State-of-the-BART: Simple Bayesian Tree Algorithms for Prediction and Causal Inference

Eoghan O'Neill*
Faculty of Economics
University of Cambridge

November 14, 2019

Abstract

Bayesian Additive Regression Trees (BART) (Chipman et al. 2010) and Bayesian Causal Forests (BCF) (Hahn et al. 2017) are state-of-the-art machine learning algorithms for prediction and individual treatment effect estimation. These methods involve averaging predictions from sum-of-tree models, typically drawn using Monte Carlo Markov Chain methods.

This paper introduces conceptually and computationally simple alternatives to MCMC implementations of BART, which can exhibit comparable performance. An importance sampling based implementation of BART (BART-IS) builds on the ideas of Hernández et al. (2018) and Quadrianto & Ghahramani (2014). Unlike most BART implementations, BART-IS has a data independent model prior. This paper also contains an extension to treatment effect estimation, BCF-IS.

In addition, I describe Bayesian Causal Forests using Bayesian Model Averaging (BCF-BMA), an implementation of BCF (Hahn et al. 2017) that extends an improved implementation of BART-BMA (Hernández et al. 2018) to treatment effect estimation. [1]

*Faculty of Economics, University of Cambridge, Cambridge CB3 9DD, UK. Email: epo21@cam.ac.uk. My thanks are due to Melvyn Weeks, Alexey Onatskiy, Belinda Hernandez, Andrew Parnell, and seminar participants at the Maynooth University Hamilton Institute. The latest version of this paper is available at https://eoghanoneill.com/research/ .

[1] **R** packages implemented in C++ for the methods described in this paper are available at https://github.com/EoghanONeill .

Contents

1 Introduction

Prediction and treatment effect estimation are key tasks for policy makers (Kleinberg et al. 2015). Economists are increasingly applying machine learning methods for treatment effect estimation (Wager & Athey 2017, Athey 2015, 2018).

BART and BCF are Bayesian machine learning methods for prediction and treatment effect estimation. In this paper, a set of different implementation algorithms are described for these methods that make use of a Bayesian Model Averaging framework (Hernández et al. 2018). BART and BCF can be interpreted as model averages of Bayesian linear regressions with the sets of covariates equal to binary variables indicating if observations fall in terminal nodes of decision trees. The covariates are defined by decision tree structures, and priors on the tree structure define the prior on a space of models. This interpretation of BART provides a link to the existing econometric literature on Bayesian Model Averaging of linear models (Steel 2017, Fernandez et al. 2001*a*,*b*, Brock & Durlauf 2001).

BART has been applied in forecasting macroeconomic data with many predictors (Prüser 2019), and has been shown to outperform other macroeconomic forecasting methods (Prüser 2019, Behrens et al. 2019). BART has been used to describe international uncertainty links (Gupta et al. 2016). Pierdzioch et al. (2016) studied the extent to which precious metals are a hedge against exchange rate movements. Pierdzioch et al. (2019) investigated the extent to which inflation predicts returns on Real Estate Investment Trusts.

Applications of BCF include the estimation of the effect of firm size on returns while controlling for book-to-market (Fisher et al. 2019) and multilevel models applied to large experimental datasets (Yeager et al. 2019).

The key contributions of this paper are: 1. An improved implementation of BART-BMA, 2. Bayesian Causal Forests using Bayesian Model Averaging (BCF-BMA), and 3. Simple importance sampling based implementations of BART and BCF (referred to in this paper as BART-IS and BCF-IS), closely following the approach for single classification trees described by Quadrianto & Ghahramani (2014).

These methods are conceptually simple, in that conjugate priors give a tractable closed form for the predictive distribution (e.g. of the Average Treatment Effect). Furthermore the implementations are simple relative to existing MCMC approaches.

The output of BCF-BMA contains relatively few sum-of-trees models. Under the default settings, each model include five trees describing the treatment effect function and each tree contains at most five splits. Therefore the output is more interpretable than that of standard MCMC implementations, which usually draw thousands of models, each of which contains a sum of a hundred or more trees.

The appeal of BART-IS and BCF-IS is that they are straightforward to implement and very parallelizable.

2 Review of BART and BART-BMA

In this section, I describe the BART model (Chipman et al. 2010), review existing BART implementations, note existing theory and applications of BART, and describe BART-BMA (Hernández et al. 2018).

2.1 BART overview

2.1.1 Description of model and priors

Suppose there are n observations, and the $n \times p$ matrix of explanatory variables, X, has i^{th} row $x_i = [x_{i1}, ..., x_{ip}]$. For the standard BART model $Y_i = \sum_{j=1}^{m} g(x_i; T_j, M_j) + \varepsilon_i$, where $g(x_i; T_j, M_j)$ is the output of a decision tree. T_j refers to decision tree $j = 1, ..., m$, where m is the total number of trees in the model. M_j are the terminal node parameters of T_j, and $\varepsilon_i \overset{i.i.d}{\sim} N(0, \sigma^2)$.

For BART (Chipman et al. 2010), prior independence is assumed across trees T_j and across terminal node means $M_j = (\mu_{1j}...\mu_{b_j j})$ (where $1, ..., b_j$ indexes the terminal nodes of tree j). The form of the prior used by Chipman et al. (2010) is:

$$p(M_1, ..., M_m, T_1, ..., T_m, \sigma) \propto \left[\prod_j \prod_k p(\mu_{kj}|T_j) p(T_j) \right] p(\sigma)$$

In standard BART, $\mu_{kj}|T_j \overset{i.i.d}{\sim} N(0, \sigma_0^2)$ where $\sigma_0 = \frac{0.5}{e\sqrt{m}}$ and e is a user-specified hyper-parameter.

Chipman et al. (2010) set a regularization prior on the tree size and shape $p(T_j)$ to discourage any one tree from having undue influence over the sum of trees. The probability that a given node within a tree T_j is split into two child nodes is $\alpha(1 + d_h)^{-\beta}$, where d_h is the depth of (internal) node h and α and β are parameters which determine the size and shape of T_j respectively. Thus $p(T_j) = \prod_{h=1}^{b_j - 1} \alpha(1+d_h)^{-\beta} \prod_{k=1}^{b_j} (1-\alpha(1+d_k)^{-\beta})$, where h indexes the internal nodes of the tree T_j, and k indexes the terminal nodes.

Chipman et al. (2010) assume that the model precision σ^{-2} has a conjugate prior distribution $\sigma^{-2} \sim Ga(\frac{v}{2}, \frac{v\lambda}{2})$ with degrees of freedom v and scale λ. There are also priors on the splitting variables and splitting points in each tree. Chipman et al. (2010) use the uniform prior on available splitting variables, and the uniform prior on the discrete set of available splitting variables.

2.1.2 Existing BART implementations

Samples can be taken from the posterior distribution $p((T_1, M_1), ..., (T_m, M_m), \sigma|y)$ by a Bayesian backfitting MCMC algorithm. This algorithm is a Gibbs or Metropolis Hastings sampler, involving m successive draws from $(T_j, M_j)|T_{(j)}, M_{(j)}, \sigma, y$ for $j = 1, ..., m$ [where $T_{(j)}, M_{(j)}$ are the trees and parameters for all trees except the j^{th} tree] followed by a draw of σ from the full conditional $\sigma|T_1, ..., T_m, M_1, ..., M_m, y$.

A set of draws induces the sum of trees function $f^*(.) = \sum_{j=1}^{m} g(. ; T_j^*, M_j^*)$. After burn-in, the sequence of f^* draws, $f_1^*, ..., f_Q^*$ may be regarded as an approximate, dependent sample of size Q from $p(f|y)$. To estimate $f(x)$, a natural choice is $\frac{1}{Q} \sum_{q=1}^{Q} f_q^*(x)$, which approximates $E(f(x)|y)$. Prediction intervals can be obtained from quantiles of the draws $f_q^*(x)$.

A number of papers describe faster BART implementation algorithms and improved sampling methods, including parallelized BART (Pratola et al. 2014), particle Gibbs algorithms (Lakshminarayanan et al. 2015), more efficient Metropolis-Hastings proposals (Pratola et al. 2016), Consensus Monte Carlo (Scott et al. 2016), a likelihood-inflated sampling algorithm (Entezari et al. 2018), and Accelerated BART (X-BART, which uses a stochastic hill climbing algorithm as a greedy stochastic approximation to MCMC) (He et al. 2018).

An alternative to the MCMC BART implementation is Approximate Bayesian Computation Bayesian Forests (Liu et al. 2018), which has been shown to be consistent for variable selection under certain conditions.

BART-BMA (Hernández et al. 2018), in contrast to other BART implementations, does not involve MCMC methods. A greedy model search algorithm adds trees to sum-of-tree models by first restricting the

set of potential splitting points using a changepoint detection algorithm, and only keeping sum-of-tree models with posterior model probabilities within a distance , known as Occam'a window (Madigan & Raftery 1994), of the highest probability model currently in the set of selected models. [2]

2.1.3 BART theory

Recent papers have discussed the asymptotic properties of BART. Posterior concentration rates are derived by Rockova & van der Pas (2017), Linero & Yang (2017) and Rocková & Saha (2018). Castillo & Rockova (2019) obtain uncertainty quantification results. Asymptotic properties of variable selection are derived by Liu et al. (2018). Asymptotic results for estimating ITEs using Bayesian methods more generally are described by Alaa & van der Schaar (2018).

2.1.4 Extensions and Applications of BART

BART has been extended for a wide range of applications. Starling et al. (2018) describe a BART method for functional data analysis that parametrizes each tree's terminal nodes with smooth functions of a target covariate. Another smooth variant of BART is BART with "soft" decision trees (Linero 2018).

Some variations on the BART priors have been suggested for variable selection, including a Dirichlet hyperprior on the probability that a variable is used for a split (Linero 2018) and spike and tree priors (Rockova & van der Pas 2017, Liu et al. 2018). An overlapping group Dirichlet hyperprior has been applied to splitting probabilities for a dataset in which the variables have an overlapping group structure (Du & Linero 2019). A prior for interaction detection has been proposed by Du & Linero (2018).

BART can be applied to data without i.i.d normally distributed error terms. Heteroscedastic BART models the error as a product of trees (Pratola et al. 2017), and fully nonparametric BART (George et al. 2018) models the error using a Dirichlet process mixture.

BART has been adapted for different outcome variables, including Bayesian quantile additive regression trees (Kindo, Wang, Hanson & Peña 2016), Multiclass Bayesian Additive Classification Trees (Kindo et al. 2013), BART methods for multinomial outcomes (Agarwal et al. 2014, Kindo, Wang & Peña 2016), loglinear BART (Murray 2017), random intercept BART (Tan et al. 2016), BART for survival analysis (Bonato et al. 2010, Sparapani et al. 2016), BART for competing risks models (Sparapani et al. 2019), and BART modelling of recurrent events (Sparapani et al. 2018). A general framework for extending BART to different tasks is described by Tan & Roy (2019).

BART can also be applied to data with multiple outcomes. Chakraborty (2016) applies BART to Seemingly Unrelated Regression, and Linero et al. (2019) describe shared Bayesian Forests. BART has been used for the imputation of missing data (Xu et al. 2016, Tan et al. 2018) and the modelling of missing data in longitudinal studies (Zhou et al. 2019).

BART has been applied to treatment effect estimation (Hill 2011, Green & Kern 2012, Taddy et al. 2015, Henderson et al. 2017). Data analysis competitions (Dorie et al. 2019, Hahn et al. 2019, Carvalho et al. 2019) have shown that BART is a very accurate treatment effect estimation method. Hahn et al. (2017) introduce Bayesian Causal Forests (BCF), a BART based method for treatment effect estimation that allows the prior regularization of the treatment effect estimate to be specified separately to the prior regularization of the rest of the model for the outcome.

[2] In the original implementation of BART-BMA, a Gibbs sampler was used for constructing prediction intervals. In the new implementation, quantiles are obtained from a closed form for the model averaged posterior predictive distribution. Therefore BART-BMA provides an implementation of BART that does not require any random number generation.

Hahn et al. (2017) also note that standard BART treatment effect estimates can be improved by including the propensity score as a potential splitting variable. Santos & Lopes (2018) study the performance of this approach on sparse data using the Dirichlet hyperprior described by Linero (2018). BCF has been extended to Instrumental Variable estimation of treatment effects by Bargagli-Stoffi et al. (2019).

2.2 BART-BMA overview

BART-BMA applies the same priors as standard BART (section 2.1.1), except the variance of the terminal node parameters is proportional to the variance of the error term, $\mu_{ij}|T,\sigma \sim N(0,\frac{\sigma^2}{a})$, as suggested by Chipman et al. (1998). Integration of the likelihood with respect to the μ parameters and σ results in a closed form expression proportional to the marginal likelihood.

The marginal likelihood can be derived as follows. Let $Y = (Y_1, ..., Y_n)$ be the outcome vector. For a given sum of trees model $\mathcal{T}$, the likelihood of Y is:

$$Y|\mathcal{T}, M, \sigma^{-2} \sim N(\sum_{j=1}^{m} J_j M_j, \sigma^2 I)$$

where J_j (which depends on the original matrix of covariates X) is an $n \times b_j$ binary matrix whose (i,k) element denotes the inclusion of observation $i = 1, ..., n$ in terminal node $k = 1, ..., b_j$ of tree j.

