Adaptive Conditional Distribution Estimation with Bayesian Decision Tree Ensembles

Yinpu Li, Antonio R. Linero, and Jared Murray

Florida State University, Tallahassee, FL; University of Texas at Austin, Austin, TX

ABSTRACT

We present a Bayesian nonparametric model for conditional distribution estimation using Bayesian additive regression trees (BART). The generative model we use is based on rejection sampling from a base model. Like other BART models, our model is flexible, has a default prior specification, and is computationally convenient. To address the distinguished role of the response in our BART model, we introduce an approach to targeted smoothing of BART models which is of independent interest. We study the proposed model theoretically and provide sufficient conditions for the posterior distribution to concentrate at close to the minimax optimal rate adaptively over smoothness classes in the high-dimensional regime in which many predictors are irrelevant. To fit our model, we propose a data augmentation algorithm which allows for existing BART samplers to be extended with minimal effort. We illustrate the performance of our methodology on simulated data and use it to study the relationship between education and body mass index using data from the medical expenditure panel survey (MEPS). Supplementary materials for this article are available online.

1. Introduction

We consider the Bayesian nonparametric estimation of the conditional distribution of a response $Y_i$ based on predictors $X_i$. A common strategy is to introduce a latent variable $b$, and set $Y_i \sim h(y \mid X_i, b, \theta)$ given $b$, where $h(y \mid x, b, \theta)$ is a parametric model. This includes mixture models where $b$ is a latent class and the conditional density has the mixture form $h(y \mid x) = \sum_{k=1}^{\infty} \omega_k(x) h(y \mid x, \theta_k)$ (MacEachern 1999; Dunson and Park 2008; Chung and Dunson 2009; Rodriguez and Dunson 2011), as well as Gaussian process latent variable models where $b$ is continuous (Kundu and Dunson 2014).

A conceptually simpler approach models $f(y \mid x)$ by tilting a base model:

$$f(y \mid x) = \frac{h(y \mid x, \theta) \Phi[r(y,x)]}{\int h(y' \mid x, \theta) \Phi[r(y',x)] \, dy'},$$

(1)

where $\{h(\cdot \mid \cdot, \theta) : \theta \in \Theta\}$ is a parametric family of conditional densities. We refer to $h(y \mid x, \theta)$ as the base model and $\Phi(\mu)$ as the link function. When $r(y, x)$ is a constant, this model reduces to the base model, allowing the user to center (1) on a desired parametric model. A special case of (1) sets $\Phi(\mu) = e^\mu$ and takes $r(y, x)$ to be a Gaussian process (Tokdar, Zhu, and Ghosh 2010). In the context of (1), the posterior possesses strong theoretical properties, obtaining posterior convergence rates close to the best possible. For routine use, a default prior can be used which obtains good practical performance. Shrinkage toward the base model $h(y \mid x, \theta)$ is easy to perform, so that the model naturally adapts to the complexity of the data.

We propose a modified variant of the Bayesian additive regression trees (BART) model of Chipman, George, and McCulloch (2010) which we refer to as the SBART density sampler (SBART-DS). We choose $r(y, x)$ to be a sum of soft decision tree (Linero and Yang 2018) which smooths in a targeted fashion on the response variable $y$ (Starling et al. 2020). A benefit of the BART framework is that we are able to develop default priors based on well-known heuristics and show that these default priors perform well in practice.

To construct inference algorithms, we restrict the choice of $\Phi(\mu)$ to the logit, probit, or $t_\nu$-link functions. Our proposal is similar to the GP-DS, but is adapted to conditional distribution estimation. We construct an efficient Markov chain Monte Carlo (MCMC) algorithm to sample from the posterior distribution by combining a data augmentation scheme of Rao, Lin, and Dunson (2016) with an additional layer of data augmentation. After performing this data augmentation, we can update the parameters of the model using the same Bayesian backfitting algorithm as Chipman, George, and McCulloch (2010). Given that one has the ability to perform Bayesian backfitting, the algorithms we construct are simple to implement.
We present theoretical results which show that suitably-specified SBART-DS priors attain convergence rates which are close to the best possible. Simplifying slightly, we show that in the high-dimensional sparse setting, where only $D - 1 < P$ of the $P$ predictors are relevant, SBART-DS can obtain the oracle rate of convergence $\epsilon_n = n^{-2\alpha/(2\alpha + D)}$ up-to a logarithmic term, where $\alpha$ is related to the smoothness level of the true conditional density. In a simulation study we show that these theoretical results are suggestive of what occurs in practice, as the performance of SBART-DS is robust to the presence of irrelevant predictors.

**Related Methods.** Starting with the works of Müller, Erkanli, and West (1996) and MacEachern (1999), there is a large literature which tackles the density regression problem from the perspective of infinite mixture modeling, with $f(y \mid x) = \sum_k \omega_k(x) h(y \mid x, \theta_k)$. This allows for shrinkage toward a base model $h(y \mid x, \theta)$ by choosing the prior so that $\omega_i(x) \approx 1$ for all $x$. The weights can be modeled implicitly by jointly modeling the predictors and responses (Müller, Erkanli, and West, 1996; Shahbaba and Neal, 2009; Wade et al., 2014). Alternatively, the weights can be directly modeled. A popular approach to modeling the weights is to use various extensions of the Sethuraman stick-breaking construction of the Dirichlet process $\omega_k(x) = V_k(x) \prod_{j=1}^{K-1} [1 - V_j(x)]$ (Sethuraman, 1994) such as the kernel stick breaking process (Dunson and Park, 2008) or probit stick-breaking process (Rodríguez and Dunson, 2011). Alternatively, Antoniano-Villalobos, Wade, and Walker (2014) model $\omega_k(x) \propto w_j K(x \mid \lambda_k)$ for some choice of kernel function $K(\cdot \mid \cdot)$. In Chung and Dunson (2009) the probit stick breaking process is extended to account for potentially high-dimensional predictors. Beyond infinite mixture models, there is a rich literature on Gaussian process based density regressors, either based on the structure (1) with the exponential link (Tokdar, Zhu, and Ghosh, 2010; Riihimäki and Vehtari, 2014) or based on a latent variable structure $f(y \mid x) = \int_0^\infty \text{Normal}(y \mid r(x, b), \sigma^2) \, db$ (Kundu and Dunson, 2014). Each of the above approaches comes with challenges—implementations of these models may not be easy for nonexperts, default priors may not be reliable, and our ability to shrink toward a base model may be limited.

**Why Density Regression?** Modeling of a conditional density is admittedly a difficult task relative to, for example, quantile regression or semiparametric mean regression, and for many problems a more targeted approach might be more appropriate. However, we find specification of a large Bayesian model attractive for a variety of reasons.

