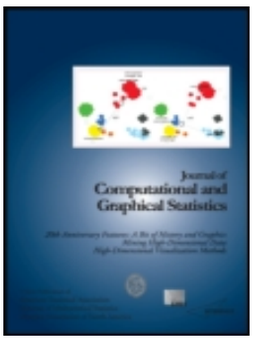

Journal of Computational and Graphical Statistics

ISSN: 1061-8600 (Print) 1537-2715 (Online) Journal homepage: http://www.tandfonline.com/loi/ucgs20

Bayesian Nonparametric Modeling for Causal Inference

Jennifer L. Hill

To cite this article: Jennifer L. Hill (2011) Bayesian Nonparametric Modeling for Causal Inference, Journal of Computational and Graphical Statistics, 20:1, 217-240, DOI: 10.1198/jcgs.2010.08162

To link to this article: http://dx.doi.org/10.1198/jcgs.2010.08162

Bayesian Nonparametric Modeling for Causal Inference

Jennifer L. HILL

Researchers have long struggled to identify causal effects in nonexperimental 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. This article proposes 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 guesswork in model fitting, handles a large number of predictors, yields coherent uncertainty intervals, and fluidly handles continuous treatment variables and missing data for the outcome variable. BART also naturally identifies heterogeneous treatment effects. BART produces more accurate estimates of average treatment effects compared to propensity score matching, propensity-weighted estimators, and regression adjustment in the nonlinear simulation situations examined. Further, it is highly competitive in linear settings with the "correct" model, linear regression. Supplemental materials including code and data to replicate simulations and examples from the article as well as methods for population inference are available online.

Key Words: Bayesian; Causal inference; Nonparametrics.

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 pretreatment or "confounding" covariates. The most appropriate modeling choices for estimation of treatment effects under ignorability are still debated (see Imbens 2004, for a comprehensive review).

Many causal methods for observational data (Rubin 1973, 1979), including many recent methods (for instance, Heckman, Ichimura, and Todd 1997; Robins and Ritov 1997; Robins, Hernan, and Brumback 2000; Rubin and Thomas 2000; Hirano, Imbens, and Ridder 2003; Kurth et al. 2006), involve fitting both a model for the treatment assignment

Jennifer L. Hill is Associate Professor of Applied Statistics, Department of Humanitites and Social Sciences, New York University Steinhardt, 246 Greene St., 3rd Floor, New York, NY 10003 (E-mail: *jennifer.hill@nyu.edu*).

Journal of Computational and Graphical Statistics, Volume 20, Number 1, Pages 217–240
DOI: 10.1198/jcgs.2010.08162

mechanism and a model for the outcome (or potential outcomes) conditional on the treatment and confounding covariates. The latter model 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, this article proposes that a robust, yet simpler, modeling approach is now available for accurately estimating causal effects in this setting. This approach focuses 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 2007, 2010).

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. Yet BART is able to detect interactions and nonlinearities in the response surface, which (among other advantages) allows it to more readily identify heterogeneous treatment effects. 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 consensus regarding appropriate interval estimation (Imbens 2004; Hill and Reiter 2006) (though Abadie and Imbens 2006 recently proposed an estimator that may be a viable solution under certain conditions). 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 nonlinear 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 simulations is almost indistinguishable from linear regression, the "correct" model for that setting. Thus BART is a simple method that appears to have the potential to be both robust and accurate in the estimation of causal effects.

Section 2 of this article discusses the estimation problem and traditional methods. Section 3 describes BART and how to use it to estimate causal effects. Section 4 presents simulations using treatment and covariate data from a real study. Section 5 explores the robustness of the simulation results to some changes in assumptions and estimation strategies. Section 6 illustrates use of BART in the context of a real data example examining dosage effects; in that section BART's performance is also compared with a wider range of propensity score and matching techniques. Section 7 concludes.

2. CAUSAL INFERENCE

Consider the causal effect of binary treatment Z, with $Z = 1$ indicating assignment to treatment and $Z = 0$ indicating assignment to control (a binary treatment variable is used for ease of exposition; however, the ideas here can be extended to multivalued or continuous treatment variables, an example of which is provided in Section 6). Following convention in the statistical causal inference literature (e.g., Rubin 1978), the causal effect for individual i is defined as a comparison of potential outcomes, $Y_i(0)$ and $Y_i(1)$; these

are the outcomes that would be observed under $Z = 0$ and $Z = 1$, respectively. This article focuses in particular on the difference between these outcomes, $Y_i(1) - Y_i(0)$.

We cannot observe both $Y_i(1)$ and $Y_i(0)$ for any individual, but only the observed outcome, $Y_i = Y_i(1)Z_i + Y_i(0)(1 - Z_i)$. For this reason, researchers generally focus on estimating average treatment effects defined either over the sample or the population. Common sample estimands are the sample average treatment effect (SATE), $\sum_{i=1}^{n}[Y_i(1) - Y_i(0)]$, and the sample average effect of the treatment on the treated (SATT), $\sum_{i:Z_i=1}[Y_i(1) - Y_i(0)]$. Common population estimands are the population average treatment effect (PATE), $E[Y(1) - Y(0)]$, and the population average effect of the treatment on the treated (PATT), $E[Y(1) - Y(0) \mid Z = 1]$. Another set of average estimands (Abadie and Imbens 2002; Imbens 2004) that retain some of the properties of the previous two are the conditional average treatment effect (CATE), $\sum_{i=1}^{n} E(Y_i(1) - Y_i(0) \mid X_i)$, and the conditional average treatment effect for the treated (CATT), $\sum_{i:Z_i=1} E(Y_i(1) - Y_i(0) \mid X_i)$. Choice of estimands is discussed below.

In observational studies potential outcomes are typically not independent of treatment assignment. We can identify the average causal effects above under assumptions such as strong ignorability of treatment assignment (though often weaker versions of this assumption are required). Strong ignorability consists of both an unconfoundedness assumption, $Y(0), Y(1) \perp\!\!\!\perp Z \mid X$, where X represents a vector of pretreatment variables or "confounding covariates" and $\perp\!\!\!\perp$ designates conditional independence, and a common support (overlap) assumption, $0 < \Pr(Z = 1 \mid X) < 1$. Estimation of these causal effects then requires evaluation of the response surfaces $E[Y(1) \mid X] = E[Y \mid X, Z = 1]$ and $E[Y(0) \mid X] = E[Y \mid X, Z = 0]$, where X is potentially high-dimensional.

How can we best estimate the relevant conditional expectations? This estimation may be difficult if $Y(0)$ and $Y(1)$ are 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: $Z \sim$ Bernoulli(0.5),

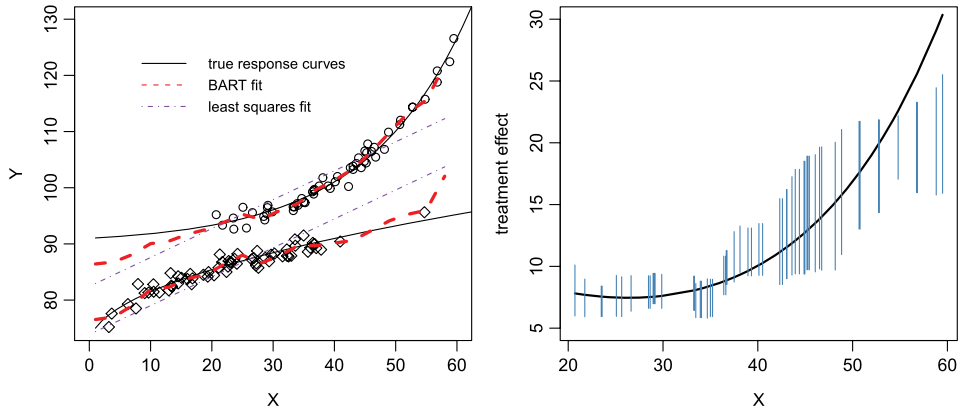

Figure 1. Left panel: simulated data with linear regression and BART fits. Right panel: BART inference for treatment effect on the treated. A color version of this figure is available in the electronic version of this article.

