BAYESIAN REGRESSION TREE MODELS FOR CAUSAL INFERENCE: REGULARIZATION, CONFOUNDING, AND HETEROGENEOUS EFFECTS

By P. Richard Hahn, Jared Murray and Carlos M. Carvalho

University of Chicago, Carnegie Mellon University and University of Texas

This paper develops a semi-parametric Bayesian regression model for estimating heterogeneous treatment effects from observational data. Standard nonlinear regression models, which may work quite well for prediction, can yield badly biased estimates of treatment effects when fit to data with strong confounding. Our Bayesian causal forests model avoids this problem by directly incorporating an estimate of the propensity function in the specification of the response model, implicitly inducing a covariate-dependent prior on the regression function. This new parametrization also allows treatment heterogeneity to be regularized separately from the prognostic effect of control variables, making it possible to informatively "shrink to homogeneity", in contrast to existing Bayesian non- and semi-parametric approaches.

1. Introduction. The success of modern predictive modeling is founded on the understanding that flexible predictive models must be carefully regularized in order to achieve good out-of-sample performance (low generalization error). In a causal inference setting, regularization is less straightforward: In the presence of confounding, regularized models originally designed for prediction can bias causal estimates towards some unknown function of high dimensional nuisance parameters (Hahn et al., 2016). That is, despite offering excellent predictive performance, the causal conclusions from a naively regularized nonlinear regression are likely to be substantially biased, leading to high estimation error of the target parameter. A key finding in this paper is that this effect will be especially pronounced in flexible models which allow for heterogeneous effects.

To mitigate these estimation problems we propose a flexible sum-of-regression-trees — a *forest* — to model a response variable as a function of a binary treatment indicator and a vector of control variables. Our work departs from existing contributions – primarily Hill (2011) and later extensions

Keywords and phrases: Bayesian; Causal inference; Heterogeneous treatment effects; Predictor-dependent priors; Machine learning; Regression trees; Regularization; Shrinkage

– in two important respects: First, we develop a novel prior for the response surface that depends explicitly on estimates of the propensity score as an important 1-dimensional transformation of the covariates (including the treatment assignment). Incorporating this transformation of the covariates is not strictly necessary in response surface modeling, but we show that it can substantially improve treatment effect estimation in the presence of strong confounding.

Second, we represent our regression as a sum of two functions: the first captures the *prognostic* impact of the control variables (the component of the conditional mean of the response that is unrelated to the treatment effect), while the second captures the treatment effect, which itself is a nonlinear function of the observed attributes (capturing possibly heterogeneous effects). We represent each function as a forest. This approach allows the degree of shrinkage on the treatment effect to be modulated *directly* and *separately* of the prognostic effect. In particular, under this parametrization, standard regression tree priors shrink towards homogeneous effects. In previous approaches, the prior distribution over treatment effects is induced indirectly, and is therefore difficult to understand and control.

Comparisons on simulated data show that the new model — which we call the Bayesian causal forest model — performs at least as well as existing approaches for estimating heterogenous treatment effects across a range of plausible data generating processes. More importantly, it performs dramatically better in many cases, especially those with strong confounding and relatively weak treatment effects, which we believe to be common in applied settings.

As we have noted, the Bayesian causal forest model directly extends ideas from two earlier papers: Hill (2011) and Hahn et al. (2016). Specifically, this paper studies the "regularization-induced confounding" of Hahn et al. (2016) in the context of nonparametric Bayesian models as utilized by Hill (2011). In terms of implementation, this paper builds explicitly on the work of Chipman, George and McCulloch (2010); see also Gramacy and Lee (2008) and Murray (2017). Other notable work on Bayesian treatment effect estimation includes Gustafson and Greenland (2006),Zigler and Dominici (2014), Heckman, Lopes and Piatek (2014), Li and Tobias (2014), and Taddy et al. (2016). A more complete discussion of how the new method relates to this earlier literature, including non-Bayesian approaches, is deferred until Section 7.

2. Problem statement and notation. Let Y denote a scalar response variable and Z denote a binary treatment indicator variable. Capital Roman letters denote random variables, while realized values appear in lower case, that is, y and z. Let x denote a length d vector of observed control

variables. Throughout, we will consider an observed sample of size n independent observations (Y_i, Z_i, x_i), for $i = 1, \ldots n$. When Y or Z (respectively, y or z) are without a subscript, they denote length n column vectors; likewise, $\mathbf{X}$ will denote the $n \times d$ matrix of control variables.

We are interested in estimating various treatment effects. In particular, we are interested in individual treatment effects (ITE) — the amount by which the response Y_i would differ between hypothetical worlds in which the treatment was set to $Z_i = 1$ versus $Z_i = 0$ – or their aggregation over the sample, or specific subgroups. This kind of *counterfactual* estimand can be formalized in the *potential outcomes* framework (Imbens and Rubin (2015), chapter 1) by using $Y(0)_i$ and $Y(1)_i$ to denote the outcomes we would have observed if treatment were set to zero or one, respectively. We observe one of the two potential outcomes: $Y_i = Z_i Y_i(1) + (1 - Z_i)Y_i(0)$.

Throughout the paper we will assume that *strong ignorability* holds, which stipulates that

$$Y(0), Y(1) \perp\!\!\!\perp Z \mid \mathbf{X}. \tag{1}$$

and also that

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

for all $i = 1, \ldots, n$. The first condition assumes we have no unmeasured confounders, and the second condition (overlap) is necessary to estimate treatment effects everywhere in covariate space. Provided that these conditions hold, it follows that $\mathrm{E}(Y_i(1) \mid \mathrm{x}_i) = \mathrm{E}(Y_i \mid \mathrm{x}_i, Z_i = 1)$ and $\mathrm{E}(Y_i(0) \mid \mathrm{x}_i) = \mathrm{E}(Y_i \mid \mathrm{x}_i, Z_i = 0)$ and our estimand may be expressed as

$$\alpha(\mathrm{x}_i) := \mathrm{E}(Y_i \mid \mathrm{x}_i, Z_i = 1) - \mathrm{E}(Y_i \mid \mathrm{x}_i, Z_i = 0). \tag{3}$$

For simplicity, we restrict attention to mean-zero additive error representations

$$Y_i = f(\mathrm{x}_i, z_i) + \epsilon_i, \tag{4}$$

so that $\mathrm{E}(Y_i \mid \mathrm{x}_i, z_i) = f(\mathrm{x}_i, z_i)$. The treatment effect of setting $z_i = 1$ versus $z_i = 0$ can therefore be expressed as

$$\alpha(\mathrm{x}_i) := f(\mathrm{x}_i, 1) - f(\mathrm{x}_i, 0).$$

In this context, (1) can be expressed equivalently as $\epsilon_i \perp\!\!\!\perp Z_i \mid \mathrm{x}_i$.

