The Bayesian causal forest model: regularization, confounding, & heterogeneous effects

September 2020

P. Richard Hahn, Arizona State University
Jared S. Murray, University of Texas at Austin
Carlos M. Carvalho, University of Texas at Austin

Problem setting

- Continuous outcome
- Binary treatment
- Observational data
- Strong ignorability
- Regression adjustment using machine learning
- Conditional average treatment effects (CATE)
- Finite sample performance

$$Y_i(0), Y_i(1) \perp\!\!\!\perp Z_i \mid \mathbf{X}_i$$
$$0 < \Pr(Z_i = 1 \mid \mathrm{x}_i) < 1$$

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

Priors on treatment effects (regularization)

$$\mathrm{E}(Y_i \mid \mathrm{x}_i, z_i) = \mu(\mathrm{x}_i) + \tau(\mathrm{w}_i) z_i$$

"separate regressions"

$$\mathrm{E}(Y_i \mid \mathrm{x}_i, z_i = 1) = f_1(\mathrm{x}_i)$$

$$\mathrm{E}(Y_i \mid \mathrm{x}_i, z_i = 0) = f_0(\mathrm{x}_i)$$

"treatment is just another covariate"

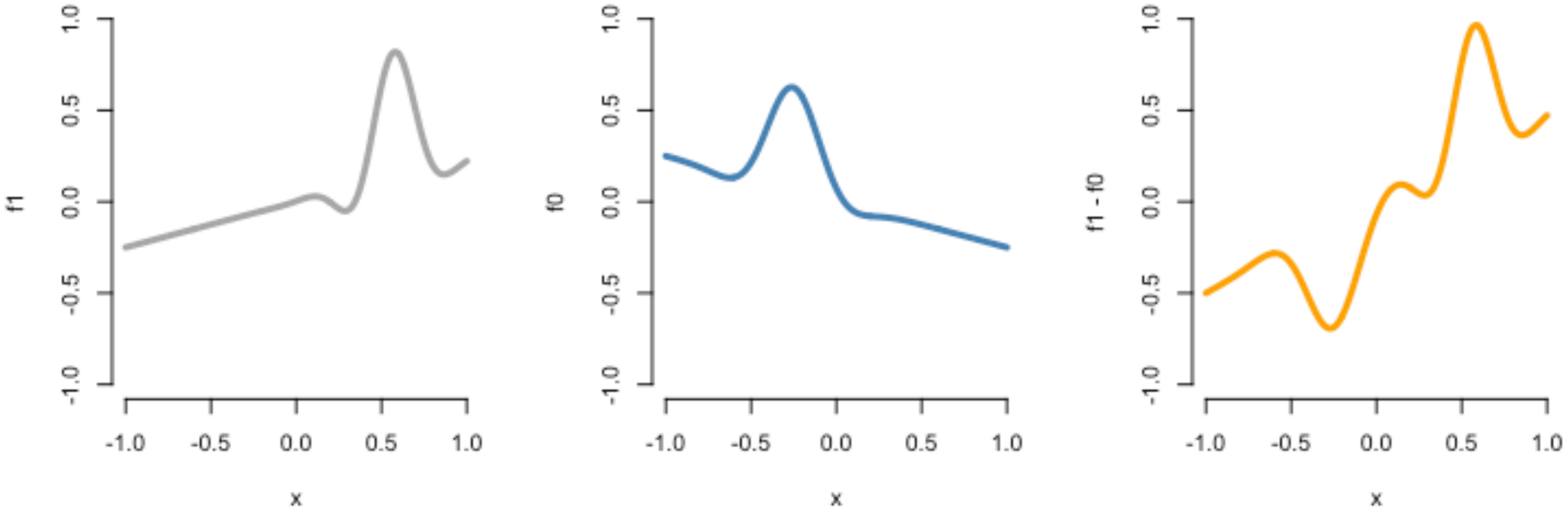
$$\mathrm{E}(Y_i \mid \mathrm{x}_i, z_i) = f(\mathrm{x}_i, z_i)$$

J. Hill in JCGS 2011

A problem with independent regressions

$$\mathrm{E}(Y_i \mid \mathrm{x}_i, z_i = 1) = f_1(\mathrm{x}_i)$$
$$\mathrm{E}(Y_i \mid \mathrm{x}_i, z_i = 0) = f_0(\mathrm{x}_i)$$

Independent priors imply treatment effect heterogeneity that is more complex *a priori* than either potential outcome mean function.



