Bayesian Causal Forests
for estimating personalized treatment effects

P. Richard Hahn, Jared Murray, and Carlos Carvalho
December 8, 2016
Overview

We consider

- observational data,
- assuming conditional unconfoundedness,
- and are interested in individual treatment effects.

Three key statistical concepts will be: prediction, regularization, and selection.

Our regression model will take the novel form

\[ D_i = g(x_i) + \epsilon_i, \]
\[ Y_i = m(x_i) + \sum_{j=1}^{k} b_j(x)D_i^j + \nu_i. \]

where \( g, m \) and \( b \) are represented as regression forests.
Accurate prediction is necessary but not sufficient

Higher predictive accuracy is associated with more reliable information about the underlying data mechanism. Weak predictive accuracy can lead to questionable conclusions. Algorithmic models can give better predictive accuracy than data models, and provide better information about the underlying mechanism.

Leo Breiman

Prediction by itself is only occasionally sufficient. The post office is happy with any method that predicts correct addresses from hand-written scrawls...[But] most statistical surveys have the identification of causal factors as their ultimate goal.

Bradley Efron
Flexible models are necessary

Figure 7.3: A linear model is not useful in this nonlinear case. These data are from an introductory physics experiment.

Taken from OpenIntro Statistics 2nd Edition.
Flexible models aren’t sufficient

An expensive car doesn’t cause good teeth.
Suppose we’re interested in the treatment effect of dietary kale intake.

We want to know how effective it is at lowering cholesterol, which is our outcome variable.

Unfortunately, we have only observational data (i.e., not a randomized study).
Kale intake predicts exercise

Only gym-rats seem to eat much kale and exercise is known to lower cholesterol.

\[ Y_i = \beta_0 + \alpha D_i + \varepsilon_i, \]

Because \( \text{cov}(D_i, \varepsilon_i) \neq 0 \), we can write

\[ Y_i = \beta_0 + \alpha D_i + \omega D_i + \tilde{\varepsilon}. \]

Since \( \text{cov}(D_i, \tilde{\varepsilon}_i) = 0 \), we mis-estimate \( \alpha \) as \( \alpha + \omega \).

But if we include exercise in the model,

\[ Y_i = \beta_0 + \alpha D_i + \beta x_i + \varepsilon_i. \]

Conditional on \( x_i \), \( \text{cov}(D_i, \varepsilon_i) = 0 \).
Randomized experiments are the special case where prediction and causation are equivalent.

Holding $x$ constant, one has experimental variation from which to infer causal impacts. (We have to assume $Y$ doesn’t cause $D$.)
Regularization

\[ Y_i = \beta_0 + \alpha D_i + \beta x_i + \varepsilon_i. \]

- a flat prior on the treatment effect: \((\alpha, \sigma^2_\varepsilon) \propto 1/\sigma_\varepsilon,\)

- shrinkage prior on \(\beta\) (e.g., the horseshoe prior).

- Now “turn the Bayesian crank”...
It turns out that this “obvious” approach is really bad at getting reasonable estimates of the treatment effect $\alpha$. 
The bias of the treatment parameter under ridge regression is:

\[
\text{bias}(\hat{\alpha}_{rr}) = - \left( (D^tD)^{-1}D^tX \right) (I_p + X^t(X - \hat{X}_D))^{-1} \beta.
\]

This is a function of all the other unknown regression parameters.
Our reparametrization: a latent error approach

We reparametrize as

\[
\begin{pmatrix}
\alpha \\
\beta + \alpha \gamma \\
\gamma
\end{pmatrix}
\rightarrow
\begin{pmatrix}
\alpha \\
\beta_d \\
\beta_c
\end{pmatrix}.
\]

which gives the new equations

Selection Eq.: \( D = x^t \beta_c + \epsilon, \quad \epsilon \sim N(0, \sigma^2_\epsilon) \),

Response Eq.: \( Y = \alpha(D - x^t \beta_c) + x^t \beta_d + \nu, \quad \nu \sim N(0, \sigma^2_\nu) \).

We may now shrink \( \beta_d \) and \( \beta_c \) with impunity.
The expression for ridge treatment effect bias is now:

$$\text{bias}(\hat{\alpha}) = - \left((R^t R)^{-1} R^t X\right) \left(I_p + X^t (X - \hat{X}_R)\right)^{-1} \beta_d.$$ 