$X \mid Z = 1 \sim \mathrm{N}(40, 10^2)$, $X \mid Z = 0 \sim \mathrm{N}(20, 10^2)$, $Y(0) \mid X \sim \mathrm{N}(72 + 3\sqrt{X}, 1)$, $Y(1) \mid X \sim \mathrm{N}(90 + \exp(0.06X), 1)$. In the left panel, the upper dark solid curve represents $E[Y(1) \mid X]$ and the lower one $E[Y(0) \mid X]$. The circles close to the upper curve are the treated observations and the diamonds close to the lower curve are the untreated. In this simulation PATT is 12.3, CATT is 12.0, and SATT is 11.9.

The parallel dot-dash lines display a linear regression fit to the data that yields an underestimate, 8.6 (standard error 0.7), of all of the treatment on the treated estimands as well as the PATE, which is equal to 10.4. Propensity score strategies (described in Section 4) were also used to estimate the effect of the treatment on the treated for this example. Matching yields an estimate of 11.8 (standard error 0.7) while the inverse-probability-of-treatment weighted estimator performs worse here with an estimate of 8.4 (standard error 0.9). 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 gray dashed line in the left panel of Figure 1 displays the BART fit to the data which is quite close to the true conditional expectation for most of the support. The right panel displays the BART inference for each treated unit (which can be averaged to estimate the effect of the treatment on the treated for this sample). In the right panel, the true treatment effect as it varies with X, $E[Y(1) - Y(0) \mid X]$, 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 do not observe empirical counterfactuals for each data point (e.g., $X > 40$). The marginal intervals nicely cover the true conditional treatment effects except at values of X far from the area of strong overlap. 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, 13.2). As discussed below, this interval best corresponds to inference with respect to CATT; however, here it covers SATT and PATT as well. Computation of these intervals is discussed in Section 3.

The difficulty in estimating the conditional expectations required for causal inference 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 huge number of covariates to justify ignorability and still is critiqued for not including enough, see Bingenheimer, Brennan, and Earls 2005). Even propensity score techniques, which simplify evaluation of the conditional expectations by replacing the vector of covariates that are conditioned on with a single estimated propensity score, can struggle when there is a very large number of covariates. In this setting the propensity score estimation can quickly suffer from problems of separation yielding estimated scores that suggest insufficient common support when in fact this may not actually be the case.

A host of new methods 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 (e.g., Rosenbaum and Rubin 1983; Gu and Rosenbaum 1993; Abadie and Imbens 2006), parametric inverse probability weighting approaches (e.g., Rosenbaum 1987; Robins, Hernan, and Brumback 2000), or semiparametric inverse probability weighting (e.g., Hahn 1998). The motivation

here is that if the treatment assignment mechanism is properly specified, one can avoid modeling the response surface (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 response surface is correctly specified, we do not have to worry about correctly specifying the assignment mechanism. In this spirit, many articles have focused on the potential advantages with respect to robustness (and potentially precision) that may be achieved by modeling both the treatment assignment mechanism *and* the response surface. Rubin and Thomas (2000) and Kurth et al. (2006) presented relatively simple approaches. The weighting estimators in this family have also been extended to address efficiency issues (e.g., Robins and Rotnitzky 1995; Hahn 1998; Scharfstein, Rotnitzky, and Robins 1999; Hirano, Imbens, and Ridder 2003). Moreover, semiparametric estimators of this sort, such as those discussed by Robins and Rotnitzky (1995) and Rotnitzky, Robins, and Scharfstein (1998), were shown by 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" and are discussed further by Carpenter, Kenward, and Vansteelandt (2005) and Kang and Schafer (2007).

This approach is different from most others because it focuses solely on precise estimation of the response surface (see Hahn 1998, for a different approach to modeling the response surface). While some competing parametric methods have been shown to be consistent under appropriate conditions (for instance, Rosenbaum 1987; Robins and Rotnitzky 1995; Heckman, Ichimura, and Todd 1997), this requires correct specification of the parametric model. Nonparametric and semiparametric versions of these methods are more robust but require a higher level of researcher sophistication to understand and implement (e.g., to specify smoothing parameters such as number of terms in a series estimator or bandwidth for a kernel estimator). This article proposes that the benefits of the BART strategy in terms of simplicity, precision, robustness, and lack of required researcher interference outweigh the potential benefit of having an estimator that is strictly consistent under certain sets of conditions.

3. BART AND ESTIMATING CAUSAL EFFECTS

BART is designed to estimate a model for the outcome Y, specified very generally as $Y = f(z, x) + \epsilon$, where z denotes the assigned treatment, and x denotes the observed confounding covariates, and ϵ are iid $N(0, \sigma^2)$. The BART model assumes additive errors; however, its inference for f is very flexible. If ignorability holds conditional on x, that is, $Y(0), Y(1) \perp Z|X = x$, then I posit $E[Y(0) \mid X = x] = E[Y \mid Z = 0, X = x] = f(0, x)$ and $E[Y(1) \mid X = x] = E[Y \mid Z = 1, X = x] = f(1, x)$. In principle, any method that flexibly estimates f could be used to model these conditional expectations. BART has some potentially important advantages over alternative methods such as random forests, boosting, and neural nets (see sec. 8.7 and chaps. 10, 11 of Hastie, Tibshirani, and Friedman 2003), for instance with regard to choosing tuning parameters and calculating uncertainty,

as described in greater detail below. BART is developed in the work of Chipman, George, and McCulloch (2007, 2010) (henceforth CGM07 and CGM10, and CGM generally).

Section 3.1 provides an overview of BART (details are in CGM07). Section 3.2 discusses its advantages for causal inference. Section 3.3 details how BART can be used to estimate a variety of average causal effects.

3.1 OVERVIEW OF BART

BART builds upon tree models. Notation for a single tree model is established first. Let T denote a binary tree. The tree T consists of the tree structure and all the decision rules leading down to a bottom node. The left panel of Figure 2 depicts a single tree model fit to the data in Figure 1. All of the interior nodes of T have decision rules which send a (z, x) pair either left or right. The jth bottom node has a parameter μ_j associated with it that represents the mean response of the subgroup of observations that fall in that node. Let $M = \{\mu_1, \mu_2, \ldots, \mu_b\}$ where b is the number of bottom nodes. The decision rules in the figure are the criteria for sending an observation with (z, x) left. Note that z is encoded as a 0–1 indicator variable; thus the first decision rule sends all the untreated left. In Figure 2, M is the set of six means 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. For example, 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 fit from a single tree; that is, they plot function $g(z, x; T, M)$ against x, with each segment corresponding to a bottom node. The dashed curve is the BART (sum-of-trees) fit. It is so close to the true response surface that it is difficult to see, except when there is no support in x.