Let $W = [J_1 ... J_m]$ be an $n \times \omega$ matrix , where $\omega = \sum_{j=1}^{m} b_j$ and $O = (M_1^T ... M_m^T)^T$ be a vector of size ω of terminal nodes assigned to trees $T_1, ..., T_m$. We can then write $WO = \sum_{j=1}^{m} J_j M_j$, [3] and therefore the likelihood can be rewritten as:

$$Y|O, \sigma^{-2} \sim N(WO, \sigma^2 I)$$

which, with $O \sim N(0, \frac{\sigma^2}{a} I_\omega)$, where I_ω is a $\omega \times \omega$ identity matrix, implies that $Y \sim MVST_v(0, \lambda(I_n + \frac{1}{a} W W^T))$ (where MVST denotes a multivariate student t-distribution), and the marginal likelihood is:

$$p(Y) = \frac{\Gamma(\frac{\nu+n}{2})(\lambda v)^{\frac{\nu+n}{2}}}{\Gamma(\frac{v}{2}) v^{\frac{n}{2}} \pi^{\frac{n}{2}} \lambda^{\frac{n}{2}} (\frac{1}{a})^{\frac{\omega}{2}} det\,(aI_\omega + W^T W)^{\frac{1}{2}}} \left[\lambda v + Y^T Y - Y^T W (aI_\omega + W^T W)^{-1} W^T Y\right]^{-\frac{\nu+n}{2}}$$

Then, noting that anything that does not depend on W or ω will cancel out when calculating the model weights, we can calculate:

$$\propto \frac{1}{(\frac{1}{a})^{\frac{\omega}{2}} det\,(aI_\omega + W^T W)^{\frac{1}{2}}} \left[\lambda v + Y^T Y - Y^T W (aI_\omega + W^T W)^{-1} W^T Y\right]^{-\frac{\nu+n}{2}}$$

And the log of this expression is:

$$\frac{\omega}{2}\log(a) - \frac{1}{2}\log(det(M)) - \frac{\nu+n}{2}\log(\lambda v + Y^T Y - Y^T W M^{-1} W^T Y)$$

where $M = aI_\omega + W^T W$

A deterministic model search algorithm first reduces the set of potential splitting variables by a change-point detection algorithm, and then recursively adds splits to trees that are potentially to be appended to models in the set of currently selected sum of tree models. After a set of single tree models are selected,

[3] $WO = \sum_{j=1}^{m} J_j M_j$ is analogous to $X\boldsymbol{\beta}$ in standard linear regression notation.

changepoints in the residuals are used as potential splitting variables for constructing the next set of trees to potentially append to the selected models.[4]

The set of models to be averaged over are those with posterior probability within some distance of the highest probability model found by the model search algorithm. i.e. For all proposed models, $\mathcal{T}_\ell$, indexed by ℓ, the algorithm obtains

$$p(Y|\mathcal{T}_\ell, X)p(\mathcal{T}_\ell) \propto p(\mathcal{T}_\ell|Y, X) = \frac{p(Y|\mathcal{T}_\ell, X)p(\mathcal{T}_\ell)}{p(\mathbf{y})}$$

And keeps the models such that

$$\log(p(\mathcal{T}_\ell|Y, X)) - \arg\min_{\ell'}(\log(p(\mathcal{T}_{\ell'}|Y, X))) \leq \log(o)$$

where o is Occam's window, and the minimum is over the set of all proposed models.

The original BART-BMA algorithm derived prediction intervals by Gibbs sampling from full conditionals for the model parameters for each selected model. However, the posterior predictive distributions for the selected models are multivariate t-distributions, as the models are Bayesian linear regressions with covariates equal to indicator variables for terminal node parameters. Therefore posterior distributions and prediction intervals can be obtained without random number generation (see section 3 for further details).

3 Extensions of the BART-BMA Algorithm

3.1 Summary of improvements

In this section, I describe some improvements made to the BART-BMA algorithm.

The BART-BMA algorithm searches for trees to add to sum-of-tree models. The set of potential splitting points to be used in searching for a tree is restricted by a grid search algorithm or Pruned Exact Linear Time changepoint detection algorithm (Killick et al. 2012, Hernández et al. 2018). [5] First, the residual from a sum-of-tree model currently in Occam's window is obtained, then the grid search approach considers a fixed number of equally spaced splitting points for each covariate, and orders the potential splitting points by squared error of the predictions of the residual resulting from a binary split. A percentage of splitting points, set by the user, are kept for constructing trees. The original BART-BMA algorithm approximated the residuals of each sum-of-tree model by subtracting single tree predictions each time a tree was appended to the model. The new implementation uses residuals from the fulll sum-of-tree models instead of an approximation.

Other improvements include bug fixes and more precise calculations of the marginal likelihood and prior. I describe below how to obtain prediction intervals from a closed form for the posterior predictive distribution.

For a given sum-of-tree model the posterior distribution for the vector of terminal node parameters is

$$O|Y, \mathcal{T} \sim MVSt_{\nu+n}\left(M^{-1}W^TY\ ,\ \frac{1}{\nu+n}[\nu\lambda + Y^TY - Y^TWM^{-1}W^TY]M^{-1}\right)$$

[4] In the original paper, Hernández et al. (2018) construct residuals by subtracting from the outcomes the sum of single tree model predictions. In this paper I present the results of an improved algorithm that calculates the residuals by subtracting the predictions obtained from the posterior mean of the whole sum-of-tree model.

[5] For details on how the PELT algorithm is used in BART-BMA, see Hernández et al. (2018),

where $M = aI_\omega + W^T W$.[6] The posterior distribution of $OW = f(x)$ (for in-sample estimates) is:

$$WO|Y, \mathcal{T} \sim MVSt_{\nu+n}\left(WM^{-1}W^T Y \ , \ \frac{1}{\nu+n}[\nu\lambda + Y^T Y - Y^T W M^{-1} W^T Y] W M^{-1} W^T\right)$$

The posterior predictive (i.e. out-of-sample) distribution for a sum-of-tree model is:

$$\tilde{Y}|Y, W, \tilde{W}, \mathcal{T}_\mu, \mathcal{T}_\tau \sim MVSt_{\nu+n}\left(\tilde{W}M^{-1}W^T Y, \ \frac{1}{\nu+n}[\nu\lambda + Y^T Y - Y^T W M^{-1} W^T Y](I_{\tilde{n}} + \tilde{W} M^{-1} \tilde{W}^T)\right)$$

where the tilde notation indicates numbers or random variables relating to out-of-sample data.

Therefore, unconditional on the model, the posterior predictive distribution for BART-BMA is a posterior model probability weighted mixture of multivariate t-distributions. Often pointwise prediction intervals are of interest. Therefore, for each individual, we only need to obtain the marginal posterior (predictive) distribution, which is (for each model) a univariate t-distribution with location and scale. Then the marginal mixture distribution has a closed form PDF, and a CDF that can be evaluated by numerical integration methods. Prediction intervals can therefore be constructed by obtaining the quantiles of the (marginal) mixture distribution's CDF by a root finding algorithm (e.g. bisection).[7] This approach is also used to obtain prediction intervals for BART-IS and BCF-IS, which involve averaging of a much larger set of predictive distributions.

3.2 Potential extensions of BART-BMA

The closed form for the predictive distribution suggests a number of possible improvements and variations on BART-BMA.

- The a parameter could be set by a full Bayesian approach, Empirical Bayes approach, cross-validation, or other methods.
- The BART-BMA predictions are a probability weighted average of ridge regressions. Methods for fast estimation of ridge regressions can be applied for improved computational speed.
- Different priors can be applied to the terminal node parameters and the error variance. This was discussed to an extent in the original single Bayesian tree context by Chipman et al. (1998). For example, data-informed priors can be applied to the error variance, as outlined in the context of standard Bayesian linear regression by Sala-i Martin et al. (2004).
- Different model weights can be applied, for example, weights can be based on in-sample sum-of-squared errors. This was discussed to an extent by Chipman et al. (1998). For BMA of linear regressions, Sala-i Martin et al. (2004) suggest model weights equal to $p(M_j)n^{-k/2}SSE_j^{n/2}$, where M_j is the model and n is the number of observations. Another option is a BIC approximation, $p(M_j)[n\ln(\frac{1}{n}SSE_j) + k\ln(n)]$.
- BART-BMA outputs a relatively small number of Bayesian linear regressions. The covariates are indicator variables for terminal nodes. In principle, any Bayesian model combination method can be

[6] An alternative would be to draw $\{O^{(q)}, \sigma^{2(q)}\}_{q=1}^{Q}$ from $O^{(q)} \sim MVN(M^{-1}W^T Y, \sigma^{2(q)} M^{-1})$ and $\sigma^{2(q)} \sim \Gamma^{-1}\left(\frac{\nu+n}{2}, \frac{\nu\lambda}{2} + \frac{1}{2}[Y^T Y - Y^T W M^{-1} W^T Y]\right)$

[7] It is also possible to directly sample from the mixture of multivariate t-distributions and obtain pointwise quantiles, or to separately sample from a mixture of univariate t-distributions for each individual in the out-of-sample data.

applied to the set of selected models. For example, Bayesian Stacking (Yao et al. 2018) might give a more accurate predictive distribution.

The spike and tree prior (Rockova & van der Pas 2017) can also be applied to the space of sum-of-tree models instead of the standard BART prior. Details for this prior are included in appendix D.

3.3 BART-BMA for Treatment Effect Estimation

This subsection outlines how BART-BMA can be applied to treatment effect estimation in an approach similar to that described by Hill (2011), but using the choice of conjugate priors in BART-BMA to obtain a closed form posterior distribution for Individual Treatment Effects (ITEs) and the Conditional Average Treatment Effect CATE.

Following the approach of Hill (2011), let the BART-BMA ITE estimate be defined as $\hat{\tau}(x) = \hat{f}_1(x) - \hat{f}_0(x)$, where $\hat{f}(x)$ is obtained from fitting a BART-BMA regression of the outcome on the covariate and treatment (i.e. include the treatment indicator as a covariate). $\hat{f}_1(x)$ ($\hat{f}_0(x)$) is the estimate obtained for covariate vector x when the treatment status is set to 1 (0).

Let $OW_1 = f_1(x)$, and $OW_0 = f_0(x)$, where W_1 is the W matrix obtained if all Z values are reset to 1 (and similarly W_0 is obtained by setting $Z = 0$). Note that some splits can be on Z, and this determines how W changes with Z.

Consider the posterior predictive distribution of the ITE for a given sum-of-trees model.

$$ITE = f_1(x) - f_0(x) = W_1O - W_0O = (W_1 - W_0)O$$

Let $W_{diff} = W_1 - W_0$ Then the in-sample posterior distribution of the vector of ITEs for all individuals (in the sample) is:

$$W_{diff}O|Y, \mathcal{T} \sim MVSt_{\nu+n}\left(W_{diff}M^{-1}W^TY, \frac{1}{\nu+n}[\nu\lambda + Y^TY - Y^TWM^{-1}W^TY]W_{diff}M^{-1}W_{diff}^T\right)$$

[8] The in-sample posterior distribution of the CATE, $\frac{1}{n}\sum_{i=1}^{n}\tau(x)$, is:

$$\frac{1}{n}\mathbf{1}^TW_{diff}O|Y, \mathcal{T} \sim MVSt_{\nu+n}\left(\frac{1}{n}\mathbf{1}^TW_{diff}M^{-1}W^TY, \frac{1}{\nu+n}[\nu\lambda + Y^TY - Y^TWM^{-1}W^TY]\frac{1}{n}\mathbf{1}^TW_{diff}M^{-1}W_{diff}^T\frac{1}{n}\mathbf{1}\right)$$

$\mathbf{1}$ is a vector of ones of length n. [9]

The distribution for the Conditional Average Treatment Effect on the Treated (CATT) can be obtained by replacing $\frac{1}{n}\mathbf{1}^T$ with $\frac{1}{n_{treated}}\mathbf{z}^T$ and the Conditional Average Treatment Effect on the Not Treated (CATNT) distribution can be obtained using $\frac{1}{n_{control}}(\mathbf{1}-\mathbf{z})^T$.

[8] For out-of-sample ITEs, let $\tilde{W}_{diff} = \tilde{W}_1 - \tilde{W}_0$. Then $\tilde{W}_{diff}O|Y, \mathcal{T} \sim MVSt_{\nu+n}\left(\tilde{W}_{diff}M^{-1}W^TY, \frac{1}{\nu+n}[\nu\lambda + Y^TY - Y^TWM^{-1}W^TY]\tilde{W}_{diff}M^{-1}\tilde{W}_{diff}^T\right)$. Note that the error term does not enter $f_1(x) - f_0(x)$ and therefore there is no $I_{\tilde{n}}$ term in the variance of the out-of-sample posterior distribution.

[9] The out-of-sample posterior distribution of the CATE is: $\frac{1}{\tilde{n}}\tilde{\mathbf{1}}^T\tilde{W}_{diff}O|Y, \mathcal{T} \sim MVSt_{\nu+n}\left(\frac{1}{\tilde{n}}\tilde{\mathbf{1}}^T\tilde{W}_{diff}M^{-1}W^TY, \frac{1}{\nu+n}[\nu\lambda + Y^TY - Y^TWM^{-1}W^TY]\frac{1}{\tilde{n}}\tilde{\mathbf{1}}^T\tilde{W}_{diff}M^{-1}\tilde{W}_{diff}^T\frac{1}{\tilde{n}}\tilde{\mathbf{1}}\right)$ where $\tilde{\mathbf{1}}$ is a vector of ones of length $\tilde{n}$.

4 BART-IS

This section presents BART-IS, which extends the importance sampling approach described by Quadrianto & Ghahramani (2014) from single classification trees to sums of regression trees by utilising the conjugate priors of BART-BMA.

