg(\max_{u=1,\ldots,N} |x_u - \tilde{x}_u| \leqslant C\tau^{2d}, \ \max_{u=1,\ldots,T} |\tau_u - \tau| \leqslant C\tau^{d+1}\delta, \\ &\qquad \max_{u=1,\ldots,N+T} |\mu_u - \tilde{\mu}_u| \leqslant CT^{-1}\tau^d\delta | \mathcal{T} = \tilde{\mathcal{T}}, T = \tilde{T} \Bigg) \\ &\geqslant \exp[-C\tau^{-d}\log^d(\varepsilon^{-1})\log\{(\tau\delta)^{-1}\}],\end{aligned}$$

where in the last step we applied assumptions 2, parts (b) and (d), and used the fact that $\Sigma_u|\tilde{\mu}_u| \leqslant C\tau^{-d}$ for some constant C only depend on f_0 (because of lemma 1). Putting all pieces together and using assumption 2, part (a), and the properties of $\tilde{f}$, we obtain

$$\Pi\{\|f-f_0\|_\infty \leqslant C(\delta+\tau^\alpha+\varepsilon/\tau^d)\} \geqslant \Pi(T=\tilde{T})\,\Pi(\mathcal{T}=\tilde{\mathcal{T}}|T=\tilde{T})\,\Pi(\|f-\tilde{f}\|_\infty \leqslant C\delta|\mathcal{T}=\tilde{\mathcal{T}},T=\tilde{T})$$
$$\geqslant \exp[-C\tau^{-d}\log^d(\varepsilon^{-1})-Cd\log(p)-C\tau^{-d}\log^d(\varepsilon^{-1})\log\{(\tau\delta)^{-1}\}].$$

Therefore, by choosing $\tau=\{\log^{d+1}(n)/n\}^{-1/(2\alpha+d)}$, $\delta=\tau^\alpha$ and $\varepsilon=\tau^{d+\alpha}$, we can obtain the claimed prior concentration probability lower bound as $\Pi(\|f-f_0\|_\infty \leqslant C\varepsilon_n) \geqslant \exp(-Cn\varepsilon_n^2)$.

Appendix C: Proof of theorem 2

Using theorem 3.2 in Bhattacharya *et al*. (2016) (see also section 4.1 therein), it suffices to show that $\Pi(\|f-f_0\|_\infty \leqslant C\varepsilon_n) \geqslant \exp(-Cn\varepsilon_n^2)$. The proof of this is almost the same as that of theorem 1; the only difference is that now we apply lemma 1 to find V functions $\{\tilde{f}_v : v=1,\ldots,V\}$, where $\tilde{f}_v$ contains $\tilde{T}_v$ trees and approximates the vth additive component $f_{0,v}$ in f_0 for $v=1,\ldots,V$, and set $\tilde{f}=\sum_{v=1}^V \tilde{f}_v$. Because of the additive structure in our sum of soft decision tree model, we can always write $f=\sum_{v=1}^V f_v$ where f_v collects $\tilde{T}_v$ trees and has the same sum of soft decision trees prior structure when conditioning on the total number of trees $T=\sum_{v=1}^V \tilde{T}_v$, and the conditional priors of $(f_1,\ldots,f_v)$ given $T=\sum_{v=1}^V \tilde{T}_v$ and the splitting proportion vector s are independent. Let $\mathcal{S}=\{s_j \geqslant (2d)^{-1}$ for $j=1,\ldots,d$, and $\sum_{j=d+1}^p s_j \leqslant d^{-1}\}$ denote the event in inequality (9) with $d:=\sum_{v=1}^V d_v$. Therefore, we obtain, by applying assumption 2, the prior concentration bound (9) for s, and theorem 1 for a single f_v (choose parameters τ_v, δ_v and ε_v for each f_v as in the proof of theorem 1) that

$$\Pi\left(\|f-f_0\|_\infty \leqslant C\sum_{v=1}^V \varepsilon_{n,v}\right) \geqslant \Pi\left(T=\sum_{v=1}^V \tilde{T}_v\right)\Pi(s\in\mathcal{S})\sup_{s\in\mathcal{S}}\left\{\prod_{v=1}^V \Pi\left(\|f_v-f_{0,v}\|_\infty \leqslant C\varepsilon_{n,v}|T=\sum_{v=1}^V \tilde{T}_v,s\right)\right\}$$
$$\geqslant \exp\left\{-Cn\sum_{v=1}^V \varepsilon_{n,v}^2 - C\sum_{v=1}^V d_v\log(p)\right\} \geqslant \exp\left\{-C'n\left(\sum_{v=1}^V \varepsilon_{n,v}\right)^2\right\}$$

where constants $C, C'>0$, $\varepsilon_{n,v}=n^{-\alpha_v/(2\alpha_v+d_v)}\log^{t_v}(n)+\sqrt{\{n^{-1}d_v\log(p)\}}$ and $t_v \geqslant \alpha_v(d_v+1)/(2\alpha_v+d_v)$.

References

Alaa, A. M. and van der Schaar, M. (2018) Bayesian nonparametric causal inference: information rates and learning algorithms. *IEEE J. Selectd Top. Signal Process*., to be published, doi 10.1109/JSTP.2018.2848230.

Athreya, K. B. and Ney, P. E. (2004) *Branching Processes*. New York: Dover Publications.

Bhattacharya, A., Pati, D. and Yang, Y. (2016) Bayesian fractional posteriors. *Ann. Statist*., to be published.

Bleich, J., Kapelner, A., George, E. I. and Jensen, S. T. (2014) Variable selection for BART: an application to gene regulation. *Ann. Appl. Statist.*, **8**, 1750–1781.

Breiman, L. (2001) Random forests. *Mach. Learn.*, **45**, 5–32.

