Bayesian Nonparametric Modeling for Causal Inference

Jennifer L. Hill, Columbia University[1]

Robert E. McCulloch, University of Chicago

June 7, 2007

Abstract

Researchers have long struggled to identify causal effects in non-experimental settings. Many recently-proposed strategies assume ignorability of the treatment assignment mechanism and require fitting two models – one for the assignment mechanism and one for the response surface. We propose a strategy that instead focuses on very flexibly modeling just the response surface using a Bayesian nonparametric modeling procedure, Bayesian Additive Regression Trees (BART). BART has several advantages: it is far simpler to use than many recent competitors, requires less guess-work in model fitting, handles a large number of predictors, yields coherent uncertainty intervals, fluidly handles continuous treatment variables and missing data for the outcome variable. BART produces more efficient estimates in the non-linear situations tested in our simulations compared to propensity score matching, propensity-weighted estimators, and regression adjustment. Further, it is highly competitive in linear settings with the "correct" model, linear regression.

[1] Correspondence should be addressed to the first author at Columbia University, School of International and Public Affairs, 420 West 118th St., 1425 IAB, New York NY 10027

1 Introduction

Causal inference is challenging in the absence of a controlled experiment or natural experiment that randomizes treatment assignment. Often researchers assume ignorability of the treatment assignment conditional on observed pre-treatment or “confounding” covariates. The most appropriate modeling choices for estimation of treatment effects under ignorability are still debated however (see Imbens 2004, for a review).

Many causal methods for observational data (Rubin 1973, 1979), including many methods currently discussed (for instance, Robins *et al.* 1995; Heckman *et al.* 1997; Robins and Ritov 1997; Robins *et al.* 2000; Rubin and Thomas 2000; Hirano and Imbens 2001; Hirano *et al.* 2003; Sato and Matsuyama 2003; Abadie and Imbens 2006), involve fitting models for both the treatment assignment mechanism and the distribution of the outcome (or potential outcomes) conditional on the treatment and confounding covariates. The latter distribution will henceforth be referred to as the response surface. Appropriately adjusting for the treatment assignment mechanism reduces reliance on modeling assumptions for the response surface.

Given recent advances in Bayesian nonparametric models with extremely flexible functional form, we propose that a simpler, yet robust, modeling approach is now available for efficiently estimating causal effects in this setting. We focus solely on precise modeling of the response surface using a nonparametric (or, equivalently, very highly parametric) modeling strategy called Bayesian Additive Regression Trees, or BART (Chipman, George, and McCulloch 2006, 2007).

The BART algorithm is straightforward to implement and requires the researcher only to input the outcome, treatment assignment, and confounding covariates, but requires no information about how these variables are parametrically related. Also, BART naturally produces coherent posterior intervals in contrast to methods such as propensity score matching and subclassification, for instance, for which there is still no agreement regarding appropriate interval estimation (Du 1998; Tu and Zhou 2003; Imbens 2004; Hill and Reiter 2006). Finally, treatment effect point estimates calculated using BART appear to be substantially more accurate (for instance as measured by root mean squared error) in the non-linear settings considered here than estimates from equally-accessible competitors such as linear regression, propensity-weighted estimators, and propensity score matching with regression adjustment. Even when the response surface is linear with additive treatment effects, BART's performance in our simulations is almost indistinguishable from linear regression, the "correct" model for that setting. Therefore we believe BART is a simple method that has the potential to be both robust and efficient in the estimation of causal effects.

Section 2 of this paper discusses the estimation problem and traditional methods. Section 3 describes BART and how we will use it to estimate causal effects. Section 4 evaluates the method using a classic constructed observational study. Section 5 presents simulations using treatment and covariate data from a real study. Section 6 briefly discusses the added complications of continuous treatment variables and outcome missing data. Section 7 concludes.

2 Causal Inference

We consider the causal effect of binary treatment Z, with $Z = 1$ indicating assignment to treatment and $Z = 0$ indicating assignment to control (the ideas here can be extended to multi-valued or continuous treatment variables; we discuss in the context of a binary treatment variable for ease of exposition). We follow convention in the statistical causal inference literature (e.g. Rubin 1978) and define a causal effect of Z for individual i as $Y_i(1) - Y_i(0)$. $Y_i(0)$ and $Y_i(1)$ are the *potential outcomes* for individual i, that is, the outcomes that would be observed under $Z = 0$ and $Z = 1$, respectively. Since we cannot observe both $Y_i(1)$ and $Y_i(0)$ for any individual we generally focus on estimating average treatment effects such as the average treatment effect for the entire population, $E[Y(1) - Y(0)]$, the average effect of the treatment on the treated, $E[Y(1) - Y(0) \mid Z = 1]$, or the average effect of the treatment on the controls, $E[Y(1) - Y(0) \mid Z = 0]$. Since we assume observations are independently and identically distributed we can drop the subscript notation for simplicity.

In observational studies the potential outcomes are generally not independent of treatment assignment. However we can estimate any of the average causal effects above under the assumption of strong ignorability of treatment assignment, which requires $Y(0), Y(1) \perp Z \mid X$, and $0 < Pr(Z = 1 \mid X) < 1$, where X represents a vector of observed pre-treatment variables or "confounding covariates" (the effect of the treatment on the treated and the effect of the treatment on the controls require slightly weaker versions of this assumption). However, estimation of these causal effects now requires evaluation of the response surfaces $E[Y(1)] = E[E[Y \mid X, Z = 1]]$ and $E[Y(0)] = E[E[Y \mid X, Z = 0]]$, where X is potentially high-dimensional.

Given this framework the question becomes how best to estimate the relevant conditional expectations. This estimation may be difficult if Y is not linearly related to X and if the distribution of X is quite different across treatment groups. A simple hypothetical example of such a scenario is illustrated in Figure 1 where the X needed to satisfy ignorability is one dimensional. These 120 data points were generated independently as follows: $Z \sim \text{Bernoulli}(.5)$, $X \mid Z = 1 \sim \text{N}(40, 10^2)$, $X \mid Z = 0 \sim \text{N}(20, 10^2)$, $Y(0) \sim \text{N}(72 + 3\sqrt{X}, 1)$, $Y(1) \sim \text{N}(90 + \exp(.06X), 1)$. In the left panel, the upper

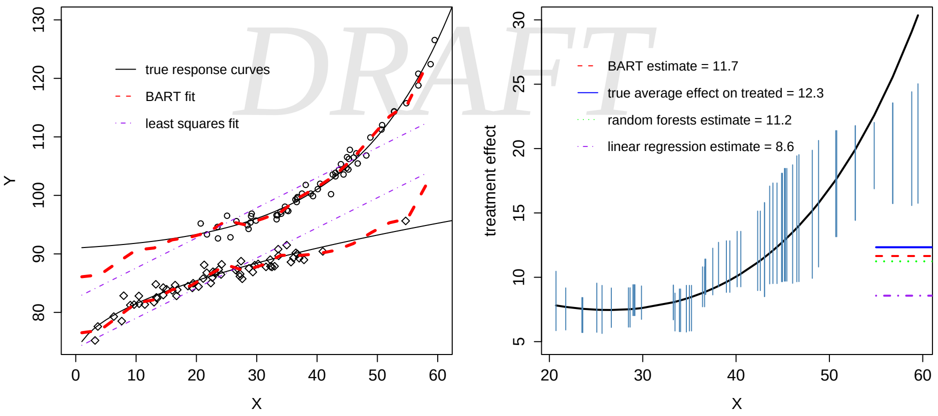

Figure 1: Left panel: simulated data with linear regression and BART fits. Right panel: BART inference for treatment effect on the treated.

dark solid curve represents $E[Y(1) \mid X]$ and the lower one $E[Y(0) \mid X]$. The dots close to the upper curve are the treated observations and the triangles close to the lower curve are the untreated. $E[Y(1) - Y(0)|Z = 1]$ is a typical estimand of interest. Its true value in this simulation is 12.3.

