Some Problems on BART and Posterior Summarization

Antonio R. Linero
University of Texas at Austin

Goal of This Talk

- Discuss some areas where I wish BART was more developed.
- Discuss some variants of BART I think are potentially useful.
- Discuss some problems in model summarization.
- Hoping to stimulate some discussion.
  - Open to being wrong on all counts!
  - Maybe converge on some ideas worth pursuing.
- Roughly ordered from "practical" to "abstract", but I don't value purely abstract topics.

Usability of BART

Holes in Software Ecosystem

Vast majority of applications just use the usual semiparametric normal model

$$Y_i = r(X_i) + \epsilon_i, \qquad \epsilon_i \sim \text{Normal}(0, \sigma^2).$$

Adding models on the next slide would form part of a complete ecosystem, which we are far away from.

Holes in Software Ecosystem

Vast majority of applications just use the usual semiparametric normal model

$$Y_i = r(X_i) + \epsilon_i, \qquad \epsilon_i \sim \text{Normal}(0, \sigma^2).$$

Adding models on the next slide would form part of a complete ecosystem, which we are far away from.

All of these need good interfaces as well! Not glamorous, but I think important if we care about people using BART.

- Diagnostics
- Automatic model comparison
- Basic S3 methods (`plot`, `summary`, `coef`, etc.)
- Posterior summaries

List of Methods

Model Class	Implemented	Published	Unpublished
Normal Regression (lm)	Semiparametric Gaussian	Heteroskedastic BART, Linked mean/variance, DP-Mixture BART	skew-t_ν
Generalized Linear Models (glm)	Binomial	Poisson, Gamma, Negative Binomial	Quasi-Binomial, Quasi-Poisson
Mixed Models	I think BCF does this?	–	This is needed for everything
Quantile Regression	–	Asym Laplace	Anything Better???
Survival	Fully Nonparametric, AFT BARTs	Cox PH, Submodel Shrinkage, Weibull Regression	–
Ordinal Outcomes	Continuation Ratio (via survival hack)	Ordinal Probit	–
Vector GLMs	Multinomial Logit	Multivariate Normal	Multivariate skew-t_ν
Fully-Nonparametric	–	Tilting models, Latent BART	Stick-Breaking Models

For reference, the **mediation** package covers most of these models.

Soft BART

Decision Tree

A decision tree can be represented as

$$g(x; \mathcal{T}, \mathcal{M}) = \sum_{\ell} \phi_\ell(x)\, \mu_\ell,$$

where $\phi_\ell(x) = I(x \text{ goes to leaf } \ell)$. *Not smooth!*

Soft BART

Decision Tree

A decision tree can be represented as

$$g(x; \mathcal{T}, \mathcal{M}) = \sum_{\ell} \phi_{\ell}(x)\, \mu_{\ell},$$

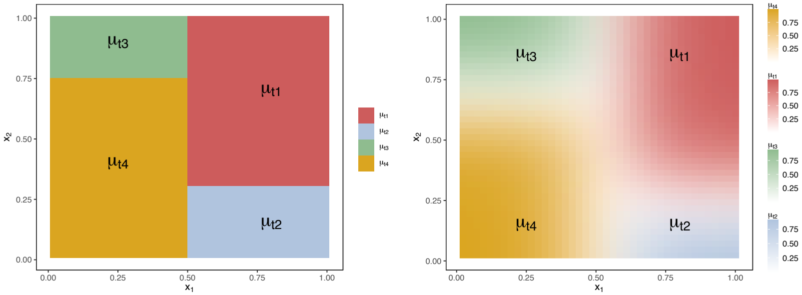
where $\phi_{\ell}(x) = I(x \text{ goes to leaf } \ell)$. *Not smooth!*

Idea: replace step function $\phi_{\ell}(x)$'s with a partition of unity:

$$\phi_{\ell}(x) = \prod_{b \in A(\ell)} \psi_b(x)^{I(\text{path to } \ell \text{ goes left})} \times \{1 - \psi_b(x)\}^{I(\text{path to } \ell \text{ goes right})}$$

where, e.g., $\psi_b(x) = [1 + \exp\{-(x - c_b)/\tau_b\}]^{-1}$.

