Stochastic Tree Ensembles for Estimating Heterogeneous Effects

Nikolay Krantsevich [1] **Jingyu He** [2] **P. Richard Hahn** [1]

Abstract

Determining subgroups that respond especially well (or poorly) to specific interventions (medical or policy) requires new supervised learning methods tailored specifically for causal inference. Bayesian Causal Forest (BCF) is a recent method that has been documented to perform well on data generating processes with strong confounding of the sort that is plausible in many applications. This paper develops a novel algorithm for fitting the BCF model, which is more efficient than the previously available Gibbs sampler. The new algorithm can be used to initialize independent chains of the existing Gibbs sampler leading to better posterior exploration and coverage of the associated interval estimates in simulation studies. The new algorithm is compared to related approaches via simulation studies as well as an empirical analysis.

1. Background

1.1. Estimating heterogeneous effects

This paper considers the use of supervised machine learning for estimating conditional average treatment effects (CATE), the treatment effect averaged across subpopulations defined in terms of measured attributes.

Let Y_i represent the scalar response variable, Z_i denote a binary treatment variable, and $\mathbf{x}_i$ represent a length d row vector of observed control variables for observation i. Let Y and Z be length n column vectors comprising variables Y_i and Z_i respectively; let $\mathbf{X}$ denote the $n \times d$ matrix of control variables. We will use lower case Roman letters, such as y and z, to denote the values assumed by variables. Our data will consist of n independent observations $(Y_i, Z_i, \mathbf{x}_i)$.

Following the potential outcomes framework (Imbens & Rubin, 2015), let $Y_i(1)$ and $Y_i(0)$ represent the outcomes under treatment and control respectively; each observed response may be expressed as $Y_i = Z_i Y_i(1) + (1 - Z_i) Y_i(0)$.

Throughout, we assume the following standard conditions licensing regression estimates of treatment effects:

1. **SUTVA (Stable Unit Treatment Value Assumption)** implies that no treatment assignment to a particular individual should affect the observed outcomes on other individuals and that there is no variation in treatment.

2. **Strong ignorability assumption** implies that, first, there are no unmeasured confounders:

$$Y_i(0), Y_i(1) \perp\!\!\!\perp Z_i \mid \mathbf{X}_i, \tag{1}$$

and, second, that every individual has a non-zero probability of being assigned to treatment:

$$0 < \Pr(Z_i = 1 \mid \mathbf{x}_i) < 1. \tag{2}$$

Under these assumptions, the conditional average treatment effect of units with covariates $\mathbf{x}$ may be estimated as the differences of two identified conditional expectations:

$$\tau(\mathbf{x}) := \mathbf{E}(Y \mid \mathbf{x}, Z = 1) - \mathbf{E}(Y \mid \mathbf{x}, Z = 0). \tag{3}$$

Further assuming a mean-zero additive error,

$$Y_i = f(\mathbf{x}_i, Z_i) + \epsilon_i, \quad \epsilon_i \sim \mathbf{N}(0, \sigma^2), \tag{4}$$

it follows that $\mathbf{E}(Y_i \mid \mathbf{x}_i, Z_i = z_i) = f(\mathbf{x}_i, z_i)$ and

$$\tau(\mathbf{x}_i) := f(\mathbf{x}_i, 1) - f(\mathbf{x}_i, 0). \tag{5}$$

Here, these conditional expectations will be estimated using a Bayesian tree ensemble method (He et al., 2019) related to a well-known method called Bayesian additive regression trees, or BART (Chipman et al., 2010).

1.2. BART model and prior

BART represents the outcome of interest as a sum of an unknown function $f(\cdot)$ and an error term,

$$Y_i = f(\mathbf{x}_i) + \epsilon_i, \quad \epsilon_i \sim \mathbf{N}(0, \sigma^2) \tag{6}$$

[1]School of Mathematical and Statistical Sciences, Arizona State University, Tempe, Arizona, USA [2]City University of Hong Kong, Hong Kong SAR. Correspondence to: Nikolay Krantsevich <krantsevich@asu.edu>.

Proceedings of the 38^{th} International Conference on Machine Learning, PMLR 139, 2021.

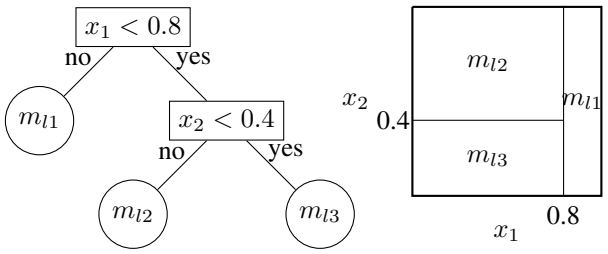

Figure 1. (Left) An example binary tree, with internal nodes labelled by their splitting rules and terminal nodes labelled with the corresponding parameters m_{lb}. (Right) The corresponding partition of the sample space and the step function. Here $\mathbf{m}_l = (m_{l1}, m_{l2}, m_{l3})$.

The mean function $f(\mathbf{x})$ is represented as a sum of many piecewise constant binary regression trees

$$f(\mathbf{x}) = \sum_{l=1}^{L} g_l(\mathbf{x}; T_l, \mathbf{m}_l) \tag{7}$$

where T_l denotes a regression tree, which represents a partition of the covariate space (say $\mathcal{A}_1, \ldots, \mathcal{A}_{B(l)}$) and consists of a set of internal decision nodes and a set of terminal nodes (or leaves) which correspond to each element of the partition. Each element of the partition $\mathcal{A}_b$ is associated a leaf parameter value, m_{lb}, and $\mathbf{m}_l = (m_{l1}, \cdots, m_{lB(l)})$ denotes a vector corresponding to all leaf parameters of the l-th tree, T_l. The piecewise constant function comprising the partition and the leaf parameters is defined as $g_l(\mathbf{x}) = m_{lb}$ if $\mathbf{x} \in \mathcal{A}_b$; see Figure 1 for demonstration.