The parallel dot-dash lines display a linear regression fit to the data that yields an underestimate, 8.6 (standard error .7) not only of the treatment on the treated but the intended estimand, $E[Y(1) - Y(0)]$ (10.4) as well. Propensity score strategies were used to estimate the effect of the treatment on the treated for this example. Matching yields an estimate of 11.8 (standard error .7) while the inverse-probability-of-treatment weighted estimator described in Section 5 yields an estimate of 8.4 (standard error .9). Using random forests to flexibly fit the response surface yields an estimate of 11.2.

As a quick introduction to the appeal of BART, the dashed line in the left panel of Figure 1 displays the BART fit to the data and the right panel displays the BART inference for the treatment effect on the treated. In the right panel, the true treatment effect as it varies with X is plotted as the solid curve. The vertical segments are marginal 95% posterior intervals for the treatment effect at each X value from a treated observation. Notice that the uncertainty bounds grow much wider in the range where there is no overlap across treatment groups – that is, where we don't observe empirical counterfactuals for each data point (e.g. $X > 40$). The intervals nicely cover the true treatment effect except at very large values of X. The BART point estimate (posterior mean) of the average effect of the treatment on the treated is 11.7 with 95% posterior interval (10.3, 13.0). The computation of these intervals and the random forests estimate is discussed in section 3.

The difficulty in estimating these conditional expectations is exacerbated when, more plausibly, there are many confounding covariates or uncertainty about which predictors are needed to satisfy ignorability (for a study that conditions on a very large number of covariates in order to justify ignorability see Bingenheimer *et al.* 2005).

A host of new methodological strategies have been proposed in the past three decades to address this estimation problem. Some methods focus primarily on appropriately controlling for the treatment assignment mechanism such as with subclassification or matching methods (for example Rosenbaum and Rubin 1983, 1985; Gu and Rosenbaum 1993; Abadie and Imbens 2006), parametric inverse probability weighting approaches (for example Rosenbaum 1987; Robins *et al.* 2000), or semi-parametric inverse probability weighting (for example, Hahn 1998). The intuitive motivation here is that if the treatment assignment mechanism is properly specified we don't have to model the response surface at all (just as in a randomized experiment).

Not only is correct specification of the assignment mechanism sufficient, the same holds for the response surface. That is, if the the response surface is correctly specified we don't have to worry about correctly specifiying the assignment mechanism. In this spirit, many recent articles have focused on the potential advantages in robustness (and potentially efficiency) that may be achieved by modeling both the treatment assignment mechanism *and* the response surface. Rubin and Thomas (2000); Kurth, Walker, Glynn, Chan, Gaziano, Berger, and Robins (2006) present relatively simple approaches. The weighting estimators in this family have also been extended to address efficiency issues (Robins and Rotnitzky 1995; Robins *et al.* 1995; Rotnitzky and Robins 1997; Hahn 1998; Scharfstein *et al.* 1999; Hirano *et al.* 2003). Moreover, semi-parametric estimators of this sort, such as those discussed in Robins, Rotnitzky, and Zhao (1994); Robins and Rotnitzky (1995); Rotnitzky, Robins, and Scharfstein (1998), were shown in Scharfstein, Rotnitzky, and Robins (1999) to be consistent as long as either the treatment assignment mechanism *or* the response surface is correctly specified. Estimators with this property are called "doubly robust" or "doubly protected" (Carpenter, Kenward, and Vansteelandt 2005).

Our approach is different from most others because we focus solely on precise estimation of the response surface (see Hahn 1998, for another approach to modeling the response surface). We believe that the benefits in terms of simplicity, efficiency, robustness, and lack of required researcher interference outweigh the potential for gains in consistency that may be achieved by competing methods (Rosenbaum 1987; Robins and Rotnitzky 1995; Heckman *et al.* 1997, for instance,) that require more choices regarding model specification (correct specification of a parametric model or specification of nonparametric smoothing parameters such as the number of terms in a series estimator or the bandwidth for a kernel estimator). The semi-parametric versions of these methods are more robust but also require a higher level of researcher sophistication to understand and implement.

3 BART and Estimating Causal Effects

BART is designed to estimate a model specified very generally as $Y = f(z, x) + \epsilon$, where ϵ are iid $N(0, \sigma^2)$, z denotes the assigned treatment, and x denotes predictors. The BART model assumes additive errors, however its inference for f is very flexible. If ignorability holds conditional on x, that is $Y(0), Y(1) \perp Z \mid X$, then we posit

$$\begin{aligned} E[Y(0) \mid X = x] &= E[Y \mid Z = 0, X = x] = f(0, x) \\ E[Y(1) \mid X = x] &= E[Y \mid Z = 1, X = x] = f(1, x). \end{aligned}$$

In principle, any method that flexibly estimates f could be used. We believe BART has important advantages over alternative methods such as random forests, boosting, and neural nets (see section 8.7, chapter 10, and chapter 11 of Hastie, Tibshirani, and Friedman (2001)). BART is developed in Chipman, George, and McCulloch (2006,2007) (henceforth CGM06 and CGM07, and CGM generally).

In section 3.1, we give an overview of BART (mostly from CGM07) emphasizing those aspects essential to this paper. In section 3.2 we discuss its advantages in causal inference and in section 3.3 we detail how BART is used to estimate causal effects.

3.1 Overview of BART

BART builds upon tree models. First we establish notation for a single tree model. Let T denote a binary tree. All of the interior nodes of T have decision rules which send a (z, x) pair either left or right. The i^{th} bottom node has a parameter μ_i associated with it and we let $M = \{\mu_1, \mu_2, \ldots, \mu_b\}$ where b is the number of bottom nodes. The left panel of Figure 2 depicts a tree model fit to the data in Figure 1. The decision rules in the figure are the criteria for sending a (z, x) pair left. Note that z is encoded as a 0-1 dummy variable so the first decision rule sends all the untreated left. The tree T

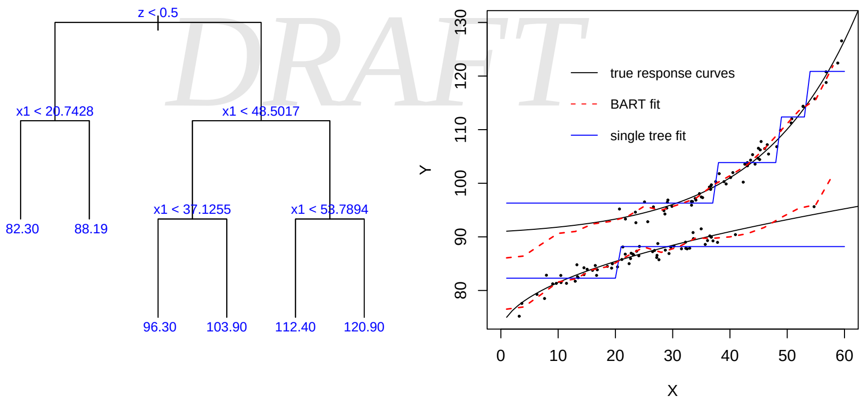

Figure 2: Left panel: the binary tree fit to the data from Figure 1. Right panel: single-tree fits (solid lines) and BART fits (dashed lines).

consists of the tree structure and all the decision rules leading down to a bottom node. In Figure 2, M is the six numbers associated with the six bottom nodes. Given the tree model (T, M) and a pair (z, x) we define $g(z, x; T, M)$ as the value obtained by first dropping (z, x) down the tree until it hits a bottom node and then reporting the μ associated with that bottom node.

Using the (T, M) in Figure 2, $g(1, 40; T, M) = 103.9$. In the right panel of Figure 2 the flat segments plot the function $g(z, x; T, M)$ against x, with each segment corresponding to a bottom node. The dashed curve is the BART fit. It is so close to the true response surface that it is difficult to see, except when there is no overlap.

The sum-of-trees model

BART consists of two pieces: a sum-of-trees model and a regularization prior. The sum-of-trees model lets,

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

where each (T_i, M_i) denotes a single sub-tree model and $\epsilon \sim N(0, \sigma^2)$. We have $Y = f(z, x) + \epsilon$ with $f(z, x) = \sum g(z, x; T_i, M_i)$.