The problem applies generically to this parametrization: It is a problem for any model with independent priors over f_1 and f_0.

A problem with “just another covariate”

Modeling a single response surface $\mathrm{E}(Y_i \mid \mathrm{x}_i, z_i) = f(\mathrm{x}_i, z_i)$ implies that the prior over the treatment effect function $\tau(\mathrm{x}_i)$ depends on $\{(\mathrm{x}_i, z_i)\}_{1 \leq i \leq n}$.

As a function of the empirical distribution of the treatment variable and the other covariates, this prior can be difficult to understand and calibrate.

The problem applies generically to this parametrization: It is a problem for any model or prior over a single response surface.

Solution: parametrize in terms of the difference

In simple models, there is a common solution: independent priors in a reparametrized model.

$$\begin{aligned} Y_{i1} &\overset{\text{iid}}{\sim} \mathrm{N}(\mu_1, \sigma^2) \\ Y_{j0} &\overset{\text{iid}}{\sim} \mathrm{N}(\mu_0, \sigma^2) \end{aligned} \qquad \text{vs.} \qquad \begin{aligned} Y_{i1} &\overset{\text{iid}}{\sim} \mathrm{N}(\mu + \tau, \sigma^2) \\ Y_{j0} &\overset{\text{iid}}{\sim} \mathrm{N}(\mu, \sigma^2) \end{aligned}$$

The same intuition applies in the (nonlinear) regression case, but it requires new code.

$$\mathrm{E}(Y_i \mid \mathrm{x}_i, z_i) = \mu(\mathrm{x}_i) + \tau(\mathrm{w}_i) z_i$$

Virtues of the treatment effect parametrization

$$\mathrm{E}(Y_i \mid \mathrm{x}_i, z_i) = \mu(\mathrm{x}_i) + \tau(\mathrm{w}_i) z_i$$

- Can explicitly regularize treatment effects to taste.
- Perfectly general: $Y_i(1) = Y_i(0) + \tau_i Z_i$
- Distinct variables as moderators!

An important covariate transformation for causal inference

We also propose including an approximate propensity function as an additional coordinate in our response surface model.

$$\hat{\pi}(\mathrm{x}_i) \approx \Pr(Z_i = 1 \mid \mathrm{x}_i)$$

$$\mathrm{E}(Y_i \mid \mathrm{x}_i, z_i) = f(\mathrm{x}_i, \hat{\pi}(\mathrm{x}_i), z_i)$$

In the treatment effect parametrization used in BCF this looks like:

$$\mathrm{E}(Y_i \mid \mathrm{x}_i, z_i) = \mu(\mathrm{x}_i, \hat{\pi}(\mathrm{x}_i)) + \tau(\mathrm{x}_i) z_i$$