Importance sampling of Bayesian linear regression models involves constructing weights by dividing the prior model probability by the model sampling probability. Therefore the model prior and importance sampler probabilities do not need to be calculated when the models are sampled from the prior. This approach is used by Quadrianto & Ghahramani (2014) in safe-Bayesian Random Forests for classification, and by Sala-i Martin et al. (2004) in their implementation of BMA of linear regressions. For completeness, the option of using different samplers and priors is provided by the **safeBart** package.[10]

Bayesian Model Averaging tends towards one model as the number of observations tends to infinity. However, when the model space does not contain the true model, more accurate predictions can be obtained from Bayesian Model Combination, which tends towards a combination of models. Quadrianto & Ghahramani (2014) apply a standard model combination approach by raising the model likelihoods to a power. This makes the approach "safe" in the sense that it does not tend towards one possibly wrong model. The option of raising the likelihood to a power is provided in the **safeBart** package

Preprocessing involves a probability integral transformation of each covariate, with the distribution equal to the empirical cumulative distribution function. The BART-IS algorithm randomly samples all trees in each sum-of-tree model from the independent tree prior, and calculates the marginal likelihood and predictions for each sum-of-tree model. The likelihoods can be raised to a power for a safe-Bayesian approach. The final predictions are a marginal-likelihood weighted average.

The BART-IS algorithm is generalizable in the sense that the prior tree model distribution can be replaced by any prior on partitions of the covariate space. The partitions do not need to be representable in binary tree structures. As long it is possible to (quickly) draw partitions and construct indicator variables for sets in the partitions, this approach is applicable. Then any conjugate Bayesian linear regression priors can be applied given a set of indicator variables as covariates.

BART-IS is applicable to ITE estimation using the distributions outlined in 3.3. [11] In principle, BART-IS can also be applied to data with multiple outcomes by applying standard conjugate priors for Bayesian multivariate linear regression. I outline this approach in appendix B.

4.1 Description of the BART-IS algorithm

1. Sample sets of trees from a prior. The prior can be the standard BART prior (Chipman et al. 2010), the prior described by Quadrianto & Ghahramani (2014), or the spike-and-tree prior (Rockova & van der Pas 2017). If the predictive distribution and/or variable importance measures are desired, then the matrices representing the tree structure should be saved at this point.

2. Obtain the model predictions. If computational speed is desired, particularly for a large number of samples, or for models with many trees, a fast ridge regression algorithm can be applied for model predictions.

[10] The package is publicly available at https://github.com/EoghanONeill/safeBart .

[11] This approach to ITE estimation is available in the **safeBart** package available at https://github.com/EoghanONeill/safeBart .

3. Obtain model weights. This could involve raising the marginal likelihood to a power, as described by Quadrianto & Ghahramani (2014). This is because it is possible that none of the set of models is the true model, but BMA tends towards placing all the weight on one model. In practice a Bayesian Model Combination approach, such as the power likelihood approach, might be more accurate. If importance sampling is not from the prior, then the likelihood should be multiplied by the ratio of the model prior probability to the importance sampler model probability. Weights constructed from residuals instead of the marginal likelihood might also increase computational speed.

4. Obtain the predictive distribution, which is a mixture of multivariate t-distributions.

It is possible to quickly sample from the prior described by Quadrianto & Ghahramani (2014) and the standard BART prior (Chipman et al. 2010). In Appendix D.2, I outline how to sample from a spike and tree prior.

5 Results for BART-BMA and BART-IS

5.1 High-dimensional data

Figure 1 presents the results obtained by applying the following methods to to the commonly used simulations introduced by Friedman et al. (1991): BART-BMA with the standard BART model prior, BART-BMA with the spike and tree prior,[12], BART-IS, BART with 1000 and 10,000 MCMC draws, Dirichlet BART with 1000 and 10,000 MCMC draws, and Random Forests. [13]

The outcome depends on 5 uniformly distributed predictor variables $x_1, x_2, ..., x_5$:

$$y = 10\sin(\pi x_1 x_2) + 20(x_3 - 0.5)^2 + 10x_4 + 5x_5 + \varepsilon$$

where $\varepsilon \sim \mathcal{N}(0, 1)$. Variables $x_6, ..., x_p$ are uniformly distributed. The number of observations is 500. I considered 5 different values of p, the number of covariates: $p = (100, 1000, 5000, 10000, 15000)$. The RMSE were obtained using fivefold cross-validation. The default parameter values were used for RF, BART, and Dirichlet BART (DART).[14]

The results for the new variations of BART-BMA compare favourably to the results obtained for the original implementation (Hernández et al. 2018), which gives RMSE between 2.9 and 3. [15] BART-BMA with the Gridpoint changepoint detection algorithm and standard priors has RMSE which does not appear to deteriorate as the number of variables increases. BART-BMA variations that update the set of potential splitting points within the construction of individual trees exhibit deteriorating performance as the number of variables is increased.

BART and Dirichlet BART under default parameter settings do not perform very well when the number of variables is increased to 5000. However, the default implementation of BART and DART includes only

[12]The BART-BMA results presented here are for BART-BMA with the gridpoint method for changepoint detection. Another option is the Pruned Exact Linear Time algorithm (Killick et al. 2012). The results presented here are for BART-BMA with no within-tree updating of potential split points. Another option is to update potential splitting points after a split is added to each tree.

[13]BART-IS was implemented with 1,000,000 draws of sum-of-tree models each containing 30 trees. Each of the 10,000 BART and DART sum-of-tree models contained 200 trees.

[14]Random Forest was implemented using the **R** package **ranger**. BART was implemented using the **wbart** function in the **R** package **BART**. DART was implemented using the **wbart** function in the **R** package **BART** with the following parameter setting sparsity = TRUE

[15]See original paper by Hernández et al. (2018).

1000 draws from the posterior with 100 burn-in draws. Figure 1 demonstrates that BART and DART exhibit much better performance in high dimensional data when the number of MCMC samples is increased to 10,000.

BART can be used for variable selection (Linero 2018, Bleich et al. 2014). The Brier scores for the BART, DART, and BART-BMA posterior variable inclusion probabilities (PIP) are given in table 1. [16] The Brier score is defined as $\frac{1}{P}\sum_{p=1}^{P}(I_p - PIP_p)^2$ where p indexes the covariates, $I_p = 1$ for truly important variables $x_1, ..., x_5$ and $I_p = 0$ otherwise. The results suggest that BART-BMA outperforms BART and DART in terms of variable selection. The spike-and-tree prior outperforms the standard BART prior.

Number of Variables	BART-BMA Prior		BART	DART
	Standard	Spike-and-tree		
100	2.000×10^{-3}	3.560×10^{-32}	7.005×10^{-1}	3.161×10^{-3}
1000	4.000×10^{-4}	1.350×10^{-31}	3.974×10^{-2}	7.210×10^{-5}
5000	2.000×10^{-4}	2.700×10^{-32}	3.236×10^{-3}	1.513×10^{-4}
10000	1.000×10^{-4}	1.350×10^{-32}	1.150×10^{-3}	1.448×10^{-4}
15000	8.000×10^{-5}	9.000×10^{-33}	6.648×10^{-4}	1.120×10^{-4}

Table 1: Brier Scores for Friedman data simulations

Prediction intervals obtained directly from the closed form for the point-wise predictive distributions were obtained for the Friedman data simulations, and the results for 95% prediction intervals are presented for BART-BMA, BART, and DART in tables 2 and 3. The average coverage of prediction intervals is comparable to, or slightly better than that originally obtained by Hernández et al. (2018), and the average interval width is narrower. BART-BMA gives more precise prediction intervals than BART and DART, although the intervals for DART and BART are notably narrower for low dimensional simulations. [17]

Number of Variables	BART	DART	BART-BMA Standard	BART-BMA Spike and tree	BART-IS
100	95.0	94.4	95.4	94.6	97.4
1000	97.4	96.8	95.8	94.6	97.4
5000	97.0	97.0	95.4	94.6	95.4
10000	97.6	98.4	94.8	94.6	97.6
15000	98.8	98.2	94.8	94.6	94.0

Table 2: Average 95% prediction interval coverage for Friedman data simulations

Number of Variables	BART	DART	BART-BMA Standard	BART-BMA Spike and tree	BART-IS
100	6.74	4.77	9.62	10.11	12.65
1000	9.24	5.07	9.61	10.24	15.81
5000	10.70	6.00	9.64	10.24	16.10
10000	12.11	8.15	9.61	10.24	16.61
15000	12.89	10.13	9.61	10.24	16.39

Table 3: Average 95% prediction interval width for Friedman data simulations

[16] The results in table 1 are for BART and DART with 10,000.

[17] Also, the BART and DART results might improve with more MCMC draws as this would allow for convergence of the Markov Chain, and more accurate estimation of quantiles of the posterior distribution. The chosen number of MCMC draws for BART and DART is 10,000. For each draw of a sum-of-tree model, 10 draws of the additive error term, ε were made from a normal distribution. It is likely that more accurate intervals could be obtained with a greater number of draws of the error.