where $R = D - x^t \beta_c$. By construction, $(R^t R)^{-1} R^t X$ will be close to the zero vector, because $R$ is independent of $x$. 
<table>
<thead>
<tr>
<th>$\rho^2$</th>
<th>Bias</th>
<th>Coverage</th>
<th>I.L.</th>
<th>MSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.1</td>
<td>New Approach</td>
<td>0.0021</td>
<td>0.919</td>
<td>0.6073</td>
</tr>
<tr>
<td></td>
<td>OLS</td>
<td>-0.0119</td>
<td>0.93</td>
<td>0.8888</td>
</tr>
<tr>
<td></td>
<td>Naive Regularization</td>
<td>-0.0291</td>
<td>0.443</td>
<td>0.2207</td>
</tr>
<tr>
<td></td>
<td>Oracle OLS</td>
<td>-0.0019</td>
<td>0.952</td>
<td>0.6138</td>
</tr>
<tr>
<td>0.3</td>
<td>New Approach</td>
<td>-0.0033</td>
<td>0.909</td>
<td>0.6918</td>
</tr>
<tr>
<td></td>
<td>OLS</td>
<td>-0.0038</td>
<td>0.944</td>
<td>1.0294</td>
</tr>
<tr>
<td></td>
<td>Naive Regularization</td>
<td>-0.0651</td>
<td>0.402</td>
<td>0.237</td>
</tr>
<tr>
<td></td>
<td>Oracle OLS</td>
<td>-0.0022</td>
<td>0.952</td>
<td>0.696</td>
</tr>
<tr>
<td>0.5</td>
<td>New Approach</td>
<td>-0.011</td>
<td>0.894</td>
<td>0.8191</td>
</tr>
<tr>
<td></td>
<td>OLS</td>
<td>0.0071</td>
<td>0.927</td>
<td>1.2041</td>
</tr>
<tr>
<td></td>
<td>Naive Regularization</td>
<td>-0.1354</td>
<td>0.349</td>
<td>0.233</td>
</tr>
<tr>
<td></td>
<td>Oracle OLS</td>
<td>-0.0026</td>
<td>0.952</td>
<td>0.8235</td>
</tr>
<tr>
<td>0.7</td>
<td>New Approach</td>
<td>-0.028</td>
<td>0.904</td>
<td>1.0842</td>
</tr>
<tr>
<td></td>
<td>OLS</td>
<td>-8e-04</td>
<td>0.938</td>
<td>1.5533</td>
</tr>
<tr>
<td></td>
<td>Naive Regularization</td>
<td>-0.2752</td>
<td>0.217</td>
<td>0.2474</td>
</tr>
<tr>
<td></td>
<td>Oracle OLS</td>
<td>-0.0033</td>
<td>0.952</td>
<td>1.0632</td>
</tr>
<tr>
<td>0.9</td>
<td>New Approach</td>
<td>-0.1078</td>
<td>0.948</td>
<td>1.8128</td>
</tr>
<tr>
<td></td>
<td>OLS</td>
<td>0.0291</td>
<td>0.942</td>
<td>2.6708</td>
</tr>
<tr>
<td></td>
<td>Naive Regularization</td>
<td>-0.6045</td>
<td>0.015</td>
<td>0.2576</td>
</tr>
<tr>
<td></td>
<td>Oracle OLS</td>
<td>-0.0058</td>
<td>0.952</td>
<td>1.8415</td>
</tr>
</tbody>
</table>

**Table 1:** $n = 50$, $p = 30$, $k = 3$. $\kappa^2 = 0.05$. $\phi^2 = 0.05$. $\sigma^2_{\nu} = 0.9$. 
Consider the nonlinear model

\[ Y_i = f(x_i, D_i) + \varepsilon_i. \]

Now our de-confoundedness condition is stronger

\[ D_i \perp \perp \varepsilon_i \mid x_i. \]

And our causal estimand is more general

\[
\begin{align*}
\alpha(x_i, D_i, D'_i) &= \mathbb{E}(Y_i \mid x_i, D'_i) - \mathbb{E}(Y_i \mid x_i, D_i), \\
&= f(x_i, D'_i) - f(x_i, D_i).
\end{align*}
\]
Regression Trees

Tree $T_h$

Leaf/End node parameters
$$M_h = (\mu_{h1}, \mu_{h2}, \mu_{h3})$$

Partition $A_h = \{A_{h1}, A_{h2}, A_{h3}\}$

$$g(x, T_h, M_h) = \mu_{ht} \text{ if } x \in A_{ht} \text{ (for } 1 \leq t \leq b_h).$$
Bayesian additive regression trees (BART) (Chipman, George and McCulloch, 2010):

\[
y_i = f(x_i) + \epsilon_i, \quad \epsilon_i \sim N(0, \sigma^2)
\]

\[
f(x) = \sum_{h=1}^{m} g(x, T_h, M_h)
\]

\(g\) are basis functions determined by a binary tree \(T_h\) and vector of parameters \(M_h\).

