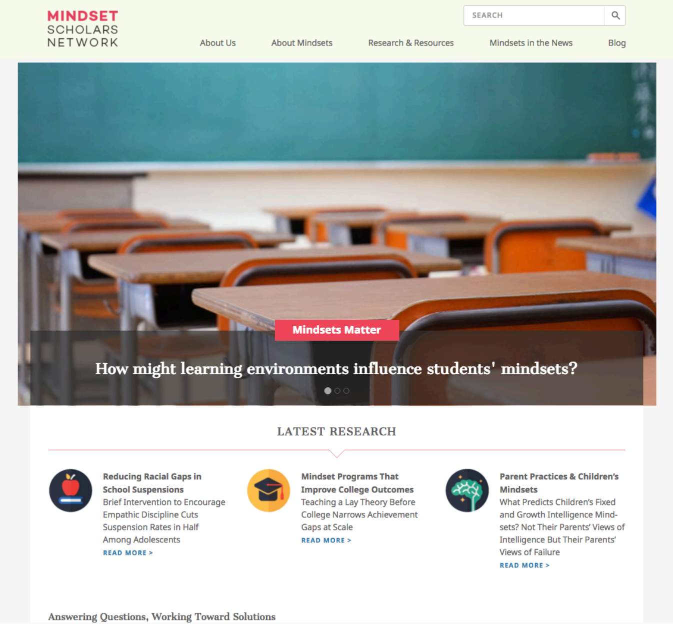
Recent developments in model specification, regularization, and summarization for nonparametric Bayesian models of heterogeneous treatment effects

Jared S. Murray – The University of Texas at Austin.
Joint work with Carlos Carvalho, P. Richard Hahn, David Yeager, et al.

June, 2019

National Study of Learning Mindsets



National Study of Learning Mindsets

- National Study of Learning Mindsets (Yeager et. al., 2017): Randomized controlled trial of a low-cost mindset intervention
- Probability sample, 65 schools in this analysis ($> 11,000$ 9th grade students)
- Specifically designed to assess treatment effect heterogeneity

National Study of Learning Mindsets

What do we hope to gain with BNP?

- Avoid (explicit) model selection/specification search
- Flexible models of treatment effect heterogeneity
- Appropriate measures of uncertainty

Where does BNP need a little help?

- Summarizing complex posteriors to communicate results

Our (generic) assumptions

Strong ignorability:

$$Y_i(0), Y_i(1) \perp\!\!\!\perp Z_i \mid \mathsf{X}_i = \mathsf{x}_i,$$

Positivity:

$$0 < \Pr(Z_i = 1 \mid \mathsf{X}_i = \mathsf{x}_i) < 1$$

for all i. Then

$$P(Y(z) \mid \mathsf{x}) = P(Y \mid Z = z, \mathsf{x})$$

,

and the conditional average treatment effect (CATE) is

$$\begin{aligned}\tau(\mathsf{x}_i) :&= \mathrm{E}(Y_i(1) - Y_i(0) \mid \mathsf{x}_i) \\ &= \mathrm{E}(Y_i \mid \mathsf{x}_i, Z_i = 1) - \mathrm{E}(Y_i \mid \mathsf{x}_i, Z_i = 0).\end{aligned}$$

Parameterizing Nonparametric Models of Causal Effects

Let's forget confounding and covariates for a second.

A simple model:

$$
\begin{aligned}
(Y_i \mid Z_i = 0) &\overset{iid}{\sim} \mathrm{N}(\mu_0, \sigma^2) \\
(Y_i \mid Z_i = 1) &\overset{iid}{\sim} \mathrm{N}(\mu_1, \sigma^2)
\end{aligned}
$$

where the estimand of interest is $\tau \equiv \mu_1 - \mu_0$.

If $\mu_j \sim N(\phi_j, \delta_j)$ independently then $\tau \sim N(\phi_1 - \phi_0, \delta_0 + \delta_1)$

Often we have stronger prior information about τ than μ_1 or μ_0 – in particular, we expect it to be small.

Parameterizing Nonparametric Models of Causal Effects

A more natural parameterization:

$$
\begin{aligned}
(Y_i \mid Z_i = 0) &\overset{iid}{\sim} \mathrm{N}(\mu, \sigma^2) \\
(Y_i \mid Z_i = 1) &\overset{iid}{\sim} \mathrm{N}(\mu + \tau, \sigma^2)
\end{aligned}
$$

where the estimand of interest is still τ.