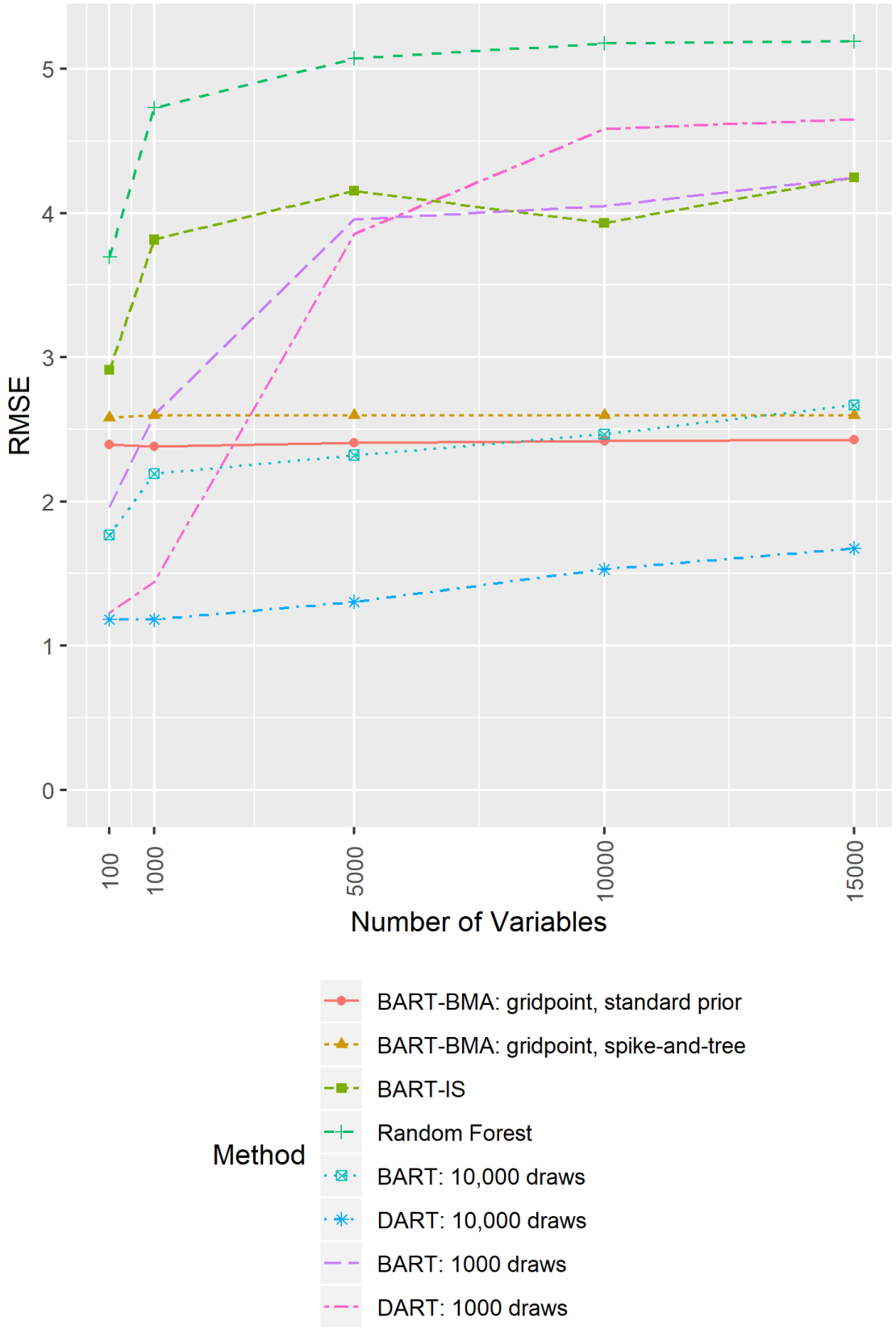

Figure 1: RMSEs for Friedman Data Simulations

5.2 Low-dimensional data

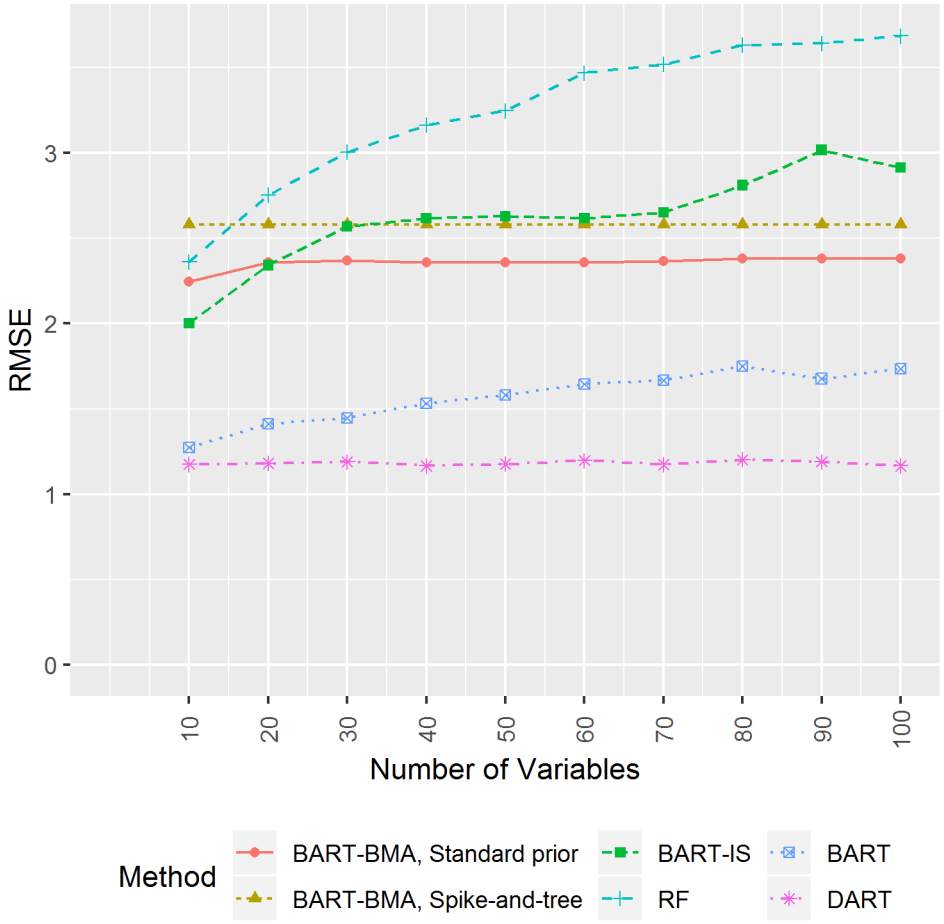

Figure 2: RMSEs for Friedman Data Simulations

Figure 2 presents the results for the Friedman simulations described in section 5.1, with the number of covariates, p equal to 10, 20, 30 ,40, 50, 60, 70, 80, 90, and 100. The RMSE was averaged across five simulations for each value of p.

BART and DART outperform other methods in terms of RMSE. It can be observed that the predictions of RF, BART, and BART-IS become less accurate as the number of covariates increases. However, it is likely that the accuracy of these methods when applied to high dimensional data would improve with a greater number of draws of models.

Tables 4 and 5 present the results for 95% prediction intervals.[18] BART-BMA obtains better coverage and average interval width than BART-IS.

[18]The chosen number of MCMC draws for BART and DART is 10,000. For each draw of a sum-of-tree model, 10 draws of the additive error term, ε were made from a normal distribution. It is likely that more accurate intervals could be obtained with a greater number of draws of the error.

Number of Variables	BART	DART	BART-BMA Standard	BART-BMA Spike and tree	BART-IS
10	95.4	94.4	95.6	94.6	98.8
20	94.0	95.0	94.6	94.6	97.8
30	93.6	94.2	94.4	94.6	96.8
40	95.4	94.0	94.4	94.6	96.8
50	94.4	94.6	94.4	94.6	97.8
60	93.2	94.6	94.4	94.6	98.6
70	94.8	94.8	94.4	94.6	97.0
80	94.2	94.6	94.4	94.6	96.6
90	94.2	94.6	94.4	94.6	96.6
100	96.2	94.8	94.4	94.6	97.4

Table 4: Average 95% prediction interval coverage for low-dimensional Friedman data simulations

Number of Variables	BART	DART	BART-BMA Standard	BART-BMA Spike and tree	BART-IS
10	5.04	4.72	8.93	10.11	9.59
20	5.49	4.73	8.97	10.11	10.47
30	5.75	4.72	9.02	10.11	11.07
40	5.98	4.77	9.03	10.11	10.91
50	6.17	4.78	9.03	10.11	11.64
60	6.34	4.76	9.03	10.11	12.01
70	6.51	4.78	9.03	10.11	11.62
80	6.45	4.78	9.08	10.11	12.48
90	6.73	4.76	9.08	10.11	12.74
100	6.81	4.79	9.08	10.11	12.65

Table 5: Average 95% prediction interval width for low-dimensional Friedman data simulations

6 BCF-BMA and BCF-IS

This section introduces a combination of the BCF parametrization of BART for treatment effect estimation (Hahn et al. 2017) and the BART-BMA model search implementation of BART (Hernández et al. 2018). [19] First I describe BCF and BCF-BMA model and prior, then I describe the BCF-BMA posterior and algorithm. Bayesian Causal Forests using Importance Sampling (BCF-IS) is also briefly described. Finally I present results for ITE estimation on simulated data, giving a comparison between BCF-BMA, other algorithms introduced in this paper, and existing state-of-the-art methods BCF, BART, and causal forests (Wager & Athey 2017, Athey et al. 2017).

6.1 BCF

BCF controls for confounding by including the estimated propensity score as a splitting variable, and allows the treatment effect function to be regularized separately to the rest of the model.

6.1.1 BCF summary

Hill (2011) proposed the use of BART to estimate treatment effects by including the treatment variable Z in the set of splitting variables, and estimating the model $Y_i = f(x_i, Z_i) + \epsilon_i,\ \epsilon \sim N(0, \sigma^2)$. The treatment

[19] The **R** package for BCF-BMA is publicly available at https://github.com/EoghanONeill/bcfbma .

effect can be expressed as $\tau(x_i) = f(x_i, 1) - f(x_i, 0)$. If an individual has a vector of covariates x, then the difference in predictions for $(X = x,\ Z = 1)$, and $(X = x,\ Z = 0)$ is the estimated treatment effect. [20]

However, the implications of the prior on $f(x, z)$ for the induced prior on τ are difficult to understand, and the induced prior on τ will vary with the number of covariates. Furthermore, the estimates can be biased in the presence of confounding. Hahn et al. (2017) propose an alternative approach, and elaborate on an issue referred to as "Regularization Induced Confounding" (Hahn et al. 2018). Regularization priors tend to adversely bias treatment effect estimates by over-shrinking control variable regression coefficients. In the presence of confounding, the finite sample bias of the treatment effect estimator will be influenced by the prior regularization, and it is desirable to directly control regularization of the treatment effect function.

Confounding can be mitigated by including the estimated propensity score as a potential splitting variable (Hahn et al. 2017). Hahn et al. (2017) propose a re-parametrization that allows for an independent prior to be placed on τ and also include the estimated propensity score, $\hat{\pi}_i$, as a potential splitting variable.

$$f(x_i, z_i) = \mu(x_i, \hat{\pi}_i) + \tau(x_i) z_i$$

where $\mu(x_i, \hat{\pi}_i)$ and $\tau(x_i)$ are both sums of trees.

Different BART parameters (e.g. the number of trees, depth penalty, splitting probability, scale of terminal node outputs) are used for the sums of trees denoted by $\mu(x_i)$ and $\tau(x_i)$, and $\tau(x_i)$ priors are set such that it is more strongly regularized than $\mu(x_i)$.

Hahn et al. (2017) demonstrate that BCF can perform well in simulations in terms of MSE of individual treatment effect estimates relative to: BART (Hill 2011), including the propensity score estimates in standard BART, fitting BART separately to treated and control groups, and causal forests (Wager & Athey 2017, Athey et al. 2017).

6.1.2 BCF priors

Chipman et al. (2010) assume that the model precision σ^{-2} has a conjugate prior distribution $\sigma^{-2} \sim Ga(\frac{v}{2}, \frac{v\lambda}{2})$ with degrees of freedom v and scale λ. The same prior is used for the model precision in BCF.

The probability of a single tree structure is $p(T_j) = \prod_{h=1}^{b_j - 1} \alpha(1 + d_h)^{-\beta} \prod_{k=1}^{b_j} (1 - \alpha(1 + d_k)^{-\beta})$, where h indexes the internal nodes of the tree T_j, and k indexes the terminal nodes. Different splitting probabilities are applied to $\mu(x)$ and $\tau(x)$ trees. In particular, $\alpha = 0.95$ and $\beta = 2$ for $\mu(x)$ trees, and $\alpha = 0.25$ and $\beta = 3$ for $\tau(x)$ trees. This regularizes the treatment effect function to a greater extent than the rest of the model.

6.1.3 ps-BART

Hahn et al. (2017) also suggest simply including the estimated propensity score as a potential splitting variable in standard BART, and then using the approach introduced by Hill (2011). It would be of interest to similarly compare BCF-BMA with standard BART-BMA including the estimated propensity score as a potential splitting variable and following the approach used by Hill (2011).

[20] Another common strategy is to separately fit a model on observations for which $z_i = 1$, and on observations for which $z_i = 0$, and let $\hat{\tau}_i$ be the difference in the predictions of these two models

6.2 Outline of BCF-BMA

For BCF (Hahn et al. 2017), we are interested in the model $f(x_i, z_i) = \mu(x_i, \hat{\pi}_i) + \tau(x_i)z_i$, where $\mu(x_i, \hat{\pi}_i)$ and $\tau(x_i)$ are separate sum of tree models.[21] Let $T_{\mu j}$ and $T_{\tau j}$ denote trees in $\mu(x_i, \hat{\pi}_i)$ and $\tau(x_i)$ respectively, and let $M_{\mu j}$ and $M_{\tau j}$ denote the terminal node parameters for $T_{\mu j}$ and $T_{\tau j}$ respectively. The BCF prior can be written as:

$$p(M_{\mu 1}, ..., M_{\mu m_\mu}, T_{\mu 1}, ..., T_{\mu m_\mu}, M_{\tau 1}, ..., M_{\tau m_\tau}, T_{\tau 1}, ..., T_{\tau m_\tau}, \sigma)$$
$$\propto \left[\prod_j \prod_i p(\mu_{ij}|T_{\mu j})p(T_{\mu j})\right] \left[\prod_j \prod_i p(\tau_{ij}|T_{\tau j})p(T_{\tau j})\right] p(\sigma)$$

For BCF-BMA, I suggest placing the prior $\mu_{ij}|T_\mu, \sigma \sim N(0, \frac{\sigma^2}{a_\mu})$ and $\tau_{ij}|T_\tau, \sigma \sim N(0, \frac{\sigma^2}{a_\tau})$. These priors are somewhat different to those proposed by Hahn et al. (2017) who place different priors on the scales of μ_{ij} and τ_{ij}. However, it is possible to directly specify different scales through the choice of a_μ and a_τ. The BCF-BMA prior, like the BART-BMA prior, provides a closed form for the marginal likelihood and a multivariate t-distribution for posterior predictions.

6.2.1 BCF-BMA marginal likelihood

Let $Z = (Z_1, ..., Z_n)$ be the treatment indicator variable. Let $J_{\mu j}$ and $J_{\tau j}$ be matrices denoting inclusion of observations in tree j in $\mu(x)$ and $\tau(x)$ respectively. The BCF-BMA likelihood can be written as:

$$Y|\mathcal{T}_\mu, M_\mu \mathcal{T}_\tau, M_\tau, \sigma^{-2} \sim N\left((\sum_{j=1}^{m_\mu} J_{\mu j} M_{\mu j}) + Diag(Z)(\sum_{j=1}^{m_\tau} J_{\tau j} M_{\tau j}), \sigma^2 I\right)$$

Now, let $W_\mu = [J_{\mu 1} ... J_{\mu m_\mu}]$ (which is a $n \times \omega_\mu$ matrix, where $\omega_\mu = \sum_{j=1}^{m_\mu} b_{\mu j}$), and let $O_\mu = [M_{\mu 1}^T ... M_{\mu m_\mu}^T]^T$ (which is a $\omega_\mu \times 1$ vector). Similarly let $W_\tau = [J_{\tau 1} ... J_{\tau m_\tau}]$ (which is a $n \times \omega_\tau$ matrix, where $\omega_\tau = \sum_{j=1}^{m_\tau} b_{\tau j}$) , and let $O_\tau = [M_{\tau 1}^T ... M_{\tau m_\tau}^T]^T$ (which is a $\omega_\tau \times 1$ vector). Then we can write

$$Y|O_\mu, O_\tau, \sigma^{-2} \sim N\left(W_\mu O_\mu + Diag(Z) W_\tau O_\tau, \sigma^2 I\right)$$

Now let $W_{BCF} = [W_\mu \ (Diag(Z)W_\tau)]$ (which is a $n \times (\omega_\mu + \omega_\tau)$ matrix), and let $O_{BCF} = [O_\mu^T O_\tau^T]^T$ (which is a $(\omega_\mu + \omega_\tau) \times 1$ matrix).

Then $Y|O_{BCF}, \sigma^2 \sim N(W_{BCF} O_{BCF}, \sigma^2 I)$, and we can write the BCF-BMA likelihood as:

$$p(Y|X, \mathcal{T}_\mu, \mathcal{T}_\tau) = \int \int p(Y|O_{BCF}, \sigma^{-2}) p(O) p(\sigma^{-2}) dO d\sigma^{-2}$$

Note that $Y|O_{BCF}, \sigma^{-2} \sim N(W_{BCF} O_{BCF}, \sigma^2 I)$ and the first ω_μ elements of O_{BCF} have independent prior distributions $\mu \sim N(0, \frac{\sigma^2}{a_\mu})$, and the last ω_τ elements of O_{BCF} also have independent normal priors, with different variance, $\tau \sim N(0, \frac{\sigma^2}{a_\tau})$. This implies that $O_{BCF}|\sigma^{-2} \sim N(0, \sigma^2 A^{-1})$ where $A = \begin{pmatrix} a_\mu I_{\omega_\mu} & 0 \\ 0 & a_\tau I_{\omega_\tau} \end{pmatrix}$ is a diagonal matrix with the first ω_μ diagonal elements equal to a_μ, and the next ω_τ diagonal elements equal to a_τ.