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

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

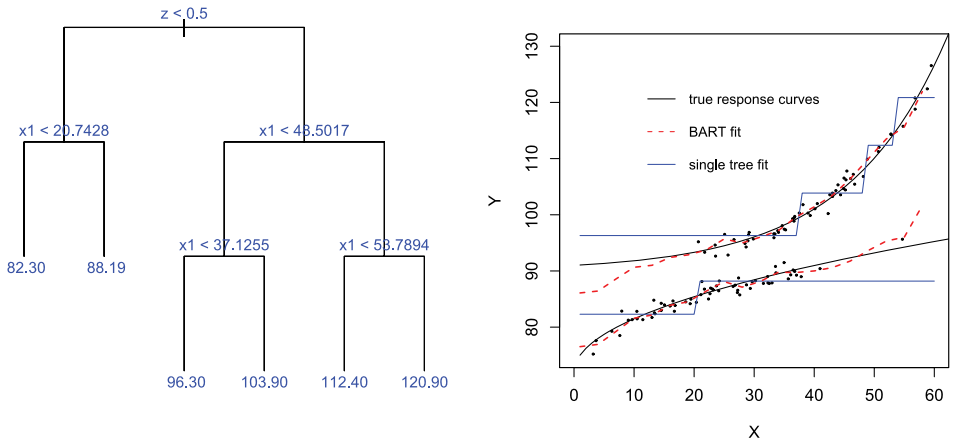

Figure 2. Left panel: the binary tree fit to the data from Figure 1. Right panel: single-tree fits (solid lines) and BART fits (dashed lines). A color version of this figure is available in the electronic version of this article.

where each (T_j, M_j) denotes a single subtree model and $\epsilon \sim \mathrm{N}(0, \sigma^2)$. Then $Y = f(z, x) + \epsilon$ with $f(z, x) = \sum g(z, x; T_j, M_j)$.

The motivation for the sum-of-trees model is not immediately obvious. To provide intuition for this model consider taking the fit from the first weak-learning (small) tree, $g(z, x; T_1, M_1)$, and subtracting it off from the observed response y to form residuals. Then imagine fitting the next tree to these residuals. This process would be performed a total of m times. In the spirit of boosting, CGM allow the number of subtree models, m, to be large. However, as with boosting, we want to avoid overfitting. This is achieved through a regularization prior which holds back the fit of each (T_j, M_j) tree allowing each to contribute only a small part to the overall fit.

In the CGM approach, the (T_j, M_j) 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 (MCMC). At each iteration of the MCMC algorithm all of the (T_j, M_j) and σ are redrawn. Note that only the parameter σ is identified. As the MCMC runs the fit is swapped amongst the trees as the stochastic search seeks out a good f, with each tree attempting to capture fit not realized by the others. At each iteration, each tree may grow in size or shrink as its fit is swapped to other trees. The contribution of a particular tree is not identified. At a minimum, the tree parameters have a label-switching problem in that any (T, M) pair may be swapped with another without changing the f. The lack of identification actually leads to a MCMC algorithm which is typically remarkably stable and rapidly finds good fit (CGM07).

Given the flexibility of the sum-of-trees model and the lack of identification, the prior is essential (for more details see CGM07, CGM10). It both regularizes the overall fit so we do not overfit and limits the contribution of each (T_j, M_j). The prior has three components: (1) a prior preference for trees T_j with only a few bottom nodes, (2) a prior which shrinks each M_j toward zero (the response is centered so that zero is the typical value), (3) a prior which suggests σ is smaller than that given by least squares, but not dramatically so. CGM provide a default prior setting. While the prior pushes the model in a certain direction, no hard limits are fit. If the data strongly suggest that larger trees are needed, the prior may be overcome.

3.2 ADVANTAGES OF BART FOR CAUSAL INFERENCE

Chipman, George, and McCulloch (2007) showed that BART's performance using the default prior is highly competitive in terms of prediction with other methods that rely on cross-validation to tune algorithm parameters. As a practical matter, not having to use cross-validation makes the method far easier to use; it is much more automatic. Moreover, given the use of the default prior, Bayesian posterior measures of uncertainty are readily available. Quantifying the uncertainty associated with the fit of data-mining tools that rely on cross-validation, on the other hand, is relatively difficult. Of course, Bayesian posterior intervals, based on a proper prior, may not have the frequentist coverage some users may desire. However, the simulation results, all based on use of the default prior, show that

the frequentist properties of this BART-based strategy are remarkably competitive with common approaches in the causal inference literature.

The sum-of-trees specification is adept at capturing both nonlinearities and interactions without the researcher having to explicitly add interaction terms or transformations of x or specify a limit on the level of interaction (as is traditionally required when boosting regression trees). To understand how BART captures nonlinearities, consider the case where each T_j uses only one component of (z, x) in all of the interior decision nodes. In this case, the overall fit is like a generalized additive model. By combining the contributions of many trees (all using only a common variable) any nonlinear $f(z_i, x_i)$ can be approximated. Interactions are captured whenever a T_j involves more than one component of (z, x) in its decision rules. In each iteration of the CGM MCMC algorithm, each T_j can grow or become smaller, allowing BART to naturally infer the level of interaction. This allows for greater flexibility in identifying heterogeneous treatment effects across different types of people.

The CGM MCMC appears to be quite stable; the method run twice (with different seeds) yields virtually the same results (CGM07, CGM10). Convergence of the chain can be assessed by plotting draws of the one interpretable parameter, σ. Initially, the draws of σ tend to 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 quantification 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 the uncertainty. BART combines the advantage (good prediction) of boosting weak-learning models with that of working within a formal Bayesian modeling framework to quantify uncertainty.

BART can handle very large numbers of predictors. If a variable is not useful it simply does not get used (or not often). Therefore researchers can include greater numbers of potential covariates even than propensity score methods which can suffer from problems of separation when estimating propensity scores using large numbers of covariates. In extreme cases propensity scores can even end up estimated at 0 for all controls and at 1 for all treated; this makes them useless for finding similar observations. The ability to include many potential confounding covariates as predictors can be quite helpful when trying to satisfy ignorability.

An overall benefit is the lack of "tinkering" required with BART. Researchers tend to condition on variables and use functional forms that fit their existing theory (for a discussion, see 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 do not 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.

3.3 Estimating Causal Effects

BART can be used to estimate average causal effects (in theory, individual-level causal effects could be estimated as well but these would likely be far less robust). Given the

nature of the algorithm, which conditions on the X values in the sample, a natural set of estimands are the conditional average treatment effect (CATE)

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

and the conditional average treatment effect for the treated (CATT)

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

Some additional notation is warranted. Define $c(x, f) \equiv f(1, x) - f(0, x)$ as the treatment effect at $X = x$. Interest is in the joint posterior distribution of $C(f) = (c(x_1, f), c(x_2, f), \ldots, c(x_K, f))$ over some subset (possibly proper) of the sample, indexed by i. Recall that each iteration of the BART Markov chain generates a new draw of f from the posterior distribution. Let f^l denote the lth draw of f. Thus for every draw, we obtain a draw from the joint posterior of $C(f)$: $C^l = C(f^l)$. 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.

Finally, let $\{x_i\}_1^K$ denote a set of x representing the empirical distribution of x over which we wish to learn the average treatment effect. For instance, if we want inference for the CATE, we simply compute the average of the vector C^l at each l, $\bar{C}^l = \frac{1}{K}\sum_i^K c(x_i, f^l)$. This gives draws $\bar{C}^l$ from the posterior distribution of the conditional average treatment effect that reflect the full joint dependent distribution of C^l. If we were interested in CATT we would focus only on $\{i : z_i = 1\}$.