We propose new prior distributions on f that improve estimation of the parameter of interest, namely α. Previous work (Hill, 2011) advocated using a Bayesian additive regression tree (BART)

prior for $f(\mathrm{x}_i, z_i)$ directly. We instead recommend expressing the response surface as

$$\mathrm{E}(Y_i \mid \mathrm{x}_i, Z_i = z_i) = m(\mathrm{x}_i, \hat{\pi}(\mathrm{x}_i)) + \alpha(\mathrm{x}_i, \hat{\pi}(\mathrm{x}_i))z_i, \tag{5}$$

where the functions m and α are given independent BART priors and $\hat{\pi}(\mathrm{x}_i)$ is an estimate of the propensity score $\pi(\mathrm{x}_i) = \Pr(Z_i = 1 \mid \mathrm{x}_i)$. The following sections motivate this model specification and provide additional context; further modeling details are given in section 5.

3. Bayesian additive regression trees for heterogeneous treatment effect estimation. Hill (2011) observed that under strong ignorability, treatment effect estimation reduces to response surface estimation. That is, provided that a sufficiently rich collection of control variables are available (to ensure strong ignorability), treatment effect estimation can proceed "merely" by estimating the conditional expectations $\mathrm{E}(Y \mid \mathrm{x}, Z = 1)$ and $\mathrm{E}(Y \mid \mathrm{x}, Z = 0)$. Noting its strong performance in prediction tasks, Hill (2011) advocates the use of the Bayesian additive regression tree (BART) model of Chipman, George and McCulloch (2010) for estimating these conditional expectations. BART is particularly well-suited to detecting interactions and discontinuities, can be made invariant to monotone transformations of the covariates, and typically requires little parameter tuning. BART has been used successfully in applications, for example Green and Kern (2012), Hill et al. (2013), Kern et al. (2016), and Sivaganesan, Müller and Huang (2017). It has subsequently been demonstrated to successfully infer treatment effects in multiple, independent simulation studies (Hill et al., 2017; Wendling et al., 2017), often outperforming competitors.

The BART prior expresses an unknown function $f(\mathrm{x})$ as a sum of many piecewise constant binary regression trees. (In this section, we suppress z in the notation; implicitly z may be considered as a coordinate of x.) Each tree $T_l,\ 1 \leq l \leq L$, consists of a set of internal decision nodes which define a partition of the covariate space (say $\mathcal{A}_1, \ldots, \mathcal{A}_{B(l)}$), as well as a set of terminal nodes or leaves corresponding to each element of the partition. Further, each element of the partition $\mathcal{A}_b$ is associated a parameter value, μ_{lb}. Taken together the partition and the leaf parameters define a piecewise constant function: $g_l(x) = \mu_{lb}$ if $x \in \mathcal{A}_b$; see Figure 1.

Individual regression trees are then additively combined into a single regression *forest*: $f(\mathrm{x}) = \sum_{l=1}^{L} g_l(\mathrm{x})$. Each of the functions g_l are constrained by their prior to be "weak learners" in the sense that the prior favors small trees and leaf parameters that are near zero. Each tree follows (independently) the prior described in Chipman, George and McCulloch (1998): the probability that a node at depth h splits is given by $\eta(1+h)^{-\beta},\ \ \eta \in (0,1),\ \beta \in [0,\infty)$.

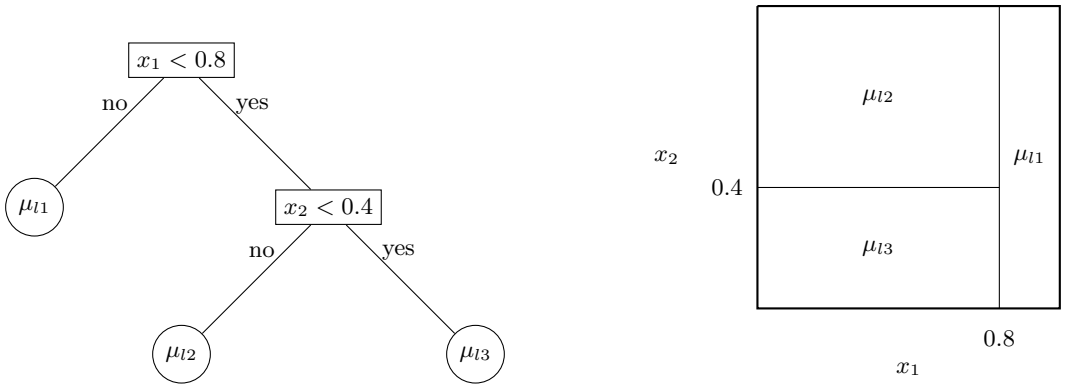

FIG 1. *(Left) An example binary tree, with internal nodes labelled by their splitting rules and terminal nodes labelled with the corresponding parameters* μ_{lb}*. (Right) The corresponding partition of the sample space and the step function.*

A variable to split on, as well as a cut-point to split at, are then selected uniformly at random from the available splitting rules. Large, deep trees are given extremely low prior probability by taking $\eta = 0.95$ and $\beta = 2$ as in Chipman, George and McCulloch (2010). The leaf parameters are assigned independent priors $\mu_{lb} \sim \mathrm{N}(0, \sigma_\mu^2)$ where $\sigma_\mu = \sigma_0/\sqrt{L}$. The induced marginal prior for $f(\mathrm{x})$ is centered at zero and puts approximately 95% of the prior mass within $\pm 2\sigma_0$ (pointwise), and σ_0 can be used to calibrate the plausible range of the regression function. Full details of the BART prior and its implementation are given by Chipman, George and McCulloch (2010).

Here we are concerned with the impact that the prior over $f(\mathrm{x}, z)$ has on estimating $\alpha(\mathrm{x}) = f(\mathrm{x}, z = 1) - f(\mathrm{x}, z = 0)$. The choice of BART as a prior over f has particular implications for the induced prior on the treatment effect that are difficult to understand: In particular, the induced prior will vary with the dimension of x and the degree of correlation with z. The next section develops a framework for investigating the prior impact on the treatment effect estimate.

4. Regularization-induced confounding. Because treatment effects may be deduced from the conditional expectation function $f(\mathrm{x}_i, z_i)$, a likelihood perspective suggests that the conditional distribution of Y, given x and Z, is sufficient to estimate treatment effects. While this is true in terms of *identification* of treatment effects, the question of estimation based on finitely many observations is more nuanced. In particular, many functions in the support of the prior will yield approximately equivalent likelihood evaluations, but may imply substantially different treatment effects. This is particularly true in a strong confounding-modest treatment effect regime, where the conditional expectation of Y is largely determined by x rather than Z.

Accordingly, the posterior estimate of the treatment effect is apt to be substantially influenced by

the prior distribution over f for realistic sample sizes. This issue was explored by Hahn et al. (2016) in the narrow context of linear regression with continuous treatment and homogenous treatment effect. In the linear regression setting an exact expression for the bias on the treatment effect under standard regularization priors is available in closed form, which is a function of the unknown regression coefficients on the control variables; see Appendix A for details. Hahn et al. (2016) call this phenomenon "regularization-induced confounding" (RIC). With more complicated semiparametric models, a closed-form expression of the bias is infeasible — nonetheless, the problem persists. The following toy example provides a striking illustration of the RIC phenomenon, which we use to provide a heuristic understanding of the problem and motivate our solution.