Now we can express prior beliefs on τ directly.

Parameterizing Nonparametric Models of Causal Effects

How does this relate to models for heterogeneous treatment effects? Consider (mostly) separate models for treatment arms:

$$y_i = f_{z_i}(\mathrm{x}_i) + \epsilon_i \quad \epsilon_i \sim N(0, \sigma^2)$$

$$(Y_i \mid Z_i = 0, \mathrm{x}_i) \overset{iid}{\sim} \mathrm{N}(f_0(\mathrm{x}), \sigma^2)$$
$$(Y_i \mid Z_i = 1, \mathrm{x}_i) \overset{iid}{\sim} \mathrm{N}(f_1(\mathrm{x}), \sigma^2)$$

Independent priors on $f_0, f_1 \to$ prior on $\tau(\mathrm{x}) \equiv f_1(\mathrm{x}) - f_0(\mathrm{x})$ has larger variance than prior on f_0 or f_1

No direct prior control – simple f_0, f_1 can compose to complex τ.

Every variable in x is a potential effect modifier.

Parameterizing Nonparametric Models of Causal Effects

What about the “just another covariate” parameterization?

$$y_i = f(\mathrm{x}_i, z_i) + \epsilon_i \quad \epsilon_i \sim N(0, \sigma^2)$$
$$(Y_i \mid Z_i = z_i, \mathrm{x}_i) \overset{iid}{\sim} N(f(\mathrm{x}_i, z_i), \sigma^2)$$

Then the heterogeneous treatment effects given by

$$\tau(\mathrm{x}) \equiv f(\mathrm{x}, 1) - f(\mathrm{x}, 0)$$

and every variable in x is still a potential effect modifier, and we have no direct prior control...

Parameterizing Nonparametric Models of Causal Effects

Set $f(\mathbf{x}_i, z_i) = \mu(\mathbf{x}_i) + \tau(\mathbf{w}_i)z_i$, where $\mathbf{w}$ is a subset of $\mathbf{x}$:

$$y_i = \mu(\mathbf{x}_i) + \tau(\mathbf{w}_i)z_i + \epsilon_i, \quad \epsilon_i \sim N(0, \sigma^2)$$

$$(Y_i \mid Z_i = z_i) \overset{iid}{\sim} N(\mu(\mathbf{x}_i) + \tau(\mathbf{w}_i)z_i, \sigma^2)$$

The heterogeneous treatment effects are given by $\tau(\mathbf{w})$ directly

In Hahn et. al. (2017) we use independent BART priors on μ and τ ("Bayesian causal forests")

Tweaking priors for causal effects

Several adjustments to the BART prior on τ:

- Higher probability on smaller τ trees (than BART defaults)
- Higher probability on "stumps", trees that never split (all stumps = homogeneous effects)
- $N^+(0, v)$ Hyperprior on the scale of leaf parameters in τ

What about observational data?

What changes when adding confounding?

Not much, but we should include an estimated propensity score as a covariate:

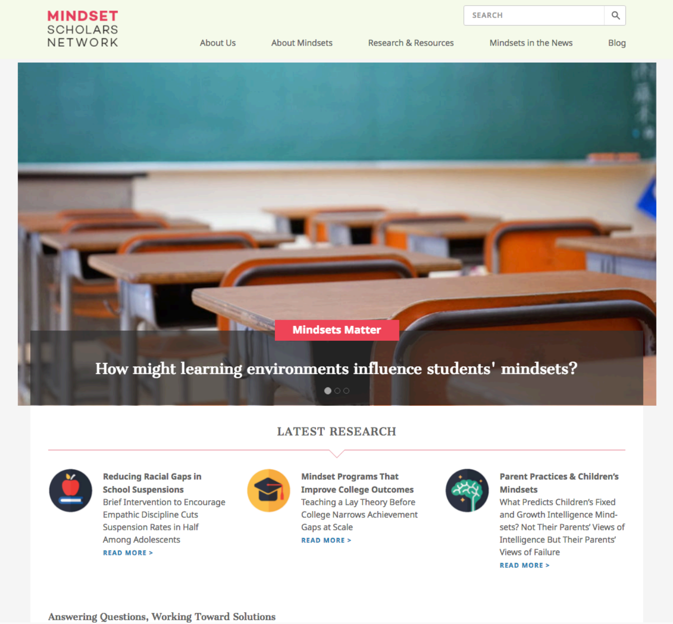
$$y_i = \mu(\mathrm{x}_i, \hat{\pi}(\mathrm{x}_i)) + \tau(\mathrm{w}_i)z_i + \epsilon_i, \quad \epsilon_i \sim N(0, \sigma^2)$$

$$(Y_i \mid Z = z_i) \overset{iid}{\sim} N(\mu(\mathrm{x}_i, \hat{\pi}(\mathrm{x}_i)) + \tau(\mathrm{w}_i)z_i, \sigma^2)$$