The vertical intervals in the right panel of Figure 1 were obtained by computing the 2.5% and 97.5% quantiles of the set of draws $c(x_i, f^l)$ for each fixed i. These intervals display the marginal posterior distributions of each $c(x_i, f)$. To obtain inference for the conditional average treatment effect, we average $c(x_i, f^l)$ for each fixed l.

3.3.1 Generalizability

Researchers sometimes focus on population estimands such as the population average treatment effect (PATE), $E(Y(1) - Y(0))$, and the population average treatment effect on the treated (PATT), $E(Y(1) - Y(0) \mid Z = 1)$. However, observational samples used to draw causal inferences are often not random samples from the purported population of interest. Moreover, when heterogeneous treatment effects exist, it is unclear why an average of these across a population will be particularly useful. Therefore, many estimators in the causal world, particularly those that eschew parametric assumptions, are specifically geared toward sample inference (Rosenbaum 2002) and only generalize to population inference under special circumstances (for instance, additive treatment effects). Nevertheless, given the popularity of population-based estimands, online Appendix A proposes two approaches to population causal inference using BART and evaluates them using extensions of the simulations presented in this article.

3.3.2 BART and Missing Outcome Data

Missing outcome data have a straightforward solution under the assumption of a missing-at-random missing data mechanism. Simply fit the BART model to the complete case sample but make predictions for the full sample.

4. SIMULATIONS BASED ON REAL DATA

Simulations often suffer from too little connection to "real-world" data analysis. The simulations used here are based on covariate data from a real study; only outcomes are simulated. These are generated using response surfaces that ensure ignorability has been satisfied (because they only condition on observed covariates). This also allows us to know the true treatment effect. Treatment effects are estimated under two different conditions with respect to overlap.

I use 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 treated children relative to controls at the end of the intervention (Brooks-Gunn, Liaw, and Klebanov 1991). The study collected data on many pretreatment variables. I 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. There are 6 continuous covariates and 19 binary covariates.

Experimental data are used because they provide a setting within which we can create an observational study but still be guaranteed to satisfy the overlap assumption for certain estimands. Starting with the experimental data, an observational study is created by throwing away a nonrandom portion of the treatment group: all children with nonwhite mothers. This leaves 139 children. The control group remains intact with 608 children. Thus the treatment and control groups are no longer balanced and simple comparisons of outcomes would lead to biased estimates of the treatment effect. Ethnicity was chosen as the variable used to partition the data because it led to subgroups that were more distinct than those yielded by the other categorical variables.

This design ensures that the overlap assumption is satisfied for the treatment group (i.e., the support of X for the treated is a subset of the support of X for the controls); however, overlap is not ensured for the control group (the support of X for the controls is not necessarily a subset of the support of X for the treated). Therefore simulations that target the conditional average treatment effect on the treated (CATT) are referred to as the "overlap" setting. Simulations that target the conditional average treatment effect on

the controls, (CATC), $\sum_{i:Z_i=0} E(Y_i(1) - Y_i(0) \mid X_i)$, are referred to as the "incomplete overlap" setting.

4.1 RESPONSE SURFACES

The $p = 25$ confounding covariates (ethnicity excluded) are used to generate two different response surfaces (for results from a third response surface see online Appendix A). Because the response surface is known, ignorability can be satisfied by appropriately conditioning on the confounding covariates used to generate the response surfaces.

Response surface A takes the form $Y(0) \sim \mathrm{N}(X\beta_A, 1)$ and $Y(1) \sim \mathrm{N}(X\beta_A + 4, 1)$, 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 (0.5, 0.2, 0.15, 0.1, 0.05) that make smaller coefficients more likely. This response surface is linear and parallel across treatment groups; thus all the conditional and population estimands equal 4 (there is no treatment effect heterogeneity). 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 are satisfied. The R^2 from a linear regression of Y on X for this response surface is about 0.95 on average across simulations.

Response surface B is nonlinear and not parallel across treatment conditions, with $Y(0) \sim \mathrm{N}(\exp((X + W)\beta_B), 1)$ and $Y(1) \sim \mathrm{N}(X\beta_B - \omega_B^s, 1)$, where W is an offset matrix of the same dimension as X with every value equal to 0.5, β_B is a vector of regression coefficients (0, 0.1, 0.2, 0.3, 0.4) randomly sampled with probabilities (0.6, 0.1, 0.1, 0.1, 0.1). For the sth simulation, ω_B^s was chosen in the overlap setting, where we estimate the effect of the treatment on the treated, such that CATT equals 4; similarly it was chosen in the incomplete setting, where we estimate the effect of the treatment on the controls, so that CATC equals 4. R^2 values from the linear regression of the outcome on the X values are about 0.78 on average.

4.2 METHODS COMPARED

Inference for CATT or CATC using BART as described in Section 3 is compared with estimates from linear regression, propensity score matching, and a propensity-score-based weighting estimator (described below). The performance of random forests was also tested but the root mean squared error performance of this method was so poor that it is not discussed further. For all BART runs, the default settings were used and no attempt was made to tune the parameters to give optimal results. Linear regression estimates were obtained simply by regressing the outcome on a treatment assignment indicator and the confounding covariates.

Matching estimates relied on one-to-one matching with replacement using the estimated propensity score, $\hat{e}(x)$. The propensity score $e(x) = \Pr(Z = 1 \mid X) = E[Z = 1 \mid X]$ was estimated using logistic regression that was additive in the covariates. This specification yielded mean balance that appeared adequate given the current standards of most applied researchers (details available upon request). More to the point, 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 the goal of helping the researcher to avoid arbitrary judgment calls. Nevertheless, we discuss the implications of alternative propensity score specifications in Sections 5.3 and 6.

Matching required finding a match for each treatment group member from among the controls when estimating CATT and finding a match for each control group member from among the treated when estimating CATC. A weighted regression of outcomes on treatment assignment and all confounders was fit with weights equal to the number of times each child was represented in the matched sample (1 for the inferential group, 0 for unmatched units, and number of times chosen for matched comparison units). "Huber" standard errors (Huber 1967; Hill and Reiter 2006) were calculated.

The second propensity-score-based estimator is a straightforward extension (Imbens 2004; Kurth et al. 2006) of standard inverse-probability-of-treatment-weighted (IPTW) estimators (Rosenbaum 1987; Robins 1999). To estimate the average effect of the treatment on the treated, a weighted regression is run with weights equal to 1 for those in the treatment group and equal to $\hat{e}(x)/(1-\hat{e}(x))$ for those in the control group. To estimate the effect of the treatment on the controls, the controls are weighted as 1 and the treated are assigned weights equal to $(1-\hat{e}(x))/\hat{e}(x)$. As with the matching estimator, robust (Huber) standard errors are calculated. Henceforth this inverse-probability-of-treatment-weighted estimator is referred to simply as the IPTW estimator.

For each setting 1000 simulations were run in R on a Dell Precision Workstation 670n IntelR XeonT with a 3.00 GHz 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.