Within each leaf, the mean parameters are given independent normal priors, $m_{lb} \sim \mathbf{N}(0, \nu)$. The prior over trees $p(T_l)$ is specified by the probability of a node having children at depth d (Chipman et al., 1998) as

$$\alpha(1+d)^{-\beta}, \qquad \alpha \in (0,1), \beta \in [0, \infty) \tag{8}$$

BART explores the posterior of the trees by random walk Metropolis-Hastings Markov chain Monte Carlo (MCMC) algorithm, which can be slow to converge and limits the broader adoption of BART for large datasets.

1.3. XBART

XBART, short for Accelerated Bayesian Additive Regression Trees (He & Hahn, 2020), was introduced to improve the fitting time of BART-like models. XBART blends regularization and stochastic search strategies from Bayesian modeling with computationally efficient techniques from recursive partitioning approaches to tree-fitting. XBART fits the same sum-of-trees ensemble model as BART but regrows each tree *recursively* at each iteration according to a stochastic process inspired by Bayesian updating.

We review the stochastic tree-growing approach of XBART (Algorithm 1). Let $\mathcal{C}$ denote a matrix of cutpoint candidates, with each element c_{jk} where $j = 1, \cdots, p$ indexes a variable and k indexes a candidate cutpoint. Assume the leaf parameter m has prior $N(0, \nu)$. At each node, the probability of splitting at cutpoint c_{jk} is proportional to

$$\begin{aligned} L(c_{jk}) \propto \exp\Bigg\{ \frac{1}{2} \Bigg[& \log\left(\frac{\sigma^2}{\sigma^2 + \nu n^l_{jk}}\right) + \\ & + \frac{\nu}{\sigma^2(\sigma^2 + \nu n^l_{jk})} \left(s^l_{jk}\right)^2 + \log\left(\frac{\sigma^2}{\sigma^2 + \nu n^r_{jk}}\right) \\ & + \frac{\nu}{\sigma^2(\sigma^2 + \nu n^r_{jk})} \left(s^r_{jk}\right)^2 \Bigg] \Bigg\}, \end{aligned} \tag{9}$$

where σ^2 is the residual variance as in equation (6), n^l_{jk} and n^r_{jk} correspond to the number of data observations on left or right child node if split at the splitting rule c_{jk}, and s^l_{jk} and s^r_{jk} are the corresponding sufficient statistics for the children nodes:

$$\begin{aligned} s^l_{jk} = \sum_{x_i \in \mathcal{A}^{\text{left}}_{jk}} y_i, \quad s^r_{jk} = \sum_{x_i \in \mathcal{A}^{\text{right}}_{jk}} y_i \\ s_{\text{all}} = s^l_{jk} + s^r_{jk} = \sum_{i=1}^{n} y_i, \end{aligned} \tag{10}$$

where $n = n^l_{jk} + n^r_{jk}$ is number of observations in the current node. Similarly, the probability of not splitting anywhere is proportional to

$$\begin{aligned} L(\emptyset) \propto |\mathcal{C}| \left(\frac{(1+d)^\beta}{\alpha} - 1 \right) \times \\ \exp\left\{ \frac{1}{2} \left[\log\left(\frac{\sigma^2}{\sigma^2 + \nu n}\right) + \frac{\nu}{\sigma^2(\sigma^2 + \nu n)} s^2_{\text{all}} \right] \right\}. \end{aligned} \tag{11}$$

where $|\mathcal{C}|$ is the total number of candidate splitting rules, d is the depth of the current node in the tree. The tree is fitted recursively where at each node, a cutpoint (or the stop-splitting option) is randomly drawn from a multinomial distribution using probabilities of (9) and (11). If stop-splitting is sampled, or other pre-set stopping conditions are satisfied, the current node becomes a terminal (leaf) node, and its associated leaf parameter m is updated by conjugate Gaussian sampling. To form an ensemble of trees, XBART uses a similar strategy as Bayesian backfitting, residualizes the data with respect to the partial fit corresponding to the forest. Specifically, the h-th tree is grown to fit the partial residual of all other trees: $\mathrm{y} - \sum_{l \neq h} g_l(\mathbf{x}; T_l, \mathbf{m}_l)$.

He & Hahn (2020) details how these strategies contribute to the improved efficiency of XBART over BART as well as improved posterior coverage of interval estimates obtained by initializing multiple Markov chains at XBART estimates. Here, we adapt the XBART approach to the BCF model and demonstrate comparable performance gains in the heterogeneous treatment effect setting.

Algorithm 1 Grow From Root (GFR)

Input: GFR$(\mathrm{y}, \mathbf{X}, \sigma, d, T, \texttt{node})$.
Calculate full sufficient statistics s_{all} by (10).
for $c_{jk} \in \mathcal{C}$, partition data to left and right sides **do**
  Calculate s^l_{jk} and s^r_{jk} by equation (10).
  Calculate $L(c_{jk})$ by equation (9).
end for
Calculate probability of no-split $L(\emptyset)$ by equation (11).
Draw a cutpoint or no-split using probability $L(c_{jk})$ and $L(\emptyset)$.
if no-split is chosen or stop conditions are met **then**
  Update leaf parameter $m_{\texttt{node}}$.
  return.
else
  Create two new nodes as children of `node`, denoted `left_node` and `right_node`.
  Sift the data into `left_node` and `right_node`.
  GFR$(y_{\text{left}}, \mathbf{X}_{\text{left}}, \sigma, d+1, T, \texttt{left_node})$
  GFR$(y_{\text{right}}, \mathbf{X}_{\text{right}}, \sigma, d+1, T, \texttt{right_node})$
end if
Output: The grown tree T, including the vector of sampled leaf parameters, $\mathbf{m}$.

1.4. Bayesian Causal Forest

Via simulation studies, Hahn et al. (2020) demonstrate the inability of BART to handle confounding for certain simple data generative processes (DGPs). They refine the BART model to overcome this limitation with several modifications. First, rather than representing $f(\mathbf{x}, z)$ as a single BART model (as in Hill (2011)), they propose using the representation

$$f(\mathbf{x}_i, z_i) = \mu(\mathbf{x}_i) + \tau(\mathbf{x}_i) z_i, \tag{12}$$

where μ and τ are prognostic and treatment functions respectively; both are given independent BART priors, permitting control and treatment effects to be regularized independently. Second, they propose including an estimate of the propensity score $\widehat{\pi}_i = P(Z_i = 1 \mid \mathbf{x}_i)$ as a crucial additional feature to combat unintended bias of treatment effects due to the regularization of μ. See Hahn et al. (2020) for more details on this phenomenon, which the authors refer to as *regularization induced confounding* (RIC).