Therefore $Y|\sigma^{-2} \sim \mathrm{W_{BCF}} \ \varepsilon_2 + \varepsilon_1$, where $\varepsilon_1 \sim N(0, \sigma^2 I)$ and $\varepsilon_2 \sim N(0, \sigma^2 A^{-1})$. This implies that

[21] The BCF-BMA package allows for the inclusion of zero, one, or more than one set of propensity score estimates as potential splitting variables in the $\mu(x)$ function.

$Y|\sigma^{-2} \sim N(0, \sigma^2(I_n + WA^{-1}W^T))$ and therefore, marginalizing over σ, we obtain

$$Y \sim MVSt_\nu \left(0, \lambda(I_n + WA^{-1}W^T)\right)$$

$$p(Y) = \frac{1}{(det(A))^{-\frac{1}{2}}\, det(I_n + WA^{-1}W^T)^{\frac{1}{2}}} \left[\lambda\nu + Y^TY - Y^TW(A + W^TW)W^TY\right]^{-\frac{\nu+n}{2}}$$

And the log of this expression is:

$$\frac{\omega_\mu}{2}\log(a_\mu) + \frac{\omega_\tau}{2}\log(a_\tau) - \frac{1}{2}\log(det(M)) - \frac{\nu+n}{2}\log\left(\lambda\nu + Y^TY - Y^TWM^{-1}W^TY\right)$$

where $M = A + W^TW$.

6.2.2 BCF-BMA posterior ITE distribution

Let $V = [\mathbf{0}_{n\times\omega_\mu}\ W_\tau]$, where $\mathbf{0}_{n\times\omega_\mu}$ is a matrix of zeros of dimensions equal to those of W_μ. The posterior distribution of $\tau(x)$ is:

$$VO|Y, \mathcal{T}_\mu, \mathcal{T}_\tau \quad \sim \quad MVSt_{\nu+n}\left(VM^{-1}W^TY,\ \frac{1}{\nu+n}[\nu\lambda + Y^TY - Y^TWM^{-1}W^TY]VM^{-1}V^T\right)$$

where $M = A + W^TW$. For out of sample predictions, replace V with $\tilde{V} = [\mathbf{0}_{\tilde{n}\times\omega_\mu}\ \tilde{W}_\tau]$.

6.2.3 BCF-BMA CATE Posterior Distribution

The posterior distribution of $\tau(x)$ given in the previous subsection is the posterior distribution of what is often referred to as the Individual Treatment Effect (ITE). However, $\tau(x)$ can also be referred to as the Conditional Average Treatment Effect (CATE) Function. In this paper, the term CATE refers to the expectation of the average of the ITEs, i.e. $\frac{1}{n}\sum_{i=1}^n \tau(x)$.

The posterior distribution of $\frac{1}{n}\sum_{i=1}^n \tau(x)$ for a given model in BCF-BMA is:

$$\frac{1}{n}\mathbf{1}^TVO|Y, \mathcal{T}_\mu, \mathcal{T}_\tau \sim$$
$$MVSt_{\nu+n}\left(\frac{1}{n}\mathbf{1}^TVM^{-1}W^TY\ ,\ \frac{1}{\nu+n}[\nu\lambda + Y^TY - Y^TWM^{-1}W^TY]\frac{1}{n}\mathbf{1}^TVM^{-1}V^T\frac{1}{n}\mathbf{1}\right)$$

where $M = A + W^TW$ and $\mathbf{1}$ is a vector of 1s of length n. Note that this is a univariate t-distribution with location and scale. For out of sample predictions, replace V with $\tilde{V} = [\mathbf{0}_{\tilde{n}\times\omega_\mu}\ \tilde{W}_\tau]$ and replace $\frac{1}{n}\mathbf{1}$ with $\frac{1}{\tilde{n}}\tilde{\mathbf{1}}$, where $\tilde{\mathbf{1}}$ is a vector of 1s of length $\tilde{n}$.

6.3 Description of the BCF-BMA algorithm

The BCF-BMA model search algorithm is similar to the improved BART-BMA algorithm, except in each round either $\mu(x)$ trees or $\tau(x)$ trees can be appended to existing models.[22] The model selection criterion is the posterior model probability.

[22] An option is also provided in the BCF-BMA package for adding a mu tree, then a tau tree, and then a mu tree, and so on in an alternating sequence.

In constructing $\tau(x)$ trees to be potentially appended to the model, potential splitting points are selected from a changepoint detection algorithm applied to treated observations only. [23]

Pseudocode for the BCF-BMA algorithm is given in Appendix E.

6.4 BCF-IS

The BCF-IS algorithm is the algorithm outlined in section 4.1, with some adjustments. $\mu(x)$ trees and $\tau(x)$ trees are drawn from separate priors. The marginal likelihood calculated is that described in 6.2.1. The output is the average of (and credible intervals for) the treatment effect, not the outcome.

For the standard BART prior, different priors can be applied to $\mu(x)$ trees and $\tau(x)$ trees as described for BCF-BMA. Similarly, for the prior described by Quadrianto & Ghahramani (2014), different splitting probabilities can be specified for $\mu(x)$ trees and $\tau(x)$ trees. For the Spike and Tree prior (Rockova & van der Pas 2017), different prior parameters can be specified for the Poisson distribution for the number of terminal nodes, and different hyperparameters can be specified for the beta hyperprior distribution on the variable inclusion probabilities.

6.5 BCF-BMA results

6.5.1 Simulation from bcf R package

In this section I present a brief comparison of BCF-BMA and standard BCF (Hahn et al. 2017) results obtained using a simulation example from the **bcf** package in **R**.

Let n be the number of observations, and p be the number of variables. Let $x_1, ..., x_p$ be the set of covariates.

Let

$$\mu(x) = -\mathbb{I}\{x_1 > x_2\} + \mathbb{I}\{x_1 < x_2\}$$

where $\mathbb{I}$ is an indicator function. Let the probability of treatment be $\pi(x) = \Phi(\mu(x))$. i.e. there is confounding. The treatment variable is Z.

Let the treatment effect function be

$$\tau(x) = 0.5\mathbb{I}\{x_3 > -0.75\} + 0.25\mathbb{I}\{x_3 > 0\} + 0.25\mathbb{I}\{x_3 > 0.75\}$$

and the outcome equals

$$Y = \mu(x) + \tau(x) + \epsilon$$

where $\epsilon \sim \mathcal{N}(0, \sigma)$ and $\sigma = max(\mu(x_i) + \tau(x_i)\pi(x_i)) - min(\mu(x_i) + \tau(x_i)\pi(x_i))$. Suppose the propensity score estimates are exact, i.e. the true propensities are known $\hat{\pi}(x) = \pi(x)$.

The results for one simulation of the data generating process outlined above with $n = 250$ and $p = 3$ are included in Figure 3. It can be observed that BCF and BCF-BMA yield similar predictions. BCF-BMA has the added advantage that the output contains a small number of models, and each model (under the default settings) contains only 5 $\mu(x)$ trees and 5 $\tau(x)$ trees, each of which contains relatively few splits.[24] Therefore it is possible to directly observe the important splitting variables and splitting points.

[23] Another option, provided in the **R** package **bcfbma** available at https://github.com/EoghanONeill/bcfbma, is to apply the changepoint detection algorithm to all a Horowitz-Thompson transformation of all residuals

[24] The maximum number of splits under the default settings for BCF-BMA is 5 per tree.

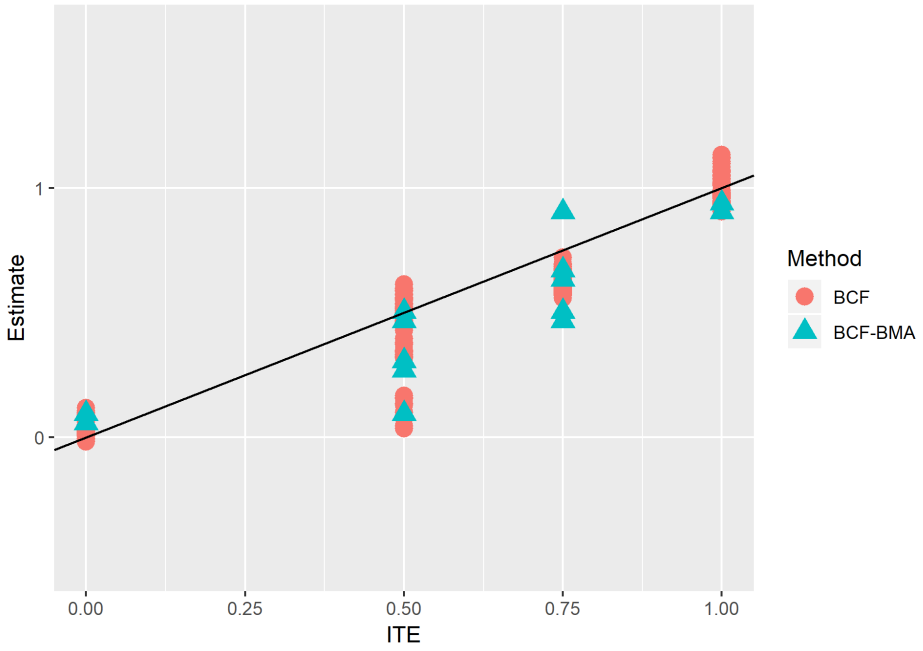

Figure 3: Example results for BCF and BCF-BMA

6.5.2 Simulations used by Hahn et al. (2017)

Hahn et al. (2017) simulate the following eight data generating processes, corresponding to the various combinations of three two-level settings: homogeneous versus heterogeneous treatment effects, a linear versus nonlinear conditional expectation function, and two different sample sizes ($n = 250$ and $n = 500$).

Five variables comprise x; the first three are continuous, drawn as standard normal random variables, the fourth is a dichotomous variable and the fifth is unordered categorical, taking three levels (denoted 1,2,3). The treatment effect is either $\tau(x) = 3$ (i.e. homogenous) or $\tau(x) = 1 + 2x_2x_5$ (i.e. heterogeneous). The prognostic function is either $\mu(x) = 1 + g(x_4) + x_1x_3$ (linear) or $\mu(x) = -6 + g(x_4) + 6|x_3 - 1|$ (nonlinear) where $g(1) = 2$, $g(2) = -1$, and $g(3) = -4$, and the propensity function is

$$\pi(x_i) = 0.8\Phi(3\mu(x_i)/s - 0.5x_1) + 0.05 + u_i/10$$

where s is the standard deviation of μ taken over the observed sample and $u_i \sim Uniform(0, 1)$.

Comparisons with other methods are made in tables 6 to 9 in terms of RMSE, coverage, and interval length for Average Treatment Effect (ATE) and Conditional Average Treatment Effect (CATE) estimates.

Tables 6 to 9 contain results for $n = 250$. The set of methods includes: BCF,[25] BCF-BMA without updates of potential splitting varaibles within trees (BCF-BMA 1), BCF-BMA with updates of potential splitting variables within trees (BCF-BMA 2), BART-BMA [26], BART,[27] BCF-IS,[28], BART-IS, [29] and

[25]Implemented with the **R** package **bcf** with default parameter values.
[26]See section 3.3
[27]Implemented with the **R** package **BART** with default parameter values.
[28]Implemented with 100,000 draws of models from the importance sampler. Each model contains 50 $\mu(x)$ trees and 25 $\tau(x)$ trees.
[29]Implemented with 100,000 draws of models from the importance sampler, each model includes 30 trees, with the propensity score and treatment included as potential splitting variables. See section 3.3 for details on the posterior distributions for individual models.

(non-Bayesian) causal forests (Athey et al. 2017, Wager & Athey 2017). [30]

For all methods except causal forest, the propensity score is estimated using the function **pbart** in the **R** package **BART**. [31]

	ATE			ITE		
	rmse	coverage	length	rmse	coverage	length
BCF	0.08	0.55	0.16	0.29	0.93	0.39
BCF-BMA 1	0.39	0.47	0.76	0.55	0.68	0.97
BCF-BMA 2	0.20	0.88	0.67	0.28	0.88	0.79
BART-BMA	0.37	0.62	0.83	0.53	0.76	1.37
BART	0.07	0.73	0.20	0.19	0.92	0.53
BCF-IS	0.26	0.83	0.72	0.30	0.97	1.25
BART-IS	0.25	0.83	0.76	0.32	0.93	1.18
CF	0.40	0.47	0.83	0.53	0.58	1.02

Table 6: Hahn et al. (2018) simulations, $\tau(x) = 3$, $\mu(x) = 1 + g(x_4) + x_1 x_3$, $n = 250$, 200 replications.

	ATE			ITE		
	rmse	coverage	length	rmse	coverage	length
BCF	0.10	0.70	0.22	0.48	0.97	0.50
BCF-BMA 1	0.20	0.98	1.15	0.74	0.86	1.68
BCF-BMA 2	0.20	0.97	1.14	0.73	0.86	1.62
BART-BMA	0.39	0.76	1.24	1.01	0.82	2.00
BART	0.08	0.83	0.30	0.29	0.95	0.79
BCF-IS	0.24	1.00	1.14	0.32	1.00	1.74
BART-IS	0.12	1.00	1.16	0.26	1.00	2.38
CF	0.57	0.38	1.06	0.62	0.57	1.29

Table 7: Hahn et al. (2018) simulations, $\tau(x) = 3$, $\mu(x) = -6 + g(x_4) + 6|x_3 - 1|$, $n = 250$, 200 replications.