4.3 RESULTS

Figure 3 displays results from the 1000 simulation runs corresponding to response surface A, which is linear and parallel across the treatment groups. The points represent a random subset of runs (limited to 500 to avoid obscuring key features of the plot) where the deviation of the treatment effect point estimates (posterior means) from the true parameter value (4) is plotted against the 95% interval length. Separate plots are provided for each combination of method (columns) and overlap setting: overlap (top row) or incomplete overlap (second row). The dashed horizontal line represents no deviation from the effect for the treated population. Solid lines that pass through the origin are plotted with slopes 0.5 and -0.5 to discriminate between points with 95% intervals that cover the true value and those that do not. The solid line segments on the right side and bottom of each plot display the bias and average interval length, respectively. 95% posterior intervals for BART were formed 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 and root mean squared error ("RMSE") are displayed at the bottom along with the percentage of times ("Missed") that the

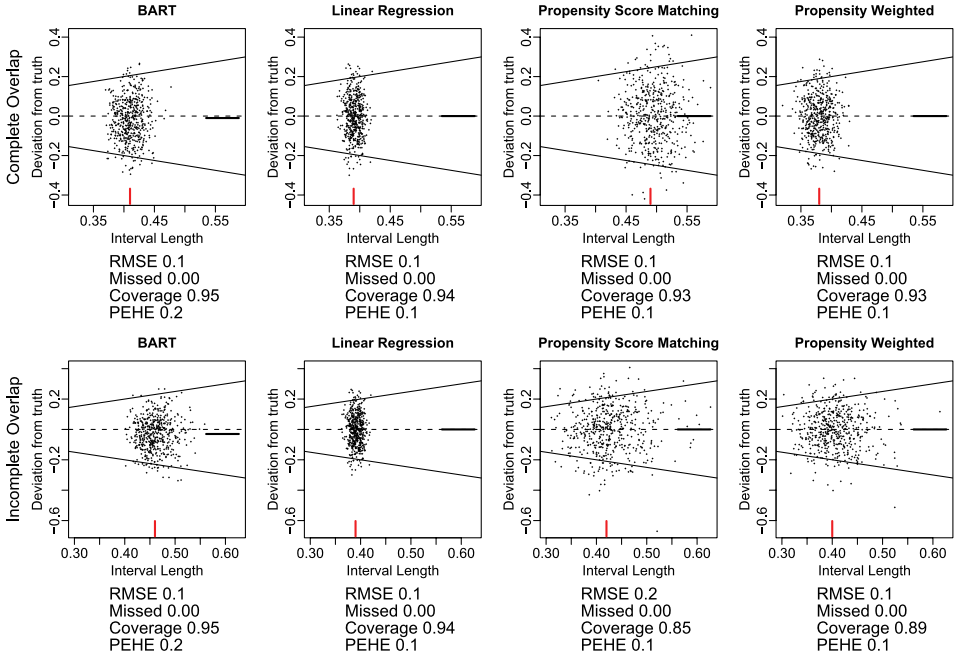

Figure 3. Results from 1000 simulations using linear, parallel response surface A. The deviation of the treatment effect estimates from the true effect for 500 simulation draws (randomly sampled from the 1000) are plotted (y-axis) against interval length on the x-axis, with separate plots for each combination of method (columns) and overlap setting: overlap. Solid 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 parameter value. The short solid line segment on the far right of each plot displays the bias across the 1000 simulations. The short solid line segment at the bottom displays the average interval length. Summary statistics calculated across all 1000 simulations include: root mean squared error (RMSE), percentage of such intervals that would exclude 0 ("Missed"), and coverage rates for 95% intervals. The final summary statistic, precision in estimation of heterogeneous effects ("PEHE"), reflects the ability to capture individual variation in treatment effects, as described in the text. A color version of this figure is available in the electronic version of this article.

method did not detect a significant treatment effect (Type II error) at a 5% significance level. Precision in estimation of heterogeneous effects ("PEHE") is operationalized as $\sqrt{\frac{1}{n_t}\sum_{i:z_i=1}(\bar{c}_i - E[Y_i(1) - Y_i(0) \mid X = x_i])^2}$ for the BART estimator, where $\bar{c}_i = (1/L)\sum_{l=1}^{L} c(x_i, f^l)$. It evaluates the ability of each method to capture treatment effect heterogeneity. For the other estimators the expectation on the right is replaced by the single treatment effect estimate (while in theory separate pair-specific estimates could be used for the propensity matched strategy, this substantially increased the PEHE value on average). These summaries are calculated across all 1000 simulations.

All four methods estimate the treatment effect with virtually no bias and have good coverage properties in the overlap scenario displayed in Figure 3, though the propensity score methods suffer a bit with regard to coverage in the incomplete overlap scenario. All have reasonably similar uncertainty (e.g., average interval length of 0.39 for regression and 0.41 for BART in overlap setting). The matched estimates have a bit more variability overall. All methods detect the significant treatment effect in all cases. PEHE is slightly

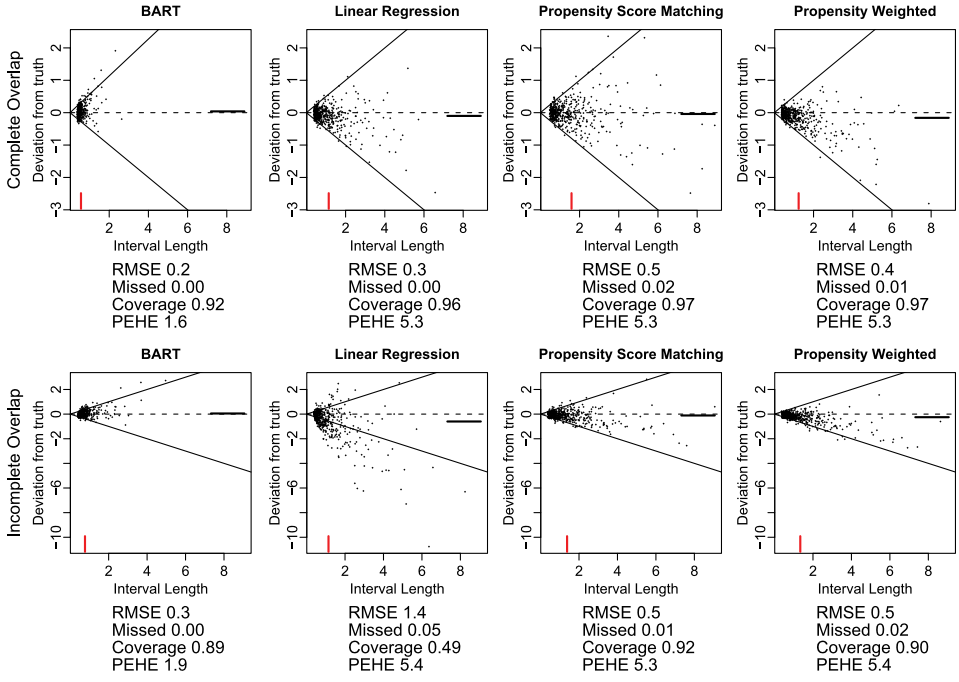

Figure 4. Results from a set of 1000 simulations using nonlinear, not parallel response surface B. See further details in the caption to Figure 3. In the overlap plot with propensity-score-matching results, however, two observations with substantially higher interval lengths have been deleted in order to maintain plot scale in a range where differences between methods can be observed in greater detail. A color version of this figure is available in the electronic version of this article.

bigger for BART than the other methods. However, the differences for all measures are so small relative to the size of the treatment effect 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 4 displays the results for each method from the 1000 simulations corresponding to response surface B. Bias is not much of an issue for any of the methods with the exception of linear regression in the incomplete overlap setting. All three competitors display noticeably higher RMSE than BART in both settings. BART interval lengths are far smaller on average than the regression and propensity-score-matching interval length though more strongly in the complete overlap setting (with ratios that are on average from about one-half to one-third the size of its competitors' intervals). Again, BART uncertainty increases as we move to the setting with less overlap.