Finally, Hahn et al. (2020) observe that Bayesian treatment effect estimation is not invariant with respect to treatment encoding – choosing different pairs of values as treatment indicators for treated and control groups implies different priors, which lead to different treatment effect estimates. By adding scaling factors b_0 and b_1 as parameters in the model, the priors are made invariant to which group is designated as the treated group. An additional scaling factor, a, is added to enhance the learning of the prognostic term. Putting these modifications together, the Bayesian causal forest (BCF) model is

$$\begin{aligned} y_i &= a\mu(\mathbf{x}_i, \widehat{\pi}_i) + b_{z_i}\tilde{\tau}(\mathbf{x}_i) + \epsilon_i, \quad \epsilon_i \sim \mathbf{N}(0, \sigma^2) \\ a &\sim \mathbf{N}(0,1), \quad b_0, b_1 \sim \mathbf{N}(0, 1/2) \end{aligned} \tag{13}$$

According to this parametrization, treatment effects are given by $\tau(\mathbf{x}_i) = (b_1 - b_0)\tilde{\tau}(\mathbf{x}_i)$.

The Bayesian Causal Forest model has been documented to perform well in a number of separate, rigorous simulation studies (Hahn et al., 2019; Dorie et al., 2018; Wendling et al., 2018). It was recently used to estimate CATEs in the high-profile Growth Mindset intervention (Yeager et al., 2019) as well as other applied work (Ghosh et al., 2020; King et al., 2019; Bail et al., 2020; Bryan et al., 2019).

Computationally, BCF is built upon the same random walk Metropolis-Hastings algorithm that underpins BART. As such, it suffers from the same slow fitting time on large data sets and the same slow posterior exploration. The next section seeks to address these limitations by applying the computational strategies of XBART to a BCF model.

2. XBCF

2.1. The model

The XBCF model differs in one substantive respect from the model presented in Hahn et al. (2020): The error standard deviations σ_0 and σ_1 are allowed to differ between the control and treatment groups, respectively, whereas the original BCF model had a common shared residual standard deviation. Thus, the XBCF model is

$$\begin{aligned} y_i &= a\mu(\mathbf{x}_i, \widehat{\pi}_i) + b_{z_i}\tilde{\tau}(\mathbf{x}_i) + \epsilon_i, \quad \epsilon_i \sim \mathbf{N}(0, \sigma^2_{z_i}), \\ a &\sim \mathbf{N}(0,1), \quad b_0, b_1 \sim \mathbf{N}(0, 1/2), \end{aligned} \tag{14}$$

or, in more detail, as

$$y_i = a\sum_{l=1}^{L} u_l(\mathbf{x}_i, \widehat{\pi}_i; T_l, \mathbf{m}^T_l) + b_{z_i}\sum_{k=1}^{K} v_k(\mathbf{x}_i; S_k, \mathbf{m}^S_k) + \epsilon_i,$$

where L, K represent the number of trees, T_l, S_k represent individual trees, $\mathbf{m}^T_l, \mathbf{m}^S_k$ denote vectors of scalar means associated with the leafs nodes of T_l and S_k respectively. We will reference the forests of trees as $T = \{T_l, \mathbf{m}^T_l\}^L_{l=1}$ and $S = \{S_k, \mathbf{m}^S_k\}^K_{k=1}$ for prognostic and treatment terms, respectively. Following BCF, we include a column vector of (estimated) propensity scores $\widehat{\pi}$ as an additional covariate for the prognostic term.

2.2. Modeling fitting procedure

The XBCF fitting algorithm uses a similar "backfitting" strategy as BART and XBART, iterating tree-by-tree through two forests (corresponding to the prognostic and treatment terms) rather than just one. The tree and parameter updates

at each iteration are based on the following "residuals":

$$\begin{aligned}\text{Prognostic residual: } \mathrm{v} &\equiv \mathrm{y} - a\sum_{l=1}^{L} u(\mathbf{X},\hat{\pi};T_l,\mathbf{m}_l^T),\\ \text{Treatment residual: } \mathrm{t} &\equiv \mathrm{y} - \mathrm{b}\cdot\sum_{k=1}^{K} v(\mathbf{X};S_k,\mathbf{m}_k^S),\\ \text{Total residual: } \mathrm{r} &\equiv \mathrm{y} - a\sum_{l=1}^{L} u(\mathbf{X},\hat{\pi};T_l,\mathbf{m}_l^T)\\ &\quad - \mathrm{b}\cdot\sum_{k=1}^{K} v(\mathbf{X};S_k,\mathbf{m}_k^S).\end{aligned} \tag{15}$$

where b is a length n vector with i-th component equal to b_{z_i}, and '$\cdot$' denotes element-wise multiplication. The update steps for trees, T_l or S_k, depend on the vectors of *partial* residuals, which subtracts off the partial fit corresponding to the forests *without* the current tree from the observed response variable:

$$\begin{aligned}\mathrm{r}_{-l}^T &\equiv \mathrm{r} + au(\mathbf{X},\hat{\pi};T_l,\mathbf{m}_l^T), \quad l=1,\ldots,L,\\ \mathrm{r}_{-k}^S &\equiv \mathrm{r} + \mathrm{b}\cdot v(\mathbf{X};S_k,\mathbf{m}_k^S), \quad k=1,\ldots,K.\end{aligned} \tag{16}$$

With these terms defined, the sequence of stochastic updates is as follows:

1. **Stage 1: update prognostic forest.** We first grow L trees comprising the forest for the prognostic term $\mu(\mathbf{x}_i,\widehat{\pi}_i)$. For each of the trees ($l=1,\ldots,L$) the sequence of updates is the following:
   (a) $T_l,\mathbf{m}_l^T \mid \mathrm{r}_{-l}^T,\sigma_0^2,\sigma_1^2,a,b_0,b_1$, which is done compositionally as
      i. $T_l \mid \mathrm{r}_{-l}^T,\sigma_0^2,\sigma_1^2$
      ii. $\mathbf{m}_l^T \mid T_l,\sigma_0^2,\sigma_1^2,a,b_0,b_1$
   (b) $a \mid \mathrm{t},T_l$
   (c) $b_0,b_1 \mid \mathrm{v},T_l$
   (d) $\sigma_0^2,\sigma_1^2 \mid \mathrm{r}$.
2. **Stage 2: update treatment forest.** We then grow K trees comprising the forest for the treatment term $\tau(\mathbf{x}_i)$. The sequence of updates for each tree ($k=1,\ldots,K$) is similar:
   (a) $S_k,\mathbf{m}_k^S \mid \mathrm{r}_{-k}^S,\sigma_0^2,\sigma_1^2,a,b_0,b_1$, which is done compositionally as
      i. $S_k \mid \mathrm{r}_{-k}^S,\sigma_0^2,\sigma_1^2$
      ii. $\mathbf{m}_k^S \mid S_k,\sigma_0^2,\sigma_1^2,a,b_0,b_1$
   (b) $a \mid \mathrm{t},S_k$
   (c) $b_0,b_1 \mid \mathrm{v},S_k$
   (d) $\sigma_0^2,\sigma_1^2 \mid \mathrm{r}$,

These two stages are repeated I times, which we refer to as "sweeps". Pseudocode is given in Algorithm 2. Although we use conditioning notation, note that these stochastic updates are *not* full conditional distributions in the usual Gibbs sampling sense. The tree-growing updates (Stage 1(a) and Stage 2(a)) are given in Algorithm 1, applied to the partial residuals defined in expression 16. Parameter updates are detailed in the next subsection.

After I sweeps, the CATE estimate for individuals with features $\mathbf{x}$ is calculated as an average of the $(b_1-b_0)\tilde{\tau}(\mathbf{x})$ samples, as if one were taking a traditional posterior mean.

Algorithm 2 Accelerated Bayesian Causal Forest (XBCF)

1: **Input:** $\mathrm{y},\mathbf{X},L,K,I$
2: Initialize $\mathrm{r},\mathrm{v},\mathrm{t}$, partial residuals $\mathrm{r}_{-l}^T,\mathrm{r}_{-k}^S$ and scale parameters $a,b_0,b_1,\sigma_0,\sigma_1$.
3: **for** iter in 1 to I **do**
4: **for** l in 1 to L **do**
5: Compute partial residual r_{-l}^T by equation (16).
6: Create `new_node` to initialize tree T_l^{iter} with root node.
7: GFR($\mathrm{r}_{-l}^T,\mathbf{X},\sigma_0^2,\sigma_1^2,d=0,T_l^{\text{iter}},$`new_node`).
8: Update leaf parameter $\mathbf{m}_h^{T,\text{iter}}$ for T_l^{iter}.
9: Update full residual r,v by equation (15).
10: Sample $a,b_0,b_1,\sigma_0,\sigma_1$ based on $\mathrm{r},\mathrm{v},\mathrm{t}$.
11: **end for**
12: **for** k in 1 to K **do**
13: Compute partial residual r_{-k}^S by equation (16).
14: Create `new_node` to initialize tree S_k^{iter} with root node.
15: GFR($\mathrm{r}_{-k}^S,\mathbf{X},\sigma_0^2,\sigma_1^2,d=0,S_k^{\text{iter}},$`new_node`).
16: Update leaf parameter $\mathbf{m}_k^{S,\text{iter}}$ for S_k^{iter}.
17: Update full residual r,t by equation (15).
18: Sample $a,b_0,b_1,\sigma_0,\sigma_1$ based on $\mathrm{r},\mathrm{v},\mathrm{t}$.
19: **end for**
20: **end for**
21: **output:** $\{\{T_l^{\text{iter}},\mathbf{m}_l^{T,\text{iter}}\}_{l=1}^L,\{S_k^{\text{iter}},\mathbf{m}_k^{S,\text{iter}}\}_{k=1}^K\}_{\text{iter}=1}^I$, I posterior draws of the prognostic and treatment forests, and $\{a^{\text{iter}},b_0^{\text{iter}},b_1^{\text{iter}},\sigma_0^{\text{iter}},\sigma_1^{\text{iter}}\}_{\text{iter}=1}^I$, I posterior draws of other model parameters.

2.2.1. Parameter updates

If the no-split option is selected, or other pre-set stopping conditions are satisfied, the current node becomes a leaf node and the associated leaf parameter is updated as follows (line 8 and 16 in Algorithm 2). This update corresponds to a conditionally conjugate Gaussian mean update; we incorporate the control group and treatment group data sequentially to accommodate their differing variances (σ_0^2 and σ_1^2):

$$\nu_{n_0} = \left(\frac{1}{\nu}+\frac{n_0}{d_0^2}\right)^{-1}, \quad \beta_{n_0} = \frac{\bar{y}_0}{d_0^2}\nu_{n_0},$$

followed by

$$\nu_n = \left(\frac{1}{\nu_{n_0}}+\frac{n_1}{d_1^2}\right)^{-1}, \quad \beta_n = \left(\frac{\beta_{n_0}}{\nu_{n_0}}+\frac{\bar{y}_1}{d_1^2}\right)\nu_n,$$

where ν is the prior variance over the mean, $d_0=\frac{\sigma_0}{b_0}, d_1=\frac{\sigma_1}{b_1}$; n_0,n_1 are the number of individuals in control and

treatment groups respectively for this leaf node, and $\bar{y}_0, \bar{y}_1$ are the corresponding partial residual means of these two groups in this leaf node. The leaf mean parameter is then sampled according to $\mathbf{m} \sim \mathbf{N}(\beta_n, \nu_n^2)$.

Model parameters $a, b_0, b_1, \sigma_0, \sigma_1$ are sampled after each tree update, for a total of $L + K$ times per sweep. After updating trees, the model parameters are sampled based on the residual vectors in equation (15) – the prognostic residual v, the treatment residual t and the total residual r (lines 9 and 17 in Algorithm 2). Since the general update sequence is similar for the two stages above, we will provide an explicit update scheme of each step for only Stage 2.

