Introduction to Bayesian nonparametrics for causal inference

M. Daniels, A. Linero, & J. Roy

U of Florida, U of Texas-Austin, & Rutgers U.

ISBA 2022, Montreal

Outline (Parts 1 and 2)

- Part 1
  - 9.00-10.00: Review of causal inference
  - 10.00-10.30: Identifiability and sensitivity analysis
  - 10.30-10.45: Break
- Part 2: Bayesian nonparametric (BNP) models
  - 10.45-11.30; BART
  - 11.30-12.30: Dirichlet process mixtures (DPM)
- 12.30-1.30: Lunch

Outline

- Part 2: BNP models (cont.)
  - 1.30-2.15: Dependent Dirichlet processes (DDP) and Gaussian processes (GP)
- Part 3: Case studies
  - 2.15-2.45: 1: EHRs and enriched DPM
  - 2.45-3.00: Break
  - 3.00-4.00: 2: Causal mediation and BART
  - 4.00-4.45: 3: Semi-competing risks and DDP
- 4.45-5: Final questions

Learning Objectives

- Understand advantages of BNP for causal inference
- Understand key concepts in causal inference
- Compute causal quantities using G-computation
- How to choose BNP models for causal inference in different settings
- Gain insights into algorithms to fit BNP models and estimate causal quantities

Part I: Review of Causal Inference

Observed data

- Treatment: A
  - Often, $A = 1$ for treated and $A = 0$ for control
- Pre-treatment variables: L
- Outcome: Y

- Data: $\{A_i, L_i, Y_i; i = 1 \cdots, n\}$

Potential outcomes and counterfactuals

Potential outcomes:

- $Y(a)$: outcome if treatment set to $A = a$
- *Example 1*: ACE Inhibitor and blood pressure
  - $Y(1)$: SBP 3 months from now if take ACE Inhibitor
  - $Y(0)$: SBP 3 months from now if no medication
- *Example 2*: kidney transplant and survival time
  - $Y(1)$: survival time if receive kidney transplant
  - $Y(0)$: surivival time if receive dialysis

If actually receive treatment A, then $Y(A)$ is observed and $Y(1-A)$ is *counterfactual*.

Causal effects

Causal effects are contrasts between population-level summaries of potential outcomes on common populations, e.g.,

- Average causal effect: $\mathrm{E}\{Y(a=1) - Y(a=0)\}$
- Causal effect of treatment on the treated: $\mathrm{E}\{Y(a=1) - Y(a=0)|A=1\}$
- Quantile causal effect: $F_1^{-1}(p) - F_0^{-1}(p)$
  - $F_a^{-1}(p)$ is the pth quantile of the cumulative distribution function $P(Y(t) \leq y)$

Not a causal effect: $E\{Y(1)|A=1\} - E\{Y(0)|A=0\}$

- different populations

Fundamental problem of causal inference

We only observe *one* potential outcome for each subject - the others are counterfactual.

We cannot simply estimate $E\{Y(1) - Y(0)\}$ as

$$\frac{1}{n}\sum_{i=1}^{n}\{Y_i(1) - Y_i(0)\},$$

because *half* of those variables are *unobserved*.

Causal assumptions are needed.

Causal assumptions

Ignorability

- $Y(a) \perp\!\!\!\perp A|L$
- also called exchangability

Implies

$$E\{Y(1)|A=1,L\} \equiv E\{Y(1)|A=0,L\}$$

Causal assumptions (cont.)

Positivity

- $P(A = a|L) > 0$ for all a, L
- every type of subject (defined by L) in the population has a chance at getting assigned any treatment

Consistency

- $Y = Y(a)$ if $A = a$

Causal effects from observational data

Suppose we want $E(Y(a))$ and we have some discrete covariates L.

$$E(Y(a)) = \sum_{\ell} E(Y(a)|L = \ell)p(\ell)$$

Causal effects from observational data

Suppose we want $E(Y(a))$ and we have some discrete covariates L.

$$\begin{aligned} E(Y(a)) &= \sum_{\ell} E(Y(a)|L=\ell)p(\ell) \\ &= \sum_{\ell} E(Y(a)|A=a, L=\ell)p(\ell) \end{aligned}$$

Causal effects from observational data

Suppose we want $E(Y(a))$ and we have some discrete covariates L.

$$\begin{aligned} E(Y(a)) &= \sum_{\ell} E(Y(a)|L=\ell)p(\ell) \\ &= \sum_{\ell} E(Y(a)|A=a, L=\ell)p(\ell) \\ &= \sum_{\ell} E(Y|A=a, L=\ell)p(\ell) \end{aligned}$$

Causal effects from observational data

$$E(Y(a)) = \sum_{\ell} E(Y|A = a, L = \ell)p(\ell)$$

- Sum is over all possible values of covariates ℓ
  - This is just *standardization*.

g-formula

The g-formula is a more general way to find obtain causal effects when the observed data distributions are known.

$$E(Y(a)) = \int E(Y|A = a, L = \ell)p(\ell)d\ell$$

or

$$E(Y(a)|A = 1) = \int E(Y|A = a, L = \ell)p(\ell|A = 1)d\ell$$

or

$$P(Y(a) \leq y) = \int_{-\infty}^{y} \int p(Y|A = a, L = \ell)p(\ell)dyd\ell$$

Notice: LHS potential outcomes; RHS observables

Estimation

Suppose $E(Y|A = a, L = \ell)$ is known up to a parameter vector θ, i.e., $E(Y|A = a, L = \ell; \theta)$.

- we could estimate θ
- and then compute $\widehat{E}(Y(t)) = \frac{1}{n}\sum_{i=1}^{n} E(Y_i|A = a, L_i = \ell_i; \widehat{\theta})$ for each a

This implicitly uses the empirical distribution of L.

Estimation (cont.)

Alternatively, suppose we also know $p(\ell)$ up to a parameter vector η, i.e., $p(\ell; \eta)$.

- we could estimate θ and η
- we could generate m draws, $\ell_1, \cdots, \ell_m$, from $p(\ell|\widehat{\eta})$
- and then compute $\widehat{E}(Y(a)) = \frac{1}{m}\sum_{j=1}^{m} E(Y|A = a, L_j = \ell_j; \widehat{\theta})$ for each t

This involves Monte Carlo integration. This approach is known as *g-computation*.

Bayesian g-computation

The Bayesian version of g-computation will be similar to previous 2 slides, except:

- prior distribution for θ and (possibly) η
- g-computation step at draws from posterior distribution of the parameters
- obtain *full posterior* for the causal effects of interest

More on this later.

Overview

Interest is often in understanding the impact that a treatment has on the outcome through intermediate variable or variables.

- need to define causal effects
- we focus here on *natural* direct and indirect effects
- consider different sets of causal assumptions

Data and Notation

- Y: outcome
- M: mediator
- A: treatment
- L: confounders
- Data: $\{L_i, A_i, M_i, Y_i; i = 1, \ldots, n\}$

Potential outcomes

$M(a)$

- value of mediator that would be observed if A was set to a
- Consistency: $M = M(A)$

$Y(a, M(a'))$

- Outcome that would be observed if A was set to a and M was set to the value that it would have taken if A was set to a'
- Consistency: $Y = Y(A, M(A))$

Natural direct effects

Natural direct effect

$$E\{Y(1, M(0))\} - E\{Y(0, M(0))\}$$

- Imagine setting the mediator to the value it would take under no treatment ($M(0)$ – its *natural* value) and then comparing the potential outcomes if treatment was set to 1 versus if it was set to 0

Natural indirect effects

Natural indirect effect

$$E\{Y(1, M(1))\} - E\{Y(1, M(0))\}$$

- Imagine setting the treatment A to 1 and then comparing the potential outcomes if mediator was set to what it would be if treatment 1 versus if treatment 0

Decomposition

We can write the total effect as the sum of the natural direct and indirect effects:

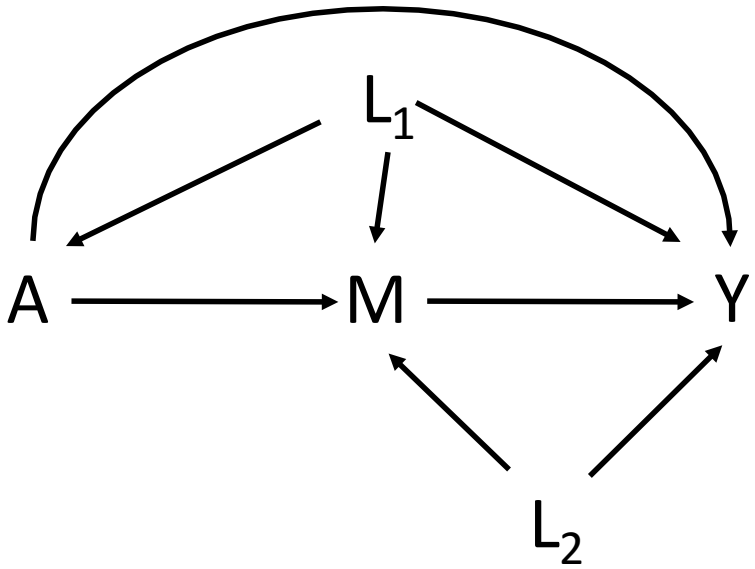
$$E\{Y(1)\} - E\{Y(0)\} = E\{Y(1, M(0))\} - E\{Y(0, M(0))\} + E\{Y(1, M(1))\} - E\{Y(1, M(0))\}$$

Cross-worlds

A drawback of the natural direct and indirect approach is they involve cross-world counterfactuals. For example, counterfactuals like: $Y(1, M(0))$

- these are counterfactuals that are unobservable even from experimental data
- e.g., suppose treatment is pre-exposure prophylaxis (PrEP), the mediator is number of times having condomless sex, and the outcome is STI. $Y(1, M(0))$ is the STI status that would be observed if a subject received PrEP, but their sexual behavior was what it would have been had they not been on PrEP.

Mediation DAG with confounding



Causal assumptions: sequential ignorability

1. Ignorability of the assignment mechanism:

$$A_i \perp\!\!\!\perp \{Y_i(a, m), M_i(a')\} | L_i = \ell$$

for all (ℓ, a, a') (i.e., no unmeasured confounding between exposure and potential outcomes/mediators)

2. Ignorability of the mediator process:

$$Y_i(a, m) \perp\!\!\!\perp M_i(a') | L_i = \ell, A_i = a',$$

for all (ℓ, a, a') (i.e., no unmeasured confounding between potential outcome and mediator)

3. Positivity: $P(A_i = a | L_i = \ell) > 0$ and $P(M_i(a) = m | A_i = a, L_i = \ell) > 0$ for all (a, ℓ, m)

Identifiability

(Conditional) natural direct effect:

$$E\{Y(1, M(0))|L\} - E\{Y(0, M(0))|L\}$$
$$= \int \{E(Y|A = 1, M = m, L) - E(Y|A = 0, M = m, L)\} dF_{M|A=0,L}(m)$$

Alternative causal assumptions

Mediator induction equivalence:

Assumption 1:

$$f(Y(1, M(0))|M(0) = m, M(1), V = v) = f(Y(1, M(1))|M(0), M(1) = m, V = v)$$

Assumption 2: Joint distribution of $M(0), M(1)|V$ follows Gaussian copula with rank correlation ρ

Note: V might be different than L

Identifiability

Under mediator induction equivalence assumptions:

$$E\{Y(1, M(0))|V\} =$$
$$\int E\{Y(1, M(1))|M(1) = m_0, V\} f(m(0), m(1)|V) dm_0 dm_1$$

Traditional approach

Baron and Kenny (1986)

Fit two regression models:

$$E(M|A, L) = \beta_0 + \beta_1 A + \beta_2^T L$$

$$E(Y|A, M, L) = \theta_0 + \theta_1 A + \theta_2 M + \theta_3^T L$$

They proposed:

- direct effect: θ_1
- indirect effect: $\beta_1 \theta_2$

Requires strong parametric (statistical) assumptions!

Alternative approach

In the identification formulae on previous slides, need either mean functions or distributions.

- Use BNP to model the appropriate mean functions and/or distributions
- Use MC integration to 'compute' causal effects (this is g-computation)
- This approach avoids making strong parametric assumptions
- Can use informative priors on sensitivity parameters

Principal Stratification

In some situations, there is a post-treatment variable S that has an important role in defining the causal effects of A on Y.

Examples:

- Randomized trials with non-compliance
- Censoring by death
- Mediation (not covered today)

Non-compliance

Randomized trial where A is treatment assignment, S is treatment actually taken, and Y is the outcome.

- $S(a)$ is treatment taken if randomized to treatment group $A = a$
- e.g., If $S_i(1) = 1$, then if subject i is assigned treatment $A = 1$ and they will actually take treatment 1 (they will be compliant with their assigned treatment)
- e.g., If $S_i(1) = 0$, then if subject i is assigned treatment $A = 1$ they will actually take treatment 0 (they will not be compliant with their assigned treatment)

Principal strata

Description	$S(0), S(1)$	observed data
always-takers	$1, 1$	$A = 0, S = 1$ or $A = 1, S = 1$
compliers	$0, 1$	$A = 0, S = 0$ or $A = 1, S = 1$
defiers	$1, 0$	$A = 0, S = 1$ or $A = 1, S = 0$
never-takers	$0, 0$	$A = 0, S = 0$ or $A = 1, S = 0$

e.g., if we observe $A_i = 0, S_i = 1$, we know subject i is either an always-taker or a defier

Principal strata are latent classes

Causal effects

Assume *exclusion restriction*: $Y(a,s) = Y(a',s)$ (treatment received is what matters).

Interest is in causal effects such as

$$E\{Y|A=1,S(0)=0,S(1)=1\} - E\{Y|A=0,S(0)=0,S(1)=1\}$$

- intention-to-treat effect for the subpopulation of compliers
- $\rightarrow$ effect of treatment received on the outcome among compliers (complier average causal effect)

Identification

One way to achieve identification is with a causal assumption such as no defiers: i.e. $P(S(0) = 1, S(1) = 0) = 0$

- This would be reasonable, for example, if people in the placebo group $A = 0$ do not have access to the treatment

With no defiers assumption, the *complier average causal effect* is

$$CACE = \frac{E(Y|A = 1) - E(Y|A = 0)}{P\{S(0) = 0, S(1) = 1\}}$$

Censoring by death

Suppose we are interested in an outcome Y some time after treatment A:

- it is possible that a subject could die $S = 1$ before the outcome is observed
- the risk of death might itself be affected by treatment

Challenges:

- naively controlling for death could be adjusted for a post-treatment variable (i.e., we'd be adjusting away some of the treatment effect)
- if a subject died, we do not observe Y (and it is not exactly missing)

Censoring by death

Principal stratification approach

- There is a subgroup of individuals who would survive regardless of treatment assignment: $\{i : S_i(0) = 0, S_i(1) = 0\}$
- For these *always survivors*, Y is observed

We can then target the average causal effect of treatment among the always survivors:

$$E\{Y|A = 1, S(0) = 0, S(1) = 0\} - E\{Y|A = 0, S(0) = 0, S(1) = 0\}$$

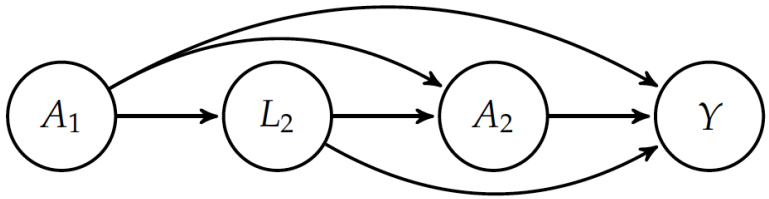
- We will explore this in a case study.

Motivating example

- A study of new treatment of high blood sugar
  - Patients were randomized to either treatment or placebo groups at baseline
  - Patients were followed over time and blood sugar level was measured at the end of study
  - Patients have different levels of compliance depending on the health status, both may change over time
- The study objective is to estimate the treatment effect taking into account non-adherence

Motivating example (cont'd)

- Consider a DAG for two-period study:
  - A_1: treatment at time 1
  - L_2: health status at time 2
  - A_2: treatment at time 2
  - Y: blood sugar



- Important feature: consider same treatment received at different times as separate treatment; separate nodes in DAG

Notation and Definitions

Observed data:

- Treatment: A_1, A_2 (at times 1 and 2)
  - Often, $A_t = 1$ for treated and $A_t = 0$ for untreated at time t
- Confounders: L_2 (at time 2)
- Outcome: Y
- Assume temporal ordering is A_1 then L_2 then A_2 then Y
- History notation: use overbar
  - $\overline{a} = (a_1, a_2)$

Potential outcomes:

- $Y(g)$: outcome if intervention set to g
  - More on this shortly

Static Interventions

A static intervention is one where what treatment is set to does not depend on post-treatment variables. The intervention g only involves treatment $\bar{a}$.

Possible static interventions:

- Never treat: $Y(a_1 = 0, a_2 = 0) \equiv Y(0, 0)$
- Always treat: $Y(1, 1)$
- Treat, then don't: $Y(1, 0)$
- Don't treat, then do: $Y(0, 1)$

Static Interventions: Causal Effects

We might be interested in contrasting the treatment strategies $g =$'always treat' and $g =$'never treat' in the population:

$$E(Y(1,1)) - E(Y(0,0))$$

- This is the average difference in outcomes if everyone in the population was treated at times 1 and 2 versus if no one was treated at time 1 and time 2

Dynamic Interventions

For dynamic interventions, the treatment strategy depends, in part, on time-updated covariates.

Consider the question of when to intensify treatment for diabetes (Neugebauer et al. 2012). Dynamic treatment strategies can take the form: "intensify treatment the first time at which their A1c level is θ% or above and remain on the intensified therapy thereafter."

g_θ: strategy of intensifying when A1c$> \theta$%

- Note: even though this is dynamic, it is well defined ahead of time

Dynamic Interventions: Causal Effects

We might be interested in comparing the average outcomes for the population under various dynamic interventions. e.g.,

$$E(Y(g_7)) - E(Y(g_8))$$

- This is the average difference in outcomes if everyone in the population had their treatment intensified the first time that their A1c exceeded 7% versus having their treatment intensified the first time A1c exceeded 8%

Causal Assumptions

Sequential ignorability (no unmeasured confounding)

- $Y(a_1, a_2) \perp\!\!\!\perp A_1$, for all a_1, a_2
- $Y(a_1, a_2) \perp\!\!\!\perp A_2 | L_2, A_1$, for all a_1, a_2

Positivity

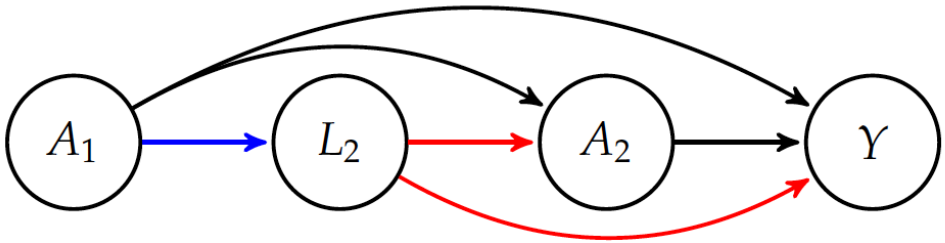
- $P(A_1 = a_1) > 0$ for all a
- $P(A_2 = a_2 | L_2, A_1) > 0$ for all a, L_2

Consistency

- $Y = Y(a_1, a_2)$ if $A_1 = a_1$ and $A_2 = a_2$

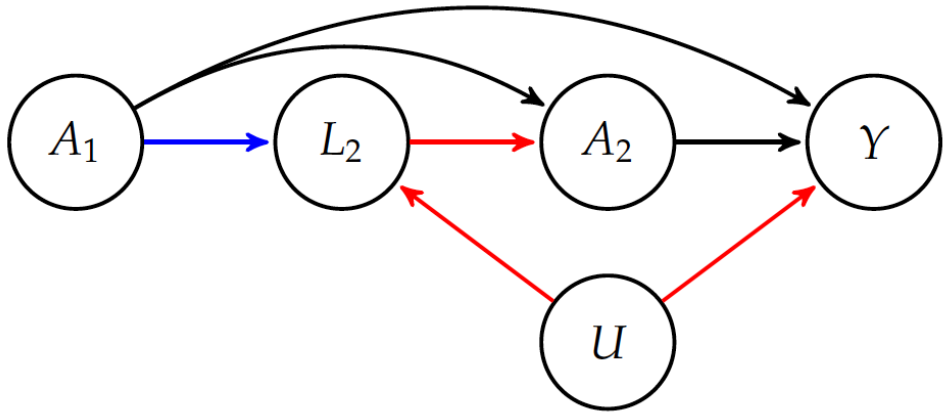
Time-dependent confounding

- Time dependent confounding occurs if both conditions satisfy:
  - L_2 confounds the effect of A_2 on Y ($L_2 \rightarrow A_2, L_2 \rightarrow Y$)
  - L_2 is on intermediate pathway from A_1 to Y ($A_1 \rightarrow L2 \rightarrow Y$)



Time-dependent confounding (cont'd)

- Time dependent confounding can also be present if L_2 is not intermediate but is affected by A_1 and associated with Y



Time-dependent confounding (cont'd)

In presence of time dependent confounding, standard regression/stratification approach is NOT valid to estimate the joint treatment effect, even if there is no unmeasured confounding

g-methods

g-methods can be used to estimate causal effects when there is time-dependent confounding:

- g-computation (likelihood-based approach)
- Inverse probability of treatment weighting (IPTW) of marginal structural models
- g-estimation of structural nested mean models

g-computation

$$
\begin{aligned}
E(Y(a_1, a_2)) &= \sum_{l_2} E(Y(a_1, a_2)|L_2 = \ell_2)p(l_2) \\
&= \sum_{l_2} E(Y(a_1, a_2)|L_2 = \ell_2, A_1 = a_1, A_2 = a_2)p(l_2|a_1) \\
&= \sum_{l_2} E(Y|L_2 = \ell_2, A_1 = a_1, A_2 = a_2)p(l_2|a_1)
\end{aligned}
$$

- 2 regression models
- fit those, then compute $E(Y(a_1, a_2))$ for any a_1, a_2

Review

- Untestable causal assumptions are necessary
- If distributions known, can use g-formula to obtain causal effects
- Bayesian version of g-formula
  - likelihood, prior, computation
- Later today
  - sensitivity to uncheckable assumptions: sensitivity parameters
  - flexible models for distributions: weak assumptions about observed data

Part I: Identifiability and Sensitivity Analysis

June 26, 2022

Priors for unidentified parameters I

- in causal inference settings, assumptions are required that cannot be 'checked' by the data
  - as such, inferences about certain parameters may not become more precise as more data is collected.
  - Such a parameter (called a *sensitivity parameter*, SP), of which the estimand of interest is likely a function, will be completely 'determined' by the prior.
- Specification of SPs will typically involve tradeoffs between allowing a realistic range of types of violations of the assumption and keeping the number of SPs low and interpretable.

Priors for unidentified parameters II

- introduce a simple example next and then provide details about sensitivity parameters for several of the assumptions introduced in the review of causal inference

Example: Sensitivity parameters for ignorability I

- Consider a point treatment setting under the *ignorability* assumption,

$$Y(a) \perp\!\!\!\perp A|L$$

- this implies

$$[Y(1) \mid A = 0, L] \equiv [Y(1) \mid A = 1, L],$$

and is (clearly) uncheckable from the observed data

Example: Sensitivity parameters for ignorability II

- assume $[Y(1) \mid A = 1, L]$ is a normal distribution with mean $\beta_0 + \beta_1 L$ and variance σ^2
- assume $[Y(1) \mid A = 0, L]$ is also a normal distribution with the same mean structure $\alpha_0 + \alpha_1 L$, but possibly different parameters, and variance σ^2
- under ignorability, $\alpha_j = \beta_j : j = 0, 1$.

Example: Sensitivity parameters for ignorability III

- now embed ignorability in a more general assumption

$$\alpha_j = \beta_j + \Delta_j$$

where Δ_j is not identifiable from the observed data (embedded sensitivity parameter)

Example: Sensitivity parameters for ignorability IV

- To quantify our prior uncertainty in how far the model deviates from ignorability, we can put an informative prior on Δ_j
  - note the ignorability assumption implicitly assumes Δ_j has prior

$$\Pi(\Delta_j) = I\{\Delta_j = 0\}$$

Example: Sensitivity parameters for ignorability V

- The impact of Δ_j on the conditional mean of the potential outcome, $Y(1)$, can be seen as

$$E[Y(1)|L=\ell] = \sum_a E[Y(1)|A=a, L=\ell]p(A=a|L=\ell)$$

where

$$E[Y(1) \mid A=1, L=\ell] = \beta_0 + \beta_1 \ell$$

and

$$E[Y(1)|A=0, L=\ell] = \beta_0 + \Delta_0 + (\beta_1 + \Delta_1)\ell.$$

- as such, the ACE is a function of the unidentified parameters, (Δ_0, Δ_1) (more in next section)

How to calibrate and/or specify priors for sensitivity parameters I

1. specify an anchoring restriction such as ignorability
2. embed that restriction in a family with substantively-meaningful sensitivity parameters ξ,
3. decide on a plausible range (and/or a prior) for ξ to investigate.

We just did the first two steps

How to calibrate and/or specify priors for sensitivity parameters II

- Consider three general approaches to determine plausible values of ξ
  1. perform a *tipping point* analysis to identify the values of ξ which cause our substantive inference to change; if the tipping point region is far away from the values of ξ which are substantively plausible then we conclude that our analysis is robust
  2. calibrate based on observed data assuming the sensitivity parameter might be bounded based on observed data summaries (e.g., proportion of variability explained or standard deviations of the equivalent quantities in the observed data) and potentially give it a 'default' prior

How to calibrate and/or specify priors for sensitivity parameters III

3. work with a subject matter expert to attempt to construct a realistic informative prior $\pi(\xi)$ for ξ. By incorporating this prior (which, due to non-identifiability, will also be the posterior of ξ) we can arrive at a single inference which combines all possible assumptions in a principled fashion.

How to calibrate and/or specify priors for sensitivity parameters IV

- The first two strategies have the advantage of not requiring subject-matter input about ξ prior to fitting the model, and we do not have to engage in a possible complicated elicitation process.
- The second and third strategy have the potential advantage that we reduce the range of possible inferences to a single inference which averages over our uncertainty in ξ.

Sensitivity to the ignorability assumption with a point treatment I

- A key assumption for identifying the average causal effect in the point treatment setting is ignorability: $Y(a) \perp\!\!\!\perp A \mid L$.
- here we expand (over the simple illustration earlier) on an approach to carrying out a sensitivity analysis (and priors) for possible violations of the ignorability assumption.

Sensitivity to the ignorability assumption with a point treatment II

- recall, we can identify $E\{Y(a)\}$ as follows:

$$E\{Y(a)\} = \int E\{Y(a)|L=\ell\}dF_{L=\ell}(\ell) \quad (1)$$

$$= \int E\{Y(a)|L=\ell, A=a\}dF_{L=\ell}(\ell) \quad (2)$$

$$= \int E\{Y|L=\ell, A=a\}dF_{L=\ell}(\ell) \quad (3)$$

where (2) holds because of ignorability and (3) because of consistency.