- Conditional density estimation provides estimates of arbitrary functionals of the conditional distribution, allowing for better translations of scientific questions to statistical questions. In addition to conditional means/quantiles, we can perform inference on higher order moments such as the conditional standard deviation, or the AUC describing the effect of a binary covariate on an outcome of interest. We demonstrate this on data from the Medical Expenditure Panel Survey, where we find that education has an effect on BMI which is stronger in the left tail than the right. We also perform inference on the probability that an individual with a given education level has a BMI which is 10% lower than the BMI of an otherwise-identical individual with a different education level. This flexibility allows for more creativity on the part of the analyst, as we need not restrict attention to simple functionals which can be conveniently estimated. Another example where this additional flexibility is desirable is Angrist, Chernozhukov, and Fernández-Val (2006), where the estimand is a nonlinear functional of the quantiles.
- Even if our goal is to handle a relatively simple task like estimating conditional quantiles, flexibly modeling the conditional density provides robustness to model misspecification, both in terms of consistency and uncertainty quantification. For example, a standard approach to performing quantile regression is to use an asymmetric Laplace error distribution (Yu and Moyeed, 2001), however, Yang, Wang, and He (2016) show that misspecification of the error distribution leads to poor uncertainty quantification. More advanced methods which avoid this issue, such as the joint quantile modeling approaches of Yang and Tokdar (2017) or Reich, Fuentes, and Dunson (2011), are in a sense tantamount to modeling the conditional distribution implicitly through linear models for the quantiles, rather than nonparametrically as we have here. By modeling $f(y \mid x)$ nonparametrically we guard ourselves against misspecification of the error distribution. This extends to other aspects of the model as well; for example, even if we were only interested in the conditional mean on the MEPS data, there is evident heteroscedasticity which would not be handled if we specified a semiparametric BART regression model, and while it is possible to design heteroscedastic BART models (see, e.g., Pratola et al., 2020) we avoid the issue of misspecification entirely by using density regression.
- By committing to a single density regression model up front, we are guaranteed that all of our inferences are coherent. If we used different models to analyze different features (say, quantile regression for quantiles and semiparametric regression for the mean) then it may be difficult to reconcile the results in a principled fashion; an extreme example of this is that separately analyzing quantiles may result in quantile curves which cross each other. We are also free to probe the posterior for interesting features without a-priori specifying them. For example, the shape/scale effects of the predictors in Section 4.3 were not known to us a-priori, and we would not have thought to look for them. By using a single Bayesian nonparametric model, we are able to examine the posterior for interesting features while still remaining principled from the Bayesian perspective. This feature of the logical coherence of Bayesian methods has recently been used to generate explainable Bayesian inference using loss functions (Hahn and Carvalho, 2015; Woody, Carvalho, and Murray, 2020). Finally, while it does not feature prominently in our application in this article, if we are fitting the model primarily to make predictions then a flexible model for the conditional density is important when the conditional distribution changes in complex ways as a function of covariates. It allows us to quantify uncertainty in our predictions, choose point predictions under different loss functions, and make other decisions based on our predictive distribution over future outcomes.
Outline. In Section 2 we review BART and describe a naive version of SBART-DS; we then describe our approach for targeted smoothing which centers our prior for $r(\cdot, x)$ on a Gaussian process. In Section 3 we present our theoretical results. In Section 4 we conduct a simulation study and apply SBART-DS to data from the Medical Expenditure Panel Survey (MEPS) to study the relationship between educational attainment and body mass index (BMI) in adult women. We conclude in Section 5 with a discussion. Proofs, detailed algorithms, and additional simulation results are given in the supplementary materials.

2. Model Description

2.1. Review of Bayesian Additive Regression Trees

The additive regression trees (BART) framework models a function $r(x)$ as a sum of regression trees $r(x) = \sum_{m=1}^{M} g(x; T_m, M_m)$ where $T_m$ denotes the topology and splitting rules of a binary decision tree and $M_m = \{\mu_{m1}, \ldots, \mu_{ml_m}\}$ gives a prediction for each of the $L_m$ terminal (leaf) nodes of $T_m$. Figure 1 gives a schematic which shows how predictions are obtained from a decision tree. Chipman, George, and McCulloch (2010) specify priors $\pi_T$ and $\pi_M$ for the tree topologies $T_m$ and the $\mu_{ml}$’s given $T_m$, respectively. We write $r \sim BART(\pi_T, \pi_M)$ to denote that $r$ has the associated BART prior. Typically we set $\mu_{ml} \sim \text{Normal}(0, \sigma^2_{ml}/M)$ so that $\text{var}(r(x)) = \sigma^2_x$ regardless of the number of trees used in the model.

A problem with methods based on decision trees is that realizations of $r \sim \text{BART}(\pi_T, \pi_M)$ will not be continuous in $x$. This is particularly problematic for density estimation, as we generally prefer estimates of the density to be smooth. A smooth variant of BART called soft BART (SBART) was introduced by Linero and Yang (2018). SBART takes the tree $T_m$ to be a smooth decision tree, where observations are assigned a weight $\varphi_{ml}(x)$ to leaf node $\ell$ of tree $m$. As a point of comparison, nonsmooth decision trees use the weights $\varphi_{ml}(x) = \prod_{b \in A_m} I(x_{jb} \leq C_b) (1 - R_b) I(x_{jb} > C_b) R_b$, where $A_m$ denotes the collection of branches which are ancestors of leaf $\ell$ of tree $m$, $j_b$ denotes the coordinate along which $b$ splits, $C_b$ denotes the cutpoint of branch $b$, and $R_b$ is the indicator that the path from the root to the leaf goes right at $b$. A soft decision tree instead takes $\varphi_{ml}(x) = \prod_{b \in A_m} \psi(x; C_b, R_b) 1 - \psi(x; C_b, R_b) R_b$, where $\psi(x; c, \tau)$ is the cumulative distribution function of a location-scale family with location $c$ and scale $\tau$. If $\psi(x) = I(x \leq 0)$ (or, equivalently, as $\tau \to 0$) we get a standard decision tree. If we instead take $\psi(x)$ to be a smooth function then $r(y, x)$ will also be smooth. The parameter $\tau$ is analogous to a bandwidth parameter, with larger values of $\tau$ giving smoother functions. Like Linero and Yang (2018) we will take $\psi(x) = (1 + e^{-x})^{-1}$ and use tree-specific bandwidths $\tau_m$. We write $r \sim \text{SBART}(\pi_T, \pi_M)$ to denote that $r$ has an SBART prior, where $\pi_T$ is now a prior over the soft trees $T_m$.

For completeness, we describe the prior over the tree structures we will use. We assume that each coordinate $x_j$ of the predictors has been scaled to lie in $[0, 1]$. This can be done, for example, by applying the empirical quantile transform to a subset of the observed values for each predictor and then interpolating the remaining values. A tree $T_m$ can be sampled from the prior as follows:

1. Initialize $T_m$ with an single node of depth $D_m = 0$.
2. For all nodes of depth $D_m$, make that node a branch node, with a left and right child of depth $D_m + 1$, with probability $\alpha(1 + D_m)^{-\beta}$; otherwise, make the node a leaf node.
3. For all branch nodes $b$ of depth $D_m$, sample the splitting coordinate $j_b \sim \text{Categorical}(s)$ and a splitting point $C_b \sim \text{Uniform}(L_{jb}, U_{jb})$, where $\prod_{b \in A_m} I(j_b, U_j)$ denotes the hyperrectangle of $x$-values which lead to node $b$.
4. If all nodes of depth $D_m$ are leaf nodes, terminate; otherwise, set $D_m \leftarrow D_m + 1$ and return to Step 2.

The distribution of the splitting coordinate $j_b \sim \text{Categorical}(s)$ determines how relevant a-priori we expect each predictor to be; for example, if $s_1 = 0.99$ we expect most splitting rules to use $x_1$, whereas if $s_1 = 10^{-10}$ we expect none of the splitting rules to use $x_1$. Linero (2018) took advantage of this fact to perform variable selection for BART models by using a sparsity-inducing Dirichlet prior $s \sim \text{Dirichlet}(a/P, \ldots, a/P)$ for some $a \ll P$. We also use this prior for the splitting proportion, which will allow us to perform automatic relevance determination in the density regression setting. This prior is crucial for proving that the posterior adapts to the presence of irrelevant predictors.

2.2. The Soft BART Density Sampler

Our modeling strategy is based on representation (1), where $\Phi(\mu)$ is a continuous, nonnegative, monotonically increasing link function. Such an $r(y, x)$ is guaranteed to exist whenever

---

**Figure 1.** Schematic which shows how a decision tree induces a regression function. Associated to the decision tree is a partition of $[0, 1]^2$ with the function $g(x; T, M)$ returning $\mu_1$, $\mu_2$, or $\mu_3$. 

Why BART? At the point of writing down model (1), we might have chosen a different specification for \( r(y, x) \), such as a Gaussian process. We find BART or BART with targeted smoothing to be an attractive model for the following reasons. First, in our experience, algorithms based on BART priors are surprisingly robust; we made little effort to tune the hyperparameters when designing our default prior, and have not yet run into any problems. Second, by using targeted smoothing, we ensure both that the estimated densities are smooth (this is true for both BART and SBART) and that we can control the interactions of \( x \) with \( y \) in a more transparent fashion. Third, as argued by Ročková and van der Pas (2020) and Linero and Yang (2018), BART shrinks our fitted model toward a parsimonious structure consisting of low-order interactions, which are common in practice. Fourth, using the Dirichlet prior described in Section 2.4, it is particularly convenient for implementing sparsity. Lastly, computations for the BART and SBART models are straight-forward and scale favorably with the sample size; additionally, while we have not pursued this here, there is scope for very fast implementations of our model based on BART without using soft decision trees, and we see no obvious reason why such an implementation would be much slower than existing BART software for fitting semiparametric regression.

