Bayesian Additive Regression Trees*

Adam Schaffroth

John Hopkins University

Abstract. The identification of causal effect can be difficult enough in a designed experiment, let alone from observational data. This paper demonstrates applications of a nonparametric model proposed by Hill [5] called Bayesian Additive Regression Trees (BART). This model is a good choice compared to others because of its simple implementation, ability to handle higher dimensional data, outputs clear uncertainty intervals, adeptly handles continuous treatment variables, and interpolates response variables well. In the paper BART will be compared against the performance of several other models.

1 Introduction

Collecting data from a controlled and randomized experiment is not always possible. Often times we must settle for an experiment with a treatment and observed "confounding" data that is assessed before treatment is applied. Understandably, this can make the identification of causal effects much more difficult and perhaps complicated. Recent developments have proposed semi-parametric and non-parametric models (e.g. matching methods) [4] with decent results . Many of these other models require fitting a model for both the treatment assignment and the response surface.

It is proposed that BART models are flexible and robust, while being simple to implement relative to the complexity of the problem[5]. The BART model[2] is an iterative Bayesian backfitting Markov Chain Monte Carlo algorithm that produces samples from a posterior distribution. BART performs well even in scenarios where the "perfect" model for the distribution is linear regression, thus showing that it is flexible enough to compete with more specific models in both the linear and nonlinear cases.

2 Primer on Causal Inference

Let Z be a binary, non-random, treatment variable where $Z = 1$ inidicates that the subject has been assigned a given treatment, and $Z = 0$ indicates no treatment will be applied. Note that a binary treatment variable is used for the sake of simplicity but non-binary treatments may also be analyzed (these are called dosage effects). There are several ways to analyze the causal effect of a

* Based on "Bayesian nonparametric modeling for Causal Inference" - Hill, 2011

treatment. Here, the focus will be on estimating the difference $Y_i(1)-Y_i(0)$, where $Y_i(1)$ and $Y_i(0)$ are the outcomes for observation i when treatment is applied or not applied. Notice that $Y_i(1) - Y_i(0)$ can only be estimated, since it would be impossible to observe both treatments on the same subject. In essence, we are only observing $Y_i = Y_i(1)Z_i+Y_i(0)(1-Z_i)$. There are a couple of common metrics used to estimate average treatment effects. In these papers, it is demonstrated that a "correct" treatment assignment similarly to a randomized experiment, thus the response surface does not need modeling. It can also be shown that the converse is also true; that is, if the response surface can be accurately modeled, the treatment assignment does not need to be modeled. This becomes on of the advantages of BART models that makes implementation so simple.

1. sample average treatment effect (SATE) $\sum_{i=1}^{n}[Y_i(1) - Y_i(0)]$.
2. sample average effect of the treated (SATT) $\sum_{i:Z_i=1}[Y_i(1) - Y_i(0)]$
3. population average treatment effect (PATE) $E[Y(1) - Y(0)]$
4. population average effect of the treated (PATT) $E[Y(1) - Y(0)|Z = 1]$
5. conditional average treatment effect (CATE) $\sum_{i=1}^{n} E(Y_i(1) - Y_i(0) \mid X_i)$
6. conditional average treatment effect for the treated (CATT) $\sum_{i:Z_i=1} E(Y_i(1)-Y_i(0) \mid X_i)$

Typically, observational data has potential outcomes that are dependent on treatment assignment. Average causal effects can be estimated with the assumption of strong ignorability of treatment assignment. This strong ignorability assumption is made up of two parts:

1. Unconfoundedness Assumption: $Y(0), Y(1) \perp\!\!\!\perp Z \mid X$ where X is a matrix of confounding covariates observed before treatment
2. Support/Overlap Assumption: $0\Pr(Z = 1 \mid X)1$

From these assumptions, we are able to estimate

$$E[Y(1) \mid X] = E[Y \mid X, Z = 1]$$

and

$$E[Y(0) \mid X] = E[Y \mid X, Z = 0]$$

Estimating these conditional expectations becomes more difficult as the number of confounding covariates grows (i.e. the dimension of X) or if it is unknown which predictors are required to satisfy the ignorability assumption. Propensity scoring models can be especially weak as the number of confounding covariates grows. More recent developments in include using matching models[1] and multivariate adjustment by subclassification[9] to estimate the conditional expectation with higher dimensional data. Estimators that are consistent when either the treatment assignment or the response surface is modeled are called "doubly robust" estimators[6]. While semiparametric and nonparametric models tend to be more robust than their parametric counterparts, most of the former require more intimate knowledge of the research space. This emphasizes another advantage of the BART models; they are easier to implement.