Example: $d = 2$, $n = 1,000$, homogeneous effects. Consider the following simple data generating process:

$$\begin{aligned} Y_i &= \mu_i + Z_i + \epsilon_i, \\ \mu_i &= \mathbb{1}(x_{i1} < x_{i2}) - \mathbb{1}(x_{i1} \geq x_{i2}) \\ P(Z_i = 1 \mid x_{i1}, x_{i2}) &= \Phi(\mu_i), \\ \epsilon_i \overset{\text{iid}}{\sim} \mathrm{N}(0, 0.7^2), &\quad x_{i1}, x_{i2} \overset{\text{iid}}{\sim} \mathrm{N}(0, 1). \end{aligned} \tag{6}$$

To aid intuition, this data generating process can be thought of in terms of the following narrative. Imagine that Y is a continuous biometric measure of heart distress, Z is an indicator for having received a new heart medication, and x_1 and x_2 are systolic and diastolic blood pressure (in standardized units), respectively. Suppose that due to prescription guidelines, patients with $x_{i1} < x_{i2}$ are approximately five times as likely to receive the new drug (given by the ratio $\Phi(1)/\Phi(-1)$). Underlying these guidelines is a biological process such that individuals with $x_{i1} < x_{i2}$ also have higher Y measurements on average; specifically $\mu_i = \mathrm{E}(Y \mid Z_i = 0, x_{i1}, x_{i2})$ is either 1 or -1 according to whether $x_{i1} < x_{i2}$ or not. That is, the guidelines target individuals who, based on their blood pressure measurements, are more likely to have the symptom (high levels of heart distress) that the drug is intended to treat.

Next, we fit three BART models to $n = 1,000$ observations from this process. The first model uses the "vanilla" BART prior over $f(\mathrm{x}_i, z_i)$, the one endorsed by Hill (2011). The second model is the infeasible "oracle" BART model that includes the true (unknown) propensity score π among its predictor variables. The third model is an empirical approximation to oracle BART, which uses a plug-in estimate for $\hat{\pi}$. In this example – and throughout the rest of the paper – we obtain $\hat{\pi}$ as

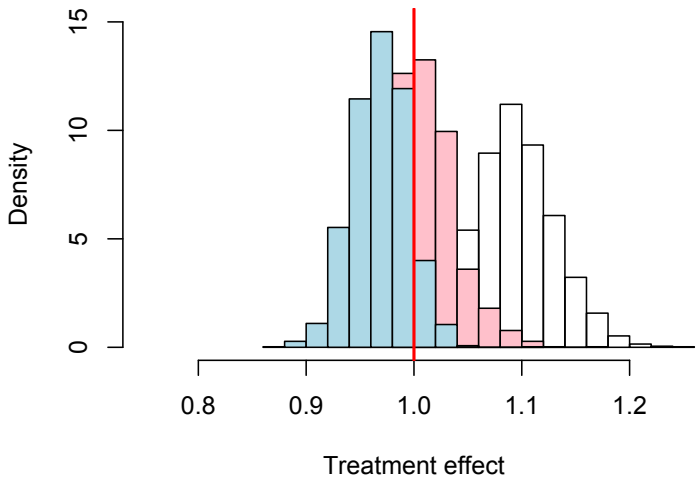

FIG 2. *The vanilla BART prior (white) exhibits substantial bias in estimating the treatment effect. Modified BART priors allowing splits in either the true propensity score (blue) or an estimated propensity score (pink) perform markedly better. The true treatment efffect is* $\alpha = 1$*, depicted by the vertical red line.*

TABLE 1

The vanilla BART prior exhibits substantial bias in estimating the treatment effect. Modified BART priors allowing splits in either the true propensity score (Oracle ps-BART) or an estimated propensity score (ps-BART) perform markedly better.

Prior	Bias	Coverage	RMSE
Vanilla BART	0.14	31%	0.15
Oracle ps-BART	0.00	98%	0.05
ps-BART	0.06	85%	0.08

the posterior mean from a BART fit to the treatment variable (using a probit link formulation); we discuss alternative options in the discussion section.

Figure 2 shows histograms based on posterior draws for each of the three models. The vanilla BART model (white) is substantially off-center of the truth, which is $\alpha = 1$. The two models which explicitly include the selection process within the regression model are markedly closer to the truth; both of the "propensity adjusted" models contain the truth within their 95% posterior credible interval, while the vanilla model does not. More troublesome for the vanilla BART approach is that this behavior is typical across data sets; the bias, coverage and root mean squared error of the three models across 100 replications are reported in Table 1.

In Hahn et al. (2016), regularization-induced confounding was most evident when the sample size was small relative to the number of control variables; this example shows that the RIC problem is

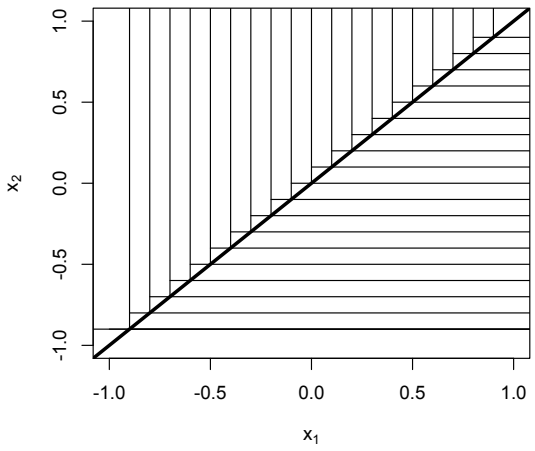

FIG 3. *Many axis-aligned splits are required to approximate a step function along the diagonal; in our example* $m(\mathrm{x}) = 1$ *above the diagonal and* $m(\mathrm{x}) = -1$ *below the diagonal. If these two regions correspond to disparate rates of treatment, regularized regression is apt to overstate the treatment effect.*

more pronounced in nonlinear models, even with a modest number of control variables ($d = 2$) and relatively large sample sizes ($n = 1,000$).

How do we explain these results? In this example, the problem arises because the propensity function requires a substantial number of axis-aligned splits to be represented as a tree/forest (see Figure 3); meanwhile, the response surface itself can be parsimoniously approximated with just a few coordinate splits in the treatment variable. Thus, a prior over f that penalizes the total number of splits will, under this particular confounding structure, tend to attribute changes in the expected value of Y entirely to a treatment effect. In general, when there is strong confounding and the nature of the confounding is such that the contribution of x to the conditional expectation of Y is similar in form to some function of the propensity score, a prior that encourages parsimony may favor attributing this effect to the univariate treatment Z rather than attempting to capture a complicated function of the covariates.

Fortunately, this problem is mitigated by including an estimate of the propensity score as a predictor variable within the BART model — doing so guarantees that splits in Z (treatment effect) are penalized exactly the same as splits in $\hat{\pi}$ (confounding). See section 7 for a discussion of how this modification can be viewed as a covariate dependent prior over f.

Finally, while the above example considered a homogeneous treatment effect, the problem does not become any easier when estimating heterogeneous effects. The next section extends the insights

of this section to a model specifically designed for estimating heterogeneous treatment effects. This new model offers demonstrable improvements over the ps-BART in terms of estimating heterogeneous effects.