Sensitivity to the ignorability assumption with a point treatment III

- If we are not confident in the ignorability assumption, we can consider how to weaken it or account for uncertainty about it (as above).
- What allowed us to go from (1) to (2) is the fact that under ignorability,

$$E\{Y(a)|L=\ell\} = E\{Y(a)|L=\ell, A=1\} = E\{Y(a)|L=\ell, A=0\}$$

- the difference

$$\Delta_a(\ell) = E\{Y(a)|L=\ell, A=1\} - E\{Y(a)|L=\ell, A=0\}$$

might not equal 0; this is a potential sensitivity parameter

Causal estimand as function of sensitivity parameter I

- Denote by Ψ the average causal effect $\Psi = E\{Y(1)\} - E\{Y(0)\}$.
- The contrast in standardized means can be written as the true causal effect plus a bias term:

$$\int E\{Y|L=\ell, A=1\}dF_{L=\ell}(\ell) - \int E\{Y|L=\ell, A=0\}dF_{L=\ell}(\ell) = \Psi + \xi$$

where

$$\xi = \int [\Delta_1(\ell)e(\ell) + \Delta_0(\ell)\{1-e(\ell)\}]dF_{L=\ell}(\ell)$$

and

$$e(\ell) = P(A=1|\ell)$$

Causal estimand as function of sensitivity parameter II

- sensitivity analysis involves tradeoffs between allowing a realistic range of the types of violations of ignorability while also keeping the number of sensitivity parameters low and interpretable.
- For example, if we specified $\Delta_a(\ell)$ as a complex function of a and ℓ with many parameters, there would be no realistic way to carry out a sensitivity analysis
- Alternatively, simple functions with, say, 1 to 3 interpretable parameters allows for the possibility of having a sensitivity analysis that can be understood by a subject matter experts and/or specified as a function of the observed data

Causal estimand as function of sensitivity parameter III

- In our example, suppose we simplify the sensitivity parameters by assuming that people who actually received treatment had potential outcome $Y(a)$ that was Δ units different, on average, than people who did not actually receive treatment.
  - Suppose the amount Δ does not depend on a or on the values of the confounders L,

$$\Delta = \Delta_1(\ell) = \Delta_0(\ell)$$

  - Now, suppose, for example, that there was an unmeasured confounder, independent from L, that lead to healthier people being more likely to receive treatment.
    - This could be viewed as a worst case scenario, because our observed L's tell us nothing about the unmeasured confounder.
    - So, although we have simplified the problem, we did so in such a way that could be viewed as conservative.

Calibration of sensitivity parameters I

- specify an informative prior distribution to capture our uncertainty about Δ (where if $\Delta = 0$ then ignorability holds)
- calibrate it using strategy 2
- Let σ be the residual standard deviation of $[Y|L]$.
- we might assume that unmeasured confounding leads to no larger than a k standard deviation deviation from ignorability: i.e.,

$$|\Delta| < k\sigma$$

Calibration of sensitivity parameters II

- 'default' informative priors might include a uniform distribution over this interval or triangular priors that place more weight on the non-zero values of Δ
- for the latter, consider a mixture of a triangular prior on $(-k\sigma, 0)$ and $(0, k\sigma)$ with the max at the non-zero ends of the intervals
- we could also place a prior on k

Calibration of sensitivity parameters III

- an alternative observed data summary (strategy 2) would be to use R^2 the total amount of variability in Y that is explained by L (we implement this approach for causal mediation)

Sensitivity to monotonicity in principal stratification I

- recall, principal stratification is an approach to causal inference with post-treatment variables.
- consider, a principal stratification estimand, the survivor average causal effect
- Define $S(a)$ to be the potential survival outcome under treatment a.
- a survivor average causal effect for a binary outcome,

$$\text{SACE} = \frac{Pr[Y(1) = 1 \mid S(1) = 1, S(0) = 1]}{Pr[Y(0) = 1 \mid S(1) = 1, S(0) = 1]},$$

Sensitivity to monotonicity in principal stratification II

- for identification, often use a monotonicity assumption (among other assumptions)
- the (deterministic) monotonicity assumption specifies $S(1) \geq S(0)$, i.e.,

$$\Pr\{S(1) = 1 \mid S(0) = 1)\} = 1$$

  - any individual who survived without the treatment would also survive if they had received the treatment

Sensitivity to monotonicity in principal stratification III

- To quantify uncertainty about this assumption, a stochastic monotonicity assumption can be used instead:

$$\begin{aligned}\Pr(S(1)=1 \mid S(0)=1) &= \Pr(S(1)=1) + \\ &\quad \rho\left[\min\left\{1, \frac{\Pr(S(1)=1}{\Pr(S(0)=1}\right\} - \Pr(S(1)=1)\right]\end{aligned}$$

- this generalizes the deterministic assumption with an embedded sensitivity parameter ρ.
- if $\pi(\rho) = I\{\rho = 1\}$ (and $P(S(1)=1) > P(S(0)=1)$), the deterministic monotonicity assumption results

Sensitivity to monotonicity in principal stratification IV

- Uncertainty about ρ (and this assumption) can be done by placing a non-degenerate prior over $[0, 1]$; e.g., a triangular prior with mode at either zero or one (based on whether want more weight on deterministic monotonicity or the max deviation from it)

Sensitivity to the sequential ignorability assumption for causal mediation I

- Recall, sequential ignorability (SI):

$$\{Y(a', m), M(a)\} \quad \perp\!\!\!\perp \quad A \,|\, L = \ell \tag{4}$$

$$Y(a', m) \quad \perp\!\!\!\perp \quad M(a) \,|\, A = a, L = \ell \tag{5}$$

- sensitivity for the first part of sequential ignorability (4) can be done using the previous approach
- but what about sensitivity to violations of (5), the second part in the sequential ignorability (mediator-outcome confounding)
- Recall (5) is based on an untestable, no unobserved confounding relationships between the mediator and the outcome.

Sensitivity to the sequential ignorability assumption for causal mediation II

- To introduce a method for sensitivity to (5), we first restate the identification results from Imai et al. (2010)

Sensitivity to the sequential ignorability assumption for causal mediation III

$$
\begin{aligned}
E(Y(1, M(0)) \mid L = \ell) &= \int E(Y(1, m) \mid M_0 = m, A = 0, L = \ell) dF_{M_0 \mid A=0, L=\ell}(m) \\
&= \int E(Y(1, m) \mid A = 0, L = \ell) dF_{M_0 \mid A=0, L=\ell}(m) \qquad (6) \\
&= \int E(Y(1, m) \mid A = 1, L = \ell) dF_{M_0 \mid A=0, L=\ell}(m) \qquad (7) \\
&= \int E(Y(1, m) \mid M_1 = m, A = 1, L = \ell) dF_{M_0 \mid A=0, L=\ell}(m) \qquad (8) \\
&= \int E(Y \mid M_1 = m, A = 1, L = \ell) dF_{M \mid A=0, L=\ell}(m),
\end{aligned}
$$

- both (6) and (8) follow from (5), the second part of SI.

Sensitivity to the sequential ignorability assumption for causal mediation IV

- Equality (7) follows from (4), first part of SI (ignorability).
- this identification result suggests a way to introduce sensitivity parameters

Sensitivity to the sequential ignorability assumption for causal mediation V

Let

$$g_a(m,\ell) = E(Y(1,m) \mid M_0 = m, A = a, L = \ell) - E(Y(1,m) \mid A = 0, L = \ell)$$

- then $E(Y(1, M(0))) \mid L = \ell)$ can be re-expressed without using (5), second part of SI, as follows,

Sensitivity to the sequential ignorability assumption for causal mediation VI

$$
\begin{aligned}
& E(Y(1, M(0)) \mid L = \ell) \\
&= \int E(Y(1, m) \mid M_0 = m, A = 0, L = \ell) dF_{A_0 \mid Z=0, L=\ell}(m) \\
&= \int \Big\{ g_0(m, \ell) + E(Y(1, m) \mid A = 0, L = \ell) \Big\} dF_{M_0 \mid A=0, L=\ell}(m) \\
&= \int \Big\{ g_0(m, \ell) + E(Y(1, m) \mid A = 1, L = \ell) \Big\} dF_{M_0 \mid A=0, L=\ell}(m) \\
&= \int \Big\{ g_0(m, \ell) - g_1(m, \ell) + E(Y(1, m) \mid M_1 = m, A = 1, L = \ell) \Big\} dF_{M_0 \mid A=0, L=\ell}(m) \\
&= \int \Big\{ g_0(m, \ell) - g_1(m, \ell) + E(Y \mid M_1 = m, A = 1, L = \ell) \Big\} dF_{M \mid A=0, L=\ell}(m),
\end{aligned}
$$

where under (5), $g_0(m, \ell) = 0$ and $g_1(m, \ell) = 0$ for all m and ℓ.

Thus, $g_0(m, \ell)$ and $g_1(m, \ell)$ are (embedded) sensitivity parameters.

Sensitivity to the sequential ignorability assumption for causal mediation VII

- to calibrate the sensitivity parameters, we again use the second strategy (based on observed data summaries)

Sensitivity to the sequential ignorability assumption for causal mediation VIII

- we first note that, from (4), the following relationships hold

$$\begin{aligned} Y(a', m) \perp A \quad &| \quad L = \ell \\ Y(a', m) \perp A \quad &| \quad M_a = m', L = \ell. \end{aligned}$$

So $g_0(m, \ell)$ and $g_1(m, \ell)$ can be re-expressed as

$$\begin{aligned} g_0(m, \ell) &= E(Y(1, m) \mid M_0 = m, L = \ell) - E(Y(1, m) \mid L = \ell), \\ g_1(m, \ell) &= E(Y(1, m) \mid M_1 = m, L = \ell) - E(Y(1, m) \mid L = \ell). \end{aligned}$$

Sensitivity to the sequential ignorability assumption for causal mediation IX

- suggest to calibrate $g_0(m, \ell)$ and $g_1(m, \ell)$ using the total amount of variability of the outcomes under $A = 0$ and $A = 1$ explained by ℓ

Sensitivity to the sequential ignorability assumption for causal mediation X

- For $g_1(m, \ell)$, we estimate the coefficient of determination among the treated $(A = 1)$ with ℓ as covariates (but not M_1) and denote it as R_1.
- Then, we set the absolute value of the difference between conditional expectations, $|g_1(m, \ell)|$, to be less than

$$\sqrt{\text{Var}(Y|A = 1) \times (1 - R_1) \times k_1}$$

where the square root of k_1 is the percent of the total variance that is not explained by ℓ.

Sensitivity to the sequential ignorability assumption for causal mediation XI

- Similarly, we assume $|g_0(m, \ell)|$ to be less than

$$\sqrt{\text{Var}(Y|A=1) \times (1 - R_1) \times k_0}.$$

Sensitivity to the sequential ignorability assumption for causal mediation XII

- further we might assume $k_0 \leq k_1$ since we expect that M_0 does not explain as much of the variance of $Y(1, m)$ as M_1.
- Thus, we have two sensitivity parameters, k_0 and k_1, bounded in the unit square.
- $k_1 = k_0 = 0$ implies second part of SI
- We can specify uniform or triangular (default) priors as noted earlier with the restriction that $k_0 \leq k_1$

Summary

- untestable assumptions are necessary for causal inference
- here we introduced strategies for embedding sensitivity parameters in these assumptions and illustrated this in three common settings
- we also introduced strategies for specifying ranges and/or priors for sensitivity parameters

Part 2: Bayesian Additive Regression Trees

Nonparametric Regression

Given data $\mathcal{D}_n = \{(X_1, Y_1), \ldots, (X_n, Y_n)\}$ with

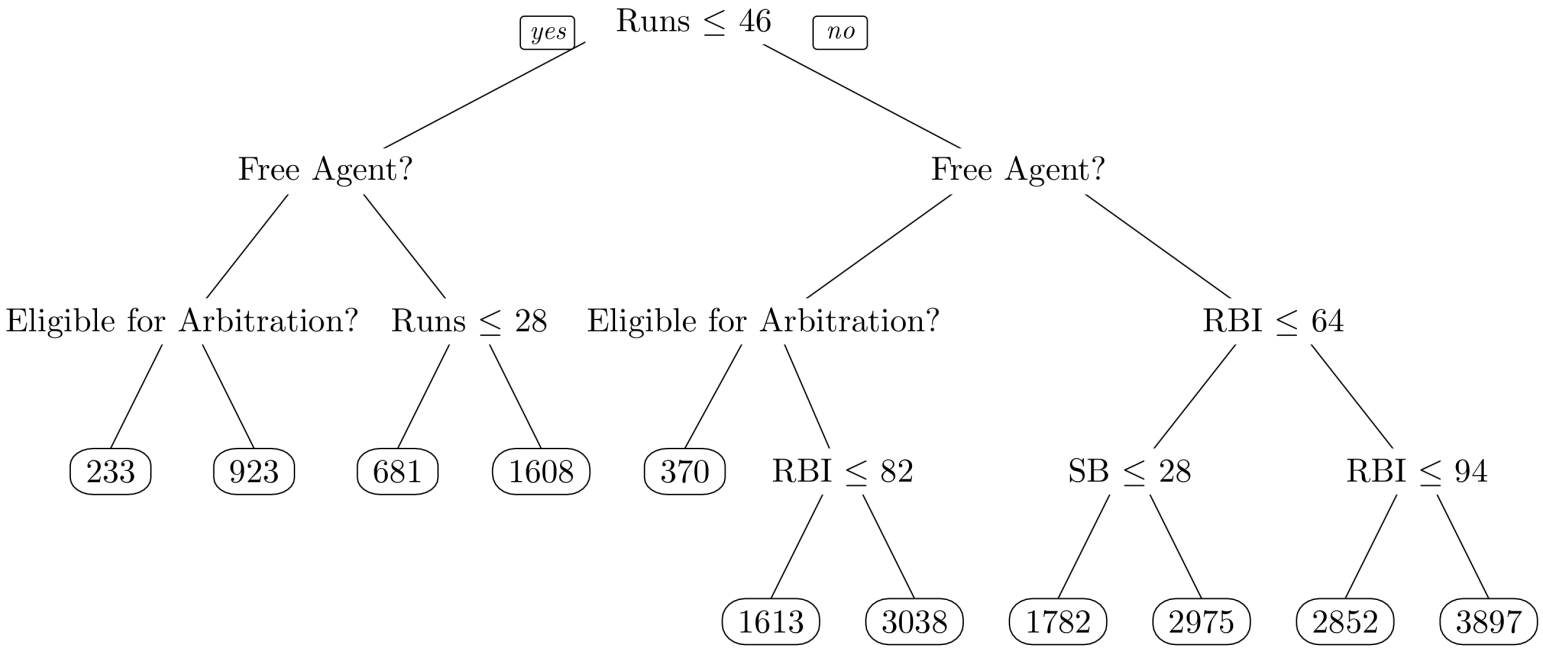
$$Y_i = \mu_0(X_i) + \epsilon_i, \qquad \epsilon_i \overset{\text{iid}}{\sim} \text{Normal}(0, \sigma_0^2),$$

how do we recover relevant features of $f_0(x)$?

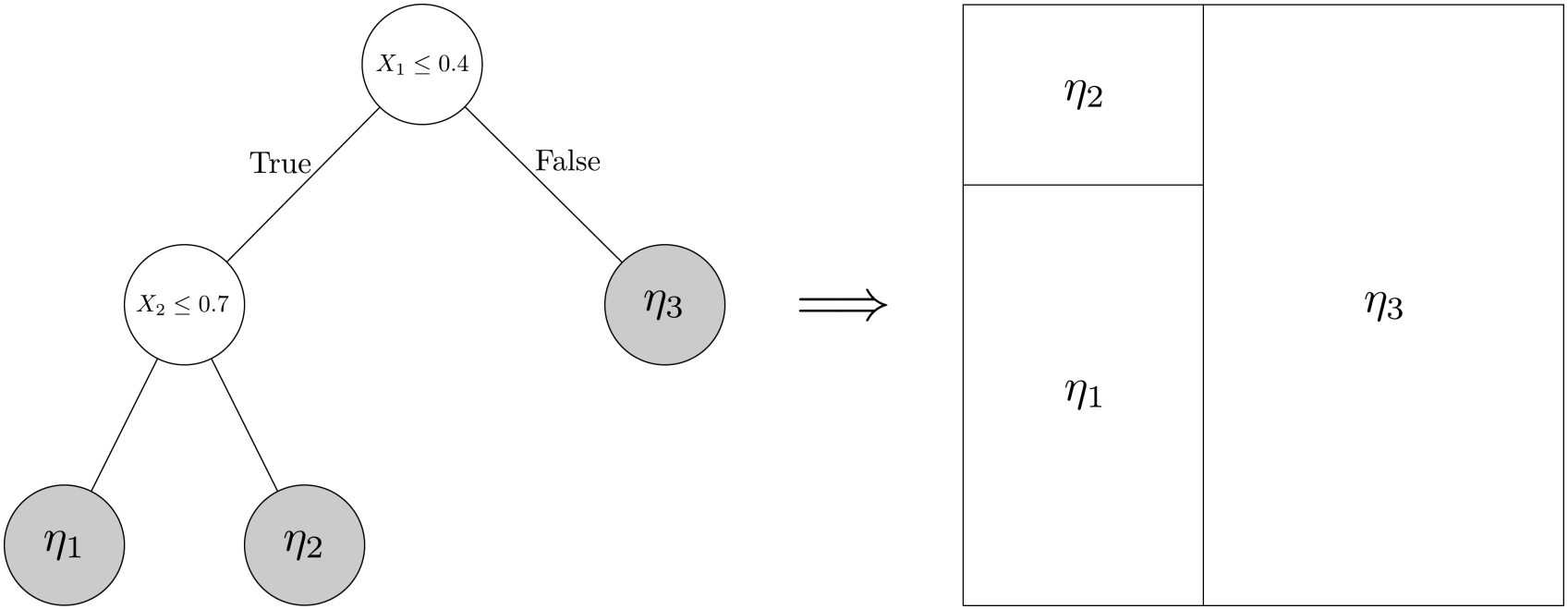
Relevant areas:

- Machine learning
- Causal inference (where μ_0 is often a nuisance parameter)

Decision Trees



Decision Trees



Tree Ensembling

Sampling methods

- Bagging (Breiman, 1996)
- Random forest (Breiman, 2001)
- Bayesian model averaging (Chipman et al., 1998)

Tree Ensembling

Sampling methods

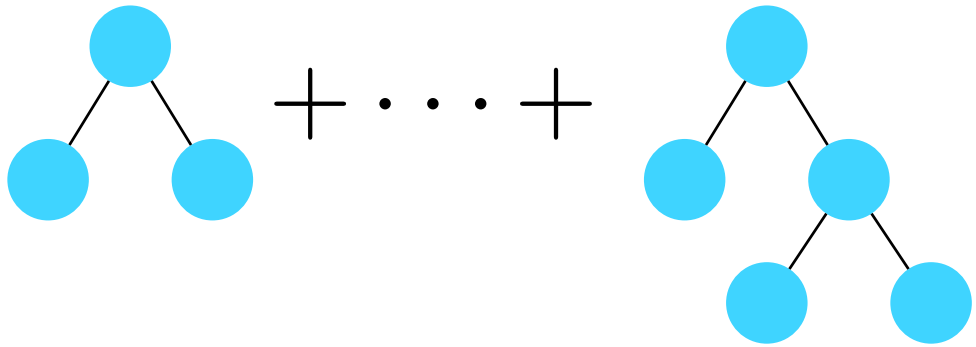
- Bagging (Breiman, 1996)
- Random forest (Breiman, 2001)
- Bayesian model averaging (Chipman et al., 1998)

Combining Weak Learners

- Boosting (Freund et al., 1999)
- BART (Chipman et al., 2010)

Bayesian additive regression trees

$$f(x) = \quad + \cdots +$$

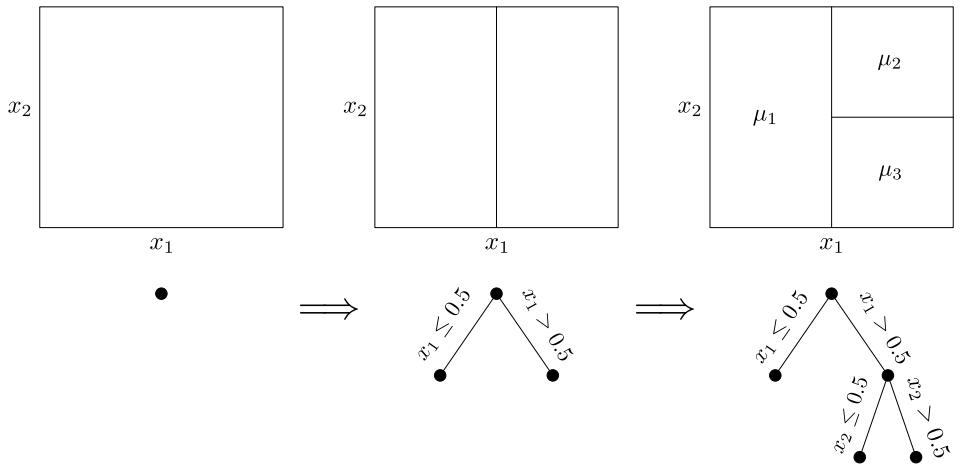


Independent priors placed on $(\mathcal{T}_t, \Theta_t)$ (Chipman et al., 2010).

Increasing in popularity

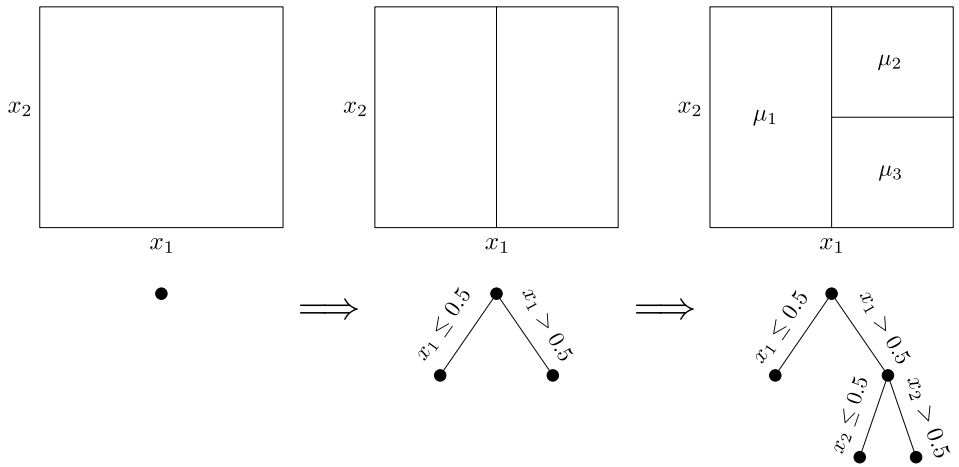
Successful in causal inference (Dorie et al., 2017)

A Prior Distribution on Trees



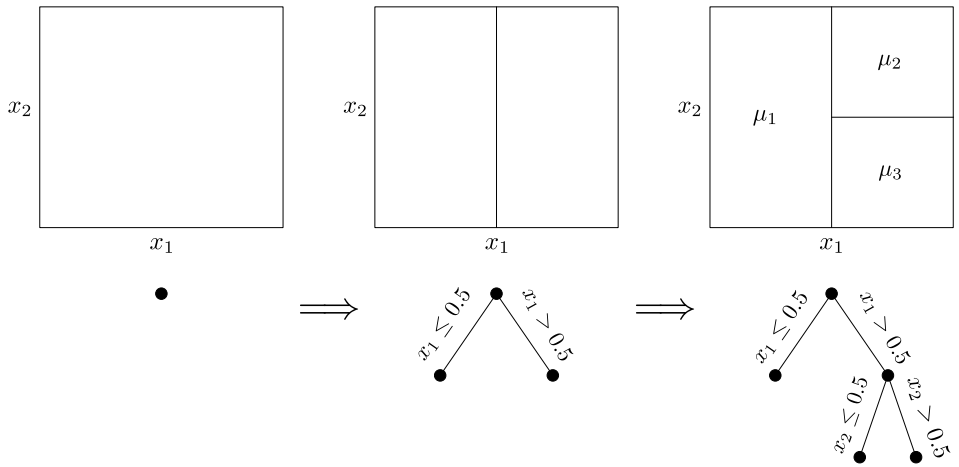
- Prior probability of splitting node n: $\frac{\alpha}{(1+\text{depth}(n))^{\beta}}$ $(\alpha = 0.95, \beta = 2)$

A Prior Distribution on Trees



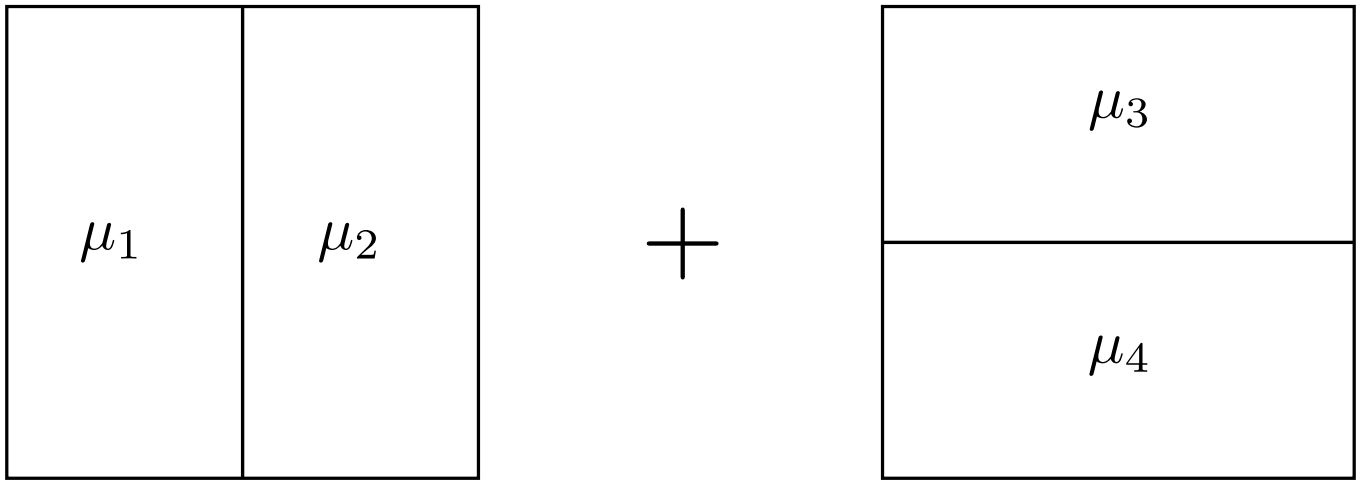
- Prior probability of splitting node n: $\frac{\alpha}{(1+\text{depth}(n))^{\beta}}$ $(\alpha = 0.95, \beta = 2)$
- Prior distribution for the leaf node parameters: $\mu_n \sim \text{Normal}(0, \sigma_\mu^2)$ where $\sigma_\mu^2 = \frac{k^2}{T}$ (k a tuning parameter, usually $k \in \{1, 2, 3\}$ if outcome is standardized, usually I put down a hyperprior).