The relative strength of the sum-of-trees model is not immediately obvious. Why not just make the first tree big enough to capture f? In this model, the first tree is not designed to fit y, as was the tree depicted in Figure 2, but rather only the part of y not explained by the other trees. Consequently this model can capture a great deal of complexity while at the same time finding additive structures far more easily than large regression trees.

In the spirit of boosting, CGM let the number of sub-tree models, m, be large, allowing each to contribute only a small part to the overall fit. This is achieved through a regularization prior which holds back the fit of each sub-tree. This prior also helps BART avoid over-fitting. In the CGM approach, the (T_i, M_i) and σ are treated as parameters in a formal statistical model rather than just algorithmically, as in much of the data-mining literature. A prior is put on the parameters, and the posterior is computed using Markov Chain Monte Carlo.

The Prior

CGM simplify the prior specification by letting the T_i be i.i.d, the $\mu_{i,b}$ (node b of tree i) be i.i.d conditional on the set of T_i, and σ independent of all T and μ. Now it is only neccessary to specify marginal priors for a single T, a single μ, and σ. In order to "hold back" the fit of each single sub-tree model, the prior on T puts prior weight on smaller trees and the prior on μ shrinks the fit of each node. The amount of shrinkage increases with m, the number of sub-trees, since the more sub-trees we use, the smaller the contribution of each should be. Thus, the sub-trees are related in that the more there are of them, the smaller the contribution of each. The prior on σ represents beliefs about the strength of the overall fit.

CGM provide default settings for the prior. They show that BART, with the default prior settings, is competitive with alternative methods (using cross-validation to choose parameter settings) in terms of out-of-sample prediction. The default prior allows BART to fit and not over-fit. This finding greatly enhances BART's usefulness as a tool for the inference of causal effects. All of the results reported in this paper our based the default prior. The default value for m is 200. The prior on T is somewhat complex and the reader is again referred to CGM. This prior choice is not influential. The priors on μ and σ are simple and important. We briefly outline CGM's choices and explain the default setting.

To place a prior on the μ, CGM standardize y so that $E(Y \mid z, x)$ is in the interval $[-.5, .5]$ with high probability and let $\mu \sim N(0, \sigma_\mu^2)$. Conditional of the set of T_i, $E(Y \mid z, x) = f(z, x) = \sum_{i=1}^{m} g(z, x; T_i, M_i) = \sum_{i=1}^{m} \mu_i$ where each μ_i corresponds to a bottom node from a different one of the sub-trees. Since the μ_i are iid $N(0, \sigma_\mu^2)$, the sum has standard deviation $\sqrt{m}\,\sigma_\mu$. Set $\sigma_\mu = \frac{.5}{k\sqrt{m}}$. The idea is that $E(Y \mid z, x)$ is then k standard deviations away from ±.5. The CGM default choice is $k = 2$. The default standardization makes the range of the observed y equal to $[-.5, .5]$. Note that the prior depends on m.

To place a prior on σ, CGM start with the usual $\sigma^2 \sim \frac{\nu\lambda}{\chi_\nu^2}$, where χ_ν^2 denotes a chi-squared random variable with ν degrees of freedom. They choose ν and λ so that $P(\sigma < \hat{\sigma}) = q$. The default choices set $\hat{\sigma}$ to be the usual linear regression estimate, $q = .9$, and $\nu = 3$, reflecting a 90% "chance" that σ is less than the least-squares value.

Finally, to compute the posterior, CGM develop a Markov chain monte carlo algorithm. MCMC iteratively produces (dependent) draws from the posterior. At iteration r of the algorithm we have a draw from the posterior of all of the model parameters: $\{T_i^r\}$, $\{M_i^r\}$, and σ^r. This gives us draws from the marginal posterior of f through the relation $f^r(z, x) = \sum g(z, x; T_i^r, M_i^r)$. Our use of BART in estimation of causal effects depends only on the draws f^r.

3.2 Advantages of BART for Causal Inference

Now that the essential elements of BART have been laid out, we motivate its potential advantages for use in causal estimations.

The basic sum-of-trees specification is wonderfully adept at capturing both nonlinearities and interactions without the researcher having to explicity add interaction terms or transformations of x or specify a limit on the level of interaction. To see how BART captures nonlinearities, consider the case where each sub-tree uses only one component of (z, x) in all of the interior decision nodes. In this case, the overall fit is additive (like a generalized additive model). By combining a large number of single-tree fits of the kind depicted in Figure 2, just about any kind of f can be nicely fit. Interactions are captured whenever a sub-tree model involves more than one component of (z, x). In each iteration of the CGM MCMC algorithm, each sub-tree can grow or become smaller allowing BART to naturally infer the level of interaction.

The CGM MCMC is very stable. If you run the method twice (with two different seeds) you get the virtually the same results. In practice, it is very easy to see when the chain has burned in by looking at the draws of the one interpretable parameter σ. Initially, the draws of σ will fall, as the algorithm searches through the space of f finding the fit. After these draws level off, the remaining variation represents the natural Bayesian quantifiction of our posterior uncertainty. The stability of the CGM MCMC and the success of the default prior (or, equivalently, the insensitivity of the results to reasonable changes in the prior) allow us to view this posterior variation as a reasonable assessment of our uncertainty. BART combines the advantages of boosting weak learning models with that of working within a statistical framework.

While other modern data-mining methods can provide excellent fit, they typically involve the use of cross-validation to find suitable values for tuning parameters used to avoid overfitting. Consequently, to obtain a measure of uncertainty, the whole procedure would have to be bootstrapped resulting in an overall methodology which we feel would be daunting to most practitioners. In fact, it is extremely rare to see applied researchers reporting uncertainty associated with fits from single trees, neural nets, random forests or boosting while it is a natural by-product of a BART run.

A residual benefit of this strategy is the lack of "tinkering" required. Researchers tend to condition on variables and use functional forms that fit their existing theory (see, for example, Leamer 1983). Moreover, when searching for the "best" model it may be difficult not to stop as soon as the model meets prior expectations or to bypass models that don't meet prior expectations. Such model-searching strategies, while understandable, particularly given strong theory regarding a phenomenon, have the potential to mask important but unexpected results and bias estimates.

In CGM07, BART with the default prior is shown to perform competitively with a variety of alternative methods in over 40 data sets. Of course, CGM do not claim the BART will outperform all other methods in all applications. For example, when there there is complex interaction we might expect random forests (a version of bagging with trees Breiman (2001)) to do better than BART which shrinks towards additive structure. We tried using random forests (randomForest package in R) and plugging in the resulting estimate of f. In the simple example, we get a good estimate (see the right panel of Figure 1). In the much more realistic Lalonde example, the estimates from random forests are very poor (see Figure 3).

3.3 Estimating causal effects

Primarily we will use BART to estimate average causal effects such as

$$E(Y(1) \mid X = x) - E(Y(0) \mid X = x) = f(1, x) - f(0, x).$$

Each iteration of the BART Markov Chain generates a new draw of f from the posterior distribution. Let f^r denote the r^{th} draw of f. Let $\{x_i\}_1^K$ denote a set of x representing the distribution of x over which we wish to learn the average treatment effect. For instance we would choose $\{i : z_i = 1\}$ if we wanted to estimate the effect of the treatment on the treated.

Let $c(x, f) = f(1, x) - f(0, x)$. Our interest is then in the joint posterior distribution of $C(f) = (c(x_1, f), c(x_2, f), \ldots, c(x_K, f))$. For every draw, we obtain a draw from the joint posterior of $C(f)$: $C^r = C(f^r)$. As f varies, the entire vector C will vary. The Bayesian MCMC technology gives us a relatively simple way to uncover the joint posterior of C which will exhibit dependence inherited from f.

If we want inference for the average treatment effect, we simply compute the average of the vector C^r at each r, $\bar{C}^r = \frac{1}{K}\sum_i^K c(x_i, f^r)$. This gives draws $\bar{C}^r$ from the posterior distribution of the average treatment effect that reflect the full joint dependent distribution of C^r.