Soft Decision Trees



Linero and Yang (2018)

Faster SoftBart

Claim

The *soft* version of BART gives superior performance to standard BART. I'm aware of no problem where Soft BART is worse than BART, but there are settings where it is meaningfully better.

Faster SoftBart

Claim

The *soft* version of BART gives superior performance to standard BART. I'm aware of no problem where Soft BART is worse than BART, but there are settings where it is meaningfully better.

Problem

Soft BART is too slow to be practical in many settings, especially for larger N.

Accelerating Soft BART?

- Ideas depend a bit too much on the technical details.
- Possibly can be acceleration through:
  - Smarter choice of $\psi_b(x)$ that allows caching computations.
  - Better bookkeeping.
  - 2x+ speedup possible from making my code less redundant.
  - XBART-type extensions?
- **Unrelated problem:** Poisson regression (or similar) for Soft BART?

Possibly scooped on this: Ran and Bai (2023) report 10x speedup! (Can Drew add to package?!)

Robust Inference With BART

Model Robustness Problem

The Problem

BART models are usually restricted to inference in *parametric families* such as Gaussian, binomial, or Poisson models. How can we adapt BART to work in general settings when we are not confident in parametric assumptions?

Model Robustness Problem

The Problem

BART models are usually restricted to inference in *parametric families* such as Gaussian, binomial, or Poisson models. How can we adapt BART to work in general settings when we are not confident in parametric assumptions?

Possible Solutions

- Build really flexible nonparametric models?
- Use robust pseudo-likelihood methods?

Why I Care About Robustness

- Bayesian inference usually assumes parametric models.
- When parametric assumptions fail, point estimates are maybe still good.
- Error bars, on the other hand, are bad!
  - Confidence intervals for, e.g., causal effects.
  - Prediction intervals
- Sometimes, we want to estimate non-standard things:
  - Quantiles and CDFs
  - Higher order moments
  - Etc.

Really Flexible Models: DPMs

Idea 1: Maybe some model with a "really flexible" error distribution? E.g.,

$$Y_i = r(X_i) + \epsilon_i, \qquad f(\epsilon) = \sum_{k=1}^{\infty} \pi_k \, \text{Normal}(\epsilon \mid \mu_k, \sigma_k)$$

with $f(\epsilon)$ modeled using a Dirichlet process mixture model. (George et al. 2019)

Really Flexible Models: DPMs

Idea 1: Maybe some model with a "really flexible" error distribution? E.g.,

$$Y_i = r(X_i) + \epsilon_i, \qquad f(\epsilon) = \sum_{k=1}^{\infty} \pi_k \, \text{Normal}(\epsilon \mid \mu_k, \sigma_k)$$

with $f(\epsilon)$ modeled using a Dirichlet process mixture model. (George et al. 2019)

Flexible errors, but not covariate dependent, e.g., cannot capture heteroskedasticity.

Really Flexible Models: Tilting

Idea 2: Take some desired parametric model and "tilt" it:

$$f(y \mid x) \propto \text{Normal}\{y \mid r(x), \sigma^2\} \times \Phi\{\ell(y, x)\}.$$

Really flexible! Directly modifies a desired model as well! (Li, Linero, and Murray 2022)

Really Flexible Models: Tilting

Idea 2: Take some desired parametric model and "tilt" it:

$$f(y \mid x) \propto \text{Normal}\{y \mid r(x), \sigma^2\} \times \Phi\{\ell(y, x)\}.$$

Really flexible! Directly modifies a desired model as well! (Li, Linero, and Murray 2022)

Pretty hard to deal with computationally; no direct access to quantities of interest like the mean; just seems sort of ridiculous.

Really Flexible Models: More Parametric

Idea 3: Specify a really flexible parametric model like

$$Y_i \sim \texttt{skew-t}_\nu \left(\mu(X_i), \sigma(X_i), \underbrace{\alpha(X_i)}_{\text{skew}} \right).$$

Called a *location-scale-skewness* (LSS) model (Stasinopoulos, Rigby, and Bastiani 2018; Um et al. 2022).