A Prior Distribution on Trees

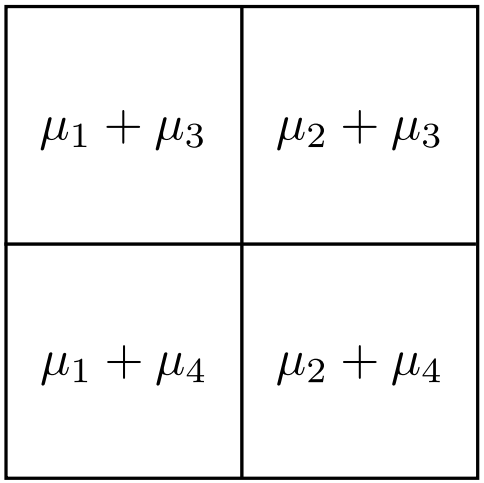


- Prior probability of splitting node n: $\frac{\alpha}{(1+\text{depth}(n))^{\beta}}$ $(\alpha = 0.95, \beta = 2)$
- Prior distribution for the leaf node parameters: $\mu_n \sim \text{Normal}(0, \sigma_\mu^2)$ where $\sigma_\mu^2 = \frac{k^2}{T}$ (k a tuning parameter, usually $k \in \{1, 2, 3\}$ if outcome is standardized, usually I put down a hyperprior).
- For regression, $\sigma^{-2} \sim \text{Gam}(a, b)$ with (a, b) chosen in a default fashion (details aren't worth going into, but it's easy).

Bayesian Additive Regression Trees



$$=$$



Why BART? Why not GPs?

Bart is successful because:

1. It is a *flexible, high quality, prediction engine*.
2. It *provides principled uncertainty quantification*.
3. It is *easy to use.*

Why BART? Why not GPs?

Bart is successful because:

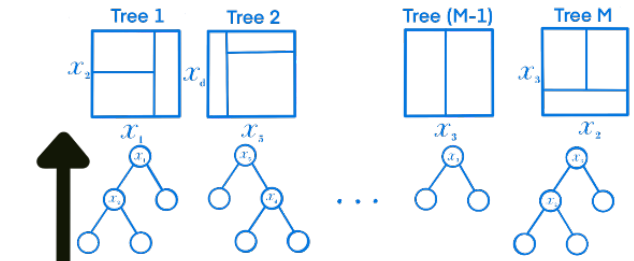
1. It is a *flexible, high quality, prediction engine*.
2. It *provides principled uncertainty quantification*.
3. It is *easy to use.*

Theorem (Heuristic, Linero and Yang, 2018; Rockova and van der Paas 2020). Suppose that $\mu_a(x)$ decomposes additively as

$$\mu_a(x) = \sum_{v=1}^{V} f_v(a, x)$$

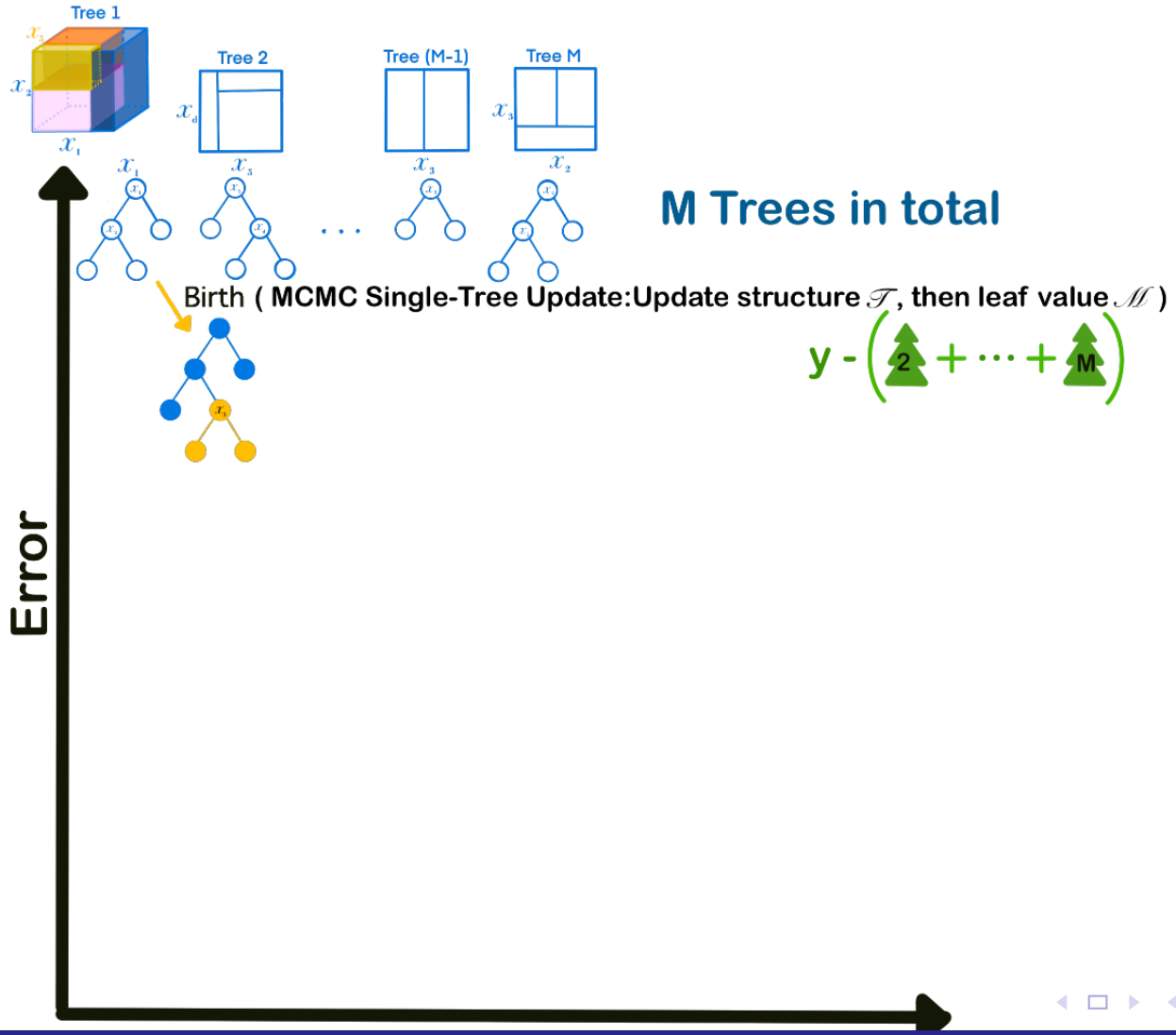
where each f_v is *sparse* (i.e., each f_v represents a low-order interaction). Then BART is optimal in the sense that it attains the minimax-optimal rate of convergence.