5. Bayesian causal forests for heterogeneous treatment effects. In much the same way that a direct BART prior on f does not allow careful handling of confounding, it also does not allow separate control over the discovery of heterogeneous effects because there is no explicit control over how f varies in Z.

Our solution to this problem is a simple re-parametrization that avoids the indirect specification of a prior over the treatment effects:

$$f(\mathrm{x}_i, z_i) = m(\mathrm{x}_i) + \alpha(\mathrm{x}_i) z_i, \tag{7}$$

which is a linear regression in Z with covariate-dependent functions for the slope and the intercept. Writing the model this way sacrifices nothing in terms of expressiveness, but permits an independent prior to be placed on α, which is precisely the treatment effect:

$$\mathrm{E}(Y_i \mid \mathrm{x}_i, Z_i = 1) - \mathrm{E}(Y_i \mid \mathrm{x}_i, Z_i = 0) = \{m(\mathrm{x}_i) + \alpha(\mathrm{x}_i)\} - m(\mathrm{x}_i) = \alpha(\mathrm{x}_i). \tag{8}$$

Based on our observations in the previous section, we further propose specifying the model as

$$f(\mathrm{x}_i, z_i) = m(\mathrm{x}_i, \hat{\pi}_i) + \alpha(\mathrm{x}_i, \hat{\pi}_i) z_i, \tag{9}$$

where $\hat{\pi}_i$ is an estimate of the propensity score .

With a default BART prior on α we have made some progress — shrinking α towards small, additive functions with zero-centered priors on the leaf parameters (as the BART prior does) implies the desired prior bias. However, some further modification is useful. We propose two modifications.

First, note that homogeneous treatment effects correspond to "stumps", trees that are only root nodes. So it is reasonable to put more probability on trees of depth 0 than the default BART prior does. An alternative way to motivate this choice is through the observation that $\alpha(\mathrm{x}_i, \hat{\pi}_i) z_i$ is structurally equivalent to a forest of trees which all have at least one initial split on Z, meaning that the trees in α all effectively start at depth 1. Specifically, we adjust the BART prior by setting $\beta = 3$ ($\beta = 2$ for the default BART prior). Moreover, we may choose the splitting probability parameter η by fixing the probability of effect homogeneity (the probability of no splits) to some constant $a_0 \in (0, 1)$, solving for $a_0 = (1 - \eta)^{L_\alpha}$, where L_α are the number of trees comprising α.

TABLE 2
BART vs ps-BART vs BCF on RMSE on the heterogeneous effect vector $\alpha(\mathrm{x})$ over 250 replicates.

Prior	coverage of ATE	ave. RMSE ATE	ave. RMSE ITE
BART	3%	0.53	0.63
ps-BART	96%	0.11	0.34
BCF	94%	0.10	0.25

Finally, we decompose L, the total number of trees, as $L = L_\alpha + L_m$, with the prior on the leaf parameters as before, but now L depends on the total number of trees in both α and m.

Second, we us a hyperprior on the scale of α. In the usual BART prior, $\alpha(\mathrm{x}_i) \sim \mathrm{N}(0, v_\alpha)$ marginally, with v_α fixed. Instead, we assume the scale follows a half-Cauchy distribution: $v \sim C(0, v_0)_+$ where v_0 is a fixed hyperparameter. This provides a guard against spurious inferences in the presence of null or undetectable effects.

5.1. *Computation.* Posterior sampling of this model is possible with a straightforward modification of the Bayesian back-fitting algorithm described in Chipman, George and McCulloch (2010). In particular, when updating the trees defining α one simply restricts the algorithm to consider only observations with $Z_i = 1$. The half Cauchy prior on v_α is implemented using the augmented parameterization $\alpha(\mathrm{x}_i) = s \times b(\mathrm{x}_i)$ with $s \sim \mathrm{N}(0, 1)$ and $b \sim$ BART (with the above modifications); see Gelman et al. (2006) for additional details. Under this augmented parametrization, the Gibbs step for s is a conjugate Gaussian update based on "data" $r_i := \{y_i - m(\mathrm{x}_i, \hat{\pi}(\mathrm{x}_i))\}/\{b(x_i, \hat{\pi}(\mathrm{x}_i))z_i\}$.

Example: $d = 3$, $n = 250$, heterogeneous effects. Here we compare naive BART, EPS-BART, and our new Bayesian causal forest (BCF) model on the same data generating process as the previous example, except now the treatment effect varies depending on an observable covariate $x_3 \sim \mathrm{N}(0, 1)$:

$$\alpha_i = \mathbb{1}(x_{i3} > 1/4) + \frac{1}{4}\mathbb{1}(x_{i3} > 1/2) + \frac{1}{2}\mathbb{1}(x_{i3} > 3/4),$$

so $\alpha_i \in \{0, 1, 1.25, 1.5\}$ according to the level of x_{i3}. We consider a smaller sample size, $n = 250$, to mimic more closely the situation in practice where the sample size is modest compared to the number of covariates, while keeping the number of variables as small as possible for ease of understanding. Table 2 confirms that, as anticipated, the causal forest model has reduced estimation error on the vector of heterogeneous effects.

6. Empirical analysis.

6.1. *Background and data.* As an empirical demonstration of the Bayesian causal forest model, we consider the question of how smoking affects medical expenditures. This question is of interest as it relates to lawsuits against the tobacco industry. The lack of experimental data speaking to this question motivates the reliance on observational data. This question has been studied in several previous papers; see Zeger et al. (2000) and references therein. Here, we follow Imai and Van Dyk (2004) in analyzing data extracted from the 1987 National Medical Expenditure Survey (NMES) by Johnson et al. (2003). The NMES records many subject-level covariates and boasts third-party-verified medical expenses. Specifically, our regression includes the following nine patient attributes:

- `age`: age in years at the time of the survey
- `smoke_age`: age in years when the individual started smoking
- `gender`: male or female
- `race`: other, black or white
- `marriage_status`: married, widowed, divorced, separated, never married
- `education_level`: college graduate, some college, high school graduate, other
- `census_region`: geographic location, Northeast, Midwest, South, West
- `poverty_status`: poor, near poor, low income, middle income, high income
- `seat_belt`: does patient regularly use a seat belt when in a car

The response variable is the natural logarithm of annual medical expenditures, which makes the normality of the errors more plausible. Under this transformation, the treatment effect corresponds to a multiplicative effect on medical expenditure. Following Imai and Van Dyk (2004), we restrict our analysis to smokers who had non-zero medical expenditure. After making these restrictions, our sample consists of $n = 7,752$ individuals. Our treatment variable is an indicator for heavy smoking, where we define heavy smoking as smoking at least half a pack (10 cigarettes) per day.