Mitigates regularization induced confounding (RIC); see Hahn et. al. (2017) for details.

Lots of independent empirical evidence this is important (cf Dorie et al (2019), Wendling et al (2019)).

The National Study of Learning Mindsets



Analysis of National Study of Learning Mindsets

- A new analaysis of effects on math GPA in the overall population of students
- Interesting moderators are baseline level of mindset norms, school achievement, and minority composition
- Many, many other controls

Multilevel BCF

Mindset study has students nested within schools:

$$y_{ij} = \alpha_j + \mu(\mathbf{x}_{ij}) + \left[\phi_j + \tau(\mathbf{w}_{ij})\right] z_{ij} + \epsilon_{ij}, \quad \epsilon_{ij} \sim N(0, \sigma^2)$$

α_j, ϕ_j have standard random effect priors (normal with half-*t* hyperpriors on scale)

$\mathbf{w} =$ school achievement, baseline mindset norms, minority composition

We fit this model, now what?

How do we present our results? Average treatment effect (ATE), school ATE, plots of (school) ATE by x_j....

Two questions from our collaborators:

- For which groups was the treatment most/least effective?
- What is the impact of posited treatment effect modifiers, holding other modifiers constant?

We fit this model, now what?

How do we present our results? Average treatment effect (ATE), school ATE, plots of (school) ATE by x_j....

Two questions from our collaborators:

- For which groups was the treatment most/least effective?
- What is the impact of posited treatment effect modifiers, holding other modifiers constant?

Subgroup finding as a decision problem

The action γ is choosing subgroups, here represented by a recursive partition (tree)

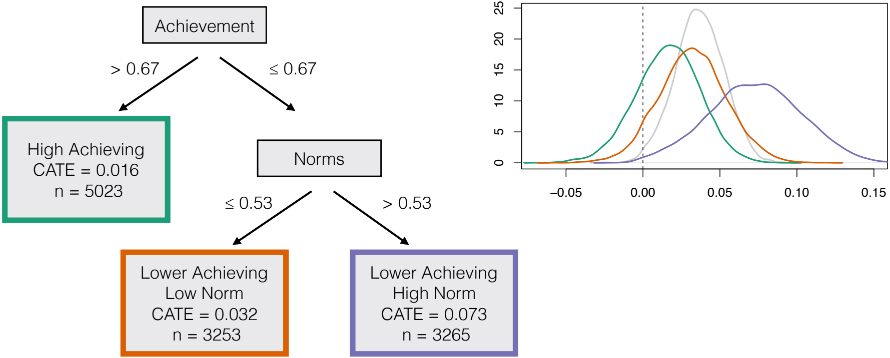
- Minimize the posterior expected loss

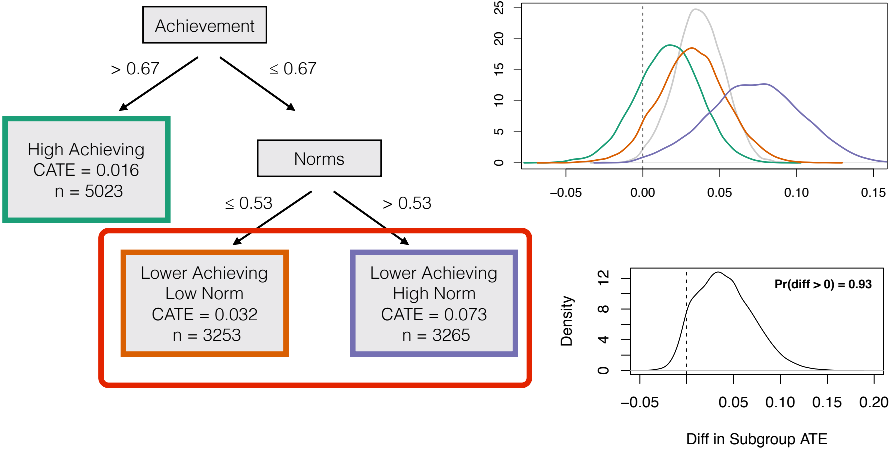
$$\hat{\gamma} = \underset{\tilde{\gamma} \in \Gamma}{\arg\min}\, E_\tau[d(\tau, \tilde{\gamma}) + p(\tilde{\gamma}) \mid Y, \mathbf{x}]$$