Bayesian Backfitting

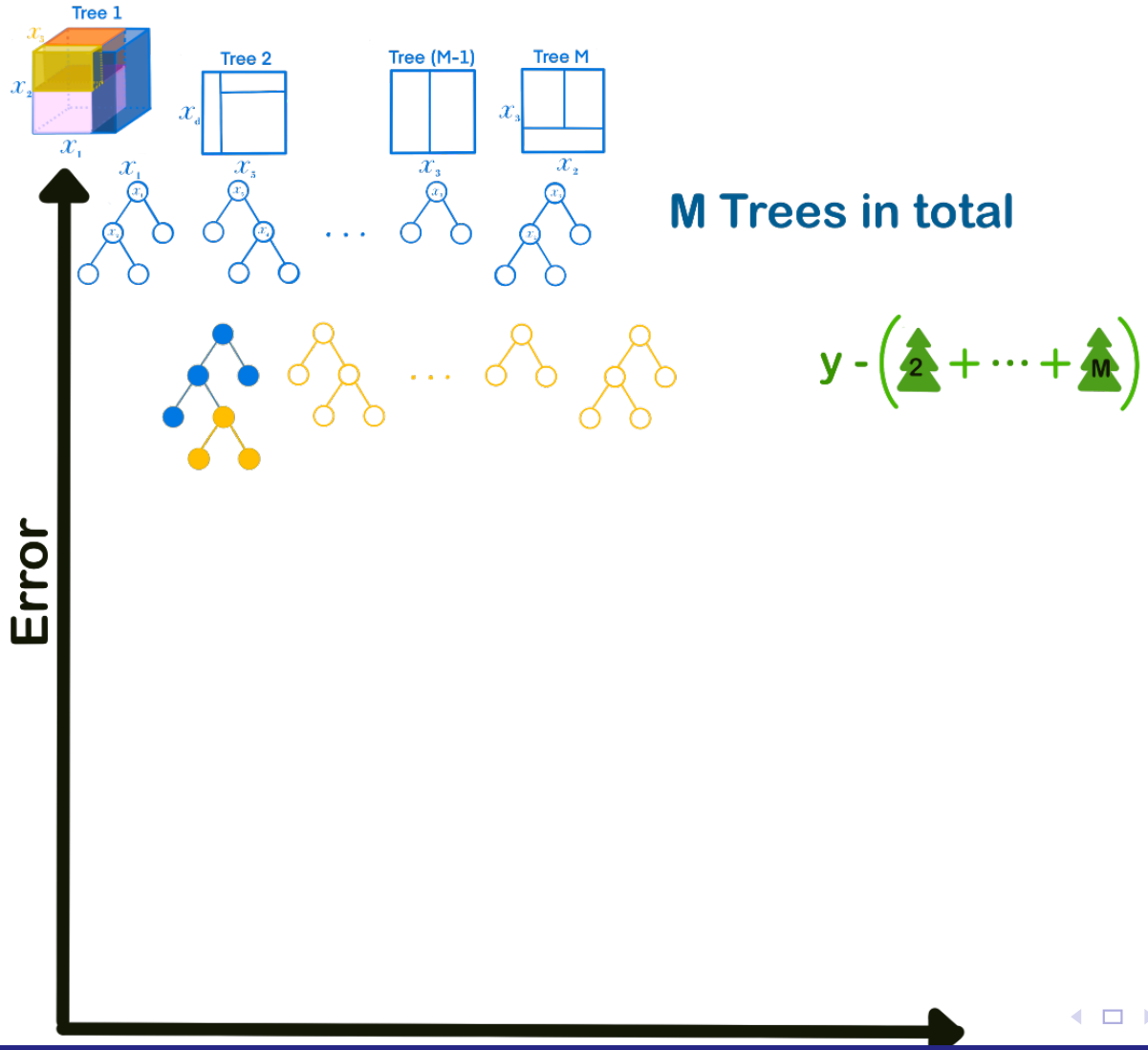


$$y - \left(2 + \cdots + M \right)$$

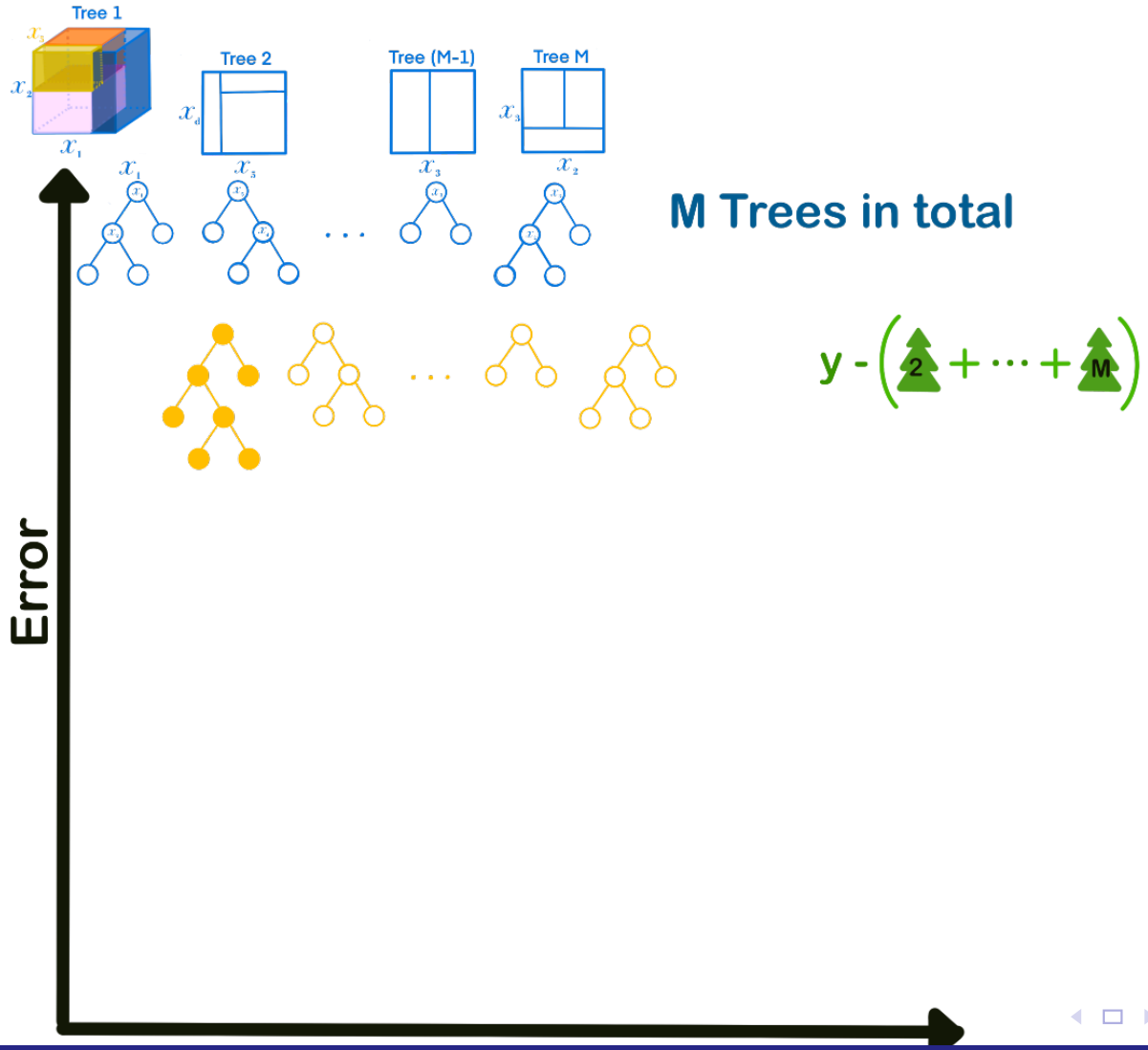
Bayesian Backfitting



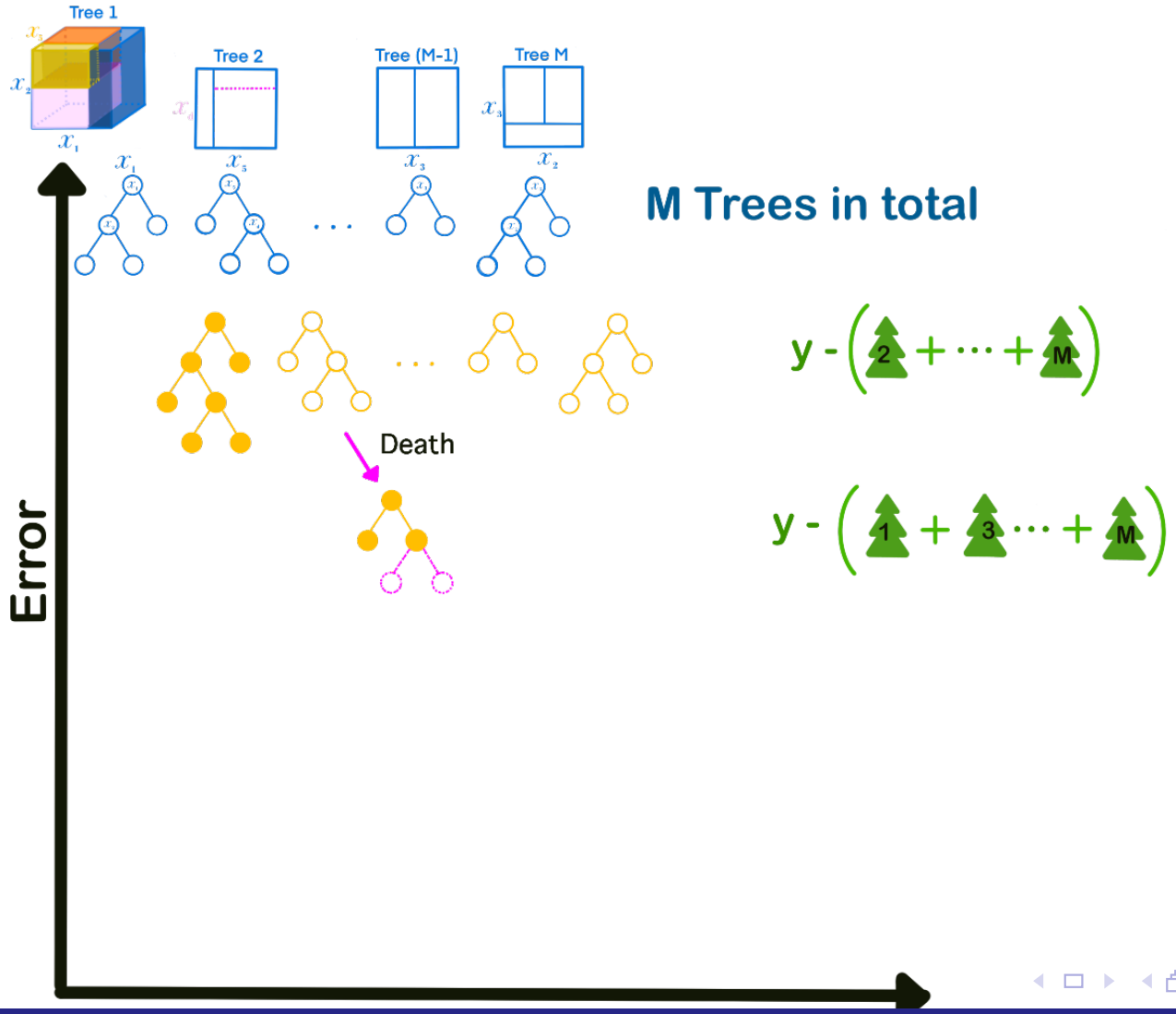
Bayesian Backfitting



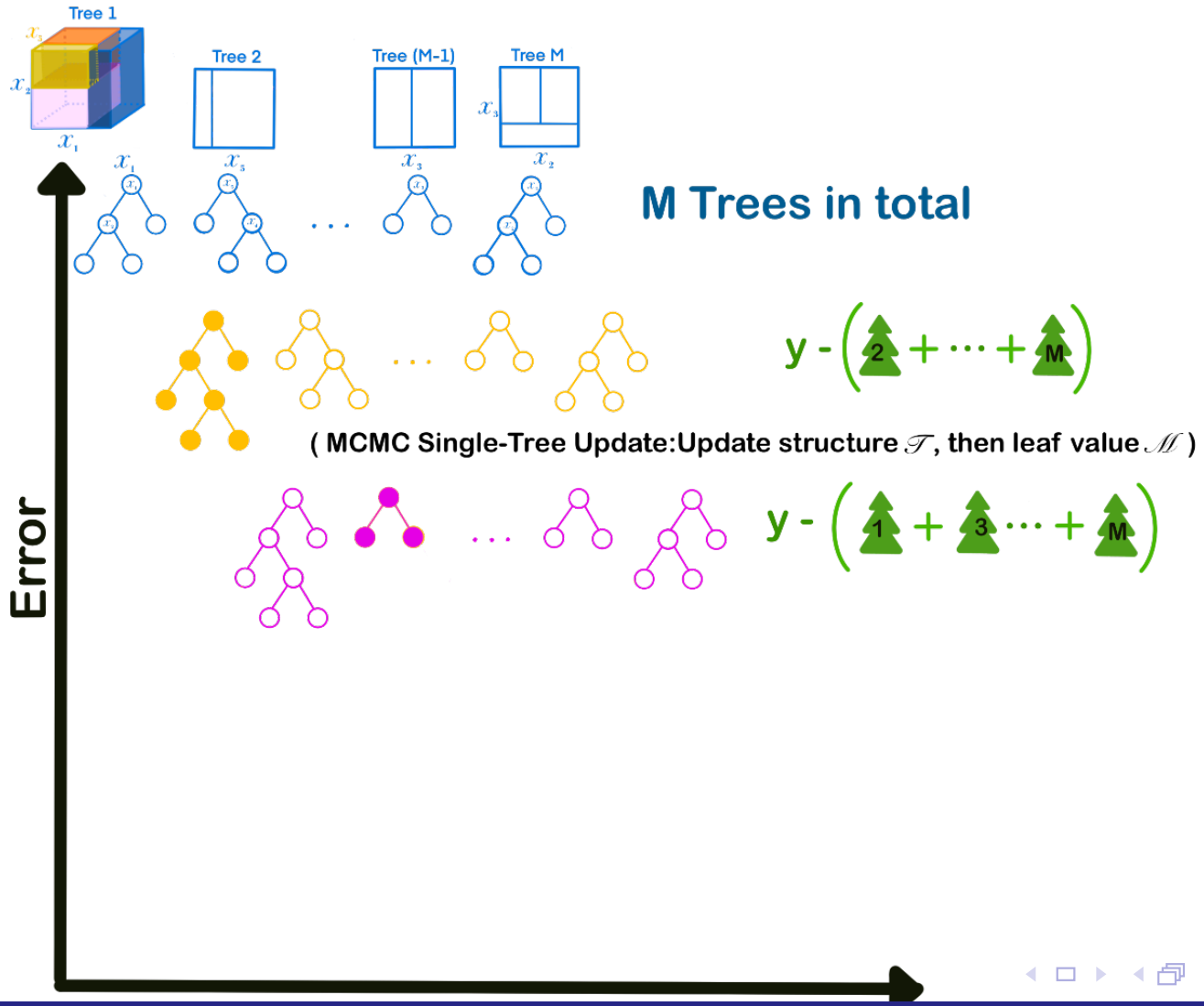
Bayesian Backfitting



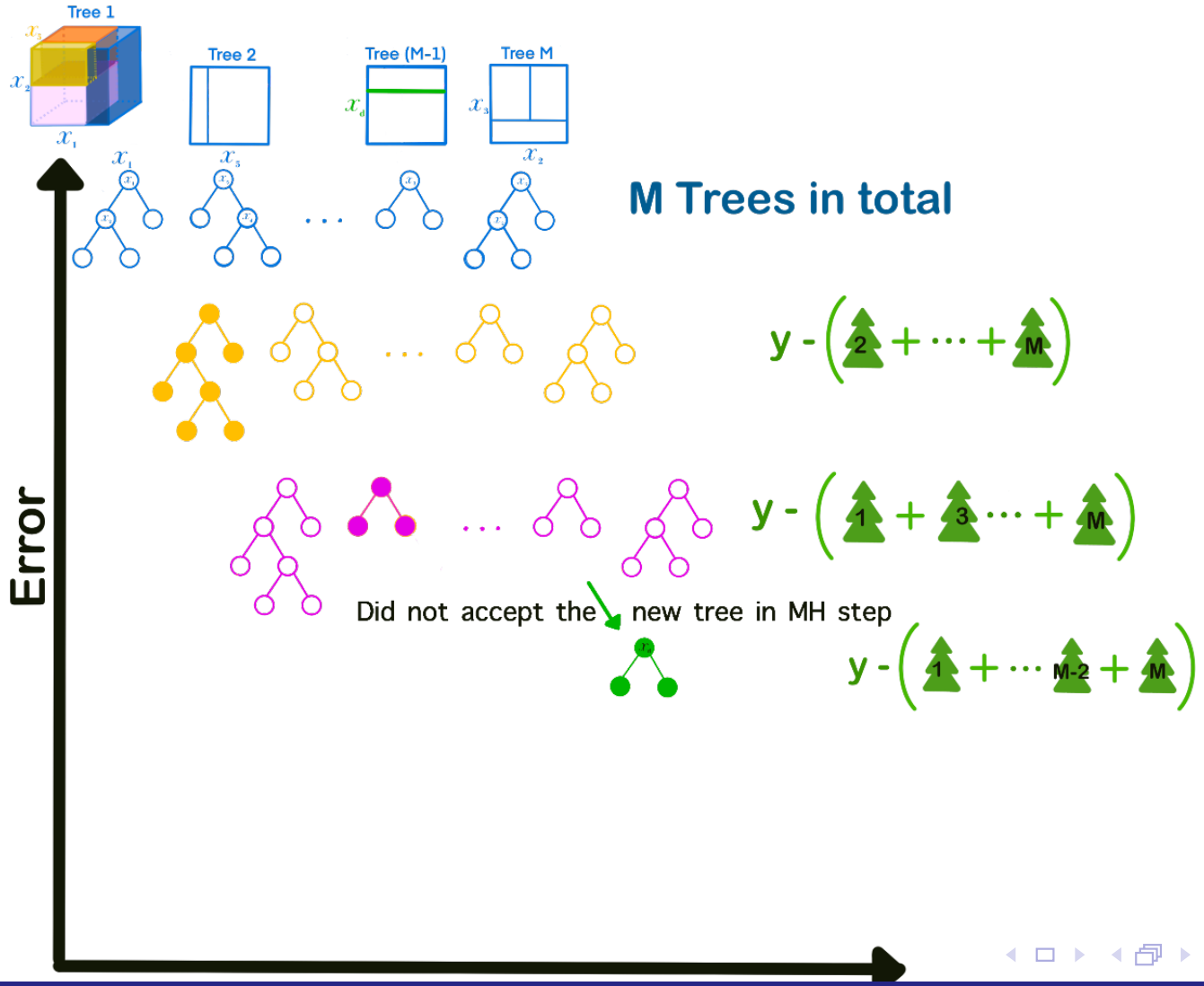
Bayesian Backfitting



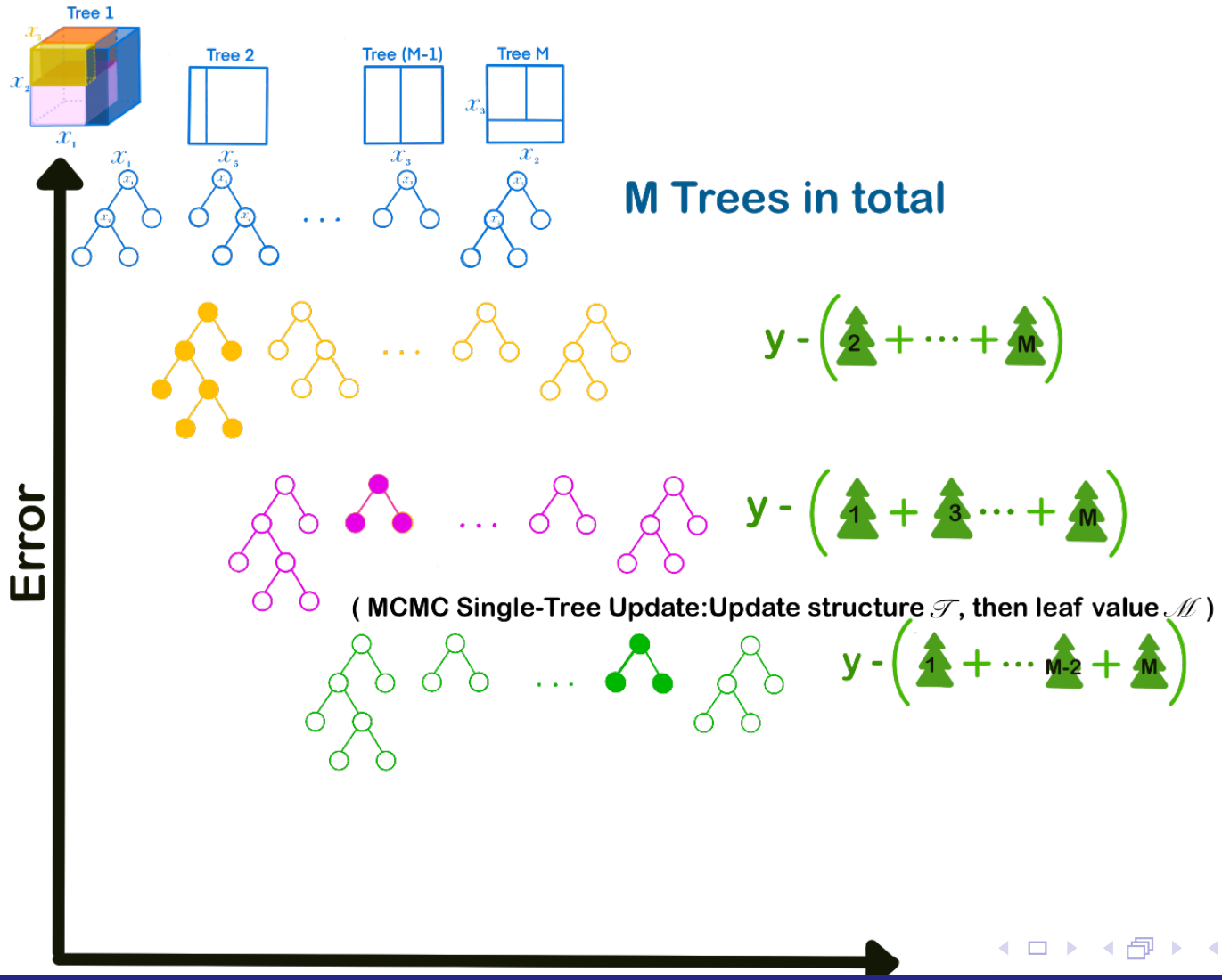
Bayesian Backfitting



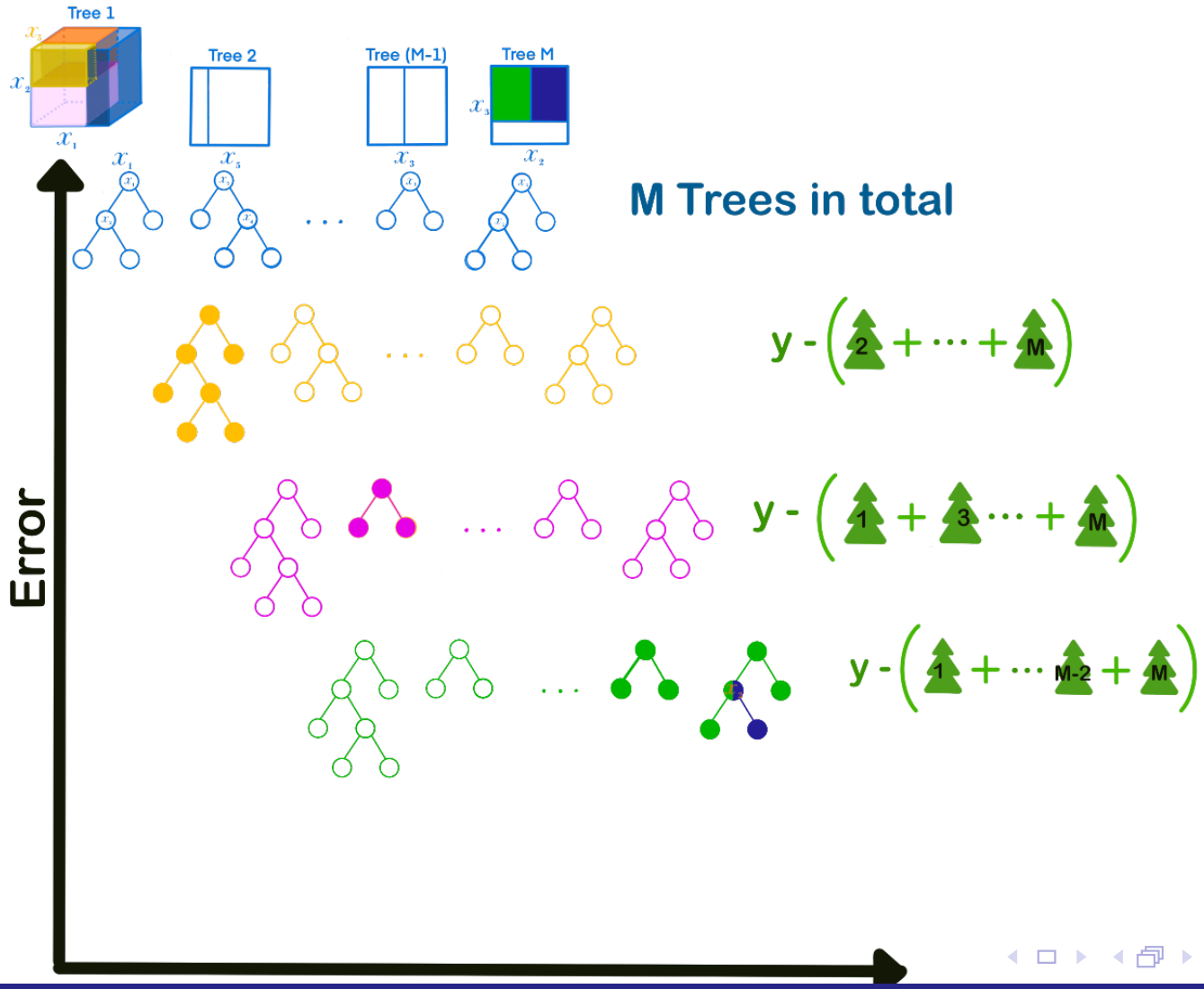
Bayesian Backfitting



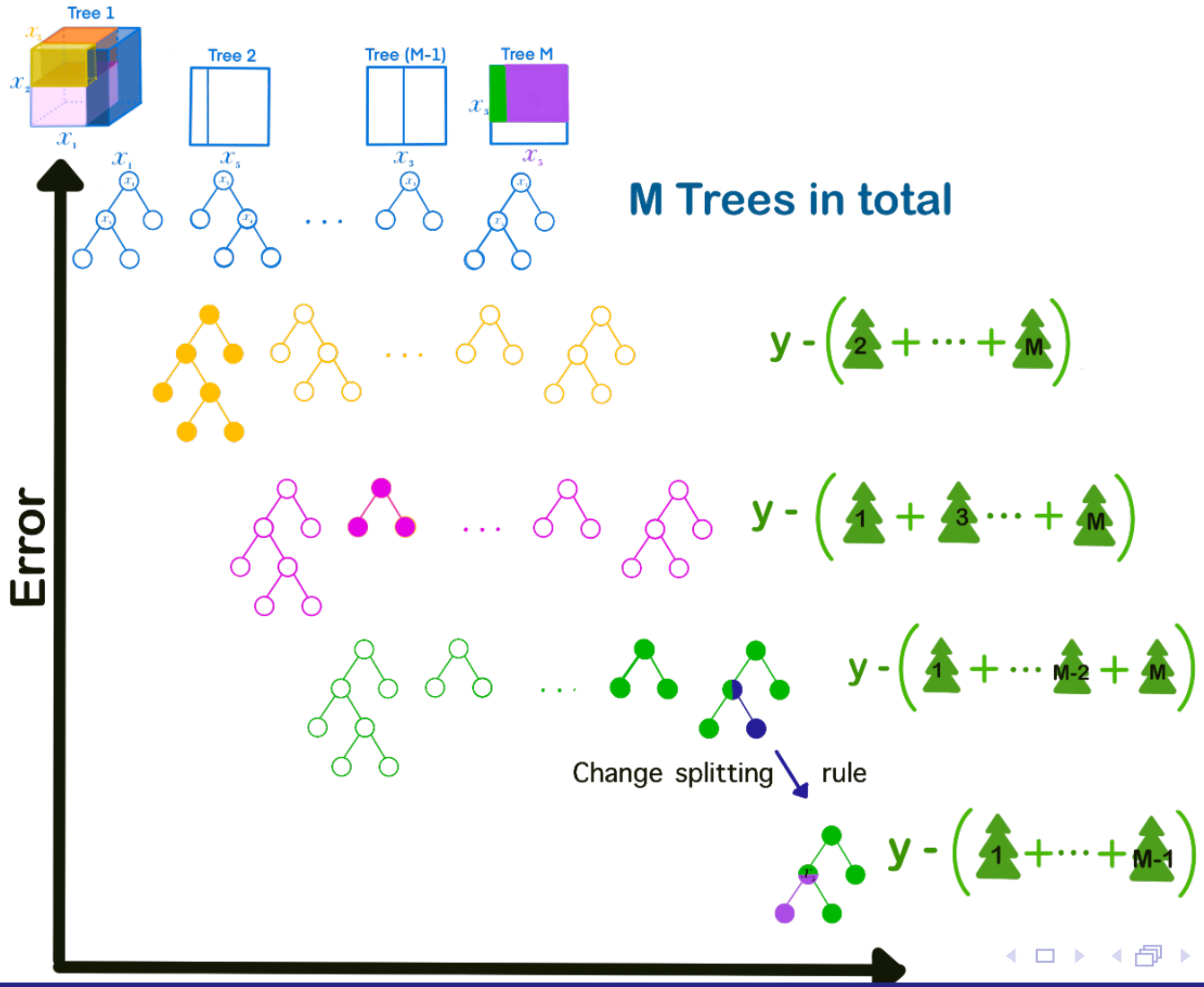
Bayesian Backfitting



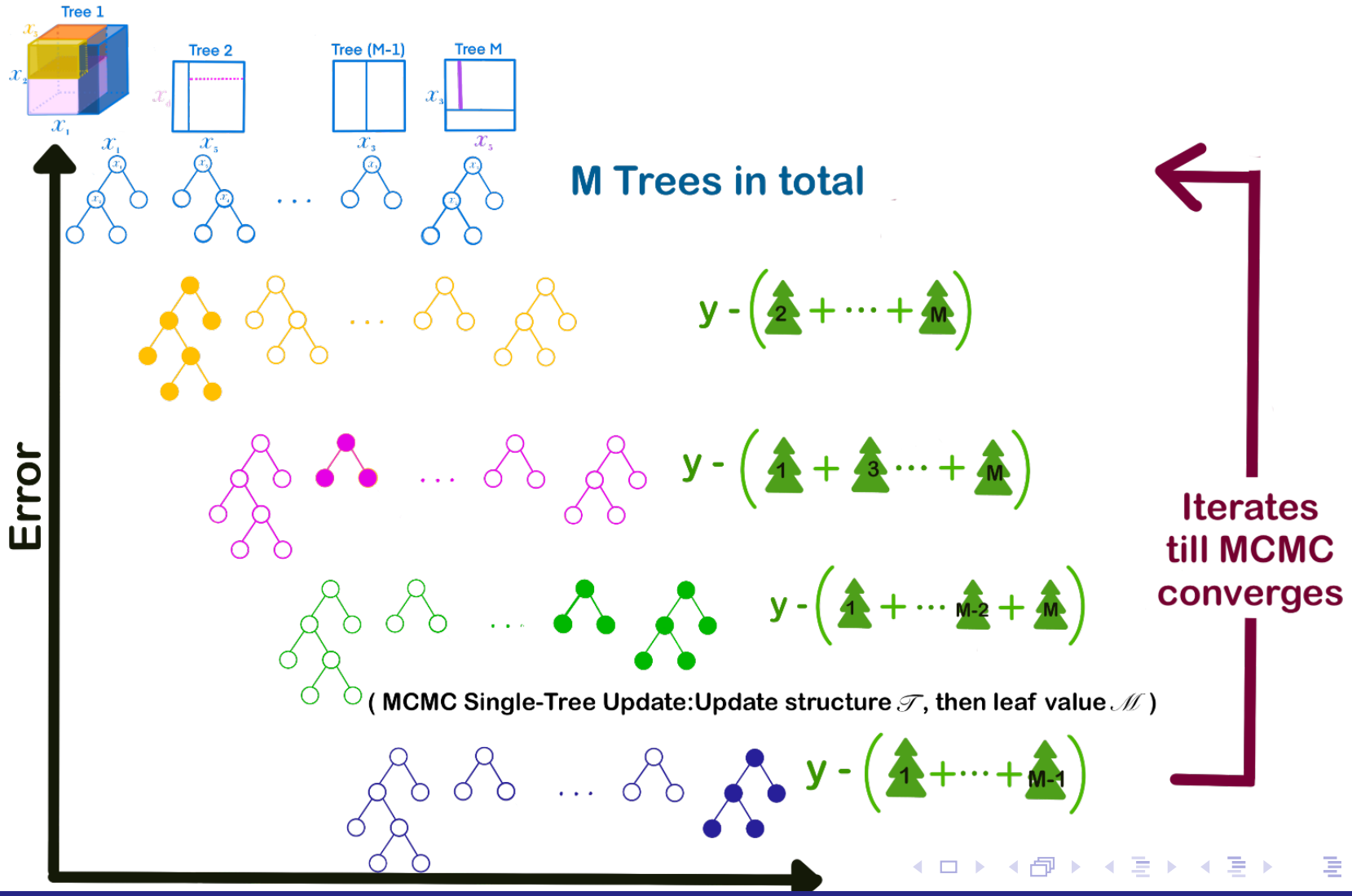
Bayesian Backfitting



Bayesian Backfitting



Bayesian Backfitting

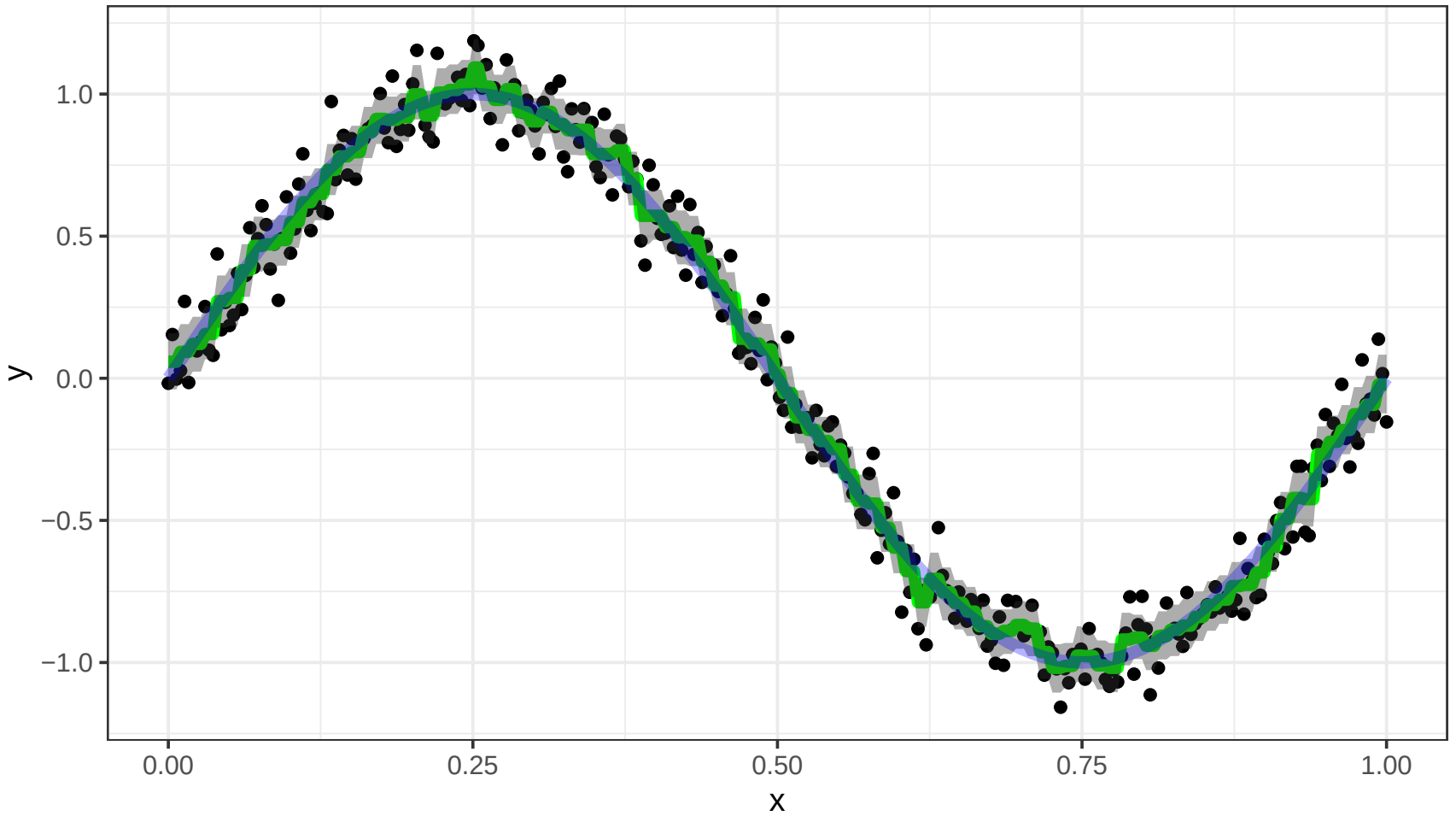


A Simple Illustration

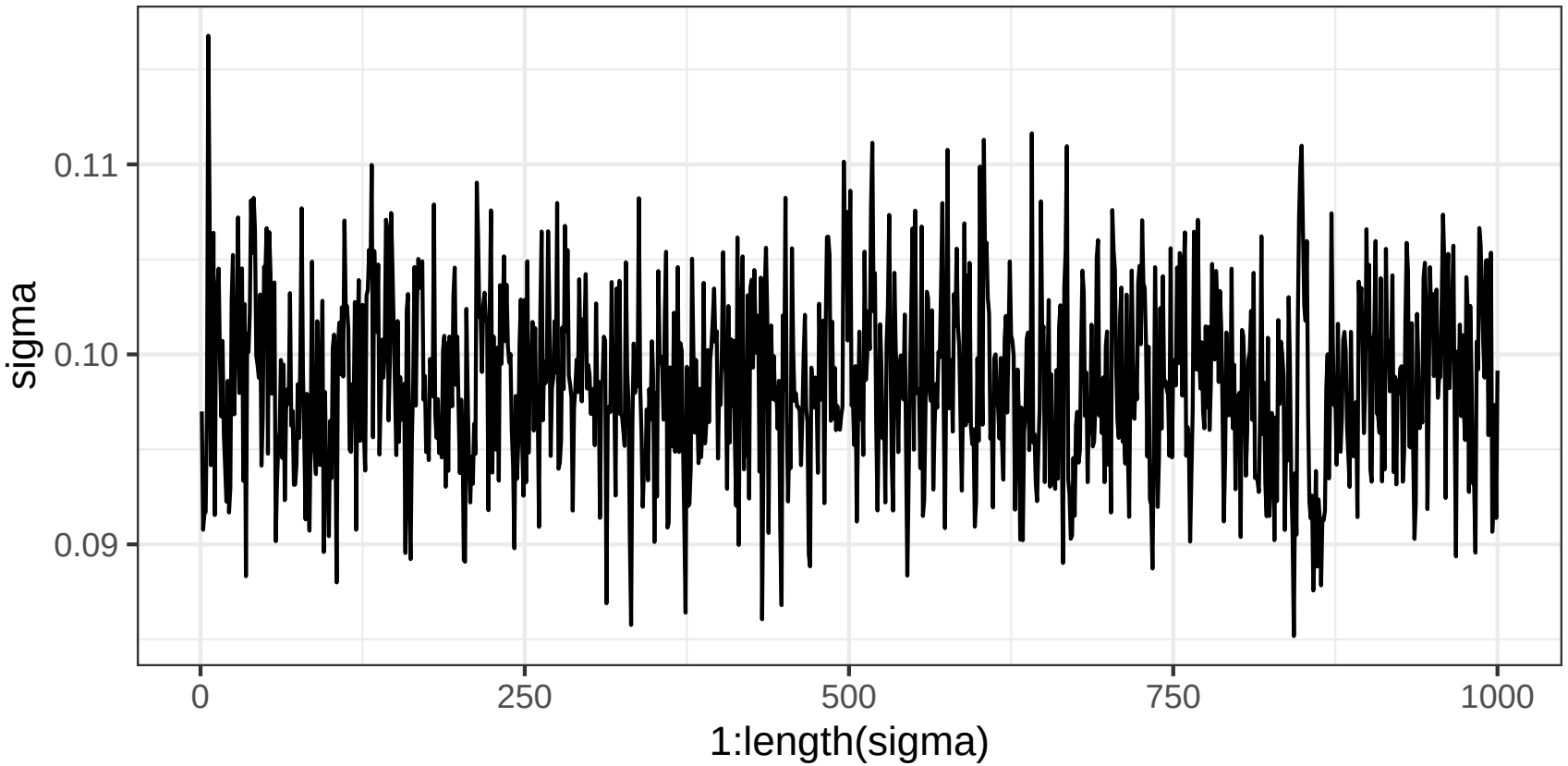
$$Y_i = \sin(2\pi X_i) + 0.1\,\epsilon_i, \qquad \epsilon_i \overset{\text{iid}}{\sim} \text{Normal}(0, 1)$$

Go over code

A Simple Illustration



Mixing



A More Complicated Example

Friedman's Function:

$$\mu_0(x) = 10 \sin(\pi x_1 \, x_2) + 20(x_3 - 0.5)^2 + 10x_4 + 5x_5$$

A More Complicated Example

Friedman's Function:

$$\mu_0(x) = 10 \sin(\pi x_1 \, x_2) + 20(x_3 - 0.5)^2 + 10x_4 + 5x_5$$

- $N = 250$
- If $P > 5$ then $P - 5$ predictors are irrelevant
- We use a sparsity inducing prior (L., 2018 JASA) which encourages the ensemble to reuse the same predictors for different splits rather than selecting uniformly.
- For causally-appropriate variable selection: Caron et al. (2021) or Kim et al. (2022) use this idea.

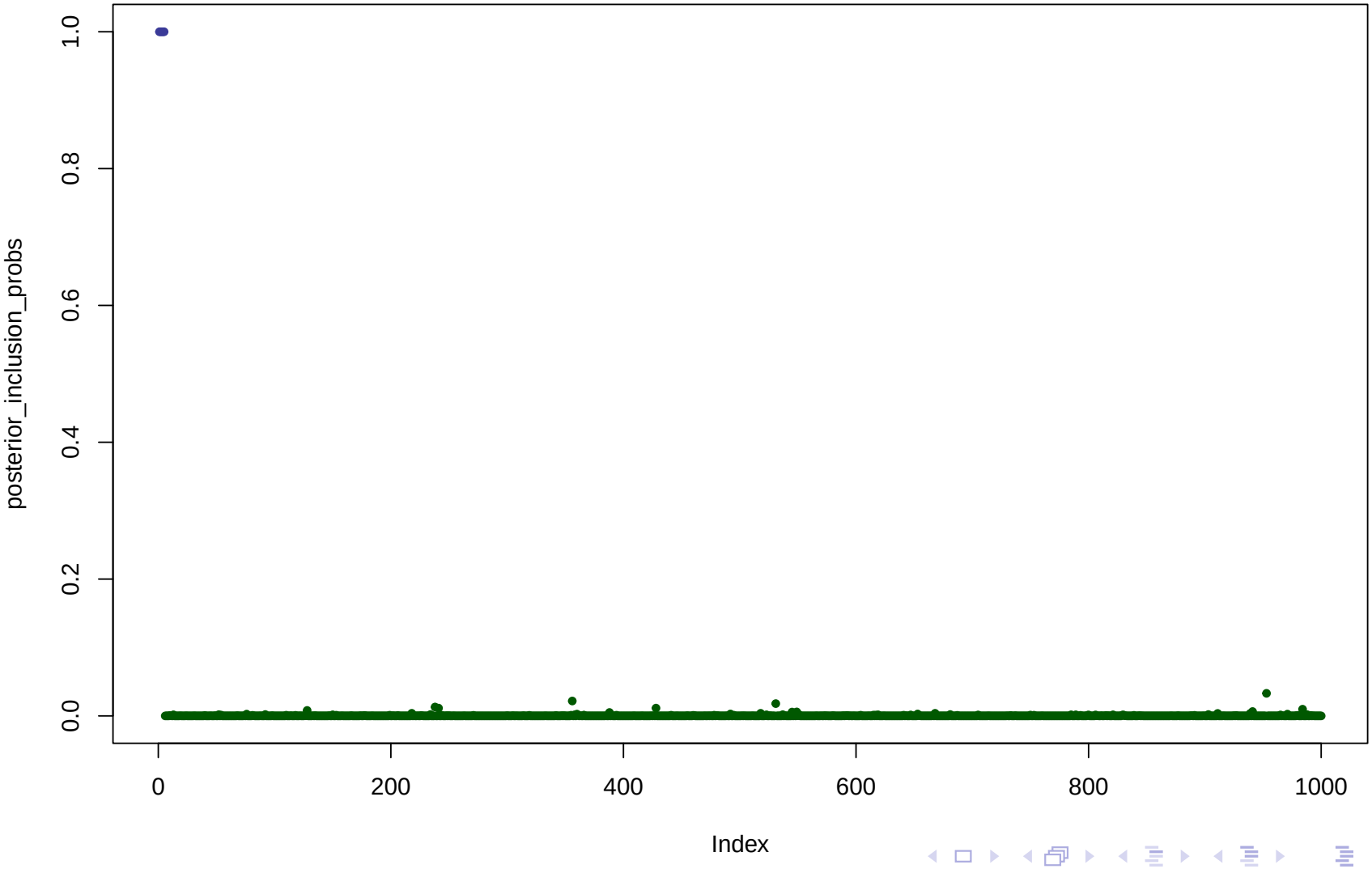
A More Complicated Example

Friedman's Function:

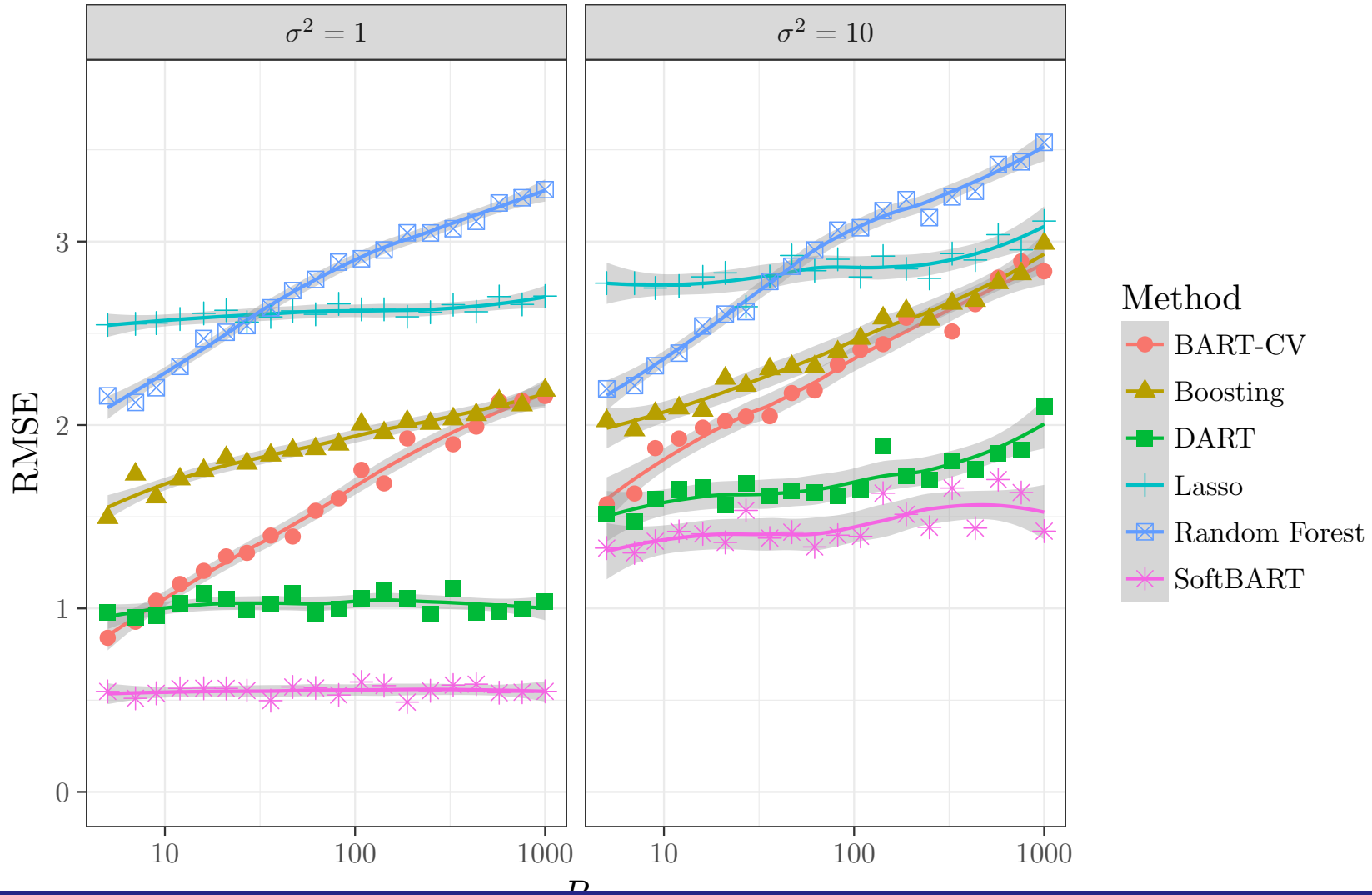
$$\mu_0(x) = 10 \sin(\pi x_1 \, x_2) + 20(x_3 - 0.5)^2 + 10x_4 + 5x_5$$

- $N = 250$
- If $P > 5$ then $P - 5$ predictors are irrelevant
- We use a sparsity inducing prior (L., 2018 JASA) which encourages the ensemble to reuse the same predictors for different splits rather than selecting uniformly.
- For causally-appropriate variable selection: Caron et al. (2021) or Kim et al. (2022) use this idea.
- Go over code

Automatic Relevance Determination



Accuracy for Large Numbers of Predictors



Some Positive Evidence of BART Goodness

Data	BART-CV	DART	SBART	RF	XGB	SBART-CV
ais	**1.00** (1)	**1.00** (1)	**1.00** (1)	1.01 (5)	1.03 (6)	**1.00** (1)
abalone	1.03 (4)	1.03 (4)	**1.00** (1)	1.02 (3)	1.03 (4)	**1.00** (1)
bbb	1.07 (6)	1.04 (4)	**0.99** (1)	1.01 (3)	1.05 (5)	1.00 (2)
cpu	0.98 (2)	1.01 (5)	1.01 (4)	**0.97** (1)	1.02 (6)	1.00 (3)
diamonds	1.15 (4)	1.07 (3)	1.01 (2)	2.29 (6)	1.43 (5)	**1.00** (1)
hatco	1.14 (3)	1.15 (4)	1.10 (2)	1.39 (6)	1.20 (5)	**1.00** (1)
servo	1.02 (3)	1.02 (3)	**0.99** (1)	1.17 (6)	1.06 (5)	1.00 (2)
tecator	1.87 (4)	1.63 (4)	**0.98** (1)	1.95 (6)	1.56 (3)	1.00 (2)
triazines	0.98 (3)	0.99 (4)	0.99 (4)	**0.92** (1)	0.94 (2)	1.00 (6)
wipp	1.19 (4)	1.14 (3)	1.03 (2)	1.43 (6)	1.28 (5)	**1.00** (1)
Average RMPE	1.14 (4)	1.11 (3)	1.01 (2)	1.32 (6)	1.16 (5)	**1.00** (1)
Average Rank	3.4 (3)	3.5 (4)	**1.9** (1)	4.5 (5)	4.6 (6)	2 (2)

Cross-validated estimate of root mean-squared prediction error from 20 replications of 5-fold CV. All scores are relative to the performance of SBART-CV; the rank of each method is given in the parenthesis.

Applying BART in Causal Inference

Consider a nonparametric regression model for an observational study:

$$Y_i(a) = \mu_a(X_i) + \epsilon_i(a).$$

Let $\tau(x) = \mu_1(x) - \mu_0(x)$. Possible targets of inference include:

- The *population average treatment effect*

$$\int \tau(x)\, f_0(x)\; dx.$$

  τ is a nuisance.

Applying BART in Causal Inference

Consider a nonparametric regression model for an observational study:

$$Y_i(a) = \mu_a(X_i) + \epsilon_i(a).$$

Let $\tau(x) = \mu_1(x) - \mu_0(x)$. Possible targets of inference include:

- The *population average treatment effect*

$$\int \tau(x)\, f_0(x)\; dx.$$

  τ is a nuisance.

- The *conditional average treatment effect*

$$\tau(x) = E_0\{Y_i(1) - Y_i(0) \mid X_i = x\}.$$

  τ is of primary interest.

Applying BART in Causal Inference

Consider a nonparametric regression model for an observational study:

$$Y_i(a) = \mu_a(X_i) + \epsilon_i(a).$$

Let $\tau(x) = \mu_1(x) - \mu_0(x)$. Possible targets of inference include:

- The *population average treatment effect*

$$\int \tau(x)\, f_0(x)\, dx.$$

  τ is a nuisance.

- The *conditional average treatment effect*

$$\tau(x) = E_0\{Y_i(1) - Y_i(0) \mid X_i = x\}.$$

  τ is of primary interest.

In either case we need a flexible model for the functions $\mu(x) = \mathbb{E}_\theta\{Y_i(0) \mid X_i = x\}$ and $\tau(x) = \mathbb{E}_\theta\{Y_i(1) - Y_i(0) \mid X_i = x\}$.

Early BART Models for Causal Inference

The first BNP causal inference models for the semiparametric regression model (Hill, 2011) took

$$Y_i = \mu_a(x) + \epsilon_i(a), \qquad \mu_\bullet(\cdot) \sim \text{BART}$$

Call this the one BART model.

Early BART Models for Causal Inference

The first BNP causal inference models for the semiparametric regression model (Hill, 2011) took

$$Y_i = \mu_a(x) + \epsilon_i(a), \qquad \mu_\bullet(\cdot) \sim \text{BART}$$

Call this the one BART model.

Estimation of sample ATE is easy as pie! Sample m from the Markov chain is

$$\text{ATE}^m = \frac{1}{N}\sum_i [\mu_1^m(X_i) - \mu_0^m(X_i)].$$

But this approach has some issues...

What Properties Should a Prior Have?

Special precautions need to be taken when using high-dimensional and nonparametric priors!!

In subjective terms, what prior information do we have?

What Properties Should a Prior Have?

Special precautions need to be taken when using high-dimensional and nonparametric priors!!

In subjective terms, what prior information do we have?

1. We expect (or, at least, do not want to rule out a-priori) the existence of confounding.

What Properties Should a Prior Have?

Special precautions need to be taken when using high-dimensional and nonparametric priors!!

In subjective terms, what prior information do we have?

1. We expect (or, at least, do not want to rule out a-priori) the existence of confounding.
2. We expect (or, at least, do not want to rule out a-priori) that the treatment effect is small relative to the effects of the covariates. It might even be zero.

What Properties Should a Prior Have?

Special precautions need to be taken when using high-dimensional and nonparametric priors!!

In subjective terms, what prior information do we have?

1. We expect (or, at least, do not want to rule out a-priori) the existence of confounding.
2. We expect (or, at least, do not want to rule out a-priori) that the treatment effect is small relative to the effects of the covariates. It might even be zero.
3. We expect (or, at least, do not want to rule out a-priori) that the treatment is *approximately homogeneous* - effect modifiers should exert a small influence, particularly when compared to their main effects. Most controls may not even be effect modifiers.

What Properties Should a Prior Have?

Special precautions need to be taken when using high-dimensional and nonparametric priors!!

In subjective terms, what prior information do we have?

1. We expect (or, at least, do not want to rule out a-priori) the existence of confounding.
2. We expect (or, at least, do not want to rule out a-priori) that the treatment effect is small relative to the effects of the covariates. It might even be zero.
3. We expect (or, at least, do not want to rule out a-priori) that the treatment is *approximately homogeneous* - effect modifiers should exert a small influence, particularly when compared to their main effects. Most controls may not even be effect modifiers.

Naively-specified priors will not necessarily reflect *any* of the above!

Regularization Induced Confounding

Definition: *Regularization induced confounding refers to the phenomenon where using regularization to stabilize estimation of* $\mu_a(x)$ *results in shrinkage of the model towards models where there is no confounding (either measured or unmeasured).*

Regularization Induced Confounding

Definition: *Regularization induced confounding refers to the phenomenon where using regularization to stabilize estimation of $\mu_a(x)$ results in shrinkage of the model towards models where there is no confounding (either measured or unmeasured).*

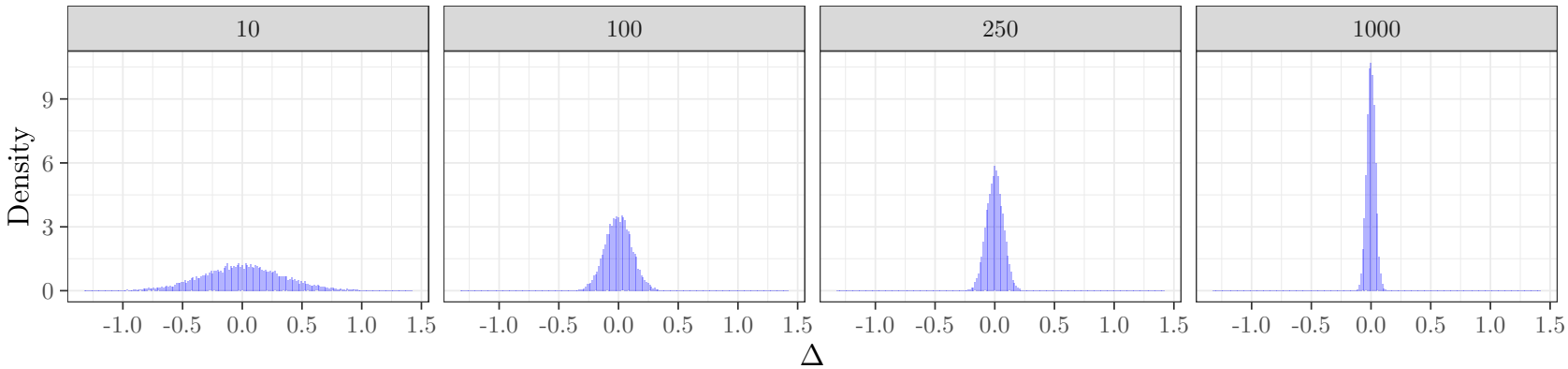
Why does this happen? Consider the models

$$Y_i(a) = X_i^\top \beta + \gamma\, a + \epsilon_i(a) \qquad \text{and} \qquad \Pr(A_i = 1 \mid X_i) = \Phi(x^\top \theta).$$

A "naive" estimate of the causal effect of treatment is $\widehat{E}(Y_i \mid A_i = 1) - \widehat{E}(Y_i \mid A_i = 0)$. Let Δ denote the bias of this estimator, with $\Delta = 0$ corresponding to confounding being irrelevant.

What is the prior on Δ? Assume a normal prior on the regression coefficients.

Regularization Induced Confounding



Bayesian Causal Forests

Take

$$Y_i(a) = \mu(X_i) + a\,\tau(X_i) + \epsilon_i(a)$$

where

- $\mu(\cdot) \sim \text{BART}_T(\alpha, \beta, \sigma_\mu^2)$
- $\tau(\cdot) \sim \text{BART}_{T_\tau}(\alpha_\tau, \beta_\tau, s_\tau^2)$

Also, include an estimate $\widehat{e}_i$ of the propensity score as a predictor in the model.

Bayesian Causal Forests

Take

$$Y_i(a) = \mu(X_i) + a\,\tau(X_i) + \epsilon_i(a)$$

where

- $\mu(\cdot) \sim \text{BART}_T(\alpha, \beta, \sigma_\mu^2)$
- $\tau(\cdot) \sim \text{BART}_{T_\tau}(\alpha_\tau, \beta_\tau, s_\tau^2)$

Also, include an estimate $\widehat{e}_i$ of the propensity score as a predictor in the model.

Model output: directly gives samples of $\tau(X_i)$ for all i, which can be converged to samples of the average causal effect $\tau = \sum_i \omega_i\,\tau(X_i)$ where $\omega \sim \text{Dirichelt}(1, \ldots, 1)$ is BB posterior of covariate distribution.

A Quick Example: The MEPS

The medical expenditure panel survey (MEPS) is an ongoing survey of healthcare uses, medical providers, and insurance companies. Available here.

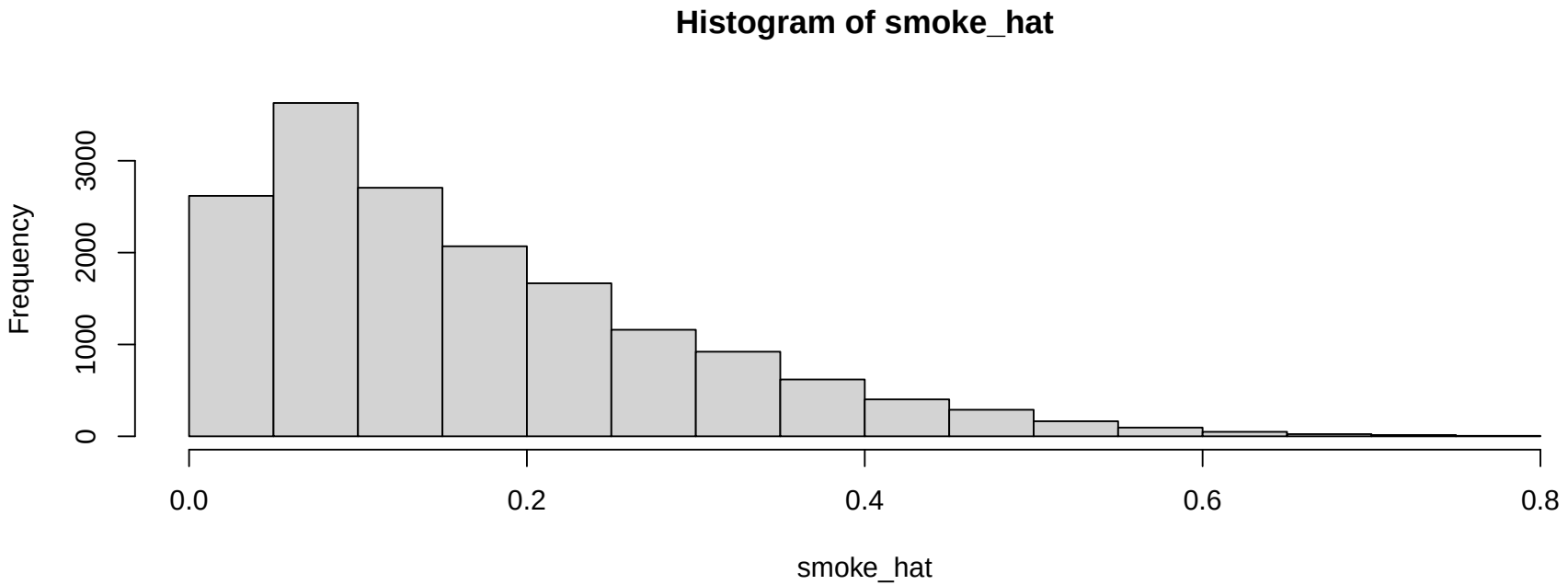
Among individuals who incurred some medical expenditure, what is the causal effect of smoking on the net medical expenditure?

- Y_i is log-medical expenditure
- A_i is whether an individual is a smoker or not
- X_i is a vector of covariates (age, sex, education level, income, seat-belt usage, body mass index)

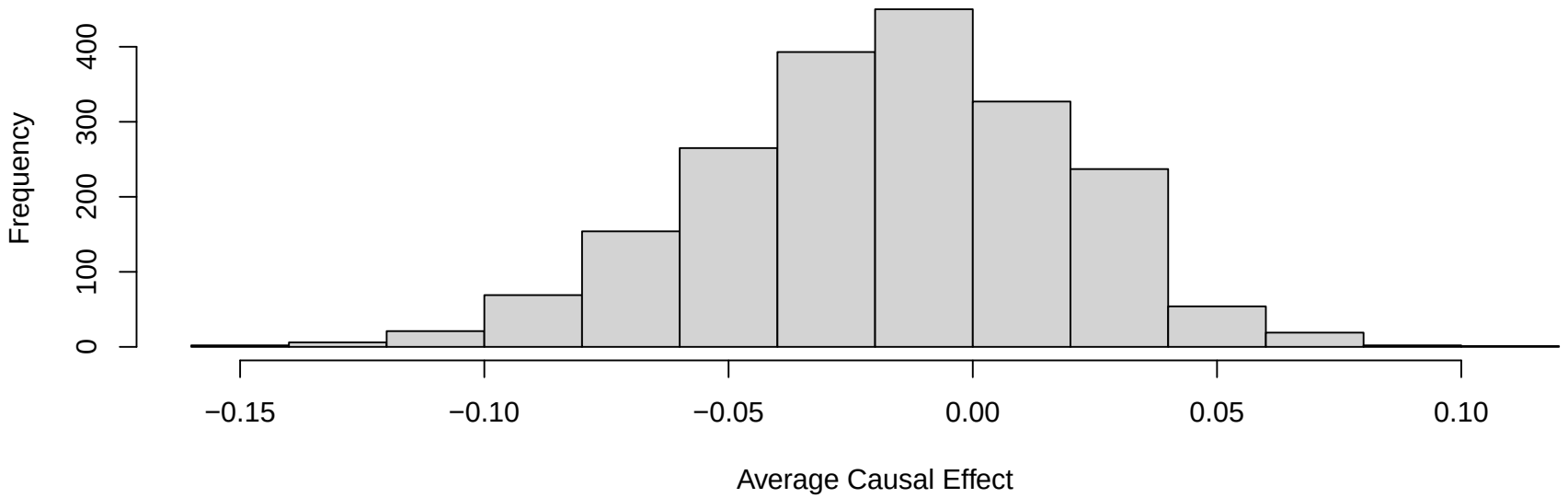
MEPS

age	bmi	edu	income	povlev	sex	marital	race	smoke	phealth	totexp
30	39.1	14	78400	343.69	Male	Married	White	No	Fair	40
53	20.2	17	180932	999.30	Male	Married	Multi	No	Very Good	429
81	21.0	14	27999	205.94	Male	Married	White	No	Very Good	14285
77	25.7	12	27999	205.94	Female	Married	White	No	Fair	7959
31	23.0	12	14800	95.46	Female	Divorced	White	No	Excellent	5017
28	23.4	9	45220	171.07	Female	Separated	PacificIslander	Yes	Excellent	13

Propensity Scores



Posterior Distribution of Treatment Effect

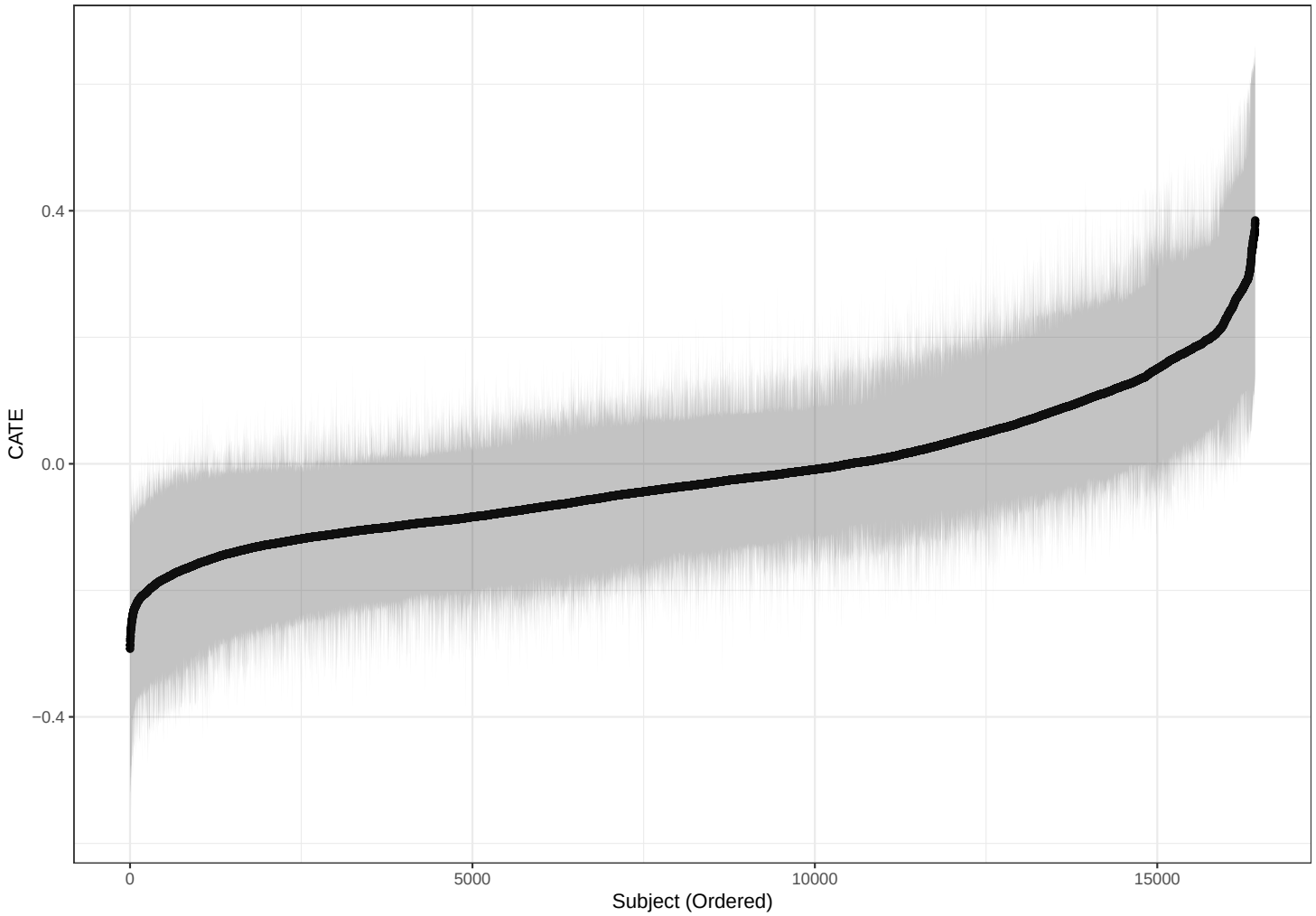


Conditional Average Treatment Effect

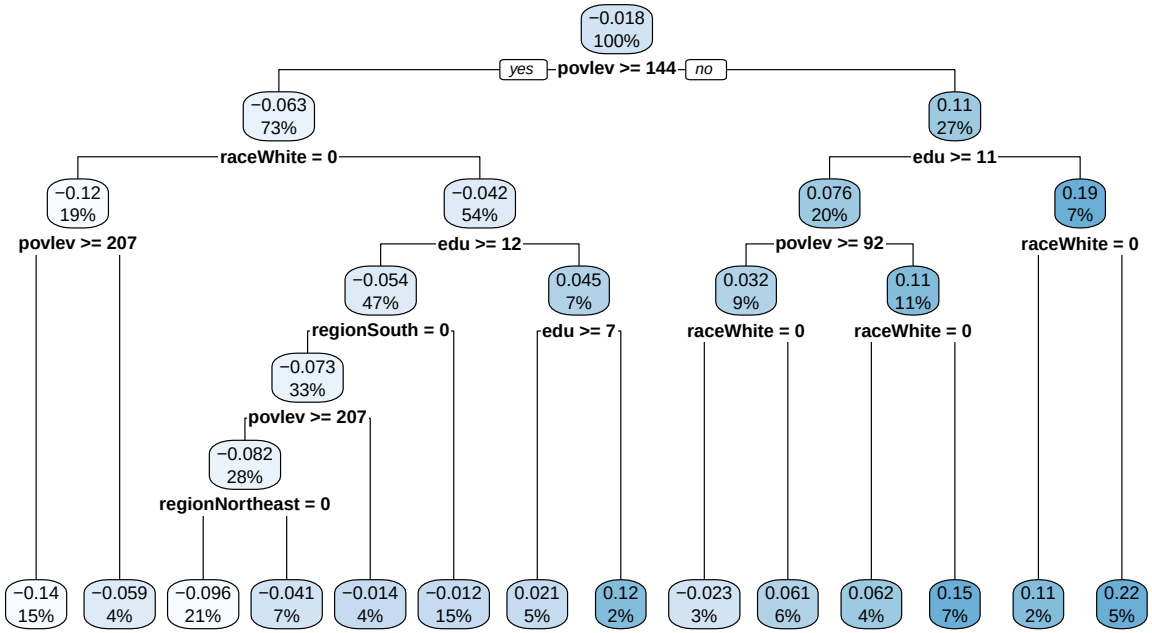
Results are a bit surprising? We'll return to this a bit later when we discuss mediation.

For now, let's see if there are meaningful subgroups where there is evidence of a treatment effect.

Conditional Average Treatment Effect



Posterior Summarization



Model thinks that for higher levels of poverty smoking decreases expenditures and for lower levels of poverty smoking increases expenditures.