Really Flexible Models: More Parametric

Idea 3: Specify a really flexible parametric model like

$$Y_i \sim \texttt{skew-t}_\nu \left(\mu(X_i), \sigma(X_i), \underbrace{\alpha(X_i)}_{\text{skew}} \right).$$

Called a *location-scale-skewness* (LSS) model (Stasinopoulos, Rigby, and Bastiani 2018; Um et al. 2022).

Strikes a good balance in terms of flexibility, capturing many features of distributions we care about, while being easy-ish to fit and easy-ish to interpret.

Pseudo-Likelihoods: Quasi Models

Idea 4: Use the quasi-likelihood

$$L_\phi(\mu) = \prod_i \exp\left\{\int_{Y_i}^{\mu(X_i)} \frac{Y_i - t}{\phi V(t)}\, dt\right\},$$

where $V(t)$ is a user-specified *variance function* and ϕ is a *dispersion parameter*. Combine this with a BART prior on $\mu(\cdot)$.

Pseudo-Likelihoods: Quasi Models

Idea 4: Use the quasi-likelihood

$$L_\phi(\mu) = \prod_i \exp\left\{\int_{Y_i}^{\mu(X_i)} \frac{Y_i - t}{\phi\, V(t)}\, dt\right\},$$

where $V(t)$ is a user-specified *variance function* and ϕ is a *dispersion parameter*. Combine this with a BART prior on $\mu(\cdot)$.

Looks reasonable to me if we are happy with $V(t)$!

Pseudo-Likelihoods: Quasi Models

Idea 4: Use the quasi-likelihood

$$L_\phi(\mu) = \prod_i \exp\left\{ \int_{Y_i}^{\mu(X_i)} \frac{Y_i - t}{\phi\, V(t)}\, dt \right\},$$

where $V(t)$ is a user-specified *variance function* and ϕ is a *dispersion parameter*. Combine this with a BART prior on $\mu(\cdot)$.

Looks reasonable to me if we are happy with $V(t)$!

Problem: Quasi-likelihood carries no information on ϕ.

Pseudo-Likelihoods: Quasi Models

Hack to infer ϕ: update ϕ based on the sampling distribution

$$\frac{1}{N}\sum_{i=1}^{N}\frac{\{Y_i - \mu(X_i)\}^2}{V\{\mu(X_i)\}} \stackrel{\bullet}{\sim} \text{Gam}\left(\frac{N}{2}, \frac{N}{2\phi}\right),$$

with $\phi^{-1} \sim \text{Gam}(a, b)$ for approximate conjugacy.

Pseudo-Likelihoods: Quasi Models

Hack to infer ϕ: update ϕ based on the sampling distribution

$$\frac{1}{N}\sum_{i=1}^{N}\frac{\{Y_i-\mu(X_i)\}^2}{V\{\mu(X_i)\}} \stackrel{\bullet}{\sim} \operatorname{Gam}\left(\frac{N}{2},\frac{N}{2\phi}\right),$$

with $\phi^{-1} \sim \operatorname{Gam}(a, b)$ for approximate conjugacy.

Have not tried this! But it seems like a reasonable way to introduce quasi-Poisson and quasi-Binomial models into the toolkit.

Pseudo-Likelihoods: Quasi Models

Hack to infer ϕ: update ϕ based on the sampling distribution

$$\frac{1}{N}\sum_{i=1}^{N}\frac{\{Y_i - \mu(X_i)\}^2}{V\{\mu(X_i)\}} \stackrel{\bullet}{\sim} \text{Gam}\left(\frac{N}{2}, \frac{N}{2\phi}\right),$$

with $\phi^{-1} \sim \text{Gam}(a, b)$ for approximate conjugacy.

Have not tried this! But it seems like a reasonable way to introduce quasi-Poisson and quasi-Binomial models into the toolkit.

Problem: Existence of stationary distribution? Does it actually work?