2.3. Targeted Smoothing via Random Basis Function Expansions

We use the “targeted smoothing” approach of Starling et al. (2020) to overcome the problems of the naive SBART-DS prior. They set \( r(y, x) = \gamma + \sum_{m=1}^{M} B_m(y) g(x; T_m, M_m) \) where each leaf node is associated with a Gaussian process; that is, for fixed \( x \), we have \( g(\cdot; T_m, M_m) \sim GP[0, \Sigma(\cdot, \cdot)] \) where \( GP[0, \Sigma(\cdot, \cdot)] \) denotes a mean-0 Gaussian process with covariance function \( \Sigma(y, y') \).

Starling et al. (2020) consider the case where the number of unique values \( y \) takes, \( N_y \), is small. When \( N_y \) is large this is not practical due to the need to store and invert an \( N_y \times N_y \) matrix for all \( m \) trees. For SBART-DS we cannot guarantee that this is the case. Instead, we set \( r(y, x) = \gamma + \sum_{m=1}^{M} B_m(y) g(x; T_m, M_m) \) where each \( B_m \) is a random basis function.

To construct our approximation, consider the case where \( T_m \) is a nonsmooth decision tree. For fixed \( x \) we can write \( r(y, x) = \gamma + M^{-1/2} \sum_{m=1}^{M} \mu_m B_m(y) \) where the \( \mu_m \)'s are iid Normal(0, \( \sigma^2_d/M \)) random variables. Under mild regularity conditions on the distribution of the \( B_m \)'s, as \( M \to \infty \) a functional central limit theorem will hold and this will converge weakly to a Gaussian process with mean \( \gamma \) and covariance function \( \Sigma(y, y') = \sigma_d^2 K_{\nu}(\beta s_{\nu}) \). This is the same as the distribution of \( r(\cdot, x) \) used by Starling et al. (2020); for fixed \( x \), this approximation is likely to be accurate for the values of \( M \) we use due to the fact that \( y \) is one-dimensional. Rather than directly choose the basis functions \( B_m \), we specify \( \Sigma(y, y') \) and derive a distribution for \( B_m \) which matches. We make use of the following proposition, which follows from Bochner’s Theorem, and is stated for completeness.

**Proposition 1.** Let \( \Sigma(y, y') = \sigma^2_d \delta(y-y') \) be a shift-invariant kernel with \( \delta(0) = 1 \). Then there exists a distribution \( P(d\omega) \) such that \( \Sigma(y, y') = \sigma^2_d \mathbb{E}[2 \cos(\omega y + b) \cos(\omega y' + b)] \) where \( \omega \sim P(d\omega) \) and \( b \sim \text{Uniform}(0, 2\pi) \). Moreover, \( \delta(\cdot) \) is the characteristic function of \( P(d\omega) \), that is, \( \delta(t) = \int e^{it\omega} P(d\omega) \).

The approach of using random Fourier features in this fashion was popularized by Rahimi and Recht (2008). It follows from Proposition 1 that we can take \( B_m(y) = \sqrt{2 \cos(\omega_m y + b_m)} \) where \( \omega_m \sim P(d\omega) \) and \( b_m \sim \text{Uniform}(0, 2\pi) \). We list some possible choices below.

- Setting \( \omega_m \sim \text{Normal}(0, \rho^{-2}) \) gives the squared exponential covariance \( \Sigma(y, y') = \sigma^2_d \exp(-(y-y')^2/(2\rho^2)) \).
- Setting \( \omega_m \sim t_1(0, \rho^{-1}) \) gives \( \Sigma(y, y') = \frac{\rho^2}{\pi} \frac{1}{K_{\nu}(\sqrt{\rho^2-y^2})} \) where \( K_{\nu}(\cdot) \) is a modified Bessel function of the second kind. The exponential kernel \( \Sigma(y, y') = \sigma^2_d \exp(-|y-y'|/\rho) \) is a special case (\( \nu = 1 \)).
Other choices of random basis functions are also possible. For example, a referee suggested the use of local kernels as basis functions, with $B_{m}(y; s, \omega) = s^{-1}k \left( \frac{y - \omega}{s} \right)$ where $k$ is (say) a Gaussian kernel. While these basis functions make it more difficult to control the induced covariance function, the localization of the kernel may be more desirable than the oscillatory behavior of the Fourier basis functions. These basis functions are also easier to analyze theoretically, as the use of kernel functions dovetails nicely with the kernel-based theory used to establish results for SBART.

### 2.4. Default Prior Specification

In our illustrations we use the following default prior specification. Following Chipman, George, and McCulloch (2010), we fix the parameters $\alpha = 0.95$ and $\beta = 2$ in the prior for $\pi_{\mathcal{T}}$ and set $\mu_{\text{ef}} \sim \text{Normal}(\sigma_{\mu}^{2}/M)$. We fix $M = 50$; in general, we recommend trying multiple values of $M$. We set $\sigma_{\mu} \sim \text{Half-Cauchy}(0,1.5)$ to learn an appropriate value of $\sigma_{\mu}$ from the data. By having mass near $\sigma_{\mu} = 0$, this also allows us to revert to the base model $h(y \mid x, \theta)$. To induce further shrinkage to the base model, we set $\gamma \sim \text{Normal}(1,1)$. This has the additional benefit of shrinking toward models with $\Phi(r(y,x)) \approx 1$, which reduces the number of latent variables used when fitting the model by MCMC (see Section 2.5). We use tree-specific bandwidths $\tau_{m}$ which are exponentially distributed with mean 0.1. To perform variable selection we specify $s \sim \text{Dirichlet}(a/P, \ldots, a/P)$ and use a hyperprior $a/(a + P) \sim \text{Beta}(0.5, 1)$. For targeted smoothing, we approximate the squared exponential kernel by setting $\rho_{m} \sim \text{Normal}(0, \rho^{-2})$. We set $\rho^{2} \sim \text{Gamma}(\alpha_{\rho}, \beta_{\rho})$ to allow the length scale to be learned from the data. As a default, we set $\alpha_{\rho} = 1$ and $\beta_{\rho} = \pi^{2}/4$ after scaling the $Y_{i}$’s to have unit variance. This choice is based on the number of times a Gaussian process with length-scale $\rho$ is expected to cross 0 on the interval $(-1, 1)$: by Rice’s formula, the expected number of crossings is $2/(\pi \rho)$ so that if $\rho^{2} = 4/\pi^{2}$ the expected number of crossings is 1 (Adler, Taylor, and Worsley 2015, Exercise 2.8.23). Smaller values of $\rho^{2}$ correspond to more wiggly functions. Because our prior has positive density at 0, setting $\alpha_{\rho} = 1$ allows for the possibility that the function is very wiggly while defaulting to the prior belief that it is not.

In this article, we consider two possible choices for the base model. The first is the Gaussian linear model with $h(y \mid x, \theta) = \text{Normal}(y \mid a_{\theta} + x^{T}\beta_{\rho}, \sigma_{\theta})$ where $\text{Normal}(y \mid \mu, \sigma)$ is the density of a Gaussian random variable with mean $\mu$ and variance $\sigma^{2}$. We use this model with the MEPS dataset. We recommend a weakly-informative prior such as $\sim \text{Normal}(0,5I)$ and $\sigma_{\theta} \sim \text{Half-Cauchy}(0,1)$ after the response and predictors. We can also shrink toward a semiparametric Gaussian model by setting $h(y \mid x, \theta) = \text{Normal}(y \mid r_{\theta}(x), \sigma_{\theta})$ where $r_{\theta}(x) \sim \text{SBART}(\pi_{T}, \pi_{M})$ and the default prior of Linero and Yang (2018) is specified for $(\pi_{T}, \pi_{M})$. While the SBART base model does not add any flexibility beyond SBART-DS, it can result in faster computations if the true data generating process has Gaussian errors but a nonlinear mean. By modeling the nonlinearity in the base model, fewer augmented data points are required by our MCMC algorithm. We use the SBART base model in the simulation study of Section 4.1.

### 2.5. Posterior Computation

We use a two-layer data augmentation scheme which removes both the intractable integral in the denominator of (1) and the link function $\Phi(\mu)$ from the likelihood. Our approach is based on the following method for sampling from $f(y \mid x)$.

**Proposition 2.** Suppose that we sample $Y_{1}, Y_{2}, Y_{3}, \ldots \sim h(y \mid x, \theta)$ and independently sample $A_{ij} \sim \text{Bernoulli}[\Phi(r_{ij}(x))]$. Let $Z$ denote the $Y_{i}$ associated with the smallest index $J + 1$ for which $A_{J+1} = 1$. Then conditional on $(J, A_{J} : 1 \leq j \leq J + 1)$, $Z$ is a draw from $f(y \mid x) \propto h(y \mid x, \theta)\Phi(r_{y,x})$ and $Y_{1}, \ldots, Y_{J}$ are draws from $f(y \mid x) \propto h(y \mid x, \theta)(1 - \Phi(r_{y,x}))$.