In view of previous slide, probably don't want to go too crazy interpreting this. . .

Part 2: Dirichlet Process Mixtures and Extensions

June 26, 2022

Dirichlet process mixtures (DPMs) I

- DPMs for a
  - flexible joint model, e.g., for causal mediation, (Y, M, L)
  - flexible regression model, $Y|L$ (flexible mean AND 'residual'; really just flexible conditional distribution)
- allows for estimation of any (not just mean) causal effects; so any functional of the distribution of potential outcomes
- we will first introduce the Dirichlet process

Dirichlet Processes I

- The *Dirichlet process* (DP) is most easily understood in terms of the *stick-breaking construction*
- Let F be a random probability distribution on a space Θ.
- If F is a Dirichlet process then it can be represented as a countably-supported discrete distribution

$$F = \sum_{j=1}^{\infty} w_j \delta_{Z_j}, \tag{1}$$

where the point-mass distributions δ_{Z_j} are such that $Z_j \sim H$ for some *base distribution* H.

Dirichlet Processes II

- The Dirichlet process $F \sim DP(\alpha, H)$ is defined by (1) and the distribution of the *weights*, w_j.
- The weights, w_j have the following *stick-breaking* form:
  - $w_1 = \beta_1$
  - $w_k = (1 - \beta_1)(1 - \beta_2) \cdots (1 - \beta_{k-1})\beta_k : k \geq 2$

  where β_k are independent Beta$(1, \alpha)$ random variables.

Dirichlet Processes III

- The term 'stick-breaking' comes from the following conceptualization:
  - start with a 'stick' of length 1 and remove $100\beta_1\%$ of the stick and assign this to w_1;
  - then, from the remaining stick of length $(1 - \beta_1)$, break off $100\beta_2\%$ of it and assign this piece to w_2;
  - and so forth.

Dirichlet Processes IV

- If $Y \sim F$ then the weights w_k correspond to the probability,

$$\Pr(Y \neq Z_j \text{ for all } j < k) = (1 - \beta_1) \cdots (1 - \beta_{k-1})$$

  times the probability

$$\Pr(Y = Z_k \mid Y \neq Z_j \text{ for all } j < k) = \beta_k$$

- unfortunately even if H is smooth, F will be discrete (motivates Dirichlet process mixtures of continuous distributions)

Dirichlet Process mixtures (DPM) I

- Dirichlet process mixture of distributions

$$\begin{aligned} Y_i &\sim p(y_i; \theta_i) \\ \theta_i &\sim F \\ F &\sim DP(\alpha, H) \end{aligned}$$

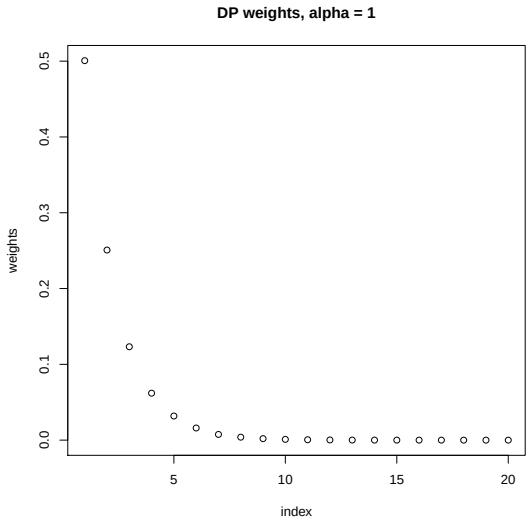
Dirichlet Process mixtures (DPM) II

- this model can also be re-written as an infinite mixture

$$p(y;\boldsymbol{\theta}) = \sum_{j=1}^{\infty} w_j p(y;\boldsymbol{\theta}_j),$$

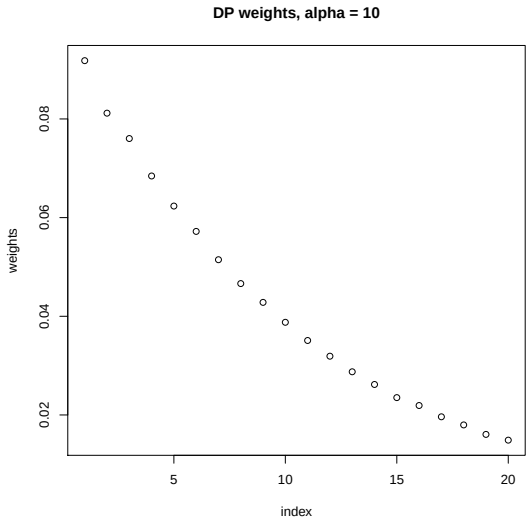
  where $w_j = \beta_j \prod_{l=1}^{j-1}(1-\beta_j)$ and $\beta_j \sim \text{Beta}(1,\alpha)$ and $\theta_j \sim H$.
- note that the weights decay fairly quickly so typically just 'need' first I components of the mixture

Dirichlet Process mixtures (DPM) III



- $\alpha = 1$ - decays very quickly - first 20 components of weights add up to .9999

Dirichlet Process mixtures (DPM) IV



- $\alpha = 10$ - decays more slowly - first 20 components of weights add up to about .85

DPM: Truncation approximation I

$$\sum_{j=1}^{\infty} w_j p(y;\theta_j) \approx \sum_{j=1}^{I} w_j p(y;\theta_j)$$

- recall $w_1 = \beta_1$, $w_j = \beta_j \prod_{l=1}^{j-1}(1-\beta_j)$ for $j = 2, \ldots, I-1$, $\beta_j \sim \text{Beta}(1, \alpha)$ for $j = 1, \ldots, I-1$ and $\beta_I = 1$ (the rest of the stick)
- Ishwaran and James (2001) note how to choose I (function of α) to minimize error
- w_j for the truncation approximation are said to follow a GEM(α) distribution

DPM: Truncation approximation II

- Dirichlet process mixture (DPM) of distributions can then be written as a (finite) latent class model (under the truncation approximation)

$$\begin{aligned} w|\alpha \sim & \quad GEM(\alpha) \\ z_i|w \sim & \quad Mult(w) \\ \theta_k^\star|H \sim & \quad H \\ y_i|z_i, \{\theta_k^\star\} \sim & \quad F(\theta_{z_i}^\star) \end{aligned}$$

where $\theta_i = \theta_{z_i}^\star$

  - suggests way to fit in JAGS or Stan

DPM: Truncation approximation III

- Simple example: Dirichlet process mixture (DPM) of normal distributions

$$
\begin{aligned}
w|\alpha \sim & \quad GEM(\alpha) \\
z_i|w \sim & \quad Mult(w) \\
\theta_k^\star|H \sim & \quad H = N(\mu, \tau^2) \\
y_i|z_i, \{\theta_k^\star\} \sim & \quad N(\theta_{z_i}^\star, \sigma^2)
\end{aligned}
$$

  where $\theta_i = \theta_{z_i}^\star$.
- so just a finite mixture of normals with weights following GEM

Example: DPM of MVN I

- Consider a joint model (Y, L) using DPM of multivariate normals,

$$(Y, L) \sim \sum_{j=1}^{\infty} w_j N(\boldsymbol{\mu}_j, \Sigma_j)$$

where $(\boldsymbol{\mu}_j, \Sigma_j) \overset{iid}{\sim} H$, $w|\alpha \sim GEM(\alpha)$ and H is normal-inverse Wishart distribution

Example: DPM of MVN II

- DPM of normals induces the following conditional distribution of $Y|L$,

$$Y|L \sim \sum_{j=1}^{\infty} w_j(\ell) N(Y|\beta_{0j} + \beta_{1j}\ell, \sigma_j^2)$$

where

$$w_j(\ell) = \frac{w_j N(\mu_{j\ell}, \sigma_{j\ell}^2)}{\sum_{j'=1}^{\infty} w_{j'} N(\mu_{j'\ell}, \sigma_{j'\ell}^2)}$$

and $\beta_{0j} = \mu_{jY} - \frac{\sigma_{jY}}{\sigma_{j\ell}}\rho_j\mu_{j\ell}$, $\beta_{1j} = \frac{\sigma_{jY}}{\sigma_{j\ell}}\rho_j$, $\sigma_j^2 = (1-\rho_j^2)\sigma_{jY}^2$.

  - $E[Y|L]$ is nonlinear and nonadditive in L and a non-normal distribution

Example: Causal inference using the propensity score I

- Recall DPM's allow for estimation of any (i.e., not just average) causal effects which can be expressed in terms of the marginal distributions of the potential outcomes $f\{Y(a) = y\}$
- Using the propensity score in a regression setting, we can specify the joint distribution of the outcome and the propensity score using a DPM of bivariate normals
- We can use the DPM of bivariate normals with (Y, L) replaced by the outcome and the estimated propensity score, $(Y, \hat{e}(L))$.

Example: Causal inference using the propensity score II

- Causal effects can be computed using the distribution of potential outcomes, which are computed using g-computation under ignorability,

$$\Pr\{Y(a) < y\} = \int \int_{-\infty}^{y} f\{t \mid A = a, \hat{e}(\ell)\}\ dt\ F_L(d\ell),$$

- the propensity score, $e(\ell)$ can be estimated nonparametrically using BART
- distribution of the confounders can be estimated using the Bayesian bootstrap (see next slide)

Bayesian bootstrap I

- a simple prior on the distribution $f(\ell)$
- when distributions are not modelled explicitly, the empirical distribution is often used
- The Bayesian bootstrap can be used to incorporate uncertainty in the empirical distribution

Bayesian bootstrap II

- The empirical distribution (implicitly) estimates the distribution of confounders with a multinomial distribution with fixed weight $1/n$ for each observed set of confounders
  - so the support of $f(\ell)$ is assumed to be just the n observed sets of confounders
- The empirical distribution of the confounders can be represented as

$$f_n(\ell) = \sum_{i=1}^{n} \varpi_i \delta_{\ell_i},$$

where δ_{ℓ_i} is a degenerate distribution at ℓ_i and $\{\ell_1, \ldots, \ell_n\}$ are the observed values of the L_i's.

Bayesian bootstrap III

- The Bayesian bootstrap is similar to using the empirical distribution, except that the weights $\varpi = (\varpi_1, \ldots, \varpi_n)$ are now considered unknown parameters and given a non-informative prior $\Pi_{i=1}^{n} \varpi_i^{-1}$.
- The resulting posterior for ϖ is Dirichlet$(1, \ldots, 1)$.
- Given the simple form for the posterior and the finite support of the distribution, integration over the distribution of the covariates only involves computing a weighted average of the observed covariate sets for each sample (of weights) from the Bayesian bootstrap.

Example: Causal inference using g-computation with confounders I

- Causal estimands can also be obtained from a DPM of multivariate normals in the case of a continuous response and continuous covariates (separately for each value of A) using the g-formula,

$$\Pr\{Y(a) < y\} = \int \Pr(Y < y \mid A = a, L = \ell)\ F_L(d\ell),$$

Example: Causal inference using g-computation with confounders II

- $A \sim \text{Bernoulli}(\pi)$ with a Beta prior on π
- the marginal distribution of L takes the form,

$$p(\ell) = \sum_{a} \sum_{j=1}^{\infty} w_j \text{Normal}(\ell \mid \mu_j, \Sigma_j) \pi^a (1-\pi)^{1-a}.$$

- now to sample F_L,
  1. sample π from its posterior
  2. sample A from a Bernoulli distribution given π
  3. sample from $[L \mid A]$
  4. use the sampled L to compute the integral (ignoring the sampled A).

Issues

- there are several restrictions in the previous example
  1. how to address non-continous covariates (including a binary treatment)?
  2. within the components of the mixture, some explicit dependence between Y and L would likely lead to better small sample properties; note this is easily addressed without computational difficulties with a continuous response and covariates (just a multivariate normal) but not with a binary response
- To address these issues, Shahbaba and Neal (2009) introduced a modified DPM of models

Shahbaba and Neal DPM I

- DPM of (multivariate) normals do not easily handle continuous and categorical predictors.
- As an alternative, Shahbaba and Neal introduced the following DPM:

$$
\begin{aligned}
[Y_i \mid L_i, \theta_i] &\sim p(y \mid \ell, \theta_i) \\
[L_{i,j} \mid \omega_i] &\sim p_j(\ell_j \mid \omega_i), j = 1, \ldots, p \\
[(\theta_i, \omega_i) \mid F] &\sim F \\
F &\sim DP(\alpha, H_\theta \times H_\omega).
\end{aligned}
$$

Shahbaba and Neal DPM II

- $p(y \mid \ell, \theta_i)$ is a generalized linear model as opposed to $p(y \mid \theta_i)$ and here the covariates, L_j are now assumed locally (within component) independent.
- This joint DP mixture model is quite flexible allowing the outcome to be continuous or discrete and covariates to be continuous or discrete.
- The local independence of covariates makes it easy to specify conjugate priors.

Example: Binary regression with a continuous and binary covariate I

- Consider a binary regression with two covariates, one continuous and one binary,

$$
\begin{aligned}
[Y_i \mid L_i, \theta_i] &\sim p_y(y \mid \ell, \theta_i) \\
[L_{i,1} \mid \omega_{1i}] &\sim p_1(\ell_1 \mid \omega_{1i}) \\
[L_{i,2} \mid \omega_{2i}] &\sim p_2(\ell_2 \mid \omega_{2i}) \\
[(\theta_i, \omega_i) \mid F] &\sim F \\
F &\sim DP(\alpha, H_{0\theta} \times H_{0\omega})
\end{aligned}
$$

where $\omega_i = (\omega_{1i}, \omega_{2i})$

Example: Binary regression with a continuous and binary covariate II

- p_y is a Bernoulli distribution with success probability, $\Phi(\theta_{1i} + \theta_{2i}\ell_1 + \theta_{3i}\ell_2)$
- p_1 is a normal distribution, $\omega_{1i} = (\mu_i, \sigma_i^2)$,
- p_2 is a Bernoulli distribution, $\omega_{2i} = \pi_i$,
- H_θ is a normal distribution,
- H_ω is the product of a normal-inverse gamma distribution and a Beta distribution.

Example: Binary regression with a continuous and binary covariate III

- this model does not require local independence for all covariates, L.
  - For example, one could still use a multivariate normal for the continuous covariates and/or a large multinomial for the binary covariates or more complex specifications (Murray and Reiter, 2016)

Priors on parameters of base measure

- For computational reasons, conjugate priors for the base measure are preferred.
- and it is preferred for the hyperparameters to be weakly data dependent
- recommendations for different situations can be found in Taddy (2008), Linero and Daniels (2015), and Roy et al. (2018)

Implementation of DPMs I

- R package: *Dirichletprocess*
- implement in JAGS/WinBUGS using finite latent class model formulation based on the truncation approximation
- also implement in Stan using the truncation approximation

DPM models

Advantages:

- Outcome can be continuous or discrete
- Covariates can be continuous or discrete
- Local independence of covariates makes it easy to specify conjugate priors

Problem:

- the likelihood for cluster (mixture component) k is $p(y|\ell;\theta_k)\prod_{j=1}^{p} p(\ell_j;\omega_k)$
- the outcome model gets about $(1/p)$th of the weight of the covariates
  - if p is large prediction model might suffer (important for G-computation)

Enriched DPM model (EDPM)

Wade et al. (2011, 2014) proposed a way to address this issue with an enriched Dirichlet process mixture (EDPM):

$$\begin{aligned} Y_i|L_i,\theta_i &\sim p(y|\ell,\theta_i) \\ L_{i,j}|\omega_i &\sim p(\ell_j|\omega_i), \\ (\theta_i,\omega_i)|F &\sim F \\ F &\sim EDP(\alpha_\theta,\alpha_\omega,H). \end{aligned}$$

$F \sim EDP(\alpha_\theta,\alpha_\omega,H)$ is defined as $F_\theta \sim DP(\alpha_\theta,H_\theta)$ and $F_{\omega|\theta} \sim DP(\alpha_\omega,H_{\omega|\theta})$ with base measures $H = H_\theta \times H_{\omega|\theta}$.

Enriched DP mixture model (cont.) I

- The joint distribution of (Y, L) has the following *square-breaking* construction

$$p(y;\theta) = \sum_{j=1}^{\infty} \left\{ \gamma_j p(y \mid \ell;\theta_j) \sum_{k}^{\infty} \gamma_{k|j} p(\ell;\omega_{k|j}) \right\}$$

where $\gamma_j = \gamma'_j \prod_{\ell=1}^{j-1}(1-\gamma'_\ell)$ and $\gamma'_\ell \sim \text{Beta}(1,\alpha_\theta)$ and $\theta_j \sim H_\theta$ and where $\gamma_{k|j} = \gamma'_{k|j} \prod_{\ell=1}^{k_j-1}(1-\gamma'_{\ell|j})$ and $\gamma'_{\ell|j} \sim \text{Beta}(1,\alpha_\omega)$ and $\omega_{k|j} \sim H_{\omega|\theta}$.

Enriched DP mixture model (cont.) II

- The EDPM induces the following conditional distribution:

$$p(y \mid \ell) = \sum_{j=1}^{\infty} w_j(\ell) p(y \mid \ell, \theta_j), \tag{2}$$

where

$$w_j(\ell) = \frac{\gamma_j \sum_{l=1}^{\infty} \gamma_{l|j} p(\ell \mid \omega_{l|j})}{\sum_{h=1}^{\infty} \gamma_h \sum_{l=1}^{\infty} \gamma_{l|h} p(\ell \mid \omega_{l|h})}.$$

which have similar flexibility to DPM conditionals.

Why EDPM for causal inference? I

Suppose we have data (Y, A, L), where L is $p \times 1$.

- Allows many L-clusters (important for local independence) without having to create additional Y-clusters
- Simple models for L makes it easy to include many covariates
- Imputation/Data augmentation of missing covariates is straightforward (under ignorability)
- for causal settings with many confounders, the EDPM should provide improved inference based on $Y \mid A, L$.

Example: Causal inference with point treatment and many confounders I

- Consider a binary response, q_c continuous covariates and q_b binary covariates where $p = q_c + q_b$.
- The EDPM takes the following form:

$$
\begin{aligned}
[Y_i \mid L_i, \theta_i] &\sim p_y(y \mid \ell, \theta_i) \\
[L_{i,j} \mid \omega_i] &\sim p_c(\ell_j \mid \omega_{ci}) : j = 1, \ldots, q_c, \\
[L_{i,k} \mid \omega_i] &\sim p_b(\ell_k \mid \omega_{bi}), k = q_c + 1, \ldots, q_c + q_b \\
[A_i \mid \omega_i] &\sim p_b(a \mid \omega_{q_c+q_b+1}) \\
[(\theta_i, \omega_i) \mid F] &\sim F \\
F &\sim EDP(\alpha_\theta, \alpha_\omega, H).
\end{aligned}
$$

Example: Causal inference with point treatment and many confounders II

- p_y a Bernoulli distribution with mean $g(L_i^T \theta_i)$, (probit link facilitates computations)
- p_c is a normal distribution with mean μ_{ij} and variance τ_{ij}^2,
- p_b is a Bernoulli distribution with mean π_{ij}
- $\omega_i = (\{\mu_{ij}, \tau_{ij}^2 : j = 1, \ldots, q_c\}, \{\pi_{ij} : j = q_c + 1, \ldots, q_c + q_b\})$.

Example: Causal mediation with a single mediator I

- implement an EDPM with $[Y \mid A, L]$ replaced with $[Y \mid M, A, L]$ and (A, L) replaced by $[M \mid A, L]$ and (A, L).
- The EDPM in this setting takes the following form:

$$
\begin{aligned}
[Y_i \mid M_i, A_i, L_i, \theta_i] &\sim p_y(y \mid m, a, \ell, \theta_i) \\
[M_i \mid A_i, L_i, \omega_i] &\sim p_m(m \mid a, \ell, \omega_{mi}) \\
[L_{i,j} \mid \omega_i] &\sim p_c(\ell_j \mid \omega_{ci}) : j = 1, \ldots, q_c, \\
[L_{i,k} \mid \omega_i] &\sim p_b(\ell_k \mid \omega_{bi}), k = q_c + 1, \ldots, q_c + q_b \\
[A_i \mid \omega_i] &\sim p_b(a \mid \omega_{bi}) \\
[(\theta_i, \omega_i) \mid F] &\sim F \\
F &\sim EDP(\alpha_\theta, \alpha_\omega, H).
\end{aligned}
$$

Example: Causal mediation with a single mediator II

- Under the sequential ignorability assumptions, we can identify the mean potential outcome needed to compute natural direct and indirect effects using a 'mediational' g-formula where $E[Y\{a, M(a')\}]$ is equal to

$$\int E(Y \mid M = m, A = a, L = \ell)\ F_{M|A=a',L=\ell}(dm)\ F_L(d\ell).$$

for $a, a' \in \{0, 1\}$

Example: Causal mediation with a single mediator III

- To compute the NIE, we compute
  $NIE = E[Y\{1, M(1)\}] - E[Y\{1, M(0)\}]$
- With a flexible model specification, under randomized treatment, the total effect should be approximately the difference of the sample means among those randomized to $A = 1$ and those randomized to $A = 0$

$$TE = E[Y\{1, M(1)\}] - E[Y\{0, M(0)\}] \approx \bar{Y}_1 - \bar{Y}_0$$

Example: Causal mediation with a single mediator IV

- We could extend this approach to multiple mediators and a three level extension of EDPM (paper submitted) with M clusters within Y clusters and L clusters within M clusters. This would provide similar advantages to the two level EDPM where the covariates L do not dominate the clustering of the mediators (M) and mediators not dominate clustering of Y.

Summary and Comparison

- joint modeling of outcome and confounders easily handles (ignorable missing) confounders
- DPMs can be used for jointly modeling outcome and confounders
- EDPM better when many confounders for the outcome regression model

Part 2: Dependent Dirichlet Processes and Gaussian Processes

June 26, 2022

Nonparametric regression

Given data $\mathcal{D}_n = \{(X_1, Y_1, \cdots, (X_n, Y_n)\}$ with

$$Y_i = \mu_0(X_i) + \varepsilon_i, \ \varepsilon_i \sim N(0, \sigma_0^2),$$

how do we recover $\mu_0(x)$?

- Earlier saw that $\mu_0 \sim BART$ is one approach
- Gaussian process (GP) priors is an alternative

First, we will motivate GPs by looking at parametric regression

Parametric Bayesian regression

Suppose

$$Y_i = \mu_0(x_i) + \varepsilon_i,\ \varepsilon_i \sim N(0, \sigma_0^2)$$

Parametric model:

$$\mu_0(x_i; \beta) = x^T \beta$$

- linear in x

Priors:

- $\beta \sim \text{MVN}(0, \Sigma_0)$
- $\sigma^2 \sim \text{IG}(a, b)$

Parametric Bayesian regression

Functional form assumed known: $x^T \beta$

- No uncertainty

Priors reflect uncertainty about the values of the parameters β, σ^2

Nonparametric Bayesian regression

$$Y_i = \mu_0(x_i) + \varepsilon_i,\ \varepsilon_i \sim N(0, \sigma_0^2)$$

$\mu_0(x_i)$ unknown, so specify prior distribution for it.

One popular prior for functions is Gaussian process (GP) priors

Gaussian process

Overview:

- μ_0 is a random function
- $\mu_0(x_i)$ at some fixed point $x_i \in \mathcal{R}^p$ is a random variable
- $\mu_0(x_1), \ldots, \mu_0(x_n)$ for some fixed set of points $x_1, \ldots, x_n$ is a random vector

Definition: if the distribution of $\mu_0(x_1), \ldots, \mu_0(x_n)$ is Gaussian for each finite set $x_1, \ldots, x_n$, then μ_0 is a Gaussian process (GP)

If μ_0 is a GP, then for each finite set $x_1, \ldots, x_n$, $\mu_0(x_1), \ldots, \mu_0(x_n)$ has a multivariate normal distribution

GP overview

- Nonparametric: infinitely many parameters characterizing $\mu_0(x)$ when you consider all possible values of x
- We will only work with a finite dimensional object: just the function at the data points $x_1, \ldots, x_n$

Gaussian process

Suppose we have model $y = \mu_0(x) + \varepsilon$

- A linear regression model assumes $\mu_0(x) = x^T\beta$
- A Gaussian process model involves specifying a Gaussian distribution for the unknown function $\mu_0(x)$

Gaussian process: $\mu_0 \sim GP(m, k)$

- m is a mean function and k is a covariance function

Gaussian process

Suppose we have points $x_1, \cdots, x_n$. Then, the values of μ at those points is normally distributed. That is,

$$\mu_0(x_1), \cdots, \mu_0(x_n) \sim N\{(m(x_1), \cdots, m(x_n)), K(x_1, \cdots, x_n)\}$$

You could think of m as your prior guess as to the form of the mean function, and k as capturing your uncertainty about it.

We have to choose m and k

Choice of m and k

If we set $m(x) = x^T\beta$, then our prior guess for μ is a linear model.

A popular choice for k is

$$k(x_i, x_j) = \eta \exp\left(-\sum_{k=1}^{p} \rho_k^R |x_{ik} - x_{jk}|^R\right) + b\delta_{ij}$$

where

- η and ρ are parameters
- $0 < R \leq 2$
- b is a small value (e.g., 0.01)
- δ_{ij} is an indicator function taking value of 1 if $i = j$.

Covariance matrix k example

$$k(x_i, x_j) = \eta \exp\left(-\rho||x_i - x_j||^2\right) + 0.01\delta_{ij}$$

- $\text{var}(\mu_0(x_i)) = \eta + 0.01$
  - Large value of η implies μ very different from linear
  - But, η penalized in the likelihood: $\log|k(x)|$
- ρ affects the degree to which the means of subjects who have similar x will have similar $\mu_0(x)$
- $\text{cov}(x_i, x_j) = \eta$ if $x_i = x_j$, $i \neq j$ (this is why the δ term is needed)
- If the x's are all binary, then $||x_i - x_j||^2$ is a count of the number of covariates where subjects i and j have different values

Posterior

If $y_i \sim \mu_0(x_i) + \varepsilon_i$, where $\varepsilon_i \sim N(0, \sigma_0^2)$ and $\mu_0 \sim GP(m, k)$, then the posterior for μ_0 is also a GP.