Moment Based Methods

Problem

The goal standard would be for me to obtain valid inference from an arbitrary *estimating equation*

$$\mathbb{E}[s\{Y_i; r(x)\} \mid X_i = x] = 0$$

such that the posterior is valid *irrespective of the data generating process.*

Moment Based Methods

Problem

The goal standard would be for me to obtain valid inference from an arbitrary *estimating equation*

$$\mathbb{E}[s\{Y_i; r(x)\} \mid X_i = x] = 0$$

such that the posterior is valid *irrespective of the data generating process.*

- Bayesian generalized method of moments?
- Bayesian generalized estimating equations?
- Bayesian exponentially tilted empirical likelihood?
- Reduction to “robust” approx-sufficient statistics?
- I have no idea how to do this effectively.

Orthogonalizing BART Models

Orthogoanlized Ensembles

Consider a *multiple forest model*:

$$Y_i = \alpha(X_i) + \beta(A_i, X_i) + \epsilon_i.$$

Examples:

- Bayesian causal forest
- Varying coefficient BART models
- Some targeted smoothing models I've used.

Orthogoanlized Ensembles

Consider a *multiple forest model*:

$$Y_i = \alpha(X_i) + \beta(A_i, X_i) + \epsilon_i.$$

Examples:

- Bayesian causal forest
- Varying coefficient BART models
- Some targeted smoothing models I've used.

Problem

It is possible that $\alpha(X_i)$ and $\beta(A_i, X_i)$ are highly correlated! This leads to all sorts of practical issues.

Can be resolved, e.g., by ensuring that $\text{Cov}\{\beta(A_i, X_i), X_i\} = \mathbf{0}$, referred to as *orthogonalization*.

Why I Care About Orthogonalizing

- **Much** better mixing.

Why I Care About Orthogonalizing

- **Much** better mixing.
- Occasionally better statistical inference.

Why I Care About Orthogonalizing

- **Much** better mixing.
- Occasionally better statistical inference.
  - Immediately suggests Robbins transform in causal inference.

Why I Care About Orthogonalizing

- **Much** better mixing.
- Occasionally better statistical inference.
  - Immediately suggests Robbins transform in causal inference.
  - Automatically incorporates “clever covariates”.

Why I Care About Orthogonalizing

- **Much** better mixing.
- Occasionally better statistical inference.
  - Immediately suggests Robbins transform in causal inference.
  - Automatically incorporates “clever covariates”.
- Better model identifiability.

Gaining Insights from Orthogonal GPs

A model that is very simple to orthogonalize is the *Gaussian process.* Suppose

$$\boldsymbol{Y} = \boldsymbol{X}\beta + \boldsymbol{r} + \epsilon, \qquad \epsilon \sim \text{Normal}(0, \sigma^2 \mathrm{I}).$$

Given a kernel matrix Σ for $\boldsymbol{r}$, can make $\boldsymbol{r}$ uncorrelated with $\boldsymbol{X}$ using the replacement kernel

$$(\mathrm{I} - \Pi)\Sigma(\mathrm{I} - \Pi)$$

where $\Pi = \boldsymbol{X}(\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{X}^\top$ is the projection onto $\mathcal{C}(\boldsymbol{X})$.

Orthogonalized GP Plus Horseshoe

Model

Generative model

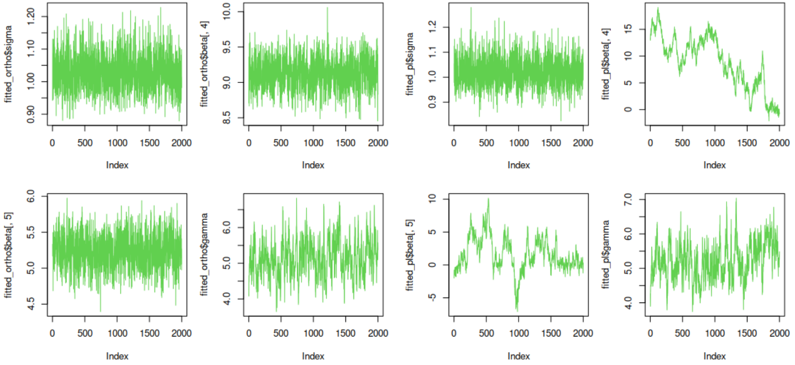
$$Y_i = X_i^\top \beta + \gamma \, r(X_i) + \epsilon_i$$

with $r(x)$ either (i) orthogonalized or (ii) not orthogonalized.