Coverage rates fall for all three nonregression methods when there is not complete overlap; it drops dramatically though for linear regression with an average coverage rate of 49%. The propensity score estimators fail to detect the significant treatment effect from 1% to 2% of the time across settings and linear regression fails to detect about 5% of the time in the incomplete overlap setting. PEHE results reflect BART's superiority in capturing heterogeneous treatment effect estimates in both overlap settings. In theory, the

models used by the other methods could be augmented to allow for heterogeneity; however, this would require either prior user knowledge regarding which covariates moderate the treatment effect or an empirical search to uncover such patterns.

5. ROBUSTNESS OF SIMULATION FINDINGS

This section investigates how changes to the distribution and magnitude of the error term or adoption of a more careful propensity score modeling procedure can affect our original simulation results. The findings are summarized here. Full results are available from the author.

5.1 SIMULATIONS WITH t-DISTRIBUTED ERRORS

To explore the implications of the normality assumption in the sum-of-trees model, the error term for the outcomes was drawn instead from a t_2 distribution. The results were basically the same in the overlap setting, except that for response surface A the propensity score matching performed markedly worse than the other three methods with regard to RMSE. For response surface B all methods performed worse but the ratios of RMSEs stayed basically the same. For the incomplete overlap setting simulations, however, BART is now the clear winner for response surface A with a RMSE that is approximately half as big as the others. For response surface B, all the RMSEs increase but the relative rates are very similar.

5.2 SIMULATIONS WITH INCREASED NOISE

To explore the effect of increased sampling variability on our results, the original simulations were rerun increasing the standard deviation first to 5 and then to 10. One important consequence of the increased noise is that, using standard residual plots, it is virtually impossible to detect the nonlinearities for response surface B in either situation. Results of these simulations are available from the author and tell a similar story to the results already presented.

5.3 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 criteria for acceptable balance, including what moments of the covariate distributions should be examined and what statistics should be used, are still not widely understood. Yet treatment effect estimates can be quite sensitive to these choices.

I investigated sensitivity of the propensity score model specification by examining results from two alternative strategies. First, I used standard balance diagnostics (simple standardized differences in means) to choose the propensity score model. This change yielded slightly *worse* results than the simple additive model choice originally used. I also reran 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 for either overlap setting. For response surface B in the complete overlap setting, matching results were quite similar but the IPTW 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 IPTW estimators performed worse than in the original simulations.

It is possible that another model for the propensity score or different matching algorithm might produce more competitive results (see, for instance, McCaffrey, Ridgeway, and Morral 2004; Hansen and Klopfer 2006; Diamond and Sekhon 2008). A broader range of strategies is explored in the Section 6 example.

6. ESTIMATING DOSAGE EFFECTS

This section compares use of BART with that of a range of strong competitors in the context of a real example. I again make use of the data from the IHDP evaluation; however, rather than a simulated outcome, I now use an IQ test (Stanford Binet) score measured at the end of the intervention (age 3). The treatment in this example is level of participation in IHDP child development centers in the two years preceding this test. Although assignment to participate in the program was random, families in the intervention arm self-selected into different participation levels. Therefore if we examine the impact of days attended we have returned, in effect, to an observational setting.[1]

The covariates available in this analysis overlap strongly with those used in the simulation above. However, since the treatment took place starting a year after the study began, a slightly larger array of covariates are available for use as described in the code and data available online. Overall I used a total of nine continuous variables, five binary variables, and eight multicategory variables.

6.1 BINARY TREATMENT VARIABLE

The participation variable ranges for the intervention group (the only ones with access) from 0 days to a maximum of 468 days; the median is 372 days. The first analysis dichotomizes this variable into "high" ($>$400 days) and "low" ($\leq$400 days) participation categories. This simplification allows us to compare BART to methods that are best suited for use with a binary treatment variable. I then attempt to estimate, for those observed to participate more than 400 days, the effect of this high rate of participation as compared to no participation at all (and assignment to the control group). Since "high participation" is the treatment, the goal is to estimate CATT or SATT. After discarding the low participation group (since a value of 0 for the treatment variable for this group does not imply no participation as it does for those not in the intervention group), there are 67 children in the treatment group and 561 in the control group.

[1] A related note is that this causal identification problem cannot be solved by an instrumental variables approach because (among other reasons) access to child development centers was just one of several resources available to the families assigned to the intervention arm. With only one instrument, there is no way to identify the effect of this service isolated from the effects of the other intervention components.

6.1.1 BART Fit

I first used BART to estimate this treatment effect. I provided only three inputs to BART. I provided the outcome variable, `y`. I provided the training data, `xt`, which is a matrix of the treatment variable (first column) and predictor variables for the full sample. Since I also needed to make counterfactual predictions for those children who were observed to participate more than 400 days, I created a subset of `xt` called `xp` consisting of just the treated observations and then reversed their recorded treatment assignment; these were the "test" (prediction) data for the algorithm. I ran BART in `R` (package BayesTree) at the default settings by typing

```
bart.tot <- bart(x.train=xt, y.train=y, x.test=xp)
```

As with any MCMC algorithm one should check convergence, here simply by plotting the residual standard error (`plot(bart.tot$sigma)`). For the lth BART draw I averaged $c(x_i, f^l) = f^l(1, x_i) - f^l(0, x_i)$ over the 67 x_i in the high dose group giving 1000 draws from the posterior distribution of the conditional average treatment effect on the treated,

```
mndiffs = apply(bart.tot$yhat.train[,dose400==1]
          - bart.tot$yhat.test, 1, mean).
```

It is simple to use this empirical distribution to estimate the posterior mean (`mean(mndiffs)`) and standard deviation (`sd(mndiffs)`) or other quantities of interest. The estimate in this example is 12.9 and the standard deviation is 2.0.

6.1.2 Competitors: Matching and Weighting

Matching is probably the most popular strategy today for estimating causal effects for observational studies with point-in-time treatments that assume ignorability. And there are a staggering number of different ways to match. I implemented a variety of options including several of the most sophisticated and promising available with existing software. All made use of estimated propensity scores. However, by no means do I claim to do justice to the full range of possibilities within the world of matching.

I used 40 propensity-score-matching strategies in this example. These comprise combinations of the following features: (1) matching method (nearest neighbor, optimal pair-matching, and full matching (Rosenbaum 1991; Hansen and Klopfer 2006)), (2) ratio of treated to control (1:1, 1:2, 1:3), (3) matching with or without replacement, and (4) method for estimating the propensity score. The first method for estimating the propensity score was the same used in our simulation—each covariate was included additively to a logistic regression. The second method extended this to a quadratic specification with squared terms for each continuous covariate and cross-product terms among all variables. The third method used a generalized additive model with smoothing terms for most of the continuous covariates (smoothing parameters were chosen by cross-validation) as proposed by Woo, Reiter, and Karr (2008). The fourth included all predictors additively and then added cross-product and squared terms for those more predictive of the outcome. Note that not all combinations of these features are possible. Matching was performed using the package

MatchIt (Ho et al. 2010) in R though MatchIt calls Hansen's optmatch program for optimal and full matching (Hansen 2004).

I also used the GenMatch function (package Matching in R) which relies on a genetic search algorithm to determine weights for each input (here covariates and the propensity score estimated with the first method) that lead to optimal balance when used in multivariate matching (Sekhon 2010; Diamond and Sekhon 2008). The function was run two ways: (1) at its default settings (not advised by its developers) and (2) using a loss function that defines balance as minimum distance between the treatment and control groups with respect to the terms in the quadratic function discussed above (but with the propensity score added as an additional continuous variable). Since genetic matching directly includes all covariates as inputs and thus is not dependent on the estimated propensity score (though it tends to benefit from its inclusion), it is not subject to the separation problems that propensity-score-dependent methods can experience with large numbers of covariates.