In order to update parameter a we first reshape (14) in a regression problem where the treatment residual vector t, with each component divided by corresponding σ_{z_i}, is the response variable:

$$\begin{bmatrix} \frac{y_1 - b_{z_1}\tau(x_1)}{\sigma_{z_1}} \\ \vdots \\ \frac{y_n - b_{z_n}\tau(x_n)}{\sigma_{z_n}} \end{bmatrix} = \begin{bmatrix} \frac{\mu(x_1)}{\sigma_{z_1}} \\ \vdots \\ \frac{\mu(x_n)}{\sigma_{z_n}} \end{bmatrix} a + \begin{bmatrix} \frac{\epsilon_1}{\sigma_{z_1}} \\ \vdots \\ \frac{\epsilon_n}{\sigma_{z_n}} \end{bmatrix}.$$

Then updating a is essentially implemented as a two-step regression update:

$$\nu_{n_0} = \left(1 + \frac{\mu_0^t \mu_0}{\sigma_0^2}\right)^{-1}, \quad \beta_{n_0} = \frac{\mathrm{t}_0^t \mu_0}{\sigma_0^2} \nu_{n_0};$$
$$\nu_n = \left(\frac{1}{\nu_{n_0}} + \frac{\mu_1^t \mu_1}{\sigma_1^2}\right)^{-1}, \beta_n = \left(\frac{\beta_{n_0}}{\nu_{n_0}} + \frac{\mathrm{t}_1^t \mu_1}{\sigma_1^2}\right) \nu_n,$$

where μ_0 is a vector with elements corresponding to $\mu(\cdot)$ evaluated at rows of $\mathbf{X}$ for which $z_i = 0$, and similarly for μ_1; t_0 is the part of residual vector t corresponding to only individuals with $z_i = 0$, and similarly for t_1. The parameter a is then sampled according to $a \sim \mathbf{N}(\beta_n, \nu_n^2)$.

For the scaling factors b_0 and b_1, we rearrange (14) in the form of a linear regression problem where the prognostic residual vector v, with each component divided by corresponding σ_{z_i}, is the response variable:

$$\begin{bmatrix} \frac{y_1 - a\mu(x_1)}{\sigma_{z_1}} \\ \vdots \\ \frac{y_n - a\mu(x_n)}{\sigma_{z_n}} \end{bmatrix} = \begin{bmatrix} \frac{\tau(x_1)z_1}{\sigma_{z_1}} & \frac{\tau(x_1)(1-z_1)}{\sigma_{z_1}} \\ \vdots & \vdots \\ \frac{\tau(x_n)z_n}{\sigma_{z_n}} & \frac{\tau(x_n)(1-z_n)}{\sigma_{z_n}} \end{bmatrix} \begin{bmatrix} b_0 \\ b_1 \end{bmatrix} + \begin{bmatrix} \frac{\epsilon_1}{\sigma_{z_1}} \\ \vdots \\ \frac{\epsilon_n}{\sigma_{z_n}} \end{bmatrix},$$

and then we update b_0, b_1 as regression coefficients. We first update their sampling parameters as follows:

$$\nu_{n_0} = \left(\frac{1}{\frac{1}{2}} + \frac{\tau_0^t \tau_0}{\sigma_0^2}\right)^{-1}, \quad \beta_{n_0} = \frac{\mathrm{v}_0^t \tau_0}{\sigma_0^2} \nu_{n_0};$$
$$\nu_{n_1} = \left(\frac{1}{\frac{1}{2}} + \frac{\tau_1^t \tau_1}{\sigma_1^2}\right)^{-1}, \quad \beta_{n_1} = \frac{\mathrm{v}_1^t \tau_1}{\sigma_1^2} \nu_{n_1},$$

where τ_0 is a vector with elements corresponding to $\tau(\cdot)$ evaluated at rows of $\mathbf{X}$ for which $z_i = 0$, and similarly for τ_1; m_0 is the part of residual vector m corresponding to only individuals with $z_i = 0$, and similarly for m_1. Then b_0 and b_1 are sampled as $b_0 \sim \mathbf{N}(\beta_{n_0}, \nu_{n_0}^2), b_1 \sim \mathbf{N}(\beta_{n_1}, \nu_{n_1}^2)$.

Lastly, updating the residual variances σ_0^2 and σ_1^2 is a conditionally conjugate inverse-Gamma update:

$$\sigma_0^2 \sim \mathbf{IG}\left(\frac{n_0 + \kappa_0}{2}, \frac{2}{\mathrm{r}_0^t \mathrm{r}_0 + s_0}\right)$$
$$\sigma_1^2 \sim \mathbf{IG}\left(\frac{n_1 + \kappa_1}{2}, \frac{2}{\mathrm{r}_1^t \mathrm{r}_1 + s_1}\right),$$

where n_0, n_1 are the total number of individuals fit in the control and treatment groups respectively, $\mathrm{r}_0, \mathrm{r}_1$ are the total residuals for the same corresponding groups; $\kappa_0, \kappa_1, s_0, s_1$ are hyperparameters of the inverse-Gamma prior.

2.3. Warm-start BCF

The simulation studies presented in Section 3 reveal that coverage of both BCF and XBCF often do not reach the desired nominal rate. On the one hand, complex Bayesian models do not guarantee a nominal coverage rate of credible intervals. On the other hand, very poor coverage is obviously undesirable. One contributor to under-coverage is inadequate Monte Carlo exploration of the posterior distribution, resulting in artificially narrow reported intervals. Because XBCF provides a fast approximation to the BCF posterior, initializing BCF MCMC at XBCF trees rather than roots is a promising strategy to improve the posterior exploration. Specifically, we propose the following: First, use XBCF (*s* sweeps, *b* burn-in) to obtain the tree draws for each of the *s*−*b* sweeps after the burn-in period. Second, initialize *s*−*b* BCF Markov chains at the forests obtained from XBCF. Initializing BCF on the trees obtained from XBCF substantially reduces the necessary burn-in period for the BCF MCMC algorithm. Furthermore, the separately initialized chains can be run in parallel. We call this initialization strategy warm-start BCF or ws-BCF.

In order to compare the performance and computational speed of XBCF, warm-start BCF, and the original BCF, we generated data with 50 covariates (25 continuous and 25 binary) as the input matrix and stratified treatment effects. The size of the sample is $n = 5000$ and it is unbalanced on average, with approximately $\frac{2}{3}$ data points in the control group. Full details of the DGP are available in the supplement; here the time comparisons are the main interest as we expect these methods will concur on any data set given sufficient run time.