Chipman, H. A., George, E. I. and McCulloch, R. E. (2010) Bart: Bayesian additive regression trees. *Ann. Appl. Statist.*, **4**, 266–298.

De Jonge, R. and Van Zanten, J. (2010) Adaptive nonparametric Bayesian inference using location-scale mixture priors. *Ann. Statist.*, **38**, 3300–3320.

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

Freund, Y. and Schapire, R. (1999) A short introduction to boosting. *J. Jpn. Soc. Artif. Intell.*, **14**, 771–780.

Friedman, J. H. (1991) Multivariate adaptive regression splines. *Ann. Statist.*, **19**, 1–67.

Ghosal, S., Ghosh, J. K. and van der Vaart, A. W. (2000) Convergence rates of posterior distributions. *Ann. Statist.*, **28**, 500–531.

Ghosal, S. and van der Vaart, A. W. (2001) Entropies and rates of convergence for maximum likelihood and Bayes estimation for mixtures of normal densities. *Ann. Statist.*, **29**, 1233–1263.

Ghosal, S. and van der Vaart, A. (2007a) Convergence rates of posterior distributions for noniid observations. *Ann. Statist.*, **35**, 192–223.

Ghosal, S. and van der Vaart, A. (2007b) Posterior convergence rates of Dirichlet mixtures at smooth densities. *Ann. Statist.*, **35**, 697–723.

Ghosal, S. and van der Vaart, A. (2017) *Fundamentals of Nonparametric Bayesian Inference*. Cambridge: Cambridge University Press.

Green, P. J. (1995) Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika*, **82**, 711–732.

Grünwald, P. (2012) The safe Bayesian. In *Proc. Int. Conf. Algorithmic Learning Theory* (eds N. H. Bshouty, G. Stoltz, N. Vayatis and T. Zeugmann), pp. 169–183. New York: Springer.
Györfi, L., Kohler, M., Krzyzak, A. and Walk, H. (2006) *A Distribution-free Theory of Nonparametric Regression*. New York: Springer Science and Business Media.
Hahn, P. R., Murray, J. S. and Carvalho, C. M. (2017) Bayesian regression tree models for causal inference: regularization, confounding, and heterogeneous effects. *Preprint arXiv:1706.09523*. Department of Mathematics and Statistics, Arizon State University, Tucson.
Hastie, T., Tibshirani, R. and Friedman, J. (2009) *The Elements of Statistical Learning*, 2nd edn. New York: Springer.
Hill, J. L. (2011) Bayesian nonparametric modeling for causal inference. *J. Computnl Graph. Statist.*, **20**, 217–240.
Hill, J. L. (2016) Atlantic Causal Inference Conference Competition results. New York University, New York. (Available from `http://jenniferhill7.wixsite.com/acic-2016/competition`.)
Irsoy, O., Yildiz, O. T. and Alpaydin, E. (2012) Soft decision trees. In *Proc. Int. Conf. Pattern Recognition*, pp. 1819–1822. New York: Institute of Electrical and Electronics Engineers.
Kapelner, A. and Bleich, J. (2016) bartMachine: machine learning with Bayesian additive regression trees. *J. Statist. Softwr.*, **70**, no. 4, 1–40.
Kim, H., Loh, W.-Y., Shih, Y.-S. and Chaudhuri, P. (2007) Visualizable and interpretable regression models with good prediction power. *IIE Trans.*, **39**, 565–579.
Linero, A. R. (2018) Bayesian regression trees for high-dimensional prediction and variable selection. *J. Am. Statist. Ass.*, **113**, 626–636.
Miller, J. W. and Dunson, D. B. (2018) Robust Bayesian inference via coarsening. *J. Am. Statist. Ass.*, to be published.
Murray, J. S. (2017) Log-linear Bayesian additive regression trees for categorical and count responses. *Preprint arXiv:1701.01503*. Department of Information, Risk, and Operations Management, McCombe School of Business, University of Texas at Austin, Austin.
Rockova, V. and van der Pas, S. (2017) Posterior concentration for Bayesian regression trees and their ensembles. *Preprint arXiv:1078.08734*. Booth School of Business, University of Chicago, Chicago.
Shen, W., Tokdar, S. T. and Ghosal, S. (2013) Adaptive Bayesian multivariate density estimation with Dirichlet mixtures. *Biometrika*, **100**, 623–640.
Sparapani, R. A., Logan, B. R., McCulloch, R. E. and Laud, P. W. (2016) Nonparametric survival analysis using Bayesian additive regression trees (BART). *Statist. Med.*, **35**, 2741–2753.
Vehtari, A., Gelman, A. and Gabry, J. (2017) Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statist. Comput.*, **27**, 1413–1432.
Walker, S. and Hjort, N. L. (2001) On Bayesian consistency. *J. R. Statist. Soc.* B, **63**, 811–821.
Watanabe, S. (2013) A widely applicable Bayesian information criterion. *J. Mach. Learn. Res.*, **14**, 867–897.
Yang, Y. and Dunson, D. B. (2014) Minimax optimal Bayesian aggregation. *Preprint arXiv:1403.1345*. Department of Statistics, University of Illinois at Urbana–Champaign, Champaign.
Yang, Y. and Tokdar, S. T. (2015) Minimax-optimal nonparametric regression in high dimensions. *Ann. Statist.*, **43**, 652–674.
Yuan, M. and Lin, Y. (2006) Model selection and estimation in regression with grouped variables. *J. R. Statist. Soc.* B, **68**, 49–67.

Supporting information

Additional 'supporting information' may be found in the on-line version of this article:

'Supplementary material to Soft Bayesian additive regression trees: ensembles that adapt to smoothness and sparsity'.