The intervals in the right panel of Figure 1 were obtained by computing the 2.5% and 97.5% quantiles of the set of draws $c(x_i, f^r)$ for each fixed i. These intervals display the marginal posterior distributions of each $c(x_i, f)$. In practice, displaying the full joint distribution of the high dimensional C is difficult. The dashed horizontal line in the right panel is the the average of the $\bar{C}^r$ over r which is the posterior mean of the average treatment effect on the treated. The histograms in Figure 3 (section 4) depict the entire set of values $\bar{C}^r$, and hence the entire posterior distribution of the average treatment effect.

4 LaLonde Example

We illustrate using a now classic example first constructed by LaLonde (1986) to test the efficacy of the then state-of-the-art econometric methods for causal inference in observational studies in the early 1980's. LaLonde created two potential observational studies by combining the treatment group from the randomized evaluation of the National Supported Work demonstration with control groups obtained from each of two large-scale public-use surveys from the same time period, the Panel Survey of Income Dynamics (PSID) and the Current Population Survey (CPS). Since the true experimental control group from the NSW also exists, causal estimates from the constructed observational studies were compared with the experimental benchmarks to judge efficacy. The conclusion from this paper was that none of the non-experimental econometric methods provided reliable answers. More recently researchers have used this dataset to demonstrate the efficacy of propensity score matching (Dehejia and Wahba 1999) and genetic matching algorithms (Diamond and Sekhon 2006).

The National Supported Work intervention provided job training to disadvantaged men. Using a randomized experiment, the effect of the program on 1978 earnings was estimated to be about $1800. Pre-treatment covariates available are ethnicity (black, hispanic, other), marital status, years of education, age, and 1974 and 1975 earnings. As in Dehejia and Wahba (1999) we use the subset of participants with two years of pre-treatment earnings available due to strong evidence in the economics literature (see, for example, Ashenfelter 1978; Ashenfelter and Card 1985) that relying on only one period of pre-treatment earnings generally will be insufficient to satisfy ignorability. This randomized study has 185 treated individuals and 260 controls.

We applied the BART methodology (keeping every tenth draw from the 10,000 draws after burn-in of 2000) to the two data sets obtained by combining the 185 treated observations with either 2490 PSID controls or the the 15,922 CPS controls. For the r^{th} kept BART draw we average $c(x_i, f^r) = f^r(1, x_i) - f^r(0, x_i)$ over the 185 x_i in the treated group giving 1,000 draws from the posterior distribution of the average treatment effect on the treated. For all BART runs, the default settings were used and no attempt was made to tune the parameters to give optimal results.

The two panels in Figure 3 display histograms of the draws from the posterior distribution of the average effect of the treatment on the treated (top: PSID controls, bottom: CPS controls). In each histogram a solid line is drawn at the estimate from the original study (the gold standard), a dashed line is drawn at the posterior mean, a dotted line is drawn at the estimate obtained using random forests, and a dot-dash line is drawn at the regression estimate (estimate of treatment coefficient in the linear regression of 1978 earnings on the the treatment indicator and the covariates listed above). In both cases, the BART estimate is very close to the gold standard, while the linear regression estimate and the one obtained using random forests to estimate f underestimate the effect. Quite reasonably, the posteriors indicate that the there is more information using the CPS controls ($n = 15,922$) than there is using the PSID controls ($n = 2,490$).

5 Simulations based on real data

Constructed observational studies have the weakness that we never know if ignorability is satisfied or whether sufficient overlap exists for valid estimation (if overlap exists for the marginal distributions of observed covariates but not the joint distributions it may be difficult to detect). Therefore we follow with a set of simulations based on real data

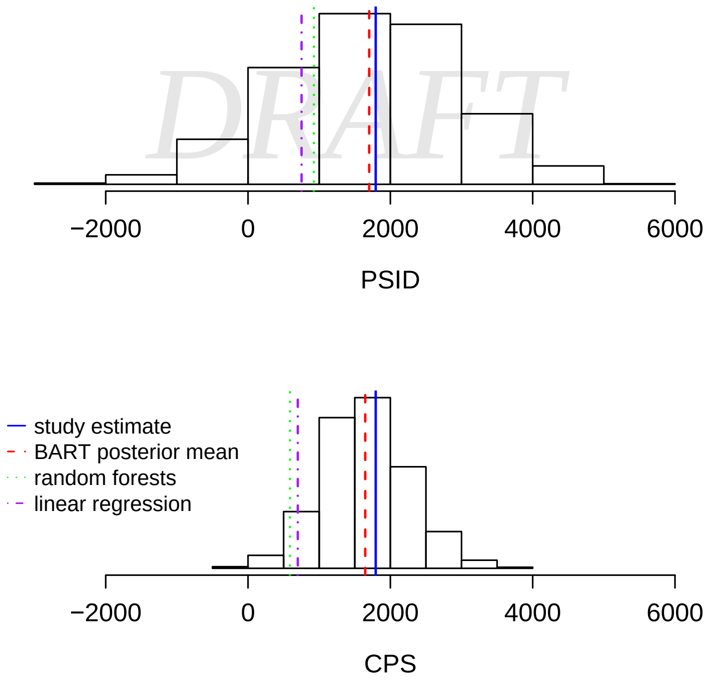

Figure 3: Posterior distributions of the mean effect of the treatment on the treated for the PSID controls (top), and the CPS controls (bottom).

in which we know that ignorability has been satisfied. We estimate treatment effects both when complete overlap exists and when there is lack of complete overlap.

We start with experimental data from the Infant Health and Development Program (IHDP), a randomized experiment that began in 1985, targeted low-birth-weight, premature infants, and provided the treatment group with both intensive high-quality child care and home visits from a trained provider. The program was highly successful at significantly raising cognitive test scores of the treatment children relative to controls at the end of the intervention when children were 3 years old (Infant Health and Development Program 1990; Brooks-Gunn, Liaw, and Klebanov 1991). The study collected data on many pre-treatment variables. We use measurements on the child – birth weight, head circumference, weeks born preterm, birth order, first born, neonatal health index (see Scott and Bauer 1989), sex, twin status – as well as behaviors engaged in during the pregnancy – smoked cigarettes, drank alcohol, took drugs – and measurements on the mother at the time she gave birth – age, marital status, educational attainment (did not graduate from high school, graduated from high school, attended some college but did not graduate, graduated from college), whether she worked during pregnancy, whether she received prenatal care – and the site (8 total) in which the family resided at the start of the intervention.

We use experimental data because it provides a setting within which we can create an observational study but still be guaranteed to have complete overlap for our treatment group. Starting with this experimental data we can create an observational study by throwing away a non-random portion of our treatment group, in particular, all children with non-white mothers. The control group remains intact. Thus our treatment and control groups are no longer balanced and in the absence of the indicator for mother's race comparison between these groups would lead to biased estimates both of the effect of the treatment on the full sample (the average treatment effect) and the effect of the treatment on the children of white mothers. The question is whether we can estimate either of these effects with the remaining data.

We chose ethnicity as the variable used to subset the data because it led to subgroups that were more distinct than those yielded by the other categorical variables. This leaves 139 children in our treatment group and 608 in our comparison group. We then use the $p = 25$ confounding covariates (ethnicity excluded) to generate three different response surfaces. Because the response surface is known, ignorability can be satified by appropriately conditioning on the confounding covariates used to generate the response surfaces.

This simulation design ensures overlap in the setting when we estimate the effect of the treatment on the treated and lack of overlap in the setting when we estimate the effect of the treatment on the controls. Moreover, while the response surface is known in each simulation design, the treatment assignment mechanism as a function of our 25 confounding covariates (i.e. excluding ethnicity), remains unknown.