Results reported in Table 1 show that warm-start BCF with default parameters (100 iterations over 40 sweeps) performs better than the original BCF MCMC in all estimands of interest, and especially improves in coverage. In general,

Method	RMSE		Coverage		I.L.		Time
	ATE	CATE	ATE	CATE	ATE	CATE	
ws-BCF	0.021	0.101	0.960	0.920	0.095	0.376	14
XBCF	0.020	0.105	0.900	0.754	0.091	0.256	4
BCF(4)	0.027	0.130	0.840	0.675	0.092	0.229	42
BCF(20)	0.024	0.125	0.900	0.731	0.092	0.262	202

Table 1. Results of root mean squared error (RMSE), interval coverage (Coverage) and interval length (I.L.) for ATE and CATE estimators for the simulation study with 5000 datapoints and 50 covariates. The number in parenthesis for BCF indicates the number of burn-in and follow-up iterations. The column Time is running time in seconds. The results are averaged over 50 independent replications.

MCMC methods need to be run for long enough in order to converge, and when we run the original BCF for a significantly larger amount of iterations (20000 after 20000 iterations of burn-in), we still see that it does not match the performance of warm-start BCF, despite taking 10 times longer.

3. Simulation Study

We reproduce the simulation study of (Hahn et al., 2020), focusing on estimation of conditional average treatment effects on the basis of three metrics: average root mean square error, coverage and average interval length. The data are generated according to four different processes: the conditional expectation can be linear or nonlinear, and the treatment effect can be homogeneous or heterogeneous. The covariate vector $\mathbf{x}$ contains five variables, three of which are continuous, standard normal random variables, one is dichotomous, and one is unordered categorical with three levels (denoted 1,2,3). Specifically, the treatment effect is either

$$\tau(\mathbf{x}) = \begin{cases} 3 & \text{homogeneous} \\ 1 + 2x_2x_5 & \text{heterogeneous,} \end{cases}$$

and the prognostic function is defined as either

$$\mu(\mathbf{x}) = \begin{cases} 1 + g(x_4) + x_1x_3 & \text{linear} \\ -6 + g(x_4) + 6|x_3 - 1| & \text{nonlinear,} \end{cases}$$

where $g(1) = 2$, $g(2) = -1$ and $g(3) = -4$, and the propensity function is given by

$$\pi(\mathbf{x}_i) = 0.8\Phi(3\mu(\mathbf{x}_i)/s - 0.5x_1) + 0.05 + u_i/10,$$

where s is the standard deviation of $\mu(\mathbf{x})$ taken over the observed sample, with $u_i \sim \text{Uniform}(0, 1)$. The inclusion of μ in defining the treatment probability is to induce strong confounding.

The set of methods which we use to estimate treatment effects on this data include: the two methods proposed in this paper, XBCF and warm-start BCF; the original BCF method; a naive version of BART with binary treatment assignment added as a non-distinguished covariate; ps-BART, which in addition to the treatment assignment also incorporates propensity score estimates as another covariate; BART-f_0f_1, which fits two separate BART models for the treatment and control groups; Causal Random Forest (Athey et al., 2018), which also incorporates propensity score estimates (Tibshirani et al., 2017); and a Bayesian linear model with a horseshoe prior (Carvalho et al., 2010) on the regression coefficients.

		Homogeneous Treatment							Heterogeneous Treatment						
Prognostic Term	Method	RMSE		Coverage		I.L.		Time	RMSE		Coverage		I.L.		Time
		ATE	CATE	ATE	CATE	ATE	CATE		ATE	CATE	ATE	CATE	ATE	CATE	
Linear	ws-BCF	0.21	0.28	0.90	0.98	0.93	1.57	0.99	0.23	1.09	0.92	0.92	0.99	3.35	1.08
	XBCF	0.20	0.24	0.88	0.94	0.84	1.13	0.23	0.23	1.26	0.86	0.77	0.86	2.63	0.24
	BCF	0.23	0.34	0.88	0.97	0.92	1.62	4.64	0.22	1.14	0.92	0.81	0.96	2.93	4.92
	ps-BART	0.26	0.49	0.87	0.98	0.99	2.52	12.44	0.27	1.21	0.90	0.93	1.07	3.67	12.62
	CRF	0.35	0.54	0.76	0.86	1.09	1.58	0.47	0.40	1.41	0.78	0.76	1.23	2.64	0.44
	BART	0.37	0.59	0.70	0.95	0.96	2.48	12.77	0.40	1.25	0.72	0.92	1.03	3.63	13.03
	BART-f_0f_1	0.56	0.98	0.44	0.95	0.99	3.99	15.00	0.55	1.39	0.44	0.93	1.07	4.91	15.46
	lm	0.18	0.31	0.96	0.99	0.87	1.73	2.30	0.22	0.38	0.92	0.98	0.97	1.98	2.14
Nonlinear	ws-BCF	0.35	0.44	0.95	0.99	1.63	2.56	0.88	0.38	1.53	0.90	0.90	1.59	4.54	0.97
	XBCF	0.37	0.44	0.87	0.94	1.49	2.00	0.22	0.40	1.67	0.84	0.78	1.45	3.57	0.24
	BCF	0.36	0.52	0.94	0.97	1.61	2.65	4.52	0.37	1.54	0.90	0.86	1.57	4.35	4.71
	ps-BART	0.43	0.89	0.88	0.99	1.72	4.70	12.38	0.45	1.61	0.86	0.93	1.68	5.54	12.67
	CRF	0.50	0.73	0.83	0.89	1.64	2.53	0.44	0.58	1.66	0.74	0.78	1.75	3.58	0.45
	BART	0.59	0.97	0.74	0.97	1.62	4.44	12.90	0.58	1.62	0.70	0.92	1.58	5.31	12.90
	BART-f_0f_1	1.38	2.50	0.14	0.85	1.70	7.54	15.02	1.30	2.65	0.20	0.86	1.67	7.86	15.38
	lm	1.82	2.12	0.02	0.46	1.73	4.03	2.07	1.73	2.09	0.04	0.55	1.72	4.30	1.95

Table 2. Results of root mean squared error (RMSE), interval coverage (Coverage) and interval length (I.L.) for ATE and CATE estimators with different combinations of treatment term and prognostic term types. Sample size is 500. The column Time is running time in seconds.

