Bayesian nonparametric models for treatment effect heterogeneity: model parameterization, prior choice, and posterior summarization

Jared S. Murray – The University of Texas at Austin.

March 10, 2022

Motivation

Bayesian nonparametric modeling is an effective tool for inferring heterogenous causal effects.

Bayes estimates from these models can have excellent frequentist properties – no need to drink the Kool-Aid.

Some insights about model and prior specification apply to flexible estimation of effect heterogeneity more generally

Putting BNP to work for inference about effect heterogeneity

Three considerations:

- **Model parameterization:** When you can, isolate your estimand as a parameter
- **Prior specification:** Priors are important for encoding beliefs but also for applying regularization. Regularization that ignores selection can be disasterous.
- **Posterior summarization:** “Solving” the Bayesian analogue of the post-selection inference problem, focusing on stable estimands, and giving actionable insights from complex models.

Some generic identifying 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

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

Model Parameterization

Parameterizing Nonparametric Models of Causal Effects

Forget confounding and covariates and consider estimating average treatment effect for a binary treatment in a randomized trial.

A simple model:

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

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

If $\mu_0, \mu_1 \sim N(\phi_j, \delta_j)$ 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:

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

where the estimand of interest is still τ.

Now we can express prior beliefs on τ directly and independent of nuisance parameters.

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 $f_0, f_1 \rightarrow$ 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 $\rightarrow$ simple f_0, f_1 can compose to complex τ (e.g. Künzel et al (2019)).

In addition, 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 are given by

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

and we still (generally) have no direct prior control!

Parameterizing Nonparametric Models of Causal Effects

For binary treatments, set $f(\mathbf{x}_i, z_i) = \mu(\mathbf{x}_i) + \tau(\mathbf{w}_i)z_i$, where $\mathbf{w}$ is (possibly) 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 $\tau(\mathbf{w})$ so we have direct prior control!

In Hahn et. al. (2020), we use independent BART priors on μ and τ ("Bayesian causal forests").

Prior Selection

Tweaking priors on τ

Several adjustments to the BART prior on τ in BCF:

- Higher probability on smaller τ trees (than BART defaults)
- Higher probability on "stumps" (all stumps = homogeneous effects)
- $N^+(0, v)$ Hyperprior on the scale of leaf parameters in τ

Other nonparametric priors for τ have similar "knobs" (scale, smoothness, sparsity, etc.)

For observational data, we need to adjust the prior on $\mu(\mathrm{x})$ as well, to avoid *regularization induced confounding* (Hahn et al (2016, 2020))

Regularization can induce confounding (bias)

Let's return to a linear model with homogeneous effects:

$$\begin{aligned} y_i &= f(\mathsf{x}_i, z_i) + \varepsilon_i \\ &= \tau z_i + \beta^t \mathsf{x}_i + \varepsilon_i \end{aligned}$$

and suppose x_i is high dimensional.

Assume $\beta \sim N(0, \lambda^{-1} I)$ (ridge prior) and $p(\tau) \propto 1$

What effect does the prior (regularization) have on estimating τ using the posterior mean?

Regularization can induce confounding (bias)

The bias of $\tilde{\tau} = E(\tau \mid Y, z, \mathsf{x})$ is

$$\text{bias}(\tilde{\tau}) = \lambda \hat{\delta}^t [\lambda \mathsf{I} + \mathsf{X}^t (\mathsf{I} - \mathsf{P}_z) \mathsf{X}]^{-1} \beta \tag{1}$$

where $\hat{\delta}_j$ = the OLS estimate of $x_{ij} = \delta_j z_i + \epsilon_{ij}$. Alternatively:

$$\text{bias}(\tilde{\tau}) = \lambda [\mathrm{z}^t (\mathrm{z} - \tilde{\mathrm{z}}_\lambda)]^{-1} \tilde{\gamma}_\lambda^t \beta \tag{2}$$

where $\tilde{\gamma}_\lambda = [\lambda \mathsf{I} + \mathsf{X}^t \mathsf{X}]^{-1} \mathsf{X}^t \mathrm{z}$ and $\tilde{\mathrm{z}}_\lambda = \mathsf{X} \tilde{\gamma}_\lambda$

In general, if z and x are correlated the bias is nonzero and depends on the nuisance parameter!