5.1 Response surfaces

We consider three response surfaces. Response surface A takes the form

$$\begin{aligned} Y(0) &\sim N(X\beta_A, 1) \\ Y(1) &\sim N(X\beta_A + 4, 1) \end{aligned}$$

where X represents a matrix of standardized (mean 0 and standard deviation 1) covariate values (with first column equal to a vector of ones) and the coefficients in the vector β (length 25) are randomly sampled values (0,1,2,3,4) with probabilities (.5,.2,.15,.1,.05) that make smaller coefficients more likely. This response surface is linear and parallel across treatment groups. In this setting linear regression should trump both BART and the propensity-score-based methods (discussed in the next section) in terms of both closer estimates and reduced uncertainty because the strong parametric assumptions implicit in this model will be satisfied. The R^2 from a linear regression of Y on X for this response surface is about .95 on average across simulations.

Response surface B is non-linear and not parallel across treatment conditions.

$$\begin{aligned} Y(0) &\sim N(\exp((X + W)\beta_B), 1) \\ Y(1) &\sim N(X\beta_B - \omega_B^s, 1) \end{aligned}$$

where W is an offset matrix with the same dimension as X with every value equal to .5, β_B is a vector of regression coefficients (0,.1,.2,.3,.4) randomly sampled with probabilities (.5,.125,.125,.125,.125) for the 6 continuous covariates and (.6,.1,1,.1,.1) for the 18 binary covariates. For simulation s, ω_B^s was chosen so that $\text{Avg}_{i:z=1}(y_i(1) -$

$y_i(0)) = 4$ in the (overlap) setting where we focus on the effect of the treatment on the treated and $\text{Avg}_{i:z=0}(y_i(1) - y_i(0)) = 4$ in the (non-overlap) setting where we focus on the effect of the treatment on the controls. In this way the sample treatment effect for this response surface is always 4. R^2 values from the linear regression of outcomes on the X values is about .77 on average.

Response surface C is nonlinear and not parallel across treatment groups.

$$\begin{aligned} Y(0) &\sim N(Q\beta_{C0}, 1) \\ Y(1) &\sim N(Q\beta_{C1} + \omega^s_C, 1) \end{aligned}$$

where Q is the matrix of confounding covariates, squared terms for all continuous covariates, and all pairwise interactions. The vectors β_{C0} and β_{C1} are sampled independently from (0,2.5,5) with probabilities (.6,.3,.1). For the s^{th} simulation ω^s_C was chosen so that $\text{Avg}_{i:z=1}(y_i(1) - y_i(0)) = 4$ in the (overlap) setting where we estimate the effect of the treatment on the treated and $\text{Avg}_{i:z=0}(y_i(1) - y_i(0)) = 4$ in the (non-overlap) setting where we estimate the effect of the treatment on the controls. R^2 values from the linear regression of outcomes on the X values is about .77 on average.

5.2 Methods compared

We compare estimates using BART as described in Section 3 with estimates from linear regression, propensity score matching and a propensity-score-based weighting estimator. Linear regression estimates were obtained simply by regressing the outcome on a treatment assignment indicator and the confounding covariates.

Propensity scores were estimated using a logistic regression on all covariates. This specification yielded balance that appeared adequate given the current standards of applied researchers (details available upon request from the authors). Most importantly, however, the goal in these simulations is to make comparisons with implementations of these methods that are as simply specified as BART. This conforms with our goal of helping the researcher to avoid arbitrary judgement calls. Nevertheless, we discuss the implications of alternative propensity score specifications at the end of this section.

Matching estimates relied on one-to-one matching with replacement (matching controls to the intact treatment group for the effect of the treatment on the treated and matching treated to the intact control group for the effect of the treatment on controls). Then we ran a weighted regression of outcomes on treatment assignment and all confounders with weights equal to the number of times each child was represented in the matched sample (1 for all treated, 0 for unmatched controls, and number of times chosen for matched controls). "Huber" standard errors (Huber 1967; Hill and Reiter 2006) were calculated.

The weighted estimator is a straightforward extension (Sato and Matsuyama 2003; Imbens 2004; Kurth *et al.* 2006) of inverse-probability-of-treatment-weighted (IPTW) estimators (Rosenbaum 1987; Robins 1999) that instead estimates the average effect of the treatment on the treated (not the average treatment effect). This estimator is simply a weighted regression estimator with weights equal to 1 for those in the treatment group and with weights equal to $\hat{e}(x)/(1 - \hat{e}(x))$ for those in the control group. As with the matching estimator, robust (Huber) standard errors are calculated. We will refer to this estimator simply as the propensity-weighted estimator.

For each setting 1000 simulations were run on a Dell Precision Workstation 670n IntelR XeonT with a 3.00GHz Processor. Each BART run used 1100 draws with the first 100 discarded as burn-in (convergence was determined as described in Section 3.2 based on several sample runs) and each took less than 90 seconds to run.

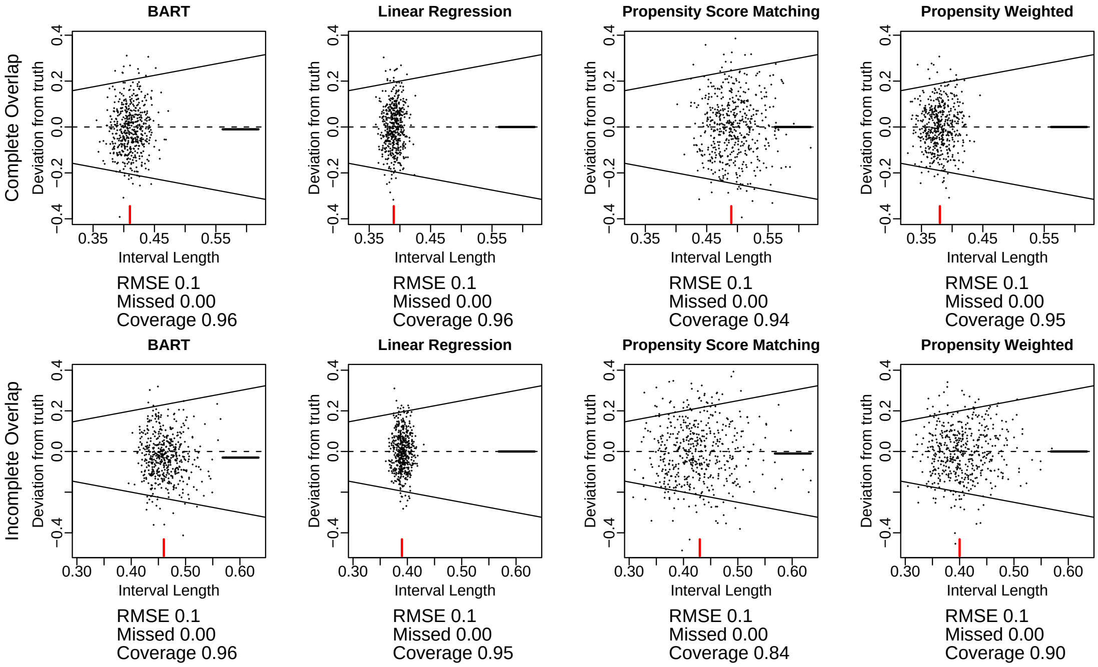

Figure 4: Results from 1000 simulations using linear, parallel Response Surface A. The deviation of the treatment effect estimates from the true effect for 500 simulation draws (randomly sampled from the 1000) are plotted (y-axis) against interval length on the x-axis, with separate plots for each combination of method (columns) and overlap setting: complete (top row) or partial (2nd row). Lines (passing through the origin) with slopes -0.5 and 0.5 display the boundary beyond which a 95% interval corresponding to each point will fail to cover the true population value. The short solid line segment on the far right of each plot displays the bias across our 1000 simulations. The short solid line segment at the bottom displays the average interval length. Summary statistics calculated across all 1000 simulations are: coverage rates for 95% intervals, % of such intervals that would exclude 0 ("Missed"), and the root mean squared error (RMSE).

5.3 Results

Figure 4 displays results from the 1000 simulation runs corresponding to Response Surface A, which is linear and parallel. The points represent a random subset of 500 runs (we limited to 500 runs so that features of the plot wasn't obscured by so many points) where the deviation of the treatment effect point estimates (posterior means) from the true population value (4) is plotted against the 95% interval length. Separate plots are provided for each combination of method (columns) and overlap setting: complete (top row) or partial (2nd row). The dashed horizontal line represents no deviation from the effect for the treated population. Lines that pass through the origin are plotted with slopes .5 and -.5 to discriminate between points with 95% intervals that cover the true value and those that do not.