These models were specifically advocated for causal inference in Hill (2012).
BART exhibits regularization-induced confounding

Figure 1: Left: BART predictions. Middle: BCF predictions. Right: treatment effect estimates; BART (blue) and BCF (orange).
Taylor expanding $f(x, D)$

We expand $f(x, D)$ about the point $D = E(D \mid x) = g(x)$:

$$f(x, D) = \sum_{j=0}^{\infty} \frac{f^{(j)}(x, g(x))(D - g(x))^j}{j!}$$

where $f^{(j)}$ denotes the $j$th derivative of $f$ with respect to $D$, and reparametrize as

$$f(x, D) = f(x, g(x)) + \sum_{j=1}^{k} b_j(x, g(x))(D - g(x))^j$$

$$= m(x, g(x)) + \sum_{j=1}^{k} b_j(x, g(x))(D - g(x))^j.$$
Taylor Shifts

One can “transform back” via

\[ \tilde{b}_j(x) = \sum_{j \leq h \leq k} \binom{h}{j} b_h(x, g(x))g(x)^{h-j}. \]

for \( 0 \leq j \leq k \) and \( b_0(x) := m(x, g(x)) \).

In this parametrization

\[ \alpha(x_i, D_i, D'_i) = f(x_i, D'_i) - f(x_i, D_i), \]

\[ = \sum_{j=0}^{k} \tilde{b}_j(x)D'_j - \sum_{j=0}^{k} \tilde{b}_j(x)D^j, \]

\[ = \sum_{j=0}^{k} \tilde{b}_j(x) (D'_j - D^j). \]

This suggests \( \sum_{j=0}^{k} \| \tilde{b}_j(x) \| \) is a nice summary of the causal effect at \( x \).
Bayesian causal forests

Take $k < \infty$, give $g$, $m$ and $b$ independent BART priors and assume normal additive errors:

$$D_i = g(x_i) + \epsilon_i,$$
$$Y_i = m(x_i, g(x_i)) + \sum_{j=1}^{k} b_j(x, g(x))(D_i - g(x_i))^j + \nu_i.$$ 

This model allows for the various roles of control variables to be regularized independently:

- $g(x)$ represents **selection**, 
- $m(x, g(x))$ represents **direct association**, and 
- $b_j(x, g(x))$ represents **nonlinearity** as it varies over $j$, 
- $b_j(x, g(x))$ represents **heterogeneous effects** as it varies in $x$.

Confusion of these roles is **confounding**.
Advantages of BCF

The analogies with the linear case are clear,

\[ D_i = g(x_i) + \epsilon_i, \]
\[ Y_i = m(x_i, g(x_i)) + \sum_{j=1}^{k} b_j(x, g(x_i))(D_i - g(x_i))^j + \nu_i. \]

Advantages are that

- Independent regularization of \( g \), \( m \), and \( b \).
- easy summarization of treatment effects,
- transparent interpolation/extrapolation.

By comparison, BART’s regularization is **implicit**, treatment effects are difficult to extract, and interpolation/extrapolation is poor.
1987 National Medical Expenditure Survey (NMES)

What is the effect of smoking on medical expenditures?

- outcome variable $Y$ is medical expenses (verified, log transformed),
- treatment variable $D$ is cigarettes smoked per day,
- $n = 7.7k$ complete-case analysis for $Y > 0$,
- covariates include age at time of the survey (19-94), age started smoking, gender (male, female), race (hispanic, black, other), marriage status (married, widowed, divorced, separated, never married), education level (college graduate, some college, high school graduate, other), census region (Northeast, Midwest, South, West), poverty status (poor, near poor, low income, middle income, high income), and seat belt usage (rarely, sometimes, always/almost always)
Figure 2: The average treatment effect is statistically significant, but small.
Figure 3: Individual level treatment effects are difficult to estimate with any precision.
Fitting a **single** regression tree to **posterior summaries** of \( g, m, \) and 
\[
\sum_{j=0}^{k} \| \tilde{b}_j(x) \|
\] allows us to find interesting sub-groups for follow-up study.

- Save posterior samples of each subject-specific parameter,
- compute \( \hat{g}, \hat{m}, \) and \( \tilde{b}_j, \)
- fit single regression trees (with little or no regularization) with these summaries as the outcome,
- Leaf nodes define subgroups for further examination (such as subgroup average treatment effects).
Figure 4: Age, gender and race appear to be the dominant determinants of treatment/smoking level.
Tree fit to $\pi(m_i)$

Figure 5: Age appears to be the dominant driver of medical expenditures.
Tree fit to $\pi(b_i)$

Figure 6: Race and educational level appear to modulate the treatment effect. Less educated whites seem to have the largest treatment effect, but it is...
What about BART?

BART and BCF give similar predictions.

Same predictions, different conclusions.
Summary

- Regularization-induced confounding can adversely bias treatment effect estimates.
- Explicitly modeling de-confoundedness allows regularization to be imposed robustly.
- The intuitive parametrization makes posterior exploration quite convenient.
- Bayesian causal forests permit finer control of regularization.
- Easily modified to handle categorical outcomes and treatments.
- Future work considers non-normal errors.

Thank you for your time.
Some related work


Our work differs from these works in its explicit focus on regularization and its novel forested-linear-model structure.