Solution: Don't penalize variation in $f(\mathsf{x}, z)$ along $E(Z \mid \mathsf{x})$

Expand the model to include $\hat{z}_i$ (a function of z and X) that estimates $E(Z \mid \mathsf{x})$:

$$\begin{aligned} y_i &= f(\mathsf{x}_i, z_i) + \varepsilon_i \\ &= \tau z_i + \phi \hat{z}_i + \beta^t \mathsf{x}_i + \varepsilon_i \end{aligned}$$

Keep $\beta \sim N(0, \lambda^{-1} I)$ (ridge prior) with $p(\tau, \phi) \propto 1$, so that variation in the direction of $\hat{z}_i$ is unregularized

$$\text{bias}(\tilde{\tau}) = \lambda \hat{\delta}^t [\lambda \mathsf{I} + \mathsf{X}^t (\mathsf{I} - \mathsf{P}_z) \mathsf{X}]^{-1} \beta \tag{3}$$

where $\hat{\delta}_j =$ the OLS estimate of $x_{ij} = \alpha_j \hat{z} + \delta_j z_i + \epsilon_{ij}$ (≈ 0).

Solution: Don't penalize variation in $f(\mathsf{x}, z)$ along $E(Z \mid \mathsf{x})$

Expand the model to include $\hat{z}_i$ (a function of z and X) that estimates $E(Z \mid \mathsf{x})$:

$$\begin{aligned} y_i &= f(\mathsf{x}_i, z_i) + \varepsilon_i \\ &= \tau z_i + \phi \hat{z}_i + \beta^t \mathsf{x}_i + \varepsilon_i \end{aligned}$$

Keep $\beta \sim N(0, \lambda^{-1} I)$ (ridge prior) with $p(\tau, \phi) \propto 1$, so that variation in the direction of $\hat{z}_i$ is unregularized

$$\text{bias}(\tilde{\tau}) = \lambda \hat{\delta}^t [\lambda \mathsf{I} + \mathsf{X}^t (\mathsf{I} - \mathsf{P}_z) \mathsf{X}]^{-1} \beta \tag{4}$$

where $\hat{\delta}_j =$ the OLS estimate of $x_{ij} = \alpha_j \hat{z}_i + \delta_j z_i + \epsilon_{ij}$ (≈ 0).

Regularization induced confounding is a general phenomenon

There is nothing special about the ridge prior or the linear model – RIC is easy to produce with nonlinear models and nonlinear data generating processes. (Hahn et al (2020))

In essence: Since Z is a proxy for $E(Z \mid \mathrm{x})$, if the prior on $f(\mathrm{x}, z)$ strongly penalizes variation in the "direction" of $E(Z \mid \mathrm{x})$ (and not Z) the prior encourages misattributing that variation in f to Z.

This is not a Bayes problem; it's a generic regularization problem.

How to avoid penalizing variation in $f(\mathrm{x}, z)$ along $E(Z \mid \mathrm{x})$

Including $\hat{z}_i$ as an extra coordinate/feature/covariate is often enough to mitigate regularization induced confounding.

Depending on the model, there may be easier/more efficient ways to accomplish this (e.g. residualization).

In Hahn et al (2020) we evaluate BART priors on $f(\mathrm{x}, z)$ with and without $\hat{z}$ and BCF:

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

The latter two are often *much* better and rarely worse, especially when selection into treatment is based on expected outcomes under control ("targeted selection").

Posterior Summarization

Posterior summaries, or: I fit this model, now what?