6.2. *Comparing BCF to vanilla BART.* We begin by comparing the estimated individual causal effects from the Bayesian causal forest model to those obtained using vanilla BART, as described in Hill (2011). The top panel of Figure 6.2 shows a scatterplot of the predictions of each model, indicating close agreement between the two. Despite the similarity in the fitted values, the two models differ markedly in their estimates of treatment effects. The bottom panel of Figure 6.2 shows the analogous scatterplot for the estimated causal effect. The estimates deviate sharply from the diagonal identity line. Further, the overall range of ITEs estimated by the vanilla BART model is

much wider and includes a substantial number of negative estimates. This pattern is consistent with BART suffering from the RIC phenomenon. In addition the wide variability in the ITE estimates, the posterior distribution of the sample average treatment effect under vanilla BART has a higher mean than under BCF (0.24 compared to 0.12) and greater dispersion; see Figure 6.2.

6.3. *Identifying heterogeneous subgroups.* As depicted in the histograms in Figure 6.2, there is substantial posterior evidence of a positive, but small, sample average treatment effect (SATE). Specifically, under the BCF model the posterior mean SATE is 0.11. The posterior probability that the SATE is positive is greater than 99%, and the 95% credible interval is $(0.03,\ 0.20)$.

Figure 6.3 shows a substantial range in point estimates of the individual effects, ranging from 0.02 to 0.2, however each of these individual estimates comes with substantial posterior uncertainty: only 2,003 out of 7,752 individuals have 95% credible intervals not containing zero. This result squares with the intuition that estimating an treatment effect for any single person cannot be done with great certainty.

However, just as the SATE posterior exhibited far greater concentration, averaging subgroups of individuals to obtain subgroup average treatment effects one might still expect to uncover convincing evidence of heterogeneity. We explore this by fitting a classification tree to an indicator variable recording which individuals have posterior credible intervals not containing zero. Once a subgroup of interest is isolated, the posterior of the subgroup ATE can be examined, as can the posterior of the *difference* in treatment effect between a given subgroup and everyone else.

Our exploratory analysis suggests that less educated (high school graduate), married individuals with a relatively high propensity for smoking ($\hat{\pi}(\mathrm{x}_i) > 0.63$) have a higher treatment effect than everyone else. Indeed, these roughly 2,000 individuals have a posterior mean subgroup ATE of 0.16, whereas the overall ATE has posterior mean of 0.11. Figure 6.3 shows a histogram of posterior samples of $\Delta := \mathrm{ATE}_s - \mathrm{ATE}_{-s}$, where s refers to the subgroup described immediately above. The posterior probability that $\Delta > 0$ is 85%, providing fairly strong evidence that this subgroup differs meaningful from other subjects in terms of treatment effect. One possibility is that education and marriage status affect smokers' perceived risk of smoking, which in turn directly impacts their consumption of medical services. It is additionally possible that an unmeasured variable is driving the heterogeneity. For example, it may be that differences in the presence or quality of health insurance drive utilization of medical services, and education and marriage status are simply correlates of the unobserved insurance coverage status. When estimating heterogeneous treatment

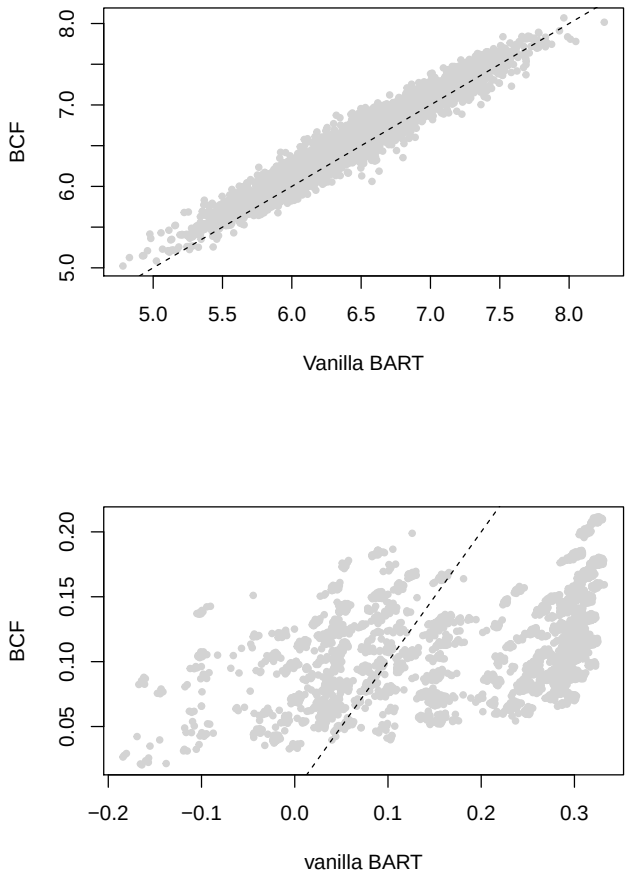

FIG 4. *Top panel: A scatterplot of predicted responses for each individual in the sample from the vanilla BART model and the Bayesian causal forest model. The points closely hug the diagonal identity line, revealing that the two models make very similar predictions about the response surface. Bottom panel: A scatterplot of estimated individual causal effects from the vanilla BART model and the Bayesian causal forest model. The points deviate markedly from the diagonal identity line, revealing that the two models draw very different causal conclusions, despite estimating very similar response surfaces.*

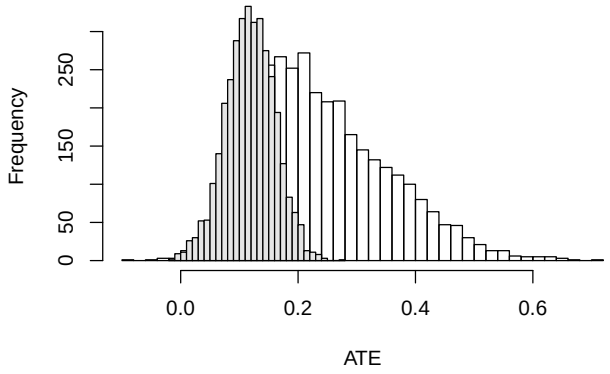

FIG 5. *A histogram of the vanilla BART (white) and Bayesian causal forest (gray) average treatment effect posteriors, based on 3,500 posterior samples.*

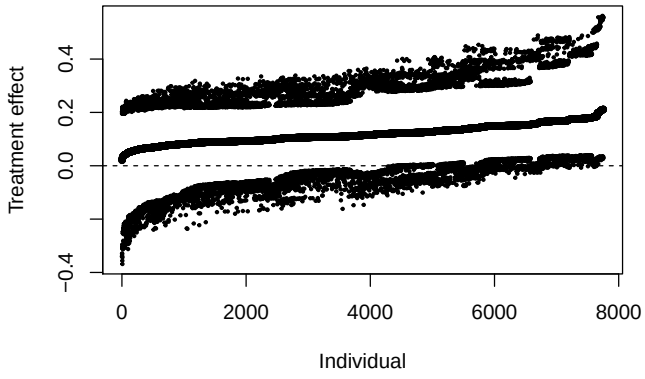

FIG 6. *A plot of all* $n = 7,752$ *estimated individual causal effects, rank ordered from lowest to highest posterior means. Each point estimate is sandwiched by corresponding upper and lower 95% credible endpoints. Observe that most individuals have an interval containing zero, shown by the dashed horizontal line.*