Targeted selection

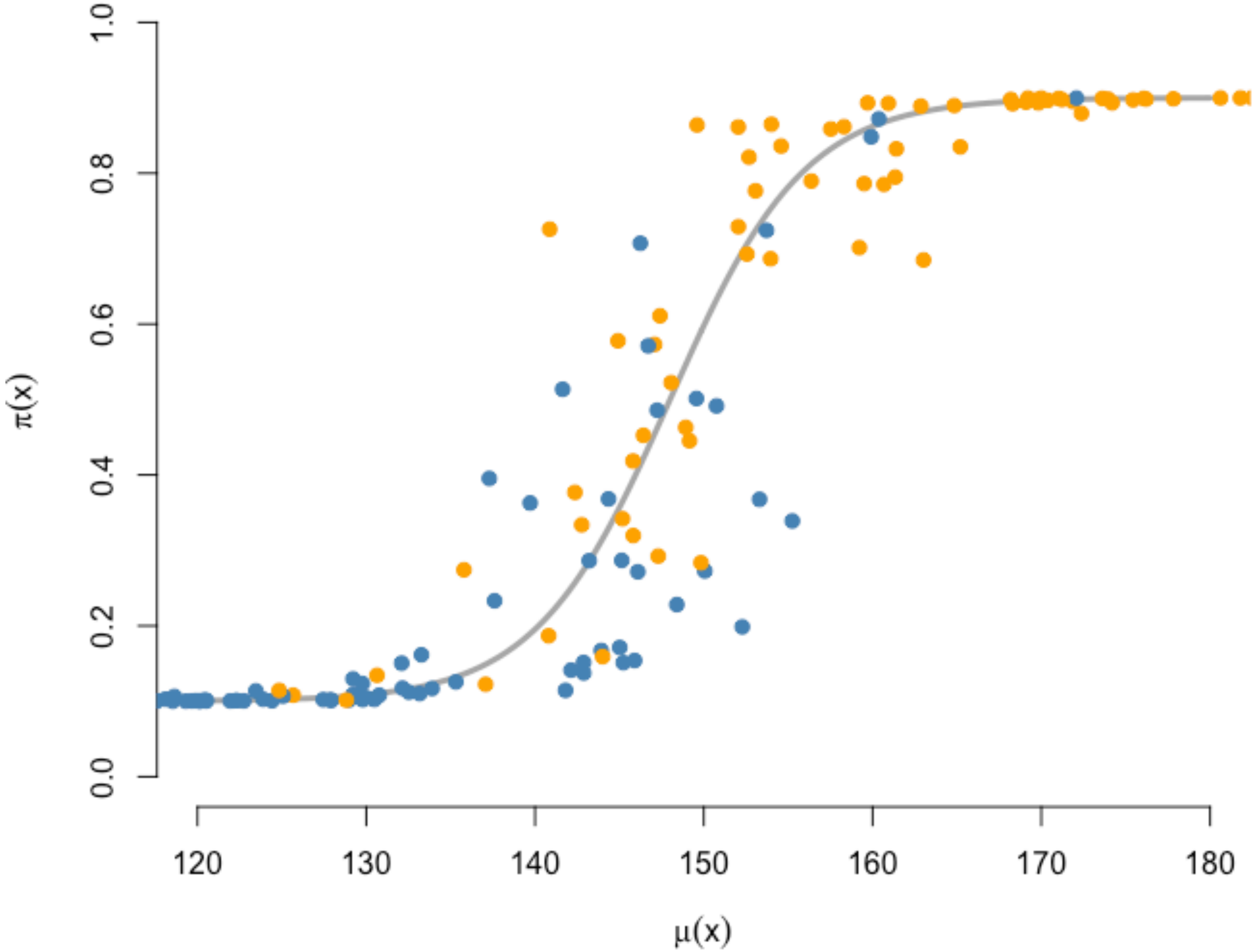
Doctors treat the patients who they think need it.

$$\Pr(Z_i = 1 \mid \mathrm{x}_i) = \Pr(Z_i = 1 \mid \hat{Y}_i(0), \mathrm{x}_i)$$

$$\hat{Y}_i(0) \approx \mathrm{E}(Y_i(0) \mid \mathrm{x}_i) = \mu(\mathrm{x}_i)$$

- Probability of treatment is increasing (or decreasing) in the expected outcome under no treatment.
- The idea is more widely applicable than this motivating example.
- We don’t argue that it *always* happens.
- But we think sometimes it does.
- **So, what are the implications of targeted selection for inference?**

Targeted selection



Regularization-induced confounding (RIC)