We make use of Proposition 2 by augmenting the latent index $J$ and the sequence of rejected points. Associated to each observation $Y_{i} = Y_{i\theta}$ we independently sample $Y_{ij} \sim \text{Normal}(X_{i}, \theta, \sigma_{\theta})$ and $A_{ij} \sim \text{Bernoulli}[\Phi(r_{ij}(x))]$ until we reach the first iteration $J_{1} + 1$ such that $A_{i(J_{1}+1)} = 1$. We then work with the augmented state $\{Y_{ij} : 1 \leq i \leq N, 0 \leq j \leq J_{1}\}$, which has likelihood

$$
\prod_{i = 1}^{N} \prod_{j = 0}^{J_{i}} h(Y_{ij} \mid X_{i}, \theta) \times \prod_{i = 1}^{N} \left( \Phi(r_{Yi\theta,X_{i}}) \prod_{j = 1}^{J_{i}} (1 - \Phi(r_{Yij,X_{i}})) \right).
$$

For more details on the derivation of this expression, see Rao, Lin, and Dunson (2016), who consider the GP-DS model. At this stage Rao, Lin, and Dunson (2016) propose the use of Hamiltonian Monte Carlo to sample from the posterior distribution. This is not an option for us, as the $\mathcal{T}_{m}$’s are discrete parameters.

From here, one can follow Chipman, George, and McCulloch (2010) and apply Bayesian backfitting after data augmentation (Albert and Chib 1993): for example, for the probit link, we can augment $Z_{ij} \sim \text{Normal}(r_{Yij,X_{i}}, 1)$ truncated to $(-\infty, 0)$ for $j > 0$ and to $(0, \infty)$ for $j = 0$. A detailed description of this approach and associated algorithms for the probit, logit, and Student’s $t_{m}$ links are given in the supplementary materials.

### 3. Theoretical Results

#### 3.1. Rates of Convergence

We show that SBART-DS attains close to the minimax-optimal concentration rate for $(P + 1)$-dimensional functions $r(y, x)$ in the high-dimensional sparse setting. All proofs are deferred to the supplementary materials. We consider the case where $r(y, x)$ depends on only $D$ coordinates of $(y, x)^{T}$ where the relevant coordinates are unknown and must be learned from the data. Following Pati, Dunson, and Tokdar (2013) we study concentration with respect to the integrated Hellinger distance. Let $H(f, f_{0}) = [\int (\sqrt{f_{0}}(y \mid x) - \sqrt{f(y \mid x)})^{2} dy F_{X}(dx)]^{1/2}$ denote
the $F_X$-integrated Hellinger distance between $f_0(y \mid x)$ and $f(y \mid x)$. The covariates $X_i$ are assumed to be iid from $F_X$, which is not assumed to be known. We similarly define an $F_X$-integrated Kulback–Leibler neighborhood. Define $K(f_0, f) = \int f_0 \log \frac{f_0}{f} dy dF_X$ and $V(f_0, f) = \int f_0 (\log \frac{f_0}{f})^2 dy dF_X$. Then the integrated Kulback–Leibler neighborhood is given by $K(\epsilon) = \{ f : K(f_0, f) \leq \epsilon^2 \}$ and $V(f_0, f) \leq \epsilon^2 \}$. Let $D_n$ denote the number of predictors $X_i, Y_i : i = 1, \ldots, n$ and let $\Pi$ denote a prior distribution on $r$ and additional hyperparameters. We say that the posterior has a convergence rate of at least $\epsilon_n$ if there exists a constant $C > 0$ such that $\Pi[H(f_0, f_0) \geq C\epsilon_n \mid D_n] \rightarrow 0$ in probability.

To simplify the theoretical results, we assume that $\tau_1 > 0$, that is, $\Pi(\{y : \tau_1 > 0\} \mid D_n) \rightarrow 0$ in probability. The true conditional density is bounded and bounded away from 0.

Condition F (on $F_0$): The true conditional density $f_0(y \mid x)$ can be written as $f_0(y \mid x)$ for some $r_0 \in C^{\alpha}([0, 1]^{D+1})$ where $C^{\alpha}([0, 1]^{D+1})$ is the ball of radius $R$ in the space of $\alpha$-Hölder smooth functions on $[0, 1]^{D+1}$, where $f_0(y \mid x)$ is defined as $\Phi(y \mid x)$ for some density $h(y)$ on $[0, 1]$. Additionally, we can write $r_0(y, x) = \nu(y, \tau_1)$ where $\nu_\infty = \nu_\infty$ and $S$ is a subset of $\{1, \ldots, P\}$ of cardinality $D - 1$. That is, $r_0(y, x)$ depends on at most $D$ coordinates of $(y, x)^T$. The number of predictors $P = P_n$ depends on $n$ but is such that $\log(P + 1) \leq C_n n^\eta$ for some $\eta (0, 1)$.

Remark 1. For simplicity, we consider $\tau$ and $S$ to be independent of $n$; in particular, we do not consider $D$ diverging with $n$. As noted in Section 2, there will exist some $r_0$ such that $f_0 = f_0$ provided that $R(y, x) = f(y \mid x)$ is nonzero and $sup_{y} R(y, x) < \infty$ for all $x$. When $h(y) = 1, r_0$ being continuous on $[0, 1]^{P+1}$ implies that $C^{-1} \leq f_0(y \mid x) \leq C$ for some constant $C > 0$, that is, $f_0(y \mid x)$ is bounded and bounded away from 0.

Condition L (on $\Phi$): The link function $\Phi(\mu)$ is strictly increasing and is the cumulative distribution function of a random variable $Z$ which is symmetric about 0 and has density $\phi(\mu)$ satisfying $\phi(\mu)/\Phi(\mu) \leq K$ for all $\mu$ and some constant $K$.

Remark 2. We show in the supplementary materials that Condition L holds for the logit ($K = 1$) and $t_\nu (K = \sqrt{\nu})$ links, but fails for the probit link.

Condition P on $\Pi$: The function $r$ is given an SBART($\pi_T, \pi_M$) prior with $M$ trees conditional on ($\pi_T, \pi_M, M$). Additionally, the prior $\Pi$ satisfies the following conditions:

(P1) There exists positive constants ($C_M1, C_M2$) such that the prior on the number of trees $M$ in the ensemble is $\Pi(M = t) = C_M1 \exp(-C_M2 t \log t)$.

(P2) A single bandwidth $r_m \equiv r$ is used and its prior satisfies $\Pi(\tau \geq \tau) \leq C_{r1} \exp(-x^{r_{12}})$ and $\Pi(\tau^{-1} \geq \tau) \leq C_{r3} \exp(-x^{r_{14}})$ for some positive constants $C_{r1}, \ldots, C_{r4}$ for all sufficiently large $x$, with $C_{r2}, C_{r4} < 1$. Moreover, the density of $\tau^{-1}$ satisfies $\tau_{\pi_1}(x) \geq C_{r5} \exp(-C_{r6} x)$ for large enough $x$ and some positive constants $C_{r5}$ and $C_{r6}$.

(P3) The prior on the splitting proportions is $\sim$ Dirichlet $(a/P^\xi, \ldots, a/P^\xi)$ for some $\xi > 1$ and $a > 0$. The function $\psi$ is a probability density function on $[0, 1]$. Moreover, $\psi$ is such that for all sufficiently large $x$, $\psi(x) \rightarrow 0$ in probability, where $\epsilon_n$ is the $\epsilon$-covering number of $\calF$ with respect to $H$. Suppose that there exist positive constants $C, C_M$ such that for all sufficiently large $n$ there exist sets of conditional densities $\calF_n$ satisfying the following conditions:

(a) Entropy Bound: $\log N(\epsilon_n, \calF_n, H) \leq C_n n \epsilon_n^2$.
(b) Support Condition: $\Pi(\calF_n^0) \leq \exp(-C \epsilon_n^2)$.
(c) Prior Thickness: $\Pi\{f \in K(\epsilon_n)\} \geq \exp(-C \epsilon_n^2)$.