3 BART Models

The BART model can be generally expressed as $Y = f(z, x) + \epsilon$ where z represents the treatment, x represents the confounding covariates, and the errors ϵ are assumed to be independent and identically distributed with distribution $N(0, \sigma^2)$. Note that additive errors are assumed, that is ϵ is independent of $f(z, x)$. The ignorability assumption makes the contenders possible models flexible, however when compared to other models such as random forests, boosting, bagging, and deep learning (as in "Elements of Statistical Learning[3] BART models stand out. BART offers a simpler parameter optimization process as well as more inference into the model's confidence.

3.1 BART as a Sum-of-Trees Model

BART is a sum-of-tree model. Let T denote a binary tree where the splits leading down the leaf node are decision rules. At each split, a tuple (z, x) is input and a direction is taken. The bottom leaf node j has parameter μ_j representing the mean response of that subgroup of observations for that node. Let $M = \mu_1, ..., \mu_b$ where b is the number of leaf nodes. Let the function $g(z, x; T, M)$ output the value μ of the leaf node obtained by inputting (z, x) at the top of the tree. The BART model is a sum of these trees with an added regularization term (or prior).

$$Y = g(z, x; T_1, M_1) + g(z, x; T_2, M_2) + \cdots + g(z, x; T_m, M_m) + \epsilon,$$

If we calculated the residual of Y whilst leaving out the single tree $g(z, x; T_1, M_1)$ and repeated this m times, we would have a boosted model that risks overfitting. For this reason the regularization term is added. This term essentially "prunes" each tree. In Bayesian terms, (T_j, M_j) and are parameterized with a prior and a posterior is calculated via Markov Chain Monte Carlo. Each iteration of MCMC redraws (T_j, M_j) and . Trees are stochastically swapped out in an effort to find an optimal estimating function f. The prior is made up of three properties:

1. prior that prefers trees with minimal leaf nodes
2. M_j is shrunk towards zero (the centered mean response) by the prior
3. a prior that recommends a σ that is less than that of a least squares estimate

The boosting aspect of this model eliminates the need for cross-validation. Additionally, when a default prior is provided, Bayesian posterior uncertainty measure become available. The sum-of-trees models accurately captures non-linearities and interaction effects, which in other models are typically required input on by the model developer. Essentially, the BART model is a combination of a boosted model of weak-learners within a Bayesian framework. The Bayesian component is necessary for uncertainty inference.

BART models also shine in their ability to deal with higher dimension confounding matrices. Propensity scoring[7] is well known for handling decently large numbers of covariates as well but can struggle once this number gets too

high. BART can outclass propensity scoring in this regard which is significant because when a larger number of covariates are included in the model, the ignorability assumption becomes more likely to be actually true.

3.2 Causal Effect Estimation

The two estimates deemed most relevant for causal estimation are conditional average treatment effect (CATE) and the conditional average treatment effect for the treated (CATT):

$$\textbf{CATE} = \frac{1}{n}\sum_{i=1}^{n} E(Y_i(1) \mid X_i) - E(Y(0) \mid X_i) = \frac{1}{n}\sum_{i=1}^{n} f(1, x_i) - f(0, x_i),$$

$$\textbf{CATT} = \frac{1}{n_t}\sum_{i,Z_i=1} E(Y_i(1) \mid X_i) - E(Y(0) \mid X_i) = \frac{1}{n_t}\sum_{i,Z_i=1} f(1, x_i) - f(0, x_i).$$

More notation: Let $c(x, f) \equiv f(1, x) - f(0, x)$ be the treatment effect where $X = x$. Let $C(f) \equiv (c(x_1,\, f),\, c(x_2,\, f),\, \ldots\, c(x_K,\, f))$ be the joint posterior distribution over part of the sample space. Let f^l be the l^{th} draw from f. Let $\{x_i\}_1^k$ be the empirical distribution used to derive CATE or CATT. The average vector C^l at each draw l would be $\bar{C}^l = \frac{1}{K}\sum_i^K c(x_i,\, f^l)$. This $\bar{C}^l$, should theoretically represent the CATE of the entire jointly dependent distribution. Uncertainty intervals are calculated by finding quantiles from the marginal posterior distributions (i.e. $c(x_i, f^l)$).