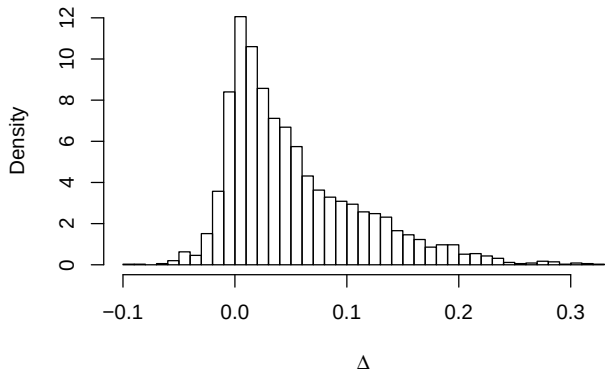

FIG 7. *A posterior histogram of the difference in subgroup average treatment effects between high school educated, married individuals with a high propensity to smoke and others. This subpopulation may be of special interest in future studies because the posterior suggests that the impact of smoking on their medical expenses is appreciably greater than that of other groups.*

effects it is worth emphasizing: strong ignorability does not imply that we have measured all the variables driving heterogeneity, merely that we have measured all those that are also involved in the selection process. While this analysis is far from definitive, it is promising that the BCF analysis has revealed an interesting subgroup to investigate in future studies.

7. Discussion. We have demonstrated the utility of our new model for the estimation of individual and subgroup causal effects. We conclude by reviewing the contributions made here and positioning them in the existing literature.

7.1. *Zellner priors for non- and semiparametric Bayesian causal inference.* In Section 4 we showed that the current gold standard in nonparametric Bayesian regression models for causal inference (BART) is susceptible to regression induced confounding as described by Hahn et al. (2016). The solution we propose is to include an estimate of the propensity score as a covariate in the outcome model. This induces a prior distribution that treats Z_i and $\hat{\pi}_i$ equitably, discouraging the outcome model from erroneously attributing the effect of confounders to the treatment variable. Here we justify and collect arguments in favor of this approach. We discuss an argument against, namely that it does not incorporate uncertainty in the propensity score, in a later subsection.

Conditioning on an estimate of the propensity score is readily justified: Because our regression model is conditional on Z and $\mathbf{X}$, it is perfectly legitimate to condition our prior on them as well.

This approach is widely used in linear regression, the most common example being Zellner's g-prior (Zellner, 1986) which parametrizes the prior covariance of a vector of regression coefficients in terms of a plug-in estimate of the predictor variables' covariance matrix. Nodding to this heritage, we refer to general predictor-dependent priors as Zellner priors.

In the Bayesian causal forest model, we specify a prior over f by applying an independent BART prior that includes $\hat{\pi}(\mathrm{x}_i)$ as one of its splitting dimensions. That is, because $\hat{\pi}(\mathrm{x}_i)$ is a fixed function of x_i, f is still a function $f : (\mathcal{X}, \mathcal{Y}) \mapsto \mathbb{R}$; the inclusion of $\hat{\pi}(\mathrm{x}_i)$ among the splitting dimensions does not materially change the support of the prior, but it does alter which functions are deemed more likely. Therefore, although writing $f(\mathrm{x}_i, z_i, \hat{\pi}(\mathrm{x}_i))$ is suggestive of how the prior is implemented in practice, we prefer notation such as

$$\begin{aligned} Y_i &= f(\mathrm{x}_i, z_i) + \epsilon_i, \quad \epsilon_i \overset{\text{iid}}{\sim} \mathrm{N}(0, \sigma^2), \\ f &\sim \mathrm{BART}(\mathbf{X}, Z, \hat{\pi}), \end{aligned} \tag{10}$$

where $\hat{\pi}$ is itself a function of $(\mathbf{X}, Z)$. Viewing BART as a prior in this way highlights the fact that various transformations of the data could be computed beforehand, prior to fitting the data with the default BART priors; the choice of transformations will control the nature of the regularization that is imposed. In conventional predictive modeling there is often no particular knowledge of which transformations of the covariates might be helpful. However, in the treatment effect context the propensity score is a natural and, in fact, critical choice.

7.2. *Why not use only the propensity score? vs. Why use the propensity score at all?.* It has long been recognized that regression on the propensity score is a useful dimension reduction tactic (Rosenbaum and Rubin, 1983). For the purpose of estimating average treatment effects, a regression model on the one-dimensional propensity score is sufficient for the task, allowing one to side-step estimating high dimensional nuisance parameters. In our notation, if π is assumed known, then one need only infer $f(\pi)$. That said, there are several reasons one should include the control vector x_i in its entirety (in addition to the propensity score).

The first reason is pragmatic: If one wants to identify heterogeneous effects, one needs to include any potential modulating variables anyway, precluding any dimension reduction at the outset.

Second, if we are to take a conventional Bayesian approach to inference and we do not in fact believe the response to depend on $\mathbf{X}$ strictly through the propensity score, we simply must include the covariates themselves and model the conditional distribution $p(Y \mid Z, \mathbf{X})$. The justification for

making inference about average treatment effects using regression or stratification on the propensity score alone is entirely frequentist; this approach is not without its merits, and we do not intend to argue frequency calibration is not desirable, but a fully Bayesian approach has its own appeal.

Third, if our propensity score model is inadequate (misspecified or otherwise poorly estimated), including the full predictor vector allows for the possibility that the response surface model remains correctly specified.

The converse question, *Why bother with the propensity score if one is doing a high dimensional regression anyway?*, has been answered in the main body of this paper. Incorporating the propensity score (or another balancing score) yields a prior that can more readily adapt to complex patterns of confounding. In fact, in the context of response surface modeling for causal effects, failing to include an estimate of the propensity score (or another balancing score) can lead to additional bias in treatment effect estimates, as shown by the simple, low-dimensional example in Section 4.

7.3. *Why not joint response-treatment modeling and what about uncertainty in the propensity score?.* Using a presumptive model for Z to obtain $\hat{\pi}$ invites the suggestion of fitting a joint model for (Y, Z). Indeed, this is the approach taken in Hahn et al. (2016) as well as earlier papers, including Rosenbaum and Rubin (1983), Robins, Mark and Newey (1992), McCandless, Gustafson and Austin (2009), Wang, Parmigiani and Dominici (2012), and Zigler and Dominici (2014). While this approach is certainly reasonable, the Zellner prior approach would seem to afford all the same benefits while avoiding the distorted inferences that would result from a joint model if the propensity score model is misspecified (Zigler and Dominici, 2014).