Then $\Pi[H(f_0, f) \geq A\epsilon_n \mid D_n] \rightarrow 0$ in probability for some constant $A > 0$.

We prove Theorem 1 by checking (a)-(c) in Proposition 3, which are analogous to conditions of Ghosal, Ghosh, and van der Vaart (2000).
More generally, one expects that Theorem 1 can be improved to allow for additive decompositions $r_n(x, y) = \sum_{j=1}^{J} r_j(x, y)$ where the $r_j$’s are functions which are $D_j$-sparse and $\alpha_j$-H"{o}lder continuous. Results in this framework (Yang and Tokdar 2015; Linero and Yang 2018; Ro"{c}ková and van der Pas 2020) suggest that we should be able to obtain a rate $\epsilon_n = \sum_{j=1}^{J} h_{\alpha_j/(2\alpha_j+1)}(\log(n))^{1/2} + \sqrt{n^{1-D_j}\log(P+1)}$, which is a substantial improvement on Theorem 1. One difficulty with extending these results is that Condition P2 only allows a single bandwidth, while different $r_j$’s will be optimal for different $\alpha_j$’s. Unlike the nonparametric regression setting, however, it is unclear how one would interpret the additive assumption for SBART-DS. We leave examining the additive framework to future work.

3.2. Topological Support

Condition F implies the restrictive condition $\sup_x |\log f_0(y | x)/h(y)| < \infty$ for all $x$. For example, when $h(y)$ is a uniform distribution, this implies that $C^{-1} \leq f_0(y | x) \leq C$ for some constant $C$, prohibiting for example the Beta($\alpha, \alpha$) distribution for all $\alpha \neq 1$.

Given the restrictiveness of Condition F, it is desirable to understand which densities can be estimated consistently (at any rate). This is intimately connected with the Kullback–Leibler (KL) support $K(\Pi) = \{f_0 : \Pi(K(f_0, f_0) < \epsilon > 0 \text{ for all } \epsilon > 0\}$ (Ghosal, Ghosh, and Ramamoorthi 1999). Roughly speaking, $f_0$ will be consistently estimable if (i) $f_0 \in K(\Pi)$ and (ii) there exists a sieve satisfying the conditions of Proposition 3 for some sequence $\epsilon_n \downarrow 0$ with $n\epsilon_n \rightarrow \infty$. As (ii) is established in the proof of Theorem 1, it suffices to characterize $K(\Pi)$.

In the supplementary materials we establish the following result, which shows that $f_0 \in K(\Pi)$ for a substantially wider class of densities than suggested by Condition F. In particular, the KL support contains all uniformly bounded smooth densities, and a very large class of unbounded densities as well.

**Theorem 2.** Suppose that $f_0(y | x) = \alpha$-H"{o}lder smooth on $(0, 1) \times [0, 1]^p$ for some $\alpha > 0$. Suppose that there exists a constant $B$ such that, for every $\delta \in (0, 1)$, there exists an $a \leq \delta$ and $b \geq 1 - \delta$ such that

\[
\begin{align*}
(C1) & \sup_x a f_0(a | x) \leq \delta \text{ and } \sup_x (1 - b) f_0(b | x) \leq \delta; \\
(C2) & \sup_x |f_0(a | x) - \inf_{y \leq a} f_0(y | x)| \leq B \text{ and } \sup_x [f_0(b | x) - \inf_{y \geq b} f_0(y | x)] \leq B; \text{ and} \\
(C3) & \int |f_0(y | x)| \log f_0(y | x) \, dy F_X(dx) < \infty.
\end{align*}
\]

Then $f_0 \in K(\Pi)$ if $\Pi$ satisfies Condition L and Condition P with $h(y) \equiv 1$.

**Remark 4.** Conditions C1, C2, and C3 are all mild. For example, all bounded continuous conditional densities on $(0, 1) \times [0, 1]^p$ can be shown to satisfy C1 and C2. A large class of unbounded conditional densities can also be shown to satisfy C1 and C2; for example, we show in the supplementary materials that the conditional density Beta$(y | \alpha(x), \beta(x))$ satisfies C1, C2, and C3 for all continuous choices of $\log \alpha(x)$ and $\log \beta(x)$.

3.3. Beyond the Naive Model

For simplicity, Sections 3.1 and 3.2 consider a naive SBART-DS model rather than our targeted smoothing variant. Beyond the need to establish analogs of Theorem 2 and Lemma S.1 of Linero and Yang (2018) to accommodate targeted smoothing, there is no fundamental theoretical difference between the naive and targeted smoothing variants of SBART-DS. To illustrate this, we prove the following result in the supplementary materials. This result takes the random basis functions to be a collection of logistic kernels $B_m(y) = \exp(y-c_m)/[1+\exp(y-c_m)]^2$ with $c_m \sim \text{Uniform}(0,1)$ and a prior on $\sigma$ satisfying the same restrictions that Condition P2 places on $\tau$.

**Theorem 3.** Suppose that Condition F, Condition L, and Condition P’ (in the supplementary materials) hold. Then the conclusion of Theorem 1 holds.

4. Illustrations

4.1. A Simple Simulation Illustration

We now assess the performance of the SBART-DS using the simulation example described by Dunson, Pillai, and Park (2007). The response $Y_i$ is sampled from a mixture model

$$Y_i \sim e^{-2x} \text{Normal}(x, 0.3^2) + (1 - e^{-2x}) \text{Normal}(x^4, 0.2^2)$$

given $X_{i1} = x$.

We set $N = 500$ and have $P - 1$ additional predictors which do not influence the response. The marginal density of the $X_i$’s is uniform on $[0, 1]^P$. For SBART-DS we use the default prior with $M \equiv 50$ and the probit link. We do not make any attempt to tune the hyperparameters $(\alpha, \sigma, \rho, \alpha, \beta, \gamma)$ beyond this. We take the base model to be the normal linear regression model $h(y | x, \theta) = \text{Normal}(y | \alpha + \beta^T x, \sigma^2)$.

We consider moderate dimensions $P$ for illustrative purposes, but in higher dimensions one might wish to induce sparsity $	heta_0$.

Figure 2 displays a scatterplot of the relationship between $Y_i$ and $X_{i1}$ as well as the posterior mean and credible band for the function $r(x) = \mathbb{E}(Y_i | X_i = x)$ with $P = 5$. We compare SBART-DS to a Dirichlet process mixture model described by Jara et al. (2011) as implemented in the function DPdensity in the DPPackage package in R. This model uses the joint specification $(X_1, Y_1) \sim \text{DP}(\alpha G_0)$ where $G \sim \text{DP}(\alpha G_0)$ is a Dirichlet process with a normal-inverse-Wishart base measure $G_0 \equiv \text{Normal}(\mu, \Sigma) \text{IW}(\Sigma | \nu, \Psi)$. The conditional density of $(Y_i | X_i = x)$ can be estimated from an infinite mixture model as $f(y | x) = \sum_{k=1}^{\infty} \omega_k(x) \text{Normal}(y | \mu_{y|x}, \Sigma_{y|x})$, where $\omega_k(x) \propto \pi_k \text{Normal}(x | \mu_x, \Sigma_{x|x})$, $\mu_{y|x} = \mu_x + \Sigma_{x|x}^{-1}(x - \mu_x)$, and $\Sigma_{y|x} = \Sigma_{y|x} - \Sigma_{y|x} \Sigma_{x|x}^{-1} \Sigma_{y|x}$.

We use the same prior specification as Jara et al. (2011) but with a larger value of $\nu$ to accommodate the fact that $\nu > P - 1$ is required.

Figure 3 shows the fitted density for several fixed values of $X_{i1}$ with all other predictors frozen at the value $X_{ij} = 0.5$ (as these predictors were correctly filtered out of the model, their particular value is irrelevant). To compute the density, we evaluated the numerator $h(y | x, \theta) \Phi(r(y, x))$ on a grid of $y$ values and applied the trapezoidal rule with these evaluations.

5. Extensions

5.1. Flexible Nonparametric Regression

The SBART-DS framework can be viewed as a nonparametric regression model, with $\theta_0$ representing the true mean function. Here we discuss an extension that allows for nonparametric regression.”
to approximate the denominator $\int h(y \mid x, \theta) \Phi \{ r(y, x) \} \, dy$. We see SBART-DS successfully captures variability in the location, shape, and scale of the densities, and produces 95% credible bands which accurately account for uncertainty in the estimates. SBART-DS also captures the mean response accurately (left panel of Figure 2).