Suppose we are interested in the posterior of $f(\widetilde{x})$ for some new set of points $\widetilde{x}$. Denote by $\widetilde{\mu}_0$, $\mu_0(\widetilde{x})$

$$\begin{pmatrix} y \\ \widetilde{\mu}_0 \end{pmatrix} \sim N\left(\begin{pmatrix} m(x) \\ m(\widetilde{x}) \end{pmatrix}, \begin{pmatrix} k(x,x) + \sigma_0^2 I, k(\widetilde{x}, x) \\ k(x, \widetilde{x}), k(\widetilde{x}, \widetilde{x}) \end{pmatrix} \right).$$

Therefore, $\widetilde{f}|x, y, \eta, \rho, \sigma$ is distributed as normal with mean

$$m(\widetilde{x}) + k(\widetilde{x}, x)[k(x,x) + \sigma^2 I]^{-1}(y - m(x))$$

and variance

$$k(\widetilde{x}, \widetilde{x}) - k(\widetilde{x}, x)[k(x,x) + \sigma^2 I]^{-1} k(x, \widetilde{x})$$

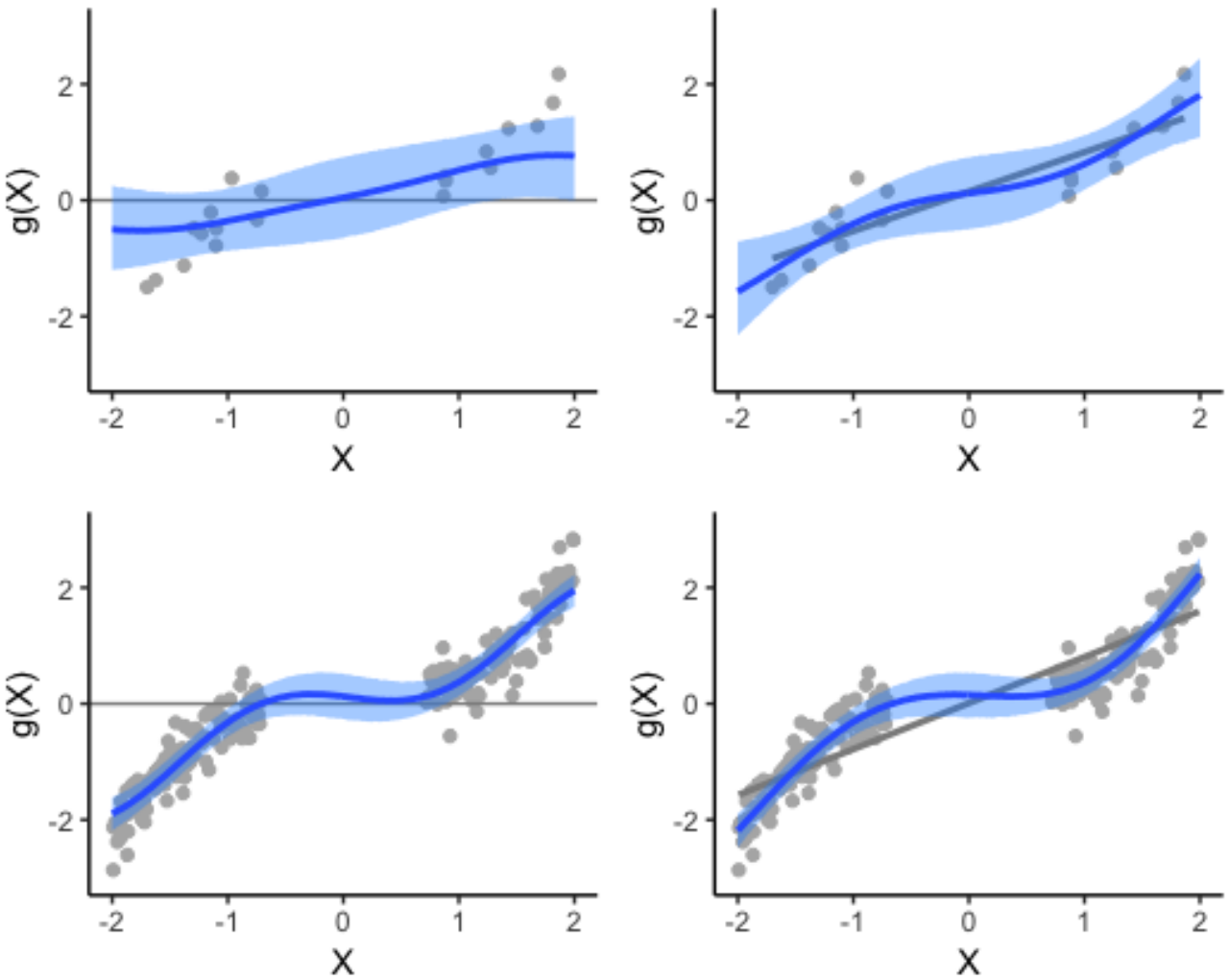
Example

Suppose truth is $y = 0.3x^3 + \varepsilon$

Instead we fit model

$$
\begin{aligned}
y &= \mu_0(x) + \varepsilon \\
\mu_0(x) &\sim GP(m(x), K(x)) \\
k(x_i, x_j) &= \exp\left(-\rho(x_i - x_j)^2\right) + 0.01\delta_{ij} \\
p(\beta_0, \beta_1, \sigma^2) &\sim N(0, s) \times N(0, s) \times IG(a, b)
\end{aligned}
$$

Plot on next slide for $n = 20$ and $n = 200$. In one case we set $m(x) = \beta_0 + \beta_1 x$ and in another $m(x) = 0$



Plots on left prior mean 0; plots on right prior mean linear model

Motivation

Previously we showed how DP priors can be used to estimate marginal or joint distributions.

- $F \sim DP(\alpha, H)$

A condition distribution could be estimated indirectly (from the joint).

Now suppose we would like to directly estimate a conditional distribution, $p(y|x)$.

- We are interested in a collection of distributions $\{P_x : x \in \mathcal{X}\}$

Infinite mixture

Recall that we can write a DP mixture

$$
\begin{aligned}
Y_i &\sim p(y_i; \theta_i) \\
\theta_i &\sim F \\
F &\sim DP(\alpha, H)
\end{aligned}
$$

as

$$
p(y; \theta) = \sum_{j=1}^{\infty} w_j p(y; \theta_j)
$$

where $w_j = \gamma_j \prod_{l=1}^{j-1}(1 - \gamma_l), \gamma_j \sim$Beta$(1, \alpha)$, and $\theta_j \sim H$

DDP as infinite mixture

Similarly, we can write a dependent DP as an infinite mixture

$$p(y|x;\theta) = \sum_{j=1}^{\infty} w_j(x) p(y; g_j(x))$$

where $g_j(x)$ is a regression function

- fixed weight DDPs: $w_j(x) = w_j$ - weights do not depend on x
- recall deriving from a DPM for (y, x) gave $w_j(x)$ with 'known' $g_j(x)$

DDP-GP

Because the form of the regression function $g_j(x)$ is unknown, we could specify a Gaussian process prior for it.

Thus, the conditional distribution of $p(y|x)$ can be specified with a DDP (distribution of outcome around mean) and a GP (for mean function)

Continuous outcome example

$$p(y|x;\theta) = \sum_{j=1}^{\infty} w_j N(y; g_j(x), \sigma_j^2)$$

$$g_j \sim GP(m, k)$$

$$m(x) = x^T \beta$$

$$k(x_i, x_j) = \eta \exp\left(-\rho||x_i - x_j||^2\right) + 0.01\delta_{ij}$$

$w_j = \gamma_j \prod_{l=1}^{j-1}(1-\gamma_l), \gamma_j \sim$Beta$(1, \alpha)$, and $\sigma_j^2 \sim IG(a, b)$
$\beta_j \sim N(0, s)$
(also priors for η, ρ, α)

DDP+GP for Causal Inference

- Data: Outcome Y_i, treatment A_i, confounders L_i
- interest is in average causal effect

$$\Delta = \int \{\mu_1(\ell) - \mu_0(\ell)\} \; F_L(d\ell)$$

Specify a DDP+GP model for $[Y_i \mid A_i, L_i, \theta]$, which implies that

$$\mu_a(\ell_i) = \sum_{j=1}^{\infty} w_j \, g_j(a, \ell_i).$$

DDP+GP for Causal Inference (cont'd)

- define $g = (g_j(a_i, \ell_i) : i = 1, \ldots, n)$
- sample $\widetilde{g} = (g_j(1 - a_i, \ell_i) : i = 1, \ldots, n)$ from its conditional distribution given g
- Denote the mth draw of $\mu_a(\ell_i)$ by $\mu_a^{(m)}(\ell_i)$

If use the Bayesian bootstrap for $p(\ell)$, at each MCMC step we obtain $\omega^{(m)} \sim (1, \ldots, 1)$.

Then compute $\Delta^{(m)}$ as

$$\Delta^{(m)} = \sum_{i=1}^{n} \omega^{(m)} \{\mu_1^{(m)}(\ell_i) - \mu_0^{(m)}(\ell_i)\}.$$

Causal inference

If we have a flexible model for $p(y|a, I)$, then we can use it to compute the causal effects of interest.

- DDP-GP is one approach to modeling $p(y|a, I)$
- Avoids making parametric modeling assumptions
- The DDP-GP combination is computationally friendly because of properties of multivariate normals

Summary

- Gaussian process models are priors for functions
- Dependent Dirichlet process priors are priors for conditional distributions
- DDP-GP can be used for full nonparametric modeling of conditional distributions
- Useful for causal inference, because we often need distribution of outcome given treatment and confounders; and for mediation, $Y|A, M, L$ and $M|A, L$

CS 1: EDP Mixtures and EHR Data

June 26, 2022

Study background

Adults living with HIV infection are coinfected with chronic Hepatitis C virus (HCV) in 10-30% of cases.

Antiretroviral therapy (ART) has been shown to help stop progression of HIV disease and death.

It has also been shown to slow progression of HCV-associated liver fibrosis.

As a result, current guidelines suggest initiating ART for all HIV/HCV coinfected patients, regardless of CD4 cell count.

Study background

Nucleoside reverse transcriptase inhibitors (NRTIs) are a class of antiretrovirals used to treat HIV infection.

Combinations of drugs usually include at least 3 drugs from at least 2 different drug classes.

Certain NRTIs associated with mitochrondrial toxicity (mtNRTIs).

- These include didanosine, stavudine, zalcitabine, zidovudine.

Hypothesis: use of these mtNRTIs in a HAART regimen increase the risk of death in HIV/HCV patients compared to patients on a HAART regimen including other NRTIs

Design and data

Data from Veteran's Aging Cohort Study (VACS), 2002-2009

- Treatment naive and HIV/HCV coinfected
- Initiating HAART regimen with NRTI
- $n = 1747$
- exposure: mtNRTI versus other NRTI
- outcome: death within 2 years of starting ART (165 total events)
- many confounders: age, race, BMI, CD4, viral load, AST, ALT, fib4, etc.
- Some laboratory variables had missing values and needed to be imputed

Year of Study Entry

1996	1997	1998	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011
149	865	550	505	426	337	249	232	163	89	42	23	24	12	9	2
2	4	9	12	7	5	15	37	63	145	158	161	180	151	163	147

Potential outcomes and causal effects

$Y(a)$: indicator died within 2 years if $A = a$

Causal effect of interest

$$\psi_{rr} = \frac{E\{Y(1)\}}{E\{Y(0)\}}$$

Causal assumptions

Consistency: $Y(a) = Y$ among subjects with $A = a$

Positivity: $P(A = a|L) > 0$ if $p(L) > 0$

Ignorability: $Y(a) \perp\!\!\!\perp A|L$

EDPM

Let $X_i = (A_i, L_i)$

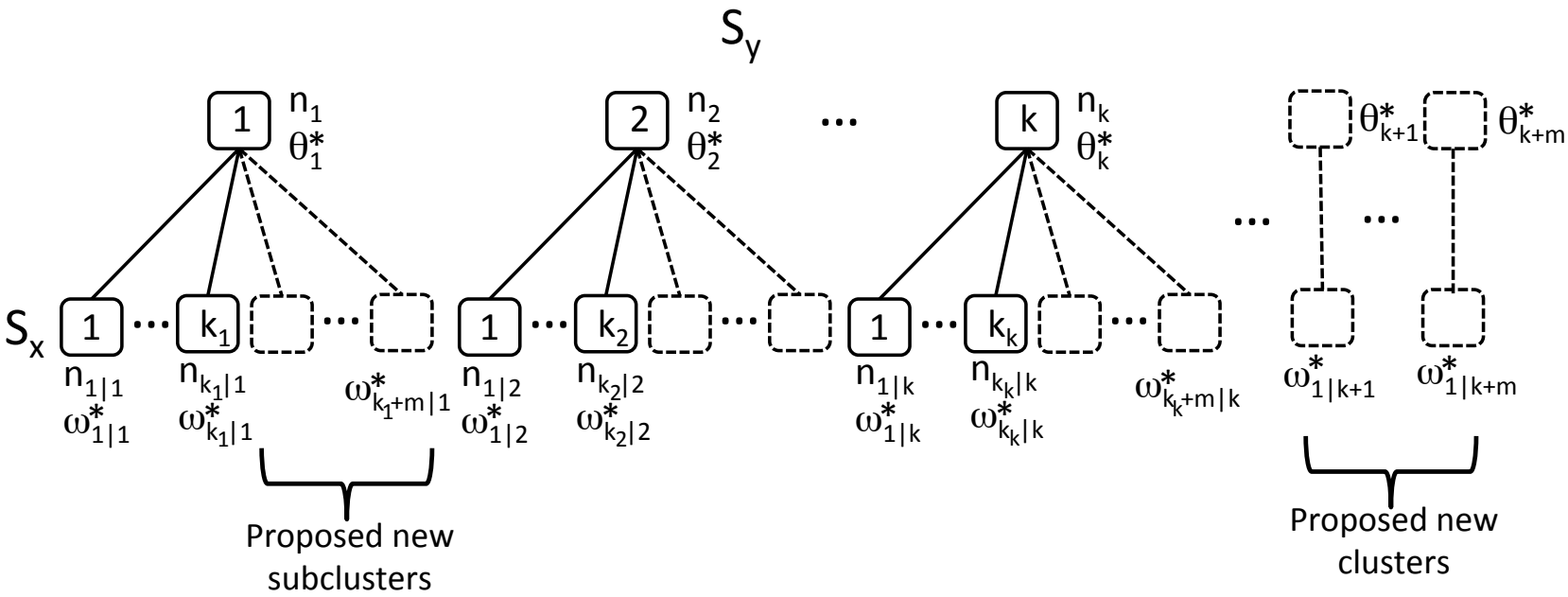
$$
\begin{aligned}
Y_i | X_i, \beta_i &\sim \text{Bern}\{\text{logit}^{-1}(X_i \beta_i)\} \\
X_{i,r} | \pi_i^r &\sim \text{Bern}(\pi_r^r), \ r = 1, \ldots, p_1 \\
X_{i,r} | \mu_i^r, \tau_i^{2,r} &\sim N(\mu_i^r, \tau_i^{2,r}), \ r = p_1 + 1, \cdots, p_1 + p_2 \\
(\beta_i, \pi_i, \mu_i, \tau_i^2, \sigma_i^2) &\sim P \\
P &\sim EDP(\alpha_\theta, \alpha_\omega, P_0)
\end{aligned}
$$

Priors

$$
\begin{aligned}
p_{0\theta}(\beta) &= N(\beta_0, c\Sigma^0_\beta) \\
p_{0\theta}(\sigma^2) &= \text{Scale Inv} - \chi^2(1,1) \\
p_{0\omega}(\pi^r) &= \text{Beta}(1,1) \\
p_{0\omega}(\tau^{2,r}) &= \text{Scale Inv} - \chi^2(2,1) \\
p_{0\omega}(\mu^r) &= N(0, 2\tau^{2,r}) \\
p(\alpha_\theta) &\sim \text{Gam}(1,1) \\
p(\alpha_\omega) &\sim \text{Gam}(1,1)
\end{aligned}
$$

We set β_0 and Σ^0_β to the MLEs from an ordinary logistic regression of Y on X and set $c = 300 \approx n/5$

Recall EDP structure



MCMC algorithm

Steps:

1. update cluster membership
2. update parameters, given clusters
3. impute missing covariates, conditional on cluster membership and parameters
4. repeat above steps many times

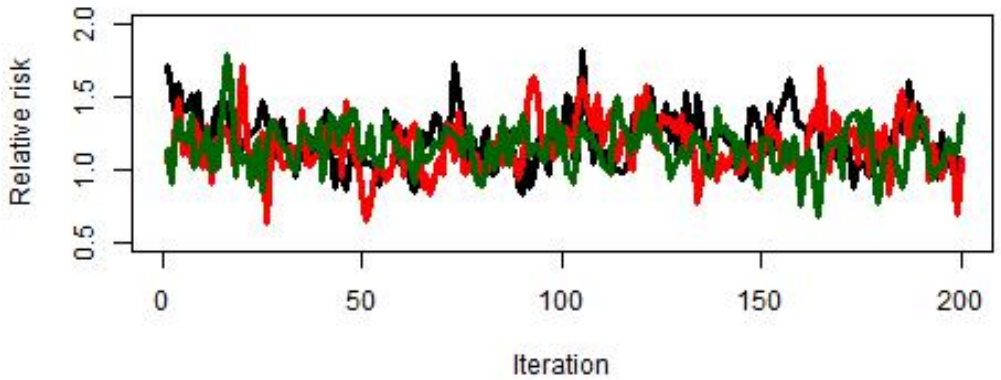
EDPM Analysis

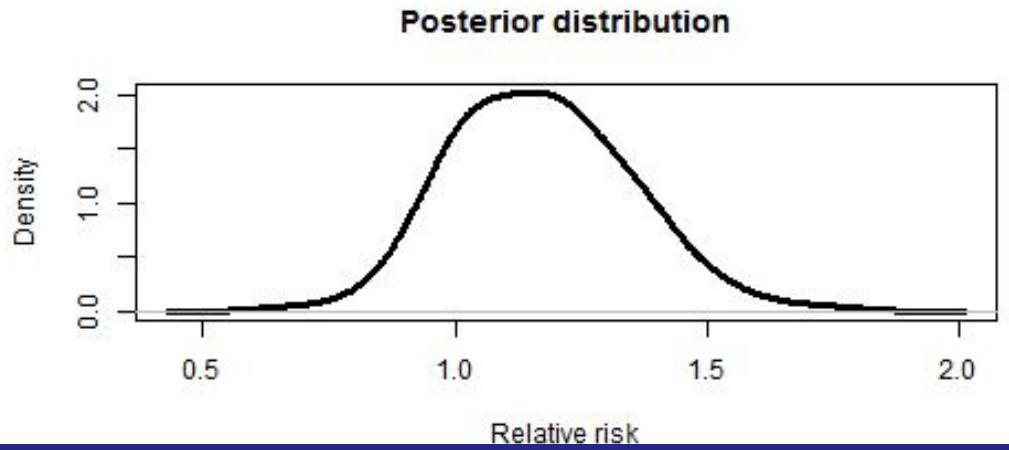
At last iteration of first chain:

$k = 5$

- $s_y = 1$: $(36, 164, 134, 45, 32, 38, 76, 1)$
- $s_y = 2$: $(171, 211, 131, 68, 18, 1)$
- $s_y = 3$: $(171, 281, 172, 50, 28)$
- $s_y = 4$: $(137, 30, 2, 1)$
- $s_y = 5$: (2)

Results


Relative risk
0.5
1.0
1.5
2.0
0
50
100
150
200
Iteration


Posterior distribution
Density
0.0
1.0
2.0
0.5
1.0
1.5
2.0
Relative risk

Data Analysis Comparisons

We also applied IPTW and TMLE methods to the data. To be able to do that, we first needed to deal with missing covariates:

- multiple imputation using predictive mean matching
- implement IPTW and TMLE to each, then combine with Rubin rules

Results:

Method	Est (LCL, UCL)
BNP	1.16 (0.87,1.54)
IPTW	1.02 (0.97, 1.08)
TMLE	1.22 (1.06, 1.47)

Summary

- Problem: causal inference with missing confounders
- Use EDPM to model joint distribution of observed data
- Impute covariates (under ignorability/MAR) within the MCMC
- Applied method to HIV study - used 4-6 y-clusters and additional x-subclusters
  - did not collapse to parametric logistic regression

CS 2: Mediation Analysis with Decision Tree Ensembles

Mediation Analysis

Definition (McKinnon)

A mediator is a variable that is intermediate in the causal process relating an independent to a dependent variable.

Mediation Analysis

Definition (McKinnon)

A mediator is a variable that is intermediate in the causal process relating an independent to a dependent variable.

1. Cognitive Behavioral Therapy $\Longrightarrow$ Exercise $\Longrightarrow$ Lower Depression.

Mediation Analysis

Definition (McKinnon)

A mediator is a variable that is intermediate in the causal process relating an independent to a dependent variable.

1. Cognitive Behavioral Therapy $\Longrightarrow$ Exercise $\Longrightarrow$ Lower Depression.
2. Tobacco use $\Longrightarrow$ Poor Health $\Longrightarrow$ High Medical Expenditures.

Mediation Analysis

Definition (McKinnon)

A mediator is a variable that is intermediate in the causal process relating an independent to a dependent variable.

1. Cognitive Behavioral Therapy $\Longrightarrow$ Exercise $\Longrightarrow$ Lower Depression.
2. Tobacco use $\Longrightarrow$ Poor Health $\Longrightarrow$ High Medical Expenditures.
3. Governmental policies vs. COVID $\Longrightarrow$ societal norms regarding social distancing and masking $\Longrightarrow$ lower COVID rates.

Mediation Analysis

Definition (McKinnon)

A mediator is a variable that is intermediate in the causal process relating an independent to a dependent variable.

1. Cognitive Behavioral Therapy $\Longrightarrow$ Exercise $\Longrightarrow$ Lower Depression.
2. Tobacco use $\Longrightarrow$ Poor Health $\Longrightarrow$ High Medical Expenditures.
3. Governmental policies vs. COVID $\Longrightarrow$ societal norms regarding social distancing and masking $\Longrightarrow$ lower COVID rates.

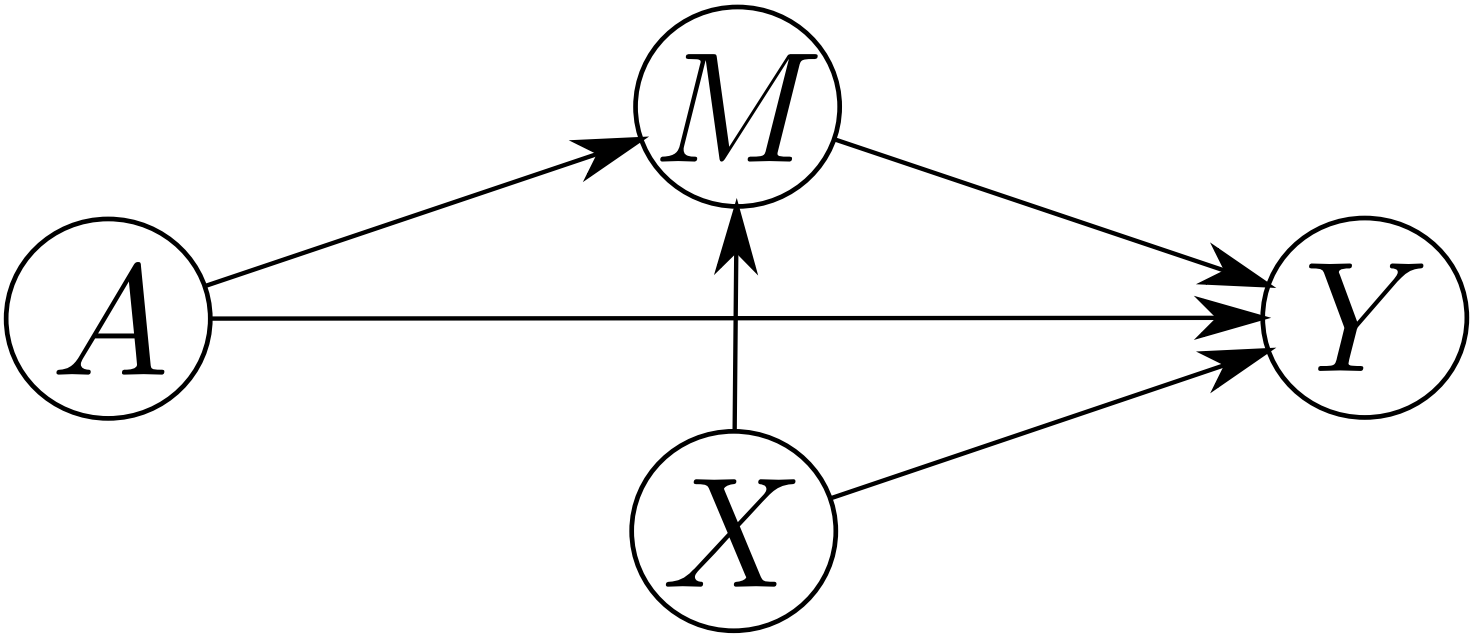
Many examples across economics, epidemiology, social sciences, etc. Many questions of central importance concern mediating processes.

Running Example

Question: does smoking increase medical expenditures?

- We make use of public data from the Medical Expenditure Panel Survey (MEPS).
- We expect an indirect effect: smoking $\Longrightarrow$ bad health $\Longrightarrow$ medical costs.
- But maybe smokers are less likely to incur medical expenditures for other reasons?

A Graphical Perspective



(In non-randomized studies, add an edge between X and A.)

Potential Outcomes for Mediation

- A_i = the exposure of interest which is observed for unit i (assumed binary)
- $M_i(a)$ = the value of the mediator if unit i receives the exposure
- $Y_i(a, m)$ = potential outcome of outcome under exposure a and mediator fixed at m
- X_i = a collection of *confounders* and *effect moderators* of interest

Potential Outcomes for Mediation

- A_i = the exposure of interest which is observed for unit i (assumed binary)
- $M_i(a)$ = the value of the mediator if unit i receives the exposure
- $Y_i(a, m)$ = potential outcome of outcome under exposure a and mediator fixed at m
- X_i = a collection of *confounders* and *effect moderators* of interest

Definition

The NDE is $\zeta(a) = E_\theta(Y_i(1, M_i(a)) - Y_i(0, M_i(a)))$.
The NIE is $\delta(a) = E_\theta(Y_i(a, M_i(0)) - Y_i(a, M_i(1)))$.
The ATE of the treatment is $\tau = E_\theta(Y_i(1, M_i(1)) - Y_i(0, M_i(0)))$; decomposes as $\tau = \delta(a) + \zeta(1 - a)$.

Causal Assumptions

Analogous to ignorability in observational studies, we invoke sequential ignorability (SI):

1. Ignorability of the assignment mechanism:

$$[A_i \perp (Y_i(a, m)M_i(a')) \mid X_i = x]$$

for all (x, a, a') (i.e., no unmeasured confounding between exposure and potential outcomes/mediators)

2. Ignorability of the mediator process:

$$[Y_i(a, m) \perp M_i(a') \mid X_i = x, A_i = a', X_i = x]$$

for all (x, a, a') (i.e., no unmeasured confounding between potential outcome and mediator)

3. Positivity: $\Pr_\theta(A_i = a \mid X_i = x) > 0$ and $f_\theta(M_i(a) = m \mid A_i = a, X_i = x) > 0$ for all (a, x, m).