One might argue that our Zellner prior approach gives under-dispersed posterior inference in the sense that it fails to account for the fact that $\hat{\pi}$ is simply a point estimate (and perhaps a bad one). However, this objection is somewhat misguided. First, as discussed elsewhere (e.g. Hill (2011)), inference on individual or subgroup treatment effects follows directly from the conditional distribution $(Y \mid Z, \mathbf{X})$. To continue our analogy with the more familiar Zellner g-prior, to model $(Y \mid Z, \mathbf{X})$ we are no more obligated to consider uncertainty in $\hat{\pi}$ than we are to consider uncertainty in $(\mathbf{X}'\mathbf{X})^{-1}$ when using a g-prior for on the coefficients of a linear model. Second, $\hat{\pi}$ appears in the model *along with* the full predictor vector x: it is provided as a hint, not as a certainty, and this model is at least as capable of estimating a complex response surface as the corresponding model without $\hat{\pi}$, and the cost incurred by the addition of a one additional "covariate" can be more than offset by the bias reduction in the estimation of treatment effects.

On the other hand, we readily acknwoledge that one might be interested in what inferences would obtain if we used different $\hat{\pi}$ estimates. One might consider fitting a series of BCF models with different estimates of $\hat{\pi}$, perhaps from alternative models or other procedures. This is a natural form of sensitivity analysis in light of the fact that the adjustments proposed in this paper only work if $\hat{\pi}$ accurately approximates π. However, it is worth noting that the available (z, x) data speak to this question: a host of empirically proven prediction methods (i.e. neural networks, support vector machines, random forests, boosting, or any ensemble method) can be used to construct candidate $\hat{\pi}$ and cross-validation may be used to gauge their accuracy. Only if a "tie" in generalization error (predicting Z) is encountered must one turn to sensitivity analysis.

7.4. *Related non-Bayesian work.* The Bayesian causal forest model is a flexible semi-parametric prediction model for estimating causal effects. Other recent work also occupies this intersection between "machine learning" and causal inference, each with a somewhat different focus. Targeted maximum likelihood estimation (TMLE) (van der Laan, 2010a,b), double machine learning (Chernozhukov et al., 2016), and generalized boosting (McCaffrey, Ridgeway and Morral, 2004; McCaffrey et al., 2013) all focus on estimation of average treatment effects, whereas our focus is on individual (heterogeneous, subgroup) treatment effects. Like us, Taddy et al. (2016) focuses on estimating heterogeneous effects, but they analyze data from experiments, whereas our data are observational. As we have discussed, this has significant implications for how we should specify prior distributions. Other related contributions include Su et al. (2012) and Lu et al. (2017). Finally, Wager and Athey (2015) prove theoretical results for random forest-based estimation of heterogeneous effects from observational data. However, Wager and Athey (2015) offer little practical guidance on how to deploy their method in practice, in particular on how much regularization to impose. Although it was not our focus here, we find that in finite samples their method exhibits regularization induced bias similar to what we report for BART. Given this landscape, we believe the Bayesian causal forest model presented in this paper represents a beneficial new tool for causal inference from observational data, especially when confounding is suspected to be strong and the general magnitude of treatment effects is thought to be relatively modest.

7.5. *Extensions.* Using ideas from Imai and Van Dyk (2004), the Bayesian causal forest model can be readily extended to the case of continuous treatment. Likewise, using ideas from Albert and Chib (1993), the Bayesian causal forest model can be extended to the case of a binary response

variable. Incorporating variable selection priors, as in Linero (2016), would further refine the regularization of the causal forest model to accommodate many spurious controls. Using ideas from Murray (2017), the causal forest model can perhaps be extended to cases with non-additive errors in the response model. These are all areas of active development.

References.

Albert, J. H. and Chib, S. (1993). Bayesian Analysis of Binary and Polychotomous Response Data. *Journal of the American Statistical Association* **88** 669-679.

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C. et al. (2016). Double machine learning for treatment and causal parameters. *arXiv preprint arXiv:1608.00060*.

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

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

Diaconis, P. and Zabell, S. L. (1982). Updating subjective probability. *Journal of the American Statistical Association* **77** 822–830.

Gelman, A. et al. (2006). Prior distributions for variance parameters in hierarchical models (comment on article by Browne and Draper). *Bayesian analysis* **1** 515–534.

Giles, D. and Rayner, A. (1979). The mean squared errors of the maximum likelihood and natural-conjugate Bayes regression estimators. *Journal of Econometrics* **11** 319–334.

Gramacy, R. B. and Lee, H. K. (2008). Bayesian treed Gaussian process models with an application to computer modeling. *Journal of the American Statistical Association* **103**.

Green, D. P. and Kern, H. L. (2012). Modeling heterogeneous treatment effects in survey experiments with Bayesian additive regression trees. *Public opinion quarterly* nfs036.

Gustafson, P. and Greenland, S. (2006). Curious phenomena in Bayesian adjustment for exposure misclassification. *Statistics in medicine* **25** 87–103.

Hahn, P. R., Puelz, D., He, J. and Carvalho, C. M. (2016). Regularization and confounding in linear regression for treatment effect estimation. *Bayesian Analysis*.

Heckman, J. J., Lopes, H. F. and Piatek, R. (2014). Treatment effects: A Bayesian perspective. *Econometric reviews* **33** 36–67.

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

Hill, J., Su, Y.-S. et al. (2013). Assessing lack of common support in causal inference using Bayesian nonparametrics: Implications for evaluating the effect of breastfeeding on children's cognitive outcomes. *The Annals of Applied Statistics* **7** 1386–1420.

Hill, J., Cervone, D., Dorie, V., Scott, M. and Shalit, U. (2017). Causal inference data analysis challenge Technical Report, New York University.

Imai, K. and Van Dyk, D. A. (2004). Causal inference with general treatment regimes: Generalizing the propensity score. *Journal of the American Statistical Association* **99** 854–866.

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

Johnson, E., Dominici, F., Griswold, M. and Zeger, S. L. (2003). Disease cases and their medical costs attributable to smoking: an analysis of the national medical expenditure survey. *Journal of Econometrics* **112** 135–151.

Kern, H. L., Stuart, E. A., Hill, J. and Green, D. P. (2016). Assessing methods for generalizing experimental impact estimates to target populations. *Journal of Research on Educational Effectiveness* **9** 103–127.

Li, M. and Tobias, J. L. (2014). Bayesian analysis of treatment effect models. *Bayesian inference in the social sciences* 63–90.

Linero, A. R. (2016). Bayesian Regression Trees for High Dimensional Prediction and Variable Selection. *Journal of the American Statistical Association* **just-accepted**.

Lu, M., Sadiq, S., Feaster, D. J. and Ishwaran, H. (2017). Estimating Individual Treatment Effect in Observational Data Using Random Forest Methods. *arXiv preprint arXiv:1701.05306*.

McCaffrey, D. F., Ridgeway, G. and Morral, A. R. (2004). Propensity score estimation with boosted regression for evaluating causal effects in observational studies. *Psychological methods* **9** 403.

McCaffrey, D. F., Griffin, B. A., Almirall, D., Slaughter, M. E., Ramchand, R. and Burgette, L. F. (2013). A tutorial on propensity score estimation for multiple treatments using generalized boosted models. *Statistics in medicine* **32** 3388–3414.

McCandless, L. C., Gustafson, P. and Austin, P. C. (2009). Bayesian propensity score analysis for observational data. *Statistics in Medicine* **28** 94–112.

Murray, J. S. (2017). Log-linear Bayesian additive regression trees for categorical and count responses Technical Report, Carnegie Mellon University.