For each of the methods, we averaged the results on the three metrics over 200 independent replications. The results on a sample of $n = 500$ data points are presented in Table 2. For this simulation study, we used default recommended settings for all of the methods. Two methods, warm-start BCF and Causal Random Forest took advantage of parallelization on eight cores.

Broadly, we recapitulate the findings of Hahn et al. (2020). Their key takeaways are that one, the propensity score is an important feature for accurate estimation of treatment effects in problems with strong confounding, and two, separate regularization of μ and τ improves estimation accuracy. Here, we highlight the differences between BCF, XBCF, and warm-start BCF.

- XBCF provides the best CATE estimation for homogeneous treatment effect case.
- XBCF provides the most narrow credible interval length, and often under covers compared to BCF and warm-start BCF.
- warm-start BCF always performs better than regular BCF in CATE estimation in terms of both RMSE and coverage.
- Overall, warm-start BCF provides the best coverage among all three methods for both ATE and CATE.

All experiments in this paper were performed on a Linux machine with Intel(R) Core(TM) i7-8700K CPU @ 3.70GHz processor and 64GB RAM; eight cores were used for parallelization whenever it was applicable.

4. Empirical demonstration

As an empirical demonstration, we analyze data on student classroom performance in language arts class collected from two public schools in Portugal during the 2005-2006 school year (Cortez & Silva, 2008). This data set is publicly available at the UCI Machine Learning Repository and was used in Cortez & Silva (2008) to predict students' final grades using supervised learning methods. The rich covariates in this data set make it possible to pose several questions regarding the causal impact of student's attributes on their final scores. Here, we focus on estimating the treatment effect of which school was attended, Gabriel Pereira (GP) or Mousinho da Silveira (MS). The course grade is an award on a 20-point scale.

From the original data set, which contained information on 649 students, we omit students whose final score is 0. We also restrict our analysis to those who state that they intend to pursue higher education, bringing the sample size to $n = 570$ students. We control for the following fifteen variables:

Method	ATE	CI length	Time
ws-BCF	0.68	1.02	1.50
XBCF	0.62	0.91	0.52
BCF	0.67	1.02	8.02
ps-BART	0.67	0.98	11.07
BART	0.68	0.99	11.26
BART-$f_0 f_1$	0.73	1.04	13.26
CRF	0.64	1.18	0.45

Table 3. ATE estimates and respective lengths of the 95% credible/confidence intervals for the set of methods we considered.

- age: age in years at the time of the survey (numeric)
- address: indicator whether student lives in a city or in a rural area (binary)
- famrel: quality of family relationship (5 levels)
- famsize: indicator whether student's family has more than 3 members or not (binary)
- famsup: family educational support (binary)
- Fedu: father's education level (5 levels)
- Fjob: father's job (5 categories)
- health: student's current health status (5 levels)
- internet: internet access at student's home (binary)
- Medu: mother's education level (5 levels)
- Mjob: mother's job (5 categories)
- nursery: indicator of attending nursery school (binary)
- Pstatus: parent's cohabitation status (binary)
- reason: reason to choose this school (4 categories)
- sex: student's gender (binary)

4.1. Treatment effect estimation

All methods considered in the simulation study are used here as well, except for the linear model. Table 3 reports point estimates and interval lengths (for 95% credible intervals for the Bayesian methods and for the 95% confidence interval for the Causal Random Forest method). All methods estimate the ATE to be in the range 0.6-0.8, with interval estimates lying above zero, suggesting a small positive average treatment effect.

Despite the ATE estimates broadly concurring, CATE estimates vary substantially across methods. Table 4 shows the correlation matrix of CATE estimates obtained from different methods. As desired, BCF and warm-start BCF are strongly positively correlated.

4.2. Subgroup analysis

Posterior inference for subgroup average treatment effects can be obtained directly from the posterior draws sampled from warm-start BCF.

	CRF	BART	BART-$f_0 f_1$	ps-BART	BCF	XBCF
BART	0.65					
BART-$f_0 f_1$	0.63	0.88				
ps-BART	0.63	0.87	0.99			
BCF	0.73	0.62	0.73	0.71		
XBCF	0.63	0.63	0.64	0.61	0.49	
ws-BCF	0.76	0.73	0.83	0.82	0.98	0.57

Table 4. The correlation matrix of CATE estimates obtained from different methods.

To discover subgroups of interest, we fit a regression tree to the posterior point estimates of the CATE, using the set of all covariates available from the original dataset; the resulting tree defines subgroups for which the CATE estimates differ. This should be considered a convenient form of posterior exploration and not a separate inference procedure. Posterior inferences are obtained simply as the sample average effects calculated according to each posterior draw. Of particular interest is the posterior difference between subgroup treatment effects: posterior credible intervals of this quantity allow us to determine if the difference between subgroups is statistically convincing.

The left panel in Figure 2 represents the fitted tree to posterior point estimates obtained from warm-start BCF. Subgroup 1, which benefited most from the treatment, with the subgroup ATE estimate of 1.3 points, consisted of 50 students with the following characteristics: mother doesn't have a higher education degree (Medu < 4); family relationship is perceived by the student as average or lower (famrel < 4); there is educational support coming from the family (famsup ≥ 2).

At the other end of the spectrum we have Subgroup 2, which benefited the least from the treatment, with the subgroup ATE estimate of -0.46 points, consisting of 11 students with the following characteristics: mother has a higher education degree (Medu ≥ 4); father's job is teacher; there is no educational support from the family (famsup < 2).

The posterior difference in subgroup ATE is shown in the middle panel of Figure 2. The majority of the computed differences is above 0 and the 95% posterior credible interval is $(-0.2, 4.7)$.

Although it makes intuitive sense that students whose parents have less education may stand to benefit more from better in-school instruction, the fact that those students are receiving at-home support while the children of teachers are not defied expectation. We speculate that the reason a pupil whose father is a teacher would not receive at-home support is if the student is not in need of assistance. If this were the case, it would suggest that better in-school instruction benefits students who are not already excelling; this is consistent with the estimated subgroup average prognostic effects (see right panel in Figure 2) as well as with previous literature on educational interventions (Yeager et al., 2019).