We formed 95% posterior intervals for BART as the posterior mean plus or minus 1.96 times the posterior standard deviation. An alternative would be to use draws from the BART posterior distribution to form an empirical interval. The two strategies yielded extremely similar intervals for these simulations.

Coverage of corresponding 95% intervals is displayed at the bottom along with the percentage of times ("Missed") that the method would not detect a significant treatment effect (Type II error probability) at a 5% significance level as well as root mean squared error ("RMSE"). These summaries are calculated across all 1000 simulations.

All four methods estimate the treatment effect with virtually no bias and all have good coverage properties in the overlap scenario though the propensity-score methods suffer a bit in the non-overlap scenario. All have similar uncertainty (e.g. average length of 95% intervals is .39 for regression and .41 for BART) in the overlap scenario. The BART uncertainty estimates (see short lines at the bottom indicating average interval length) naturally increase when there is less overlap. Propensity score matched estimates have a bit more variability overall. All methods detect the significant treatment effect in all cases. These differences are so small relative to the size of the treatment effect, however, that they are of negligible *practical* importance. Results for the non-BART methods are not surprising given they all use linear regression which is the correct model. What is surprising is that BART is still amazingly competitive without "knowing" the true model.

Figure 5 displays the results for each method from the 1000 simulations corresponding to Response Surface B (treatment effect is 4). The BART estimates are centered on the true treatment effect while the linear regression and both propensity score estimates are a bit biased on average and display greater variation in both settings. All methods achieve nominal or near-nominal coverage but this time the BART interval lengths are far smaller on average than the regression and propensity-score-matching interval length (less than half the size on average of the former and one third the length of the latter). Again, BART uncertainty increases as we move to the setting with less overlap. The non-BART estimators fail to detect the significant treatment effect from 1-3% of the time and the linear regression fails to detect about 6% of the time in the partial ovelap setting. In general linear regression falls apart when there is not complete overlap and the coverage rates drop to 46%.

The results from Response Surface C (treatment effect equal to 4) displayed in Figure 6 are striking. The point estimates from BART are centered on and clustered fairly tightly around the true treatment effect. The linear regression and propensity score matching estimates on the other hand are more unstable with wider variation in estimates. All methods have decent coverage when there is complete coverage though

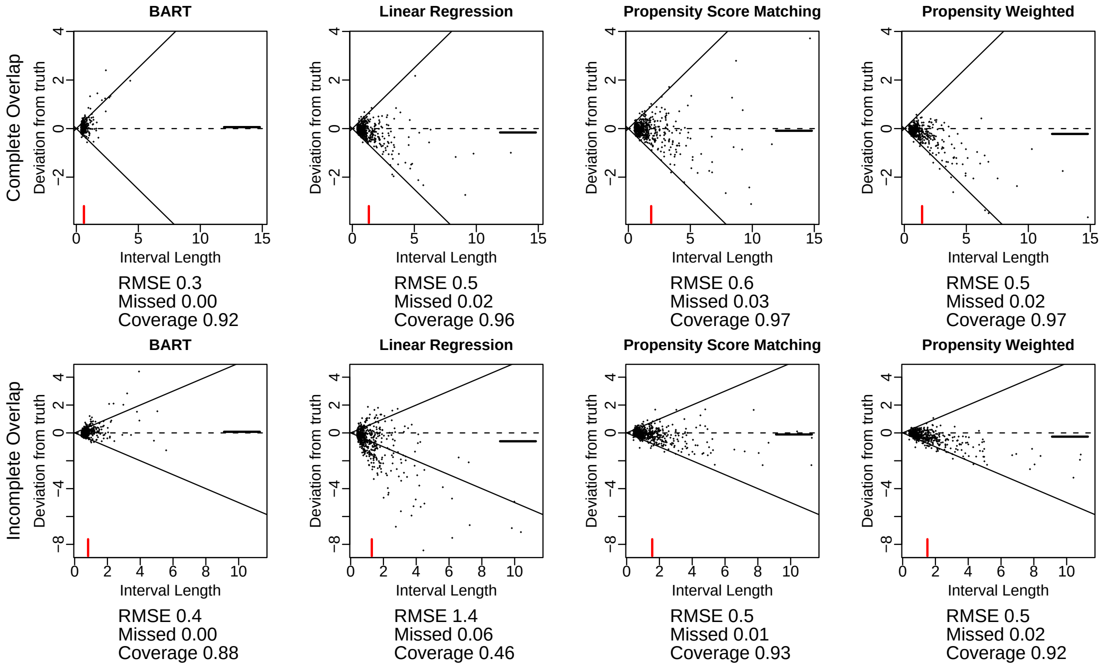

Figure 5: Results from a set of 1000 simulations using non-linear, not parallel Response Surface B. See further details in the caption to Figure 4. In the overlap plot with propensity score matching results, however, two observations with substantially higher interval lengths have been deleted in order to maintain plot scale in a range where differences between methods can be observed in greater detail.

BART achieves this with much less uncertainty (interval lengths smaller than the other approaches by about the same ratios as the last simulation). Coverage drops for all methods when there is not complete overlap but most noticeably for regression (54%). Regression, propensity score matching, and propensity-weighted approaches fail to detect the significant treatment effect a substantial proportion of the time (36%, 56%, and 42%, respectively) in the overlap setting whereas BART detects it in every case tested. When there is not complete overlap BART misses about 11% of the time but the others all miss at least 40% of the time.

Following in the pattern of the other simulations, but more noticeably in this set-

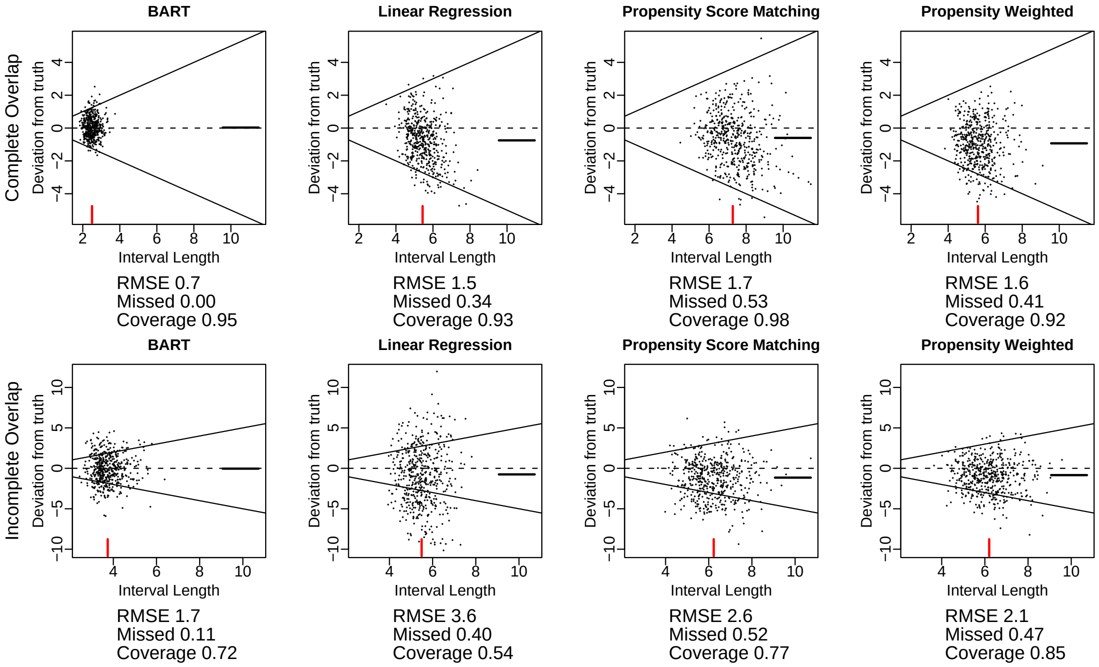

Figure 6: Results from a set of 1000 simulations using non-linear, not parallel Response Surface C. See further details in the caption to Figure 4.

ting, BART uncertainty increases as overlap decreases. Regression uncertainty stays nearly the same, propensity matched uncertainty *decreases*, and propensity-weighted uncertainty increases, but not to the same degree as BART when we do not have complete overlap.