I also calculated estimates using inverse probability of treatment weighting, as in the simulations above. All four methods for calculating propensity scores were tested for this strategy.

Balance statistics were calculated for each method. Summaries used to distinguish methods were based on both standardized difference in means and maximum distance between empirical QQ plots across groups (Sekhon 2007). These statistics were calculated for each covariate (including the propensity score) as well as for all terms in the quadratic function described above. Balance was summarized separately for three groups: binary variables (all multicategory were converted into separate binary indicators), continuous variables, and all the interactions and squared terms.

Results are presented in Figure 5 in the form of covariance adjusted treatment effect estimates. For each row, only estimates from matching methods that satisfied specified balance criteria are presented. Balance criteria for each row are depicted in the legend (as explained in the table caption). Only the "intelligent" genetic matching choice achieves the highest specified level of balance. The BART estimate of 12.9 is indicated by the solid vertical line on the plot. It is very close to both the genetic matching and full matching estimates.

6.1.3 Comparison of the Results

The BART results and the results from the matching strategies with the best balance are quite similar. However, in practice there is evidence that matching practitioners rarely rigorously test balance beyond simple differences in means. For example, in his review of 47 articles published in the medical literature between 1996 and 2003 that use propensity score matching, Austin (2008) found that 17 did not present any balance results whatsoever and that all but two used methods that arguably were inappropriate for at least some variables. Even in this example, just coercing the existing software to provide balance statistics for the desired terms took a bit of work. BART implementation, on the other hand, was exceedingly simple and yet in this example yielded estimates similar to the matching estimates with the best balance without requiring choosing the best matching method, determining appropriate balance criteria, or performing balance checks.

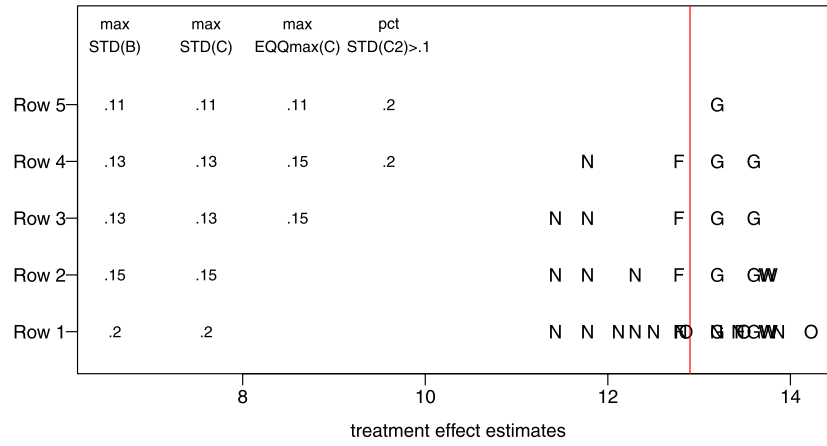

Figure 5. Treatment effect estimates for a variety of matching and weighting strategies. Letters indicate the type of strategy: N indicates nearest neighbor, O indicates optimal pair-matching, F indicates full matching, W indicates weighting, and G indicates genetic matching. The legend on the left displays the balance criteria required of each method for its estimate to be displayed. "max STD(B)" and "max STD(C)" refer to the maximum standardized difference in means between treatment groups for the binary and continuous variables, respectively. "max EQQmax(C)" refers to the maximum across covariates of the maximum difference (for a given covariate) in empirical QQ plots across groups; this is only applicable for the continuous variables. "pct STD(C2)>.1" refers to the percent of standardized difference in means that exceed 0.1 for all the quadratic terms (the squared terms for the continuous and interactions across all variables). The numbers in each row describe the threshold required for admission into that row. A color version of this figure is available in the electronic version of this article.

6.2 CONTINUOUS TREATMENT VARIABLE

A far more difficult challenge for matching estimators is a situation with a continuous treatment variable. It is not completely straightforward to extend matching or weighting estimators to accommodate continuous treatment variables and there is not sufficient space in this article to do justice to the approaches that have been proposed. In this section, therefore, I simply compare the BART results to a linear regression where quadratic terms are added to try to capture the nonlinearity.

BART is easily extended to accommodate a continuous treatment variable. First I fit BART to all the data and then used it to make two sets of predictions of age-3 IQ test scores; these are displayed in Figure 6. The upper line reflects (smoothed) predictions for the intervention group ($I = 1$) at their observed level of participation ($Y(Z = a) \mid I = 1, Z = a$). The lower line reflects (smoothed) predictions for these same children had they not been in the intervention group and thus had 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 receive the intervention (which all these children had been). The right panel plots the (smoothed) differences between these lines and associated uncertainty intervals.

It bears emphasizing that accommodating 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. One such

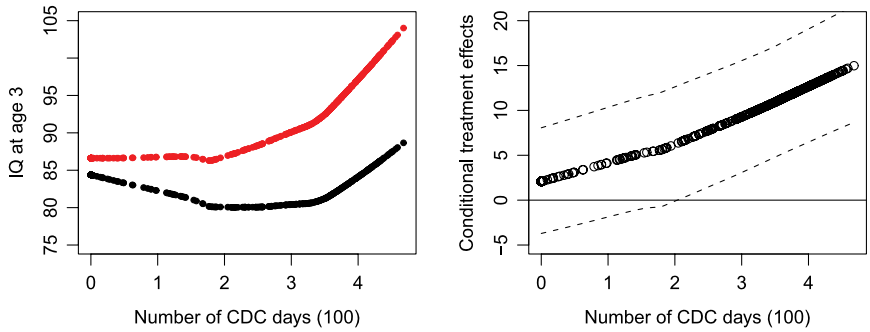

Figure 6. Left panel displays plot of BART-predicted 3-year IQ test scores against CDC participation (in hundreds of days) for children in the treatment group (upper line). The lower line shows predicted scores for the same children if they had not attended any CDC days. Lines were smoothed using lowess. 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. A color version of this figure is available in the electronic version of this article.

set of assumptions is $Y(Z=0) \perp Z \mid X$ and $0 < P(Z=0 \mid X, Z=a) < 1$, for all $a = 0, \ldots, A$.

Under this set of assumptions the difference between the response surfaces at each level of 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 (lowess) function of these treatment effects is displayed, with corresponding uncertainty bounds, in the right panel of Figure 6. This suggests that only children who selected into more than 200 CDC days demonstrated significant effects for 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 attended either the fewest days (perhaps because their parents do not think they need it) or the most days (perhaps because their parents think they are strong or healthy enough to attend regularly).

A linear regression fit to these data (including the covariates) cannot pick up these nonlinearities using quadratic terms. When the participation variable is added linearly the coefficient is strong and statistically significant. When participation squared is also included, however, neither term is statistically significant. These results hold across various specifications (inclusion/exclusion of treatment variable, full sample, or just the treated). Residual plots from these regressions do not demonstrate any discernible pattern. The regression results suggest a linear relationship between participation days and age-3 IQ, whereas BART detects a nonlinear relationship.