	ATE			ITE		
	rmse	coverage	length	rmse	coverage	length
BCF	0.11	0.67	0.26	0.55	0.76	0.89
BCF-BMA 1	0.44	0.41	0.81	1.24	0.40	1.27
BCF-BMA 2	0.28	0.63	0.65	0.92	0.56	1.11
BART-BMA	0.48	0.41	0.86	1.34	0.33	1.10
BART	0.11	0.71	0.31	0.52	0.82	1.05
BCF-IS	0.31	0.76	0.83	1.12	0.67	1.81
BART-IS	0.24	0.89	0.84	1.11	0.74	2.07
CF	0.44	0.56	0.99	1.20	0.48	1.48

Table 8: Hahn et al. (2018) simulations, $\tau(x) = 1 + 2x_2 x_5$, $\mu(x) = 1 + g(x_4) + x_1 x_3$, $n = 250$, 200 replications.

[30] Causal forests are estimated using the **R** package **grf** and 4000 trees.

[31] Results obtained by using the true propensity score used instead of an estimated propensity score are presented in Appendix A.

	ATE			ITE		
	rmse	coverage	length	rmse	coverage	length
BCF	0.11	0.75	0.32	0.75	0.78	1.05
BCF-BMA 1	0.26	0.94	1.16	1.29	0.57	1.81
BCF-BMA 2	0.23	0.95	1.13	1.23	0.63	1.82
BART-BMA	0.40	0.66	0.96	1.57	0.42	1.23
BART	0.11	0.86	0.39	0.64	0.82	1.32
BCF-IS	0.25	0.99	1.24	1.28	0.74	2.33
BART-IS	0.16	0.99	1.18	1.20	0.80	2.64
CF	0.60	0.42	1.19	1.26	0.53	1.73

Table 9: Hahn et al. (2018) simulations, $\tau(x) = 1 + 2x_2x_5$, $\mu(x) = -6 + g(x_4) + 6|x_3 - 1|$, $n = 250$, 200 replications.

The results suggest that standard BART generally performs best in terms of RMSE, followed by BCF, although BART-IS and BCF-IS also perform well. [32] In some cases the coverage of credible intervals, particularly for the ATE, is better for the new algorithms described in this paper than for BART or BCF, although it should be noted that the 100% or nearly 100% coverage observed, for example in Table 7 is not desirable, and the prediction intervals for the new methods are notably wider than those of BART and BCF.

The RMSE of ITE estimates for simulations with heterogeneous treatment effects is worse for BCF-BMA than for BART and BCF. This is expected because the default setting for BCF-BMA are 5 $\mu(x)$ trees, and 5 $\tau(x)$ trees, each of which has a maximum of 5 splits. Therefore the estimates are less heterogeneous than those produced by BART and BCF with many trees. However, the relatively small set of simpler models averaged by BCF-BMA is more interpretable and still performs reasonably well, particularly for ATE estimation.

6.5.3 Data Challenge Datasets

The annual Atlantic Causal Inference Conference (ACIC) has run a data analysis competition for treatment effect estimation methods. BART and BCF have performed well in this competition (Dorie et al. 2019, Hahn et al. 2019).

Table 10 presents a comparison between BCF, BCF-IS, BART-IS, BART, and CF applied to the publicly available data from the 2019 ACIC Data Challenge.[33] The results are restricted to the 1200 datasets in the low-dimensional category with less than 1000 observations and a continuous dependent variable.[34]. In all cases the estimates and intervals are produced for $\frac{1}{N}\sum_{i=1}^{N}\tau(x_i)$, and the RMSE and coverage are calculated using the true population ATE. BCF-IS attains the lowest RMSE. BART-IS achieves the most accurate coverage of prediction intervals.

[32] The performance of BART-IS and BCF-IS improves with the number of samples drawn. There is therefore a trade-off between computational time and accuracy, although this is less of an issue when the draws are parallelized across many threads. The extent to which the results would improve with a greater number of draws is a potential topic for future research.

[33] Results are not presented for BCF-BMA or BART-BMA, because the current implementations can require a large quantity of RAM.

[34] The current implementations of BART-IS and BCF-IS are slow when applied to datasets with many observations. The methods are also only currently designed for data with a continuous dependent variable.

	ATE		
	rmse	coverage	length
BCF	0.18	0.88	0.67
BCF-IS	0.17	0.91	0.69
BART-IS	0.19	0.95	0.93
BART	0.23	0.93	0.99
CF	0.22	0.93	1.01

Table 10: Results for ACIC Data challenge low-dimensional datasets with less than 1000 observations and a continuous dependent variable.

7 Conclusion

Many MCMC implementations of BART have been demonstrated to be effective in a variety of applications. In this paper, I improve an alternative BART implementation, BART-BMA (Hernández et al. 2018), and describe an extension to treatment effect estimation, BCF-BMA. Unlike other BART and BCF implementations, BART-BMA and BCF-BMA are entirely deterministic.

I also describe BART-IS and BCF-IS, which, in notable contrast with BART-BMA and BCF-BMA, do not involve any model search, but rather are implementations that involve importance sampling from a data independent model prior.

BART-IS and BCF-IS provide a simple framework for implementing BART, but the sampling scheme is unlikely to be as effective as MCMC methods, despite the marginalization and possibly safe-Bayesian approach. However, the framework allows for simple testing of different priors and other variations.

Interesting topics for further research include faster implementations of BART-IS, multivariate BART-IS, [35] BART-IS plus a linear model, [36] BART Monte Carlo Markov Chain Model Composition, and Bayesian stacking of sum-of-tree models.

[35] See appendix B.
[36] See appendix C.

A Supplementary Simulation Results

A.1 Hahn et al. (2017) Simulations, $n = 500$

	ATE			ITE		
	rmse	coverage	length	rmse	coverage	length
BCF	0.04	0.42	0.06	0.21	0.93	0.15
BCF-BMA 1	0.29	0.29	0.46	0.53	0.64	0.69
BCF-BMA 2	0.11	0.84	0.37	0.28	0.83	0.54
BART-BMA	0.25	0.58	0.52	0.37	0.69	0.87
BART	0.04	0.63	0.10	0.15	0.94	0.29
BCF-IS	0.21	0.48	0.41	0.25	0.82	0.71
BART-IS	0.22	0.43	0.43	0.31	0.69	0.72
CF	0.33	0.29	0.54	0.51	0.52	0.81

Table 11: Hahn et al. (2018) simulations, $\tau(x) = 3$, $\mu(x) = 1 + g(x_4) + x_1 x_3$, $n = 500$, 200 replications.

	ATE			ITE		
	rmse	coverage	length	rmse	coverage	length
BCF	0.04	0.64	0.09	0.32	0.97	0.17
BCF-BMA 1	0.15	0.90	0.64	0.89	0.59	1.09
BCF-BMA 2	0.15	0.94	0.64	0.85	0.61	1.12
BART-BMA	0.35	0.72	0.76	0.84	0.78	1.31
BART	0.04	0.86	0.14	0.20	0.97	0.37
BCF-IS	0.15	0.96	0.63	0.25	0.97	0.98
BART-IS	0.07	1.00	0.58	0.24	0.99	1.42
CF	0.22	0.52	0.51	0.35	0.83	0.76

Table 12: Hahn et al. (2018) simulations, $\tau(x) = 3$, $\mu(x) = -6 + g(x_4) + 6|x_3 - 1|$, $n = 500$,200 replications.

	ATE			ITE		
	rmse	coverage	length	rmse	coverage	length
BCF	0.04	0.63	0.10	0.33	0.71	0.40
BCF-BMA 1	0.33	0.29	0.52	1.16	0.31	0.92
BCF-BMA 2	0.19	0.51	0.37	0.70	0.54	0.81
BART-BMA	0.42	0.16	0.58	1.27	0.25	0.73
BART	0.05	0.78	0.15	0.34	0.81	0.63
BCF-IS	0.25	0.53	0.51	1.02	0.53	1.19
BART-IS	0.19	0.70	0.50	1.02	0.60	1.33
CF	0.34	0.50	0.68	0.99	0.51	1.28

Table 13: Hahn et al. (2018) simulations, $\tau(x) = 1+2x_2x_5$, $\mu(x) = 1+g(x_4)+x_1x_3$, $n = 500$, 200 replications.

	ATE			ITE		
	rmse	coverage	length	rmse	coverage	length
BCF	0.05	0.71	0.11	0.43	0.71	0.42
BCF-BMA 1	0.17	0.89	0.69	1.30	0.37	1.26
BCF-BMA 2	0.17	0.90	0.63	1.00	0.56	1.32
BART-BMA	0.36	0.66	0.65	1.53	0.40	0.87
BART	0.04	0.83	0.18	0.39	0.81	0.75
BCF-IS	0.16	0.92	0.70	1.23	0.50	1.42
BART-IS	0.10	0.99	0.65	1.14	0.57	1.63
CF	0.23	0.71	0.66	0.87	0.64	1.30

Table 14: Hahn et al. (2018) simulations, $\tau(x) = 1 + 2x_2x_5$, $\mu(x) = -6 + g(x_4) + 6|x_3 - 1|$, $n = 500$, 200 replications.

A.2 Hahn et al. (2017) Simulations, True Propensity Scores

Tables 15 to 18 present the results for the simulations outlined in section 6.5.2, but with the known propensity score included as a covariate instead of an estimated propensity score being included as a covariate.

	ATE			ITE		
	rmse	coverage	length	rmse	coverage	length
BCF	0.06	0.67	0.14	0.26	0.93	0.35
BCF-BMA 1	0.12	0.99	0.70	0.28	0.94	0.84
BCF-BMA 2	0.09	1.00	0.65	0.18	0.97	0.74
BART-BMA	0.13	0.96	0.76	0.32	0.96	1.32
BART	0.05	0.88	0.19	0.16	0.95	0.50
BCF-IS	0.09	1.00	0.71	0.17	1.00	1.22
BART-IS	0.09	1.00	0.75	0.13	1.00	1.03

Table 15: Hahn et al. (2018) simulations, $\tau(x) = 3$, $\mu(x) = 1 + g(x_4) + x_1x_3$, $n = 250$, 200 replications.

	ATE			ITE		
	rmse	coverage	length	rmse	coverage	length
BCF	0.08	0.74	0.21	0.38	0.97	0.48
BCF-BMA 1	0.17	0.99	1.09	0.68	0.88	1.56
BCF-BMA 2	0.21	0.97	1.09	0.65	0.90	1.52
BART-BMA	0.31	0.86	1.24	0.85	0.86	2.04
BART	0.06	0.89	0.27	0.23	0.97	0.71
BCF-IS	0.14	1.00	1.13	0.26	1.00	1.73
BART-IS	0.14	1.00	1.13	0.26	1.00	1.73

Table 16: Hahn et al. (2018) simulations, $\tau(x) = 3$, $\mu(x) = -6 + g(x_4) + 6|x_3 - 1|$, $n = 250$, 200 replications.

	ATE			ITE		
	rmse	coverage	length	rmse	coverage	length
BCF	0.08	0.79	0.26	0.52	0.76	0.85
BCF-BMA 1	0.16	0.91	0.77	1.10	0.43	1.18
BCF-BMA 2	0.14	0.91	0.63	0.82	0.63	1.08
BART-BMA	0.15	0.94	0.78	1.26	0.42	1.05
BART	0.08	0.90	0.30	0.50	0.84	1.04
BCF-IS	0.14	0.96	0.81	1.02	0.75	1.76
BART-IS	0.13	0.99	0.80	0.98	0.80	2.04

Table 17: Hahn et al. (2018) simulations, $\tau(x) = 1+2x_2x_5$, $\mu(x) = 1+g(x_4)+x_1x_3$, $n = 250$, 200 replications.

	ATE			ITE		
	rmse	coverage	length	rmse	coverage	length
BCF	0.11	0.78	0.32	0.73	0.77	1.04
BCF-BMA 1	0.21	0.97	1.11	1.25	0.55	1.74
BCF-BMA 2	0.22	0.95	1.09	1.17	0.65	1.81
BART-BMA	0.40	0.71	1.00	1.52	0.45	1.34
BART	0.10	0.87	0.37	0.60	0.83	1.31
BCF-IS	0.21	0.99	1.23	1.26	0.75	2.36
BART-IS	0.14	1.00	1.16	1.17	0.81	2.65

Table 18: Hahn et al. (2018) simulations, $\tau(x) = 1 + 2x_2x_5$, $\mu(x) = -6 + g(x_4) + 6|x_3 - 1|$, $n = 250$, 200 replications.

	ATE			ITE		
	rmse	coverage	length	rmse	coverage	length
BCF	0.03	0.63	0.06	0.18	0.92	0.14
BCF-BMA 1	0.10	0.905	0.42	0.31	0.86	0.57
BCF-BMA 2	0.06	0.97	0.36	0.19	0.90	0.46
BART-BMA	0.11	0.94	0.46	0.20	0.95	0.79
BART	0.03	0.84	0.09	0.11	0.95	0.26
BCF-IS	0.07	0.99	0.38	0.16	0.98	0.72
BART-IS	0.07	1.00	0.42	0.10	0.99	0.54

Table 19: Hahn et al. (2018) simulations, $\tau(x) = 3$, $\mu(x) = 1 + g(x_4) + x_1x_3$, $n = 500$, 200 replications.