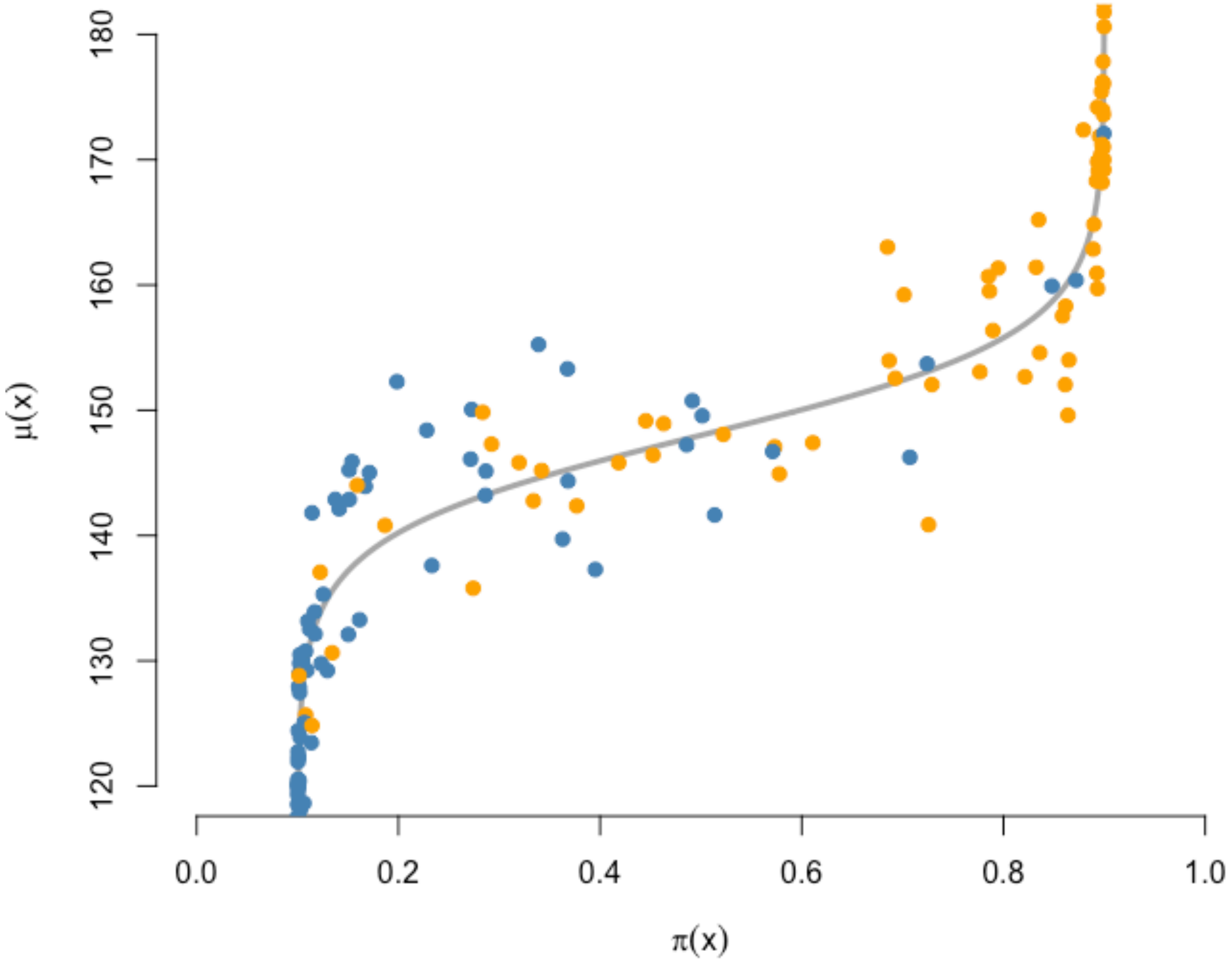
When

1) $\mu(\mathrm{x}_i)$ is **complex,**
2) $\pi(\mathrm{x}_i)$ looks like $\mu(\mathrm{x}_i)$

then misattributing $\mu(\mathrm{x}_i)$ to treatment effect can yield a similar overall fit with a much **simpler** response surface, which may be favored by a regularization prior.

Targeted selection reliably produces RIC in simulation studies.

RIC results from complexity penalties generally, rather than any particular representation of complexity.



Regularization-induced confounding

When

1) $\mu(\mathrm{x}_i)$ is **complex,**
2) $\pi(\mathrm{x}_i)$ looks like $\mu(\mathrm{x}_i)$

then misattributing $\mu(\mathrm{x}_i)$ to treatment effect can yield a similar overall fit with a much **simpler** response surface, which may be favored by a regularization prior.

Solution:

make it simple to deconfound.