7. DISCUSSION

This article has explored the capacity of a new Bayesian nonparametric modeling algorithm, BART, to estimate causal effects. The potential advantages of BART 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 article. Across several different response surface specifications and two different levels of overlap, BART performs very well compared to linear regression, propensity score matching, and an IPTW 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 nonlinear, nonadditive specifications (including those in online Appendix B), BART outperformed the competitors in nearly every combination of setting and performance criterion. A particular strength is BART's ability to identify heterogeneous treatment effects.

BART also has advantages compared to alternative nonparametric or semiparametric methods that might be used to flexibly model the assignment mechanism and the response surface. First BART can handle a large number of both continuous 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 will not 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 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 phenomenon observable in one's data. However, if we build models solely based on past theories, it is 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 regularization priors which place a limit on the amount that this uncertainty will increase and hence can hurt coverage in the case with limited overlap. Concurrent work (Hill and Su 2010) investigates this trade-off as well as ways to identify areas that lack overlap.

BART 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 very high levels of interaction (our simulation setting tested it only through three-way interactions). Finally, only a few different scenarios have been tested and 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.

SUPPLEMENTAL MATERIALS

Computer Code: Tar file containing code for Section 4 simulations and Section 6 examples. A README file in the directory describes the contents. (code.zip)

Data: Tar file containing data for Section 4 simulations and Section 6 examples. A README file in the directory describes the contents. (data.zip)

Appendices: Tar file with online Appendix A and Appendix B. (appendices.zip)

ACKNOWLEDGMENTS

The author gratefully acknowledges financial support for this project from NSF grant 0532400. I thank Andrew Gelman, James Robins, Kosuke Imai, and anonymous referees for helpful comments on the article and the methods.

[Received November 2008. Revised January 2010.]

REFERENCES

Abadie, A., and Imbens, G. W. (2002), "Simple and Bias-Corrected Matching Estimators for Average Treatment Effects," working paper, NBER. [219]

——— (2006), "Large Sample Properties of Matching Estimators for Average Treatment Effects," *Econometrica*, 74 (1), 253–267. [218,220]

Austin, P. (2008), "A Critical Appraisal of Propensity Score Matching in the Medical Literature Between 1996 and 2003," *Statistics in Medicine*, 27, 2037–2049. [234]

Bingenheimer, J., Brennan, R., and Earls, F. (2005), "Firearm Violence Exposure and Serious Violent Behavior," *Science*, 308, 1323–1326. [220]

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. [226]

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, Ser. A*, 169, 1–14. [221]

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. [218,222,223]

——— (2010), "BART: Bayesian Additive Regression Trees," *The Annals of Applied Statistics*, to appear. [218, 222]

Diamond, A., and Sekhon, J. (2008), "Genetic Matching for Estimating Causal Effects: A General Multivariate Matching Method for Achieving Balance in Observational Studies," technical report, U.C. Berkeley. [232, 234]

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. [220]

Hahn, J. (1998), "On the Role of the Propensity Score in Efficient Semiparametric Estimation of Average Treatment Effects," *Econometrica*, 66, 315–322. [220,221]

Hansen, B. (2004), "OPTMATCH, An Add-on Package for R," technical report, University of Michigan. [234]

Hansen, B., and Klopfer, S. O. (2006), "Optimal Full Matching and Related Designs via Network Flows," *Journal of Computational and Graphical Statistics*, 15, 609–627. [232,233]

Hastie, T., Tibshirani, R., and Friedman, J. (2003), *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, New York: Springer-Verlag. [221]

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. [217,221]

Hill, J., and Reiter, J. (2006), "Interval Estimation for Treatment Effects Using Propensity Score Matching," *Statistics in Medicine*, 25 (13), 2230–2256. [218,228]

Hill, J. L., and Su, Y.-S. (2010), "Assessing Common Support for Causal Inference in High-Dimensional Covariate Space," technical report, New York University. [237]

Hirano, K., Imbens, G. W., and Ridder, G. (2003), "Efficient Estimation of Average Treatment Effects Using the Estimated Propensity Score," *Econometrica*, 71, 1161–1189. [217,221]

Ho, D. E., Imai, K., King, G., and Stuart, E. A. (2010), "MatchIt: Nonparametric Preprocessing for Parametric Causal Inference," *Journal of Statistical Software*, to appear. [234]

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*, Berkeley, CA: University of California Press, pp. 221–233. [228]

Imbens, G. (2004), "Nonparametric Estimation of Average Treatment Effects Under Exogeneity: A Review," *The Review of Economics and Statistics*, 86 (1), 4–29. [217-219,228]

Kang, J., and Schafer, J. (2007), "Demystifying Double Robustness: A Comparison of Alternative Strategies for Estimating a Population Mean From Incomplete Data" (with discussion), *Statistical Science*, 22, 523–580. [221]

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. [217,221,228]

Leamer, E. (1983), "Let's Take the Con Out of Econometrics," *American Economic Review*, 73, 31–43. [224]

McCaffrey, D., Ridgeway, G., and Morral, A. (2004), "Propensity Score Estimation With Boosted Regression for Evaluating Causal Effects in Observational Studies," *Psychological Methods*, 9, 403–425. [232]

Robins, J. M. (1999), "Association, Causation, and Marginal Structural Models," *Synthese*, 121, 151–179. [228]

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. [217]

Robins, J. M., and Rotnitzky, A. (1995), "Semiparametric Efficiency in Multivariate Regression Models With Missing Data," *Journal of the American Statistical Association*, 90, 122–129. [221]

Robins, J. M., Hernan, M. A., and Brumback, B. (2000), "Marginal Structural Models and Causal Inference in Epidemiology," *Epidemiology*, 11, 550–560. [217,220]

Rosenbaum, P. R. (1987), "Model-Based Direct Adjustment," *Journal of the American Statistical Association*, 82, 387–394. [220,221,228]

——— (1991), "A Characterization of Optimal Designs for Observational Studies," *Journal of the Royal Statistical Society, Ser. B*, 53, 597–610. [233]

——— (2002), *Observational Studies*, New York: Springer. [225]

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. [220]

Rotnitzky, A., Robins, J. M., and Scharfstein, D. (1998), "Semiparametric Regression for Repeated Outcomes With Nonignorable Nonresponse," *Journal of the American Statistical Association*, 93, 1321–1339. [221]

Rubin, D. B. (1973), "The Use of Matched Sampling and Regression Adjustment to Remove Bias in Observational Studies," *Biometrics*, 29, 185–203. [217]

——— (1978), "Bayesian Inference for Causal Effects: The Role of Randomization," *The Annals of Statistics*, 6, 34–58. [218]

——— (1979), "Using Multivariate Matched Sampling and Regression Adjustment to Control Bias in Observational Studies," *Journal of the American Statistical Association*, 74, 318–328. [217]

Rubin, D. B., and Thomas, N. (2000), "Combining Propensity Score Matching With Additional Adjustments for Prognostic Covariates," *Journal of the American Statistical Association*, 95, 573–585. [217,221]

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. [221]

Scott, D., and Bauer, C. (1989), "A Neonatal Health Index for Preterm Infants," *Pediatric Research*, 25, 263A. [226]

Sekhon, J. S. (2007), "Alternative Balance Metrics for Bias Reduction in Matching Methods for Causal Inference," technical report, U.C. Berkeley. [231,234]

——— (2010), "Multivariate and Propensity Score Matching Software With Automated Balance Optimization: The Matching Package for R," *Journal of Statistical Software*, to appear. [234]

Woo, M.-J., Reiter, J. P., and Karr, A. F. (2008), "Estimation of Propensity Scores Using Generalized Additive Models," *Statistics in Medicine*, 17, 3805–3816. [233]