### 4.2. Comparison to Competing Methods

We now conduct an in-depth simulation experiment to assess the merits of SBART-DS relative to competing methods. In this simulation we took the baseline model $h(y \mid x, \theta)$ to itself correspond to an SBART model with the default prior described by Linero and Yang (2018).

**Simulation Settings** We consider a variety of different ground truths for comparison. All settings take $X_i \sim \text{Uniform}(0, 1)^p$, except for the ZK setting which has correlated predictors.

- **KD**: We use a modified version of the simulation experiment of Kundu and Dunson (2014). Specifically, we take $Y_i \sim \text{Normal}(3e^{-X_i}, Z_i + 0.05)$ where $Z_i = \sqrt{X_{i1}X_{i2}}$.

---

**Figure 2.** (Left) Plot of realized values of $X_1$ against $Y$ for a single replication of the experiment, with solid black line indicating the true mean, light blue line indicating the estimated posterior mean, and the dashed green lines indicating 95% credible bands for the mean function. (Right) The posterior distribution of the number of rejected points estimated via Markov chain Monte Carlo.

**Figure 3.** Posterior mean (blue), 95% credible bands for the density (dashed black) and true density function (green) for the simulated data, for the values $X_1 \in \{0.14, 0.28, 0.42, 0.58, 0.72, 0.86\}$.
SN: We set \( Y_i \mid X_i = x \sim \text{Skew-Normal} (\xi(x), \omega(x), \alpha(x)) \) where \( \text{Skew-Normal}(\xi, \omega, \alpha) \) is the skewed normal distribution with location \( \xi \), scale \( \omega \), and slant parameter \( \alpha \) (Azzalini 2013, sec. 2.1). We take \( \omega(x) \equiv 2 \), \( \xi(x) = 10 \sin(\pi x_1 x_2) + 20(x_3 - 0.5)^2 + 10 x_4 + 5 x_5 \) to correspond to Friedman’s test function (Friedman 1991; Chipman, George, and McCulloch 2010). We set \( \alpha(x) = (\xi(x) - 14.7)/10.0 \), which standardizes \( \alpha(X_i) \) to have mean 0 and unit variance. This choice of \( \alpha(x) \) allows for both positive and negative skewness.

DP: The simulation setting from Dunson, Pillai, and Park (2007) considered in Section 4.1.

ZK: We modify the simulation experiment of Zhu and Kosorok (2012), taking \( \{e^{X_i^T} \mid X_i = x \sim \Gamma(\alpha(x), \beta(x)) \) with \( \beta(x) \equiv 1 \) and \( \alpha(x) = 0.5 + 0.3 \sum_{k=1}^{5} |x_k| \). We take \( X_i \sim \text{Normal}(0, \Sigma) \) with \( \Sigma_{jk} = \rho^{|i-j|} \) with \( \rho = 0.75 \) to assess whether SBART-DS is robust to correlation structure in the predictors.

We consider differing values of \( P \) to assess robustness to the number of covariates, with \( P \in \{5, 20, 100\} \). All cases use \( N = 500 \).

Competing Methods We compare SBART-DS to the probit stick-breaking (PSBP) approach of Chung and Dunson (2009), the joint Dirichlet process mixture model (Joint-DP) described by Jara et al. (2011), and a random forest conditional density estimation (RFCDE) algorithm proposed by Pospisil and Lee (2018) which constructs a conditional density estimate using random forests (Breiman 2001). All methods use default hyper-prior/hyperparameter settings. PSBP is implemented using Matlab code available at github.com/david-dunson/probit-stick-breaking, Joint-DP is implemented with the DPpackage density function in DPpackage in R, and RFCDE is implemented using the RFCDE package in R.

Evaluation Criteria Methods are compared according to the following criteria.

- **TV**: The integrated total variation distance of the fitted density from the truth, \( TV(f_0, \hat{f}) = \int \int f_0(y \mid x) - \hat{f}(y \mid x) \cdot dy \cdot F_X(dx) \). This integral is approximated numerically as \( N^{*-1} \sum_{i=1}^{N^*} |f_0(y \mid X_i^*) - \hat{f}(y \mid X_i^*)| \cdot dy \), where \( \{X_i^* \}_{i=1}^{N^*} \) consists of heldout covariates sampled from the model and the \( dy \) integral is approximated numerically. We also consider \( nTV \), a normalized version of TV such that the best performing method has an average TV of 1.

- **Coverage**: For PSBP and SBART-DS, we examine the average coverage of posterior credible intervals for the quantiles \( Q_{0.25}(X_i) \), \( Q_{0.5}(X_i) \), and \( Q_{0.75}(X_i) \), where \( Q_{\alpha}(x) \) denotes the \( \alpha \)th quantile of \( Y_i \) given \( X_i = x \). Quantiles are computed numerically from the density estimates at each MCMC iteration. Credible intervals are not available for RFCDE, and are not readily accessible from the output of DPpackage.

- **RMSE**: For \( \alpha \in \{0.25, 0.5, 0.75\} \) we examine the root mean-squared error in estimating the quantiles \( Q_{\alpha}(x) \), RMSE\( _\alpha = \left\{ \int |Q_{\alpha}(x) - \hat{Q}_{\alpha}(x)|^2 F_X(dx) \right\}^{1/2} \). For PSBP and SBART-DS, we take \( \hat{Q}_{\alpha}(x) \) to be the posterior mean of \( Q_{\alpha}(x) \), while Joint- DP and RFCDE construct \( \hat{Q}_{\alpha}(x) \) from the point estimator of the conditional densities.

**Other Details** The simulation was replicated 200 times for each \( P \), setting, and method. Results for \( P = 100 \) are missing for Joint-DP and PSBP; Joint-DP was impractically slow for this large \( P \) while the PSBP software produced errors. As the software for Joint-DP was quite slow, we performed less than 200 simulations, and replicated until its performance relative to the other approaches was established.

**Results** We present partial results here, with complete results deferred to the supplementary materials. Figure 4 gives the TV and coverage for each method, setting and \( P \), and the average coverage of credible intervals for \( Q_{0.75}(X_i^*) \). For all settings and \( P \), the best performing method by TV is SBART-DS by a considerable margin, with the single exception of \( DP \) where PSBP performs best; this is expected since the ground truth of \( DP \) is a mixture model. In terms of TV and RMSE in predicting the conditional quantiles, we also see that SBART-DS is highly robust to the inclusion of irrelevant predictors, with the performance varying negligibly as \( P \) is increased. SBART-DS also performs very well in terms of coverage, attaining close to the nominal coverage rate for all settings. By comparison, PSBP does not attain close to nominal coverage even on \( DP \).

Detailed results for SN are given in Table 1, with similar tables for ZK, KD, and DP given in the supplementary materials. Surprisingly, for most choices of \( \alpha \) SBART-DS outperforms PSBP on DP; otherwise, the overall trends we observe in Figure 4 are present.

Computationally, the RFCDE method was by far the fastest, taking roughly 4 sec to fit to the ZK \( (P = 5) \) setting. PSBP with 2000 MCMC iterations and our SBART-DS sampler with 10,000 MCMC iterations each took roughly 30 min to run; we remark that both approaches used unoptimized code, and could both likely be sped up considerably. The slowest procedure was the Joint-DP, which took roughly 2 hr to fit 10,000 iterations.

In the supplementary materials we also analyze the computational burden as a function of \( N \) on the ZK illustration with \( P = 20 \). Summarizing the results, we found that computation scales roughly linearly in \( N \), but that the posterior of the number of rejected points is not well-identified, with chains occasionally using more rejected points than strictly necessary. Mixing on identified components of the model did not seem to be associated to the poor identification of the number of rejected points.

### 4.3. Analysis of MEPS Data

We apply SBART-DS to data from the Medical Expenditure Panel Survey (MEPS) from the year 2015. MEPS is an ongoing survey in the United States which collects data on families/individuals, their medical providers, and employers, with a focus on the cost and use of health care.

There is a large literature which has considered the relationship between socioeconomic status, education, and obesity. Educational attainment relates to obesity in a complex fashion, with the effect modified by the overall income of a region, gender, and other factors (Cohen et al. 2013). We examine this...
relationship on a subset of MEPS consisting of responses from 1452 women aged between 25 and 35 years old, controlling for log-income (measured as a percentage of the poverty line), age, and race. Existing research predicts that higher educational attainment will be associated with lower obesity levels in this group.

