Using Bayesian Causal Forest Models to Examine Treatment Effect Heterogeneity

Jared S. Murray
UT Austin
Multilevel Linear Models for Heterogeneous Treatment Effects

\[ y_{ij} = \alpha_j + \sum_{h=1}^{p} \beta_h x_{ijh} + \sum_{\ell=1}^{k} \tau_{\ell} w_{ij\ell} + \gamma_j + z_{ij} + \epsilon_{ij} \]

- School-specific intercepts/fixed/random effects
- School-specific “unexplained” heterogeneity
- Controls at the student and/or school level
- Moderators at the student and/or school level
Coloring outside the lines: Multilevel Bayesian Causal Forests

We replace linear terms with Bayesian additive regression trees (BART)

\[ y_{ij} = \alpha_j + \beta(x_{ij}) + [\tau(w_{ij}) + \gamma_j] z_{ij} + \epsilon_{ij} \]
We replace linear terms with Bayesian additive regression trees (BART).

\[ y_{ij} = \alpha_j + \beta(x_{ij}) + \left[ \tau(w_{ij}) + \gamma_j \right] z_{ij} + \epsilon_{ij} \]


Parameterizing treatment effect heterogeneity with BART is due to Hahn, Murray and Carvalho (2017).
Coloring outside the lines: Multilevel Bayesian Causal Forests

We replace linear terms with Bayesian additive regression trees (BART)

\[ y_{ij} = \alpha_j + \beta(x_{ij}) + \left[ \tau(w_{ij}) + \gamma_j \right] z_{ij} + \epsilon_{ij} \]


Parameterizing treatment effect heterogeneity with BART is due to Hahn, Murray and Carvalho (2017)

Allows for complicated functional forms (nonlinearity, interactions, etc) without pre-specification…

…while carefully regularizing estimates with prior distributions (shrinkage toward additive structure and discouraging implausibly large treatment effects)
Analyzing data with ML BCF

- Obtain posterior samples for all the parameters, compute treatment effect estimates for each unit/school/etc.

- The challenge: How do we summarize these complicated objects?
  - “Roll up” treatment effect estimates to ATE
  - Subgroup search
  - Counterfactual treatment effect predictions/“partial effects of moderators”
Application: A new analysis with NMS

- Same moderators (school mindset norms, achievement, and minority composition) + controls
- Different population (all students) and outcome (math GPA)
- Same basic process with limited researcher DOF
  - Weakly informative priors on $\tau(w)$ ($<0.5$ GPA points with high prior probability) and random effects
Inference for the Average Treatment Effect

95% confidence interval from ML Linear Model

95% uncertainty interval from ML BCF
Subgroup search

- Obtain posterior mean of treatment effects
- Use recursive partitioning (CART) on the posterior mean to find moderator-determined subgroups with high variation across subgroup ATE
  - Statistically kosher! We use the data once (prior -> posterior)
  - Can be formalized as the Bayes estimate under a particular loss function
Achievement

> 0.67

High Achieving
CATE = 0.016
n = 5023

≤ 0.67

Lower Achieving
Low Norm
CATE = 0.032
n = 3253

≤ 0.53

Norms

> 0.53

Lower Achieving
High Norm
CATE = 0.073
n = 3265

≤ 0.67

High Achieving
CATE = 0.016
n = 5023

≤ 0.53

Lower Achieving
Low Norm
CATE = 0.032
n = 3253

> 0.53

Lower Achieving
High Norm
CATE = 0.073
n = 3265

-0.05 0.00 0.05 0.10 0.15

0 5 10 15 20 25
Achievement

Norms

High Achieving
CATE = 0.016
n = 5023

Lower Achieving

Low Norm
CATE = 0.032
n = 3253

Lower Achieving
High Norm
CATE = 0.073
n = 3265

Pr(diff > 0) = 0.93
Achievement Norms

High Achieving
CATE = 0.016
n = 5023

Lower Achieving High Norm
CATE = 0.073
n = 3265

Pr(diff > 0) = 0.81

> 0.67
≤ 0.67

≤ 0.53
> 0.53

Low Achieving
Low Norm
CATE = 0.010
n = 1208

Mid Achieving
Low Norm
CATE = 0.045
n = 2045

Achievement

> 0.38
≤ 0.38

0.00
0.05
0.10
0.15

Pr(diff > 0) = 0.81

Diff in Subgroup ATE
Counterfactual treatment effect predictions

• How do estimated treatment effects change in lower achieving/low norm schools if norms increase, holding constant school minority comp & achievement?

• Not a formal causal mediation analysis (roughly, we would need “no unmeasured moderators correlated with norms”)
1 IQR = 0.6 extra problems on worksheet task

Original 0.032 (-0.011 0.076)
+10% 0.050 (0.005, 0.097)
+Half IQR 0.051 (0.005, 0.099)
+Full IQR 0.059 (0.009, 0.114)
Conclusion

• Flexible models + careful regularization + posterior summarization is a winning combination

• Our approach takes the best parts of linear models with lots of researcher degrees of freedom and “black box” machine learning methods that only afford bankshot regularization and summarization

  • Many “degrees of freedom” in the summarization step, but these depend on the data only through the posterior

  • Unlike many ML methods, we can handle multilevel structure and prior knowledge with ease
\[ y_i = f(x_i) + \epsilon_i, \quad \epsilon_i \sim N(0, \sigma^2) \]

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