	ATE			ITE		
	rmse	coverage	length	rmse	coverage	length
BCF	0.04	0.60	0.08	0.30	0.97	0.18
BCF-BMA 1	0.12	0.93	0.58	0.60	0.82	0.94
BCF-BMA 2	0.19	0.84	0.61	0.74	0.71	1.06
BART-BMA	0.28	0.83	0.75	0.72	0.81	1.34
BART	0.03	0.85	0.12	0.16	0.97	0.32
BCF-IS	0.10	0.99	0.63	0.19	0.99	0.96
BART-IS	0.07	1.00	0.58	0.21	1.00	1.38

Table 20: Hahn et al. (2018) simulations, $\tau(x) = 3$, $\mu(x) = -6 + g(x_4) + 6|x_3 - 1|$, $n = 500$, 200 replications.

	ATE			ITE		
	rmse	coverage	length	rmse	coverage	length
BCF	0.03	0.79	0.10	0.32	0.70	0.39
BCF-BMA 1	0.11	0.95	0.51	1.07	0.32	0.88
BCF-BMA 2	0.10	0.90	0.36	0.62	0.58	0.76
BART-BMA	0.12	0.91	0.50	1.18	0.23	0.68
BART	0.04	0.88	0.14	0.33	0.81	0.62
BCF-IS	0.11	0.93	0.49	1.00	0.59	1.17
BART-IS	0.09	0.96	0.48	0.94	0.66	1.34

Table 21: Hahn et al. (2018) simulations, $\tau(x) = 1+2x_2x_5$, $\mu(x) = 1+g(x_4)+x_1x_3$, $n = 500$, 200 replications.

	ATE			ITE		
	rmse	coverage	length	rmse	coverage	length
BCF	0.05	0.70	0.11	0.42	0.72	0.43
BCF-BMA 1	0.15	0.90	0.65	1.22	0.36	1.15
BCF-BMA 2	0.21	0.77	0.61	0.96	0.61	1.27
BART-BMA	0.38	0.59	0.64	1.51	0.38	0.87
BART	0.04	0.86	0.17	0.37	0.81	0.72
BCF-IS	0.17	0.92	0.72	1.22	0.53	1.49
BART-IS	0.09	0.99	0.64	1.15	0.57	1.62

Table 22: Hahn et al. (2018) simulations, $\tau(x) = 1 + 2x_2x_5$, $\mu(x) = -6 + g(x_4) + 6|x_3 - 1|$, $n = 500$, 200 replications.

B Multivariate BART-IS

For multivariate BART-IS, options include the use of the same tree structures for different outcomes (as in shared Bayesian Forests (Linero et al. 2019)), or different tree structures for each outcome (as in BART for Seemingly Unrelated Regression (Chakraborty 2016)). [37]

Let the vector of d outcomes for individual i be denoted by $\boldsymbol{y}_i = (y_{i,1}, ..., y_{i,d})^T$. Then, if we impose the same tree structure on the model for all outcomes, we have

$$\boldsymbol{y}_i = \begin{bmatrix} O_1^T \\ O_2^T \\ \vdots \\ O_d^T \end{bmatrix} (W)_i^T + \begin{bmatrix} \varepsilon_{i,1} \\ \varepsilon_{i,2} \\ \vdots \\ \varepsilon_{i,d} \end{bmatrix}$$

Where $O_1, ..., O_d$ are distinct terminal node coefficient vectors for each outcome, $(W)_i$ is the i^{th} row of the W matrix of terminal node indicator variables, and $\varepsilon_{i,j}$ is the error for individual i, outcome j.

Alternatively, one can allow for distinct sets of tree structures for each outcome, with corresponding matrices of terminal node indicator variables $W_1, ..., W_d$. It is also possible for splits in each sum-of-tree model to be constructed from different sets of potential splitting variables.

This gives the following model:

$$\boldsymbol{y}_i = \begin{bmatrix} O_1^T & 0 & 0 & \dots & 0 \\ 0 & O_2^T & 0 & \dots & 0 \\ 0 & 0 & O_3^T & & \vdots \\ \vdots & \vdots & & \ddots & 0 \\ 0 & 0 & \dots & 0 & O_d^T \end{bmatrix} \begin{bmatrix} (W_1)_i^T \\ (W_2)_i^T \\ \vdots \\ (W_d)_i^T \end{bmatrix} + \begin{bmatrix} \varepsilon_{i,1} \\ \varepsilon_{i,2} \\ \vdots \\ \varepsilon_{i,d} \end{bmatrix}$$

The tree drawing process for any tree is the same as in univariate BART-IS. There exist priors that give a closed form for the marginal likelihood and posterior predictive distribution (Minka 2000).

C Importance sampling of BART plus a linear model

It is straightforward to average over models that are defined by the addition of a linear combination of covariates and a sums-of-trees. First, define a set of covariates that can be included in the linear model, then define prior inclusion probabilities for these covariates [such that the prior is independent of the prior over the sum-of-trees], and priors on the coefficients of included covariates. The prior for coefficients should allow for conjugacy of the whole model, for example, the prior variance of coefficients can be set equal to a scalar multiple of the variance of the error terms.

Then, for each sampled model, we sample the variables included in the linear part of the model by a set of Bernoulli draws, and this gives a covariate matrix X. Then draw the sum-of-trees part of the model as in standard BART-IS, which gives a matrix W and define the overall model matrix as $[X\ W]$.

The resulting models are standard Bayesian linear regressions, and the marginal likelihoods and predictive distributions have closed forms. Importance sampling of BART plus a linear model can be viewed as a

[37] Code has not yet been written for Multivariate BART-IS. This appendix only outlines the idea.

combination of BART-IS and the implementation of BMA of Bayesian linear regressions used by Sala-i Martin et al. (2004).

D Spike-and-Tree prior

D.1 Definition of Spike-and-Tree prior

Results are presented in this paper for BART-BMA with the spike and tree prior described by Rockova & van der Pas (2017) (as an alternative to the standard BART splitting prior). The prior is defined as follows:

For $\alpha, q, c > 0$

$$\pi(\mathcal{S}|q) = \frac{1}{\binom{p}{q}}$$

This prior can be implemented by taking a Bernoulli draw for inclusion of each variable, with a conjugate beta prior distribution on the splitting probability. (i.e. the number of splitting variables can be given a beta-binomial distribution). A drawn variable is used at least once in the tree.

A Poisson prior is placed on the number of terminal nodes

$$\pi(k) = \frac{\lambda_0^k}{(e^{\lambda_0} - 1)k!} \quad , \quad k = 1, 2, ...$$

for some $\lambda_0 \in \mathbb{R}$. However, this should be $\pi(k|q)$ and truncated from the left and to the right so that $q \leq k \leq n-1$, where right truncation only occurs with a date-informed prior that requires every terminal node contains at least one observation.

Then, given $q, \mathcal{S}, k$, assign a uniform prior over valid tree topologies $\mathcal{T} = \{\Omega_k\}_{k=1}^K \in \mathcal{V}_c^k$. A valid tree topology must have some minimum number of training observations in each terminal node.

$$\pi(\mathcal{T}|\mathcal{S}, k) = \frac{1}{\Delta(\mathcal{V}_s^k)} \mathbb{I}(\mathcal{T} \in \mathcal{V}_s^k)$$

The number of possible valid tree constructions is $S(k-1, q)q!(n-1)!/(n-k)!$, where $S(k-1, q)$ is a Sterling number of the second kind. This can be used to account for duplications of the same tree by multiple possible tree constructions in the BART-BMA algorithm. The number of valid tree diagrams is equal to $C_{k-1}q!S(k-1, q)\binom{n-1}{k-1}$, where C_{k-1} is the $k-1^{th}$ Catalan number.

D.2 Sampling from the spike and tree prior

1. Bernoulli draws on the set of included variables. Obtain a set of variables, $\mathcal{S}$, with $|\mathcal{S}| = q$.

2. Draw number of terminal nodes, k from a Poisson distribution truncated on the left (if we require that the tree splits on each variable in $|\mathcal{S}|$ at least once) and right such that $q \leq K - 1 \leq n$, i.e. $q + 1 \leq K \leq n + 1$.

3. Draw a tree structure with the specified number of terminal nodes uniformly at random. This is an efficient algorithm created by Bacher (2016). This gives a representation of the tree structure.

4a. (If using a data-independent prior) Take a standard uniform draw for each split point. Then loop through splitting points, and adjust splitting points within the corresponding sub-tree that split on the same variable again such that it is possible for observations to fall in any terminal node.

4b. (If using the data-dependent prior) Draw a set of splitting points from the $n-1$ possible splits of the variables. Here, the splits are splits of the n observations (i.e. still in one dimension, we haven't allocated splits to the variables yet. Each split "point" just specifies the number of observations that are to the left of that split. Note that for each of these split "points" there is a possible split on each variable).

Fill in the splits in the tree. Apply the following algorithm:

While there are split points remaining:

(a) Take the lowest remaining split point.

(b) Allocate it to the leftmost remaining internal node.

(c) Remove the split point and internal node.

5. For each internal node, randomly draw a splitting variable from $\mathcal{S}$. There will be one split point on the chosen variable that results in the number of observations to the left allocated to that split in step 5.

If we want to apply the condition that each of the $|S|$ potential splitting variables must be used at least once, then we can first draw from all possible variables $(K-1)-|S|$ times with replacement, but then start restricting the number of possible draws, i.e. draw $|S|$ times without replacement. Then randomly shuffle the splitting variables among the splitting points. [An alternative would be any algorithm that creates random (ordered) partitions of the $K-1$ splitting points among the $|\mathcal{S}|$ splitting variables.]

E BCF-BMA Algorithm

.

Input: $n \times p$ matrix X with continuous response variable Y
Output: RMSE, Credible interval for $\hat{Y}$, after burn-in updates for σ
Initialise: $Tree\_Response = Y\_scaled$;
Initialise: $lowest\_BIC$, $L = 1$, Set of $\mathcal{T} = List\_ST$ = a tree stump
Initialise: $count\_mu\_trees_\ell = 1$, $count\_tau\_trees_\ell = 1$
for $j \leftarrow 1$ *to* $m_\mu + m_\tau$ **do**
  for $\ell \leftarrow 1$ *to* L **do**
    if $count\_mu\_trees_\ell \leq m_\mu$ **then**
      1. **Find Good Splitting Rules**. Run greedy search to find $numcp\%_\mu$ best split rules for each current sum of trees $\mathcal{T}_{\mu\ell}$ in $\mathcal{T}_\ell$ in Occam's window, using the partial residuals of $\mathcal{T}_\ell$ as $Tree\_response$.
      2. **Grow greedy trees based on their partial residuals to append to current sum of trees model** $\mathcal{T}_{\mu\ell}$. Set new proposal tree T^* to stump

      for $H \leftarrow 1$ *to* $max\_tree\_depth_\mu$ **do**
        for $i \leftarrow 1$ *to number of terminal nodes in* T^* **do**
          for $d \leftarrow 1$ *to* $num\_split\_rules_\mu$ **do**
            Grow proposal tree T^* using splitting rule d from list of splitting rules found in part 1. Append T^* to $\mathcal{T}_{\mu\ell}$ to make new sum of trees model $\mathcal{T}_\ell^*$. **if** *Sum of trees* $\mathcal{T}_\ell^*$ *is in Occam's window* **then**
              Append T^* to $\mathcal{T}_{\mu\ell}$ and save new sum of trees model to temporary list $temp_{OW}$, and save new values of $count\_mu\_trees := count\_mu\_trees_\ell + 1$, and $count\_tau\_trees := count\_tau\_trees_\ell$ for each element of $temp_{OW}$ in lists $temp\_count\_mu\_list$ and $temp\_count\_tau\_list$.
            end
          end
        end
      end
    end
    if $j \leq count\_tau\_trees_\ell$ **then**
      1. **Find Good Splitting Rules**. Run greedy search to find $numcp\%_\tau$ best split rules for each current sum of trees $\mathcal{T}_{\tau\ell}$ in $\mathcal{T}_\ell$ in Occam's window, using the partial residuals of $\mathcal{T}_\ell$ *for treated individuals only* as $Tree\_response$.
      2. **Grow greedy trees based on their *treated individuals'* partial residuals to append to current sum of trees** $\mathcal{T}_{\tau\ell}$. Set new proposal tree T^* to stump

      for $H \leftarrow 1$ *to* $max\_tree\_depth_\tau$ **do**
        for $i \leftarrow 1$ *to number of terminal nodes in* T^* **do**
          for $d \leftarrow 1$ *to* $num\_split\_rules_\tau$ **do**
            Grow proposal tree T^* using splitting rule d from list of splitting rules found in part 1. Append T^* to $\mathcal{T}_{\tau\ell}$ to make new sum of trees model $\mathcal{T}_\ell^*$. **if** *Sum of trees* $\mathcal{T}_\ell^*$ *is in Occam's window* **then**
              Append T^* to $\mathcal{T}_{\tau\ell}$ and add new sum of trees model to temporary list $temp_{OW}$, and save a new value of $count\_mu\_trees := count\_mu\_trees_\ell$, and $count\_tau\_trees := count\_tau\_trees_\ell + 1$ for each element of $temp_{OW}$ in lists $temp\_count\_mu\_list$ and $temp\_count\_tau\_list$.
            end
          end
        end
      end
    end
    Make sum of trees models and update residuals
    List of sum of trees models to grow further $List\_ST = temp_{OW}$
    List of all sum of trees models to date $sum\_trees\_in\_OccamsWindow+ = temp_{OW}$
    Lists of counts of mu trees and tau trees in all sum of tree models to date $count\_mu\_trees+ = temp\_count\_mu\_list$, $count\_tau\_trees+ = temp\_count\_tau\_list$.
    Update $lowest\_BIC = min(sum\_trees\_in\_OccansWindow)$
    Set $L = length(temp_{OW})$
    Set $length(temp_{OW}) = 0$
  end
end
Get total list of L sum of trees in Occam's window by deleting models from $sum\_trees\_in\_OccamsWindow$ list whose BIC is greater than $\log(o)$ from $lowest\_BIC$
$\hat{\tau}$ = **Sum of weighted predictions** $\hat{\tau}_\ell$ **over all** L **sum of trees models in Occam's window**
For prediction intervals, obtain quantiles by a root finding algorithm (or implement a post hoc Gibbs Sampler for each sum of trees accepted in Occam's window)
return:
Credible intervals for $\hat{\tau}$; Sum of trees in Occam's window;
Posterior probability of each sum of trees model.