Examine the "best" (in a user-defined sense) simple approximation to a "true" $g(\mathrm{x})$ (Woody et al (2020))

Given samples of a function $g(\mathrm{x})$,

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

$$\gamma = \underset{\tilde{\gamma} \in \Gamma}{\arg\min}\, d(g, \tilde{\gamma}, \tilde{\mathrm{X}}) + p(\tilde{\gamma})$$

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

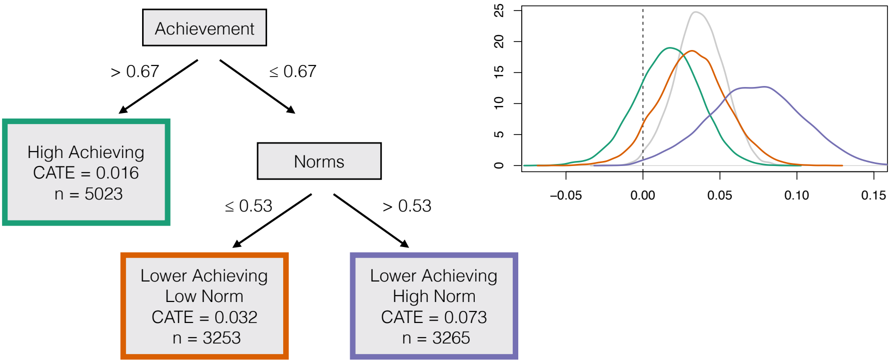
Get *draws* of γ by solving the optimization for each draw of g. Get point estimates by solving

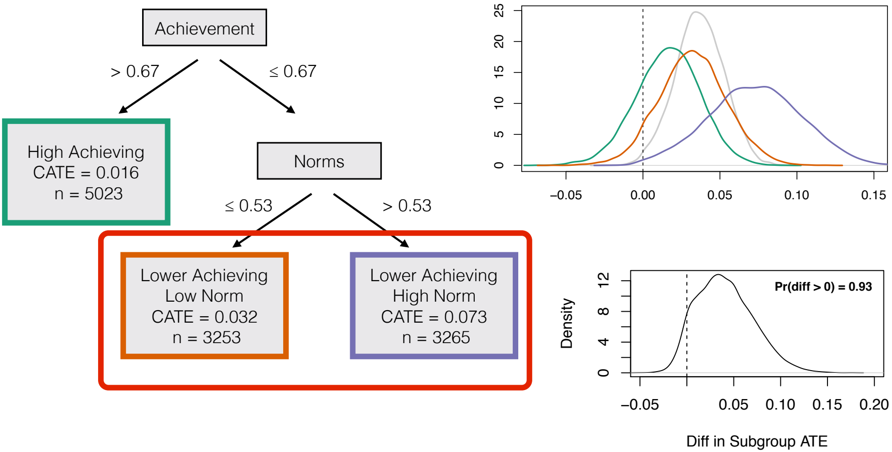
$$\hat{\gamma} = \underset{\tilde{\gamma} \in \Gamma}{\arg\min}\, E_g[d(g, \tilde{\gamma}, \tilde{\mathrm{X}}) + p(\tilde{\gamma}) \mid Y, \mathrm{x}]$$

Posterior summaries, or: I fit this model, now what?

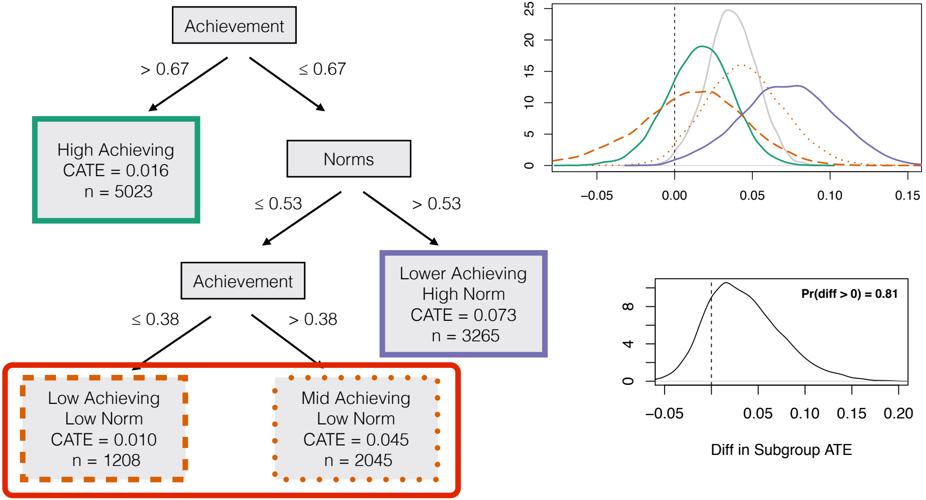
Posterior summaries:

1. Are more interpretable (subgroup analysis, linear/additive/sparse approximations) and can be targeted to scientific questions
2. Obviate the "need" to fit multiple models for different questions (Bayesians need to think about post-selection issues too) – multiple summaries use the data *once* to go prior $\rightarrow$ posterior
3. Are often more stable (coarse subgroup effects vs. individualized estimates)
4. Come with (Bayes) valid estimates of uncertainty


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
0
5
10
15
20
25
−0.05
0.00
0.05
0.10
0.15
Density
0
4
8
12
Pr(diff > 0) = 0.93
−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

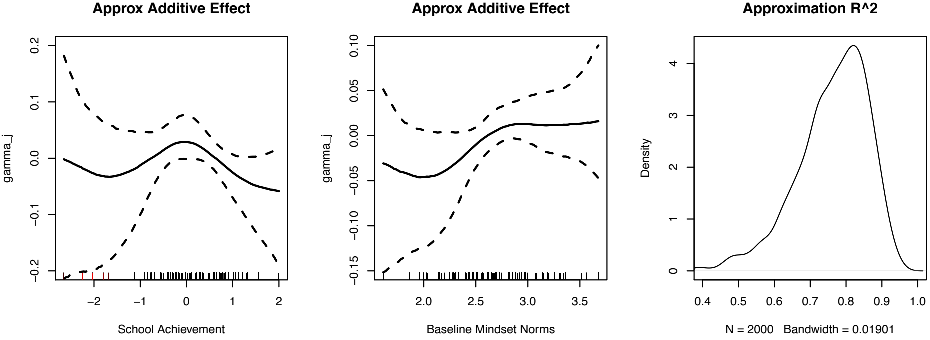
Additive summaries

We can get pproximate partial effect curves via additive summaries:

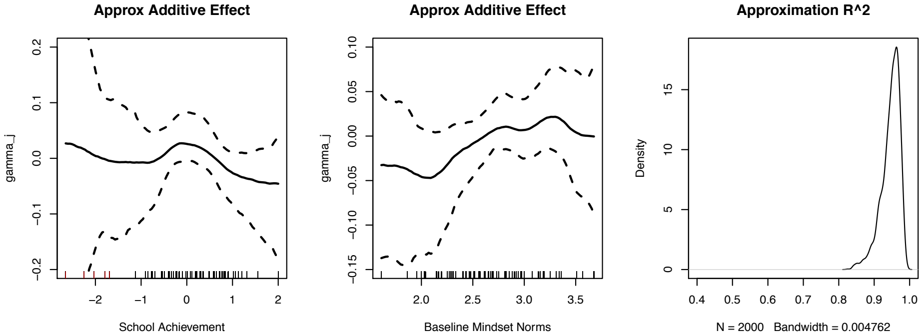
$$\tau(\mathrm{w}) \approx \gamma_0 + \sum_{j=1}^{p} \gamma_j(w_j)$$

with appropriate forms for γ_j plus smoothing penalties.

We can also get posterior on discrepancy metrics, like pseudo-R^2: $Cor^2(\gamma(\mathrm{w}_i), \tau(\mathrm{w}_i))$



(Partial effect of minority composition not shown)



(Partial effect of minority composition not shown)

Other applications of posterior summarization

- Interaction detection (Woody et al (2020)): If an additive summary is poor how do we search for missing interactions?
- Sensitivity to control function specification (Woody et al (2020b)): How do I expect removing confounders (or nonlinear/interaction terms) to change my effect estimate?
- "Explanations": Linear summaries in neighborhoods of $\mathbf{x}_i$ = LIME with uncertainty

Thank you!

jared.murray@mccombs.utexas.edu

https://jaredsmurray.github.io