5.4 Simulations with increased noise

All of the simulations above had a standard deviation of 1 for each of the response surfaces. To explore the effect of sampling variability on these results we re-ran all of the simulations above both with standard deviation of 5 and then with standard deviation of 10. One important consequence of the increased noise is that it becomes much harder to detect the non-linearities. For instance, using standard residual plots it is virtually impossible to detect the non-linearities for Response Surface B when the standard deviation equals 5 and for Response Surface C when the standard deviation equals 10. The results of these simulations are available from the authors and tell a similar story to the results already presented.

5.5 Sensitivity to specification of propensity score model

No strong theory exists regarding how best to model the propensity score (Sekhon 2007). In the matching literature the general advice is to choose a model that yields matched groups that are most "balanced". Unfortunately the definition of adequate balance, including what moments should be balanced and what criteria should be used to discriminate between acceptable and unacceptable balance, is still not widely understood. Yet treatment effect estimates can be quite sensitive to these choices.

To investigate sensitivity of our propensity score model specification we examined the results from two alterative strategies. First, we ran simulations in which we used standard balance diagnostics (simple standardized differences in means) to choose the propensity score model. These yielded slightly worse results than our simple model choice. Upon the advice of a referee we also re-ran our simulations using a generalized additive model (GAM) to fit the propensity score model. This had basically no impact on the results from Response Surface A and C for either overlap setting. For response surface B in the complete overlap setting matching results were quite similar but the weighting estimator yielded better results than in the original simulations, though performance still did not match BART. In the setting without complete overlap both matching and weighting estimators performed worse than in the original simulations.

It is also possible that a more sophisticated matching algorithm or model for the propensity scores might produce more competitive results (see, for instance, Diamond and Sekhon 2006; Hansen 2006). We do not pursue these approaches here.

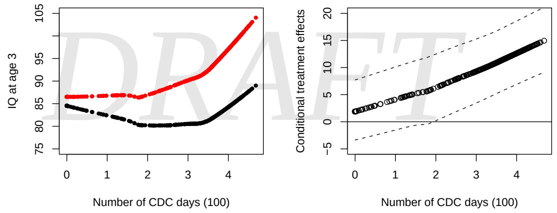

Figure 7: Left panel displays plot of BART-predicted 3 year IQ test scores against CDC participation (in hundreds of days) for children in the treatment group (upper line). The lower line shows predicted scores for the same children if they had not attended any CDC days. Lines were smoothed using lowess. The right panel displays a smoothed function of the treatment effect estimates at each level of CDC participation (conditional on having that level of participation in the treatment group). Dashed lines represent 95% uncertainty bounds.

6 More complicated analyses

BART can be easily extended to accommodate more complicated inferential challenges such as continuous treatment variables and missing outcome data. Accommodating continuous treatment variables is never completely straightforward in the absence of an experiment that randomizes treatment levels to units, because it always requires stronger assumptions to justify causal conclusions.

Consider, for example, the IHDP experiment discussed in the previous section. The primary treatment in this experiment for those randomized to the intervention group was access to intensive, high-quality child care in Child Development Centers (CDCs) in the second and third years of the child's life. The treatment group ($I = 1$) participated in this care with varying intensity, however; number of CDC days, Z, range from 0 to about 500. The control group ($I = 0$) did not have access to this care at all (everyone had number of CDC days equal to 0).

We fit a BART model to all the data and then used it to make two sets of predictions of age 3 IQ test scores which are displayed in Figure 7. The upper line reflects (smoothed) predictions for the treatment group at their observed treatment group status and level of participation ($Y(Z = a) \mid I = 1, Z = a$). The lower line reflects (smoothed) predictions for these same children had they been in the control group with no CDC days ($Y(0) \mid I = 1, Z = a$). Thus number of CDC days on the x-axis refers to the number of CDC days children *would have participated in* if they had been assigned to the treatment (which all these children had been). The right panel plots the (smoothed) differences between these lines and associated uncertainty intervals.

One set of identifying assumptions for a finite set of treatments Z ($Z = a : a \subset [0, 1, \ldots, A]$) would be

$$Y(Z = 0), Y(Z = 1), \ldots, Y(Z = A) \perp Z \mid X$$

$$0 < P(Z = a \mid X) < 1, \text{ for all } a = 0, \ldots, A.$$

Under this assumption BART can predict counterfactuals for any desired unit (preferably within the observed support of the predictor space) at any level of the treatment variable. This would allow for inferences regarding the effect of moving from any dosage level to any other dosage level (for example $E[Y(Z = a) - Y(Z = a')]$ for all combinations of a and a') for individual units or groups of units.

An example of a weaker set of assumptions is

$$Y(Z = 0) \perp Z \mid X, \text{ and}$$

$$0 < P(Z = 0 \mid X, Z = a) < 1, \text{ for all } a = 0, \ldots, A.$$

Under this set of assumptions the difference between the response surfaces at each level of a CDC days in the left panel represents the effect of moving from that dosage level to a dosage level of 0 days *for those who would have participated in that number* *of CDC days had they been assigned to the treatment.* A smoothed function of these treatment effects is displayed, with corresponding uncertainty bounds, in the right panel of Figure 7. This suggests that only children who selected into participation in more than 200 CDC days demonstrated significant effects of this intervention.

The counterfactual line is interesting in and of itself. It suggests that the children who have the highest potential scores in the absence of the treatment are most likely to have received the either least treatment (perhaps because their parents don't think they need it) or the most treatment (perhaps because their parents think they are strong or healthy enough to attend regularly).

A linear regression fit to these data cannot pick up these non-linearities using quadratic terms – these terms yield coefficients that are not statistically significant and the model fit is poor.

Missing data. Missing outcome data has a straightforward solution under the assumption of a strongly ignorable missing data mechanism. Simply fit the BART model to the complete case sample but make predictions for the full sample.

7 Discussion

We have explored the capacity of a new Bayesian nonparametric modeling algorithm, BART, to estimate causal effects. The potential advantages of BART as compared to methods that are reasonably similar in terms of difficulty in implementation and simplicity of specification have been revealed through the examples and simulations in this paper. Across several different response surface specifications and two different levels of overlap BART performs very well compared to linear regression, propensity-score matching and a propensity-weighted estimator. In linear specifications, BART's performance in terms of root mean squared error, levels of uncertainty, and 95% interval coverage is quite similar to the other estimators. In non-linear, non-additive specifications BART outperfomed the competitors in nearly every combination of setting and performance criterion.

There are also advantages of BART compared to alternative nonparametric or semiparametric methods that may have the capacity to flexibly model the assignment mechanism and the response surface. First BART can handle a large number of both continous and discrete predictors. Moreover BART overcomes a standard barrier to widespread implementation of new methodology because it requires far less researcher involvement, technical sophistication, and investment of time. The method is accessible to applied researchers who may not have a strong mathematical background and won't require days or weeks of programming to implement (particularly important given that it is difficult to know when a more sophisticated method will actually make a difference in practice).

Finally, unlike almost all methods that require strong theories about the relationships between variables in the model, BART is not constrained by prior theories. Substantive theories are extremely important as a benchmark against which to test the phenomemon observable in one's data. However, if we build models solely based on past theories is it much more difficult to advance science by uncovering unexpected relationships between model inputs.

One of the most compelling aspects of BART's performance is that its uncertainty estimates naturally increase when there is less information (for instance when there is lack of complete overlap and hence limited empirical counterfactuals). At the same time, BART's accurate treatment effect estimation relies on shrinkage which places a limit on the amount that this uncertainty will increases and hence can hurt coverage in the case with limited overlap. We plan to investigate better ways to address this trade-off between efficiency and coverage rates.