Priors on the Degree of Confounding

Consider the simplified ridge model

$$Y(a, m)_i = X_i^\top \beta_x + a\,\beta_a + m\,\beta_m + \sigma_y\,\epsilon_i,$$
$$M_i(a) = X_i^\top \alpha_x + a\,\alpha_a + \sigma_m\,\nu_i,$$
$$X_i \sim \text{Normal}(0, \Sigma).$$

Direct and indirect effects under this model can be shown to be

$$\zeta = \beta_a, \quad \text{and} \quad \delta = \alpha_a\,\beta_m.$$

Priors on the Degree of Confounding

Consider the simplified ridge model

$$Y(a,m)_i = X_i^\top \beta_x + a\,\beta_a + m\,\beta_m + \sigma_y\,\epsilon_i,$$
$$M_i(a) = X_i^\top \alpha_x + a\,\alpha_a + \sigma_m\,\nu_i,$$
$$X_i \sim \text{Normal}(0, \Sigma).$$

Direct and indirect effects under this model can be shown to be

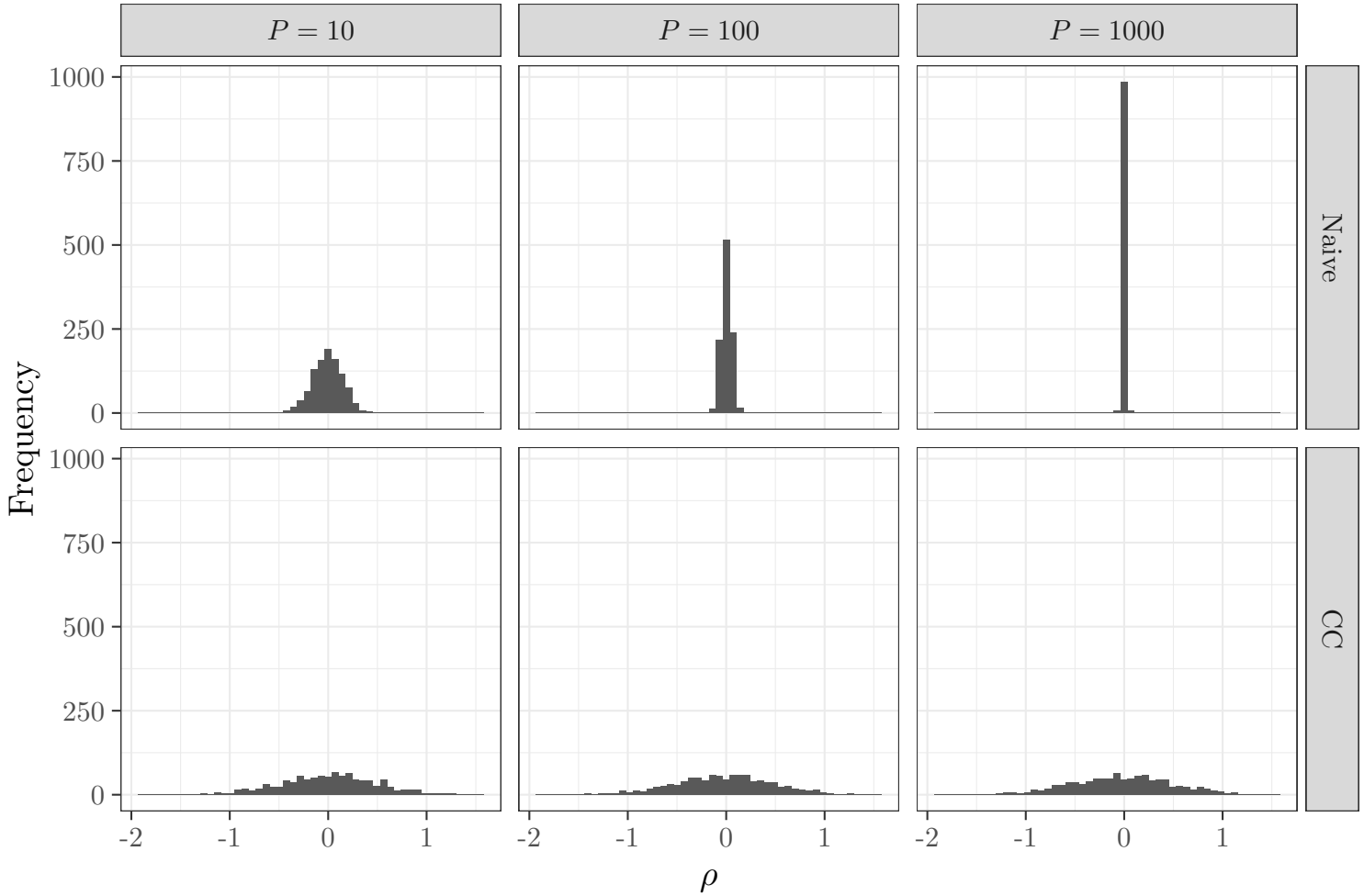
$$\zeta = \beta_a, \quad \text{and} \quad \delta = \alpha_a\,\beta_m.$$

Can be shown that if we ignore confounders and fit a linear model, we will instead be estimating

$$\zeta^\star = \beta_a - \rho\,\alpha_a, \quad \text{and} \quad \delta^\star = \alpha_a\,\beta_m + \rho\,\alpha_a,$$

where $\rho = \frac{\beta_x^\top \Sigma \alpha_x}{\sigma_m^2 + \alpha_x^\top \Sigma \alpha_x}$. If we believe ρ is zero, it means we believe that confounding is irrelevant.

Samples from the Ridge Prior



Clever Covariates

In the case of the high-dimensional ridge model, we can correct this issue by setting

$$\beta_x \sim \text{Normal}(\psi\,\alpha_x, \sigma^2_\beta)$$

instead. This is equivalent to including $X_i^\top \alpha_x$ as a *covariate* in the outcome model.

Clever Covariates

In the case of the high-dimensional ridge model, we can correct this issue by setting

$$\beta_x \sim \text{Normal}(\psi\,\alpha_x, \sigma^2_\beta)$$

instead. This is equivalent to including $X_i^\top \alpha_x$ as a *covariate* in the outcome model.

Extending this intuition to BART we include estimates of $E(M_i \mid A_i = 0, X_i)$ and $E(M_i \mid A_i = 1, X_i)$ as covariates in the outcome BART model. Additionally, if A_i is not randomized, then we also include $E(A_i \mid X_i = x)$ as a covariate.

A Bayesian Causal Mediation Forest

Use the observed data models:

- $Y_i \sim \text{Normal}\{\mu_y(M_i, A_i, X_i), \sigma_y^2\}$
- $M_i \sim \text{Normal}\{\mu_m(A_i, X_i), \sigma_m^2\}$
- $A_i \sim \text{Bernoulli}[\Phi\{\mu_a(X_i)\}]$

A Bayesian Causal Mediation Forest

Use the observed data models:

- $Y_i \sim \text{Normal}\{\mu_y(M_i, A_i, X_i), \sigma_y^2\}$
- $M_i \sim \text{Normal}\{\mu_m(A_i, X_i), \sigma_m^2\}$
- $A_i \sim \text{Bernoulli}[\Phi\{\mu_a(X_i)\}]$

Three modifications:

1. We stratify by treatment: two separate BART models are fit for both the outcome and mediator (four in total).
2. We include an estimate of the smoking propensity as a covariate in the mediator and outcome models.
3. We include an estimate of $\mu_m(0, X_i)$ and $\mu_m(1, X_i)$ as covariates in $\mu_y(m, a, X_i)$.

Computing the Causal Parameters

We can compute

$$E_\theta(Y_i(a, M_i(a'))) = \int y\, f_\theta(Y_i = y \mid M_i = m, a, x)\, f_\theta(M_i = m \mid a', x)\, f_\theta(x)\; dm\; dx.$$
$$= \int r_y(m, a, x)\, f_\theta(M_i = m \mid a', x)\, f_\theta(x)\; dm\; dx.$$

Two issues:

- Two decision tree ensembles in this $\Longrightarrow$ integral is intractable.
- We need a model for $f_\theta(x)$.

Bayesian Bootstrap

Recall the Bayesian bootstrap takes $X_i \sim \sum_{j=1}^{n} \varpi_j \, \delta_{x_j}$ where $(x_1, \ldots, x_N)$ are the realized samples of the covariates. We then use the improper prior

$$\pi(\varpi_1, \ldots, \varpi_N) = \prod_{j=1}^{N} \varpi_j^{-1}.$$

Posterior distribution is $\varpi \sim \text{Dirichlet}(1, \ldots, 1)$.

Bayesian Bootstrap

Recall the Bayesian bootstrap takes $X_i \sim \sum_{j=1}^{n} \varpi_j \, \delta_{x_j}$ where $(x_1, \ldots, x_N)$ are the realized samples of the covariates. We then use the improper prior

$$\pi(\varpi_1, \ldots, \varpi_N) = \prod_{j=1}^{N} \varpi_j^{-1}.$$

Posterior distribution is $\varpi \sim \text{Dirichlet}(1, \ldots, 1)$.
Under BB:

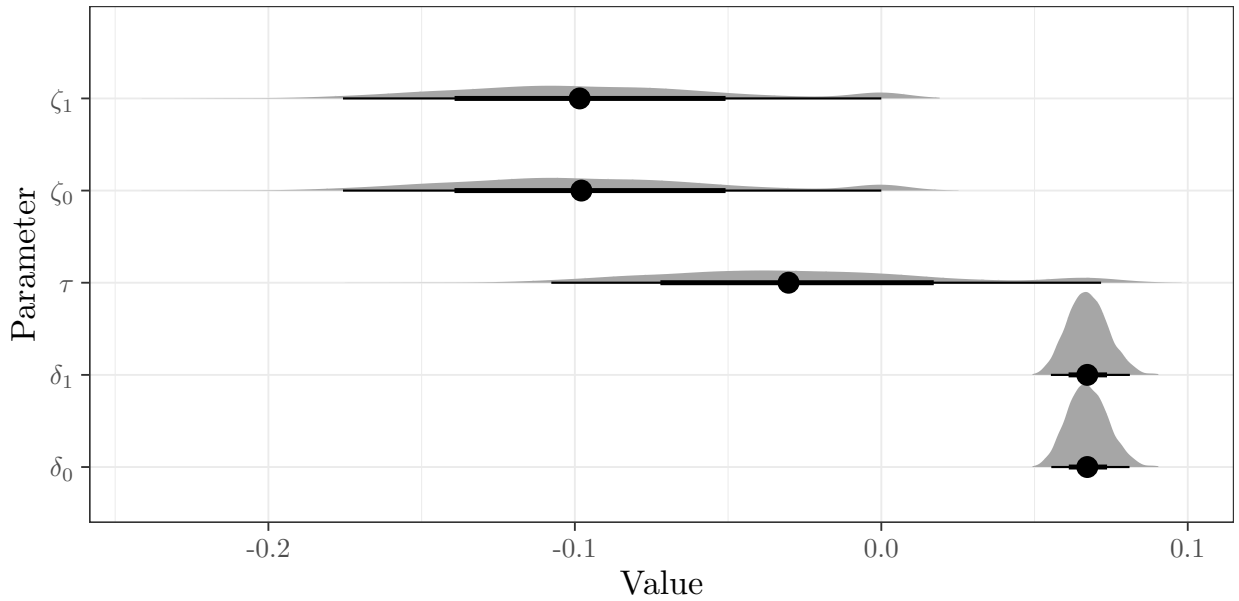
$$E_\theta(Y_i(a, M_i(a'))) = \sum_{j=1}^{U} \varpi_j \int r_y(m, a, x_j) \, f_\theta(M_i = m \mid a', x) \, dm$$

At this point, we can either (i) compute the integral numerically or (ii) via Monte Carlo integration. Linero (2021) shows how the latter can be done extremely efficiently.

Implement a Bayesian Causal Mediation Forest

Go over code.

Results from BCMF



Answer makes a lot more sense than the non-mediated version!

Sensitivity Analysis

What if SI is wrong? It is untestable!

Sensitivity Analysis

What if SI is wrong? It is untestable!
Our approach: introduce an unidentified *sensitivity parameter* that indexes how severely SI fails. L. and Zhang (2022) take

$$Y(a,m)_i \sim \text{Normal}\{r_y(m,a,x) + \lambda(M_i(a) - m), \sigma_y^2\}.$$

We can interpret λ as taking λ units of "explanation" away from the causal effect of mediation and shifting it into statistical association with $M_i(a)$.

Sensitivity Analysis

What if SI is wrong? It is untestable!
Our approach: introduce an unidentified *sensitivity parameter* that indexes how severely SI fails. L. and Zhang (2022) take

$$Y(a,m)_i \sim \text{Normal}\{r_y(m,a,x) + \lambda(M_i(a) - m), \sigma_y^2\}.$$

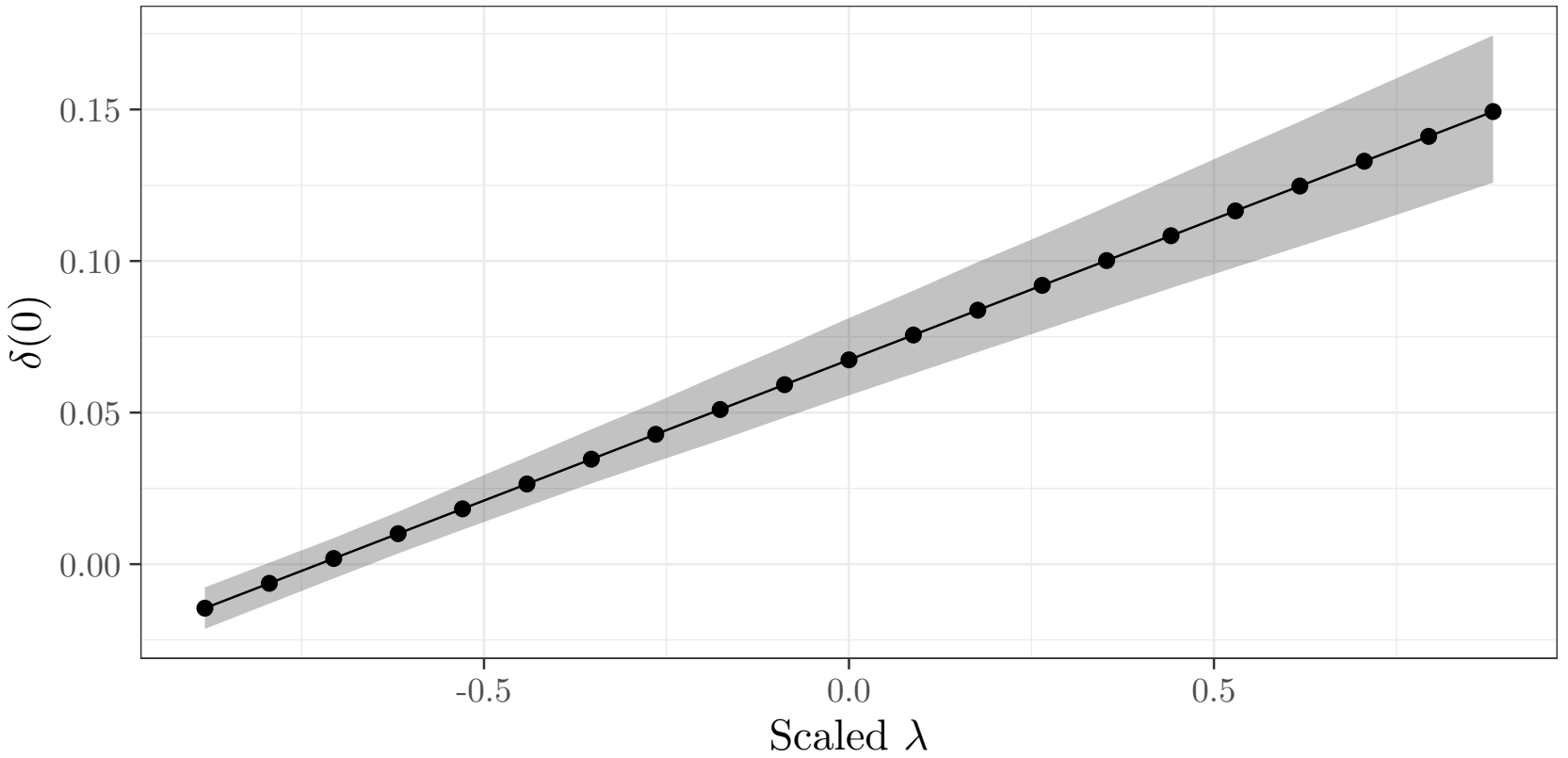
We can interpret λ as taking λ units of "explanation" away from the causal effect of mediation and shifting it into statistical association with $M_i(a)$.
Modifications to estimates are convenient!

$$\delta_\lambda(a) = \delta_0(a) - \lambda \sum_i \varpi_i \{r_m(1,X_i) - r_m(0,X_i)\},$$

$$\zeta_\lambda(a) = \zeta_0(a) + \lambda_i \sum_i \varpi_i \{r_m(1,X_i) - r_m(0,X_i)\}.$$

Results

To calibrate λ, I divided it by an OLS estimate of the regression coefficient of Y_i on M_i (conditional on confounders). Roughly, what % of apparent causal relationship between Y_i and M_i needs to be non-causal for us to get a different conclusion?



CS 3: Semi-competing risks and DDP

June 26, 2022

Semi-competing risks

- Semi-competing risks occur in studies where observation of a nonterminal event (e.g., progression) may be pre-empted by a terminal event (e.g., death), but not vice versa.
- In randomized clinical trials to evaluate treatments of life-threatening diseases, patients are often observed for specific types of disease progression and survival.
- Often, the primary outcome is patient survival, resulting in data analyses focusing on the terminal event using standard survival analysis tools
- However, there may also be interest in understanding the effect of treatment on nonterminal outcomes such as progression or readmission

Application

- randomized trial for the treatment of malignant brain tumors
  - one of the important progression endpoints is based on deterioration of the cerebellum
  - biologically plausible that a patient could die without cerebellar deterioration
  - thus, analyzing the effect of treatment on progression needs to account for the fact that progression is not well-defined after death.

Notation

- $z = 0, 1$ represents control and treatment group
- Y_P^z: progression time under treatment z.
- Y_D^z: death time under treatment z.
- C^z: censoring time under treatment z.
- Fundamental to our setting is that $Y_P^z \not> Y_D^z$ (i.e., progression cannot happen after death).

Causal estimand

The causal estimand of interest:

$$\tau(u) = \frac{Pr[Y_P^1 < u \mid Y_D^0 \geq u, Y_D^1 \geq u]}{Pr[Y_P^0 < u \mid Y_D^0 \geq u, Y_D^1 \geq u]},$$

where $\tau(\cdot)$ is a smooth function of u.

- Among patients who survive to time u under both treatments, this estimand contrasts the risk of progression prior to time u for treatment 1 relative to treatment 0.
- example of a principal stratum causal effect

Observed data

- Z denote treatment assignment
- $\boldsymbol{X}$ denote a vector of the baseline covariates.
- the observed event times and event indicators.
  - $Y_P = Y_P^Z$, $Y_D = Y_D^Z$ and $C = C^Z$.
  - $T_1 = Y_P \wedge Y_D \wedge C$,
  - $\delta = I(Y_P < Y_D \wedge C)$,
  - $T_2 = Y_D \wedge C$,
  - $\xi = I(Y_D < C)$
- The observed data for each patient are $\boldsymbol{O} = (T_1, T_2, \delta, \xi, Z, \boldsymbol{X})$.

Assumption 1

Assumption 1: Treatment is randomized, i.e.,

$$Z \perp (Y_P^z, Y_D^z, C^z, \boldsymbol{X}); \quad z = 0, 1,$$

and $0 < Pr[Z = 1] < 1$.
This holds by design in randomized trials as considered here.

Assumption 2

Assumption 2: Censoring is non-informative in the sense that

$$C^z \perp (Y_P^z, Y_D^z) \,|\, \boldsymbol{X} = \boldsymbol{x}; \quad z = 0, 1,$$

and $Pr[C^z > Y_P^z, C^z > Y_D^z | \boldsymbol{X} = \boldsymbol{x}] > 0$ for all $\boldsymbol{x}$.

Identification Results 1

- Let $\lambda^z_{\boldsymbol{X}}$ denote the conditional hazard function of Y^z_D given $\boldsymbol{X} = \boldsymbol{x}$
- Let $G^z_{\boldsymbol{X}}$ denote the conditional distribution function of Y^z_D given $\boldsymbol{X} = \boldsymbol{x}$
- Under Assumptions 1 and 2, $\lambda^z_{\boldsymbol{X}}$ and $G^z_{\boldsymbol{X}}$ are identified
- this is standard identification for survival data

Identification Results 2

- The conditional sub-distribution function of Y_P^z given Y_D^z and $\boldsymbol{X} = \boldsymbol{x}$, $V_{\boldsymbol{X}}^z$, is

$$V_{\boldsymbol{X}}^z(s|t) \quad = \quad Pr[T_1 \leq s, \delta = 1 \,|\, T_2 = t, \xi = 1, \boldsymbol{X} = \boldsymbol{x}, Z = z],$$

  where $s \leq t$.
- this sub-distribution function is also identified from Assumptions 1 and 2
  - Together $G_{\boldsymbol{X}}^z(t)$ and $V_{\boldsymbol{X}}^z(s|t)$ identify the joint subdistribution $V_{\boldsymbol{X}}^z(s,t)$ for (Y_P^z, Y_D^z) given $\boldsymbol{X} = \boldsymbol{x}$.

Assumption 3

Assumption 3: The conditional joint distribution function of (Y_D^0, Y_D^1) given $\boldsymbol{X} = \boldsymbol{x}$, $G_{\boldsymbol{X}}$, follows a Gaussian copula model, i.e.,

$$G_{\boldsymbol{X}}(v, w; \rho) = \Phi_{2,\rho}[\Phi^{-1}\{G_{\boldsymbol{X}}^0(v)\}, \Phi^{-1}\{G_{\boldsymbol{X}}^1(w)\}],$$

where Φ is a standard normal c.d.f. and $\Phi_{2,\rho}$ is a bivariate normal c.d.f. with mean 0, marginal variances 1, and correlation ρ.

- for fixed ρ, $G_{\boldsymbol{X}}$ is identified since $G_{\boldsymbol{X}}^0$ and $G_{\boldsymbol{X}}^1$ are identified
- ρ will be a sensitivity parameter here - $\rho = 0$, independence; $\rho = 1$, rank preserving assumption
- similar assumptions have been used in the causal mediation literature

Assumption 4

Assumption 4: Progression time under treatment z is conditionally independent of death time under treatment $1-z$ given death time under treatment z and covariates $\boldsymbol{X} = \boldsymbol{x}$, i.e.,

$$Y_P^z \perp Y_D^{1-z} \,\big|\, Y_D^z, \boldsymbol{X} = \boldsymbol{x}; \quad z = 0, 1.$$

Final identification Result

Lemma: Under Assumptions 1-4, the principal stratum causal effect, $\tau(\cdot)$ is identified from the distribution of the observed data

BNP model for the observed data distribution I

- need a model for the observed data, $\boldsymbol{O} = (T_1, T_2, \delta, \xi, Z, \boldsymbol{X})$.
- use a Dependent Dirichlet Process-Gaussian process (DDP-GP) for the *conditional* distribution of $\boldsymbol{V} = (Y_p, Y_D)$ given X
- specify independent DDP-GP for each treatment group z
- the prior induces priors on non-identified (ill-defined) quantities (i.e., progression after death), but these have no impact on our analysis.

Brain Cancer Data example I

- randomized (placebo-controlled) phase II trial (Brem et al, 1995)
- 222 recurrent gliomas patients, who were scheduled for tumor resection
- The data includes 11 baseline prognostic measures and a baseline evaluation of cerebellar function.

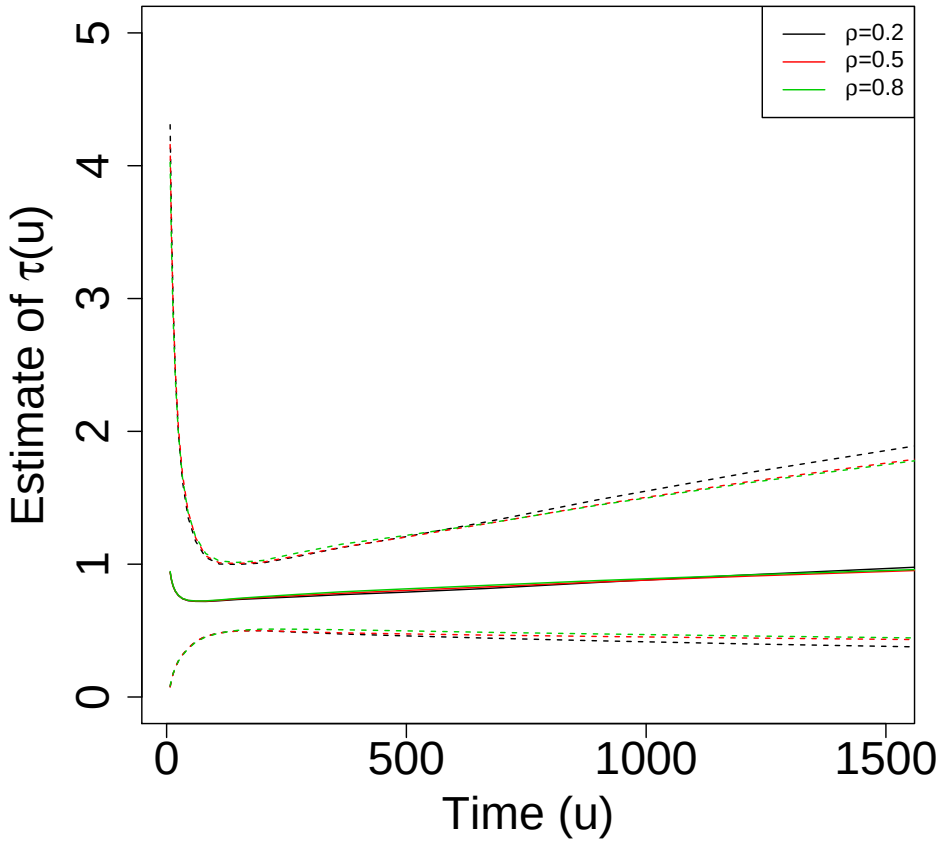
Brain Cancer Data example II

- Patient were randomized to receive surgically implanted biodegradable polymer discs *with or without* 3.85% of carmustine.
- The follow-up duration was 1 year.
- Of the 219 patients with complete baseline measures
  - 204 were observed to die
  - 100 were observed to progress prior to death
  - Of the 15 patients who did not die, 4 were observed to have cerebellar progression.
- **Goal:** estimate the causal effect of treatment on time to cerebellar progression.

Causal inference results I

- posterior inference for the causal estimand, $\tau(u)$.
- sensitivity parameter, ρ
  - fix ρ at 0.2, 0.5, and 0.8.
  - prior $\rho \sim \text{Beta}(0.1875, 0.0625)$ [mean and variance, 0.75 and 0.15]

Causal inference results II



Conclusions for semi-competing risk

- proposed a Bayesian approach for causal inference in setting of semi-competing risks
  - BNP for the observed data distribution
  - an interpretable causal estimand
  - one of uncheckable assumptions parameterized by a sensitivity parameter
- open issues
  - how to best determine values of the sensitivity parameter
  - weaken/remove/sensitivity Assumption 4
  - alternative BNP for observed data