Robins, J. M., Mark, S. D. and Newey, W. K. (1992). Estimating exposure effects by modelling the expectation of exposure conditional on confounders. *Biometrics* 479–495.

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

Sivaganesan, S., Müller, P. and Huang, B. (2017). Subgroup finding via Bayesian additive regression trees. *Statistics in Medicine*.

Su, X., Kang, J., Fan, J., Levine, R. A. and Yan, X. (2012). Facilitating score and causal inference trees for large observational studies. *Journal of Machine Learning Research* **13** 2955–2994.

Taddy, M., Gardner, M., Chen, L. and Draper, D. (2016). A Nonparametric Bayesian Analysis of Heterogenous Treatment Effects in Digital Experimentation. *Journal of Business & Economic Statistics* **34** 661–672.

van der Laan, M. J. (2010a). Targeted maximum likelihood based causal inference: Part I. *The International Journal of Biostatistics* **6**.

van der Laan, M. J. (2010b). Targeted maximum likelihood based causal inference: Part II. *The International Journal of Biostatistics* **6**.

Wager, S. and Athey, S. (2015). Estimation and Inference of Heterogeneous Treatment Effects using Random Forests. *arXiv preprint arXiv:1510.04342*.

Wang, C., Parmigiani, G. and Dominici, F. (2012). Bayesian effect estimation accounting for adjustment uncertainty. *Biometrics* **68** 661–671.

Wendling, T., Jung, K., Shah, N. and Luxan, B. G. (2017). Comparison of methods for prediction of individual treatment effects using observational data from electronic health records. *Statistics in Medicine*.

Zeger, S. L., Wyant, T., Miller, L. S. and Samet, J. (2000). Statistical testimony on damages in Minnesota v. Tobacco Industry. In *Statistical science in the courtroom* 303–320. Springer.

Zellner, A. (1986). On assessing prior distributions and Bayesian regression analysis with g-prior distributions. *Bayesian inference and decision techniques: Essays in Honor of Bruno De Finetti* **6** 233–243.

Zigler, C. M. and Dominici, F. (2014). Uncertainty in propensity score estimation: Bayesian methods for variable selection and model-averaged causal effects. *Journal of the American Statistical Association* **109** 95–107.

APPENDIX A

Here we derive an expression for the bias of the treatment effect estimate when the response and treatment model are both linear

$$\begin{aligned} Y_i &= \alpha Z_i + \beta^t \mathrm{x}_i + \varepsilon_i, \\ Z_i &= \gamma^t \mathrm{x}_i + \nu_i, \end{aligned} \tag{11}$$

where the error terms are mean zero Gaussian, and a Gaussian (ridge) prior is placed over all regression coefficients. The Bayes estimator under squared error loss is the posterior mean, so we examine the expression for the bias of $\hat{\alpha}_{rr} \equiv \mathrm{E}(\alpha \mid Y, \mathrm{z}, \mathbf{X})$. We begin from a standard expression for the ridge estimator, as given, for example, in Giles and Rayner (1979). Write $\theta = (\alpha, \beta^t)^t$,

$$\tilde{\mathbf{X}} = \begin{pmatrix} \mathrm{z} & \mathbf{X} \end{pmatrix}$$

and let $\theta \sim \mathrm{N}(0, \mathbf{M}^{-1})$, then the bias of the Bayes estimator is

$$\mathrm{bias}(\hat{\theta}_{rr}) = -(\mathbf{M} + \tilde{\mathbf{X}}^t \tilde{\mathbf{X}})^{-1} \mathbf{M} \theta \tag{12}$$

where the bias expectation is taken over Y, conditional $\mathbf{X}$ and all model parameters. Next, consider $M = \begin{pmatrix} 0 & 0 \\ 0 & \mathrm{I}_p \end{pmatrix}$, where I_p denotes a p-by-p identity matrix, which corresponds to a ridge prior (with ridge parameter $\lambda = 1$ for simplicity) on the control variables and a non-informative "flat" prior over the first element, corresponding to the treatment effect. Plugging this into the above expression and applying the block-inverse formula to matrix

$$(\mathbf{M} + \tilde{\mathbf{X}}^t \tilde{\mathbf{X}})^{-1} = \begin{pmatrix} \mathrm{z}^t \mathrm{z} & \mathrm{z}^t \mathbf{X} \\ \mathbf{X}^t \mathrm{z} & \mathbf{X}^t \mathbf{X} + \mathbf{I}_p \end{pmatrix}^{-1}$$

yields

$$\text{bias}(\hat{\alpha}_{rr}) = -\left((\mathrm{z}^t\mathrm{z})^{-1}\mathrm{z}^t\mathbf{X}\right)(\mathbf{I} + \mathbf{X}^t(\mathbf{X} - \hat{\mathbf{X}}_\mathrm{z}))^{-1}\beta. \tag{13}$$

where $\hat{\mathbf{X}}_\mathrm{z} = \mathrm{z}(\mathrm{z}^t\mathrm{z})^{-1}\mathrm{z}^t\mathbf{X}$. Notice that the leading term is a vector of regression coefficients from p univariate regression predicting X_j given z; in the presence of selection these will generally be non-zero. The middle matrix will likewise generally be non-zero, with the result that the bias on the treatment effect parameter will be a function of the unknown regression coefficients, β.

Next, we observe how including a propensity function estimate into our regularized regression can mitigate this bias. Denoting our propensity function estimate, $\hat{z}_i \approx \gamma^t\mathrm{x}_i$, we consider the redundantly parametrized regression that includes as regressors both Z and $\hat{\mathrm{z}}$; in other words we define

$$\tilde{\mathbf{X}} = \begin{pmatrix} \mathrm{z} & \hat{\mathrm{z}} & \mathbf{X} \end{pmatrix}.$$

Following the above derivation, putting a flat prior over the coefficient associated with $\hat{\mathrm{z}}$ yields bias

$$\text{bias}(\hat{\alpha}_{rr}) = -\left\{(\tilde{\mathrm{z}}^t\tilde{\mathrm{z}})^{-1}\tilde{\mathrm{z}}^t\mathbf{X}\right\}_1(\mathbf{I} + \mathbf{X}^t(\mathbf{X} - \hat{\mathbf{X}}_\mathrm{z}))^{-1}\beta \approx 0.$$

where $\tilde{\mathrm{z}} = (\mathrm{z} \;\; \hat{\mathrm{z}})$ and $\left\{(\tilde{\mathrm{z}}^t\tilde{\mathrm{z}})^{-1}\tilde{\mathrm{z}}^t\mathbf{X}\right\}_1$ denotes the top row of $\left\{(\tilde{\mathrm{z}}^t\tilde{\mathrm{z}})^{-1}\tilde{\mathrm{z}}^t\mathbf{X}\right\}$, which corresponds to the regression coefficient associated with Z in the two variable regression predicting X_j given $\tilde{\mathrm{z}}$. Finally, note that these regression coefficients will be approximately zero because $\hat{z}$ captures the association between Z and X, rendering Z itself conditionally independent, given $\hat{z}$.