4 Application

To demonstrate how BART can be applied to causal inference, the "smartpill" dataset from "se of Wireless Utility Capsule to Determine Gastric Emptying and Small Intestinal Transit Times in Critically Ill Trauma Patients"[8] will be used. This data regards gastric emptying, small bowel transit time, and total intestinal transit time in 8 patients in critical condition. A separate trial provided 87 healthy volunteers to augment the data. A motility capsule wirelessly transmitted pH, pressure, and temperature to a recorder attached to each subject's abdomen. The study's objective was to use the new motility capsule technology to asses the affect of the relationship of gastric emptying and small bowel transit times in critically ill trauma patients with intracranial hemorrhages.

The treatment variable is assigned to the "group" variable in the data set, where 0 := Critically Ill Trauma Patient and 1 := Healthy Volunteer. The response variable is GE.Time i.e. Time is time from ingestion to gastric emptying. The variables included in the confounding covariate matrix are height, weight, age, SB.Time (Small Bowel Transit Time is the time from gastric emptying to

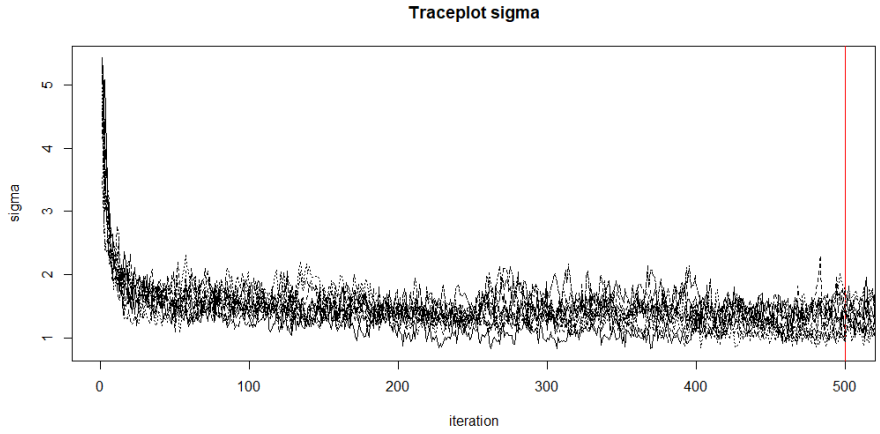
ileocecal junction), and WG.Time (Whole Gut Time is the time from ingestion to body exit). The rest of the variables are discarded due to incompleteness. The package used is the "bartCause" package in R in which the "bartc" function fits a model for estimating causal effects.