Algorithm 1: BCF-BMA Algorithm

References

Agarwal, R., Ranjan, P. & Chipman, H. (2014), 'A new bayesian ensemble of trees approach for land cover classification of satellite imagery', *Canadian Journal of Remote Sensing* **39**(6), 507–520.

Alaa, A. M. & van der Schaar, M. (2018), 'Bayesian nonparametric causal inference: Information rates and learning algorithms', *IEEE Journal of Selected Topics in Signal Processing* **12**(5), 1031–1046.

Athey, S. (2015), Machine learning and causal inference for policy evaluation, *in* 'Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining', ACM, pp. 5–6.

Athey, S. (2018), The impact of machine learning on economics, *in* 'The economics of artificial intelligence: An agenda', University of Chicago Press.

Athey, S., Tibshirani, J. & Wager, S. (2017), 'Generalized random forests', *arXiv preprint arXiv:1610.01271* .

Bacher, A. (2016), 'A new bijection on m-dyck paths with application to random sampling', *arXiv preprint arXiv:1603.06290* .

Bargagli-Stoffi, F. J., De-Witte, K. & Gnecco, G. (2019), 'Heterogeneous causal effects with imperfect compliance: a novel bayesian machine learning approach', *arXiv preprint arXiv:1905.12707* .

Behrens, C., Pierdzioch, C. & Risse, M. (2019), 'Do german economic research institutes publish efficient growth and inflation forecasts? a bayesian analysis', *Journal of Applied Statistics* pp. 1–26.

Bleich, J., Kapelner, A., George, E. I. & Jensen, S. T. (2014), 'Variable selection for bart: An application to gene regulation', *The Annals of Applied Statistics* pp. 1750–1781.

Bonato, V., Baladandayuthapani, V., Broom, B. M., Sulman, E. P., Aldape, K. D. & Do, K.-A. (2010), 'Bayesian ensemble methods for survival prediction in gene expression data', *Bioinformatics* **27**(3), 359–367.

Brock, W. A. & Durlauf, S. N. (2001), 'What have we learned from a decade of empirical research on growth? growth empirics and reality', *The World Bank Economic Review* **15**(2), 229–272.

Carvalho, C., Feller, A., Murray, J., Woody, S. & Yeager, D. (2019), 'Assessing treatment effect variation in observational studies: Results from a data challenge', *arXiv preprint arXiv:1907.07592* .

Castillo, I. & Rockova, V. (2019), 'Multiscale analysis of bayesian cart', *arXiv preprint arXiv:1910.07635* .

Chakraborty, S. (2016), Bayesian additive regression tree for seemingly unrelated regression with automatic tree selection, *in* 'Handbook of Statistics', Vol. 35, Elsevier, pp. 229–251.

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

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

Dorie, V., Hill, J., Shalit, U., Scott, M., Cervone, D. et al. (2019), 'Automated versus do-it-yourself methods for causal inference: Lessons learned from a data analysis competition', *Statistical Science* **34**(1), 43–68.

Du, J. & Linero, A. (2019), Incorporating grouping information into bayesian decision tree ensembles, *in* 'International Conference on Machine Learning', pp. 1686–1695.

Du, J. & Linero, A. R. (2018), 'Interaction detection with bayesian decision tree ensembles', *arXiv preprint arXiv:1809.08524* .

Entezari, R., Craiu, R. V. & Rosenthal, J. S. (2018), 'Likelihood inflating sampling algorithm', *Canadian Journal of Statistics* **46**(1), 147–175.

Fernandez, C., Ley, E. & Steel, M. F. (2001*a*), 'Benchmark priors for bayesian model averaging', *Journal of Econometrics* **100**(2), 381–427.

Fernandez, C., Ley, E. & Steel, M. F. (2001*b*), 'Model uncertainty in cross-country growth regressions', *Journal of applied Econometrics* **16**(5), 563–576.

Fisher, J. D. et al. (2019), Balancing model structure and flexibility in forecasting financial time series, PhD thesis.

Friedman, J. H. et al. (1991), 'Multivariate adaptive regression splines', *The annals of statistics* **19**(1), 1–67.

George, E., Laud, P., Logan, B., McCulloch, R. & Sparapani, R. (2018), 'Fully nonparametric bayesian additive regression trees', *arXiv preprint arXiv:1807.00068* .

Green, D. P. & Kern, H. L. (2012), 'Modeling heterogeneous treatment effects in survey experiments with bayesian additive regression trees', *Public opinion quarterly* **76**(3), 491–511.

Gupta, R., Pierdzioch, C. & Risse, M. (2016), 'On international uncertainty links: Bart-based empirical evidence for canada', *Economics Letters* **143**, 24–27.

Hahn, P. R., Carvalho, C. M., Puelz, D., He, J. et al. (2018), 'Regularization and confounding in linear regression for treatment effect estimation', *Bayesian Analysis* **13**(1), 163–182.

Hahn, P. R., Dorie, V. & Murray, J. S. (2019), 'Atlantic causal inference conference (acic) data analysis challenge 2017', *arXiv preprint arXiv:1905.09515* .

Hahn, P. R., Murray, J. S. & Carvalho, C. M. (2017), 'Bayesian regression tree models for causal inference: regularization, confounding, and heterogeneous effects'.

He, J., Yalov, S. & Hahn, P. R. (2018), 'Accelerated bayesian additive regression trees', *arXiv preprint arXiv:1810.02215* .

Henderson, N. C., Louis, T. A., Rosner, G. L. & Varadhan, R. (2017), 'Individualized treatment effects with censored data via fully nonparametric bayesian accelerated failure time models', *arXiv preprint arXiv:1706.06611* .

Hernández, B., Raftery, A. E., Pennington, S. R. & Parnell, A. C. (2018), 'Bayesian additive regression trees using bayesian model averaging', *Statistics and Computing* **28**(4), 869–890.

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

Killick, R., Fearnhead, P. & Eckley, I. A. (2012), 'Optimal detection of changepoints with a linear computational cost', *Journal of the American Statistical Association* **107**(500), 1590–1598.

Kindo, B. P., Wang, H., Hanson, T. & Peña, E. A. (2016), 'Bayesian quantile additive regression trees', *arXiv preprint arXiv:1607.02676* .

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

Kindo, B., Wang, H. & Pena, E. (2013), 'Mbact-multiclass bayesian additive classification trees', *stat* .

Kleinberg, J., Ludwig, J., Mullainathan, S. & Obermeyer, Z. (2015), 'Prediction policy problems', *American Economic Review* **105**(5), 491–95.

Lakshminarayanan, B., Roy, D. & Teh, Y. W. (2015), Particle gibbs for bayesian additive regression trees, *in* 'Artificial Intelligence and Statistics', pp. 553–561.

Linero, A. R. (2018), 'Bayesian regression trees for high-dimensional prediction and variable selection', *Journal of the American Statistical Association* pp. 1–11.

Linero, A. R., Sinha, D. & Lipsitz, S. R. (2019), 'Semiparametric mixed-scale models using shared bayesian forests', *Biometrics* .

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

Liu, Y., Rocková, V. & Wang, Y. (2018), 'Abc variable selection with bayesian forests', *arXiv preprint arXiv:1806.02304* .

Madigan, D. & Raftery, A. E. (1994), 'Model selection and accounting for model uncertainty in graphical models using occam's window', *Journal of the American Statistical Association* **89**(428), 1535–1546.

Minka, T. (2000), Bayesian linear regression, Technical report, Citeseer.

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

Pierdzioch, C., Risse, M., Gupta, R. & Nyakabawo, W. (2019), 'On reit returns and (un-) expected inflation: Empirical evidence based on bayesian additive regression trees', *Finance Research Letters* **30**, 160–169.

Pierdzioch, C., Risse, M. & Rohloff, S. (2016), 'Are precious metals a hedge against exchange-rate movements? an empirical exploration using bayesian additive regression trees', *The North American Journal of Economics and Finance* **38**, 27–38.

Pratola, M., Chipman, H., George, E. & McCulloch, R. (2017), 'Heteroscedastic bart using multiplicative regression trees', *arXiv preprint arXiv:1709.07542* .

Pratola, M. T., Chipman, H. A., Gattiker, J. R., Higdon, D. M., McCulloch, R. & Rust, W. N. (2014), 'Parallel bayesian additive regression trees', *Journal of Computational and Graphical Statistics* **23**(3), 830–852.

Pratola, M. T. et al. (2016), 'Efficient metropolis–hastings proposal mechanisms for bayesian regression tree models', *Bayesian analysis* **11**(3), 885–911.

Prüser, J. (2019), 'Forecasting with many predictors using bayesian additive regression trees', *Journal of Forecasting* .

Quadrianto, N. & Ghahramani, Z. (2014), 'A very simple safe-bayesian random forest', *IEEE transactions on pattern analysis and machine intelligence* **37**(6), 1297–1303.

Rocková, V. & Saha, E. (2018), 'On theory for bart', *arXiv preprint arXiv:1810.00787* .

Rockova, V. & van der Pas, S. (2017), 'Posterior concentration for bayesian regression trees and their ensembles', *arXiv preprint arXiv:1708.08734* .

Sala-i Martin, X., Doppelhofer, G. & Miller, R. I. (2004), 'Determinants of long-term growth: A bayesian averaging of classical estimates (bace) approach', *American economic review* pp. 813–835.

Santos, P. H. F. d. & Lopes, H. F. (2018), 'Tree-based bayesian treatment effect analysis', *arXiv preprint arXiv:1808.09507* .

Scott, S. L., Blocker, A. W., Bonassi, F. V., Chipman, H. A., George, E. I. & McCulloch, R. E. (2016), 'Bayes and big data: The consensus monte carlo algorithm', *International Journal of Management Science and Engineering Management* **11**(2), 78–88.

Sparapani, R. A., Logan, B. R., McCulloch, R. E. & Laud, P. W. (2016), 'Nonparametric survival analysis using bayesian additive regression trees (bart)', *Statistics in medicine* **35**(16), 2741–2753.

Sparapani, R. A., Rein, L. E., Tarima, S. S., Jackson, T. A. & Meurer, J. R. (2018), 'Non-parametric recurrent events analysis with bart and an application to the hospital admissions of patients with diabetes', *Biostatistics* .

Sparapani, R., Logan, B. R., McCulloch, R. E. & Laud, P. W. (2019), 'Nonparametric competing risks analysis using bayesian additive regression trees', *Statistical methods in medical research* p. 0962280218822140.

Starling, J. E., Murray, J. S., Carvalho, C. M., Bukowski, R. K. & Scott, J. G. (2018), 'Bart with targeted smoothing: An analysis of patient-specific stillbirth risk', *arXiv preprint arXiv:1805.07656* .

Steel, M. F. (2017), 'Model averaging and its use in economics', *arXiv preprint arXiv:1709.08221* .

Taddy, M., Chen, C.-S., Yu, J. & Wyle, M. (2015), 'Bayesian and empirical bayesian forests', *arXiv preprint arXiv:1502.02312* .

Tan, Y. V., Flannagan, C. A. & Elliott, M. R. (2016), 'Predicting human-driving behavior to help driverless vehicles drive: random intercept bayesian additive regression trees', *arXiv preprint arXiv:1609.07464* .

Tan, Y. V., Flannagan, C. A. & Elliott, M. R. (2018), '" robust-squared" imputation models using bart', *arXiv preprint arXiv:1801.03147* .

Tan, Y. V. & Roy, J. (2019), 'Bayesian additive regression trees and the general bart model', *arXiv preprint arXiv:1901.07504* .

Wager, S. & Athey, S. (2017), 'Estimation and inference of heterogeneous treatment effects using random forests', *Journal of the American Statistical Association* (just-accepted).

Xu, D., Daniels, M. J. & Winterstein, A. G. (2016), 'Sequential bart for imputation of missing covariates', *Biostatistics* **17**(3), 589–602.

Yao, Y., Vehtari, A., Simpson, D., Gelman, A. et al. (2018), 'Using stacking to average bayesian predictive distributions', *Bayesian Analysis* .

Yeager, D. S., Hanselman, P., Walton, G. M., Murray, J. S., Crosnoe, R., Muller, C., Tipton, E., Schneider, B., Hulleman, C. S., Hinojosa, C. P. et al. (2019), 'A national experiment reveals where a growth mindset improves achievement', *Nature* pp. 1–6.

Zhou, T., Daniels, M. J. & Müller, P. (2019), 'A semiparametric bayesian approach to dropout in longitudinal studies with auxiliary covariates', *Journal of Computational and Graphical Statistics* (just-accepted), 1–32.