5. Summary

This paper introduces a novel algorithm for fitting Bayesian causal forest models, which are increasingly popular and successful in causal inference problems with heterogeneous treatment effects. The new method makes BCF models capable of fitting larger data sets than could be fit with the previous random walk Metropolis-Hastings algorithm, which can under-explore the vast space of regression tree ensembles. We hope in the future to apply our approach to large observational health databases. Moreover, even on smaller data sets, the new algorithm provides better interval estimates of conditional average treatment effects in simulations, a property that we believe to hold for empirical analyses as well, as the warm-start BCF intervals tend to be longer. We hope that other researchers can build on these tools to consider other causal inference methods that call for regularized regression, such as instrumental variables approaches or regression discontinuity designs, et cetera.

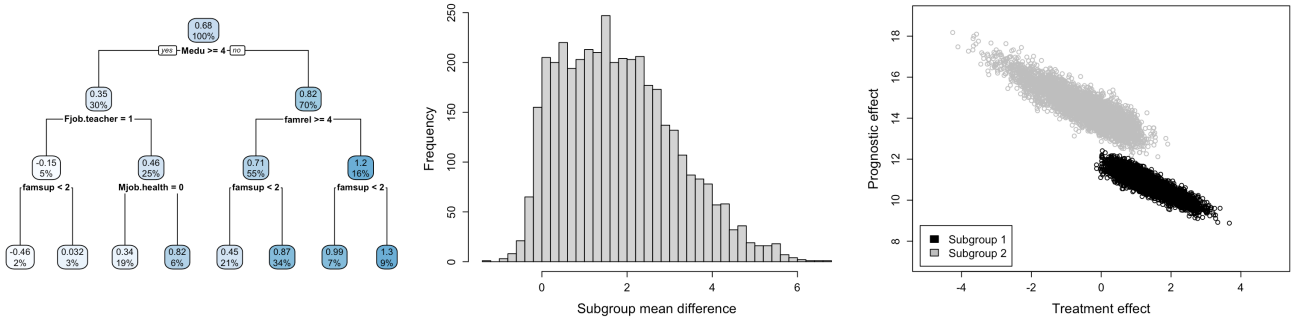

Figure 2. (Left) A single deterministic tree fit to the individual-level treatment estimates of warm-start BCF. The top number in each box is the average subgroup treatment effect, the lower number indicates the percentage of the total sample. (Middle) The histogram of difference in means of Subgroup 1 and Subgroup 2 over all posterior draws of warm-start BCF. (Right) Posterior draws of subgroup average treatment and prognostic effects for the two subgroups.

References

Athey, S., Tibshirani, J., and Wager, S. Generalized random forests, 2018.

Bail, C. A., Guay, B., Maloney, E., Combs, A., Hillygus, D. S., Merhout, F., Freelon, D., and Volfovsky, A. Assessing the russian internet research agency's impact on the political attitudes and behaviors of american twitter users in late 2017. *Proceedings of the National Academy of Sciences*, 117(1):243–250, 2020. ISSN 0027-8424. doi: 10.1073/pnas.1906420116. URL https://www.pnas.org/content/117/1/243.

Bryan, C. J., Yeager, D. S., and O'Brien, J. M. Replicator degrees of freedom allow publication of misleading failures to replicate. *Proceedings of the National Academy of Sciences*, 116(51):25535–25545, 2019. doi: 10.1073/pnas.1910951116. URL https://www.pnas.org/content/116/51/25535.

Carvalho, C. M., Polson, N. G., and Scott, J. G. The horseshoe estimator for sparse signals. *Biometrika*, 97 (2):465–480, 2010. ISSN 00063444. URL http://www.jstor.org/stable/25734098.

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

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

Cortez, P. and Silva, A. M. G. Using data mining to predict secondary school student performance. 2008.

Dorie, V., Hill, J., Shalit, U., Scott, M., and Cervone, D. Automated versus do-it-yourself methods for causal inference: Lessons learned from a data analysis competition, 2018.

Ghosh, A., Orzol, S., Dale, S., Laird, J., Fu, N., Singh, P., Kim, M.-Y. andMarkovitz, A., Swankoski, K., Duda, N., Machta, R., and Urato, C and, M. F. Independent evaluation of comprehensive primary care plus (cpc+) second annual report: Appendices to the supplemental volume, 2020.

Hahn, P. R., Dorie, V., and Murray, J. S. Atlantic causal inference conference (acic) data analysis challenge 2017, 2019.

Hahn, P. R., Murray, J. S., and Carvalho, C. M. Bayesian regression tree models for causal inference: regularization, confounding, and heterogeneous effects. *Bayesian Analysis*, 2020.

He, J. and Hahn, P. R. Stochastic tree ensembles for regularized nonlinear regression. *Technical report*, 2020.

He, J., Yalov, S., and Hahn, P. R. XBART: Accelerated Bayesian additive regression trees. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pp. 1130–1138, 2019.

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

Imbens, G. W. and Rubin, D. B. *Causal inference in statistics, social, and biomedical sciences*. Cambridge University Press, 2015.

King, C., Escallier, K., Ju, Y.-E., Lin, N., Palanca, B. J., McKinnon, S., and Avidan, M. Obstructive sleep apnoea, positive airway pressure treatment and postoperative delirium: protocol for a retrospective observational study. *BMJ Open*, 9:e026649, 08 2019. doi: 10.1136/bmjopen-2018-026649.

Tibshirani, J., Athey, S., and Wager, S. *grf: Generalized Random Forests*, 2017. R package version 1.2.

Wendling, T., Jung, K., Callahan, A., Schuler, A., Shah, N., and Gallego, B. Comparing methods for estimation of heterogeneous treatment effects using observational data from health care databases. *Statistics in Medicine*, 2018.

Yeager, D. S., Hanselman, P., Walton, G. M., Murray, J. S., et al. A national experiment reveals where a growth mindset improves achievement. *Nature*, 573(7774):364–369, 2019. doi: 10.1038/s41586-019-1466-y. URL https://doi.org/10.1038/s41586-019-1466-y.