Prior

- $r(x)$ has (orthogonalized) GP prior with squared exponential kernel
- $\beta_j, \gamma \sim \text{Normal}(0, \tau^2 \lambda_j^2)$
- $\tau, \lambda_j \sim C_+(0, 1)$

Mixing



Without orthogonalization, $r(x)$ is confounded with $x^\top \beta$.

Applications to BART

- When using the *general BART model*, orthogonalize with

$$\sum_{t,\ell}\{\phi_{t\ell}(x) - x^\top(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{\phi}_{t\ell}\}\mu_{t\ell}.$$

- In a BCF, suggests we should instead use a forest of the form

$$\mu(X_i) + \{A_i - \widehat{e}(X_i)\}\,\tau(X_i),$$

  which is already a good idea for statistical reasons.

- Also relevant for hierarchical models, where it leads naturally to “within-between” models.

Targeted Smoothing

- A *targeted smoothing* (Starling et al. 2020; Li, Linero, and Murray 2022) approach takes

$$\sum_{t,\ell} \psi_t(z)\, \phi_{t\ell}(x)\, \mu_{t\ell}.$$

The variable Z_i is the variable we want to smooth over.

- Orthogonalize by instead using

$$\sum_{t,\ell} [\psi_t(z) - \mathbb{E}\{\psi_t(Z_i) \mid X_i = x\}] \phi_{t\ell}(x)\, \mu_{t\ell}.$$

- For certain ψ_t's and models for Z_i, expectation will be easy to compute (Fourier features, for example).

Posterior Projections

Uncertainty in Projections

Posterior Project Approach

To produce an interpretable model summary, we *project* $\mu(x)$ onto an interpretable model class:

$$\mu^\star(x) = x^\top \beta \qquad \text{where} \qquad \beta = \arg\min_b \|\mu(X) - X^\top \beta\|^2$$

Uncertainty in Projections

Posterior Project Approach

To produce an interpretable model summary, we *project* $\mu(x)$ onto an interpretable model class:

$$\mu^\star(x) = x^\top \beta \qquad \text{where} \qquad \beta = \arg\min_b \|\mu(X) - X^\top \beta\|^2$$

Problem?

The definition of β is sensitive to the choice of norm, e.g.,

$$\|g\|^2_{\mathbb{F}_N} = \frac{1}{N}\sum_{i=1}^{N} g(X_i)^2 \qquad \text{or} \qquad \|g\|^2_{F_X} = \int g(x)^2 \, F_X(dx).$$

A Bad Thought Pattern

- I *probably* (not always) want to use F_X.

A Bad Thought Pattern

- I *probably* (not always) want to use F_X.
- Estimating F_X is a pain.

A Bad Thought Pattern

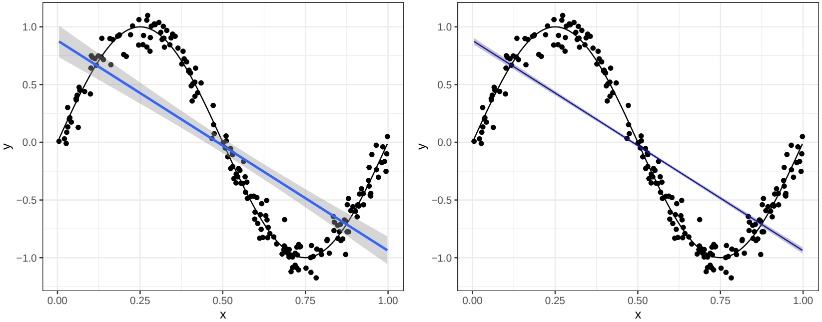
- I *probably* (not always) want to use F_X.
- Estimating F_X is a pain.
- I'll use $\mathbb{F}_N$ instead because it is easier.