We collected 55,000 samples from the posterior via MCMC, discarded the first 5000 iterations to burn-in, and saved every 10th sample for a total of 5000 saved samples. In the supplementary materials we examine the mixing of the chain using traceplots and find no evidence of poor mixing for the quantities of interest.

In Figure 5 we display the estimated density as the level of educational attainment is varied from less-than-high-school to graduate degree for white women with all other covariates frozen at their median value. We see that as educational attainment increases the bulk of the distribution remains concentrated near 0 (the overall mean level of BMI) but goes from roughly symmetric to being highly right-skewed. The nature of

Table 1. Simulation results for the SN setting.

| $P$  | Method     | $nTV$ | $Q_{0.25}$ | Coverage | RMSE | $Q_{0.5}$ | Coverage | RMSE | $Q_{0.75}$ | Coverage | RMSE |
|------|------------|-------|------------|----------|------|-----------|----------|------|------------|----------|------|
| 5    | Joint DP   | 2.05  | —          | 2.15     | —    | 1.10      | —        | —    | 1.32       | —        | —    |
|      | PSBP       | 3.71  | 0.37       | 2.55     | 0.33 | 2.43      | 0.27     | 2.85 |
|      | RFCDE      | 4.41  | —          | 3.87     | —    | 2.84      | —        | 3.03 |
|      | SBART-DS   | 1.00  | 0.97       | 0.53     | 0.98 | 0.52      | 0.97     | 0.53 |
| 20   | Joint DP   | 4.03  | —          | 3.12     | —    | 2.88      | —        | 3.09 |
|      | PSBP       | 3.73  | 0.36       | 2.52     | 0.31 | 2.38      | 0.23     | 2.85 |
|      | RFCDE      | 5.11  | —          | 4.32     | —    | 3.29      | —        | 4.46 |
|      | SBART-DS   | 1.00  | 0.97       | 0.54     | 0.98 | 0.52      | 0.98     | 0.53 |
| 100  | RFCDE      | 5.58  | —          | 5.71     | —    | 4.55      | —        | 5.84 |
|      | SBART-DS   | 1.00  | 0.98       | 0.54     | 0.98 | 0.52      | 0.98     | 0.53 |

Bold values denote the best performance among all methods.
Figure 5. Density estimates and 95% credible bands for \( f(y \mid x) \) for different educational levels for white women aged between 25 and 35, fixing log-income and age at their median values.

this relationship is that, while the modal value of BMI is fairly stable as education level changes, highly educated women are less likely to be highly obese.

Each predictor \( j \) is associated to two coefficients: the base model coefficient \( \beta_\theta j \) and the splitting proportion \( s_j \). The posterior median, density, and (66%, 95%)-credible intervals are given for each coefficient in Figure 6. Interestingly, education level is the only relevant predictor in the nonparametric component \( \Phi \{ r(y, x) \} \) so that the overall shape of the density is primarily determined by education. Intuitively, one might expect that education is only relevant through its indirect effect on income, however, our results suggest this is not the case. Log-income has a strong presence in the base model as well, while race and age have weaker effects.

We now use SBART-DS to perform statistical inference on several functionals of the conditional distribution which would not be available from existing BART models. We first examine how education affects the conditional quantiles of BMI. Let \( x \) and \( x' \) denote the values of the covariate which differ only in terms of the level of education, with all noneducation variables fixed at their respective medians and let \( \Delta_\alpha = Q_\alpha(Y_i \mid X_i = x) - Q_\alpha(Y_i \mid X_i = x') \) denote the difference in the \( \alpha \)th quantiles associated to \( f(y \mid x) \) and \( f(y \mid x') \). Figure 7 and Table 2 summarize the posterior of \( \Delta_\alpha \) for \( \alpha \in \{0.1, 0.25, 0.5, 0.75, 0.9\} \) for the comparison of (i) “Less than 9th Grade” a “Master’s/Doctorate/Professional Degree” and (ii) “Some College/Associate’s Degree” and “Bachelor’s Degree.” We see generally that the effects are estimated to be larger in the left tail of the distribution than the right, with (for example) the posterior 95% credible interval for \( \Delta_{0.1} \) including 0 for both comparisons. Generally we find that the gaps in BMI are larger as we move to larger quantiles. The effects are smaller but more precise for the Some College versus Bachelor’s comparison because these are two of the larger groups, whereas the < 9th Grade and Graduate Degree groups are the two smallest. The above analysis of the conditional quantiles illustrates the flexibility of our approach, and because we have specified a full model for the density we are guaranteed that the inferences we obtain are consistent across all quantiles.

The flexibility of SBART-DS permits us to also make inferences on more exotic functionals of \( f(y \mid x) \) than the quantiles. To illustrate, we perform inference on the functional \( P^* = \int_0^\infty \int_0^\infty f(y' \mid x) f(y' \mid x') \, dy' \, dy \) where \( x \) and \( x' \) differ only in terms of education level (all other covariates are set equal to their median value). In words, this corresponds to the probability that a randomly selected individual with \( [X_i = x] \) has a BMI which is 10% lower than a randomly selected individual with \( [X_i = x'] \). Performing inference on \( P^* \) necessarily requires estimating the entire conditional distribution \( f(y \mid x) \), and is available from neither mean nor quantile regression. Posterior summaries for the comparison Graduate Degree versus < 9th Grade and the comparison Bachelor’s Degree versus Some College, as well as their reverse comparisons, are given in Figure 8 and Table 2. We see that increasing the level of education generally corresponds
Figure 6. Left: Posterior medians, (66%, 95%)-credible intervals, and density estimates for the regression coefficients of the base model $\beta_0$. Right: posterior medians, credible intervals, and density estimates for the splitting proportions $s_j$ for each predictor.

Figure 7. The posterior distribution of the difference-in-quantiles going from (left) less than ninth grade education to a graduate degree and (right) some college education to a Bachelor’s degree.

Table 2. Posterior summaries for $P^*$ and $\Delta y$.

| Quantity | Mean | LCL | UCL | Mean | LCL | UCL |
|----------|------|-----|-----|------|-----|-----|
| $P^*$    | 0.59 | 0.49| 0.70| 0.46 | 0.42| 0.50|
| $\Delta_{0.1}$ | 1.78 | −1.08| 4.44| 0.46 | −0.07| 0.99|
| $\Delta_{0.25}$ | 3.38 | 1.06| 6.00| 0.91 | 0.35| 1.54|
| $\Delta_{0.5}$ | 5.28 | 2.78| 8.15| 1.86 | 0.85| 2.94|
| $\Delta_{0.75}$ | 6.76 | 3.35| 10.45| 3.22 | 1.31| 5.11|
| $\Delta_{0.9}$ | 6.68 | 1.19| 13.02| 2.74 | 0.86| 5.01|

NOTE: Mean denotes the posterior mean, while LCL and UCL denote the lower and upper limits of a 95% credible interval.

5. Discussion

In this article we proposed a new method for density regression based on Bayesian additive regression trees. SBART-DS is suitable for routine use—it has a simple default specification, strong theoretical properties, and can be fit using a tuning-parameter-free Gibbs sampling algorithm. On simulated data we illustrated how SBART-DS is robust to the presence of irrelevant variables, giving empirical support to the theoretical results supporting a faster posterior concentration rate when the rejection model $\Phi\{r(y, x)\}$ is sparse. Using data from MEPS we showed how SBART-DS can capture the effect of education level on the conditional distribution of body mass index.

The general strategy of defining a prior using a rejection sampling model can be used to extend that SBART-DS model to other domains. For example, in subsequent work we have extended this approach to survival analysis by modeling the hazard function $\lambda(y \mid x)$ as the hazard of a thinned Poisson
process $\lambda(y \mid x) = \lambda_0(y \mid x) \Phi(r(y,x))$ (Basak et al. 2021; Linero et al. 2021).

**Supplementary Materials**

In the Supplementary Material we provide proofs of all the results, give additional computational details and simulation results, and provide code for reproducing some of the results of our simulation study.

**Funding**

This work was partially supported by grant SES-2046896 from the National Science Foundation.