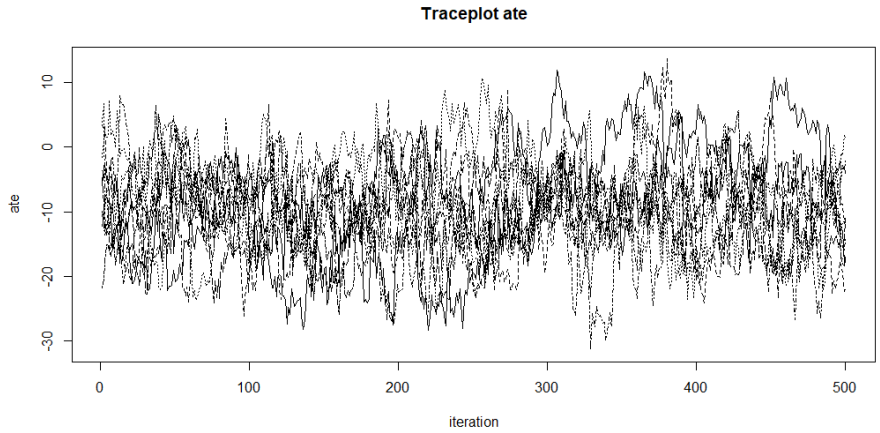
4.1 R Code

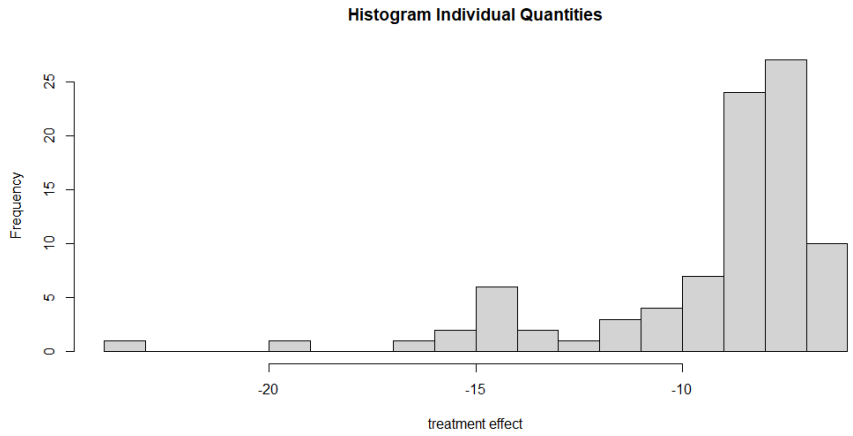
```
> library(medicaldata)
> library(bartCause)
> data("smartpill")
> keep = c('Group', 'Gender', 'Height', 'Weight',
    'Age', 'GE.Time', 'SB.Time', 'WG.Time')

> smartpill = smartpill[,keep]
> smartpill <- na.omit(smartpill)

> covar <- within(smartpill, rm(Group,GE.Time))

> fit <- bartc(response = smartpill$GE.Time,
+              treatment = smartpill$Group,
+              confounders = covar,
+              verbose = T)
fitting treatment model via method 'bart'
fitting response model via method 'bart'


> summary(fit)
Call: bartc(response = smartpill$GE.Time,
            treatment = smartpill$Group,
            confounders = covar, verbose = T)


Causal inference model fit by:
  model.rsp: bart
  model.trt: bart

Treatment effect (population average):
    estimate   sd ci.lower ci.upper
ate    -9.45 6.95   -23.07    4.171
Estimates fit from 89 total observations
95% credible interval calculated by: normal approximation
  population TE approximated by: posterior predictive distribution
Result based on 500 posterior samples times 10 chains
```

4.2 Model Analysis

The parameters of the function are rather straight forward, with the exception of the confounding covariate matrix from which the response and treatment variables must be discarded. The summary function shows that both the response surface and the treatment assignment were modeled by the bart algorithm by default (other choices include Targeted Minimum Loss based Estimation adjustment (TMLE), generalized linear models, and propensity score weightings). The "ate" shows the model's estimate for average treatment effect which is negative with seemingly substantial variation. The lower and upper bounds of the Bayesian uncertainty interval are also provided. The training set consisted of 89 observations with 500 iterations of the MCMC algorithm. The first two traceplot show the value of σ and the average treatment effect (ate), respectively over each of the 500 MCMC iterations. The third plot shows a histogram of the distribution of the estimates of each effect, averaged on their posterior samples.







References

[1] Alberto Abadie and Guido W. Imbens. "Large sample properties of matching estimators for average treatment effects". In: *Econometrica* 74.1 (2006), pp. 235–267. DOI: 10.1111/j.1468-0262.2006.00655.x.

[2] Hugh A. Chipman, Edward I. George, and Robert E. McCulloch. "Bart: Bayesian additive regression trees". In: *The Annals of Applied Statistics* 4.1 (2010). DOI: 10.1214/09-aoas285.

[3] Trevor Hastie, Jerome Friedman, and Robert Tibshirani. "The elements of Statistical Learning". In: *Springer Series in Statistics* (2001). DOI: 10.1007/978-0-387-21606-5.

[4] J. J. Heckman, H. Ichimura, and P. E. Todd. "Matching as an econometric evaluation estimator: Evidence from evaluating a job training programme". In: *The Review of Economic Studies* 64.4 (1997), pp. 605–654. DOI: 10.2307/2971733.

[5] Jennifer L. Hill. "Bayesian nonparametric modeling for Causal Inference". In: *Journal of Computational and Graphical Statistics* 20.1 (2011), pp. 217–240. DOI: 10.1198/jcgs.2010.08162.

[6] Joseph D. Kang and Joseph L. Schafer. "Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data". In: *Statistical Science* 22.4 (2007). DOI: 10.1214/07-sts227.

[7] Stephanie T. Lanza, Julia E. Moore, and Nicole M. Butera. "Drawing causal inferences using propensity scores: A practical guide for community psychologists". In: *American Journal of Community Psychology* 52.3–4 (2013), pp. 380–392. DOI: 10.1007/s10464-013-9604-4.

[8] Stefan Rauch et al. "Use of wireless motility capsule to determine gastric emptying and small intestinal transit times in critically ill trauma patients". In: *Journal of Critical Care* 27.5 (2012). DOI: 10.1016/j.jcrc.2011.12.002.

[9] Paul R. Rosenbaum and Donald B. Rubin. “The central role of the propensity score in observational studies for causal effects.” In: (1981). DOI: 10.21236/ada114514.