Including $\hat{\pi}(\mathrm{x}_i) \approx \Pr(Z_i = 1 \mid \mathrm{x}_i)$ explicitly as a feature achieves this.

This strategy can be used in any regression model.

Putting the pieces together

$$\mathrm{E}(Y_i \mid \mathrm{x}_i, z_i) = \mu(\mathrm{x}_i, \hat{\pi}(\mathrm{x}_i)) + \tau(\mathrm{x}_i) z_i$$

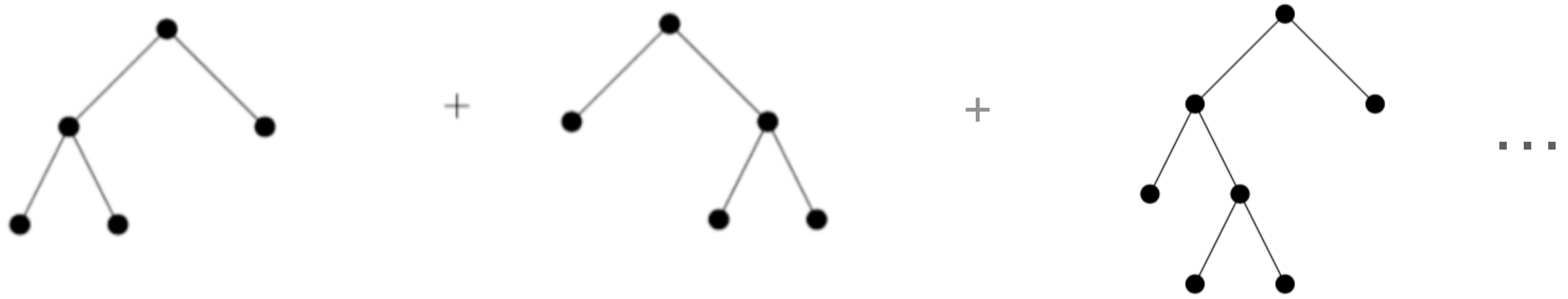
To implement these insights, we must

1) specify nonparametric priors over the the two unknown functions, and
2) chose a model for the distribution of the observed responses, given the form of the mean function above. (Our response model assumes additive, homoskedastic, Gaussian errors.)

Note: we treat the propensity score approximation as a fixed, pre-specified transformation. Its role is to prevent regularization-induced confounding.

BART as a prior over functions

$$f(\mathrm{x}_i) = \sum_{\ell=1}^{L} g(\mathrm{x}_i, T_\ell, \mathbf{m}_\ell) =$$