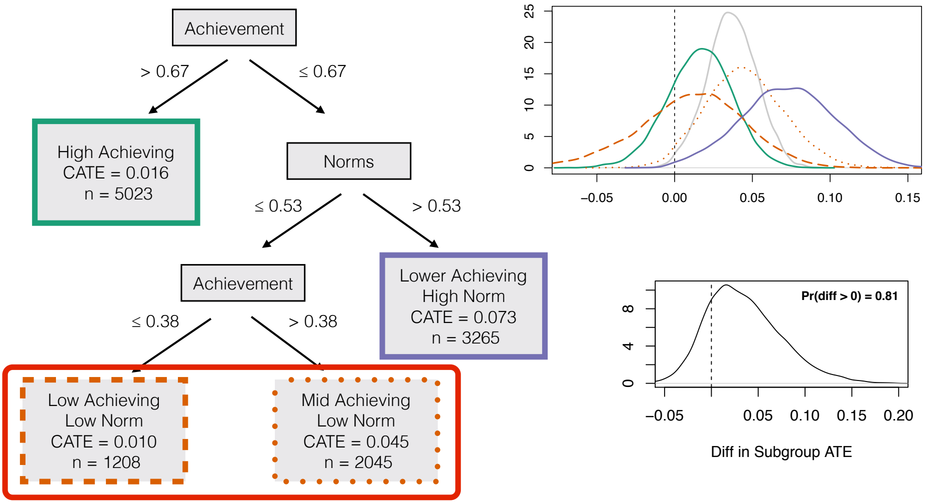
  where $p()$ is a complexity penalty and $d()$ is squared error averaged over a distribution for $\mathbf{x}$.
- With $\hat{\gamma}$ in hand we can look at the **joint** posterior distribution of subgroup ATEs

Relaxing complexity penalties in the loss function $\rightarrow$ growing a deeper tree

(Hahn et al (2017), Sivaganesan et al (2017))


Achievement
> 0.67
≤ 0.67
High Achieving
CATE = 0.016
n = 5023
Norms
≤ 0.53
> 0.53
Lower Achieving
Low Norm
CATE = 0.032
n = 3253
Lower Achieving
High Norm
CATE = 0.073
n = 3265
0
5
10
15
20
25
−0.05
0.00
0.05
0.10
0.15


Achievement
> 0.67
≤ 0.67
High Achieving
CATE = 0.016
n = 5023
Norms
≤ 0.53
> 0.53
Lower Achieving
Low Norm
CATE = 0.032
n = 3253
Lower Achieving
High Norm
CATE = 0.073
n = 3265
25
20
15
10
5
0
−0.05
0.00
0.05
0.10
0.15
Pr(diff > 0) = 0.93
Density
12
8
4
0
−0.05
0.05
0.10
0.15
0.20
Diff in Subgroup ATE



Achievement
> 0.67
≤ 0.67
High Achieving
CATE = 0.016
n = 5023
Norms
≤ 0.53
> 0.53
Achievement
≤ 0.38
> 0.38
Lower Achieving
High Norm
CATE = 0.073
n = 3265
Low Achieving
Low Norm
CATE = 0.010
n = 1208
Mid Achieving
Low Norm
CATE = 0.045
n = 2045
0
5
10
15
20
25
−0.05
0.00
0.05
0.10
0.15
Pr(diff > 0) = 0.81
0
4
8
−0.05
0.05
0.10
0.15
0.20
Diff in Subgroup ATE

We fit this model, now what?

How do we present our results? ATE, school ATE, plots of school ATE by variables....

Two questions we got from our collaborators:

- For which groups was the treatment most/least effective?
- What was the impact of posited treatment effect modifiers, holding other modifiers constant?