**References**

Adler, R. J., Taylor, J. E., and Worsley, K. J. (2015), “Applications of Random Fields and Geometry: Foundations and Case Studies,” Preprint. Available at https://web.stanford.edu/class/stats317/hrf.pdf. [2133]

Albert, J. H., and Chib, S. (1993), “Bayesian Analysis of Binary and Polychotomous Response Data,” *Journal of the American Statistical Association*, 88, 669–679. [2133]

Angrist, J., Chernozhukov, V., and Fernández-Val, I. (2006), “Quantile Regression Under Misspecification, with an Application to the US Wage Structure,” *Econometrica*, 74, 539–563. [2130]

Antoniano-Villalobos, I., Wade, S., and Walker, S. G. (2014), “A Bayesian Nonparametric Regression Model with Normalized Weights: A Study of Hippocampal Atrophy in Alzheimer’s Disease,” *Journal of the American Statistical Association*, 109, 477–490. [2130]

Azzalini, A. (2013). *The Skew-Normal and Related Families* (Vol. 3). Cambridge: Cambridge University Press. [2137]

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*. Advance online publication. [2141]

Breiman, L. (2001), “Random Forests,” *Machine Learning*, 45, 5–32. [2137]

Chipman, H. A., George, E. I., and McCulloch, R. E. (2010), “BART: Bayesian Additive Regression Trees,” *The Annals of Applied Statistics*, 4, 266–298. [2129,2131,2133,2134,2137]

Chung, Y., and Dunson, D. B. (2009), “Nonparametric Bayes Conditional Distribution Modeling with Variable Selection,” *Journal of the American Statistical Association*, 104, 1646–1660. [2129,2130,2137]

Cohen, A. K., Rai, M., Rehkopf, D. H., and Abrams, B. (2013), “Educational Attainment and Obesity: A Systematic Review,” *Obesity Reviews*, 14, 989–1005. [2137]

Dunson, D. B., and Park, J.-H. (2008), “Kernel Stick-Breaking Processes,” *Biometrika*, 95, 307–323. [2129,2130]

Dunson, D. B., Pillai, N., and Park, J. H. (2007), “Bayesian Density Regression,” *Journal of the Royal Statistical Society*, Series B, 69, 163–183. [2135,2137]

Friedman, J. H. (1991), “Multivariate Adaptive Regression Splines,” *The Annals of Statistics*, 19, 1–67. [2137]

Ghosal, S., Ghosh, J. K., and Ramamoorthi, R. V. (1999), “Posterior Consistency of Dirichlet Mixtures in Density Estimation,” *Annals of Statistics*, 27, 143–158. [2135]

Ghosal, S., Ghosh, J. K., and van der Vaart, A. W. (2000), “Convergence Rates of Posterior Distributions,” *Annals of Statistics*, 28, 500–531. [2134]

Hahn, P. R., and Carvalho, C. M. (2015). “Decoupling Shrinkage and Selection in Bayesian Linear Models: A Posterior Summary Perspective,” *Journal of the American Statistical Association*, 110, 435–448. [2130]

Jara, A., Hanson, T. E., Quintana, F. A., Müller, P., and Rosner, G. L. (2011), Dppackage: Bayesian Semi- and Nonparametric Modeling in R,” *Journal of Statistical Software*, 40, 1–30. [2135,2137]

Kundu, S., and Dunson, D. B. (2014), “Latent Factor Models for Density Estimation,” *Biometrika*, 101, 641–654. [2129,2130,2136]

Linero, A. R. (2018), “Bayesian Regression Trees for High-Dimensional Prediction and Variable Selection,” *Journal of the American Statistical Association*, 113, 626–636. [2131]

Linero, A. R., Basak, P., Li, Y., and Sinha, D. (2021), “Bayesian Survival Tree Ensembles with Submodel Shrinkage,” *Bayesian Analysis*. Advance online publication. [2141]

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. [2129,2131,2132,2133,2134,2135,2136]

MacEachern, S. N. (1999), “Dependent Nonparametric Processes,” in ASA *Proceedings of the Section on Bayesian Statistical Science*, pp. 50–55. [2129,2130]

Müller, P., Erkanli, A., and West, M. (1996), “Bayesian Curve Fitting Using Multivariate Normal Mixtures,” *Biometrika*, 83, 67–79. [2130]
Murray, I., MacKay, D., and Adams, R. P. (2009), “The Gaussian Process Density Sampler,” in *Advances in Neural Information Processing Systems* (Vol. 21), eds. D. Koller, D. Schuurmans, Y. Bengio, and L. Bottou, Curran Associates, Inc. [2129]

Pati, D., Dunson, D. B., and Tokdar, S. T. (2013), “Posterior Consistency in Conditional Distribution Estimation,” *Journal of Multivariate Analysis*, 116, 456–472. [2135]

Pospisil, T., and Lee, A. B. (2018), “RFCDE: Random Forests for Conditional Density Estimation,” arXiv preprint arXiv:1804.05753. [2137]

Pratola, M. T., Chipman, H. A., George, E. I., and McCulloch, R. E. (2020), “Heteroscedastic BART via Multiplicative Regression Trees,” *Journal of Computational and Graphical Statistics*, 29, 405–417. [2130]

Rahimi, A., and Recht, B. (2008), “Random Features for Large-Scale Kernel Machines,” in *Advances in Neural Information Processing Systems* (Vol. 20), eds. J. Platt, D. Koller, Y. Singer, and S. Roweis, Curran Associates, Inc. [2132]

Rao, V., Lin, L., and Dunson, D. B. (2016), “Data Augmentation for Models Based on Rejection Sampling,” *Biometrika*, 103, 319–335. [2129,2133]

Ravikumar, P., Liu, H., Lafferty, J., and Wasserman, L. (2007), “SPAM: Sparse Additive Models,” in *Proceedings of the 20th International Conference on Neural Information Processing Systems*, pp. 1201–1208. [2132]

Reich, R. J., Fuentes, M., and Dunson, D. B. (2011), “Bayesian Spatial Quantile Regression,” *Journal of the American Statistical Association*, 106, 6–20. [2130]

Riihimäki, J., Vehtari, A. (2014), “Laplace Approximation for Logistic Gaussian Process Density Estimation and Regression,” *Bayesian Analysis*, 9, 425–448. [2130]

Rocková, V., and van der Pas, S. (2020), “Posterior Concentration for Bayesian Regression Trees and Forests,” *The Annals of Statistics*, 48, 2108–2131. [2132,2134,2135]

Rodriguez, A., and Dunson, D. B. (2011), “Nonparametric Bayesian Models Through Probit Stick-Breaking Processes,” *Bayesian Analysis*, 6, 145–177. [2129,2130]

Sethuraman, J. (1994), “A Constructive Definition of Dirichlet Priors,” *Statistica Sinica*, 4, 639–650. [2130]

Shahbaba, B., and Neal, R. M. (2009), “Nonlinear Models Using Dirichlet Process Mixtures,” *Journal of Machine Learning Research*, 10, 1829–1850. [2130]

Starling, J. E., Murray, J. S., Carvalho, C. M., Bukowski, R. K., and Scott, J. G. (2020), “BART with Targeted Smoothing: An Analysis of Patient-Specific Stillbirth Risk,” *The Annals of Applied Statistics*, 14, 28–50. [2129,2132]

Tokdar, S. T., Zhu, Y. M., Ghosh, J. K. (2010), “Bayesian Density Regression with Logistic Gaussian Process and Subspace Projection,” *Bayesian Analysis*, 5, 319–344. [2129,2130]

van der Vaart, A. W., and van Zanten, J. H. (2008), “Rates of Contraction of Posterior Distributions Based on Gaussian Process Priors,” *The Annals of Statistics*, 36, 1435–1463. [2129]

Wade, S., Dunson, D. B., Petrone, S., and Trippa, L. (2014), “Improving Prediction from Dirichlet Process Mixtures via Enrichment,” *The Journal of Machine Learning Research*, 15, 1041–1071. [2130]

Woody, S., Carvalho, C. M., and Murray, J. S. (2020), “Model Interpretation Through Lower-Dimensional Posterior Summarization,” *Journal of Computational and Graphical Statistics*, 30, 144–161. [2130]

Yang, Y., and Tokdar, S. T. (2015), “Minimax-Optimal Nonparametric Regression in High Dimensions,” *The Annals of Statistics*, 43, 652–674. [2135]

Yu, K., and Moyeed, R. A. (2001), “Bayesian Quantile Regression,” *Statistics & Probability Letters*, 54, 437–447. [2130]

Zhu, R., and Kosorok, M. R. (2012), “Recursively Imputed Survival Trees,” *Journal of the American Statistical Association*, 107, 331–340. [2137]