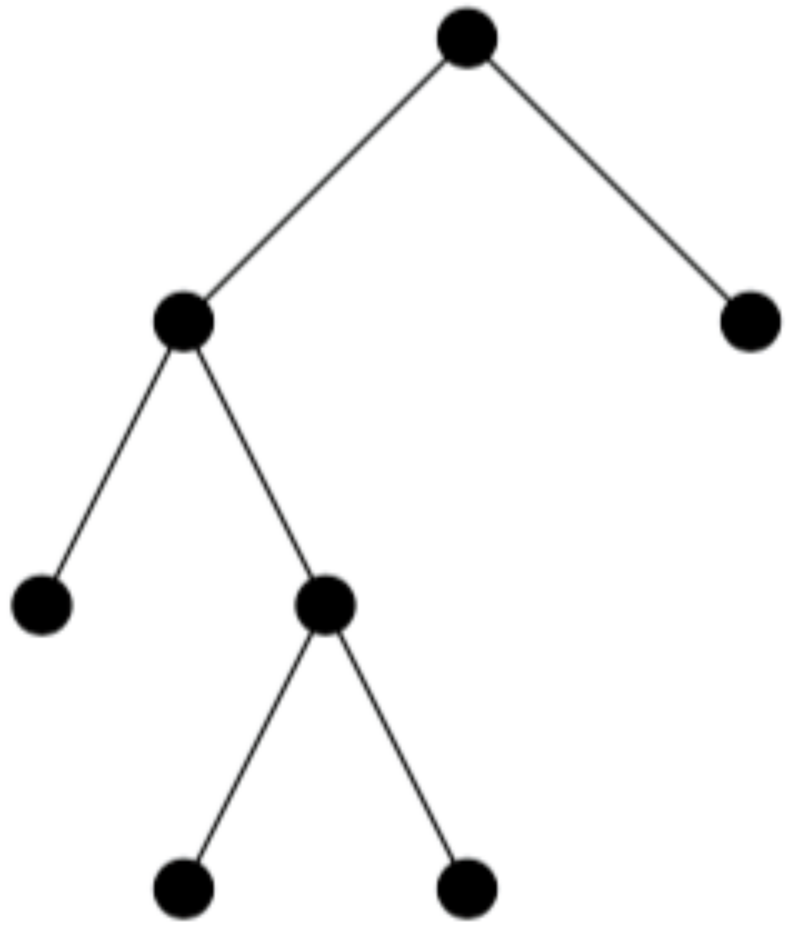


A prior over functions is then induced via

1) independent “process priors” over trees
2) independent priors over leaf parameters, given the tree.

$$\Pr(\text{split} \mid d) = \alpha(1+d)^{-\beta} \implies T_\ell \sim P_{\alpha,\beta}$$

$$m_{\ell,b} \overset{\text{iid}}{\sim} \mathrm{N}(0, v)$$



Virtues of BART priors

$$\begin{aligned} \mu \mid \mathbf{X}, \hat{\pi}(\mathbf{X}, \mathrm{z}) &\sim \mathrm{BART}(\alpha_\mu, \beta_\mu, L_\mu, \upsilon_\mu) \\ \tau \mid \mathbf{X} &\sim \mathrm{BART}(\alpha_\tau, \beta_\tau, L_\tau, \upsilon_\tau) \end{aligned}$$

- Conditional on the trees, it is a Gaussian process.
- Because the trees are estimated, we learn the implied covariance function.
- The learned covariance may be nonstationary and/or anisotropic.
- “Smoothness” can be modulated by the number of trees in the sum.
- A prior can easily be placed over the leaf scale parameter.
- Computation is not trivial, but is relatively tractable compared to similarly flexible models.

Does BCF work?

- If our model and prior are exactly right then Bayes estimators minimize Bayes risk.
- The existing theory is encouraging but incomplete and asymptotic.
- In a wide array of simulation studies BCF outperforms competitors by a variety of criteria.
- Competing methods with available theory do not attain their theoretical performance.

What does its success in simulation studies tell us about BCF?

Simulation **experiments**

A simulation experiment should

1) indicate what questions it means to address, and
2) provide a rationale for why the proposed data generating processes (DGPs) will help answer them.

A simulation study might:

- Be suggestive of real-world performance by using DGPs designed to be representative of data we are likely to observe in practice.
- Provide an understanding of which aspects of a method drive its success or failure on specific DGPs.
- Distinguish between families of DGPs where a method performs well and those where it performs poorly.

Decide on which factors to vary and attempt to control everything else.

Designing realistic causal DPGs

We pay special attention to five aspects of our causal DGPs:

- magnitude of treatment effects relative to the response surface level
- variation in treatment effects
- signal-to-noise ratio of the outcome data
- complexity of functions
- strength of confounding

These aspects are easy to monitor and control in the treatment effect parametrization and using targeted selection.

$$\mathrm{E}(Y_i \mid \mathrm{x}_i, z_i) = \mu(\mathrm{x}_i) + \tau(\mathrm{x}_i) z_i$$

$$\Pr(Z_i = 1 \mid \mathrm{x}_i) = \Pr(Z_i = 1 \mid \hat{Y}_i(0), \mathrm{x}_i)$$

$$\hat{Y}_i(0) \approx \mathrm{E}(Y_i(0) \mid \mathrm{x}_i) = \mu(\mathrm{x}_i)$$

Tricks for designing causal DPGs

- magnitude of treatment effects relative to the response surface level: Plot them!
- variation in treatment effects: Plot them!
- signal-to-noise ratio of the outcome data: Set error variance relative to function variation in-sample
- complexity of functions: Compositions permit univariate plotting of complex multivariate functions
- strength of confounding: Targeted selection with a two-parameter link function