We could generate posterior distributions of individual-level partial effects manually and try to summarize these, **or** try to approximate τ by a proxy with simple partial effects (i.e. an additive function)

Posterior projections for model interpretations with uncertainty

The goal: **Examine the "best" (in a user-defined sense) simple approximation to the "true" $\tau(\mathrm{x})$**

Given samples of τ,

1. Consider a class of simple/interpretable approximations Γ to τ
2. Make inference on

$$\gamma = \arg\min_{\tilde{\gamma} \in \Gamma} d(\tau, \tilde{\gamma}) + p(\tilde{\gamma})$$

   for an appropriate distance function d and complexity penalty $p(\gamma)$

Get draws of γ by solving the optimization for each draw of τ

Also get discrepancy metrics, like pseudo-R^2: $Cor^2[\gamma(\mathrm{x}_i), \tau(\mathrm{x}_i)]$

(Woody, Carvalho, and Murray (2019) for this approach in predictive models)

Approximate partial effects of treatment effect modifiers

Here we use:

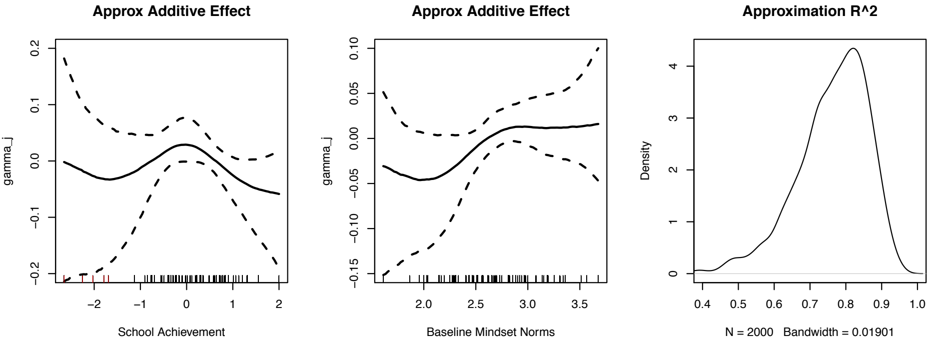
- An additive approximation $\gamma(\mathrm{x})$
- $d(\tau, \gamma) = \sum_{i=1}^{n}(\tau(\mathrm{x}_i) - \gamma(\mathrm{x}_i))^2$
- A smoothness penalty $p(\gamma)$

Don't average draws to get a point estimate! Treat

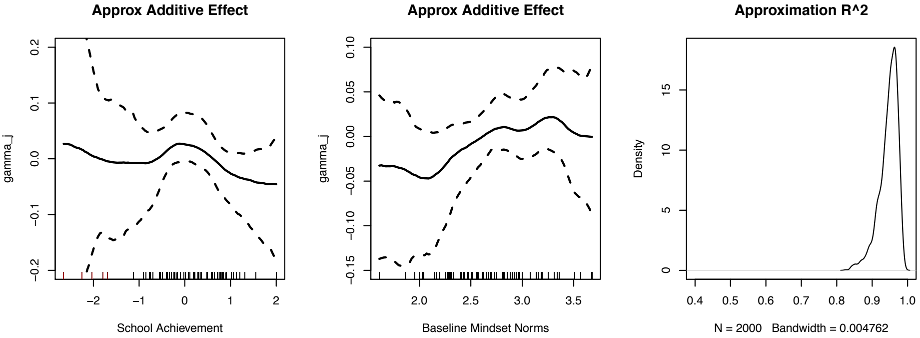
$$d(\tau, \tilde{\gamma}) + p(\tilde{\gamma}) = \sum_{i=1}^{n}(\tau(\mathrm{x}_i) - \gamma(\mathrm{x}_i))^2 + p(\tilde{\gamma})$$

as a loss function, and minimize posterior expected loss:

$$\hat{\gamma} = \operatorname*{arg\,min}_{\tilde{\gamma} \in \Gamma} E_{\tau}[d(\tau, \tilde{\gamma}) + p(\tilde{\gamma}) \mid Y, \mathrm{x}]$$



(Partial effect of minority composition not shown)



(Partial effect of minority composition not shown)

Thank you!

- P. Richard Hahn, Jared S. Murray, Carlos M. Carvalho: “Bayesian regression tree models for causal inference: regularization, confounding, and heterogeneous effects”, 2017; arXiv:1706.09523.
- Spencer Woody, Carlos M. Carvalho, Jared S. Murray: “Model interpretation through lower-dimensional posterior summarization”, 2019; arXiv:1905.07103.