A Bad Thought Pattern

- I *probably* (not always) want to use F_X.
- Estimating F_X is a pain.
- I'll use $\mathbb{F}_N$ instead because it is easier.
- *It probably won't make a big difference.*

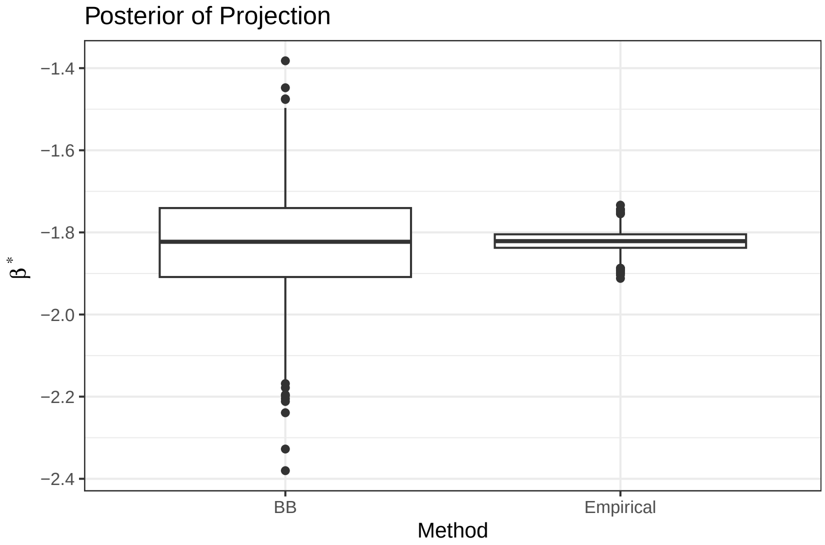
Evidence of Badness

Consider estimating linear projection $\mu^\star(x) = \beta_0 + \beta_1 x$:

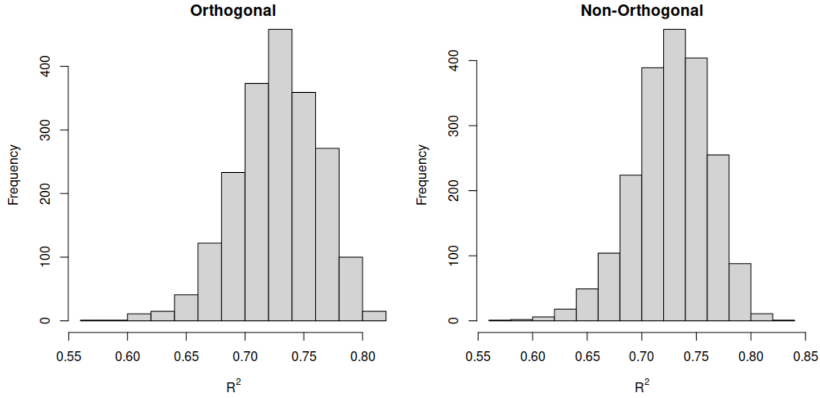


Posterior standard deviation of β_1 is roughly 7 times larger if using F_X rather than $\mathbb{F}_N$, and the difference doesn't go away with larger samples!

Evidence of Badness



Also a Problem for Summary R^2



Friedman problem with GP. When $\mathbb{F}_N$ is used instead, we get tight concentration around 0.7.

Why Does This Happen?

Suppose

$$Y_i = X_i^\top \beta + \phi(X_i)^\top \gamma + \epsilon_i,$$

where $\phi(x)$ has been orthogonalized with respect to $\mathbb{F}_N$.

- Variance of Y_i conditional on $\boldsymbol{X}$: σ^2
- Variance of Y_i unconditional on $\boldsymbol{X}$: $\gamma^\top \mathrm{Var}\{\phi(X_i)\}\gamma + \sigma^2$
- $\gamma^\top \mathrm{Var}\{\phi(X_i)\}\gamma$ gets absorbed as F_X uncertainty!
- The worse the approximation, the larger the inflation.

Is This a Real Problem?

- I can't be the first person to notice this?

Is This a Real Problem?

- I can't be the first person to notice this?
  - It seems likely Jared and Richard know this already.

Is This a Real Problem?

- I can't be the first person to notice this?
  - It seems likely Jared and Richard know this already.
  - I haven't found any reference acknowledge this as an issue.

Is This a Real Problem?