BART still could benefit from further refinement in other areas as well. For instance, as currently specified BART would not be expected to reliably uncover a response surface with high levels of interaction. Finally we have tested in only a few different scenarios, more work needs to be done to determine whether BART will work as effectively as a causal inference strategy in a still broader range of settings.

References

Abadie, A. and Imbens, G. W. (2006), "Large Sample Properties of Matching Estimators for Average Treatment Effects," *Econometrica* 74, 1, 253–267.

Ashenfelter, O. (1978), "Estimating the effects of training programs on earnings," *Review of Economics and Statistics* 60, 47–57.

Ashenfelter, O. and Card, D. (1985), "Using the longitudinal structure of earnings to estimate the effect of training programs," *Review of Economics and Statistics* 67, 648–660.

Bingenheimer, J., Brennan, R., and Earls, F. (2005), "Firearm violence exposure and serious violent behavior," *Science* 308, 1323–1326.

Breiman, L. (2001), "Random Forests," *Machine Learning* 45, 1, 5–32.

Brooks-Gunn, J., Liaw, F., and Klebanov, P. (1991), "Effects of early intervention on cognitive function of low birth weight preterm infants," *Journal of Pediatrics* 120, 350–359.

Carpenter, J., Kenward, M., and Vansteelandt, S. (2005), "A comparison of multiple imputation and doubly robust estimation for analyses with missing data," *Journal of the Royal Statistical Society, Series A* 169, 1–14.

Chipman, H., George, E., and McCulloch, R. (2006), "BART: Bayesian Additive Regression Trees," Tech. rep., University of Chicago.

Chipman, H., George, E., and McCulloch, R. (2007), "Bayesian Ensemble Learning," in *Advances in Neural Information Processing Systems 19*, eds. B. Schölkopf, J. Platt, and T. Hoffman, Cambridge, MA: MIT Press.

Dehejia, R. H. and Wahba, S. (1999), "Causal Effects in Nonexperimental Studies: Reevaluating the Evaluation of Training Programs," *JASA* 94, 1053–1062.

Diamond, A. and Sekhon, J. (2006), "Genetic Matching for Estimating Causal Effects: A General Multivariate Matching Method for Achieving Balance in Observational Studies," Tech. rep., U.C. Berkeley.

Du, J. (1998), "Valid inferences after propensity score subclassification using maximum number of subclasses as building blocks," Ph.D. thesis, Harvard University.

Gu, X. S. and Rosenbaum, P. R. (1993), "Comparison of Multivariate Matching Methods: Structures, Distances, and Algorithms," *Journal of Computational and Graphical Statistics* 2, 405–420.

Hahn, J. (1998), "On the Role of the Propensity Score in Efficient Semiparametric Estimation of Average Treatment Effects," *Econometrica* 66, 315–322.

Hansen, B. (2006), “Bias reduction in observational studies via prognosis scores,” Tech. rep., University of Michigan.

Heckman, J. J., Ichimura, H., and Todd, P. (1997), “Matching as an Econometric Evaluation Estimator: Evidence from a Job Training Programme,” *Review of Economic Studies* 64, 605–654.

Hill, J. and Reiter, J. (2006), “Interval estimation for treatment effects using propensity score matching,” *Statistics in Medicine* 25, 13, 2230–2256.

Hirano, K. and Imbens, G. W. (2001), “Estimation of Causal Effects Using Propensity Score Weighting: An Application of Data on Right Ear Catheterization,” *Health Services and Outcomes Research Methodology* 1, 259–278.

Hirano, K., Imbens, G. W., and Ridder, G. (2003), “Efficient estimation of average treatment effects using the estimated propensity score,” *Econometrica* 71, 1161–89.

Huber, P. (1967), “The behavior of maximum likelihood estimates under non-standard conditions,” in *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, pp. 221–233.

Imbens, G. (2004), “Nonparametric Estimation of Average Treatment Effects Under Exogeneity: A Review,” *The Review of Economics and Statistics* 86, 1, 4–29.

Infant Health and Development Program (1990), “Enhancing the outcomes of low-birth-weight, premature infants,” *Journal of the American Medical Association* 22, 3035–3042.

Kurth, T., Walker, A. M., Glynn, R. J., Chan, K. A., Gaziano, J. M., Berger, K., and Robins, J. M. (2006), “Results of multivariable logistic regression, propensity matching, propensity adjustment, and propensity-based weighting under conditions of non-uniform effect," *American Journal of Epidemiology* 163, 3, 262–270.

LaLonde, R. (1986), "Evaluating the Econometric Evaluations of Training Programs," *American Economic Review* 76, 604–620.

Leamer, E. (1983), "Let's take the con out of econometrics," *American Economic Review* 73, 31–43.

Robins, J. M. (1999), "Association, causation, and marginal structural models," *Synthese* 121, 151–179.

Robins, J. M., Hernan, M. A., and Brumback, B. (2000), "Marginal structural models and causal inference in epidemiology," *Epidemiology* 11, 550–60.

Robins, J. M. and Ritov, Y. (1997), "Towards a curse of dimensionality appropriate (CODA) asymptotic theory for semi-parametric models," *Statistics in Medicine* 16, 285–319.

Robins, J. M. and Rotnitzky, A. (1995), "Semiparametric Efficiency in Multivariate Regression Models with Missing Data," *JASA* 90, 122–129.

Robins, J. M., Rotnitzky, A., and Zhao, L. P. (1994), "Estimation of regression coefficients when some regressors are not always observed," *JASA* 89, 846–866.

Robins, J. M., Rotnitzky, A., and Zhao, L. P. (1995), "Analysis of Semiparametric Regression Models for Repeated Outcomes in the Presence of Missing Data," *JASA* 90, 106–121.

Rosenbaum, P. R. (1987), "Model-Based Direct Adjustment," *Journal of the American Statistical Association* 82, 387–394.

Rosenbaum, P. R. and Rubin, D. B. (1983), "The central role of the propensity score in observational studies for causal effects," *Biometrika* 70, 1, 41–55.

Rosenbaum, P. R. and Rubin, D. B. (1985), "Constructing a control group using multivariate matched sampling methods that incorporate the propensity score," *The American Statistician* 39, 33–38.

Rotnitzky, A. and Robins, J. M. (1997), "Analysis of semi-parametric regression models with nonignorable nonresponse," *Statistics in Medicine* 16, 81–102.

Rotnitzky, A., Robins, J. M., and Scharfstein, D. (1998), "Semiparametric regression for repeated outcomes with nonignorable nonresponse," *JASA* 93, 1321–1339.

Rubin, D. B. (1973), "The Use of Matched Sampling and Regression Adjustment to Remove Bias in Observational Studies," *Biometrics* 29, 185–203.

Rubin, D. B. (1978), "Bayesian Inference for Causal Effects: The role of randomization," *The Annals of Statistics* 6, 34–58.

Rubin, D. B. (1979), "Using Multivariate Matched Sampling and Regression Adjustment to Control Bias in Observational Studies," *Journal of the American Statistical Association* 74, 318–328.

Rubin, D. B. and Thomas, N. (2000), "Combining Propensity Score Matching with Additional Adjustments for Prognostic Covariates," *JASA* 95, 573–585.

Sato, T. and Matsuyama, Y. (2003), "Marginal structural models as a tool for standardization," *Epidemiology* 14, 680–6.

Scharfstein, D., Rotnitzky, A., and Robins, J. (1999), "Adjusting for nonignorable drop-out using semi-parametric nonresponse models (with discussion)," *Journal of the American Statistical Association* 94, 1096–1146.

Scott, D. and Bauer, C. (1989), "A neonatal health index for preterm infants," *Pediatric Research* 25, 263A.

Sekhon, J. S. (2007), "Alternative Balance Metrics for Bias Reduction in Matching Methods for Causal Inference," Tech. rep., U.C. Berkeley.

Tu, W. and Zhou, X.-H. (2003), "A bootstrap confidence interval procedure for treatment using propensity score subclassification," Tech. rep., University of Washington Biostatistics Working Paper Series. Working Paper 200.