- I can't be the first person to notice this?
  - It seems likely Jared and Richard know this already.
  - I haven't found any reference acknowledge this as an issue.
- Extra uncertainty comes packaged with not knowing F_X.

Is This a Real Problem?

- I can't be the first person to notice this?
  - It seems likely Jared and Richard know this already.
  - I haven't found any reference acknowledge this as an issue.
- Extra uncertainty comes packaged with not knowing F_X.
- If you *genuinely* think $\mathbb{F}_N$ (or the empirical on a test set, or whatever) is of intrinsic interest, then there is no problem: just use that and enjoy the small variance.

Is This a Real Problem?

- I can't be the first person to notice this?
  - It seems likely Jared and Richard know this already.
  - I haven't found any reference acknowledge this as an issue.
- Extra uncertainty comes packaged with not knowing F_X.
- If you *genuinely* think $\mathbb{F}_N$ (or the empirical on a test set, or whatever) is of intrinsic interest, then there is no problem: just use that and enjoy the small variance.
- If you *really cared* about F_X, acknowledge the uncertainty.

Is This a Real Problem?

- I can't be the first person to notice this?
  - It seems likely Jared and Richard know this already.
  - I haven't found any reference acknowledge this as an issue.
- Extra uncertainty comes packaged with not knowing F_X.
- If you *genuinely* think $\mathbb{F}_N$ (or the empirical on a test set, or whatever) is of intrinsic interest, then there is no problem: just use that and enjoy the small variance.
- If you *really cared* about F_X, acknowledge the uncertainty.
  - In this case, access to large quantities of unlabeled data is hugely valuable!!!

Is This a Real Problem?

- I can't be the first person to notice this?
  - It seems likely Jared and Richard know this already.
  - I haven't found any reference acknowledge this as an issue.
- Extra uncertainty comes packaged with not knowing F_X.
- If you *genuinely* think $\mathbb{F}_N$ (or the empirical on a test set, or whatever) is of intrinsic interest, then there is no problem: just use that and enjoy the small variance.
- If you *really cared* about F_X, acknowledge the uncertainty.
  - In this case, access to large quantities of unlabeled data is hugely valuable!!!
- It feels like there might be a SATE vs. PATE lesson here…

References I

George, Edward, Purushottam Laud, Brent Logan, Robert McCulloch, and Rodney Sparapani. 2019. "Fully Nonparametric Bayesian Additive Regression Trees." In, 89–110. Emerald Publishing Limited. https://doi.org/10.1108/s0731-905320190000040b006.

Li, Yinpu, Antonio R. Linero, and Jared Murray. 2022. "Adaptive Conditional Distribution Estimation with Bayesian Decision Tree Ensembles." *Journal of the American Statistical Association* 118 (543): 2129–42. https://doi.org/10.1080/01621459.2022.2037431.

Linero, Antonio R., and Yun Yang. 2018. "Bayesian Regression Tree Ensembles That Adapt to Smoothness and Sparsity." *Journal of the Royal Statistical Society Series B: Statistical Methodology* 80 (5): 1087–1110. https://doi.org/10.1111/rssb.12293.

References II

Ran, Hao, and Yang Bai. 2023. "ASBART:accelerated Soft Bayes Additive Regression Trees." https://doi.org/10.48550/ARXIV.2310.13975.

Starling, Jennifer E., Jared S. Murray, Carlos M. Carvalho, Radek K. Bukowski, and James G. Scott. 2020. "BART with Targeted Smoothing: An Analysis of Patient-Specific Stillbirth Risk." *The Annals of Applied Statistics* 14 (1). https://doi.org/10.1214/19-aoas1268.

Stasinopoulos, Mikis D, Robert A Rigby, and Fernanda De Bastiani. 2018. "GAMLSS: A Distributional Regression Approach." *Statistical Modelling* 18 (3-4): 248–73. https://doi.org/10.1177/1471082x18759144.

Um, Seungha, Antonio R. Linero, Debajyoti Sinha, and Dipankar Bandyopadhyay. 2022. "Bayesian Additive Regression Trees for Multivariate Skewed Responses." *Statistics in Medicine* 42 (3): 246–63. https://doi.org/10.1002/sim.9613.